OODML: Whole Slide Image Classification Meets Online Pseudo-Supervision and Dynamic Mutual Learning

Tingting Zheng[1], Kui Jiang[1], Hongxun Yao[1]*, Yi Xiao[2], Zhongyuan Wang[2]

[1]Harbin Institute of Technology
[2]Wuhan University
23b903051@stu.hit.edu.cn,{jiangkui, h.yao}@hit.edu.cn, xiao_yi@whu.edu.cn, wzy_hope@163.com

Abstract

Bag-label-based multi-instance learning (MIL) has demonstrated significant performance in whole slide image (WSI) analysis, particularly in pseudo-label-based learning schemes. However, due to inaccurate feature representation and interference, existing MIL methods often yield unreliable pseudo-labels, which spawn undesired predictions. To address these issues, we propose an Online Pseudo-Supervision and Dynamic Mutual Learning (OODML) framework that enhances pseudo-label generation and feature representation while exploring their mutual learning to improve bag-level prediction. Specifically, we design an Adaptive Memory Bank (AMB) to collect the most informative components of the current WSI. We also introduce a Self-Progressive Feature Fusion (SPFF) module that integrates label-related historical information from the AMB with current semantic variations, thereby enhancing the representation of pseudo-bag tokens. Furthermore, we propose a Decision Revision Pseudo-Label (DRPL) generation scheme to explore intrinsic connections between pseudo-bag representations and bag-label predictions, resulting in more reliable pseudo-label generation. To alleviate redundant and ambiguous representations, the class-wise prior of pseudo-label prediction is borrowed to facilitate label-related feature learning and to update the AMB, forming a mutual refinement between feature representation and pseudo-label generation. Additionally, a Dynamic Decision-Making (DDM) module is developed to harmonize explicit and implicit representations of bag information for more robust decision-making. Extensive experiments on four datasets demonstrate that our OODML surpasses the state-of-the-art by 3.3% and 6.9% on the CAMELYON16 and TCGA Lung datasets.

1 Introduction

Histopathology slides provide complex cellular structures and disease properties are pivotal for cancer diagnosis (Bejnordi et al. 2017). However, examining slides using light microscopy in practical diagnostics is a tedious task for pathologists while suffering from low consistency and reproducibility (Campanella et al. 2019). Recently, the advent of digital pathology to scan slides as whole slide images (WSIs) has revolutionized computational pathology, making it easier for researchers and clinicians to analyze WSIs (Brancati et al. 2022; Tomczak, Czerwińska, and Wiznerowicz 2015). Nevertheless, due to the gigapixel-sized and diverse microstructures of WSIs, conventional hand-crafted techniques fail to generate satisfied predictions (Van der Laak, Litjens, and Ciompi 2021). By contrast, deep learning (DL) methods have attracted widespread attention due to their excellent capability in exploring contextual and textural information of WSIs (Lu et al. 2021; Shao et al. 2021; Zhang et al. 2022). While showing considerable superiority over traditional algorithms, collecting fine-grained annotations from irregular tumor tissues and high-resolution WSIs is time-consuming and laborious (Bin 2021; Zheng et al. 2023).

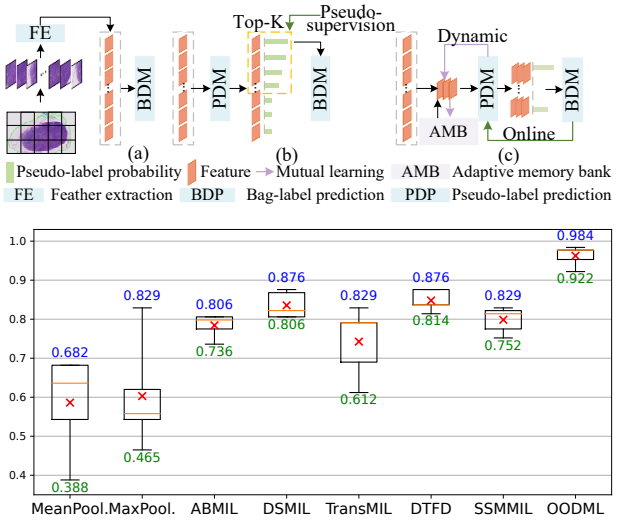

Figure 1: Comparative accuracy of bag-label-level approaches ((a): MeanPooling, ABMIL, DSMIL, TransMIL, SSMMIL), offline pseudo-label-level approaches ((b): MaxPooling, DTFD), and our online pseudo-label-level OODML (c), shown in box plots on the CAMELYON16 dataset with 5-fold cross-validation. Blue/green denote the max/min values. Our OODML demonstrates impressive performance by promoting mutual learning between feature representation and pseudo-label generation.

*Corresponding authors.

To tackle this issue, some researchers explore weakly supervised technologies for WSI analysis. In particular, the multi-instance learning (MIL) scheme (Van der Laak, Litjens, and Ciompi 2021; Ilse, Tomczak, and Welling 2018) considers each WSI as a "bag", and then learns the aggregated representation of the sampled patches (instances) from WSI. According to introducing pseudo-labels, previous MIL-based approaches can be roughly divided into two categories: bag-label-level and pseudo-label-level (Zhang et al. 2022), as illustrated in Figure 1.

The former typically depicts a high-level bag representation from all instances for bag label prediction (Shao et al. 2021; Lin et al. 2023). However, it is non-trivial to aggregate target-related components from numerous and redundant instances, especially in the context of sparse tumor regions within WSIs. Furthermore, relying solely on bag-label constraints inevitably results in the loss of fine-grained pathological structures and semantics in the bag representations (Lin et al. 2024). These limitations in feature representation significantly reduce the precision of the final predictions (Chen, Sun, and Zhao 2024).

To promote feature representation, the latter (Lu et al. 2021; Zheng, Jiang, and Yao 2024) derives merits from auxiliary priors, like bag-label, attention scores, clustering similarity, and teacher models (Qu et al. 2022a; Yu et al. 2023) to generate pseudo-labels. While pseudo-label supervision facilitates model training and performance, the offline approach for pseudo-label generation (Figure 1 (b)), directly using fixed bag labels as instance or pseudo-bag labels, may lead to unpredictable issues. Meanwhile, the unreliable pseudo-label may cause error accumulation, consequently declining model performance (Wang et al. 2023).

Overall, the aforementioned limitations in existing technologies (Qu et al. 2022b; Xiong et al. 2023; Yu et al. 2023) may originate from immature framework design and optimization schemes. *First,* these technologies commonly derive the bag label inference via single or Top-K instances of pseudo-label predictions, where the information is sparse and incomplete in characterizing the real distribution of tumor regions. *Second,* there exist significant gaps in offline pseudo-label generation using pre-trained models, since these models are dedicated to holistic bag representations rather than instance representations. Thus, these defects raise significant uncertainty in both feature representation and pseudo-label generation, leading to a performance crisis.

To promote the optimization, some efforts have been developed to co-train bag-label-level and pseudo-label-level models in separate steps, employing bag-level models to generate pseudo-labels for instance-level optimization (Qu et al. 2022b; Yu et al. 2023). However, the asynchrony between the update of feature representation and the usage of pseudo-label generation inevitably compromises pseudo-label accuracy.

