Multiple Instance Learning Framework with Masked Hard Instance Mining for Whole Slide Image Classification

Wenhao Tang[1] Sheng Huang[1]* Xiaoxian Zhang[1] Fengtao Zhou[2] Yi Zhang[1] Bo Liu[3]
[1] Chongqing University [2] The Hong Kong University of Science and Technology
[3] Walmart Global Tech
{whtang, huangsheng, zhangxiaoxian, zhangyii}@cqu.edu.cn
fzhouaf@connect.ust.hk, kfliubo@gmail.com

Abstract

The whole slide image (WSI) classification is often formulated as a multiple instance learning (MIL) problem. Since the positive tissue is only a small fraction of the gigapixel WSI, existing MIL methods intuitively focus on identifying salient instances via attention mechanisms. However, this leads to a bias towards easy-to-classify instances while neglecting hard-to-classify instances. Some literature has revealed that hard examples are beneficial for modeling a discriminative boundary accurately. By applying such an idea at the instance level, we elaborate a novel MIL framework with masked hard instance mining (MHIM-MIL), which uses a Siamese structure (Teacher-Student) with a consistency constraint to explore the potential hard instances. With several instance masking strategies based on attention scores, MHIM-MIL employs a momentum teacher to implicitly mine hard instances for training the student model, which can be any attention-based MIL model. This counter-intuitive strategy essentially enables the student to learn a better discriminating boundary. Moreover, the student is used to update the teacher with an exponential moving average (EMA), which in turn identifies new hard instances for subsequent training iterations and stabilizes the optimization. Experimental results on the CAMELYON-16 and TCGA Lung Cancer datasets demonstrate that MHIM-MIL outperforms other latest methods in terms of performance and training cost. The code is available at: https://github.com/DearCaat/MHIM-MIL.

1. Introduction

Histopathological image analysis plays a crucial role in modern medicine, particularly in the treatment of cancer, where it serves as the gold standard for diagnosis [18, 20, 24, 45]. Digitalizating pathological images into Whole Slide Images (WSIs) through digital slide scanner has opened new avenues for computer-aided analysis [9, 26]. Due to the huge size of a WSI and the lack of pixel-level annotations, histopathological image analysis is commonly formulated as a multiple instance learning (MIL) task [10, 23, 31]. In MIL, each WSI (or slide) is a bag containing thousands of unlabeled instances (patches) cropped from the slide. With at least one instance being disease positive, the bag is deemed positive, otherwise negative.

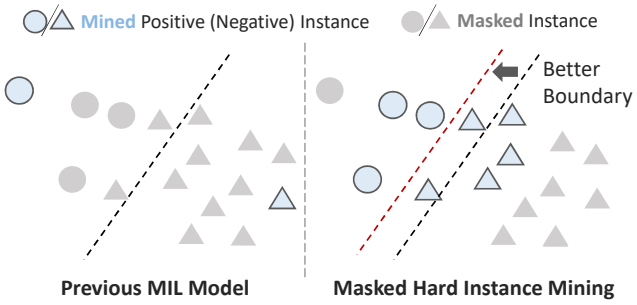

Figure 1: **Left:** Previous MIL models focus on the more salient instances. **Right:** MHIM-MIL mines an amount of hard-to-classify instances to learn a better boundary.

However, the number of slides is limited and each slide contains a mass of instances with a low positive proportion. This imbalance would hinder the inference of bag labels [16, 43]. To alleviate this issue, several WSI classification methods [6, 16–18, 26] employ an attention mechanism to aggregate salient instance features into a bag-level feature for WSI classification. Furthermore, some MIL frameworks [17, 21, 40, 43] focus on the more salient instances in the bag and leverage them to facilitate WSI classification. For instance, existing frameworks [40, 43] propose to only select the instances that correspond to the top K highest or lowest attention scores [17, 40] or patch probabilities [43] for yielding high-quality bag embedding for both training and testing.

*Corresponding Author.

These salient instances are actually "easy-to-classify" instances, which are not optimal for training a discriminative WSI classification model. In conventional machine learning, such as Support Vector Machines (SVM) [13], samples near the category distribution boundary are more challenging to classify, but are more useful for depicting the classification boundary, as illustrated in Figure 1. Moreover, other deep learning works [25, 28, 33, 34] also reveal that mining hard samples for training can improve the generalization abilities of models. By applying such an idea at the instance level, we can better highlight the "hard-to-classify" instances that facilitate MIL model training, and benefit the final WSI classification. However, the lack of instance labels poses a challenge to the direct application of traditional hard sample mining strategies at the instance level.

To address this issue, we present a novel MIL framework based on masked hard instance mining strategies (MHIM) named MHIM-MIL. The main idea of MHIM is to mask out the instances with high attention scores to highlight the hard instances for model training. Based on this, we incorporate two other instance masking strategies to enhance training efficiency and mitigate the over-fitting risk. Another key design of MHIM-MIL is an instance attention generator based on a Siamese structure (Teacher-Student) [3, 8]. In MHIM-MIL, the MIL-based WSI classification model is the student network, which aggregates hard instances mined by a momentum teacher with different instance masking strategies. The momentum teacher is updated using an exponential moving average (EMA) of the student model. Moreover, the framework is optimized by inducing a consistency constraint that explores more supervised information beyond the limited slide label. Unlike the conventional MIL frameworks [40, 43], which adopt complex cascade gradient-updating structures, our method is more simple and does not require additional parameters. It not only improves efficiency but also provides improved performance stability. The contribution of this paper is summarized as follows,

- We propose a simple and efficient MIL framework with masked hard instance mining named MHIM-MIL. It implicitly mines hard instances with instance attention for training a more discriminative MIL model. Extensive experiments on two WSI datasets validate that MHIM boosts different MIL models and outperforms other latest methods in terms of performance and training cost.

