Revisiting Out-of-distribution Detection in Multi-instance Learning

Jiahui Zhang[1], Yukun Cui[1], Mei Yang[2*], Zhen Pan[3], Xuemei Cao[4], and Yu-Xuan Zhang[5]
[1]School of Computing and Communications, Lancaster University
[2]School of Computer Science and Software Engineering, Southwest Petroleum University
[3]School of Computer Science, Chengdu Normal University
[4]School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics
[5]School of Information Science and Technology, Southwest Jiaotong University
j.zhang105@lancaster.ac.uk, y.cui8@lancaster.ac.uk, yangmei@swpu.edu.cn, 151113@cdnu.edu.cn
caoxuemei.qpz@gmail.com, inki.yinji@gmail.com

***Abstract*—Out-of-distribution (OOD) detection is a critical step in ensuring the security and reliability of machine learning systems. However, with the surge of large-scale datasets and complex applications, the efficacy of traditional OOD methods is increasingly limited under weak supervision, and their adaptation to the popular multi-instance learning (MIL) framework proves challenging. Existing MIL-OOD methods often either directly adapt traditional OOD techniques or only perform transfer learning for OOD samples (bags), thereby overlooking the inherent hierarchical structure of instances within each bag in the MIL context. As a result, they struggle to distinguish in-distribution (ID) and OOD bags effectively. To address this challenge, we propose a clustering-aware MIL-OOD detection method (CluMILOOD). This method leverages the MIL attention mechanism to jointly model both the fused intra-bag feature (FIBF) and the clustered inter-bag feature (CIBF), capturing both local and global structure. We design four OOD scoring strategies, i.e., maximum distance, minimum distance, mean distance, and weighted average distance, to evaluate the unknown bags. We conducted experiments on three OOD detection tasks and compared them with 10 state-of-the-art OOD detection methods. The results show that CluMILOOD significantly outperforms rivals in overall performance, demonstrating stronger detection capabilities and stability.**

***Index Terms*—Multi-instance learning, Out-of-distribution, Clustering-aware, Detection.**

I. INTRODUCTION

Out-of-distribution (OOD) detection is a critical step in ensuring the reliability and security of machine learning systems, particularly in high-risk scenarios and complex decision-making systems [1]. In this situation, models inevitably encounter inputs outside the training distribution, known as OOD samples. The failure to promptly detect these samples can lead to severe decision-making errors and introduce systemic risks. Traditional OOD research has made significant progress in fully supervised tasks [2], [3]. However, as data scales and application complexity continue to grow, these limitations become increasingly pronounced, particularly in the domain of weakly supervised modeling.

This work was supported by the Nanchong Municipal Government-Universities Scientific Cooperation Project (23XNSYSX0084) and Chengdu Normal University Research Projects (YJRC202449).

Multi-instance learning (MIL) is a key framework for weakly supervised learning and has been widely applied to practical tasks such as image classification [4]–[6] and pathological analysis [7]–[9]. In MIL, each sample is referred to as a bag, and the label is provided only at the bag-level, lacking instance-level supervision. While maintaining flexibility, this also poses challenges for OOD detection. Existing MIL-OOD research falls into two main categories [10], [11]: one directly transfers traditional OOD methods to bag-level tasks, and the other utilizes OOD data for transfer learning. However, these methods fail to fully utilize instance features within bags or capture the global distribution information across them. Consequently, they struggle to effectively differentiate between in-distribution (ID) and OOD bags.

To address these problems, this paper proposes a clustering-aware MIL-OOD detection method (CluMILOOD), as shown in Fig. 1. Unlike existing methods that only consider model output logits, CluMILOOD extracts fused intra-bag features (FIBFs) from bags based on the attention mechanism while simultaneously modeling the clustered inter-bag features (CIBFs). This method jointly captures local and global information and incorporates four detection strategies to measure the difference between the FIBF of the unknown bag and the CIBFs of ID bags, thereby achieving higher performance in OOD detection.

The main contributions of this paper are as follows:

1) To the best of our knowledge, we propose CluMILOOD, the first clustering-based posterior OOD detection method for the MIL framework. Unlike directly transferring traditional OOD detection criteria, CluMILOOD explicitly utilizes the bag-level structure of MIL, constructing a global reference in the bag-level space that can be used for OOD detection.
2) CluMILOOD collaboratively combines FIBFs and CIBFs for OOD scoring and designs four complementary scoring strategies to characterize the distance between the unknown bag and CIBFs from different perspectives, thereby achieving stable and generalizable OOD detection.
3) We systematically evaluated several MIL-OOD tasks,