By revisiting the above issues, two natural questions arise: 1) *whether can the mutual learning between pseudo-label prediction and feature representation facilitate representations compactness and accuracy for better bag decision-making;* 2) *how to correlate pseudo-bag representations and bag label predictions online to generate more reliable pseudo-labels?*

To answer these questions, we take the intrinsic relations among feature representation and pseudo-label generation into consideration, and harmonize the merits of Online Pseudo-Supervision and Dynamic Mutual Learning (OODML) for WSI classification tasks, as shown in Figure 1 (c). For discriminative and compact representation, we devise a Self-Progressive Feature Fusion (SPFF) module to fuse historical information of current WSI from an Adaptive Memory Bank (AMB) and instant semantics to enhance current pseudo-bag token representation. The philosophies behind SPFF involve: 1) the pseudo-label prediction provides class-wise insights to help AMB capture salient tokens; 2) AMB integrates Memory-based Cross-Attention (MCA) and Multi-Level Feature Fusion (MLFF) scheme to provide refined features facilitating current token; 3) the fine-grained token promotes pseudo-label prediction and updates AMB.

To generate reliable pseudo-labels, we propose a Decision Revision Pseudo-Label (DRPL) generation scheme to explore the potential relationships between pseudo-bag representations and bag-label predictions. Specifically, an Explicit-Implicit Representation Learning (EIRL) strategy is presented to investigate complex relations among pseudo-bag tokens and their contributions to bag-label inference via Linear Attention Module (LAM) and Multi-head Self-Attention (MSA) mechanism. Since the attention scores are inherently noisy, we use bag-label probabilities to further calibrate the scores for reliable pseudo-label generation. In addition, a Dynamic Decision-Making (DDM) module is developed to adaptively harmonize bag explicit-implicit representations and predictions for better decision-making. Experimental results demonstrate that our OODML significantly outperforms the current state-of-the-art BCL (Yu et al. 2023) by 3.3% on the CAMELYON16 and ACMIL (Zhang et al. 2025) by 17.3% on the BRACS datasets, respectively.

The contributions of the paper are summarized as follows:

- We propose an Online Pseudo-Supervision and Dynamic Mutual Learning (OODML) framework for WSI classification, where the progressive mutual representation between label prediction and feature representation facilitates the training and performance.
- We devise a Self-Progressive Feature Fusion (SPFF) module to aggregate informative components from an Adaptive Memory Bank (AMB) and instant semantics by fully exploiting pseudo-label prediction priors for accurate and compact feature representations.
- We pioneer a Decision Revision Pseudo-Label (DRPL) generation scheme to correlate pseudo-bag representation and bag-label prediction for more reliable pseudo-label generation. In addition, the Dynamic Decision-Making (DDM) module harmonizes explicit-implicit bag perspectives for better bag-label inference.

2 Related Work

In this section, we briefly review some related advances in WSI analysis based on MIL and pseudo-label generation as well as collaborative learning in the area of computer vision.

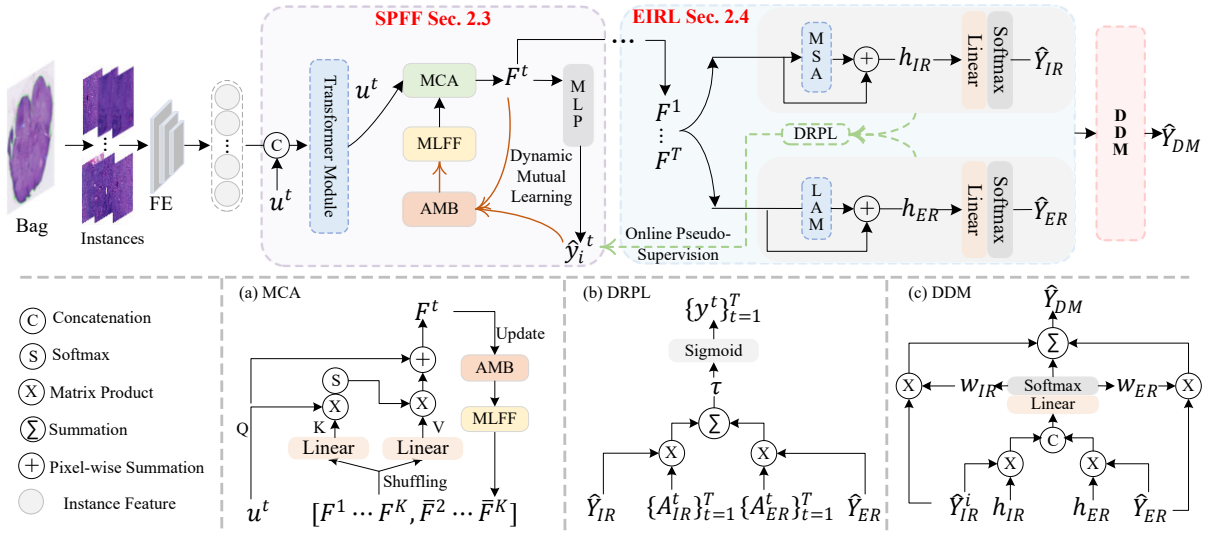

Figure 2: The architecture of our proposed Online Pseudo-Supervision and Dynamic Mutual Learning (OODML) framework. It consists of a Self-Progressive Feature Fusion (SPFF) module and an Explicit-Implicit Representation Learning (EIRL) module for feature representation, a Decision Revision Pseudo-Label (DRPL) scheme for pseudo-label generation, and a Dynamic Decision-Making (DDM) module for bag label prediction. OODML employs a Feature Extractor (FE) to embed all instances into feature vectors and randomly sample v^t at t. The SPFF takes v^t and an initial class token u^t as inputs into a Transformer module, enhancing u^t using a Memory-based Cross-Attention (MCA) mechanism to fuse label-related information from the Adaptive Memory Bank (AMB) and Multi-Level Feature Fusion (MLFF) scheme into F^t for reliable pseudo-label prediction $\hat{y}^t$ while updating the AMB. After that, the EIRL explores the explicit (h_{ER}) and implicit (h_{IR}) bag representations by employing a Linear Attention Module (LAM) and Multi-head Self-Attention (MSA), and predicts bag label probabilities $\hat{Y}_{ER}$, $\hat{Y}_{IR}$, respectively. The DRPL introduces $\hat{Y}_{ER}$, $\hat{Y}_{IR}$ and attention scores (A^t_{ER}, A^t_{IR}) to generate pseudo-labels y^t for supervision $\hat{y}^t$. Finally, the DDM takes bag representations and label probabilities as inputs for better bag decision-making $\hat{Y}_{DM}$.

2.1 Multi-instance Learning in WSI Classification

The MIL aims to predict bag labels by aggregating target-dependent features from sampling instances. As shown in Figure 1, there are two main MIL categories based on label settings: bag-label-level and pseudo-label-level methods. The former applies diverse fusion strategies, ranging from the early MeanPooling, attention to advanced Transformer and graph network (Ilse, Tomczak, and Welling 2018; Shao et al. 2021; Bin 2021; Chen, Sun, and Zhao 2024), to integrate a bag feature representation from all instances for bag label decision-making. However, due to enormous and redundant instances, models relying solely on bag-label constraints struggle to obtain sufficiently discriminative representations for accurate WSI classification. The latter tends to introduce pseudo-supervision for instances or pseudo-bags via elaborate pseudo-labels to enhance feature representation (Lu et al. 2021; Zhang et al. 2022; Qu et al. 2022b; Yu et al. 2023). However, unreliable pseudo-labeling and the integration of limited instance predictions into final bag decisions inevitably compromise model robustness and generalization. In particular, the mutual potential between representation and label prediction is not fully exploited, which may lead to spurious associations between features and disease properties. To address this issue, we propose a pseudo-label scheme and an Adaptive Memory Bank, exploring the collaborative relationships among feature representations, pseudo-label, and bag-label for better decision-making.