- We propose several hybrid instance masking strategies for indirectly mining hard instances in MIL. These strategies not only address the reliance problem of conventional methods on instance-level supervision but also enhance the training efficiency of the model and mitigate the over-fitting risk.

- With the Siamese structure, we introduce a parameter-free momentum teacher to obtain instance attention scores more efficiently and stably. Moreover, we employ a consistency-based iterative optimization to improve the discriminability of both models progressively.

2. Related Work

2.1. Multiple Instance Learning in WSI Analysis

Multiple Instance Learning (MIL) [10] has been widely used in WSI analysis with its unique learning paradigm in recent years [17, 22, 26, 35, 40, 42]. MIL is a weakly supervised learning framework that utilizes coarse-grained bag labels for training instead of fine-grained instance annotations. Previous algorithms can be broadly categorized into two groups: instance-level [4, 12, 15, 40] and embedding-level [9, 27, 38, 39, 43]. The former obtain instance labels and aggregate them to obtain the bag label, whereas the latter aggregate all instance features into a high-level bag embedding for bag prediction. Most embedding-level methods share the basic idea of AB-MIL [16], which employs learnable weights to aggregate salient instance features into bag embedding. Furthermore, some MIL frameworks [17, 21, 40, 43] mine more salient instances making classification easier and facilitating classification. For example, Lu *et al*. selected the most salient instances based on their attention scores (e.g., maximum and minimum scores) to compute instance-level loss and improve performance [21]. Zhang *et al*. proposed a class activation map (CAM) based on the AB-MIL paradigm to better mine salient instances and used AB-MIL to aggregate them into bag embedding [43]. In addition, feature clustering methods [27,37,44] computed cluster centroids of all feature embeddings and used representative embeddings for the final prediction. However, all these methods focused excessively on salient instances in training, which are easy instances with high confidence scores and can be easily classified. As a result, they overlook the importance of hard instances for training. In this paper, we intend to mine hard instances for improving WSI classification performance.

2.2. Hard Sample Mining in Computer Vision

Hard sample mining is a popular technique to speed up convergence and enhance the discriminative power of the model in many deep learning areas, such as face recognition [25], object detection [29, 36], person re-identification [1, 28, 33, 34], and deep metric learning [30, 32]. The main idea behind this technique is to select the samples which are hard to classify correctly (i.e., hard negatives and hard positives) for alleviating the imbalance between positive and negative samples and facilitating model training. There are generally three groups of approaches

for evaluating sample difficulty: loss-based [14], similarity-based [7], and learnable weight-based [41]. Typically, these strategies require complete sample supervision information. Drawing on the ideas of the above works, we propose a hard instance mining approach in MIL, mining hard examples at the instance level. In this, there are no complete instance labels, only the bag label is available. Similar to our approach, Li *et al*. utilized attention scores to identify salient instances from false negative bags to serve as hard negative instances and used them to compose the hard bags for improving classification performance [19]. A key difference is that we indirectly mine hard instances by masking out the most salient instances rather than directly locating hard negative instances.

3. Proposed Method

3.1. Background: MIL Formulation

In MIL, any input WSI X is considered as a bag with multiple instances, which can be represented as $X = \{x_i\}_{i=1}^{N}$. x_i is a patch collected from the WSI and considered as the i-th instance of X. N is the number of instances. For a classification task, there exists a known label $Y \in C$ for the bag and an unknown label $y_n \in C$ for each instance, where C is the collection of category labels. The goal of a MIL model $\mathcal{M}(\cdot)$ is to predict the bag label with all instances $\hat{Y} \leftarrow \mathcal{M}(X)$. The popular solution is to learn a bag representation F from the extracted features of instances $Z = \{z_i\}_{i=1}^{N}$ in a bag, which is also referred as the instance aggregation step. And a classifier $\mathcal{C}(\cdot)$, trained upon the F, can be used to predict the bag label $\hat{Y} \leftarrow \mathcal{C}(F)$. There are two ways to aggregate instances for achieving bag embedding. One is the attention-based aggregation [16] denoted as follows,

$$F = \sum_{i=1}^{N} a_i z_i \in \mathbb{R}^D, \tag{1}$$

where a_i is the learnable scalar weight for z_i, and D is the dimension of vector F and z_i. Many works [17, 21, 43] follow this formulation but differ in the ways they generate the attention score a_i.

Another is the multi-head self-attention (MSA) based aggregation [26]. In this fashion, a class token z_0 is embedded with the instance features to get the initial input sequence $Z^0 = [z_0, z_1, \ldots, z_N] \in \mathbb{R}^{(N+1)\times D}$ for aggregating instance features. This can be formulated as,

$$\begin{aligned} \text{head} &= A^{\ell}\left(Z^{\ell-1}W^{V}\right) \in \mathbb{R}^{N\times\frac{D}{H}}, && \ell = 1\ldots L \\ Z^{\ell} &= \text{Concat}\left(\text{head}_1, \cdots, \text{head}_H\right)W^{O}, && \ell = 1\ldots L \end{aligned} \tag{2}$$

where $W^V \in \mathbb{R}^{D\times\frac{D}{H}}$ and $W^O \in \mathbb{R}^{D\times D}$ are the learnable projection matrices of MSA. $A^{\ell} \in \mathbb{R}^{(N+1)\times(N+1)}$ is the attention matrix of the ℓ-th layer, L is the number of MSA block, and H is the number of head in each MSA block. The bag embedding F is the output class token at the final layer,

$$F = Z_0^L. \tag{3}$$

The self-attention-based bag embedding is essentially a special case of attention-based bag embedding in the multi-instance learning setting. Collectively, these approaches can be referred to as the general attention-based MIL method.

