

DGR-MIL: Exploring Diverse Global Representation in Multiple Instance Learning for Whole Slide Image Classification

Wenhui Zhu^{1*†}, Xiwen Chen^{2*}, Peijie Qiu^{3*},
Aristeidis Sotiras³, Abolfazl Razi², and Yalin Wang¹

¹ Arizona State University, AZ, USA

{wzhu59,ylwang}@asu.edu

² Clemson University, SC, USA

xiwenc@g.clemson.edu, arazi@clemson.edu

³ Washington University in St. Louis, MO, USA

{peijie.qiu,aristeidis.sotiras}@wustl.edu

Abstract. Multiple instance learning (MIL) stands as a powerful approach in weakly supervised learning, regularly employed in histological whole slide image (WSI) classification for detecting tumorous lesions. However, existing mainstream MIL methods focus on modeling correlation between instances while overlooking the inherent diversity among instances. However, few MIL methods have aimed at diversity modeling, which empirically show inferior performance but with a high computational cost. To bridge this gap, we propose a novel MIL aggregation method based on diverse global representation (DGR-MIL), by modeling diversity among instances through a set of global vectors that serve as a summary of all instances. First, we turn the instance correlation into the similarity between instance embeddings and the pre-defined global vectors through a cross-attention mechanism. This stems from the fact that similar instance embeddings typically would result in a higher correlation with a certain global vector. Second, we propose two mechanisms to enforce the diversity among the global vectors to be more descriptive of the entire bag: (i) positive instance alignment and (ii) a novel, efficient, and theoretically guaranteed diversification learning paradigm. Specifically, the positive instance alignment module encourages the global vectors to align with the center of positive instances (e.g., instances containing tumors in WSI). To further diversify the global representations, we propose a novel diversification learning paradigm leveraging the determinantal point process. The proposed model outperforms the state-of-the-art MIL aggregation models by a substantial margin on the CAMELYON-16 and the TCGA-lung cancer datasets. The code is available at <https://github.com/ChongQingNoSubway/DGR-MIL>.

Keywords: Weakly-supervised learning · Multiple Instance Learning · Histological Whole Slide Image · Transformer

* These authors contributed equally to this paper.

† Corresponding author

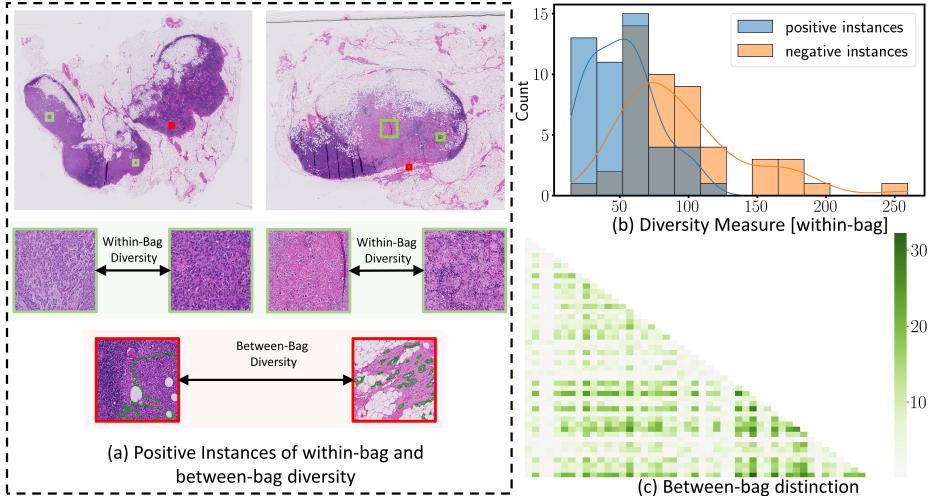


Fig. 1: (a) Examples of positive instances of with-bag and between-bag diversities measured by rate-distortion theory. (b) Histogram of the diversity measure within positive bags on the CAMELYON16 dataset. (c) The between-bag distinction measures the pair-wise similarity between bags.

1 Introduction

Histological whole slide images (WSIs) are commonly used to diagnose a variety of cancers, e.g., breast cancer, lung cancer, etc. [46]. However, the gigapixel resolution of WSIs hinders the direct translation of classic deep learning methods into WSI applications mainly due to computational intractability [4, 11, 35, 38]. Therefore, the analysis of WSIs typically starts with cropping images into small patches and then performing analysis on a per-patch basis. In addition, the absence of labor-intensive pixel/patch-level annotations poses a significant challenge for the precise localization of targets of interest (e.g., tumors in WSIs) in a fully supervised setting. As a result, Multiple Instance Learning (MIL), a weakly supervised method, is commonly employed in WSI analyses by treating an entire WSI as a bag and the cropped patches as instances.

The prevailing MIL models in analyzing WSIs have been built upon the attention-based MIL (AB-MIL) framework [28] since its introduction. However, the standard AB-MIL treats each instance independently and does not take the correlations between instances into account. Although many of its follow-ups address this challenge by a variety of means [30, 47, 58, 64], they mainly focus on modeling the correlation between instances by assigning high correlations to instances from the same category (e.g., tumor instances). However, even instances from the same category exhibit variations in phenotype, size, as well as spatial diversity marked by immune infiltration across different patients [7, 37, 66]. For example, negative instances close to the tumor boundaries typically resemble positive instances while appearing differently compared to the other negative instances [24]. As a result, instances belonging to the same category may not

be assigned high correlations; similarly, instances from different categories could also receive high correlations. This spurious correlation between instances is prone to trap the MIL model by incorrectly aggregating them when making predictions. Formally, we quantify the diversity of instances between and within bags in WSIs by leveraging the rate-distortion theory [12, 15, 63], where a higher rate indicates a less compressible but more diverse collection of samples (see details of computing the diversity measure in Appendix A). As consistent with findings in pathology, we observe that both positive and negative instances in WSIs exhibit between-bag and within-bag diversity (refer to Fig. 1). Based on this fact, we argue that the diversity of instances is important in designing MIL models. Before that, clustering/prototype-based MIL methods tried to solve the diversity by utilizing attention scores as pseudo labels to provide instance-level supervision [55, 61]. This introduces a chicken-and-egg issue. The effectiveness of pseudo-labels relies on successful MIL classification pooling, which in turn depends on precise attention localization. Especially when patch representations are inferior or MIL initially guided by poor pseudo label, leading to even misleading localization and unstable optimization [32, 68]. Among them, PMIL presents an alternative method to avoiding noise attention [62], initially selecting prototypes through clustering, followed by modeling diversity via prototype and patch representation. However, the design of the multi-stage framework empirically leads to suboptimal learning outcomes, and the restricted number of prototypes, due to high computational burden, results in diminished diversity.

To this end, we propose to jointly model this diversity through a set of learnable global vectors. The learned global vectors summarize diverse instances of interest (e.g., tumors in WSIs). As a result, the diversity between instances can be implicitly modeled by computing the correlation between instance embeddings and the global vectors through a cross-attention mechanism. To enhance the ability of the global vectors to capture the most discriminative global context for WSI classification, we introduce the concept of tokenized global vectors. It is worth mentioning that the importance map for instances can be calculated based on the attention between the tokenized global vector and the embedding of each individual instance. To learn diverse global vectors, we propose two main strategies. First, we push the global vectors toward the centers of the positive bag by a positive instance alignment mechanism. Second, we propose a low-complexity and theoretically guaranteed diversity loss to enforce the orthogonality between the global vectors by utilizing the linear algebra property of the determinantal point process (DPP). In this paper, we explore the design of diverse global representation in the MIL model to model the diversity of instances in WSI. The main contributions are four-fold: (i) We introduce a new perspective on modeling the diversity of instances in WSI. (ii) We further propose a novel MIL aggregation model, termed DGR-MIL, to model diversity in MIL through a set of learnable global vectors. (iii) To learn a diverse global representation (vectors), we propose two main mechanisms: positive instance alignment and a novel diversity loss. (iv) Experimental results on two WSI benchmarks demonstrate the proposed DGR-MIL outperforms other competing MIL aggregation methods.

2 Related Work

2.1 Multiple instance learning in WSIs

MIL has been widely applied in many fields, e.g., pathology [28, 30, 47, 64], video analysis [2, 42], time series [14, 21]. In particular, the applications of the MIL in Whole Slide Image classification can be roughly summarized into two sub-categories: i) instance-based MIL [22, 27, 60] and ii) bag embedding-based MIL. Instance-based methods typically require the propagation of the bag-level label to each of its instances to train the model. Consequently, the final bag-level prediction is obtained by aggregating instance-level predictions. However, empirical studies have proven its performance inferior to the embedding-based competitors because of the noisy instance-level supervision [54]. In contrast, bag-embedding-based methods start by projecting instances into feature embeddings and subsequently aggregate the information of these embeddings to obtain the bag-level prediction. Since the introduction of attention-based MIL (AB-MIL) [28], the prevailing applications of bag embedding-based MIL in WSI analysis have revolved around this framework. However, AB-MIL operates under the assumption that all instances within a bag are independent and identically distributed while failing to uncover inter-instance correlations. Therefore, numerous of its follow-up works centered around mitigating this limitation by taking advantage of non-local attention mechanism [30], transformer [47], pseudo bags [64], sparse coding [40], and low-rank constraints [58].