2.2 Pseudo-label Generation on WSI

The pseudo-label-based MIL approach shows significant superiority in exploring richer information via fine-grained supervision for WSI analysis, broadly divided into pseudo-instance-level and pseudo-bag-level (Campanella et al. 2019; Zhang et al. 2022). Generating pseudo-label for instances approaches has been explored from various perspectives, including threshold based (Wang et al. 2022; Tokunaga et al. 2020; Liu et al. 2023) attention scores (Yang et al. 2023; Qu et al. 2022b), clustering prototype similarity (Qu et al. 2024), and teacher-student model (Wang et al. 2023). However, besides over-reliance on threshold or pre-existing models affecting decision stability and generalization, offline or lagged update pseudo-label schemes hinder models from capturing real-time semantic variations, leading to sub-optimal feature representations. The latter generally divides the bag into multiple groups and introduces a bag-label for pseudo-bag representation (Zhang et al. 2022; Liu et al. 2024). However, these methods often overlook the latent inaccuracies in pseudo-bag-labels. Furthermore, the advanced PAMIL method (Zheng, Jiang, and Yao 2024) merges all historical pseudo-bag representations, which ex-

acerbates computational consumption and redundant information interference. In contrast, our proposed an online and dynamic Decision Revision Pseudo-Label (DRPL) generation scheme while designing a most informative bank and a multi-level feature refinement for more compact and discriminative representations.

3 Method

This section starts with the MIL formulation and the overview of our Online Pseudo-Supervision and Dynamic Mutual Learning (OODML), followed by our proposed Self-Progressive Feature Fusion (SPFF) module, Explicit-Implicit Representation Learning (EIRL) module, Decision Revision Pseudo-Label (DRPL) generation scheme, Dynamic Decision-Making (DDM) module.

3.1 Review for MIL Formulation

Taking a binary classification task as an example, given a WSI X and the corresponding labels $Y \in \{0, 1\}$, it is treated as the "bag" involving B instances $\{x_b \in \mathbb{R}^{W\times H\times 3} | 1 \leq b \leq B\}$, where H and W denote the height and width of one instance.

$$Y = \begin{cases} 1, & \text{if } \exists x_b \in X : y_b = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where y_b is the ground truth label of the x_b instance. It is interpreted as if an instance is labeled 1, the bag prediction is "positive". Otherwise, it will be labeled 0 (negative).

The common MIL methods use the pre-trained encoder (Lu et al. 2021; Zheng, Jiang, and Yao 2024) to embed x_b into a D-dimensional feature vector $v_b \in \mathbb{R}^{1\times D}$, and a feature aggregator G to integrate $\{v_b\}_{b=1}^{B}$ into a panoramic view of bag representation h, followed by a Multi-Layer Perceptron (MLP) $\mathcal{P}$ to predict the label $\hat{Y}_{DM}$. Since the labels of $\{v_b\}_{b=1}^{B}$ are unknown for G and $\mathcal{P}$ training, inaccurate feature representation and undesired decision-making can arise in MIL approaches.

3.2 Overview for OODML

Figure 2 outlines our proposed OODML. Our primary goal is to construct high-quality discriminative features and reliable pseudo-label generation to facilitate bag label inference by introducing an Adaptive Memory Bank and mutual learning between label predictions and feature representations.

To alleviate single instance information incompleteness and all instance redundancy interference, we randomly sample pseudo-bag $\{v^t = \{v_m^t\}_{m=1}^{M} | t \in [T = \frac{B}{M}]\}$ at t from bag X, where X is divided into T pseudo-bags and a pseudo-bag contains M instance features. Unlike packing all instances into the model simultaneously (Lu et al. 2021), dynamic sampling alleviates the learning difficulties across entire instances by focusing on correlations within pseudo-bags, thus enhancing pseudo-bag representation u^t capability. To promote feature representation, u^t is equipped with a SPFF module to aggregate additional valuable information from the Adaptive Memory Bank (AMB) and refined features from the Multi-Level Feature Fusion (MLFF) into the current pseudo-bag token F^t facilitating pseudo-label prediction $\hat{y}^t$ of F^t. Subsequently, the $\hat{y}^t$ assist AMB to capture and update the most K label-related tokens $\{\{F^k\}_{k=1}^{K} | 1 \leq K \leq T\}$ for providing class-aware knowledge for pseudo-bag token u^{t+1}. To deliver credible pseudo-label $\{y^t\}_{t=1}^{T}$ supervision $\{\hat{y}^t\}_{t=1}^{T}$, we devise a DRPL generation scheme by designing an EIRL module to explore complex relationships between pseudo-bag tokens $\{F^t\}_{t=1}^{T}$ and bag-label predictions $(\hat{Y}_{ER}, \hat{Y}_{IR})$. In addition, a DDM module takes the bag predictions and representations as input to adaptively weight $\hat{Y}_{IR}$ and $\hat{Y}_{ER}$ for more robust bag decision-making $\hat{Y}_{DM}$. Finally, harmonizing bag labels and pseudo-labels optimizes OODML, which promotes better feature representation and pseudo-label generation, resulting in elegant mutual learning.

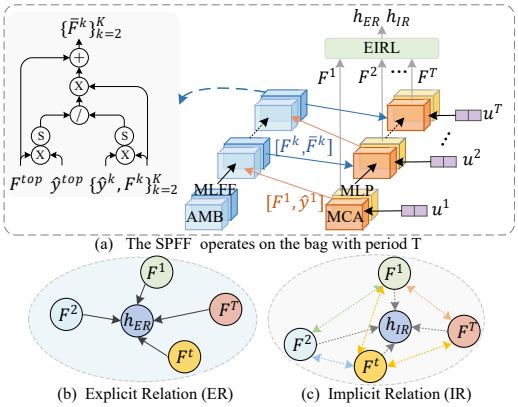

Figure 3: Illustration of SPFF, explicit and implicit relations.