3.2. MHIM-MIL for WSI Classification

In general attention-based MIL frameworks, the attention scores of instances indicate the contributions of instances to the bag classification. The salient instances with high scores are useful for classifying WSI in the testing phase but are not conducive to training a MIL model with good generalization ability. Although hard samples have been proven to enhance the generalization ability of the model in many computer vision scenarios [11, 32–34], previous MIL works focus more on exploiting the salient instances and neglecting the utilization of hard instances in model optimization.

In this paper, we propose a simple and efficient MIL framework with Masked Hard Instance Mining (MHIM-MIL) to boost the WSI classification. As illustrated in Figure 2, the MHIM-MIL framework employs a Siamese structure during the training phase. The main component of our framework is a general attention-based MIL model (Student), denoted as $\mathcal{S}(\cdot)$, for aggregating instance features. To increase the discriminatory difficulty of the student model and force it to focus on hard instances, we introduce a momentum teacher, denoted as $\mathcal{T}(\cdot)$, to score the instances with attention weights and then employ some masked hard instance mining strategies to mask the salient instances while preserving the hard instances. After hard instance mining, all the mined features are forwarded into the student model for the inference of the bag label. The teacher shares the same network structure as the student model but does not need gradient-based updates. It is worth mentioning that, due to the varying number of instances within each bag, the non-batch gradient descent algorithm (i.e., SGD with batch size 1) is typically employed to optimize the MIL model. Therefore, compared to the traditional MIL frameworks with two-tier gradient updating models [40, 43], this Siamese structure makes training more stable and efficient with fewer parameters. The proposed framework can be defined as,

$$\hat{Y} = \mathcal{S}\left(\hat{Z}\right) = \mathcal{S}\left(M_{\mathcal{T}}\left(Z\right)\right), \tag{4}$$

where $M_{\mathcal{T}}(\cdot)$ denotes a masked hard instance mining strategy through the teacher model and $\hat{Z}$ are the mined instances.

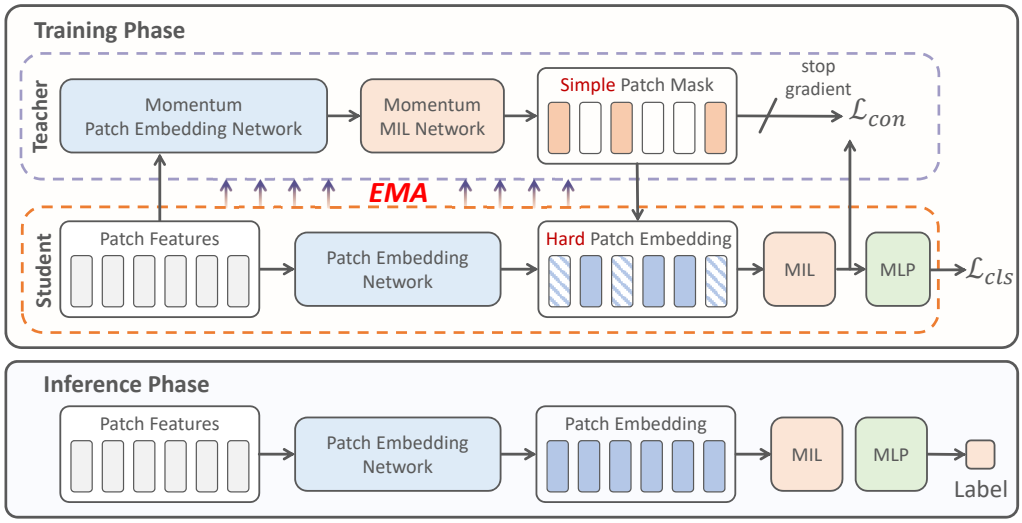

Figure 2: Overview of proposed MHIM-MIL. A momentum teacher is used to compute attention scores for all instances. We mask instances based on attention with hard mining strategies and feed the remaining to the student model. The student is updated by a consistency loss term $\mathcal{L}_{con}$ and a label error loss term $\mathcal{L}_{cls}$. The teacher parameters are updated with an Exponential Moving Average (EMA) of the student parameters without gradient updates. In the inference phase, we use the complete input instances and the student model only.

3.3. Masked Hard Instance Mining Strategy

Conventional hard sample mining strategies are difficult to apply without instance-level supervision. We address this challenge by proposing masked hard instance mining strategies that use attention scores to implicitly mine hard instances by masking out easy instances with high attention scores. More specifically, given a complete sequence of instance features $Z = \{z_i\}_{i=1}^{N}$ as the input of the teacher model $\mathcal{T}(\cdot)$, the teacher outputs the attention weight a_i for each instance as follow,

$$A = [a_1, \ldots, a_i, \ldots, a_N] = \mathcal{T}(Z). \tag{5}$$

Then, we obtain the indices of the attention sequence in descending order by applying a sorting operation on A,

$$I = [i_1, i_2, \ldots, i_N] = \text{Sort}(A), \tag{6}$$

where i_1 is the index of the instance with the highest attention score while i_N is the index of the one with the lowest score. With this index collection I, we will present several masked hard instance mining strategies to select the hard instances. We define an N-dimensional binary vector $M = [m_1, \ldots, m_i, \ldots, m_N]$ for encoding the mask flags of instances where $m_i \in \{0, 1\}$. If $m_i = 1$, the i-th instance is masked, otherwise, it is unmasked.

