Rethinking Multi-Instance Learning through Graph-Driven Fusion: A Dual-Path Approach to Adaptive Representation

Yu-Xuan Zhang[1], Zhengchun Zhou[1*], Weisha Liu[2], Mingxing Zhang[3]

[1]School of Information Science and Technology, Southwest Jiaotong University
[2]SWJTU-Leeds Joint School, Southwest Jiaotong University
[3]School of Mathematics, Southwest Jiaotong University
inki.yinji@gmail.com, zzc@swjtu.edu.cn, lws_weiss@my.swjtu.edu.cn, mxz@swjtu.edu.cn

Abstract

Multi-instance learning (MIL) has become a powerful paradigm for weakly supervised learning tasks, where each sample is a bag of unlabeled instances with only the bag-level label. While graph-based MIL methods enhance bag topological structure modeling, they often suffer from high computation costs and limited representation due to rigid graph construction and insufficient integration of intra-bag semantics. To address these challenges, we propose GDF-MIL, a novel graph-driven MIL framework, which introduces a dual-path feature fusion mechanism to adaptively balance topological structure modeling and semantic feature preservation. First, the adaptive bag mapping module (ABMM) performs soft clustering to extract compact and informative representations. Subsequently, a dynamic graph structure learning (DGSL) component efficiently learns sparse topological structures via weighted connectivity, aggregating them into a comprehensive graph-level representation. Finally, to balance fast graph construction and bag-level knowledge, dual-path feature fusion (DPFF) employs a dual-path gating mechanism to integrate both types of features, which are then passed to the classifier for bag label prediction. Extensive experiments on twenty-four datasets across four domains show that GDF-MIL significantly outperforms eighteen state-of-the-art methods on the majority of datasets.

Introduction

Multi-instance learning (MIL) has become the cornerstone of modern weakly supervised learning systems, powering critical applications such as drug activity prediction (Dietterich, Lathrop, and Lozano-Pérez 1997; Li, Li, and Eliceiri 2021), web page recommendation (Wei et al. 2019; Zhang et al. 2024), and medical image analysis (Cui, Chen, and Su 2025; Tang, Zhang, and Zhang 2024; Zhao, Chen, and Zhao 2025; Zheng et al. 2025a,b; Zhong et al. 2025). Different from traditional supervised learning that requires fine-grained ground-truth, MIL naturally organizes data in a bag of unlabeled instances with only bag-level supervision. However, despite this success, the fundamental challenge remains: how to effectively model the semantic relationships and topological structure among instances and to extract the most representative features for the bag label prediction.

*Corresponding author

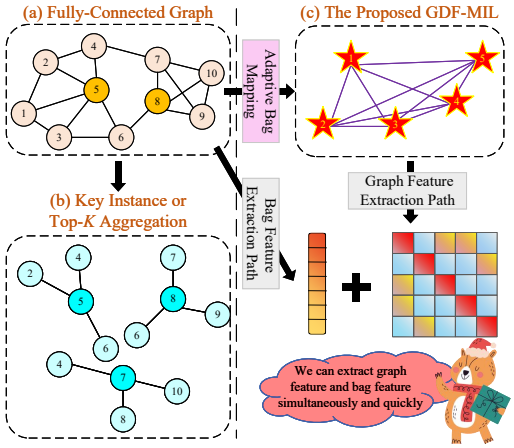


Figure 1: The differences between existing graph-driven MIL and the proposed GDF-MIL. (a) Mining fine-grained topological structure in a bag inevitably incurs high computational cost. Note that for simplicity, we have not shown all the connections; (b) Focusing on key information but ignoring potential intra-bag contextual information; (c) Our GDF-MIL: A novel graph-based MIL architecture that adaptively balances the strengths of the above methods.

Recent graph-based MIL methods aim to answer this question by mining the potential topological structure of the bag to capture latent structure and contextual semantics. However, existing methods still suffer from fundamental limitations, as shown in Figure 1: Fully connected graph method (Pal et al. 2022; Zhao et al. 2024; Wang et al. 2025) aims to capture the comprehensive topological structure of the bag. While pruning alleviates some costs, their time and memory still scale quadratically with bag size, making them impractical for high-cardinality bags frequently observed in medical images and drug screening. While key-instance-based methods (Li et al. 2024; Zhang et al. 2024) select key or top-K instances to construct graphs by employing neighborhood filtering and dynamic structure refinement. Due to the lack of instance-level supervision, the concept of key instances remains inherently ambiguous. As a result, this ambiguity can lead to the omission of crucial graph structures and the underutilization of rich intra-bag information.

In this work, we propose GDF-MIL, a novel and efficient graph MIL framework that explicitly addresses the trade-offs in topology modeling and semantic abstraction through a modular and adaptive design, as shown in Figure 5. GDF-MIL is composed of three key modules. First, to significantly reduce bag cardinality and computation, we design an adaptive bag mapping module (ABMM) that employs Gumbel-Softmax-based soft clustering to project each bag into a compact hidden space. Second, building on the efficiency of ABMM, a dynamic graph structure learning (DGSL) module constructs sparse but informative graph representations by leveraging inductive representation learning (SAGEConv) and a dual-path gating mechanism. Finally, a dual-path feature fusion (DPFF) module adaptively and dynamically integrates graph-level and bag-level representations to produce a robust bag representation for final classification.

Our contributions can be summarized as follows:

1. We propose GDF-MIL, a novel MIL framework that jointly models bag structure and semantics via an adaptive dual-path design, offering a fresh perspective on graph MIL.
2. We introduce a soft-clustering-based bag mapping strategy and a dual-path fusion mechanism, which together improve both scalability and representation quality.
3. We conduct comprehensive experiments on 24 datasets from 4 domains demonstrate the consistent superiority of GDF-MIL over 18 strong MIL baselines in terms of performance and efficiency.

Related Works

MIL originated in pharmaceutical research and was initially used to model the relationship between a drug molecule (bag) composed of various isomers (instances) and its collective drug activity (Dietterich, Lathrop, and Lozano-Pérez 1997). Over time, the core concepts of bag and instance were generalized beyond molecule and isomer, leading to the evolution of MIL into two main branches: traditional algorithms and deep learning-based methods. The former adapts classical machine learning strategies to the MIL setting. Notable examples include instance-level classifiers with hierarchical aggregation (Xiao, Liu, and Hao 2024), bag-level distance-based methods (Yang et al. 2021), and embedding techniques that map each bag into a fixed-length representation (Wu et al. 2018; Zhang, Liu, and Li 2020). Deep learning-based MIL methods build on these traditional methods and aim to leverage the powerful feature representation and reasoning capabilities of deep neural networks to improve scalability and classification performance (Ling et al. 2024; Qu et al. 2024; Wu et al. 2025; Zhang et al. 2022b).