3.3 Self-Progressive Feature Fusion

To enhance feature representation and facilitate accurate pseudo-label predictions, we propose a SPFF module to explore and aggregate discriminative information. In particular, SPFF is equipped with an Adaptive Memory Bank (AMB) to summarize high confidence tokens $\{F^k\}_{k=1}^{K}$ and a Multi-Level Feature Fusion (MLFF) to extract label-dependent properties into $\{\bar{F}^k\}_{k=2}^{K}$. Specifically, taking the operation at t as an example, the Transformer module takes a position vector v_p^t and an initial class token u^t as input to aggregate spatial and morphological features in v^t into u^t. To promote pseudo-bag token representation, u^t employs a Memory-based Cross-Attention (MCA) to fuse class-related historical information from concatenated $\{F^k\}_{k=1}^{K}$ and $\{\bar{F}^k\}_{k=2}^{K}$ into the F^t. Subsequently, F^t is passed to the MLP to infer pseudo-label $\hat{y}^t$ of F^t. The procedures in SPFF are expressed as

$$u^t = \text{Transformer}\left(u^t, (v_p^t + v^t)\right), \tag{2}$$

$$F^t = \text{MCA}\left(u^t, \text{Cat}[\{F^k\}_{k=1}^{K}, \{\bar{F}^k\}_{k=2}^{K}]\right), \tag{3}$$

$$\hat{y}^t = \text{MLP}\left(F^t\right), \tag{4}$$

Adaptive Memory Bank and Multi-level Feature Fusion. To facilitate discriminative pseudo-bag tokens, we devise the

AMB and MLFF by leveraging the pseudo-label prediction confidence to store and distill label-related information into tokens. As shown in Figure 3 (a), before u^t, AMB takes $\hat{y}^{t-1}$ and F^{t-1} as inputs, and rearranges $\mathcal{M}^{t-1}$ by comparing $\hat{y}^{t-1}$ to the stored $\{\hat{y}^k\}_{k=1}^K$ to preserve the K pivotal tokens and corresponding label score. In addition, the MLFF involves two-stage selections and a fusion feature operation. Specifically, MLFF employs $\hat{y}^k$ to serve as class-related prompts to refine $\{F^k\}_{k=1}^K$ into $\{\hat{F}^k\}_{k=1}^K$ (Eq. 6). To further highlight key contributions, the Softmax$(\cdot)$ is used to reinforce the most salient features in $\hat{F}^k$, and then $\{\hat{F}^k\}_{k=2}^K$ are quantified to the token $\hat{F}^{top}$ with the highest confidence to obtain the weights $\{\mathcal{W}\}_{k=2}^K$ (Eq. 7). Subsequently, the precise $\{\mathcal{W}^k\}_{k=2}^K$ are employed to aggregate dispersed discriminative expressions from $\{F^k\}_{k=2}^K$ and F^{top} into $\{\bar{F}^k\}_{k=2}^K$ (Eq. 8). The operations can be expressed as

$$\mathcal{M}^{t-1} = \begin{cases} \left[F^{\text{top}}, F^2, \ldots, F^K\right], \\ \left[\hat{y}^{\text{top}} \geqslant \hat{y}^2 \geqslant, \ldots, \geqslant \hat{y}^K\right], \end{cases} \tag{5}$$

$$\{\hat{F}^k\}_{k=1}^K = \{\hat{y}^k\}_{k=1}^K \times \{F^k\}_{k=1}^K, \tag{6}$$

$$\{\mathcal{W}^k\}_{k=2}^K = \frac{\text{Softmax}(\{\hat{F}^k\}_{k=2}^K)}{\text{Softmax}(\hat{F}^{top})}, \tag{7}$$

$$\{\bar{F}^k\}_{k=2}^K = F^{top} + \{\mathcal{W}^k \times F^k\}_{k=2}^K. \tag{8}$$

3.4 Explicit-Implicit Representation Learning

To promote pseudo-label generation, we present an EIRL module to explore explicit relations (ER) and implicit relations (IR) between pseudo-bag tokens $\{F^t\}_{t=1}^T$, and bag representations (h_{ER}, h_{IR}), as illustrated in Figure 3 (b) and (c). In particular, we design a DRPL scheme to generate pseudo-label $\{y^t\}_{t=1}^T$ for $\{F^t\}_{t=1}^T$ in the positive bag X. Otherwise, the $\{y^t\}_{t=1}^T$ is set to negative 0. Specifically, we employ a Linear Attention Module (LAM) to explore single-hop relations $\{A_{ER}^t\}_{t=1}^T$ between $\{F^t\}_{t=1}^T$ and h_{ER}. Meanwhile, a Multi-head Self-Attention (MSA) introduce an initial class token h_{IR} for multi-hop and cross-relations $\{A_{IR}^t\}_{t=1}^T$ between $\{F^t\}_{t=1}^T$ and h_{IR}. After that, h_{ER} and h_{IR} are transformed into the MLP to predict bag label $\hat{Y}_{ER}$ and $\hat{Y}_{IR}$, respectively. The procedures can be formulated as

$$h_{ER}, \{A_{ER}^t\}_{t=1}^T = \text{LAM}\left(\{F^t\}_{t=1}^T\right), \tag{9}$$

$$h_{IR}, \{A_{IR}^t\}_{t=1}^T = \text{MSA}\left(h_{IR}, \{F^t\}_{t=1}^T\right), \tag{10}$$

$$\hat{Y}_{ER} = \text{MLP}\left(h_{ER}\right), \hat{Y}_{IR} = \text{MLP}\left(h_{IR}\right). \tag{11}$$

Since the one-perspective attention scores are noisy, we leverage class-specific confidence scores to rectify $\{A_{ER}^t\}_{t=1}^T$ and $\{A_{IR}^t\}_{t=1}^T$ for reliable pseudo-labels. This process is expressed as

$$\{y^t\}_{t=1}^T = S\left(\frac{\{A_{ER}^t\}_{t=1}^T \times \hat{Y}_{ER} + \{A_{IR}^t\}_{t=1}^T \times \hat{Y}_{IR}}{\tau}\right), \tag{12}$$

where $S(\cdot)$ is the Sigmoid function to adjust the pseudo-label to $0-1$ and τ is set to 1 or 2 to accommodate different tumor sizes of WSI datasets.

3.5 Dynamic Decision-Making Module

DDM aims to harmonize bag explicit-implicit representations and predictions for better bag decision-making. Specifically, we employ $\hat{Y}_{IR}$ and $\hat{Y}_{ER}$ to distill class-dependent information in h_{IR} and h_{ER} for generating reliable weights, and then integrate $\hat{Y}_{IR}$ and $\hat{Y}_{ER}$ for the final bag label prediction $\hat{Y}_{DM}$. The procedure is expressed as

$$H_{ER} = (\hat{Y}_{ER} \times h_{ER}), \quad H_{IR} = (\hat{Y}_{IR} \times h_{IR}), \tag{13}$$

$$W_{ER}, W_{IR} = \text{Softmax}\left(\text{Linear}\left(\text{Cat}[H_{ER}, H_{IR}]\right)\right), \tag{14}$$

$$\hat{Y}_{DM} = W_{ER} \times \hat{Y}_{ER} + W_{IR} \times \hat{Y}_{IR}. \tag{15}$$

3.6 Model Optimization

To achieve better feature representation and pseudo-label generation facilitating bag decision-making, we explore practical composite constraints on three bag label probabilities ($\hat{Y}_{ER}$, $\hat{Y}_{IR}$ and $\hat{Y}_{DM}$), and pseudo-label prediction $\hat{y}^t$, involving two bag cross-entropy loss ($\mathcal{L}_{\text{ER}}$ and $\mathcal{L}_{\text{IR}}$), DDM cross-entropy losses $\mathcal{L}_{\text{DM}}$ and pseudo-label cross-entropy loss (PLCE) $\mathcal{L}_{\text{PLCE}}$. These loss functions are formulated as

$$\begin{aligned} &\mathcal{L}_\zeta = -\left[Y \log \hat{Y}_\zeta + (1-Y)\log(1-\hat{Y}_\zeta)\right], \\ &\zeta \in \{\text{ER}, \text{IR}, \text{DM}\}, \end{aligned} \tag{16}$$

$$\mathcal{L}_{\text{PLCE}} = -\frac{1}{T}\sum_{t=1}^{T}\left[y \log \hat{y}^t + (1-y)\log(1-\hat{y}^t)\right], \tag{17}$$

$$\mathcal{L}_{\text{OODML}^\star} = \frac{1}{N}\sum_{i=1}^{N}\left[\mathcal{L}_{\text{ER}} + \mathcal{L}_{\text{IR}} + \lambda_{\text{PLCE}}\mathcal{L}_{\text{PLCE}}\right]. \tag{18}$$

where λ_{PLCE} is adjusted from 0 to 0.5 by integrating epoch and cosine curve to progressively learn label-related information among pseudo-bags. DDM is optimized by $\mathcal{L}_{\text{DM}}$.