High Attention Masking: The simplest masked hard instance mining strategy is the High Attention Masking (HAM) strategy, which simply masks instances with the top $\beta_h\%$ highest attention scores. The instance mask flags under HAM are initialized as all zero vectors, $M_h(:) = 0$. Then we collect the indices of the instances whose scores are ranked in the top $\beta_h\%$, $I_h = [i_t]_{t=1}^{\lceil \beta_h\% \times N \rceil}$. Finally, we set the mask flags with these indices, $M_h(I_h) = 1$. To ensure that positive instances are preserved within the unmasked sequences, we also utilized techniques such as mask ratio decay.

Hybrid Masking: We combine HAM with several other instance masking strategies as hybrid masking strategies to achieve some specific properties in hard instance mining, as shown in Figure 3. We consider the obtained mask flags as a collection and employ the union operation for mask flag fusion. We design three hybrid masking strategies as follows:

- **L-HAM**: We use the same pipeline as HAM to generate the mask flags M_l for masking the instances with the top $\beta_l\%$ lowest attention scores in order to filter out the redundant uninformative instances and improve efficiency. To endower this property to HAM, we union the mask flags obtained by two strategies to get the new mask flags, $\hat{M} = M_h \cup M_l$.
- **R-HAM**: Randomness is beneficial to reduce the risk of over-fitting. We generate a random mask flag vector M_r with a given random ratio $\beta_r\%$, and combine it with M_h for introducing the randomness to the hard instance mining, $\hat{M} = M_h \cup M_r$.
- **LR-HAM**: Combining the above strategies, we can obtain completely hybrid mask flags, $\hat{M} = M_h \cup M_r \cup M_l$, which is expected to achieve both of the mentioned desirable properties.

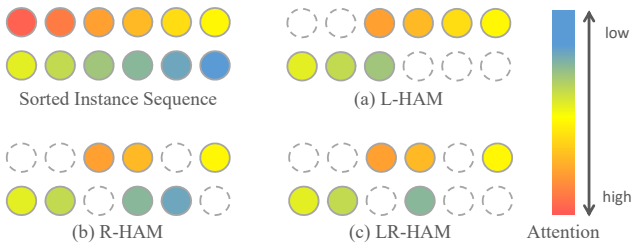

Figure 3: Illustration of proposed hybrid masking strategy for hard instance mining.

Once the final mask flag $\hat{M}$ is produced, the masked instance sequence will be obtained:

$$\hat{Z} = M_{\mathcal{T}}(Z) = \text{Mask}\left(Z, \hat{M}\right) \in \mathbb{R}^{\hat{N} \times D}, \tag{7}$$

where the $\hat{N}$ is the number of unmasked instances.

3.4. Consistency-based Iterative Optimization

Under the Siamese structure, while the teacher model guides the training of the student model, the new knowledge learned by the student model will also update the teacher model. This iterative optimization process progressively improves the mining ability of the teacher and the discriminability of the student. To further facilitate this optimization and explore additional supervised information provided by the momentum teacher, we propose a consistency loss that constrains the classification results of both models.

Student Optimization: There are two losses in student optimization. One is the cross-entropy for measuring the bag label prediction loss,

$$\mathcal{L}_{cls} = Y \log \hat{Y} + (1 - Y) \log\left(1 - \hat{Y}\right). \tag{8}$$

Another is a consistency loss between the bag representation of student F_s and momentum teacher F_t,

$$\mathcal{L}_{con} = -\text{softmax}\left(F_t/\tau\right) \log F_s \tag{9}$$

where the $\tau > 0$ is a temperature parameter. Overall, the final optimization loss is as follows:

$$\{\hat{\theta}_s\} \leftarrow \arg\min_{\theta_s} \mathcal{L} = \mathcal{L}_{cls} + \alpha \mathcal{L}_{con} \tag{10}$$

where θ_s is the parameters of $\mathcal{S}(\cdot)$, and α is scaling factor.

Teacher Optimization: The parameters of momentum teacher θ_t are updated by an exponential moving average (EMA) of the student parameters. The update rule is $\theta_t \leftarrow \lambda\theta_t + (1 - \lambda)\theta_s$, where λ is a hyperparameter. More importantly, the updated teacher is utilized in the next iteration of hard instance mining.

4. Experiments and Results

4.1. Datasets and Evaluation Metrics

CAMELYON-16 [2] is a WSI dataset proposed for metastasis detection in breast cancer. The dataset contains a total of 400 WSIs, which are officially split into 270 for training and 130 for testing, and the testing sample ratio is 13/40≈1/3. Following [6, 21, 44], we adopt 3-times 3-fold cross-validation on this dataset to ensure that each slide is used in training and testing, which can alleviate the impact of data split and random seed on the model evaluation. Each fold has approximately 133 slides. We report the mean and standard deviation of performance metrics over 3 runs.

TCGA Lung Cancer includes two sub-type of cancers, Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). There are diagnostic slides, LUAD with 541 slides from 478 cases, and LUSC with 512 slides from 478 cases. We randomly split the dataset into training, validation, and testing sets with a ratio of 65:10:25 on the patient level. 4-fold cross-validation is adopted, and the mean and standard deviation of performance metrics of the 4 test folders are reported.

We adopt the same data pre-processing as in the CLAM [21]. Following the previous work [21, 26] we leverage Accuracy, Area Under Curve (AUC), and F1-score to evaluate model performance. AUC is the primary performance metric in the binary classification task, and we only report AUC in ablation experiments. Please refer to the **Supplementary Material** for the details of these two datasets.