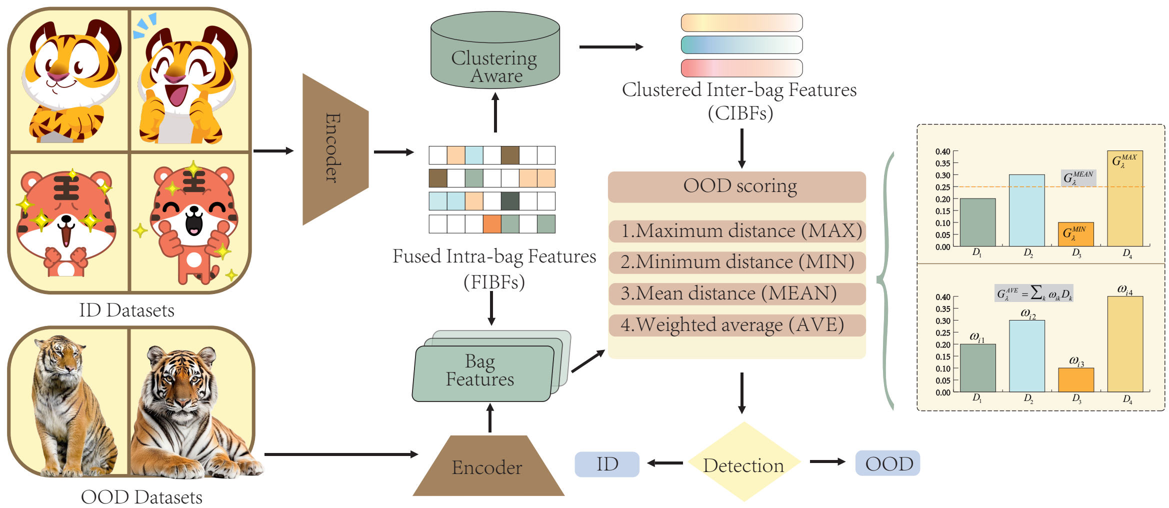


Fig. 1. The overview of the proposed CluMILOOD. To reliably detect OOD bags, we first establish a global reference on the complete ID dataset. Specifically, all ID bags are processed by an encoder and attention mechanism to obtain fused intra-bag features (FIBFs). Then, we employ a clustering-aware strategy to infer clustered inter-bag features (CIBFs) from FIBFs, thereby characterizing typical patterns in the ID data. Based on the distance relationship between these features, four complementary OOD scoring strategies are designed to more robustly detect ID and OOD bags.

covering various MIL backbones, and compared them with 10 state-of-the-art OOD detectors. The results show that CluMILOOD achieves significant advantages in most cases.

To ensure reproducibility, our code and datasets will be made publicly available upon acceptance of this paper.

II. Related Work

A. Out-of-distribution (OOD) Detection

OOD detection aims to identify input samples outside the training distribution, preventing the model from making overconfident and inaccurate predictions for unknown samples [1]. This task is particularly important in high-risk applications such as autonomous driving and medical diagnosis. In recent years, extensive studies have systematically investigated OOD detection. Salehi et al. [12] described OOD from the perspectives of anomaly detection and open set recognition. Zhang et al. [2] reformulated the logit model as cosine similarity and logit norm, recommending MaxCosine and MaxNorm. Wei et al. [3] built detectors with clearer decision boundaries by training on OOD data with virtual labels. Furthermore, OOD detection has been expanded to specific fields, such as natural language processing and OOD methods for large-scale pretrained models [13]. With the emergence of new paradigms such as test-time learning [1], the definition and application of OOD have continued to expand, and the methodological framework has continued to improve.

B. Multi-instance Learning (MIL)

MIL is a typical weakly supervised learning framework widely used in tasks such as medical analysis, video anomaly detection, and drug discovery [4], [8]. Unlike traditional supervised learning, MIL organizes samples into bags with multiple instances, provides only bag-level labels, and lacks instance-level supervision. This characteristic gives it a natural advantage in tasks where labeling is expensive or difficult to perform independently. Early methods often relied on the basic assumption that a positive bag contains at least one positive instance. With the development of deep learning, researchers began to focus on attention mechanisms. ABMIL is an approach that utilizes attention weights to select key instances [14]. MIPLGP [15] assigns each instance a candidate label set in an augmented label space and then transforms them into a logarithmic space to yield the disambiguated and continuous labels. StructMIL [16] proposed a structure-aware and prototype-driven framework for robust and interpretable cancer detection and grading. Subsequently, scholars extended this method and proposed to model graph structure, cluster mechanism, and even multimodal fusion to capture more intra-bag relationships and inter-bag distribution [17], [18]. Although several studies have attempted to extend MIL for OOD detection, two challenges still exist. Firstly, when directly extending existing methods to MIL, the hierarchical relationship of instances in a bag is usually ignored. Second, some methods attempt to directly classify OOD bags through transfer learning rather than detecting them. Therefore, this paper aims to develop a MIL-suitable OOD detection method, providing a more comprehensive methodology and detection framework for subsequent research.

III. Method

A. Problem Statement

Out-of-distribution (OOD) detection [3], [10] is a crucial method to enhance the reliability and security of machine learning systems. For instance, in the domain of autonomous driving, when the model encounters an anomalous scene or object, it is imperative to issue an early warning and defer to human judgment. In this process, these anomalous objects are classified as "unknowns" and are referred to as OOD samples, while others are categorized as in-distribution (ID) samples.

Existing OOD strategies remain underexplored within weakly supervised paradigms, particularly in multi-instance learning (MIL), which stands as one of its most representative frameworks. In MIL, each sample is a bag $B_i = \{\boldsymbol{x}_{ij}\}_{j=1}^{n_i}$ containing n_i instances $\boldsymbol{x}_{ij} \in \mathbb{R}^D$, where D is the dimension. In the MIL setting, bag-level labels Y_i are given but instance-level annotations y_{ij} are missing, a feature that makes this setting distinct for OOD detection. Current OOD detection methods in MIL are plagued by two limitations: a) they often only leverage transfer learning to enhance model robustness to OOD bags, rather than explicitly detecting them; and b) they directly migrate traditional OOD methods, a strategy that ignores the bag hierarchy and is thus unsuitable for the MIL paradigm. As a result, this paper attempts to address these deficiencies in the system by providing an OOD detection method for MIL data.

B. OOD Detection in MIL

A reliable OOD detector should be capable of effectively distinguishing ID and OOD bags:

$$\mathcal{D}_\lambda(B_i) = \begin{cases} \text{ID}, & G(B_i) \geq \lambda; \\ \text{OOD}, & otherwise, \end{cases} \tag{1}$$

where higher scores indicate ID bags, and λ is the empirical threshold, specifically set to correctly recall 95% of ID validation data when computing the FPR@95 metric. Among the existing methods, the most direct strategy is to use the BASELINE [10] in the traditional OOD method to implement Eq. (1):

$$\mathcal{D}_\lambda(B_i) = \max_c \frac{\exp(z_{ic}/T)}{\sum_{j=1}^{C} \exp(z_{ij}/T)}, \tag{2}$$