4 Experiments

To validate our OODML, we conduct extensive experiments on the CAMELYON16 (Bejnordi et al. 2017), Breast Carcinoma Subtyping (Brancati et al. 2022), TCGA Lung Cancer and TCGA Esophageal Cancer (Tomczak, Czerwińska, and Wiznerowicz 2015) datasets with representative WSI analysis MIL methods, involving *1) bag-label-level:* (MeanPooling, ABMIL (Ilse, Tomczak, and Welling 2018), SetTransformer (Lee et al. 2019), DeepAttnMIL (Yao et al. 2020), DSMIL (Bin 2021), TransMIL (Shao et al. 2021), MuRCL (Zhu et al. 2022), MHIM (Tang et al. 2023), IBMIL (Lin et al. 2023), SSMMIL (Fillioux et al. 2023), ACMIL (Zhang et al. 2025))) and *2) pseudo-label-level:* (MaxPooling, RNNMIL (Campanella et al. 2019), CLAM (Lu et al. 2021), DTFD (Zhang et al. 2022), DGMIL (Qu et al. 2022a), IAT (Xiong et al. 2023), BCL (Yu et al. 2023), CIMIL (Lin et al. 2024)).

4.1 Datasets and Metrics

CAMELYON16 Dataset consists of 270 training WSIs (159 normal, 111 tumor) and 129 testing WSIs for breast cancer lymph nodes (Bejnordi et al. 2017). **TCGA**

Methods	CAMELYON16			TCGA Lung		
	Accuracy	F1	AUC	Accuracy	F1	AUC
ABMIL (Ilse, Tomczak, and Welling 2018)	0.784±0.030	0.703±0.029	0.852±0.054	0.844±0.023	0.849±0.021	0.919±0.026
MeanPooling (Van der Laak, Litjens, and Ciompi 2021)	0.586±0.125	0.455±0.065	0.573±0.021	0.829±0.020	0.830±0.026	0.905±0.014
MaxPooling (Van der Laak, Litjens, and Ciompi 2021)	0.603±0.138	0.536±0.132	0.633±0.122	0.839±0.041	0.844±0.043	0.920±0.023
DSMIL (Bin 2021)	0.836±0.034	0.746±0.068	0.862±0.022	0.783±0.041	0.789±0.033	0.856±0.040
CLAM-MB (Lu et al. 2021)	0.826±0.024	0.748±0.033	0.867±0.024	0.844±0.023	0.849±0.021	0.919±0.026
CLAM-SB (Lu et al. 2021)	0.803±0.017	0.719±0.026	0.856±0.016	0.834±0.030	0.838±0.029	0.912±0.034
TransMIL (Shao et al. 2021)	0.804±0.016	0.716±0.035	0.771±0.046	0.819±0.038	0.823±0.032	0.885±0.030
DTFD(AFS) (Zhang et al. 2022)♯	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(MaxMin) (Zhang et al. 2022)♯	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 (Qu et al. 2022a)♯	0.802±—	—	0.837±—	0.920±—	—	0.970±—
MuRCL(CLAM-SB) (Zhu et al. 2022)♯	0.913±0.019	0.880±0.026	0.945±0.009	0.892±0.007	0.886±0.007	0.964±0.003
MuRCL(ABMIL) (Zhu et al. 2022)♯	0.902±0.022	0.864±0.026	0.953±0.009	0.889±0.013	0.887±0.012	0.958±0.004
IAT (Xiong et al. 2023)♯	0.887±0.022	0.874±0.025	0.946±0.011	0.849±0.016	0.849±0.015	0.921±0.008
MHIM (TransMIL) (Tang et al. 2023)♯	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) (Tang et al. 2023)♯	0.925±0.003	0.908±0.007	0.965±0.006	0.898±0.033	0.897±0.029	0.955±0.017
BCL (Yu et al. 2023)♯	0.945±—	—	0.956±—	0.908±—	—	0.960±—
SSMMIL (Fillioux et al. 2023)	0.798±0.033	0.843±0.027	0.715±0.043	0.843±0.033	0.847±0.034	0.919±0.022
CIMIL(TransMIL) (Lin et al. 2024)♯	0.886±—	—	0.928±—	0.913±0.019	—	0.958±0.015
CIMIL(CLAM) (Lin et al. 2024)♯	0.898±—	—	0.902±—	0.896±0.010	—	0.960±0.008
OODML⋆ (Ours)	**0.962±0.023**	**0.972±0.022**	**0.979±0.014**	**0.938±0.020**	**0.937±0.019**	**0.965±0.012**
OODML♠ (Ours)	**0.978±0.016**	**0.977±0.017**	**0.987±0.005**	**0.982±0.013**	**0.940±0.015**	**0.986±0.011**

Table 1: Quantitative comparison of our results and MIL methods on CAMELYON16 and TCGA Lung datasets. ♯ Results are derived from their papers, with others taken from official code implementations. The numbers in **black** and underline indicate the best and second performance. ⋆ and ♠ denote $\hat{Y}_{DM}$ from the average weighted and DDM weighted of $\hat{Y}_{ER}$ and $\hat{Y}_{IR}$.

Methods	Accuracy	F1	AUC
ABMIL	0.745±0.026	0.686±0.031	0.846±0.024
RNNMIL	0.739±0.062	0.682±0.089	0.818±0.058
SetTransformer	0.758±0.037	0.737±0.052	0.868±0.017
DeepAttnMIL	0.739±0.017	0.611±0.014	0.804±0.013
DSMIL	0.840±0.047	0.780±0.069	0.932±0.022
CLAM-MB	0.883±0.043	0.843±0.063	0.941±0.023
CLAM-SB	0.872±0.038	0.828±0.057	0.936±0.022
TransMIL†	0.894±0.020	0.899±0.024	0.945±0.008
DTFD†	0.827±0.024	0.835±0.029	0.866±0.027
MuRCL (CLAM)	0.887±0.014	0.861±0.016	0.938±0.012
MuRCL (ABMIL)	0.831±0.029	0.800±0.048	0.891±0.007
SSMMIL†	0.858±0.024	0.857±0.021	0.916±0.012
OODML⋆ (Ours)	**0.965±0.010**	**0.963±0.013**	**0.975±0.013**
OODML♠ (Ours)	**0.967±0.012**	**0.966±0.011**	**0.987±0.012**

Table 2: Quantitative comparison of our results and MIL results from MuRCL (Zhu et al. 2022) on the ESCA dataset. † Results follow their official codes for implementation.