Graph-based MIL has become a highly active research area, primarily due to its powerful ability to model intra-bag topology. The core idea of graph-based MIL is to construct a graph structure for each bag, where instances are represented as nodes and their relationships as edges. One prominent approach is the fully connected graph method (Kapse et al. 2024; Pal et al. 2022; Wang et al. 2025). While designed to comprehensively capture topological structures, this method inherently incurs high computational costs. The second category focuses on key or top-K instances (Zhang et al. 2024; Li et al. 2024), which build graphs faster by using the most informative instances but inevitably ignore the potential topological structure in the bag, thus harming the classification performance.

Therefore, this paper aims to share a novel graph-based MIL architecture that preserves the core ideas of the above two methods and enhances scalability by adaptively fusing bag-level and graph-level features.

Method

Overview of GDF-MIL

In the standard MIL setting, each training sample is a bag $B_i = \{\boldsymbol{x}_{ij}\}_{j=1}^{n_i} \in \mathbb{R}^{n_i \times d}$ with only bag-level label $Y_i \in \mathcal{Y} = \{c\}_{c=1}^{C}$. Here, $\boldsymbol{x}_{ij}$ is the j-th instance, n_i is the cardinality, d is the dimension, and C is the number of classes. In the MIL classification task, the goal is to learn an effective mapping function $f_c : B_i \mapsto \hat{Y}_i$ that maximizes classification performance. Among the various MIL methods, graph-based approaches (Zhao et al. 2024; Wang et al. 2025) leverage the topological structure of the dataset $\mathcal{D} = \{B_i\}_{i=1}^{N}$ or the bag B_i, incorporating learned graph matrix and node/edge representations into the model learning process, e.g., $f_g(B_i|G_i) : B_i \mapsto \hat{Y}_i$. However, fully-connected graph constructions are computationally expensive and may suffer from noisy or redundant edges. Additionally, approaches that focus on key or top-K instances can alleviate this issue but often overlook the potential intra-bag contextual information, leading to suboptimal performance (Zhang et al. 2024; Li et al. 2024).

To overcome these challenges, we proposed a novel graph-driven MIL architecture (GDF-MIL), as shown in Figure 5. The core idea of GDF-MIL is that ABMM accelerates the graph construction process in DGSL by efficiently extracting informative and compact features from B_i through soft clustering and residual gating, thereby avoiding rigid and manually predefined graph structures. Based on this, DPFF incorporates both graph- and instance-level representations and employs dual-path gating to adaptively fuse the two and preserve the bag semantics.

Adaptive Bag Mapping Module (ABMM)

A major bottleneck in graph-based MIL lies in the high cost of constructing and processing graphs over bags with large cardinality (Pal et al. 2022; Wang et al. 2025). ABMM addresses this by transforming the variable-cardinality bag into a compact representation via a soft clustering mechanism, while retaining key semantic cues. Specifically, a two-layer encoder is used to initially extract the information from B_i:

$$B_i^E = \mathcal{A}_L(B_i W_{E_1}) W_{E_2} = B_i^A W_{E_2}, \tag{1}$$

where $\mathcal{A}_L$ is the LeakyReLU activation function, $W_{E_1} \in \mathbb{R}^{d \times d_E}$ and $W_{E_2} \in \mathbb{R}^{d_E \times d_K}$ are the learnable parameters, d_E and d_K are the node numbers of the fully connected layers (FCLs), respectively. This yields an encoded representation $B_i^E \in \mathbb{R}^{n_i \times d_K}$, where the dimensionality aligns with downstream clustering and graph learning stages.

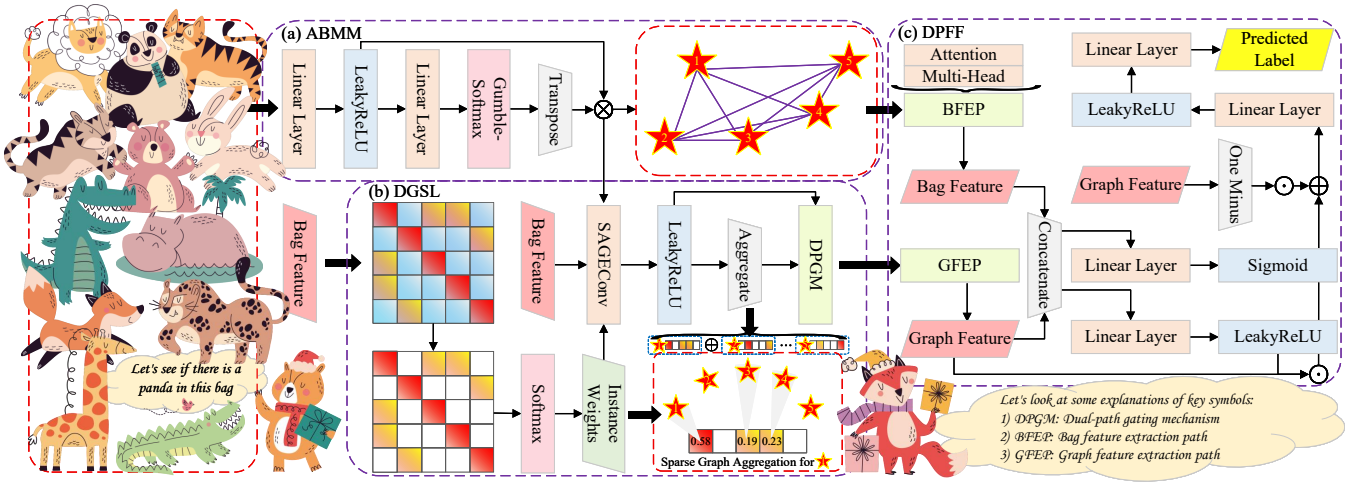


Figure 2: The overview of the proposed GDF-MIL. (a) Adaptive bag mapping module (ABMM): Performs preliminary feature extraction and generates a compact representation; (b) Dynamic graph structure learning (DGSL): Fully explores the graph topological structure based on the weighted connectivity, SAGEConv (inductive representation learning (Hamilton, Ying, and Leskovec 2017)), and DPGM; (c) Dual-path feature fusion (DPFF): Adaptively fuse bag-level and graph-level representations.