where $z_{ic} \in Z_i = \mathcal{M}(B_i) = \{z_{ic}\}_{c=1}^{C}$ is the logit of the trained MIL model $\mathcal{M}$ (e.g. ABMIL [14] and WiKG [17]), T is the temperature, C is the number of classes, Z_i is the C-dimensional feature used to predict the bag label, i.e., $\hat{Y}_i = f_c(\text{Softmax}(Z_i))$, and $f_c(\cdot)$ is the classification function. In addition, some recent studies have also used methods such as SCALE [19] to enhance the expressiveness and feature discrimination of Eq. (2) by scaling the original bag:

$$B_i := \exp\left(\frac{\sum_{j,d} \boldsymbol{x}_{ij}}{\text{Top-}K(\sum_{j,d} \boldsymbol{x}_{ij})}\right) B_i, \tag{3}$$

where $\boldsymbol{x}_{ij} = [x_{ij}^1, \ldots, x_{ij}^D]$ and Top-K means taking the top K largest values of all elements in B_i. However, these two strategies are agnostic to the inherent characteristics of MIL, relying solely on logit values to form a simplistic criterion for OOD detection.

To this end, we propose a clustering-aware OOD method (CluMILOOD), which simultaneously uses the fused intra-bag feature (FIBF) and the clustered inter-bag feature (CIBF) to output the OOD score of the bag. For instance, leveraging a prevalent MIL attention mechanism [14], the FIBF can be calculated as:

$$\boldsymbol{b}_i = \sum_{j=1}^{n_i} \alpha_{ij} \boldsymbol{x}_{ij}, \tag{4}$$

where

$$\alpha_{ij} = \text{Softmax}((\text{Tanh}(\boldsymbol{x}_{ij}U) \odot \text{Sigmoid}(\boldsymbol{x}_{ij}V))W). \tag{5}$$

Here, U, V, W are learnable parameters, $\odot$ is the element-wise multiplication, and in MIL, x_{ij} is first passed through a feature encoder to obtain its high-level representation. In the training phase of OOD methods, all training bags are IDs. Thus, for a given dataset $\mathcal{S} = \{B_i\}_{i=1}^N$, we can collect $\boldsymbol{b}_i$ corresponding to all bags, recorded as $\mathcal{B} = \{\boldsymbol{b}_i\}_{i=1}^N$, where N is the cardinality of $\mathcal{S}$. Based on the K-Means clustering algorithm, $\mathcal{B}$ can be divided into K disjoint clusters $\mathcal{C}_k \subset \mathcal{B}$, and the CIBF of $\mathcal{C}_k$ can be expressed as the center $\boldsymbol{c}_k$, which records the inter-bag features of the entire ID dataset under clustering-aware.

Having derived the FIBFs and CIBFs, we now possess two feature sets that encapsulate key intra-bag and inter-bag information, respectively. To judge ID bags and OOD bags, we propose four sub-strategies:

1) **Maximum distance (MAX)**:

$$G_\lambda^{MAX}(B_i^U) = \max_k \|\boldsymbol{b}_i^U - \boldsymbol{c}_k\|_2. \tag{6}$$

2) **Minimum distance (MIN)**:

$$G_\lambda^{MIN}(B_i^U) = \min_k \|\boldsymbol{b}_i^U - \boldsymbol{c}_k\|_2. \tag{7}$$

3) **Mean distance (MEAN)**:

$$G_\lambda^{MEAN}(B_i^U) = \text{mean}_k \|\boldsymbol{b}_i^U - \boldsymbol{c}_k\|_2. \tag{8}$$

4) **Weighted Average (AVE)**:

$$G_\lambda^{AVE}(B_i^U) = \sum_{k=1}^{K} \omega_{ik} \|\boldsymbol{b}_i^U - \boldsymbol{c}_k\|_2, \tag{9}$$

$$\omega_{ik} = \text{Softmax}(\exp(-\|\boldsymbol{b}_i^U - \boldsymbol{c}_k\|_2)). \tag{10}$$

It should be noted that during testing, the OOD properties of a given bag are unknown. Therefore, we denote this bag as B_i^U and its corresponding FIBF as $\boldsymbol{b}_i^U$.

IV. Experiments

A. Experimental Design

1) OOD Detectors for Comparison: As there is a notable absence of OOD detectors specifically designed for MIL, we follow the latest MIL-OOD strategy [20] and migrate 10 state-of-the-art traditional OOD methods to MIL, including BASELINE [10], DICE [21], EBO [22], IODIN [23], MLS [24], NCI [25], KL-Matching [24], ODIN [26], SCALE [19], and VIM [27]. In addition, only MLS has no parameters, and the other parameter settings are entirely based on OpenOOD [26].

2) Data Sets: Given the absence of dedicated benchmarks for OOD detection in MIL, we curate three pairs of ID/OOD evaluation sets from six established MIL datasets across the image and text domains [14], [28] based the work of Bescond et al. [20]: 1) Elephant (ID) and Tiger (OOD); 2) News.aa (ID) and News.rsh (OOD); and 3) MNIST-BAG (ID) and FashionMNIST-BAG (OOD). In addition, CluMILOOD is a posterior OOD algorithm for which we selected six MIL algorithms, ABMIL [14], DSMIL [29], MAMIL [18], WiKG [17], DYHG-MIL [30], GDA-MIL [31], to train ID data.

TABLE I

PERFORMANCE COMPARISON OF CLUMILOOD AND RIVAL OOD DETECTORS ON ELEPHANT-TIGER DATASET.