Methods	Res18ImageNet		ViT16SSL	
	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	0.523±0.028	0.723±0.035	0.680±0.051	0.866±0.029
DSMIL	0.511±0.052	0.751±0.028	0.577±0.028	0.816±0.028
CLAM	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	0.469±0.016	0.717±0.032	0.612±0.080	0.870±0.022
IBMIL	0.510±0.043	0.726±0.034	0.645±0.041	0.871±0.014
ACMIL	0.552±0.048	0.754±0.008	0.722±0.030	0.888±0.010
MHIM	0.511±0.022	0.774±0.021	0.625±0.060	0.865±0.017
SSMMIL†	0.549±0.052	0.759±0.018	0.658±0.031	0.853±0.015
OODML⋆ (Ours)	**0.772±0.043**	**0.899±0.006**	**0.814±0.033**	**0.934±0.023**
OODML♠ (Ours)	**0.835±0.039**	**0.947±0.033**	**0.843±0.035**	**0.960±0.026**

Table 3: Quantitative comparison of our results and MIL results from ACMIL (Zhang et al. 2025) on the BRACS. † Results follow their official codes for implementation.

Esophageal Cancer Dataset (ESCA) contains 156 diagnostic WSIs, composed of 90 squamous cell carcinoma and 66 adenocarcinomas (Tomczak, Czerwińska, and Wiznerowicz 2015). Following existing methods (Zhang et al. 2022; Zhu et al. 2022), CLAM (Lu et al. 2021) is employed to extract tissue regions from WSIs and crop non-overlapping 256×256 patches at $20\times$ magnification. **TCGA Lung Cancer Dataset** contains two cancer sub-categories: 541 Lung Adenocarcinoma (LUAD) and 512 Lung Squamous Cell Carcinoma (LUSC) WSIs (Tomczak, Czerwińska, and Wiznerowicz 2015). CLAM (Lu et al. 2021) is used to crop each WSI into 256×256 non-overlapping instances at $20\times$ magnification. **Breast Carcinoma Subtyping Dataset** (BRACS) involves three categories: 265 benign, 89 atypical, and 193 malignant breast tumor (Brancati et al. 2022). CLAM is used to extract instances of size 256×256 at $10\times$ magnification.

Evaluation Metrics. Standard metrics such as Area Under Curve (AUC), accuracy, F1 score (F1), precision, and recall are used for evaluation, with a threshold set at 0.5. Due to discrepancies in dataset splits, we follow the settings from (Zhang et al. 2022, 2023; Zhu et al. 2022; Zhang et al. 2025) for fair comparison. The CAMELYON16 official

training set is randomly divided into training and validation sets at a $9:1$ ratio. TCGA Lung and TCGA ESCA datasets are randomly split into training, validation, and testing sets with ratios of $65:10:25$ and $3:1:1$, respectively. For BRACS, we follow the official dataset split (Brancati et al. 2022; Zhang et al. 2025), with 537 WSIs available 395 for training, 65 for validation, and 87 for testing. We report the mean and standard deviation for at least 5 models in all experiments.

4.2 Implementation Details

For the CAMELYON16 dataset, following (Zhang et al. 2022), we extract 1024-dimensional feature vectors from each instance by a pre-trained ResNet50 (He et al. 2016) on ImageNet (Deng et al. 2009) (Res50ImageNet). For the TCGA Lung and TCGA ESCA datasets, we employ ImageNet with ResNet18 encoders, respectively, to obtain 512-dimensional instance feature vectors. For the BRACS dataset, we use public features from (Zhang et al. 2025), pre-trained with 512-dimensional ResNet18 on ImageNet (Res18ImageNet) and 384-dimensional ViT-S/16 self-supervised learning (ViT16SSL) using DINO (Caron et al. 2021). AdaMax optimizer (Adam et al. 2014) with a weight decay of $1e-5$ and the initial learning rate of $1e-4$ are used. With the above settings, we train OODML with 200 epochs and batch size 1 on a single NVIDIA RTX 3090Ti GPU.

4.3 Comparison with State-of-the-arts

We compare the performance of OODML with representative methods on four WSI analysis datasets. Quantitative results are presented in Tables 1, 2 and 3. As expected, our OODML* and OODML$^\spadesuit$ achieves competitive performance in terms of all metrics, surpassing pseudo-label-based (BCL (Yu et al. 2023) and CIMIL (Lin et al. 2024)) and bag-label-based TransMIL (Shao et al. 2021) methods by 3.3%, 6.9% and 7.3% in accuracy, respectively. In particular, our OODML$^\spadesuit$ outperforms the ACMIL (Zhang et al. 2025) on both pre-processing instance features on the BRACS dataset, with improvements of 28.3% and 12.1% on F1. The substantial gains benefit from DRPL generating reliable pseudo-labels and SPFF producing more discriminative pseudo-bag representations. Yet, the BCL and ACMIL employ attention scores to focus on Top-K instances for improving feature aggregators, leading to inaccurate and insufficient representations. In addition, the MHIM (Tang et al. 2023) and MuRCL (Zhu et al. 2022) achieve an improvement by capturing salient features or instances across the bag. However, due to redundant instance interference, they fall 5.3% and 6.5% behind our OODML$^\spadesuit$ in terms of accuracy on the CAMELYON16 dataset, respectively. Notably, our OODML$^\spadesuit$ significantly enhances the prediction performance compared to OODML* across all metrics and datasets. These findings validate the effectiveness and rationality of facilitating pseudo-label generation and feature representation and their mutual learning.

Model	PBCE	DRPL	AMB	MLFF	Accuracy	F1	AUC
w BCE	✗	✗	✗	✗	0.574±0.026	0.532±0.012	0.764±0.010
w PBCE	✓	✗	✗	✗	0.609±0.012	0.555±0.024	0.740±0.021
w DRPL	✗	✓	✗	✗	0.640±0.011	0.602±0.004	0.765±0.006
w/o MLFF	✗	✓	✓	✗	0.651±0.015	0.611±0.036	0.780±0.002
w/o MLFF†	✗	✓	✓	✗	0.627±0.015	0.531±0.035	0.743±0.019
w/o MLFF‡	✗	✓	✓	✗	0.667±0.014	0.642±0.026	0.799±0.011
w/o PBCE	✗	✗	✓	✓	0.685±0.021	0.649±0.032	0.816±0.034
w/o DRPL	✓	✗	✓	✓	0.712±0.017	0.693±0.023	0.881±0.004
OODML*	✗	✓	✓	✓	0.770±0.041	0.772±0.043	0.899±0.006
OODML$^\spadesuit$	✗	✓	✓	✓	0.837±0.043	0.835±0.039	0.947±0.033

(a) Validation on basic components. † and ‡ denote mixup scheme (Zhang et al. 2017) ($\bar{F}^k = \beta F^{top} + (1-\beta)F^k$) and $\mathcal{W}^k = \frac{F^k}{F^{top}}$ in Eq. 8.