4.2. Implementation Details

The details on network architectures and training are described in **Supplementary Material**.

4.3. Performance Comparison with Exiting Works

We mainly compare with AB-MIL [16], DSMIL [17], CLAM-SB [21], CLAM-MB [21], TransMIL [26], and DTFD-MIL [43], all of which are attention-based MIL methods. In addition, we compared two traditional MIL pooling operations, Max-pooling and Mean-pooling. Due to the dataset differences, the results of all other methods are reproduced using the official code they provide under the same settings.

As shown in Table 1, max-pooling and mean-pooling perform poorly on two datasets compared to other methods. We attribute this to their insufficient modeling of key instance information. Simple pooling operations are prone to be misled by limited slides that contain numerous instances. This problem is especially severe on the CAMELYON-16 dataset, where the proportion of significant instances is extremely small. For example, max-pooling lags behind DTFD-MIL [43] by 13.87% on AUC. Attention-based

Method	CAMELYON-16			TCGA Lung Cancer		
	Accuracy	AUC	F1-score	Accuracy	AUC	F1-score
Max-pooling	78.95±2.28	81.28±3.74	71.06±2.59	81.49±1.24	86.45±0.71	80.56±1.09
Mean-pooling	76.69±0.20	80.07±0.78	70.41±0.16	84.14±2.97	90.13±2.40	83.39±3.14
AB-MIL [16]	90.06±0.60	94.00±0.83	87.40±1.05	88.03±2.19	93.17±2.05	87.41±2.42
DSMIL [17]	90.17±1.02	94.57±0.40	87.65±1.18	88.32±2.70	93.71±1.82	87.90±2.50
CLAM-SB [21]	90.31±0.12	94.65±0.30	87.89±0.59	87.74±2.22	93.67±1.64	87.36±2.24
CLAM-MB [21]	90.14±0.85	94.70±0.76	88.10±0.63	88.73±1.62	93.69±0.54	88.28±1.58
TransMIL [26]	89.22±2.32	93.51±2.13	85.10±4.33	87.08±1.97	92.51±1.76	86.40±2.08
DTFD-MIL [43]	90.22±0.36	95.15±0.14	87.62±0.59	88.23±2.12	93.83±1.39	87.71±2.04
MHIM-MIL (AB-MIL)	91.81±0.82	96.14±0.52	89.94±0.70	89.64±2.25	94.97±1.72	89.31±2.19
MHIM-MIL (TransMIL)	91.98±0.89	**96.49±0.48**	90.13±1.08	**90.02±2.59**	94.87±2.17	89.65±2.63
MHIM-MIL (DSMIL)	**92.48±0.35**	96.49±0.65	**90.75±0.73**	89.83±3.37	**95.53±1.74**	**89.71±2.92**

Table 1: The performance of different MIL approaches on CAMELYON-16 (C16) and TCGA Lung Cancer (TCGA). The highest performance is in bold. The Accuracy and F1-score are determined by the optimal threshold.

Model	C16	TCGA	Para.	Time	Mem.
AB-MIL	94.00	93.17	657K	**4.0s**	2.4G
CLAM-MB	94.70	93.69	789K	4.3s	2.7G
DTFD-MIL	95.15	93.83	987K	5.2s	**2.1G**
MHIM-MIL	**96.14**	**94.97**	**657K**	4.3s	2.3G
TransMIL	93.51	92.51	2.67M	13.1s	10.6G
MHIM-MIL	**96.49**	**94.87**	**2.67M**	**10.1s**	**5.5G**

Table 2: Comparison of time and memory requirements of different MIL methods. We report the model size (Para.), the training time per epoch (Time), and the peak memory usage (Mem.) on the CAMELYON-16 dataset (C16).

methods achieve better performance on both datasets by focusing on salient instances. In particular, the representative MIL framework DTFD-MIL [43] benefits from the further exploration of significant instances and achieves the second-best performance on both datasets (95.15% AUC on CAMELYON-16 and 93.83% AUC on TCGA). However, it also suffers from overemphasizing salient instances during training, which limits its generalization. Our proposed MHIM-MIL achieves significant performance improvement on both datasets (+1.34% AUC on CAMELYON-16 and +1.70% AUC on TCGA) by mining hard instances during training, breaking the performance bottleneck. It is worth mentioning that we validate our framework on three representative MIL models, both of which can outperform the existing MIL methods.

4.4. Computational Cost Analysis

In this section, we report the training time and GPU memory requirements for running different MIL models on a 3090 GPU. The upper part of Table 2 compares some MIL frameworks that use AB-MIL [16] as a baseline. We observe that traditional MIL frameworks typically introduce additional parameters and reduce efficiency due to their complex structures. For example, the state-of-the-art framework DTFD-MIL [43] increases the parameter size by nearly twice (657K vs. 987K) and the training time by 30%. In contrast, MHIM-MIL achieves the most significant performance improvement with almost no extra computational cost due to the momentum teacher. Moreover, existing Transformer-based MIL methods are usually plagued by high computing costs due to their large number of parameters and self-attention operations. For instance, TransMIL [26], which first applies a pure Transformer MIL model to solve WSI classification problems, has 4× more parameters than AB-MIL, 3× longer training time, and almost 4.5× higher memory consumption. Furthermore, the extremely long input sequences in WSI classification degrade the stability of such complex structures (2.13% AUC standard deviation on C16, which is the highest among all embedding-level MIL methods). With the masked hard instance mining strategy, the MHIM-MIL framework significantly reduces the computational cost (-24% training time and -48% memory usage) and enhances its stability (0.48% AUC standard deviation on C16). More details are provided in **Supplementary Material**.