Metric	OOD detector	ABMIL	DSMIL	MAMIL	WiKG	DYHG-MIL	GDA-MIL
				Elephant(ID)→Tiger(OOD)			
AUROC↑	BASELINE	0.524 ± 0.189	0.490 ± 0.221	0.504 ± 0.139	0.338 ± 0.066	0.496 ± 0.054	0.424 ± 0.113
	DICE	0.031 ± 0.023	N/A	0.126 ± 0.108	0.418 ± 0.246	N/A	0.445 ± 0.069
	EBO	0.558 ± 0.130	0.571 ± 0.234	0.527 ± 0.104	0.351 ± 0.072	0.502 ± 0.050	0.397 ± 0.115
	IODIN	0.481 ± 0.194	0.489 ± 0.221	0.466 ± 0.157	0.336 ± 0.069	0.441 ± 0.050	0.396 ± 0.093
	MLS	0.522 ± 0.209	0.589 ± 0.214	0.563 ± 0.133	0.356 ± 0.074	0.476 ± 0.161	0.420 ± 0.093
	NCI	0.019 ± 0.012	0.000 ± 0.000	0.168 ± 0.125	0.296 ± 0.144	0.467 ± 0.075	0.447 ± 0.145
	KL-Matching	0.609 ± 0.108	0.570 ± 0.052	0.489 ± 0.093	0.600 ± 0.104	0.484 ± 0.032	0.528 ± 0.027
	ODIN	0.472 ± 0.174	0.490 ± 0.221	0.604 ± 0.098	0.261 ± 0.078	0.437 ± 0.087	0.447 ± 0.078
	SCALE	0.552 ± 0.117	0.571 ± 0.234	0.504 ± 0.111	0.359 ± 0.077	0.439 ± 0.076	0.370 ± 0.048
	VIM	0.507 ± 0.251	N/A	0.347 ± 0.073	0.350 ± 0.078	N/A	0.424 ± 0.168
	CluMILOOD	$\mathbf{0.822 \pm 0.024}$	$\mathbf{0.824 \pm 0.026}$	$\mathbf{0.678 \pm 0.083}$	$\mathbf{0.660 \pm 0.075}$	$\mathbf{0.576 \pm 0.101}$	$\mathbf{0.556 \pm 0.060}$
FPR@95↓	BASELINE	0.880 ± 0.123	0.925 ± 0.071	0.905 ± 0.118	0.955 ± 0.033	0.940 ± 0.058	0.950 ± 0.047
	DICE	1.000 ± 0.000	N/A	1.000 ± 0.000	0.935 ± 0.080	N/A	0.965 ± 0.042
	EBO	0.845 ± 0.089	0.780 ± 0.303	0.915 ± 0.038	0.975 ± 0.025	0.925 ± 0.085	0.955 ± 0.048
	IODIN	0.880 ± 0.096	0.925 ± 0.071	0.920 ± 0.041	0.985 ± 0.034	0.940 ± 0.055	0.960 ± 0.052
	MLS	0.870 ± 0.082	0.795 ± 0.187	0.900 ± 0.061	0.955 ± 0.048	0.915 ± 0.058	0.965 ± 0.052
	NCI	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.990 ± 0.022	0.975 ± 0.031	0.915 ± 0.058
	KL-Matching	0.865 ± 0.114	0.920 ± 0.057	0.915 ± 0.080	0.860 ± 0.249	0.955 ± 0.027	0.990 ± 0.022
	ODIN	0.865 ± 0.068	0.925 ± 0.071	0.875 ± 0.108	0.985 ± 0.014	0.945 ± 0.048	0.955 ± 0.051
	SCALE	0.850 ± 0.075	0.780 ± 0.303	0.955 ± 0.076	0.960 ± 0.029	0.945 ± 0.037	0.960 ± 0.052
	VIM	0.875 ± 0.129	N/A	0.940 ± 0.072	0.985 ± 0.022	N/A	0.980 ± 0.021
	CluMILOOD	$\mathbf{0.835 \pm 0.063}$	$\mathbf{0.875 \pm 0.079}$	$\mathbf{0.815 \pm 0.155}$	$\mathbf{0.775 \pm 0.174}$	$\mathbf{0.840 \pm 0.095}$	$\mathbf{0.905 \pm 0.097}$

TABLE II

PERFORMANCE COMPARISON OF CLUMILOOD AND RIVAL OOD DETECTORS ON NEW.AA-NEWS.RSH DATASET.