Model	Precision	Recall	Accuracy	F1
w $\{A^t_{IR}\}^T_{t=1}$	0.576±0.012	0.573±0.009	0.595±0.016	0.568±0.010
w $\hat{Y}_{IR}\{A^t_{IR}\}^T_{t=1}$	0.596±0.012	0.583±0.009	0.609±0.016	0.581±0.004
w $\{A^t_{ER}\}^T_{t=1}$	0.599±0.014	0.586±0.006	0.618±0.008	0.580±0.007
w $\hat{Y}_{ER}\{A^t_{ER}\}^T_{t=1}$	0.608±0.019	0.592±0.009	0.628±0.009	0.581±0.011
w DRPL (Ours)	**0.645±0.033**	**0.608±0.007**	**0.640±0.011**	**0.602±0.004**

(b) Effects of decision revision pseudo-label scheme.

Model	Accuracy	F1	AUC
$\hat{Y}_{IR}$	0.770±0.026	0.767±0.027	0.907±0.024
$\hat{Y}_{ER}$	0.726±0.016	0.726±0.017	0.840±0.015
Concat($h_{IR}; h_{ER}$)	0.795±0.061	0.790±0.055	0.897±0.061
Concat($H_{IR}; H_{ER}; h_{IR}; h_{ER}$)	0.832±0.048	0.829±0.046	0.935±0.040
Concat($H_{IR}; H_{ER}$) (Ours)	**0.837±0.043**	**0.835±0.039**	**0.947±0.033**

(c) Effects of dynamic decision-making module.

K	Accuracy	F1	AUC	M	Precision	Recall	F1
0	0.640	0.602	0.765	128	0.585	0.567	0.549
5	**0.770**	**0.772**	**0.899**	256	0.564	0.544	0.518
10	0.720	0.700	0.844	**512**	**0.590**	**0.571**	**0.555**
20	0.671	0.619	0.789	1024	0.554	0.551	0.529

(d) Effects of AMB size and pseudo-bag size are evaluated using OODML* and *w* PBCE models, respectively.

Table 4: Ablation study on the BRACS dataset. Apart from OODML$^\spadesuit$, the other results use the average $\hat{Y}_{IR}$ and $\hat{Y}_{ER}$.

4.4 Ablation Studies

Validation on Basic Components. We conduct ablation studies to validate the contributions of individual components, including bag label as pseudo-label for cross-entropy loss (PBCE), DRPL generation scheme, SPFF module and DDM module to the final performance. For simplicity, we denote our final model as OODML$^\spadesuit$ and devise the baseline model by removing all components apart from the loss $\mathcal{L}_{\text{IR}}$ (*w* BCE). We design four models (*w* BCE, *w* PBCE, *w* DRPL, and *w/o* DRPL) to investigate the effectiveness of pseudo-label generation. The mutual learning among pseudo-label predictions and feature representations in SPFF is verified by *w* DRPL, *w/o* PBCE, *w/o* MLFF, *w/o* MLFF†, *w/o* MLFF‡ and OODML* models. Finally, OODML* and OODML$^\spadesuit$ models are designed to evaluate the DDM.

Quantitative results are presented in Table 4 (a), revealing that the complete model OODML$^{\spadesuit}$ achieves significant improvements over its incomplete versions. Specifically, the reliable pseudo-label helps the model to fully capture class-dependent information among pseudo-bag instances, resulting in accuracy gains by 6.6% (*w* DRPL vs. *w* BCE) and 3.1% (*w* DRPL vs. *w* PBCE). Removing DRPL may significantly damage pseudo-bag representation quality and label predictions, leading to a performance drop by 5.8% (*w/o* DRPL vs. OODML*). But notably, the scarcity of disease features in individual instances or pseudo-bag struggles to deliver accurate and reliable pseudo-label predictions. To address this issue, we design the informative AMB and MLFF for better pseudo-bag representation, achieving 13.0% (*w* DRPL vs. OODML*) and 11.1% (*w* BCE vs. *w/o* PBCE) gains in accuracy. However, random fusion destroys label-related components in AMB, and removing two-stage feature refinement in MLFF suffers from imprecise and redundant features, leading to a dramatic decline of 14.3%, 11.9% and 10.3% in accuracy (referring to OODML*, *w/o* MLFF^{+}, *w/o* MLFF and *w/o* MLFF‡ models). In addition, comparing OODML*, OODML$^{\spadesuit}$ shows considerable superiority, achieving a 6.7% improvement in accuracy.

Effect of Decision Revision Pseudo-Label Scheme. As reported in Table 4 (b), comparing $\{A_{IR}^t\}_{t=1}^T$ and $\{A_{ER}^t\}_{t=1}^T$ as pseudo-labels, introducing bag label confidence scores $\hat{Y}_{IR}$ and $\hat{Y}_{ER}$ to calibrate the attention scores can effectively reduce biases and bag label predictions, resulting in 1.4% and 1.0% accuracy gains. By coupling pseudo-bag representation and bag label prediction using LAM and MSA to generate reliable pseudo-labels, our *w* DRPL achieves favorable performance across all metrics.

Effect of Dynamic Decision-Making Module. The quantitative results are listed in Table 4 (c), illustrating that harmonizing relations between bag representations and label predictions facilitates more precise decision-making, raising accuracy by 11.1% and 6.7%. While connecting explicit-implicit representations and refined features can improve performance, it results in a decrease of 5.0% and 1.2% in AUC compared to our method due to redundant and inaccurate representations.

Effect of Adaptive Memory Bank and Pseudo-bag Size. We conduct experiments to analyze the effect of AMB size K and pseudo-bag size M. The numerical comparisons are shown in Table 4 (d). Large K and M may lead to redundant interference, while small ones lack the detail to accurately distinguish tumors from common characteristics, leading to a high false-positive rate. Based on these findings, we set the AMBS and PBS to 5 and 512 respectively across all experiments and designed feature refinement schemes to effectively aggregate class-related information.

5 Conclusion

In this study, an Online Pseudo-Supervision and Dynamic Mutual Learning (OODML) framework is devised to promote feature representation and pseudo-label generation for better WSI classification. To promote feature representation, we implement a Self-Progressive Feature Fusion (SPFF) module equipped with an Adaptive Memory Bank (AMB) to fully exploit pseudo-label priors for capturing target-related features while refining and aggregating informative representations. Meanwhile, we devise a Decision Revision Pseudo-Label (DRPL) generation scheme to explore intrinsic connections between pseudo-bag representations and bag label predictions, ensuring reliable pseudo-labels. Furthermore, we design a Dynamic Decision-Making (DDM) module to boost bag decision-making. Extensive experiments on four datasets demonstrate the impressive performance of our proposed OODML. Although our OODML excels in eliminating interference and preserving high-quality salient features, one limitation is the memory mechanism. A potential improvement could involve learnable dictionary networks based on intrinsic mutual relations between predictions and representations.

Acknowledgements

This research was supported by the National Science and Technology Major Project (2021ZD0110901), the National Science Foundation of China (62476069), and in part by the National Natural Science Foundation of China (62371350).

References

Adam, K. D. B. J.; et al. 2014. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412.

Bejnordi, B. E.; Veta, M.; Van Diest, P. J.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J. A.; Hermsen, M.; Manson, Q. F.; Balkenhol, M.; et al. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22): 2199–2210.