4.5. Ablation Study

4.5.1 Importance of the Different Components

Table 3 shows the effect of different modules in MHIM-MIL on two datasets. The baseline methods are two representative attention-based MIL methods, namely AB-MIL [16] and TransMIL [26]. First, we introduce the naive masked hard instance mining strategy, which leverages the model itself to mine hard instances during training. This strategy improves AUC by 1.86% and 2.55% for the two MIL models on CAMELYON-16 respectively, indicating that concentrating on hard instances during training can as-

Module	CAMELYON-16		TCGA	
	AB.	Trans.	AB.	Trans.
Baseline	94.00	93.51	93.17	92.51
+MHIM	95.86	96.06	94.14	93.75
+MHIM+Siam.	95.82	96.24	94.55	94.13
+MHIM+Siam.+Con.	**96.14**	**96.49**	**94.97**	**94.87**

Table 3: The effect of different components in MHIM-MIL with two MIL models: AB-MIL (AB.) and TransMIL (Trans.). MHIM denotes the masked hard instance mining strategy. Siam. refers to the Siamese framework. Con. represents consistency loss.

Strategy	CAMELYON-16		TCGA	
	AB.	Trans.	AB.	Trans.
Baseline	94.00	93.51	93.17	92.51
HAM	95.68	95.90	93.83	94.54
R-HAM	**96.14**	95.88	94.79	94.60
L-HAM	95.81	**96.49**	94.33	94.67
LR-HAM	95.92	96.33	**94.97**	**94.87**

Table 4: Comparison between different masked hard instance mining strategies. The three hybrid strategies show varying performance across the benchmarks.

sist mainstream MIL models in building better classification boundaries. Further discussion on the masked hard instance mining strategy is presented in Section 4.5.2. Compared with the naive MHIM strategy, the third row of the table suggests that a Siamese structure [3,5,8] based on a momentum teacher is beneficial for more stable and effective mining of hard instances. We elaborate more on choosing the teacher model in Section 4.5.3. After adding consistency loss term to the objective function, our full MHIM-MIL framework achieves the best performance (96.49% AUC on CAMELYON-16 and 94 .97% AUC on TCGA). For subsequent ablation experiments, we include consistency loss by default to facilitate the optimization of our framework.

4.5.2 Impact of the Different MHIM Strategies

The masked hard instance mining strategy is the core design of our method. The main idea of this strategy is masking the most salient instances to indirectly mine hard instances to facilitate model training. Based on this idea, we devise three hybrid strategies (R-HAM, L-HAM, and LR-HAM) and present their impact in Table 4. The basic strategy, High Attention Masking (HAM), already boosts performance significantly, leading to AUC improvements of 1.68% and 2.39% for two MIL models on the CAMELYON-16 dataset, respectively. After introducing the other two strategies, different MIL models achieve performance improvements

Teacher	CAMELYON-16		TCGA	
	AB.	Trans.	AB.	Trans.
Baseline	94.00	93.51	93.17	92.51
Student copy	95.84	95.86	93.68	93.45
Init.	95.88	96.12	94.66	94.15
Momentum	95.96	96.11	94.65	94.45
Init.+Momentum	**96.14**	**96.49**	**94.97**	**94.87**

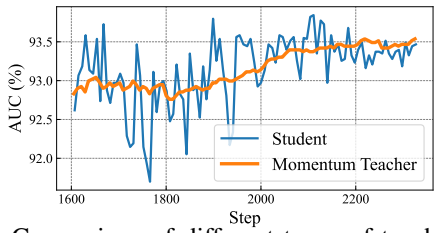

Table 5: Comparison of different types of teachers. Momentum denotes the teacher is updated by EMA strategy. Init. indicates the initialization of the teacher with pre-trained parameters. The bottom figure compares the stability of the momentum teacher and the non-batch gradient updated student during training.

on both datasets. Specifically, AB-MIL [16] shows more significant performance gains after introducing randomness (96.14% AUC on CAMELYON-16 with R-HAM) due to its better ability to filter out redundant information, while TransMIL [26] shows the reverse trend (96.49% AUC on CAMELYON-16 with L-HAM). Furthermore, the more complex three-hybrid strategy (LR-HAM) achieves the best performance on the TCGA dataset, which has a larger proportion of positive areas and more instances. Overall, our experiments validate the effectiveness of masked hard instance mining strategy, and the diversity of proposed strategies improves its applicability to different datasets and MIL models.

4.5.3 Impact of the Choice of Teacher Network

In MHIM-MIL, we employ a Teacher model to mine hard instances and facilitate training of the Student model. In Table 5, we comprehensively investigate the effects of various choices of Teacher network. First, we utilize a single-model structure, which treats the Student model as the Teacher. The student conducts masked hard instance mining prior to training. Due to the non-batch gradient update, the unstable performance of the Student model makes the strategy susceptible to noise, so the performance is not optimal. Second, we adopt a momentum teacher, which shares the same network structure as the Student model and is updated with the EMA strategy. This updating strategy enhanced the stability of momentum teachers, as shown in the figure below, and enabled MHIM-MIL to achieve 0.97% and 1.00% performance improvement in TCGA under the two MIL mod-