Metric	OOD detector	ABMIL	DSMIL	MAMIL	WiKG	DYHG-MIL	GDA-MIL
				News.aa(ID)→News.rsh(OOD)			
AUROC↑	BASELINE	0.138 ± 0.034	0.369 ± 0.210	0.179 ± 0.056	0.223 ± 0.092	0.573 ± 0.081	0.218 ± 0.051
	DICE	0.127 ± 0.040	N/A	0.661 ± 0.096	0.134 ± 0.056	N/A	$\mathbf{0.672 \pm 0.190}$
	EBO	0.143 ± 0.042	0.536 ± 0.078	0.178 ± 0.063	0.245 ± 0.063	0.575 ± 0.116	0.332 ± 0.269
	IODIN	0.140 ± 0.032	0.368 ± 0.210	0.208 ± 0.098	0.199 ± 0.067	0.599 ± 0.043	0.221 ± 0.098
	MLS	0.142 ± 0.032	0.432 ± 0.202	0.183 ± 0.057	0.227 ± 0.098	0.560 ± 0.176	0.329 ± 0.238
	NCI	0.129 ± 0.045	0.070 ± 0.034	0.226 ± 0.099	0.255 ± 0.093	0.666 ± 0.257	0.436 ± 0.199
	KL-Matching	0.816 ± 0.075	0.879 ± 0.041	$\mathbf{0.753 \pm 0.100}$	$\mathbf{0.792 \pm 0.098}$	0.492 ± 0.114	0.589 ± 0.226
	ODIN	0.135 ± 0.036	0.369 ± 0.210	0.245 ± 0.134	0.148 ± 0.061	0.489 ± 0.159	0.269 ± 0.117
	SCALE	0.135 ± 0.044	0.536 ± 0.078	0.173 ± 0.066	0.226 ± 0.090	0.561 ± 0.146	0.368 ± 0.236
	VIM	0.140 ± 0.038	N/A	0.173 ± 0.044	0.217 ± 0.078	N/A	0.221 ± 0.101
	CluMILOOD	$\mathbf{0.867 \pm 0.018}$	$\mathbf{0.958 \pm 0.021}$	$\mathbf{0.753 \pm 0.102}$	0.749 ± 0.096	$\mathbf{0.738 \pm 0.148}$	$\mathbf{0.672 \pm 0.323}$
FPR@95↓	BASELINE	1.000 ± 0.000	0.750 ± 0.154	0.960 ± 0.042	0.920 ± 0.076	0.850 ± 0.079	0.940 ± 0.042
	DICE	1.000 ± 0.000	N/A	0.500 ± 0.061	0.990 ± 0.022	N/A	0.600 ± 0.387
	EBO	0.980 ± 0.027	0.600 ± 0.146	0.960 ± 0.022	0.930 ± 0.027	0.770 ± 0.135	0.860 ± 0.258
	IODIN	0.990 ± 0.022	0.750 ± 0.154	0.930 ± 0.045	0.960 ± 0.022	0.780 ± 0.135	1.000 ± 0.000
	MLS	0.990 ± 0.022	0.680 ± 0.175	0.960 ± 0.042	0.940 ± 0.022	0.750 ± 0.203	0.930 ± 0.104
	NCI	1.000 ± 0.000	1.000 ± 0.000	0.960 ± 0.042	0.950 ± 0.050	$\mathbf{0.670 \pm 0.303}$	0.970 ± 0.045
	KL-Matching	0.460 ± 0.065	0.300 ± 0.094	0.810 ± 0.286	0.570 ± 0.249	0.910 ± 0.074	0.810 ± 0.370
	ODIN	1.000 ± 0.000	0.750 ± 0.154	1.000 ± 0.000	0.990 ± 0.022	0.900 ± 0.100	0.980 ± 0.045
	SCALE	0.980 ± 0.027	0.600 ± 0.146	0.950 ± 0.035	0.980 ± 0.027	0.790 ± 0.129	0.870 ± 0.211
	VIM	0.980 ± 0.027	N/A	0.970 ± 0.027	0.940 ± 0.065	N/A	0.940 ± 0.042
	CluMILOOD	$\mathbf{0.440 \pm 0.147}$	$\mathbf{0.150 \pm 0.087}$	$\mathbf{0.540 \pm 0.178}$	$\mathbf{0.570 \pm 0.055}$	0.750 ± 0.194	$\mathbf{0.580 \pm 0.409}$

3) Evaluation Metrics: For OOD detection, we utilize the two most commonly used metrics [10], [26]: AUROC and FPR@95. A higher AUROC (indicated by ↑) and a lower FPR@95 (indicated by ↓) indicate better performance, respectively. For MNIST-BAG dataset, which includes both training and test sets, only the AUROC is shown. For the other datasets, 5-fold cross-validation is used, and the evaluation performance values and corresponding standard deviations are shown. Note that for the value of the clustering centers K, we will analyze it specifically in the parameter analysis. For four sub-strategies of CluMILOOD, the best performance value is taken.

4) Implementation Details: Our CluMILOOD contains only one hyperparameter: the number of CIBFs, which will be enumerated in $\{2, 4, 8, 16, 32\}$ and discussed in the parameter analysis. Furthermore, we will verify the effectiveness and contribution of the four OOD scoring strategies (MAX, MIN, MEAN, AVE) through ablation analysis. All experiments were conducted on a system equipped with an AMD Ryzen 5 7500F

TABLE III
OVERALL ABLATION RESULTS ON THREE DATASET PAIRS. COLUMNS SHARE THE SAME METRICS (MAX/MIN/MEAN/AVE).

Method	Elephant(ID)→Tiger(OOD)				News.aa(ID)→News.rsh(OOD)				MNIST(ID)→FashionMNIST(OOD)			
	MAX	MIN	MEAN	AVE	MAX	MIN	MEAN	AVE	MAX	MIN	MEAN	AVE
ABMIL	**0.822**	0.277	0.514	0.216	**0.867**	0.212	0.548	0.212	**0.960**	0.820	0.580	0.630
DSMIL	**0.824**	0.278	0.528	0.230	**0.958**	0.170	0.545	0.170	0.480	**0.920**	0.600	0.750
MAMIL	0.498	0.656	0.526	**0.678**	**0.844**	0.275	0.531	0.241	**0.950**	0.780	0.570	0.460
WiKG	**0.661**	0.660	0.518	0.630	**0.749**	0.387	0.538	0.521	**1.000**	0.690	0.520	0.640
DYHG-MIL	0.553	0.568	0.514	**0.576**	**0.738**	0.336	0.535	0.342	0.758	**0.774**	0.735	0.532
GDA-MIL	**0.556**	0.502	0.530	0.512	0.616	0.557	0.532	**0.672**	0.873	**0.950**	0.610	0.610