To generate a fixed-cardinality bag, the soft clustering with the Gumble-Softmax (Jang, Gu, and Poole 2017) is used to map the bag into the hidden space:

$$B_i^S = P_i^T B_i^A = \{\boldsymbol{x}_{ik}^S\}_{k=1}^{K_C}, \tag{2}$$

where

$$P_i = \{\boldsymbol{p}_{ik}\}_{k=1}^{K_C},$$
$$\boldsymbol{p}_{ik} = \frac{\exp((\boldsymbol{x}_{ik}^E + g_k)/\tau)}{\sum_{j=1}^{K_C} \exp((\boldsymbol{x}_{ij}^E + g_k)/\tau)}. \tag{3}$$

Here, $K_C \leq n_i$ ensures cardinality reduction and computational efficiency, and the Gumbel-Softmax trick introduces noise g_k and temperature τ to allow differentiable clustering while approximating hard assignment behavior. The clustered bag B_i^S then serves as the input for efficient topology modeling in DGSL.

Dynamic Graph Structure Learning (DGSL)

DGSL is responsible for modeling the relational dependencies among the new bag from ABMM. Instead of constructing graphs over all instances, we restrict attention to a compact bag with reduced cardinality, enabling efficient and informative graph learning.

First, a fully-connected graph is constructed by computing the pairwise similarities between all instances in B_i^S:

$$S_i = \frac{\left(B_i^S W_{S_1}\right)\left(B_i^S W_{S_2}\right)^T}{\sqrt{d_{E_2}}}, \tag{4}$$

where $W_S \in \mathbb{R}^{d_K \times d_K}$ is the learnable parameters. This step is crucial to learn adaptive similarity rather than relying on static distances, allowing the model to tune instance relations during training.

Second, to focus aggregation and enforce sparsity in the learned graph, we consider only the most informative K_N instances within each instance's domain as its neighbors:

$$\mathcal{N}_i(k) = \text{TopK}(S_i, K_N), \tag{5}$$

where $K_N \leq K_C$ is the number of neighbors. ensuring that the importance of each neighbor is properly weighted during aggregation. Based on this, we normalize the selected instance weights using a softmax-like scheme:

$$W_{ik} = \frac{\exp(S_{ik})}{\sum j \in \mathcal{N}_i(k) \exp(S_{ij})}. \tag{6}$$

Next, we perform inductive graph representation learning using SAGEConv (Hamilton, Ying, and Leskovec 2017):

$$B_i^W = \{\boldsymbol{x}_{ik}^W\}_{k=1}^{K_C} \in \mathbb{R}^{K_C \times d_K}, \quad \boldsymbol{x}_{ik}^W = \sum_{j \in \mathcal{N}_i(k)} W_{ik} \boldsymbol{x}_{ik}^R,$$
$$B_i^R = \{\boldsymbol{x}_{ik}^R\}_{k=1}^{K_C} \in \mathbb{R}^{K_C \times d_K},$$
$$= \mathcal{A}_L(\text{SAGEConv}(B_i^S, \mathcal{N}_i(k), W_{ik}, d_R)). \tag{7}$$

Here, the residual path B_i^R preserves transformed instance features, while B_i^W collects the context.

Finally, we use a dual-path gating mechanism (DPGM) to capture the global context in B_i^W and model feature interactions:

$$B_i^D = \mathcal{A}_N \left(B_i^R + \mathcal{A}_L \left(\boldsymbol{g}_i^D \odot \underbrace{\left(\left(B_i^W + B_i^R\right) W_{D_1}\right)}_{\text{sum path}} + (1 - \boldsymbol{g}_i^D) \odot \underbrace{\left(\left(B_i^W \odot B_i^R\right) W_{D_2}\right)}_{\text{product path}} \right) \right),$$
$$\boldsymbol{g}_i^D = \mathcal{A}_S \left((B_i^W W_{D_3} + B_i^R W_{D_4}) W_{D_5} \right), \tag{8}$$

where $\mathcal{A}_S$ is the Sigmoid function, $\mathcal{A}_N$ is the layer norm, $W_{D_1}, W_{D_2} \in \mathbb{R}^{d_K \times d_K}$, $W_{D_3}, W_{D_4} \in \mathbb{R}^{d_K \times \lfloor d_K/2 \rfloor}$ and $W_{D_5} \in \mathbb{R}^{\lfloor d_K/2 \rfloor \times d_K}$ are the learnable parameters, and $\odot$ is the element-wise product. The dual-path gating mechanism enables the model to adaptively balance the contributions of the sum path (captures linear relationships) and the product path (models complex feature interactions). Therefore, DGSL can dynamically determine the importance of each path for each instance, enabling a flexible fusion of local and contextual semantics.

Dual-Path Feature Fusion (DPFF)

Through the combined effects of ABMM and DGSL, the proposed GDF-MIL effectively captures the bag's topological structure and integrates it with contextual information into B_i^D. However, constructing the current graph similarity matrix solely from the bag's hidden representation inevitably obscures some potential connections among the original instances. Therefore, it is essential to balance the trade-off between bag knowledge completeness and graph construction efficiency. Specifically, DPFF completes these two sub-tasks through the bag feature extraction path (BFEP) and the graph feature extraction path (GFEP).

For BFEP, we first extract an attention-based representation from the original bag embedding B_i^E:

$$\boldsymbol{r}_i^A = \sum_{j=1}^{n_i} \alpha_{ij} \boldsymbol{x}_{ij}^E \in \mathbb{R}^{1 \times d_K}, \tag{9}$$

where

$$\alpha_{ij} = \frac{\exp\{(\mathcal{A}_R(\boldsymbol{x}_{ij}^E W_{A_1}) \odot \mathcal{A}_S(\boldsymbol{x}_{ij}^E W_{A_2}))W_{A_3}\}}{\sum_{k=1}^{n_i} \exp\{(\mathcal{A}_R(\boldsymbol{x}_{ik}^E W_{A_1}) \odot \mathcal{A}_S(\boldsymbol{x}_{ik}^E W_{A_2}))W_{A_3}\}}. \tag{10}$$

Here, $\mathcal{A}_R$ is the ReLU function, $W_{A_1}, W_{A_2} \in \mathbb{R}^{d_E \times d_K}$, $W_{A_3} \in \mathbb{R}^{d_K \times 1}$. Note this formulation is derived from ABMIL (Ilse, Tomczak, and Welling 2018), which helps to make up for the information that may be lost during clustering in ABMM and graph abstraction in DGSL. It also provides explicit modeling of instance saliency for classification tasks.