Most existing mainstream MIL methods have modeled correlations mainly through similarity between instances. However, they did not consider the variability of instances between and within bags. Conversely, clustering/prototype-based MIL employs attention scores for selecting prototypes [55, 61], potentially introducing noise and misleading model decisions [32, 68]. Unlike attention-guided methods, PMIL [62] suggests a two-stage framework that first leverages clustering to identify reference prototypes and capture the sub-cluster representation among patch instances and prototypes. However, unrestricted optimization in prototype selection can easily lead to suboptimal outcomes, and a limited number of prototypes can result in a loss of diversity (limited by computational resources). In this paper, we explicitly model the diversity among instances in bag-embedding-based MIL through a learnable global representation. Although the proposed method falls into the category of transformer-based MILs, it differs from the previous transformer-based MILs [47, 58] in two main aspects. First, we model the diversity between instances by comparing instances to the proposed global vectors via a cross-attention mechanism. Second, we propose a tokenized global vector to summarize the context information of positive instances.

2.2 Transformer

The transformer [51] has been widely applied in computer vision [9, 20, 33, 52], time series modeling [57, 67], and the natural language processing fields [18, 43,

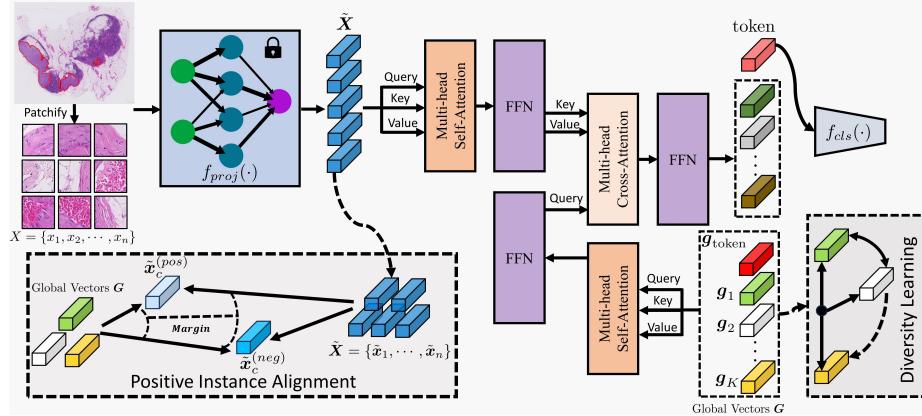


Fig. 2: Overview of the proposed DGR-MIL where the global vectors are used for modeling the diversity of instances. The diverse global vectors are learned through the positive instance alignment module and the diversity learning mechanism.

[44]. Standard transformers discover contextually relevant information by modeling the correlation between elements within a sequence through the self-attention mechanism. However, the traditional self-attention operation has quadratic time and space complexity $\mathcal{O}(n^2)$, with respect to a sequence containing n elements. In the context of MIL, sequence length typically becomes quite large since one bag often approximately comprises ten thousand instances. This extremely long sequence poses significant computational intractability. Although [23, 48, 53] demonstrate that proper approximation of standard self-attention can reduce its quadratic complexity to linear, it still struggles to capture extremely long-term dependencies of context [6, 45, 58]. In contrast, the cross-attention mechanism [49, 52], which was originally proposed to relate positions from one sequence to another, allows models to consider cross-sequence information. Inspired by this, we propose to model the diversity between and among instances through a cross-attention between instances and the proposed global vectors (see details in Section 3.1). This dramatically reduces the complexity compared to the self-attention mechanism (see Appendix C for details of model complexity) since the number of global vectors is significantly less than the sequence length.

3 Methods

The proposed DGR-MIL comprises two main parts: i) the design of the global representation in MIL pooling (Section 3.1), and ii) the strategy of learning diverse global representation (Section 3.2), where we further propose positive instance alignment and a computational-efficient diversity loss with a theoretical guarantee. The entire framework of DGR-MIL is depicted in Fig. 2.

Preliminary. Without loss of generality, we take binary MIL classification as an example: The objective is to predict the bag-level label $Y \in \{0, 1\}$, given a

bag of instances $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$, denoting a WSI with n tiled patches. However, the corresponding instance-level labels $\{y_i\}_{i=1}^n$ are unknown in most WSI analyses due to the laboriousness of obtaining patch-level annotations. This turns the WSI classification into a weakly-supervised learning scheme according to the standard MIL formulation:

$$Y = \begin{cases} 0, & \text{iff } \sum_i y_i = 0 \\ 1, & \text{otherwise.} \end{cases} \quad (1)$$

Because of the gigapixel resolution of WSIs, MIL typically cannot be performed in an end-to-end fashion [8, 34, 35] and instead necessitates a simplified learning scheme. This simplified MIL learning process comprises three main parts: i) a pre-trained feature extractor $f_{proj}(\cdot)$ that projects each instance into a L -dimensional vector, ii) a MIL pooling operator $\sigma(\cdot)$ that combines instance-level embeddings into a bag-level feature, and iii) a bag-level classifier $f_{cls}(\cdot)$ that takes the bag-level feature as input and produces the bag-level prediction as output. Mathematically, this process is given by

$$\begin{aligned} \hat{Y} &= f_{cls}(\sigma(\{\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n\})), \quad \tilde{\mathbf{x}}_i \in \mathbb{R}^L \\ \text{with } \{\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n\} &= f_{proj}(\{\mathbf{x}_1, \dots, \mathbf{x}_n\}), \end{aligned} \quad (2)$$

where \hat{Y} denotes the predicted bag-level label. In the attention-based MIL (AB-MIL) [28] framework, the typical formulation for the MIL pooling operator is as follows:

$$\sigma(\tilde{\mathbf{x}}_i) = \frac{\exp\{\mathbf{W}^T(\tanh(\mathbf{V}\tilde{\mathbf{x}}_i)) \odot \text{sigm}(\mathbf{U}\tilde{\mathbf{x}}_i)\}}{\sum_{i=1}^n \exp\{\mathbf{W}^T(\tanh(\mathbf{V}\tilde{\mathbf{x}}_i)) \odot \text{sigm}(\mathbf{U}\tilde{\mathbf{x}}_i)\}}, \quad (3)$$

where \mathbf{W} , \mathbf{V} , and \mathbf{U} are learnable parameters.

3.1 Global Representation in MIL Pooling

To accommodate the variability of the target lesions within and between bags, we develop a diverse global representation in the MIL pooling stage. Specifically, we define the global representation of the target (positive) instances as a set of learnable vectors given by $\mathbf{G} = [\mathbf{g}_1^\top, \dots, \mathbf{g}_K^\top] \in \mathbb{R}^{K \times L}$ with $\mathbf{g}_k \in \mathbb{R}^L$ where K is the number of global vectors. It is worth noting that a feed-forward network (FFN) is used to embed further both the input instance vectors $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_i\}_{i=1}^n$ and the global vectors \mathbf{G} (see Fig. 2). However, we keep using $\mathbf{G} \in \mathbb{R}^{K \times L}$ to denote global vectors for notation brevity.

Instance Correlation as Cross Attention. The standard AB-MIL framework assumes the instances are independent and identically distributed while overlooking the correlation effect between instances. Hence, the self-attention mechanism becomes a natural choice for modeling the inter-instance correlation. However, due to the large number of instances within a bag in MIL, the

quadratic time and space complexity $\mathcal{O}(n^2)$ of standard self-attention poses a significant challenge in computation. Alternatively, the previous transformer-based MIL [47] mitigates this problem by employing Nystrom-Attention [59], approximating the standard self-attention with linear complexity, which has proved effective of modeling correlation between positive and negative instances. It could be used to gather similar instances together by attention, benefiting from filtering background information. However, self-attention usage only guarantees the general separation of the positive and negative instances in a bag, which overlooks the diversity between instances and between bags.

Here, we implicitly model the diversity between instances by comparing the similarity between each instance vector and the proposed diverse global vectors. Specifically, this is achieved through a cross-attention mechanism where the global vector \mathbf{G} serves as queries, and a bag of instance vectors $\tilde{\mathbf{X}}$ is used as key-value pairs. Formally, the h -th head of the proposed cross attention is given by

$$\begin{aligned} \text{head}_h(\mathbf{G}, \tilde{\mathbf{X}}) &= \text{Attention}(\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h) \\ \mathbf{Q}_h &= \mathbf{G}\mathbf{W}_h^Q, \quad \mathbf{K}_h = \tilde{\mathbf{X}}\mathbf{W}_h^K, \quad \mathbf{V}_h = \tilde{\mathbf{X}}\mathbf{W}_h^V, \end{aligned} \quad (4)$$

where $\mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V \in \mathbb{R}^{L \times L/H}$ are learnable parameters for linear projections, where H is number of heads. For the derivation purposes, we follow the traditional definition of the attention mechanism in the transformer (i.e., $\text{Attention}(\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h) = \text{softmax}(\mathbf{Q}_h \mathbf{K}_h^\top / \sqrt{d_k}) \mathbf{V}_h$). The output of the yielding multi-head cross attention (MHCA) is the concatenation of the outputs from all heads through a linear projection:

$$\text{MHCA}(\mathbf{G}, \tilde{\mathbf{X}}) = \text{concat}(\text{head}_1; \dots; \text{head}_H)\mathbf{W}^O, \quad (5)$$

where $\mathbf{W}^O \in \mathbb{R}^{L \times L}$ is a trainable parameter. The proposed cross-attention mechanism reduces the quadratic time and space complexity $\mathcal{O}(n^2)$ in the standard self-attention mechanism to linear $\mathcal{O}(Kn)$ where $K \ll n$. In practice, we applied the Nystrom-Attention to the instance vectors and global vectors before performing the cross-attention (see Fig. 2) for two main reasons. First, applying self-attention to input instance vectors can facilitate filtering out the background. Second, applying self-attention to the global vectors can increase their discrepancies.