Bin, K. W. E., Li. Yin Li. 2021. Dual-Stream Multiple Instance Learning Network for Whole Slide Image Classification With Self-Supervised Contrastive Learning. In *CVPR*, 14318–14328.

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

Campanella, G.; Hanna, M. G.; Geneslaw, L.; Miraflor, A.; Werneck Krauss Silva, V.; Busam, K. J.; Brogi, E.; Reuter, V. E.; Klimstra, D. S.; and Fuchs, T. J. 2019. Clinical-Grade Computational Pathology Using Weakly Supervised Deep Learning on Whole Slide Images. *Nature medicine*, 25(8): 1301–1309.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *CVPR*, 9650–9660.

Chen, K.; Sun, S.; and Zhao, J. 2024. CaMIL: Causal Multiple Instance Learning for Whole Slide Image Classification. In *AAAI*, 1120–1128.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Fillioux, L.; Boyd, J.; Vakalopoulou, M.; Cournède, P.-H.; and Christodoulidis, S. 2023. Structured state space models for multiple instance learning in digital pathology. In *MICCAI*, 594–604.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-Based Deep Multiple Instance Learning. In *ICML*, 2127–2136.

Lee, J.; Lee, Y.; Kim, J.; Kosiorek, A.; Choi, S.; and Teh, Y. W. 2019. Set transformer: A framework for attention-based permutation-invariant neural networks. In *ICML*, 3744–3753.

Lin, T.; Yu, Z.; Hu, H.; Xu, Y.; and Chen, C. W. 2023. Interventional Bag Multi-Instance Learning On Whole-Slide Pathological Images. In *CVPR*, 19830–19839.

Lin, W.; Zhuang, Z.; Yu, L.; and Wang, L. 2024. Boosting Multiple Instance Learning Models for Whole Slide Image Classification: A Model-Agnostic Framework Based on Counterfactual Inference. In *AAAI*, 3477–3485.

Liu, K.; Zhu, W.; Shen, Y.; Liu, S.; Razavian, N.; Geras, K. J.; and Fernandez-Granda, C. 2023. Multiple instance learning via iterative self-paced supervised contrastive learning. In *CVPR*, 3355–3365.

Liu, P.; Ji, L.; Zhang, X.; and Ye, F. 2024. Pseudo-Bag Mixup Augmentation for Multiple Instance Learning-Based Whole Slide Image Classification. *TMI*.

Lu, M. Y.; Williamson, D. F.; Chen, T. Y.; Chen, R. J.; Barbieri, M.; and Mahmood, F. 2021. Data-Efficient and Weakly Supervised Computational Pathology on Whole-Slide Images. *Nature biomedical engineering*, 5(6): 555–570.

Qu, L.; Luo, X.; Liu, S.; Wang, M.; and Song, Z. 2022a. Dgmil: Distribution guided multiple instance learning for whole slide image classification. In *MICCAI*, 24–34.

Qu, L.; Ma, Y.; Luo, X.; Guo, Q.; Wang, M.; and Song, Z. 2024. Rethinking multiple instance learning for whole slide image classification: A good instance classifier is all you need. *TCSVT*.

Qu, L.; Wang, M.; Song, Z.; et al. 2022b. Bi-directional weakly supervised knowledge distillation for whole slide image classification. *NeurIPS*, 35: 15368–15381.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; and Ji, X. 2021. Transmil: Transformer Based Correlated Multiple Instance Learning for Whole Slide Image Classification. *NeurIPS*, 34: 2136–2147.

Tang, W.; Huang, S.; Zhang, X.; Zhou, F.; Zhang, Y.; and Liu, B. 2023. Multiple Instance Learning Framework with Masked Hard Instance Mining for Whole Slide Image Classification. In *CVPR*, 4078–4087.

Tokunaga, H.; Iwana, B. K.; Teramoto, Y.; Yoshizawa, A.; and Bise, R. 2020. Negative pseudo labeling using class proportion for semantic segmentation in pathology. In *ECCV*, 430–446.

Tomczak, K.; Czerwińska, P.; and Wiznerowicz, M. 2015. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1): 68–77.

Van der Laak, J.; Litjens, G.; and Ciompi, F. 2021. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5): 775–784.

Wang, H.; Luo, L.; Wang, F.; Tong, R.; Chen, Y.-W.; Hu, H.; Lin, L.; and Chen, H. 2023. Iteratively coupled multiple instance learning from instance to bag classifier for whole slide image classification. In *MICCAI*, 467–476.

Wang, X.; Xiang, J.; Zhang, J.; Yang, S.; Yang, Z.; Wang, M.-H.; Zhang, J.; Yang, W.; Huang, J.; and Han, X. 2022. Scl-wc: Cross-slide contrastive learning for weakly-supervised whole-slide image classification. *NeurIPS*, 18009–18021.

Xiong, C.; Chen, H.; Sung, J. J.; and King, I. 2023. Diagnose Like a Pathologist: Transformer-Enabled Hierarchical Attention-Guided Multiple Instance Learning for Whole Slide Image Classification. In *IJCAI-23*.

Yang, L.; Mehta, D.; Liu, S.; Mahapatra, D.; Di Ieva, A.; and Ge, Z. 2023. TPMIL: Trainable Prototype Enhanced Multiple Instance Learning for Whole Slide Image Classification. *arXiv preprint arXiv:2305.00696*.

Yao, J.; Zhu, X.; Jonnagaddala, J.; Hawkins, N.; and Huang, J. 2020. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *MIA*, 65: 101789.

Yu, J.-G.; Wu, Z.; Ming, Y.; Deng, S.; Wu, Q.; Xiong, Z.; Yu, T.; Xia, G.-S.; Jiang, Q.; and Li, Y. 2023. Bayesian collaborative learning for whole-slide image classification. *TMI*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, H.; Meng, Y.; Zhao, Y.; Qiao, Y.; Yang, X.; Coupland, S. E.; and Zheng, Y. 2022. DTFD-MIL: Double-Tier Feature Distillation Multiple Instance Learning for Histopathology Whole Slide Image Classification. In *CVPR*, 18780–18790.

Zhang, R.; Zhang, Q.; Liu, Y.; Xin, H.; Liu, Y.; and Wang, X. 2023. Multi-Level Multiple Instance Learning with Transformer for Whole Slide Image Classification. *arXiv preprint arXiv:2306.05029*.

Zhang, Y.; Li, H.; Sun, Y.; Zheng, S.; Zhu, C.; and Yang, L. 2025. Attention-challenging multiple instance learning for whole slide image classification. In *ECCV*, 125–143.

Zheng, T.; Chen, W.; Li, S.; Quan, H.; Zou, M.; Zheng, S.; Zhao, Y.; Gao, X.; and Cui, X. 2023. Learning how to detect: A deep reinforcement learning method for whole-slide melanoma histopathology images. *CMIG*, 108: 102275.

Zheng, T.; Jiang, K.; and Yao, H. 2024. Dynamic Policy-Driven Adaptive Multi-Instance Learning for Whole Slide Image Classification. In *CVPR*, 8028–8037.

Zhu, Z.; Yu, L.; Wu, W.; Yu, R.; Zhang, D.; and Wang, L. 2022. Murcl: Multi-instance reinforcement contrastive learning for whole slide image classification. *TMI*, 42(5): 1337–1348.