To enhance the model's flexibility to handle more complex feature situations, we design a multi-head attention version of Eq. (9):

$$\boldsymbol{r}_i^M = \text{Softmax}\left(\frac{Q\left(B_i^E\right)^T}{\sqrt{d_E}}\right) B_i^E W_{A_4} \in \mathbb{R}^{1 \times d_K} \tag{11}$$

where $W_{A_4} \in \mathbb{R}^{d_E \times d_K}$. This allows the model to attend to diverse semantics across multiple subspaces, improving robustness in complex bag distributions.

On the other hand, for GFEP, we extract the graph-level representation from the bag by using a graph matching network (Li et al. 2019) to capture both structural and relational patterns not visible in the raw bag feature:

$$\boldsymbol{g}_i^G = \sum_{j=1}^{K} \left(\text{Softmax}\left(\left(\mathcal{A}_L(\boldsymbol{x}_{ij}^D W_{G_1})\right) W_{G_2}\right)\right)^T \boldsymbol{x}_{ij}^D, \tag{12}$$

where $W_{G_1} \in \mathbb{R}^{d_K \times \lfloor d_K/2 \rfloor}, W_{G_2} \in \mathbb{R}^{\lfloor d_K/2 \rfloor \times 1}$.

Finally, we robtain the fused representation of B_i for subsequent classification:

$$\begin{aligned} \boldsymbol{b}_i &= \boldsymbol{g}_i^G \odot \boldsymbol{u}_i + (1 - \boldsymbol{g}_i^G) \odot \boldsymbol{v}_i, \\ \boldsymbol{u}_i &= \mathcal{A}_S\left(\left(\boldsymbol{g}_i^G \| \boldsymbol{g}_i^B\right) W_{B_1}\right), \\ \boldsymbol{v}_i &= \mathcal{A}_L\left(\left(\boldsymbol{g}_i^G \| \boldsymbol{g}_i^B\right) W_{B_2}\right), \end{aligned} \tag{13}$$

where $\|$ is the concatenation operator, $\boldsymbol{g}_i^B \in \{\boldsymbol{g}_i^A, \boldsymbol{g}_i^M\}$ and $W_{B_1} \in \mathbb{R}^{2d_K \times d_K}$, $W_{B_2} \in \mathbb{R}^{2d_K \times d_K}$. Note that the DPGM and the final DPFF in DGSL are functionally distinct. DPGM operates at the instance level, controlling how local neighborhood features interact with residual instance features when constructing node embeddings. In contrast, DPFF operates at the representation level, balancing the complementary strengths of graph structure and raw bag semantics.

Classification and Loss Function

The final classification is performed by a simple FCL:

$$\hat{Y}_i = [p_{i1}, \cdots, p_{iC}] = \mathcal{A}_L\left(\boldsymbol{b}_i W_{C_1}\right) W_{C_2}, \tag{14}$$

where $W_{C_1} \in \mathbb{R}^{d_K \times d_K}$ and $W_{C_2} \in \mathbb{R}^{d_K \times C}$.

The loss function is defined as the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} Y_i \log p_{ic}, \tag{15}$$

Experiments

Experiment Setup

Comparison Algorithms We compared GDF-MIL against 18 state-of-the-art MIL algorithms, including 6 graph-based methods: BagGraph (Pal et al. 2022), DKMIL (Zhang et al. 2024), MIL-GNN (Tu et al. 2019), RGMIL (Zhao et al. 2024), TAD-Graph (Wang et al. 2025), and WiKG (Li et al. 2024). The remaining algorithms are ABMIL (Ilse, Tomczak, and Welling 2018), CAMIL (Fourkioti, De Vries, and Bakal 2024), CLAM-MB (Lu et al. 2021), CLAM-SB (Lu et al. 2021), DGR-MIL (Zhu et al. 2024), DSMIL (Li, Li, and Eliceiri 2021), DTFD-MIL (Zhang et al. 2022a), FRMIL (Chikontwe et al. 2024), GAMIL (Ilse, Tomczak, and Welling 2018), MaxMIL (Shao et al. 2021), MeanMIL (Shao et al. 2021), TransMIL (Shao et al. 2021). Except for uniformly setting the number of training epochs to 100, all other parameters were configured according to the optimal settings reported in the original papers.

Datasets The experiments cover 24 datasets in 4 domains: text (Zhou, Sun, and Li 2009), web (Wang et al. 2019), medicine (musk1 and musk2 (Dietterich, Lathrop, and Lozano-Pérez 1997)), and image (elephant, tiger, fox (Wei et al. 2019)), which comprehensively evaluate key indicators such as classification performance and computation cost of the proposed method and rivals. Please note that we only present comprehensive experiments on four datasets in the main text. Please refer to the Appendix for the remaining experiments.