Tokenized Global Vector. The vision transformer includes a class token to encode the globally discriminative representation associated with certain labels in image classification tasks. This token is typically added to the input token embedding by serving as a summary of the entire image. Building upon this inspiration, we propose to add a tokenized global vector $\mathbf{g}_{\text{token}}$ as a summary of all the other global vectors. Now, the yielding global vectors can be denoted as $\tilde{\mathbf{G}} = \{\mathbf{g}_{\text{token}}, \mathbf{g}_1, \dots, \mathbf{g}_K\} \in \mathbb{R}^{(K+1) \times L}$. The output of the tokenized global

vectors after the cross-attention layer (Eq.(5)) is then used for bag-level classification. Following the convention in AB-MIL, the yielded importance score of each instance can be computed as

$$\sigma(\tilde{x}_i) = \text{softmax} \left(\frac{(\mathbf{g}_{\text{token}} \mathbf{W}_h^Q)(\tilde{x}_i \mathbf{W}_h^K)^\top}{\sqrt{d_k}} \right). \quad (6)$$

At first glance, adding the token to the global vectors instead of the input instance embedding appears counterintuitive. However, an in-depth analysis reveals its favorable properties. The proposed global vectors are learned in an unsupervised way (see details in Section 3.2), which poses a significant challenge in perfectly eliminating information from negative instances in the global vectors. This may be attributed to the similarity between positive instances and their adjacent negative instances, as tumor-adjacent regions typically exhibit high-density, quantitative expression in the spatial relationships of cells [24]. Each diverse global vector encapsulates a collection of analogous tissue features. As a result, certain global vectors emphasize certain types of positive instances. Accordingly, adding tokenized global vectors facilitates the model to capture the most discriminative global representation while suppressing the information from the negative instances (as evident in Fig. 5(b)).

3.2 Learning Diverse Global Representation

Due to the weakly-supervised nature of MIL, how to learn the global representation of the target of interest remains an open problem. In this section, we introduce two strategies that can be used to learn a reliable and diverse global representation in MIL, respectively: i) positive instance alignment and ii) diversity learning via utilizing the linear algebra property of the DPP.

Positive Instance Alignment. To enforce that the global representation aligns with the instances of interest (i.e., positive instances), we push the global vectors toward the positive bag centers but away from the negative bag centers. To do so, we first define the center of the positive and negative bags as $\tilde{\mathbf{x}}_c^{(pos)} \in \mathbb{R}^L$ and $\tilde{\mathbf{x}}_c^{(neg)} \in \mathbb{R}^L$, respectively. Similar to [25], the positive and negative centers are then updated in a momentum fashion at each training iteration:

$$\begin{aligned} \tilde{\mathbf{x}}_c^{(pos)} &= m\tilde{\mathbf{x}}_c^{(pos)} + (1 - m) \frac{1}{|\mathcal{I}_{pos}|} \sum_{i \in \mathcal{I}_{pos}} \tilde{x}_i \\ \tilde{\mathbf{x}}_c^{(neg)} &= m\tilde{\mathbf{x}}_c^{(neg)} + (1 - m) \frac{1}{|\mathcal{I}_{neg}|} \sum_{i \in \mathcal{I}_{neg}} \tilde{x}_i, \end{aligned} \quad (7)$$

where m denotes the momentum update rate, which is set empirically to 0.4. \mathcal{I}_{pos} and \mathcal{I}_{neg} are the index sets of positive bags and negative bags, respectively. This indicates that the update of the positive instance center occurs only if a

positive bag is fed into the network. The same strategy is applied to the negative center update (i.e., updated if and only if a negative bag is encountered). Up to now, we can formulate a set of triplet $\{\mathbf{G}, \tilde{x}_c^{(pos)}, \tilde{x}_c^{(neg)}\}$. The triplet loss [3] is then adopted to enforce the global representation \mathbf{G} being close to the positive bag center while away from the negative bag center:

$$\mathcal{L}_{tri} = \sum_{k=1}^K [d_+(G_k, \tilde{x}_c^{(pos)}) - d_-(G_k, \tilde{x}_c^{(neg)}) + \mu]_+, \quad (8)$$

where μ is the margin parameter, and d denotes the distance measure. We use cosine similarity as the distance measure.

Diversity Learning. Although the positive instance alignment mechanism pushes the global representation to be aligned with the positive bag center, it is likely to result in a trivial solution where all the global vectors are identical. However, a diverse global representation is desired to capture the variability of positive instances. Hence, we propose our unique diversity loss inspired by DPP for data selection to maximize the diversity among global vectors and hence better summarize the instances. DPP is a well-known diversification tool [29] and is often used to select diverse subsets [10, 12, 13, 17, 50]. Inspired so, rather than use it for selection, we utilize it as a diversity measurement.

Mathematically, \mathcal{P} is an L-ensemble DPP if the likelihood of an arbitrary subset $A \subseteq \mathcal{S}$ drawn from the entire set \mathcal{S} satisfies:

$$\mathcal{P}_{\mathbf{L}}(A) \propto \det(\mathbf{L}_A), \quad (9)$$

where \mathbf{L}_A denotes a submatrix of the similarity *Gram matrix* \mathbf{L} indexed by A . In the case of prompting diversity of global vectors $\mathbf{G} = [\mathbf{g}_1^\top, \dots, \mathbf{g}_K^\top]$, the similarity matrix is given as $\mathbf{L} = \mathbf{G}\mathbf{G}^\top \in \mathbb{R}^{K \times K}$, we simply set $A = \mathcal{B} = [K]$ and each global vector \mathbf{g}_i , $i \in A$ is treated as a data point, and the total number of subsets can be calculated as $2^{|\mathcal{S}|} = 2^K$. It is worth noting that the matrix \mathbf{L} is positive semi-definite.

Lemma 1. ([29]) *From a geometric perspective, the determinants in Eq.(9) can be interpreted as the squared $|A|$ -dimensional volume spanned by its feature vectors:*

$$\mathcal{P}_{\mathbf{L}}(A) \propto \det(\mathbf{L}_A) = \text{Vol}^2(\{\mathbf{g}_i\}_{i \in A}). \quad (10)$$

Lemma 1 immediately implies that a diverse subset is more likely to span larger volumes. This is because as the similarity between two data points (i.e., $\mathbf{L}_{ij: i \neq j}$) increases, they will span fewer areas (see Fig. 3(a) and (b)), hence decreasing the probabilities of sets containing both of them (see Eq.(9)). Accordingly, feature vectors that are more orthogonal to each other span the largest volumes (see Fig. 3(a)), hence resulting in the most diverse subsets.

Theorem 1. *Given a set of global vectors $\mathbf{G} = [\mathbf{g}_1^\top, \dots, \mathbf{g}_K^\top]$ with $\|\mathbf{g}_i\| = C, \forall i \in [K]$, maximizing the DPP-based diversity (i.e. $\max \det(\mathbf{G}\mathbf{G}^\top)$) results in orthogonal global vectors with $\mathbf{g}_i \perp \mathbf{g}_j, \forall i \neq j, i, j \in [K]$.*

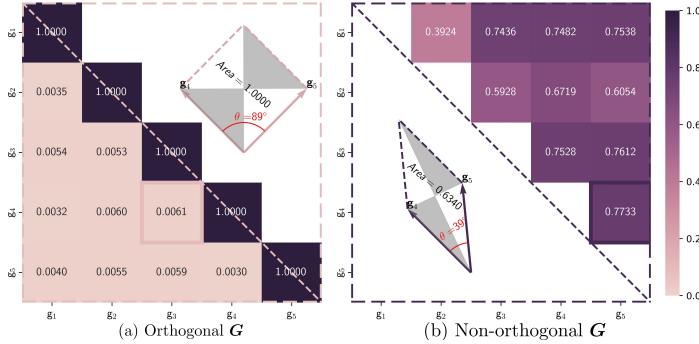


Fig. 3: The similarity matrix for the global vectors \mathbf{G} learned from the CAMELYON16 dataset in two scenarios: (a) \mathbf{G} is orthogonal and (b) \mathbf{G} is non-orthogonal. To support Lemma 1 and Remark 1, we computed the area of the parallelogram corresponding to the two highly correlated global vectors. We omitted the diagonal elements in subpanel figure (b), as $\mathbf{L}_{ii} = 1, \forall i \in [K]$.

Proof. The determinant $\det(\mathbf{L}) = \det(\mathbf{G}\mathbf{G}^\top)$ is upper-bounded according to Hadamard's inequality [39]:

$$|\det(\mathbf{L})| \stackrel{(a)}{=} \det(\mathbf{L}) \stackrel{(b)}{\leq} \prod_{i=1}^K \mathbf{L}_{ii}. \quad (11)$$

Condition (a) is fulfilled because the matrix \mathbf{L} is positive semi-definite. The equality of *Condition (b)* is achieved if and only if all non-diagonal entries of \mathbf{G} are zeros, meaning rows of the global vectors are orthogonal. The normalization constraint in Eq.(11) leads the upper bound to be the infimum, since $\mathbf{L}_{ii} = \|\mathbf{g}_i\|^2 \leq C^2$ and it can be achieved if and only if the equality of *Condition (b)* is satisfied. This completes the proof.