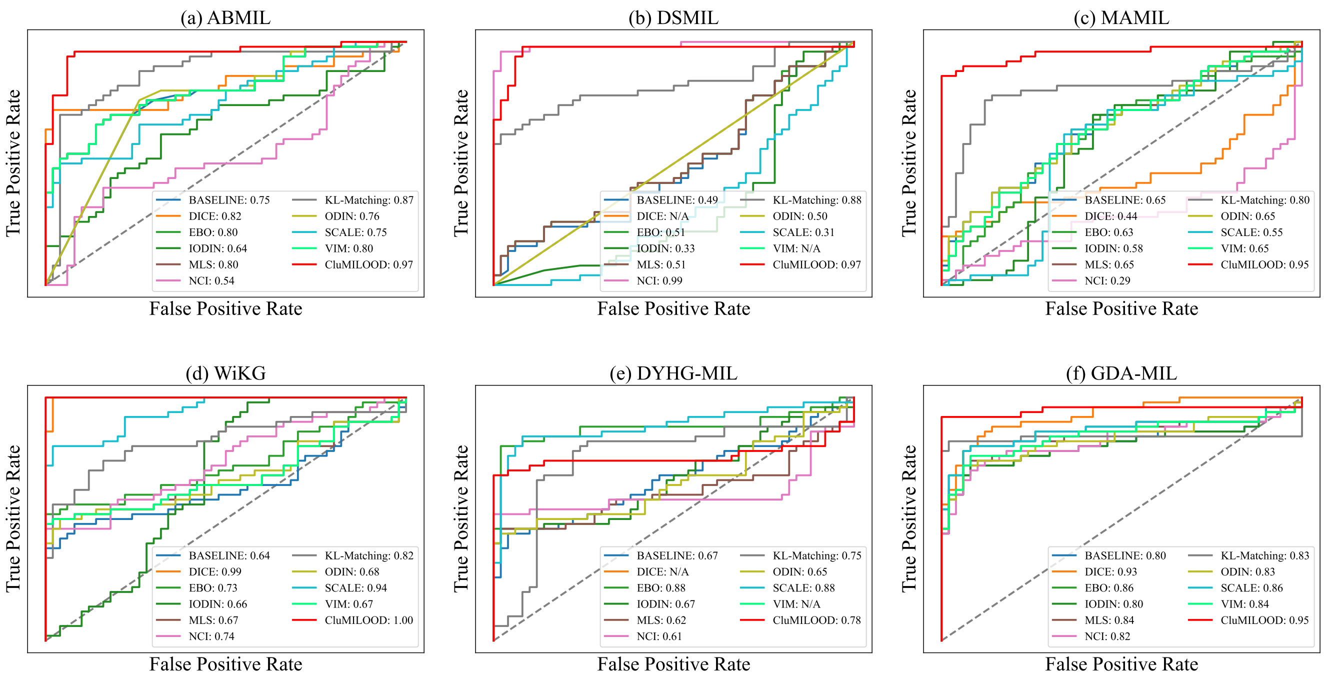


Fig. 2. Performance comparison of CluMILOOD and rival OOD detectors on MNIST-BAG and FashionMNIST-BAG datasets.

processor, 32 GB of RAM, and an NVIDIA GeForce RTX 5070 Ti GPU with 16 GB of video memory.

B. Performance Comparison

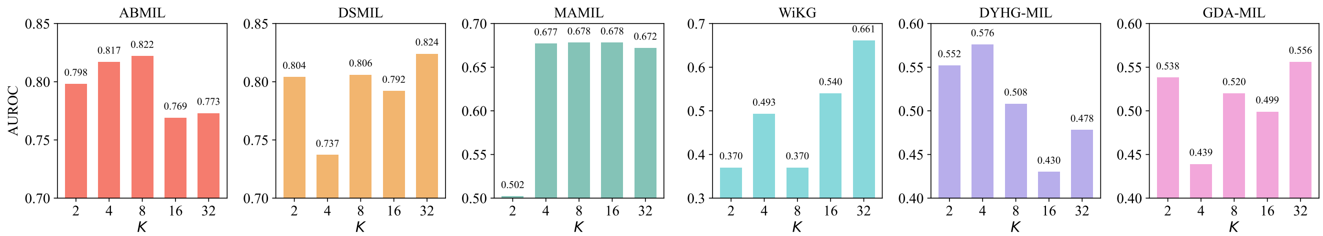
Tables I-II and Fig. 2 show the comparison results of CluMILOOD with various OOD methods on the MIL datasets of the MIL benchmark across six MIL backbone architectures. Overall, CluMILOOD demonstrates stronger ID/OOD detection capabilities across different datasets and MIL frameworks and exhibits good robustness and generalization ability. Specifically, its AUROC surpasses that of the next-best performing rivals by 21.3%, 22.5%, 7.4%, 6.0%, 7.4%, and 2.6% under six MIL trainers. For the remaining two datasets, CluMILOOD significantly outperforms other algorithms. For example, under DSMIL training, its AUROC is 7.9% higher than the second place. Furthermore, some baseline algorithms even show poor or unstable performances in certain cases. The overall experimental results have demonstrated the validity and effectiveness of the proposed method on various data modalities and MIL frameworks.

C. Ablation Analysis

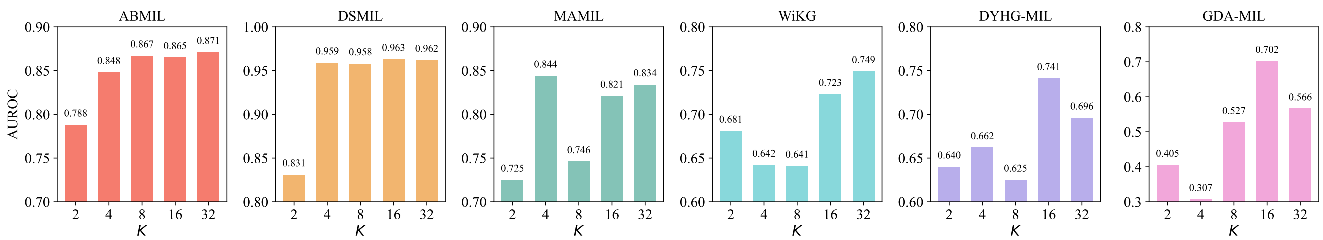
We conducted ablation analysis on four OOD scoring strategies in CluMILOOD. Specifically, keeping other steps constant, we only changed the scoring strategy to compare AUROC. Overall results show that MAX performs best in most backbone models and data settings, making it the most stable and versatile choice. MIN and AVE perform better only in a few combinations, indicating some complementary value under specific distribution patterns or backbone models. MEAN showed no advantage in all three settings, suggesting that simple averaging easily weakens key bias information, resulting in insufficient discriminatory power between ID and OOD. Therefore, we recommend MAX as default, and provide MIN/AVE as optional strategies to suit specific scenarios.