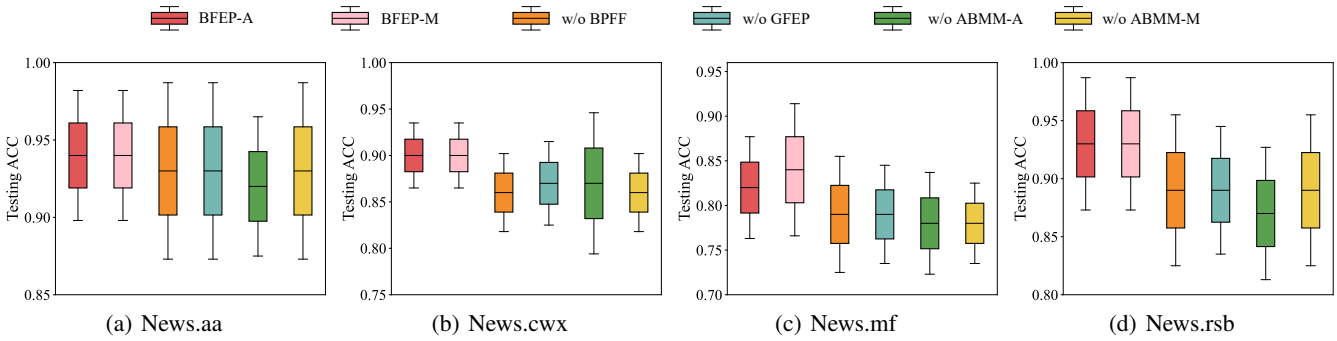
Algorithm	News.aa			News.cwx		
	ACC	F1-Score	AUC	ACC	F1-Score	AUC
ABMIL	$.870 \pm .076$	$.862 \pm .083$	$.861 \pm .087$	$.870 \pm .084$	$.866 \pm .088$	$.873 \pm .086$
ACMIL	$.570 \pm .196$	$.483 \pm .255$	$.607 \pm .153$	$.570 \pm .164$	$.518 \pm .195$	$.624 \pm .110$
BagGraph†	$.850 \pm .071$	$.846 \pm .071$	$.851 \pm .064$	$.820 \pm .067$	$.796 \pm .102$	$.799 \pm .105$
CAMIL	$.850 \pm .100$	$.843 \pm .110$	$.850 \pm .111$	$.820 \pm .076$	$.813 \pm .088$	$.825 \pm .092$
CLAM-MB	$.850 \pm .087$	$.846 \pm .088$	$.853 \pm .085$	$.770 \pm .091$	$.760 \pm .097$	$.775 \pm .096$
CLAM-SB	$.840 \pm .065$	$.834 \pm .068$	$.841 \pm .065$	$.790 \pm .074$	$.782 \pm .078$	$.803 \pm .066$
DGR-MIL	$.890 \pm .074$	$.885 \pm .081$	$.890 \pm .082$	$.870 \pm .057$	$.855 \pm .079$	$.849 \pm .084$
DKMIL†	$.670 \pm .168$	$.630 \pm .206$	$.678 \pm .119$	$.810 \pm .042$	$.802 \pm .049$	$.811 \pm .057$
DSMIL	$.510 \pm .167$	$.403 \pm .222$	$.560 \pm .133$	$.570 \pm .179$	$.449 \pm .227$	$.577 \pm .145$
DTFD-MIL	$.800 \pm .127$	$.782 \pm .142$	$.800 \pm .121$	$.800 \pm .122$	$.771 \pm .146$	$.778 \pm .135$
FRMIL	$.720 \pm .045$	$.698 \pm .047$	$.715 \pm .035$	$.710 \pm .089$	$.680 \pm .117$	$.717 \pm .099$
MaxMIL	$.890 \pm .082$	$.885 \pm .086$	$.885 \pm .090$	$.830 \pm .027$	$.822 \pm .039$	$.830 \pm .047$
MeanMIL	$.660 \pm .185$	$.558 \pm .261$	$.643 \pm .194$	$.750 \pm .272$	$.728 \pm .308$	$.796 \pm .209$
RGMIL†	$.890 \pm .055$	$.887 \pm .058$	$.892 \pm .063$	$.790 \pm .074$	$.778 \pm .085$	$.783 \pm .087$
TAD-Graph†	$.510 \pm .089$	$.354 \pm .071$	$.511 \pm .025$	$.520 \pm .115$	$.363 \pm .097$	$.514 \pm .032$
TransMIL	$.610 \pm .108$	$.577 \pm .094$	$.607 \pm .078$	$.650 \pm .061$	$.558 \pm .124$	$.601 \pm .077$
WiKG†	$.890 \pm .055$	$.887 \pm .057$	$.888 \pm .061$	$.880 \pm .027$	$.871 \pm .039$	$.871 \pm .051$
GDF-MIL (Ours)	$\mathbf{.940 \pm .042}$	$\mathbf{.938 \pm .043}$	$\mathbf{.937 \pm .045}$	$\mathbf{.900 \pm .035}$	$\mathbf{.895 \pm .043}$	$\mathbf{.897 \pm .048}$
Algorithm	**News.mf**			**News.rsb**		
	ACC	**F1-Score**	**AUC**	**ACC**	**F1-Score**	**AUC**
ABMIL	$.700 \pm .146$	$.685 \pm .163$	$.711 \pm .121$	$.860 \pm .114$	$.857 \pm .116$	$.867 \pm .099$
ACMIL	$.490 \pm .074$	$.342 \pm .048$	$.509 \pm .020$	$.680 \pm .189$	$.651 \pm .234$	$.722 \pm .149$
BagGraph†	$.720 \pm .125$	$.693 \pm .129$	$.721 \pm .106$	$.890 \pm .082$	$.889 \pm .082$	$.893 \pm .073$
CAMIL	$.710 \pm .114$	$.701 \pm .113$	$.710 \pm .115$	$.830 \pm .115$	$.826 \pm .118$	$.845 \pm .092$
CLAM-MB	$.660 \pm .129$	$.598 \pm .187$	$.654 \pm .116$	$.850 \pm .106$	$.844 \pm .109$	$.852 \pm .089$
CLAM-SB	$.650 \pm .100$	$.537 \pm .179$	$.605 \pm .110$	$.850 \pm .106$	$.847 \pm .106$	$.860 \pm .082$
DGR-MIL	$.740 \pm .114$	$.685 \pm .190$	$.709 \pm .137$	$.910 \pm .082$	$.909 \pm .082$	$.913 \pm .072$
DKMIL†	$.750 \pm .087$	$.737 \pm .089$	$.751 \pm .072$	$.860 \pm .102$	$.858 \pm .103$	$.869 \pm .082$
DSMIL	$.560 \pm .129$	$.413 \pm .172$	$.544 \pm .099$	$.520 \pm .115$	$.427 \pm .172$	$.570 \pm .103$
DTFD-MIL	$.590 \pm .178$	$.537 \pm .238$	$.633 \pm .149$	$.770 \pm .076$	$.760 \pm .077$	$.782 \pm .057$
FRMIL	$.620 \pm .076$	$.571 \pm .076$	$.618 \pm .048$	$.750 \pm .079$	$.733 \pm .075$	$.754 \pm .048$
MaxMIL	$.700 \pm .079$	$.686 \pm .090$	$.702 \pm .082$	$.860 \pm .119$	$.859 \pm .119$	$.867 \pm .103$
MeanMIL	$.580 \pm .179$	$.505 \pm .227$	$.617 \pm .143$	$.650 \pm .166$	$.617 \pm .206$	$.685 \pm .121$
RGMIL†	$.790 \pm .074$	$.779 \pm .079$	$.791 \pm .074$	$.850 \pm .061$	$.844 \pm .063$	$.846 \pm .064$
TAD-Graph†	$.480 \pm .076$	$.323 \pm .035$	$.500 \pm .000$	$.600 \pm .050$	$.437 \pm .110$	$.542 \pm .066$
TransMIL	$.650 \pm .079$	$.597 \pm .103$	$.614 \pm .094$	$.550 \pm .061$	$.496 \pm .094$	$.549 \pm .049$
WiKG†	$.740 \pm .055$	$.732 \pm .053$	$.737 \pm .058$	$.870 \pm .091$	$.869 \pm .090$	$.874 \pm .081$
GDF-MIL (Ours)	$\mathbf{.840 \pm .074}$	$\mathbf{.833 \pm .075}$	$\mathbf{.833 \pm .076}$	$\mathbf{.930 \pm .057}$	$\mathbf{.929 \pm .057}$	$\mathbf{.933 \pm .049}$

Table 1: Performance comparison of GDF-MIL with 18 rivals on text datasets. The symbols † indicate the graph-based methods. For experiments on the remaining 20 datasets, please refer to the Appendix.