According to Theorem 1, we propose a diversity loss \mathcal{L}_{div} to diversify the proposed global vectors by minimizing the negative logarithm of $\det(\mathbf{G}\mathbf{G}^\top)$:

$$\mathcal{L}_{div} = -\log \det(\mathbf{G}\mathbf{G}^\top), \quad \text{s.t. } \|\mathbf{g}_i\| = 1 = C. \quad (12)$$

Remark 1. Theorem 1 implies that optimal diversity through minimizing our loss is theoretically achievable. This is because enforcing the constraints $\|\mathbf{g}_i\| = 1$ leads the infimum of \mathcal{L}_{div} to reach zero due to $\log(\mathbf{G}\mathbf{G}^\top)_{ii} = \log(\|\mathbf{g}_i\|^2) = 0$. In contrast, the diversity loss \mathcal{L}_{div} can be arbitrarily small (up to $-\infty$) without the constraint $\|\mathbf{g}_i\| = 1$, which results in an unstable training.

We also add a small value $\epsilon = 1 \times 10^{-10}$ to prevent the logarithm of the determinant from being negative infinity (i.e. any two global vectors become collinear). The final diversity loss is given as

$$\mathcal{L}_{div} = -\log \det(\mathbf{G}\mathbf{G}^\top + \epsilon\mathbf{I}), \quad (13)$$

where \mathbf{I} denotes the identity matrix. It is noteworthy that the complexity to compute the loss is approximate $\mathcal{O}(L)$, which is negligible (see Appendix D).

3.3 Objective Function

The proposed MIL model is trained in an end-to-end fashion by jointly optimizing the weighted combination of cross-entropy (ce) loss that corresponds to the bag-level classification, triplet loss, and the proposed diversity loss:

$$\mathcal{L}_{final} = \mathcal{L}_{ce} + \lambda_{tri}\mathcal{L}_{tri} + \lambda_{div}\mathcal{L}_{div}, \quad (14)$$

where λ_{tri} and λ_{div} are balance parameters.

4 Experiments and Results

To validate the effectiveness of the proposed DGR-MIL, we conduct experiments on the CAMELYON16 dataset [5] and TCGA-lung cancer dataset (TCGA-NSCLC).

Dataset and Evaluation Metrics. The two datasets are followed the experimental data partition setting in [64]. For the CAMELYON16, the training set is further divided into training and validation sets with a 9:1 ratio. We report the mean of accuracy, F1 score, and AUC with their corresponding 95% interval on the testing dataset after running five experiments. For the TCGA lung cancer dataset, we perform 4-fold cross-validation experiments, where the dataset is partitioned into training, validation, and testing sets with a patient ratio of 65:10:25. We report the mean and standard variation of accuracy, F1 score, and AUC on the testing dataset from 4-fold cross-validation.

Experiment Setup. Three sets of instance features were extracted using different strategies to evaluate the proposed method’s adaptability across various feature embeddings. The first set provided by DTFD-MIL [64], employing OTSU’s method for patch extraction from WSIs and ResNet-50 for feature extraction, resulting in 1024-dimensional vectors per patch. For thorough validation, two additional sets of features were generated by segmenting each WSI into non-overlapping 224x224 patches using threshold filtering, resulting in 3.4 and 10.3 million patches from CAMELYON16 and TCGA lung cancer datasets [30,31,40,68], respectively. These patches were processed using ResNet-18 and Vision Transformer, pre-trained on ImageNet, to produce 512 and 768-dimensional feature vectors.

Baseline MIL Models. We compare the proposed model to eight state-of-the-art MIL methods. These models can be roughly divided into two categories: i) AB-MIL [28] and its variants, including CLAM-SB [35], DS-MIL [30], and DTFD-MIL [64]; ii) the transformer-based methods including Trans-MIL [47] and ILRA-MIL [58], iii) clustering/prototype-based MIL including PMIL [62].

Implementation Details. All the models are trained using the parameter settings provided by [30,35,47,58,64]. (See Appendix B, including our method).

Additional Experiments. We also include the experiments on using CTransPath [56] as feature extractor for CAMELYON16 dataset. Additionally, to validate the generalizability of our method on broader applications other than WSI, we conduct the experiment on MIL benchmark [1,19]. Our method demonstrates

Table 1: Main results on the CAMELYON16 dataset and TCGA-NSCLC dataset by using features extracted by different means. Our method statistically outperforms all other competitors (refer to the statistic test in Appendix E)

	CAMELYON16			TCGA-NSCLC		
	Accuracy	F1	AUC	Accuracy	F1	AUC
ResNet-50 ImageNet Pretrained						
Classic AB-MIL (<i>ICML’18</i>)	0.845 _(0.839, 0.851)	0.780 _(0.769, 0.791)	0.854 _(0.848, 0.860)	0.869 _{0.032}	0.866 _{0.021}	0.941 _{0.028}
DS-MIL (<i>CVPR’21</i>)	0.856 _(0.843, 0.869)	0.815 _(0.797, 0.832)	0.899 _(0.890, 0.908)	0.888 _{0.013}	0.876 _{0.011}	0.939 _{0.019}
CLAM-SB (<i>Nature Bio. Eng.’21</i>)	0.837 _(0.809, 0.865)	0.775 _(0.755, 0.795)	0.871 _(0.856, 0.885)	0.875 _{0.041}	0.864 _{0.043}	0.944 _{0.023}
CLAM-MB (<i>Nature Bio. Eng.’21</i>)	0.823 _(0.795, 0.850)	0.774 _(0.752, 0.795)	0.878 _(0.861, 0.894)	0.878 _{0.043}	0.874 _{0.028}	0.949 _{0.019}
PMIL (<i>MedIA’23</i>)	0.831 _(0.799, 0.863)	0.816 _(0.779, 0.853)	0.845 _(0.813, 0.876)	0.873 _{0.010}	0.875 _{0.011}	0.933 _{0.007}
Trans-MIL (<i>NeurIPS’21</i>)	0.858 _(0.848, 0.868)	0.797 _(0.776, 0.818)	0.906 _(0.875, 0.937)	0.883 _{0.022}	0.876 _{0.021}	0.949 _{0.013}
DTFD-MIL (MaxS) (<i>CVPR’22</i>)	0.864 _(0.848, 0.880)	0.814 _(0.802, 0.826)	0.907 _(0.894, 0.919)	0.868 _{0.040}	0.863 _{0.029}	0.919 _{0.037}
DTFD-MIL (MaxMinS) (<i>CVPR’22</i>)	0.899 _(0.887, 0.912)	0.865 _(0.848, 0.882)	0.941 _(0.936, 0.944)	0.894 _{0.033}	0.891 _{0.027}	0.961 _{0.021}
DTFD-MIL (AFS) (<i>CVPR’22</i>)	0.908 _(0.892, 0.925)	0.882 _(0.861, 0.903)	0.946 _(0.941, 0.951)	0.891 _{0.033}	0.883 _{0.025}	0.951 _{0.022}
ILRA-MIL (<i>ICLR’23</i>)	0.848 _(0.844, 0.853)	0.826 _(0.823, 0.829)	0.868 _(0.852, 0.883)	0.895 _{0.017}	0.896 _{0.017}	0.946 _{0.014}
Our	0.917_(0.902, 0.931)	0.913_(0.898, 0.928)	0.957_(0.951, 0.963)	0.908_{0.015}	0.911_{0.018}	0.963_{0.008}
ResNet-18 ImageNet Pretrained						
Classic AB-MIL (<i>ICML’18</i>)	0.805 _(0.772, 0.837)	0.786 _(0.757, 0.815)	0.843 _(0.827, 0.858)	0.874 _{0.005}	0.873 _{0.006}	0.937 _{0.001}
DS-MIL (<i>CVPR’21</i>)	0.791 _(0.739, 0.843)	0.776 _(0.712, 0.840)	0.814 _(0.754, 0.875)	0.831 _{0.012}	0.838 _{0.008}	0.896 _{0.009}
CLAM-SB (<i>Nature Bio. Eng.’21</i>)	0.792 _(0.769, 0.815)	0.766 _(0.746, 0.786)	0.811 _(0.777, 0.845)	0.869 _{0.010}	0.869 _{0.010}	0.931 _{0.006}
CLAM-MB (<i>Nature Bio. Eng.’21</i>)	0.786 _(0.754, 0.818)	0.770 _(0.746, 0.795)	0.825 _(0.808, 0.843)	0.880 _{0.016}	0.880 _{0.016}	0.944 _{0.012}
PMIL (<i>MedIA’23</i>)	0.800 _(0.775, 0.825)	0.784 _(0.765, 0.804)	0.829 _(0.807, 0.851)	0.856 _{0.006}	0.862 _{0.003}	0.933 _{0.010}
Trans-MIL (<i>NeurIPS’21</i>)	0.839 _(0.822, 0.856)	0.827 _(0.805, 0.848)	0.854 _(0.823, 0.886)	0.877 _{0.009}	0.879 _{0.008}	0.938 _{0.014}
DTFD-MIL (MaxS) (<i>CVPR’22</i>)	0.856 _(0.824, 0.887)	0.792 _(0.742, 0.842)	0.875 _(0.862, 0.893)	0.830 _{0.014}	0.821 _{0.020}	0.893 _{0.015}
DTFD-MIL (MaxMinS) (<i>CVPR’22</i>)	0.833 _(0.807, 0.858)	0.768 _(0.747, 0.788)	0.878 _(0.872, 0.883)	0.853 _{0.012}	0.850 _{0.021}	0.925 _{0.013}
DTFD-MIL (AFS) (<i>CVPR’22</i>)	0.817 _(0.791, 0.843)	0.734 _(0.687, 0.781)	0.868 _(0.841, 0.896)	0.870 _{0.007}	0.864 _{0.012}	0.935 _{0.010}
ILRA-MIL (<i>ICLR’23</i>)	0.831 _(0.768, 0.895)	0.819 _(0.768, 0.871)	0.852 _(0.811, 0.893)	0.878 _{0.002}	0.879 _{0.001}	0.937 _{0.004}
Our	0.873_(0.862, 0.884)	0.862_(0.852, 0.871)	0.898_(0.886, 0.909)	0.891_{0.029}	0.890_{0.021}	0.955_{0.023}
Vision Transformer ImageNet Pretrained						
Classic AB-MIL (<i>ICML’18</i>)	0.851 _(0.837, 0.865)	0.835 _(0.810, 0.860)	0.873 _(0.840, 0.906)	0.904 _{0.011}	0.904 _{0.010}	0.953 _{0.013}
DS-MIL (<i>CVPR’21</i>)	0.810 _(0.741, 0.879)	0.806 _(0.742, 0.869)	0.871 _(0.836, 0.906)	0.875 _{0.020}	0.879 _{0.016}	0.933 _{0.016}
CLAM-SB (<i>Nature Bio. Eng.’21</i>)	0.839 _(0.831, 0.847)	0.816 _(0.799, 0.834)	0.864 _(0.841, 0.887)	0.907 _{0.008}	0.907 _{0.001}	0.954 _{0.014}
CLAM-MB (<i>Nature Bio. Eng.’21</i>)	0.826 _(0.806, 0.846)	0.804 _(0.795, 0.813)	0.851 _(0.825, 0.878)	0.911 _{0.007}	0.911 _{0.007}	0.959 _{0.008}
PMIL (<i>MedIA’23</i>)	0.843 _(0.831, 0.856)	0.826 _(0.814, 0.838)	0.843 _(0.820, 0.867)	0.882 _{0.009}	0.884 _{0.006}	0.940 _{0.006}
Trans-MIL (<i>NeurIPS’21</i>)	0.862 _(0.841, 0.883)	0.846 _(0.823, 0.869)	0.860 _(0.848, 0.873)	0.909 _{0.009}	0.909 _{0.009}	0.953 _{0.006}
DTFD-MIL (MaxS) (<i>CVPR’22</i>)	0.846 _(0.832, 0.860)	0.767 _(0.746, 0.787)	0.859 _(0.842, 0.876)	0.904 _{0.011}	0.904 _{0.010}	0.953 _{0.013}
DTFD-MIL (MaxMinS) (<i>CVPR’22</i>)	0.839 _(0.826, 0.851)	0.752 _(0.742, 0.763)	0.862 _(0.836, 0.888)	0.895 _{0.013}	0.892 _{0.016}	0.952 _{0.011}
DTFD-MIL (AFS) (<i>CVPR’22</i>)	0.831 _(0.818, 0.844)	0.759 _(0.737, 0.781)	0.880 _(0.864, 0.897)	0.901 _{0.005}	0.900 _{0.008}	0.959 _{0.012}
ILRA-MIL (<i>ICLR’23</i>)	0.850 _(0.825, 0.875)	0.838 _(0.812, 0.865)	0.864 _(0.843, 0.885)	0.902 _{0.007}	0.904 _{0.007}	0.954 _{0.006}
Our	0.893_(0.889, 0.897)	0.882_(0.877, 0.886)	0.891_(0.884, 0.899)	0.926_{0.008}	0.925_{0.008}	0.969_{0.004}