D. Parameter Analysis

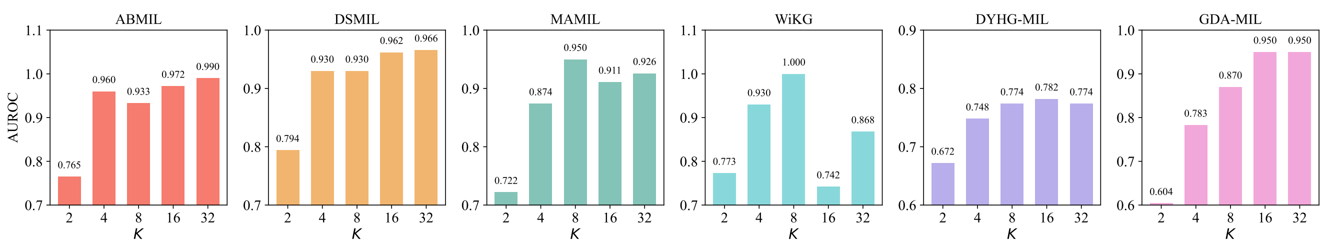
In the proposed CluMILOOD algorithm, the number of CIBFs K is a crucial parameter. It controls the number of learned clusters and their corresponding CIBFs and influences the density of global information. Therefore, we analyze



(a) **Elephant(ID)→Tiger(OOD)**



(b) **News.aa(ID)→News.rsh(OOD)**



(c) **MNIST(ID)→FashionMNIST(OOD)**

Fig. 3. Parameter analysis of the proposed CluMILOOD, focusing on the number of CIBFs K. We enumerate $K \in \{2, 4, 8, 16, 32\}$ and report its AUROC values on three ID/OOD benchmark datasets: (a) Elephant (ID) and Tiger (OOD), (b) News.aa (ID) and News.rsh (OOD), and (c) MNIST-BAG (ID) and FashionMNIST-BAG (OOD).

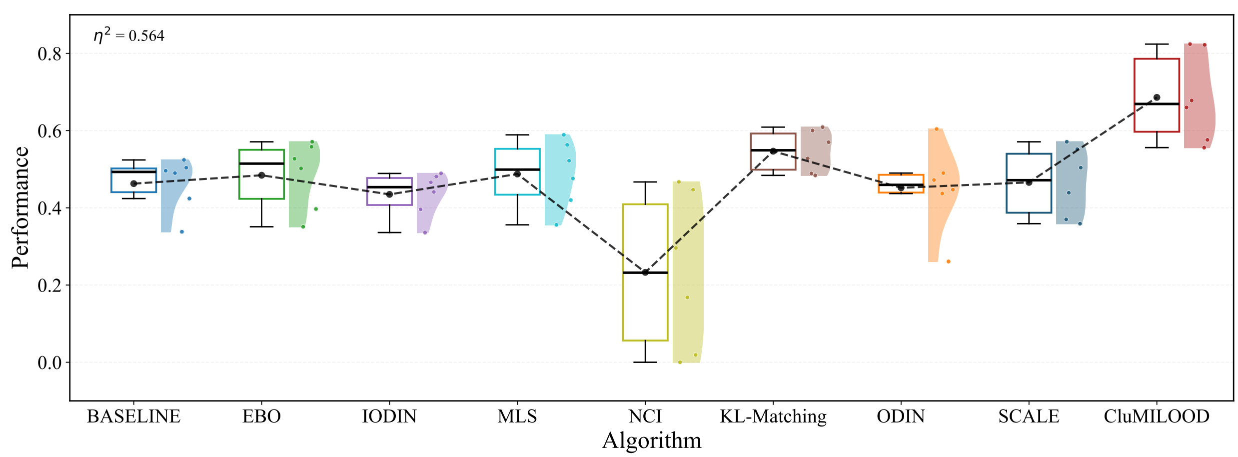


Fig. 4. Statistical analysis of CluMILOOD with state-of-the-art OOD detectors. In the figure, the box plot with violin plot shows the performance distribution within each algorithm, and the black dotted line with the mean dot shows the overall trend between algorithms.

the influence of this parameter on the performance of CluMILOOD, as shown in Fig. 3. The results show a certain correlation between the task and the MIL network. When the K value is too small (e.g., 2), the network often fails to cover the diversity of ID bags, resulting in generally low AUROC scores. As the K value increases, detection performance typically improves significantly. However, when K is further increased to 16 or 32, the performance improvement tends to saturate and, in most cases, even exhibits slight fluctuations, while also leading to higher operating costs. Therefore, considering both performance stability and efficiency, $K = 4$ or $K = 8$ yields better and more stable results in most settings and can be considered the recommended default parameter values.

E. Statistical Analysis

Analysis of variance (ANOVA) [32], [33] is a statistical method based on analysis of variance, used to determine whether there are significant differences among algorithms. It intuitively shows the performance distribution, variance range, and central tendency of different algorithms, which helps to determine whether the differences between them are statistically significant, as shown in Fig. 4. In the figure, the box plot represents the quartile statistics of the algorithm performance, the half violin plot represents the density estimate of the sample distribution, the black dashed line represents the mean trend of the algorithm, and the black dots represent the mean point of the algorithm. Results show that CluMILOOD significantly outperforms other algorithms, with performance values ranging between 0.55 and 0.85. While others may perform well on certain datasets, their overall transferability is far inferior to CluMILOOD. Furthermore, the η^2 value of ANOVA is 0.564, indicating a statistically significant difference among all algorithms.

V. Conclusion

This paper proposes CluMILOOD, an OOD detection method on the MIL framework, aiming to achieve innovative improvements for MIL-OOD detection methods. Specifically, CluMILOOD fully utilizes the hierarchical structure of MIL to achieve more discriminative OOD detection by combining the fused intra-bag features with the clustered inter-bag features. In addition, we design four complementary sub-strategies to characterize the differences in unknown bags and ID distributions from different perspectives, effectively improving the generalization ability of OOD detection. Experiments show that this method significantly outperforms existing methods on three datasets, verifying its effectiveness in MIL-OOD scenarios. Nevertheless, the selection of the clustering parameter K currently relies on experimental validation, and a more adaptive selection mechanism warrants future exploration.