Implementations For our GDF-MIL, we set the number of clusters K_C to $\{10, 20, 50, 100, 200\}$, the number of neighbors $K_N \leq K_C$ to $\{10, 20, 50, 100, 200\}$. The nunber of nodes d_E and d_K in the FCLs were fixed to 256 and 128, respectively. The learning rate was set to 0.0002. The optimizer was AdamW with a weight decay of $1e^{-5}$. The temperature parameter τ was set to 1. The random seed was set to 3407. The evaluation metrics include accuracy (ACC), F1-score, and AUC. The evaluation strategy was the five-fold cross validation, and its mean and standard deviation were recorded. The model was implemented using PyTorch and trained on a single NVIDIA RTX 5060TI GPU.

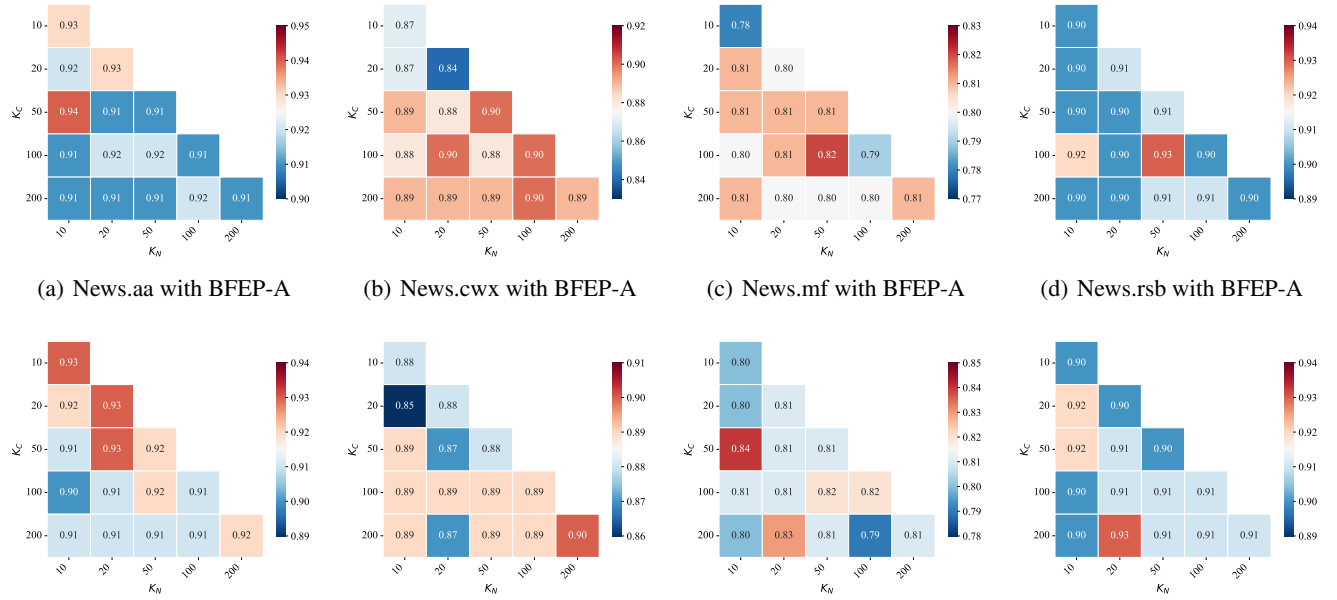
Performance Comparison

The performance of GDF-MIL and the comparison algorithms on the 4 datasets is shown in Table 1. The results show that GDF-MIL achieves the best performance on all datasets, demonstrating its strong generalization capability. For example, on the news.aa dataset, GDF-MIL outperformed all baselines by approximately 4% in terms of ACC, F1-score, and AUC. On news.rsb, it achieved perfect classification performance (100% across all metrics). The reason for such achievements is that GDF-MIL fully integrates the strengths of existing intra-bag context feature extraction and bag topology structure mining methods. By achieving



(a) News.aa (b) News.cwx (c) News.mf (d) News.rsb

Figure 3: Ablation study for the components of GDF-MIL.



(a) News.aa with BFEP-A (b) News.cwx with BFEP-A (c) News.mf with BFEP-A (d) News.rsb with BFEP-A

(e) News.aa with BFEP-M (f) News.cwx with BFEP-M (g) News.mf with BFEP-M (h) News.rsb with BFEP-M

Figure 4: Parameter analysis for the key parameters of GDF-MIL.

an adaptive balance between these two paths, it enhances classification performance and enables efficient model construction. Combined with the appendix, we can see that half of the current competitors, especially graph-based methods, have poor scalability. For example, MIL-GNN, RGMIL, and TAD-Graph cannot perform effective classification on the Web dataset. This may be due to the high dimensionality and sparsity of this dataset. For the musk dataset, the best algorithms are ABMIL and DTFD-MIL, which are relatively early MIL studies and can effectively capture the key knowledge in the bag. Nevertheless, GDF-MIL remains one of the top-performing methods on this dataset.

Ablation Study

GDF-MIL comprises three core components: ABMM for compact instance representations, DGSL (the indispensable main body) for topological structure learning, and DPFF for adaptively fusing bag-level and graph-level representations. DPFF further contains two feature extraction mechanisms: attention-based and multi-head attention-based. Therefore, to ensure clarity and validity of each component, we designed six ablation strategies, as shown in Figure 3: a) BFEP-A: GDF-MIL with the attention-based BFEP; b) BFEP-M: GDF-MIL with the multi-head attention-based BFEP; c) w/o BPFF: GDF-MIL without the BPFF; d) w/o ABMM: GDF-MIL without the GFEP; f) w/o ABMM-A: attention-based GDF-MIL without the ABMM; g) w/o ABMM-M: multi-head attention-based GDF-MIL without the ABMM.