the obvious superiority over other methods in both experiments. Please refer to Appendix F.

4.1 Experimental Results

The proposed method outperforms the other state-of-the-art MIL aggregation models by a large margin in both the CAMELYON16 and TCGA-NSCLC datasets using features extracted by three different means (see Table 1). We also show the statistical superiority of our method in Appendix E. Specifically, the proposed model outperforms the second-best models in terms of accuracy (1.7%; 1.3%), F1 score (3.1%; 1.5%), and AUC (1.1%; 1.7%) when using features extracted from ResNet-50 in CAMELYON16 and TCGA-NSCLC, respectively. A similar performance gain is observed on features extracted from ResNet-18 including accuracy (3.4%; 1.1%), F1 score (3.5%; 1.0%), and AUC (4.4%; 1.1%). We also observe an improvement in accuracy (3.4%; 1.1%), F1 score (3.5%; 1.0%), and AUC (4.4%; 1.1%) when using features extracted from the vision transformer.

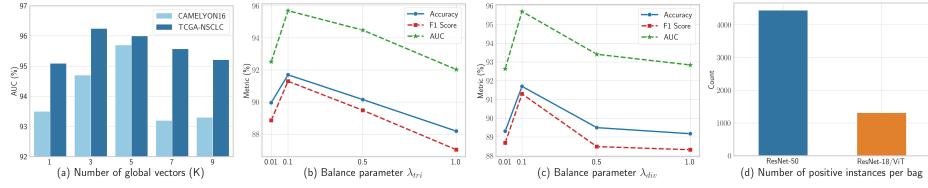


Fig. 4: Ablation studies on (a) number of non-tokenized global vectors on both CAMELYON16 and TCGA-NSCLC datasets, (b) and (c) balance parameter λ_{tri} and λ_{div} on CAMELYON16 dataset, respectively. (d) Comparison in the number of positive instances per bag.

Table 2: The ablation studies on different modules. \mathcal{P} : Positive instance alignment module. \mathcal{D} : Diversity loss.

\mathcal{P}	\mathcal{D}	CAMELYON16			TCGA-NSCLC		
		Accuracy	F1	AUC	Accuracy	F1	AUC
✗	✗	0.895	0.887	0.922	0.872	0.875	0.928
✗	✓	0.906	0.900	0.938	0.896	0.896	0.952
✓	✗	0.917	0.910	0.944	0.900	0.904	0.956
✓	✓	0.917	0.913	0.957	0.908	0.911	0.963

Table 3: The ablation studies on tokenized global representation.

g_{token}	CAMELYON16			TCGA-NSCLC		
	Accuracy	F1	AUC	Accuracy	F1	AUC
✗	0.907	0.900	0.935	0.903	0.905	0.957
✓	0.917	0.913	0.957	0.908	0.911	0.963

In general, the proposed model shows a greater performance improvement in the CAMELYON16 dataset compared to the TCGA-NSCLC dataset. This might be attributed to the fact that CAMELYON16 consists of more diverse instances than TCGA-NSCLC.

We also observe the performance of the three sets of feature embeddings varied: the ViT feature embeddings outperform the ResNet-18 features but show inferior performance compared to the ResNet-50 features. This is mainly attributed to the fact that a greater number of positive instances is extracted by the ResNet-50 (provided by DTFD-MIL) as shown in Fig. 4(d). In contrast, a smaller portion of positive instances in the extracted patches may accompany a drop in performances [41]. This phenomenon benefits the pseudo-bag partitions in DTFD-MIL, as more positive instances within a bag are prone to result in less noisy pseudo-bag labels. This accounts for the drop in DTFD-MIL performance when applied to feature embeddings that contain a lower proportion of positive instances.

4.2 Ablation Studies

We conduct ablation studies on model design variants in the CAMELYON16 dataset with features extracted by a ResNet-50, unless specified otherwise.

Effectiveness of the Proposed Global Representation. We ablate different components of the proposed model, i.e., the positive instance alignment module and the diversity loss. While the model without these two components serves as the baseline in Table 2. We first observe that incorporating the proposed

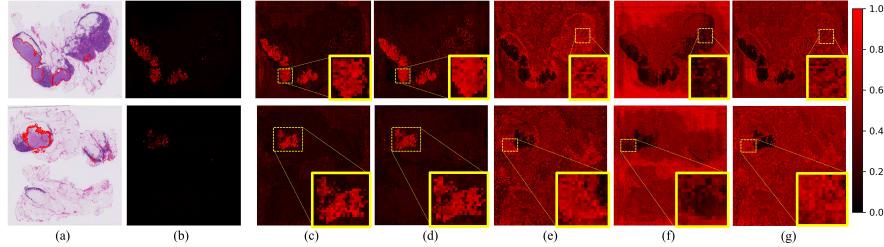


Fig. 5: Visualization of the attention map: (a) raw WSI with the ground-truth annotation, (b) the attention map computes using the tokenized global vectors, and (c-g) the attention map computes using the other ($K - 1$) global vectors with $K = 6$ in our experiment.

global vectors described in Section 3.1 (without employing any of the learning strategies in Section 3.2) yielded an AUC of 0.922 and 0.928. This AUC exceeds that of most existing MIL models, except for DTFD-MIL (MaxMinS & AFS) (see Table 1 and 2). Subsequently, by including the proposed positive instance alignment module, we observe a performance gain of (2.2%, 2.8%) in accuracy, (2.3%, 2.9%) in F1 score, and (2.2%, 2.8%) in AUC. Up to now, we outperform the DTFD-MIL in terms of accuracy and F1 score (see Table 1 and 2), and achieve a similar AUC (AUC = 0.944, 0.956) compare to the DTFD-MIL(AFS) (AUC = 0.946, 0.951). Further incorporating the proposed diversity loss into the objective function yields a performance gain of (1.3%, 0.7%) in AUC, which outperforms DTFD-MIL (AFS) by (1.1%, 1.2%).

Effectiveness of the Tokenized Global Representation. As shown in Table 3, including the tokenized global vector $\mathbf{g}_{\text{token}}$ yields a remarkable performance gain by improving accuracy by (1.0%, 0.5%), F1 score by (1.3%, 0.6%), and AUC by (2.2%, 0.6%). As consistent with the pathological findings that instances are diverse, we observe that different global vectors indeed corresponded to different instance representations, which can be depicted by the attention map produced by different global vectors in Fig. 5. However, we also observe that the learned global vectors still include non-tumor related representation, particularly around tumor boundaries, as positive instances around tumor boundaries have a similar appearance to surrounding negative instances (see Fig. 5.(c) and (d)). As a result, incorporating tokenized global vectors can mitigate this problem by capturing the most discriminative positive (tumor) regions (see Fig. 5.(b)).