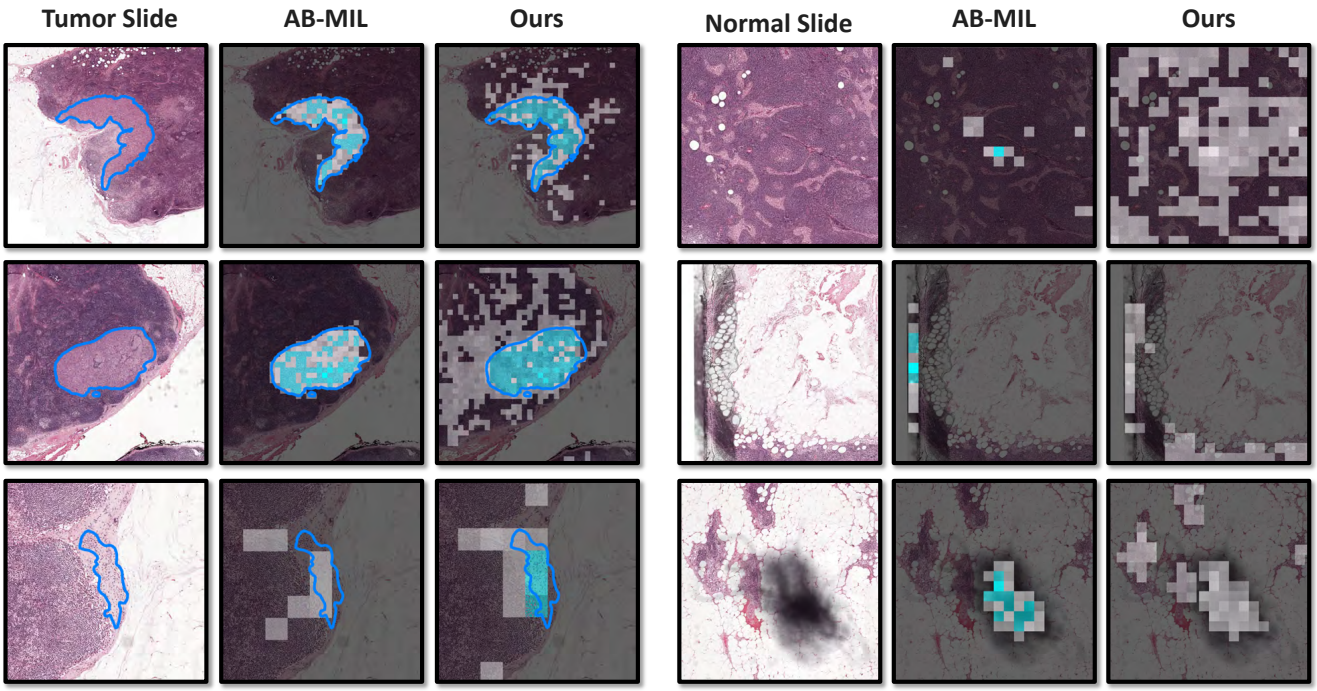

Figure 4: Patch visualization produced by AB-MIL [16] (baseline) and MHIM-MIL. The blue lines outline the tumor regions. The brighter patch indicates higher attention scores. The cyan colors indicate high probabilities of being tumor for the corresponding locations. Ideally, the cyan patches should cover only the area within the blue lines. We show that focusing only on more salient regions reduces the generalization ability of the model and that hard instances can provide useful information for more accurate and comprehensive judgments.

els, respectively. With proper initialization, the momentum teacher achieves the best performance. However, a fixed initialization teacher fails to learn new knowledge, which emphasizes the significance of iterative optimization.

4.6. Visualization

To more intuitively understand the effect of the masked hard instance mining, we visualize the attention scores (bright patch) and tumor probabilities (cyan patch) of patches produced by AB-MIL and MHIM-MIL, as illustrated in Figure 4. Here, MHIM-MIL employs AB-MIL as its baseline model. We note that attention scores only indicate the regions of interest of models and are infeasible to reflect tumor probabilities [17, 43]. First, as shown in Figure 4, AB-MIL often assigns high tumor probabilities to patches in non-tumor areas. We attribute this phenomenon to the low generalization capability of conventional attention-based MIL models, which tend to focus only on salient regions during training. In contrast, MHIM-MIL trained with hard instances shows a much better generalization ability than the baseline model for noise robustness (rows 2 and 3 on the right) and for precise detection of challenging subtle tumor areas (row 3 on the left). More significantly, we find that focusing only on tumor areas leads to missing most of them, expanding the view to include some "irrelevant areas" enables the model to make more complete judgments (rows 1 and 2 on the left). This phenomenon demonstrates how hard instances provide more useful information to help the model make more accurate and comprehensive judgments. We provide more details and an in-depth analysis of this patch visualization in **Supplementary Material**.

5. Conclusion

This paper rethinks the impact of salient instances for MIL-based WSI classification algorithms. We demonstrate that attention-based MIL methods excessively prioritizing salient instances harm the generalization ability of the model. To address this issue, we have proposed several masked hard instance mining strategies that mask out salient patches and encourage the model to attend to informative regions for better discriminative learning. Through qualitative analysis, we have demonstrated that these strategies effectively alleviate the under-fitting problem of general AB-MIL to hard instances. We have also developed the MHIM-MIL framework that leverages momentum teacher and consistency loss to further enhance hard instance mining. Our experimental results demonstrate the superiority and generality of the MHIM-MIL framework over other lat-

est methods. In future work, we plan to devise a more precise localization scheme for hard instances that can facilitate model training and convergence.

6. Acknowledgement

Reported research is partly supported by the National Natural Science Foundation of China under Grant 62176030, and the Natural Science Foundation of Chongqing under Grant cstc2021jcyj-msxmX0568.