Experiment results show that BFEP-A and BFEP-B achieve the best testing ACC on four representative datasets, along with improved model stability, as indicated by the lower standard deviations achieved across multiple independent ones. These two strategies correspond to the two

Algorithm	News.aa	News.cwx	News.mf	News.rsb
ABMIL	193.66	178.14	195.80	185.08
ACMIL	535.15	320.22	334.15	309.53
CAMIL	1148.24	842.11	854.52	1125.30
CLAMMB	480.08	478.70	495.30	451.70
CLAMSB	218.24	226.38	222.29	218.25
DGRMIL	2032.36	1284.18	1327.61	1259.72
DSMIL	300.22	262.68	273.23	255.10
DTFDMIL	909.03	934.07	952.37	952.54
FRMIL	378.53	407.54	428.41	410.38
MaxMIL	162.50	150.35	168.76	143.03
MeanMIL	157.70	153.83	172.22	140.58
TransMIL	1604.29	1236.43	1271.74	1221.36
BagGraph†	732.42	709.79	683.39	686.18
DKMIL†	1155.68	1137.49	1157.61	1141.60
RGMIL†	10404.10	9615.16	10029.53	10342.26
TADGraph†	1737.73	1761.76	1829.61	1770.48
WiKG†	3056.30	3262.06	3323.66	3273.81
GDF-MIL	773.78	708.70	725.17	704.92

Table 2: Time cost comparison on the text datasets. Note that the algorithms marked with † are graph-based algorithms. In addition, MILGNN cannot complete the classification task on the text datasets.

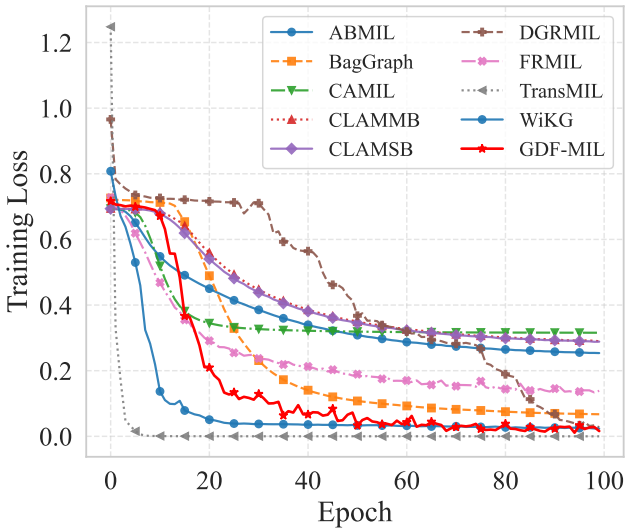


Figure 5: Convergence analysis of GDF-MIL on the News.aa dataset. WiKG and TransMIL currently have the best convergence, but suffer from serious overfitting on text datasets. Note that some algorithms with poor convergence are not shown.

branches of GDF-MIL. Among them, BFEP-B performs slightly better than BFEP-A, likely due to its use of a multi-head self-attention mechanism, which enables the extraction of more informative features. For the remaining componentes, the ACC of the algorithm decreases, especially after removing ABMM. This highlights the positive contribution of each component to the overall performance of GDF-MIL.

Parameter Analysis

The most key parameter in GDF-MIL is K_C, which controls the number of soft cluster centers and directly affects the effectiveness of compact representation extraction. Another important parameter is the number of neighbors, K_N, which controls the granularity of subsequent information filtering. Therefore, this section focuses on analyzing the impact of these two parameters on the classification performance of GDF-MIL, as shown in Figure 4. Experiment results indicate that setting both K_C and K_N to 10 generally decreases classification performance. Among eight independent experiments, only the News.aa with BFEP-M achieved optimal accuracy. However, reducing the cardinality of the bag from n_i to K_c inevitably leads to information loss. Here, DPFF mitigates this risk by directly reusing the original instance embeddings through BFEP. This explains the trend in Figure 4, where a moderate K_c can achieve a good balance between compression and semantic preservation.

Time Cost Comparison

We conducted time cost comparison experiments across 24 datasets in four domains: text, web, image, and medicine (The main paper can only show 4 data sets). Due to the varying characteristics of these datasets, such as high feature sparsity and dimensionality, we evaluated each domain separately. The experiment uses 5-fold cross validation, and each fold validation is run 100 times. The final running time is the average running time (ms) of each epoch, as shown in Table 2. Experiment results show that GDF-MIL achieves comparable runtime efficiency to BagGraph, the fastest existing graph-based MIL method, while significantly outperforms high-performance baselines such as WiKG and RGMIL. In addition, ABMIL shows an excellent time advantage over other algorithms, which is unmatched by any other algorithm. However, GDF-MIL maintains a reasonable runtime compared to other strong baselines, and the additional cost is well compensated by its significant performance gains. The time cost of GDF-MIL is significantly lower than that of existing graph-based MIL methods and is competitive with other comparison algorithms in most cases.

Conclusion

We propose a novel architecture, GDF-MIL, distinct from any existing graph-based MIL methods. This approach centers on the efficient extraction of compact representations and fully mines the graph structure, followed by the adaptive fusion of bag- and graph-level features. Experiment results across 24 datasets demonstrate the superiority of GDF-MIL, with a significant breakthrough on text datasets. These prove the scalability and transferability of GDF-MIL across multiple domains. However, GDF-MIL contains two bag-level feature extraction strategies, and we can only show its adaptability to different datasets through experiments. These will limit the broader adoption of GDF-MIL in more application fields. Consequently, in future work, we will focus on adaptive information filtering and multi-path feature fusion to propose a more comprehensive MIL framework.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62131016). The code and appendix can be found at https://github.com/InkiYinji/GDF-MIL-AAAI26. All datasets used can be found at https://palm.seu.edu.cn/zhangml/ and https://www.lamda.nju.edu.cn/CH.Data.ashx.

References

Chikontwe, P.; Kim, M.; Jeong, J.; Sung, H. J.; Go, H.; Nam, S. J.; and Park, S. H. 2024. FR-MIL: Distribution re-calibration based multiple instance learning with transformer for whole slide image classification. *IEEE Transactions on Medical Imaging*, 1–10.

Cui, X.; Chen, W.; and Su, J. 2025. A multiscale frequency domain causal framework for enhanced pathological analysis. In *ICLR*, 1–16.

Dietterich, T. G.; Lathrop, R. H.; and Lozano-Pérez, T. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2): 31–71.