Number of Global Vectors. We find that the optimal number of global vectors K in different data sets may vary due to dataset intrinsic properties. Specifically, the optimal K for the CAMELYON16 and TCGA-NSCLC dataset are $K = 5$ and $K = 3$, respectively (Fig. 4.(a)). We observe that an overly large K is likely to decrease performance as it will harden the learning task (see Fig. 4.(a)).

Loss Balance Hyperparameters. By conducting a grid search, we find that the optimal setting of the balance parameters is $\lambda_{\text{tri}} = 0.1$ and $\lambda_{\text{div}} = 0.1$ (see Fig. 4.(b) and (c)). An overly small \mathcal{L}_{tri} and \mathcal{L}_{div} (e.g., 0.01) is likely to enforce inadequate constraints on the learned global representation by deviating it from

learning meaningful information of instance of interest. While larger balance parameters (e.g., {0.5, 1.0}) distract the model from the main classification task, leading to a drop in classification performance.

5 Conclusion

Inspired by the pathological fact that instances are diverse, we propose a novel MIL model from the perspective of modeling diversity in instances through the cross-attention between instances and a set of learnable and diverse global vectors. To learn the global vectors, we propose a positive instance alignment mechanism and the DPP-driven diversity loss. Extensive experiments demonstrate that the proposed MIL model competed favorably against other existing MIL models. Importantly, our work provides an explicit way to account for the diversity in WSI. This pathology-driven approach is beneficial in capturing heterogeneity among the patient population. We also narrowed the performance gap between the diversity-drive MIL method and mainstream MIL.

6 Acknowledgement

This work was partially supported by the grants from NIH (R01EY032125, and R01DE030286), and the State of Arizona via the Arizona Alzheimer Consortium.

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. *Advances in neural information processing systems* **15** (2002) 11
2. Babenko, B., Yang, M.H., Belongie, S.: Robust object tracking with online multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence* **33**(8), 1619–1632 (2010) 4
3. Balntas, V., Riba, E., Ponsa, D., Mikolajczyk, K.: Learning local feature descriptors with triplets and shallow convolutional neural networks. In: *Bmvc*. vol. 1, p. 3 (2016) 9
4. 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.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama* **318**(22), 2199–2210 (2017) 2
5. 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.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama* **318**(22), 2199–2210 (2017) 11
6. Bhattacharya, S., Patel, A., Goyal, N.: On the computational power of transformers and its implications in sequence modeling. *arXiv preprint arXiv:2006.09286* (2020) 5
7. Burrell, R.A., McGranahan, N., Bartek, J., Swanton, C.: The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**(7467), 338–345 (2013) 2

8. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine* **25**(8), 1301–1309 (2019) 6
9. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020) 4
10. Chen, L., Zhang, G., Zhou, E.: Fast greedy map inference for determinantal point process to improve recommendation diversity. *Advances in Neural Information Processing Systems* **31** (2018) 9
11. Chen, P.H.C., Gadepalli, K., MacDonald, R., Liu, Y., Kadokami, S., Nagpal, K., Kohlberger, T., Dean, J., Corrado, G.S., Hipp, J.D., et al.: An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nature medicine* **25**(9), 1453–1457 (2019) 2
12. Chen, X., Li, H., Amin, R., Razi, A.: Rd-dpp: Rate-distortion theory meets determinantal point process to diversify learning data samples. arXiv preprint arXiv:2304.04137 (2023) 3, 9
13. Chen, X., Li, H., Amin, R., Razi, A.: Learning on bandwidth constrained multi-source data with mimo-inspired dpp map inference. *IEEE Transactions on Machine Learning in Communications and Networking* pp. 1–1 (2024). <https://doi.org/10.1109/TMLCM.2024.3421907> 9
14. Chen, X., Qiu, P., Zhu, W., Li, H., Wang, H., Sotiras, A., Wang, Y., Razi, A.: TimeMIL: Advancing multivariate time series classification via a time-aware multiple instance learning. In: Forty-first International Conference on Machine Learning (2024) 4
15. Cover, T.M.: Elements of information theory. John Wiley & Sons (1999) 3, 20
16. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research* **7**, 1–30 (2006) 24
17. Derezhinski, M., Mahoney, M.W.: Determinantal point processes in randomized numerical linear algebra. *Notices of the American Mathematical Society* **68**(1), 34–45 (2021) 9
18. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 4
19. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence* **89**(1-2), 31–71 (1997) 11
20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4
21. Early, J., Cheung, G., Cutajar, K., Xie, H., Kandola, J., Twomey, N.: Inherently interpretable time series classification via multiple instance learning. In: The Twelfth International Conference on Learning Representations (2024) 4
22. Feng, J., Zhou, Z.H.: Deep miml network. In: Proceedings of the AAAI conference on artificial intelligence. vol. 31 (2017) 4
23. Guo, M.H., Liu, Z.N., Mu, T.J., Hu, S.M.: Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(5), 5436–5447 (2022) 5, 24

24. Hannig, J., Schäfer, H., Ackermann, J., Hebel, M., Schäfer, T., Döring, C., Hartmann, S., Hansmann, M.L., Koch, I.: Bioinformatics analysis of whole slide images reveals significant neighborhood preferences of tumor cells in hodgkin lymphoma. *PLOS Computational Biology* **16**(1), e1007516 (2020) [2](#), [8](#)
25. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020) [8](#)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) [22](#)
27. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2424–2433 (2016) [4](#)
28. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018) [2](#), [4](#), [6](#), [11](#), [21](#), [22](#), [23](#)
29. Kulesza, A., Taskar, B., et al.: Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning* **5**(2–3), 123–286 (2012) [9](#)
30. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2021) [2](#), [4](#), [11](#), [21](#), [22](#)
31. Lin, T., Yu, Z., Hu, H., Xu, Y., Chen, C.W.: Interventional bag multi-instance learning on whole-slide pathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19830–19839 (2023) [11](#)
32. Liu, K., Zhu, W., Shen, Y., Liu, S., Razavian, N., Geras, K.J., Fernandez-Granda, C.: Multiple instance learning via iterative self-paced supervised contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3355–3365 (2023) [3](#), [4](#)
33. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) [4](#)
34. Lu, M.Y., Chen, R.J., Wang, J., Dillon, D., Mahmood, F.: Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. arXiv preprint arXiv:1910.10825 (2019) [6](#)
35. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering* **5**(6), 555–570 (2021) [2](#), [6](#), [11](#), [21](#)
36. Ma, Y., Derksen, H., Hong, W., Wright, J.: Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE transactions on pattern analysis and machine intelligence* **29**(9), 1546–1562 (2007) [20](#)
37. Marusyk, A., Polyak, K.: Tumor heterogeneity: causes and consequences. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* **1805**(1), 105–117 (2010) [2](#)
38. Nagpal, K., Foote, D., Tan, F., Liu, Y., Chen, P.H.C., Steiner, D.F., Manoj, N., Olson, N., Smith, J.L., Mohtashamian, A., et al.: Development and validation of a deep learning algorithm for gleason grading of prostate cancer from biopsy specimens. *JAMA oncology* **6**(9), 1372–1380 (2020) [2](#)

39. Petersen, K.B., Pedersen, M.S., et al.: The matrix cookbook. Technical University of Denmark **7**(15), 510 (2008) [10](#)
40. Qiu, P., Xiao, P., Zhu, W., Wang, Y., Sotiras, A.: Sc-mil: Sparsely coded multiple instance learning for whole slide image classification. arXiv preprint arXiv:2311.00048 (2023) [4](#), [11](#)
41. Qu, L., Yang, Z., Duan, M., Ma, Y., Wang, S., Wang, M., Song, Z.: Boosting whole slide image classification from the perspectives of distribution, correlation and magnification. In: Proceedings of the IEEE/CVF International Conference Computer Vision (ICCV). pp. 21463–21473 (October 2023) [13](#)
42. Quellec, G., Cazuguel, G., Cochener, B., Lamard, M.: Multiple-instance learning for medical image and video analysis. IEEE reviews in biomedical engineering **10**, 213–234 (2017) [4](#)
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) [4](#)
44. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019) [4](#)
45. Ruoss, A., Delétang, G., Genewein, T., Grau-Moya, J., Csordás, R., Bennani, M., Legg, S., Veness, J.: Randomized positional encodings boost length generalization of transformers. arXiv preprint arXiv:2305.16843 (2023) [5](#)
46. Schrader, T., Niepage, S., Leuthold, T., Saeger, K., Schluns, K., Hufnagl, P., Kayser, K., Dietel, M.: The diagnostic path, a useful visualisation tool in virtual microscopy. Diagnostic Pathology **1**(1), 1–7 (2006) [2](#)
47. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021) [2](#), [4](#), [7](#), [11](#), [21](#), [22](#), [23](#)
48. Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H.: Efficient attention: Attention with linear complexities. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3531–3539 (2021) [5](#), [24](#)
49. Sun, R., Li, Y., Zhang, T., Mao, Z., Wu, F., Zhang, Y.: Lesion-aware transformers for diabetic retinopathy grading. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10938–10947 (2021) [5](#)
50. Tremblay, N., Barthélémy, S., Ambard, P.O.: Determinantal point processes for coresets. J. Mach. Learn. Res. **20**, 168–1 (2019) [9](#)
51. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) [4](#)
52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) [4](#), [5](#)
53. Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H.: Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768 (2020) [5](#), [24](#)
54. Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W.: Revisiting multiple instance neural networks. Pattern Recognition **74**, 15–24 (2018) [4](#)
55. Wang, X., Xiang, J., Zhang, J., Yang, S., Yang, Z., Wang, M.H., Zhang, J., Yang, W., Huang, J., Han, X.: Scl-wc: Cross-slide contrastive learning for weakly-supervised whole-slide image classification. Advances in neural information processing systems **35**, 18009–18021 (2022) [3](#), [4](#)

56. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis* **81**, 102559 (2022) [11](#)
57. Wu, H., Xu, J., Wang, J., Long, M.: Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems* **34**, 22419–22430 (2021) [4](#)
58. Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: *The Eleventh International Conference on Learning Representations* (2023) [2](#), [4](#), [5](#), [11](#), [22](#), [23](#), [24](#)
59. Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., Singh, V.: Nyströmformer: A nyström-based algorithm for approximating self-attention. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 35, pp. 14138–14148 (2021) [7](#)
60. Xu, G., Song, Z., Sun, Z., Ku, C., Yang, Z., Liu, C., Wang, S., Ma, J., Xu, W.: Camel: A weakly supervised learning framework for histopathology image segmentation. In: *Proceedings of the IEEE/CVF International Conference on computer vision*. pp. 10682–10691 (2019) [4](#)
61. Yang, L., Mehta, D., Liu, S., Mahapatra, D., Di Ieva, A., Ge, Z.: Tpmil: Trainable prototype enhanced multiple instance learning for whole slide image classification. *arXiv preprint arXiv:2305.00696* (2023) [3](#), [4](#)
62. Yu, J.G., Wu, Z., Ming, Y., Deng, S., Li, Y., Ou, C., He, C., Wang, B., Zhang, P., Wang, Y.: Prototypical multiple instance learning for predicting lymph node metastasis of breast cancer from whole-slide pathological images. *Medical Image Analysis* **85**, 102748 (2023) [3](#), [4](#), [11](#)
63. Yu, Y., Chan, K.H.R., You, C., Song, C., Ma, Y.: Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in Neural Information Processing Systems* **33**, 9422–9434 (2020) [3](#), [21](#)
64. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfm-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*. pp. 18802–18812 (2022) [2](#), [4](#), [11](#), [21](#), [22](#), [23](#)
65. Zhang, M., Lucas, J., Ba, J., Hinton, G.E.: Lookahead optimizer: k steps forward, 1 step back. *Advances in neural information processing systems* **32** (2019) [22](#)
66. Zhao, S., Chen, D.P., Fu, T., Yang, J.C., Ma, D., Zhu, X.Z., Wang, X.X., Jiao, Y.P., Jin, X., Xiao, Y., et al.: Single-cell morphological and topological atlas reveals the ecosystem diversity of human breast cancer. *Nature Communications* **14**(1), 6796 (2023) [2](#)
67. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., Zhang, W.: Informer: Beyond efficient transformer for long sequence time-series forecasting. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 35, pp. 11106–11115 (2021) [4](#)
68. Zhu, W., Qiu, P., Dumitrascu, O.M., Wang, Y.: Pdl: Regularizing multiple instance learning with progressive dropout layers. *arXiv preprint arXiv:2308.10112* (2023) [3](#), [4](#), [11](#)

Supplementary Materials - DGR-MIL: Exploring Diverse Global Representation in Multiple Instance Learning for Whole Slide Image Classification

A Measuring Diversity Based on Rate-distortion Theory

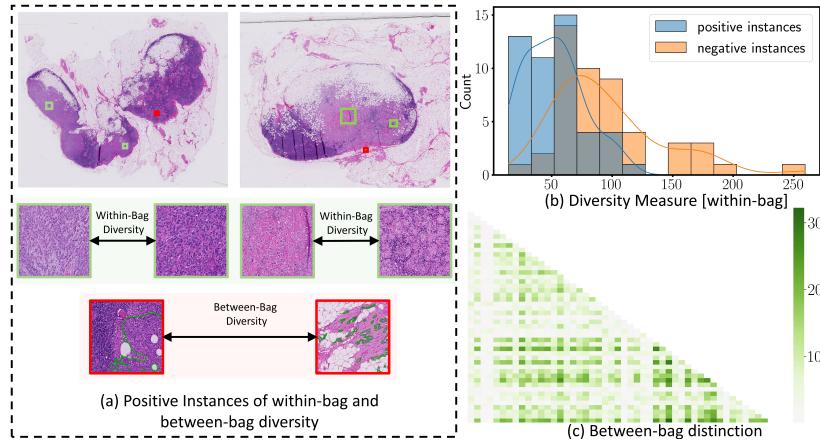


Fig. 6: (a) Examples of positive instances of within-bag and between-bag diversities measured by rate-distortion theory. (b) Histogram of the diversity measure within positive bags on the CAMELYON16 dataset. (c) The between-bag distinction measures the pair-wise similarity between bags.

Rate-distortion (RD) theory is a fundamental concept in *information theory* to describe the lossy compression for arbitrary data sources with tolerable distortion. Here, rate R refers to the number of bits or units per symbol of information required to represent the source data or signal; while distortion measures the quality of the reconstructed data compared to the original source data. Mathematically, given an arbitrary source X , we can use finite bits nR bits to encode a sequence of n samples X^n with $f_n(X^n)$ using a size codebook 2^{nR} , and then decode it with $\hat{X}^n = g_n(f_n(X^n))$. Accordingly, the reconstruction error for the sample sequence x^n can be computed as $d(x^n, \hat{x}^n) := 1/n \sum_{i=1}^n d(x_i, \hat{x}_i)$ for some distance measure $d(\cdot)$. The most commonly used distortion metric is Mean Squared Errors (MSE), which is presented as $\epsilon^2 := 1/n \sum_{i=1}^n (x_i - \hat{x}_i)^2$ and distortion D is defined as $D := \mathbb{E}[d(X^n, \hat{X}^n)]$ [15]. The rate R is computed for a sequence with infinite length ($n \rightarrow \infty$) and distortion D . For a Gaussian source, given a finite dataset $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, the theoretical coding rate with a small tolerable MSE distortion ϵ^2 , can be approximately estimated as [36],

$$R(\mathbf{X}, \epsilon) := \frac{1}{2} \log \det \left(\mathbf{I} + \frac{d}{n\epsilon^2} \mathbf{X} \mathbf{X}^\top \right), \quad (15)$$

where the unit of $R(\mathbf{X}, \epsilon)$ is bit/dimension or nat/dimension for log base 2 or e , respectively. Accordingly, the rate of the sub-space for each class i can be approximated,

$$R_i^c(\mathbf{X}, \epsilon | C_i) := \frac{1}{2} \log \det \left(\mathbf{I} + \frac{d}{|C_i|\epsilon^2} \mathbf{X}_{C_i} \mathbf{X}_{C_i}^\top \right), \quad (16)$$

where C_i is the index set of class i , c_T is the number of classes, \mathbf{X}_{C_i} is a matrix using columns of \mathbf{X} indexed by C_i ($\mathbf{X}[:, C_i]$), and $|C_i|$ is the cardinality of C_i . Having adopted the assumption in [63], we use the latent features extracted by the projector to estimate the diversity.

To better illustrate the way to compute the diversity, we copy Fig.6 from the main body to here. In Fig.6.(b), we use Eq. 16 to compute the within-bag diversity, which refers to either all negative instances or positive instances (if applicable) from the same bag. The instances are treated as a data matrix \mathbf{X} in Eq.15. We separately compute the diversity for each bag from the test set (80 negative bags and 49 positive bags), which results in a total of 129 negative within-bag diversity data points and 49 positive within-diversity data points. Then, we plot the histograms in Fig.6.(b), where the x-axis denotes the measure of diversity and the y-axis denotes the count (or frequency) of the diversity within the interval (i.g. the width and height in a bin, respectively). It evidences both positive and negative within-bag instances are diverse and on a comparable scale. In Fig.6.(c), we use the rate reduction from [63] to compute the between-bag distinction of positive instances for every two bags. A rate reduction is presented as

$$\Delta R := R(\mathbf{X}[:, C_1 \cup C_2], \epsilon) - \sum_{i=1}^2 \frac{|C_i|}{n} R_i^c(\mathbf{X}, \epsilon | C_i), \quad (17)$$

where C_1 and C_2 are the index sets of two sub-space. This concept is used to describe the difference to encode the entire space and encode the sum of all sub-spaces, and a higher value indicates two sub-spaces are more discriminative; hence, we employed it as a metric to describe the distinction between two bags. In detail, we compare the distinction for every two bags. In each computation, C_1 and C_2 denote the indices of positive instances from two different bags, respectively. $C_1 \cup C_2$ denotes all instances from the selected two bags. Fig.6.(c) denotes the pair-wise distinction matrix, and we neglected the diagonal elements since they are zero. We also neglected the upper-triangle elements since this distinction matrix is symmetric.

B Baseline Models Parameter Setting

The baseline MIL methods include AB-MIL [28], CLAM-SB, multi-attention CLAM-MB [35], DS-MIL [30], DTFD-MIL [64], Trans-MIL [47] and ILRA-

Table 4: All training parameters setting for all methods in experiments. Here, Cosine annealing* denotes cosine decay with 20 epoch linear warmup from 1e-5. AMP represents automatic mixed precision, and the grad clip was clipped gradient norm constrained of model weight. Here, BCE was BCEWithLogitsLoss, which combines a sigmoid layer and the binary cross entropy loss.