References

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3908–3916, 2015. 2

[2] 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. 5

[3] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993. 2, 7

[4] 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. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 7

[6] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 1, 5

[7] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017. 3

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 2, 7

[9] Philip Chikontwe, Miguel Luna, Myeongkyun Kang, Kyung Soo Hong, June Hong Ahn, and Sang Hyun Park. Dual attention multiple instance learning with unsupervised complementary loss for covid-19 screening. *Medical Image Analysis*, 72:102105, 2021. 1, 2

[10] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997. 1, 2

[11] Qi Dong, Shaogang Gong, and Xiatian Zhu. Class rectification hard mining for imbalanced deep learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1851–1860, 2017. 3

[12] Ji Feng and Zhi-Hua Zhou. Deep miml network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. 2

[13] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998. 2

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3

[15] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2424–2433, 2016. 2

[16] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1, 2, 3, 5, 6, 7, 8

[17] 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, 8

[18] Hang Li, Fan Yang, Yu Zhao, Xiaohan Xing, Jun Zhang, Mingxuan Gao, Junzhou Huang, Liansheng Wang, and Jianhua Yao. Dt-mil: Deformable transformer for multi-instance learning on histopathological image. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 206–216. Springer, 2021. 1

[19] Meng Li, Lin Wu, Arnold Wiliem, Kun Zhao, Teng Zhang, and Brian Lovell. Deep instance-level hard negative mining model for histopathology images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 514–522. Springer, 2019. 3

[20] Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, Melissa Zhao, Maha Shady, Jana Lipkova, and Faisal Mahmood. Ai-based pathology predicts origins for cancers of unknown primary. *Nature*, 594(7861):106–110, 2021. 1

[21] 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

[22] Siyamalan Manivannan, Caroline Cobb, Stephen Burgess, and Emanuele Trucco. Subcategory classifiers for multiple-instance learning and its application to retinal nerve fiber layer visibility classification. *IEEE Transactions on Medical Imaging*, 36(5):1140–1150, 2017. 2

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

[24] Hans Pinckaers, Bram Van Ginneken, and Geert Litjens. Streaming convolutional neural networks for end-to-end learning with multi-megapixel images. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1581–1590, 2020. 1

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[26] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *NeurIPS*, 34, 2021. 1, 2, 3, 5, 6, 7

[27] Yash Sharma, Aman Shrivastava, Lubaina Ehsan, Christopher A Moskaluk, Sana Syed, and Donald E Brown. Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification. *arXiv preprint arXiv:2103.10626*, 2021. 2

[28] Hao Sheng, Yanwei Zheng, Wei Ke, Dongxiao Yu, Xiuzhen Cheng, Weifeng Lyu, and Zhang Xiong. Mining hard samples globally and efficiently for person reidentification. *IEEE Internet of Things Journal*, 7(10):9611–9622, 2020. 2

[29] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016. 2

[30] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016. 2

[31] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67:101813, 2021. 1

[32] Yumin Suh, Bohyung Han, Wonsik Kim, and Kyoung Mu Lee. Stochastic class-based hard example mining for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7251–7259, 2019. 2, 3

[33] Han Sun, Zhiyuan Chen, Shiyang Yan, and Lin Xu. Mvp matching: A maximum-value perfect matching for mining hard samples, with application to person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6737–6747, 2019. 2, 3

[34] Zichang Tan, Ajian Liu, Jun Wan, Hao Liu, Zhen Lei, Guodong Guo, and Stan Z Li. Cross-batch hard example mining with pseudo large batch for id vs. spot face recognition. *IEEE Transactions on Image Processing*, 31:3224–3235, 2022. 2, 3

[35] Tong Tong, Robin Wolz, Qinquan Gao, Ricardo Guerrero, Joseph V Hajnal, Daniel Rueckert, Alzheimer's Disease Neuroimaging Initiative, et al. Multiple instance learning for classification of dementia in brain mri. *Medical Image Analysis*, 18(5):808–818, 2014. 2

[36] Keze Wang, Xiaopeng Yan, Dongyu Zhang, Lei Zhang, and Liang Lin. Towards human-machine cooperation: Self-supervised sample mining for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1605–1613, 2018. 2

[37] Xi Wang, Hao Chen, Caixia Gan, Huangjing Lin, Qi Dou, Efstratios Tsougenis, Qitao Huang, Muyan Cai, and Pheng-Ann Heng. Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE transactions on cybernetics*, 50(9):3950–3962, 2019. 2

[38] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018. 2

[39] Yunan Wu, Arne Schmidt, Enrique Hernández-Sánchez, Rafael Molina, and Aggelos K Katsaggelos. Combining attention-based multiple instance learning and gaussian processes for ct hemorrhage detection. In *MICCAI*, pages 582–591. Springer, 2021. 2

[40] Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, and Wei Xu. Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on computer vision*, pages 10682–10691, 2019. 1, 2, 3

[41] Lin Xu, Han Sun, and Yuai Liu. Learning with batch-wise optimal transport loss for 3d shape recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3333–3342, 2019. 3

[42] Yan Xu, Jun-Yan Zhu, I Eric, Chao Chang, Maode Lai, and Zhuowen Tu. Weakly supervised histopathology cancer image segmentation and classification. *Medical Image Analysis*, 18(3):591–604, 2014. 2

[43] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18802–18812, 2022. 1, 2, 3, 5, 6, 8

[44] Xiaoxian Zhang, Sheng Huang, Yi Zhang, Xiaohong Zhang, Mingchen Gao, and Liu Chen. Dual space multiple instance representative learning for medical image classification. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. 2, 5

[45] Yu Zhao, Zhenyu Lin, Kai Sun, Yidan Zhang, Junzhou Huang, Liansheng Wang, and Jianhua Yao. Setmil: spatial encoding transformer-based multiple instance learning for pathological image analysis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II*, pages 66–76. Springer, 2022. 1