Fourkioti, O.; De Vries, M.; and Bakal, C. 2024. CAMIL: Context-Aware Multiple Instance Learning for Cancer Detection and Subtyping in Whole Slide Images. In *ICLR*, 1–16.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *NeruIPS*, 1–11.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *ICML*, 2127–2136.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical reparametrization with Gumble-Softmax. In *ICLR*, 1–12.

Kapse, S.; Pati, P.; Das, S.; Zhang, J.; Chen, C.; Vakalopoulou, M.; Saltz, J.; Samaras, D.; Gupta, R. R.; and Prasanna, P. 2024. SI-MIL: Taming deep mil for self-interpretability in gigapixel histopathology. In *CVPR*, 11226–11237.

Li, B.; Li, Y.; and Eliceiri, K. W. 2021. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *CVPR*, 14318–14328.

Li, J.; Chen, Y.; Chu, H.; Sun, Q.; Guan, T.; Han, A.; and He, Y. 2024. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *CVPR*, 11323–11332.

Li, Y.; Gu, C.; Dullien, T.; Vinyals, O.; and Kohli, P. 2019. Graph matching networks for learning the similarity of graph structured objects. In *ICML*, 3835–3845.

Ling, X.; Ouyang, M.; Wang, Y.; Chen, X.; Yan, R.; Chu, H.; Cheng, J.; Guan, T.; Tian, S.; Liu, X.; et al. 2024. Agent aggregator with mask denoise mechanism for histopathology whole slide image analysis. In *ACM MM*, 2795–2803.

Lu, M. Y.; Williamson, D. F.; Chen, T. Y.; Chen, R. J.; Barbieri, M.; and Mahmood, F. 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6): 555–570.

Pal, S.; Valkanas, A.; Regol, F.; and Coates, M. 2022. Bag graph: Multiple instance learning using bayesian graph neural networks. In *AAAI*, 7922–7930.

Qu, L.; Ma, Y.; Luo, X.; Guo, Q.; Wang, M.; and Song, Z. 2024. Rethinking multiple instance learning for whole slide image classification: A good instance classifier is all you need. *IEEE Transactions on Circuits and Systems for Video Technology*, 9732–9744.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X.; and Zhang, Y. 2021. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *NeruIPS*, 2136–2147.

Tang, W.; Zhang, W.; and Zhang, M.-L. 2024. Exploiting conjugate label information for multi-instance partial-label learning. *IJCAI*, 1–9.

Tu, M.; Huang, J.; He, X.; and Zhou, B. 2019. Multiple instance learning with graph neural networks. In *ICML Workshop*, 1–9.

Wang, F.; Xin, J.; Zhao, W.; Jiang, Y.; Yeung, M.; Wang, L.; and Yu, L. 2025. TAD-Graph: Enhancing Whole Slide Image Analysis via Task-Aware Subgraph Disentanglement. *IEEE Transactions on Medical Imaging*, 1–13.

Wang, X.; Yan, Y.; Tang, P.; Liu, W.; and Guo, X. 2019. Bag similarity network for deep multi-instance learning. *Information Sciences*, 504: 578–588.

Wei, X.-S.; Ye, H.-J.; Mu, X.; Wu, J.; Shen, C.; and Zhou, Z.-H. 2019. Multi-instance learning with emerging novel class. *IEEE Transactions on Knowledge and Data Engineering*, 33(5): 2109–2120.

Wu, B.; Wang, Z.; Lin, X.; Xu, J.; Yu, J.; Shicheng, Z.; Chen, H.; and Hu, L. 2025. Distributed parallel gradient stacking (DPGS): Solving whole slide image stacking challenge in multi-instance learning. In *ICML*, 1–11.

Wu, J.; Pan, S.; Zhu, X.; Zhang, C.; and Wu, X. 2018. Multi-instance learning with discriminative bag mapping. *IEEE Transactions on Knowledge and Data Engineering*, 30(6): 1065–1080.

Xiao, Y.; Liu, B.; and Hao, Z. 2024. Multi-instance nonparallel tube learning. *IEEE Transactions on Neural Networks and Learning Systems*, 1–12.

Yang, M.; Zhang, Y.-X.; Wang, X.; and Min, F. 2021. Multi-instance ensemble learning with discriminative bags. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(9): 5456–5467.

Zhang, H.; Meng, Y.; Zhao, Y.; Qiao, Y.; Yang, X.; Coupland, S. E.; and Zheng, Y. 2022a. DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *CVPR*, 18802–18812.

Zhang, W.; Liu, L.; and Li, J. 2020. Robust multi-instance learning with stable instances. In *ECAI*, 1682–1689.

Zhang, W.; Zhang, X.; Deng, H.-W.; and Zhang, M.-L. 2022b. Multi-instance causal representation learning for instance label prediction and out-of-distribution generalization. In *NeurIPS*, 34940–34953.

Zhang, Y.-X.; Zhou, Z.; He, X.; Adhikary, A. R.; and Dutta, B. 2024. Data-driven knowledge fusion for deep multi-instance learning. *IEEE Transactions on Neural Networks and Learning Systems*, 8292–8306.

Zhao, X.; Dai, Q.; Bai, X.; Wu, J.; Peng, H.; Peng, H.; Yu, Z.; and Philip, S. Y. 2024. Reinforced GNNs for multiple instance learning. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15.

Zhao, Z.; Chen, K.; and Zhao, J. 2025. RPMIL: Rethinking uncertainty-aware probabilistic multiple instance learning for whole slide pathology diagnosis. In *IJCAI*, 2467–2475.

Zheng, T.; Jiang, K.; Yao, H.; Xiao, Y.; and Wang, Z. 2025a. OODML: Whole slide image classification meets online pseudo-supervision and dynamic mutual learning. In *AAAI*, 10626–10634.

Zheng, T.; Yao, H.; Jiang, K.; Xiao, Y.; and Zhao, S. 2025b. GMMamba: Group masking Mamba for whole slide image classification. In *ICCV*, 9935–9944.

Zhong, H.; Ding, M.; Zhao, C.; Zhang, Y.; Wang, T.; and Lei, B. 2025. MSMMIL: Multi-scan Mamba-based multiple instance learning for whole slide image classification. *Knowledge-Based Systems*, 324: 113871.

Zhou, Z.-H.; Sun, Y.-Y.; and Li, Y.-F. 2009. Multi-instance learning by treating instances as non-iid samples. In *ICML*, 1249–1256.

Zhu, W.; Chen, X.; Qiu, P.; Sotiras, A.; Razi, A.; and Wang, Y. 2024. DGR-MIL: Exploring diverse global representation in multiple instance learning for whole slide image classification. In *ECCV*, 333–351.