Parameters Setting	AB-MIL	CLAM-SB/MB	DS-MIL	DTFD-MIL	Trans-MIL	ILRA-MIL	Our proposed Method
optimizer	Adam	Adam	Adam	Adam	Adam	SGD	
learning rate	1e-3	1e-4	1e-4	1e-4	2e-4	1e-4	5e-4
weight decay	0.005	1e-5	5e-3	1e-4	1e-5	1e-4	1e-4
scheduler	Cosine annealing*	Cosine annealing*	Cosine annealing	MultiStepLR	LookAhead [65]	Cosine annealing	Cosine annealing*
Dropout rate	0.15	0.15	0.15	0.15	0.15	0.15	0.15
epoch	200	200	200	200	200	200	200
loss	BCE	BCE	BCE	BCE + Tier-2 loss	BCE	BCE	$\mathcal{L}_{ce} + \lambda_{tri}\mathcal{L}_{tri} + \lambda_{div}\mathcal{L}_{div}$
other settings	None	Early stop	Droppath = 0.2	grad clip = 5	AMP	Xavier initialize	Warmup training strategy

MIL [58]. We follow the optimal parameter settings outlined in their original papers. The detailed parameters that we use to train all the baselines and the proposed model are shown in Table 4. It is worth noting that our method adopts the linear learning rate warmup for the first 20 epochs, and details can be referred to B.4.

B.1 ResNet-50 ImageNet Pre-Trained

We use extracted features released by the DTFD-MIL. Each patch was embedded into a 1024-dimensional vector using a ResNet-18 pretrained on ImageNet [26]. The instance features are directly fed to MIL methods for training. In the experiments, we consistently set the middle layer (Some MIL methods including feed-forward layers before entering the aggregation method) output dimension to 512, For example, TransMIL [47], ILRA [58], DTFD [64], ABMIL [28], and the proposed method.

B.2 ResNet-18 ImageNet Pre-Trained

Different from DTFD-MIL, we employ the threshold filter method (entropy < 5 discarded) to extract patches from raw WSIs [30]. This results in fewer patches compared to DTFD-MIL. Each patch was embedded into a 512-dimensional vector as an instance feature. Here, we consistently set the middle layer output dimension to 256 in all MIL methods, including ILRA [58], DTFD [64], ABMIL [28], and the proposed method. Here, the TransMIL middle layers dimension output is 512, following the settings in its original paper [47]. Here, The TransMIL middle layers dimension output was 512, following the original paper setting [47]. The experiments section of the manuscript reveals a notable performance decline in most MIL methods.

B.3 Vision Transformer ImageNet Pre-Trained

We employ the same threshold filter technique for patch extraction as we have done in the ResNet18 scheme. Each patch is transformed into a 768-dimensional

Table 5: Comparision over efficiency among different transformer-based MIL aggregation methods in terms of the number of Parameters (M) and MACs (G) represent the model size and multiple-accumulated operation computational complexity, respectively.

Models	Params(M)	MACs(G)
ILRA-MIL [58]	1.049	1.842
Trans-MIL [47]	3.040	2.409
Our	0.642	1.054

vector using a vision transformer pre-trained on ImageNet. The middle layer output dimension in MIL methods with feed-forward layers, such as TransMIL [47], ILRA [58], DTFD [64], ABMIL [28], and the proposed method, is set to 512. In line with the TransMIL study, the output dimension of its middle layers is also established at 512.

B.4 Warm-up Training Strategy

As outlined in our paper, a warm-up training strategy is incorporated in all experiments of the proposed method. This warm-up training can be described as follows:

$$\mathcal{L}_{final} = \begin{cases} \mathcal{L}_{ce}, & \text{iff } t < 20, \\ \mathcal{L}_{ce} + \lambda_{tri}\mathcal{L}_{tri} + \lambda_{div}\mathcal{L}_{div}, & \text{iff } t > 20, \end{cases} \quad (18)$$

where t is the current epoch. The total training epoch is set to 200 for all experiments in this paper. We only employ the cross entropy classification loss to train our model at the first 20 epochs; while adding all the other losses for the latter epochs. The rationale behind this is that the randomly initialized global vectors usually lead to instability in training. The warmup training will help the global vectors to learn the meaningful instance relation in classification. This prevents poorly initialized global vectors from incorrectly misleading the modeling of instance correlations at the start.

C Efficiency Comparision of Transformer-Based MIL Aggregation Methods

Take feature vectors extracted by ResNet-18 as an example, we apply the same hidden parameters as reported in the experiments. As shown in Table 5, the proposed method demonstrates superior efficiency compared to the other two transformer-based MIL aggregation methods, exhibiting notable advantages in terms of both model size and computational complexity. The cross-attention mechanism is more computationally efficient compared to the self-attention mechanism used across all instances. This efficiency stems from the use of an extremely

short sequence of global vectors, which is substantially less in number than the total count of instances.

It is worth noting that ILRA-MIL [58] employs self-attention for modeling the correlation between instances. Similarly, it also presents a larger number of parameters and is more computationally complex than the proposed method. The main reason is that they rewrite self-attention instead of self-attention with linear, and added the non-local pooling extra module. complexity [23, 48, 53].

D Complexity of Diversity Loss

The proposed diversity loss can be computed in a linear time complexity. For a global vector $\mathbf{G} \in \mathbb{R}^{K \times L}$, $\log \det(\mathbf{G}\mathbf{G}^T) = \sum_{i=1}^K \log(\lambda_i^2)$, where the main overhead is an SVD decomposition of \mathbf{G} to get λ_i , resulting in a complexity of $\mathcal{O}(LK^2) \approx \mathcal{O}(L)$ due to K is often set a small number (e.g., 5).

E Statistical Test

We present the Wilcoxon signed-rank test and the critical difference diagram [16] in Fig. 7 with $\alpha = 0.5$ significance level. Our method statistically outperforms all other competitors.

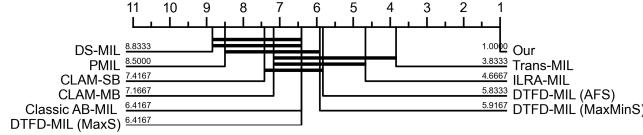


Fig. 7: Wilcoxon signed-rank test, average rank denoted by the number. No statistical difference found between methods connected with one thickness line in the critical difference diagram.

F Additional Results

We present the additional experiments on CAMELYON16 with CTransPath feature extractor and classic MIL benchmarks. The results are shown in Table 6 and Table 7, respectively.

Table 6: Results on **CTransPath** extractor. We employ the 4-fold cross-validation using data split provided by DTFD.

	CAMELYON16		
	Accuracy	F1	AUC
Classic AB-MIL (<i>ICML’18</i>)	0.940 _(0.933,0.948)	0.936 _(0.928,0.944)	0.951 _(0.932,0.970)
DS-MIL (<i>CVPR’21</i>)	0.929 _(0.898,0.959)	0.923 _(0.889,0.957)	0.942 _(0.916,0.968)
Trans-MIL (<i>NeurIPS’21</i>)	0.952 _(0.935,0.970)	0.949 _(0.930,0.968)	0.973 _(0.958,0.987)
DTFD-MIL (MaxMinS) (<i>CVPR’22</i>)	0.949 _(0.931,0.953)	0.933 _(0.906,0.937)	0.985 _(0.976,0.994)
DTFD-MIL (AFS) (<i>CVPR’22</i>)	0.942 _(0.931,0.953)	0.922 _(0.906,0.937)	0.982 _(0.969,0.995)
ILRA-MIL (<i>ICLR’23</i>)	0.940 _(0.924,0.957)	0.937 _(0.922,0.953)	0.961 _(0.946,0.975)
Our	0.972 _(0.965,0.979)	0.971 _(0.963,0.978)	0.994 _(0.991,0.996)

Table 7: Results on MIL benchmarks.

Methods	MUSK1	MUSK2	FOX	TIGER	ELEPHANT
mi-Net	0.889 ± 0.039	0.858 ± 0.049	0.613 ± 0.035	0.824 ± 0.034	0.858 ± 0.037
MI-Net	0.887 ± 0.041	0.859 ± 0.046	0.622 ± 0.038	0.830 ± 0.032	0.862 ± 0.034
MI-Net with DS	0.894 ± 0.042	0.874 ± 0.043	0.630 ± 0.037	0.845 ± 0.039	0.872 ± 0.032
MI-Net with RC	0.898 ± 0.043	0.873 ± 0.044	0.619 ± 0.047	0.836 ± 0.037	0.857 ± 0.040
ABMIL	0.892 ± 0.040	0.858 ± 0.048	0.615 ± 0.043	0.839 ± 0.022	0.868 ± 0.022
ABMIL-Gated	0.900 ± 0.050	0.863 ± 0.042	0.603 ± 0.029	0.845 ± 0.018	0.857 ± 0.027
DP-MINN	0.907 ± 0.036	0.926 ± 0.043	0.655 ± 0.052	0.897 ± 0.028	0.894 ± 0.030
NLMIL	0.921 ± 0.017	0.910 ± 0.009	0.703 ± 0.035	0.857 ± 0.013	0.876 ± 0.011
ANLMIL	0.912 ± 0.009	0.822 ± 0.084	0.643 ± 0.012	0.733 ± 0.068	0.883 ± 0.014
DSMIL	0.932 ± 0.023	0.930 ± 0.020	0.729 ± 0.018	0.869 ± 0.008	0.925 ± 0.007
Our Method	0.989 ± 0.033	0.970 ± 0.0458	0.785 ± 0.120	0.925 ± 0.055	0.950 ± 0.044