References

[1] S. Lu, Y. Wang, L. Sheng, L. He, A. Zheng, and J. Liang, "Out-of-distribution detection: A task-oriented survey of recent advances," *ACM Computing Surveys*, pp. 1–39, 2025.

[2] Z. Zhang and X. Xiang, "Decoupling MaxLogit for out-of-distribution detection," in *CVPR*, 2023, pp. 3388–3397.

[3] T. Wei, B. Wang, and M.-L. Zhang, "EAT: Towards long-tailed out-of-distribution detection," in *AAAI*, 2024, pp. 1–12.

[4] W. Tang, W. Zhang, and M.-L. Zhang, "Disambiguated attention embedding for multi-instance partial-label learning," in *NeurIPS*, 2023, pp. 1–16.

[5] H. Luo, Y.-X. Zhang, Z. Zhou, W. Liu, and M. Li, "Propensity scoring for multi-instance partial-label learning," in *APWeb-WAIM*, 2025, pp. 239–255.

[6] W. Tang, Y.-F. Yang, W. Zhang, and M.-L. Zhang, "Calibratable disambiguation loss for multi-instance partial-label learning," *arXiv*, pp. 1–16, 2025.

[7] W. Yuan, Y. Chen, B. Zhu, S. Yang, J. Zhang, N. Mao, J. Xiang, Y. Li, Y. Ji, X. Luo *et al.*, "Pancancer outcome prediction via a unified weakly supervised deep learning model," *Signal Transduction and Targeted Therapy*, vol. 10, no. 1, p. 285, 2025.

[8] T. Zheng, K. Jiang, H. Yao, Y. Xiao, and Z. Wang, "Oodml: Whole slide image classification meets online pseudo-supervision and dynamic mutual learning," in *AAAI*, 2025, pp. 10 626–10 634.

[9] J. Zhang, Y. Cui, P. Zhen, and Y. Mei, "Global context-aware multi-instance learning for whole slide image classification," in *ICASSP*, 2026.

[10] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *ICLR*, 2017, pp. 1–12.

[11] W. Zhang, X. Zhang, H.-W. Deng, and M.-L. Zhang, "Multi-instance causal representation learning for instance label prediction and out-of-distribution generalization," in *NeurIPS*, 2022, pp. 34 940–34 953.

[12] M. Salehi, H. Mirzaei, D. Hendrycks, Y. Li, M. Rohban, M. Sabokrou *et al.*, "A unified survey on anomaly, novelty, open-set, and out of-distribution detection: Solutions and future challenges," *Transactions on Machine Learning Research*, no. 234, pp. 1–81, 2022.

[13] R. Xu and K. Ding, "Large language models for anomaly and out-of-distribution detection: A survey," in *NAACL*, 2025, pp. 5992–6012.

[14] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *ICML*, 2018, pp. 2127–2136.

[15] W. Tang, W. Zhang, and M.-L. Zhang, "Multi-instance partial-label learning: Towards exploiting dual inexact supervision," *Science China Information Sciences*, vol. 67, no. 3, p. 132103, 2024.

[16] Z. Yu, Z. Xia, D. Xu, Z. Zhang, L. Zhang, P. Zhang, L. Wu, B. Wang, H. Wang, and Z. Zhao, "Structure-aware generalization for heterogeneous histopathology via prototype-based multiple instance learning," *npj Digital Medicine*, 2026.

[17] J. Li, Y. Chen, H. Chu, Q. Sun, T. Guan, A. Han, and Y. He, "Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis," in *CVPR*, 2024, pp. 11 323–11 332.

[18] A. V. Konstantinov and L. V. Utkin, "Multi-attention multiple instance learning," *Neural Computing and Applications*, pp. 1–23, 2022.

[19] K. Xu, R. Chen, G. Franchi, and A. Yao, "Scaling for training time and post-hoc out-of-distribution detection enhancement," in *ICLR*, 2024.

[20] L. L. Bescond, M. Vakalopoulou, S. Christodoulidis, F. Andre, and H. Talbot, "On the detection of out-of-distribution samples in multiple instance learning," in *ICCV Workshops*, 2023, pp. 1–6.

[21] Y. Sun and Y. Li, "DICE: Leveraging sparsification for out-of-distribution detection," in *ECCV*, 2022, pp. 691–708.

[22] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *NeurIPS*, vol. 33, 2020, pp. 21 464–21 475.

[23] S. Regmi, "Going beyond conventional ood detection," *arXiv*, pp. 1–38, 2024.

[24] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," in *ICML*, 2022, pp. 8759–8773.

[25] L. Liu and Y. Qin, "Detecting out-of-distribution through the lens of neural collapse," in *CVPR*, 2025, pp. 15 424–15 433.

[26] J. Zhang, J. Yang, P. Wang, H. Wang, Y. Lin, H. Zhang, Y. Sun, X. Du, K. Zhou, W. Zhang, Y. Li, Z. Liu, Y. Chen, and H. Li, "OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection," in *NeurIPS Workshop*, 2023, pp. 1–18.

[27] H. Wang, Z. Li, L. Feng, and W. Zhang, "VIM: Out-of-distribution with virtual-logit matching," in *CVPR*, 2022, pp. 4921–4930.

[28] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-I.I.D. samples," in *ICML*, 2009, pp. 1249–1256.

[29] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *CVPR*, 2021, pp. 14 318–14 328.

[30] Y. Chen, J. Li, L. Zhu, Y. Xu, T. Guan, H. Shi, Y. He, and A. Han, "Dynamic hypergraph representation for bone metastasis analysis," *Computer Methods and Programs in Biomedicine*, vol. 271, pp. 1–12, 2026.

[31] Y.-X. Zhang, Z. Zhou, W. Liu, and M. Zhang, "Rethinking multi-instance learning through graph-driven fusion: A dual-path approach to adaptive representation," in *AAAI*, 2026, pp. 28 510–28 518.

[32] L. St and S. Wold, "Analysis of variance (anova)," *Chemometrics and Intelligent Laboratory Systems*, vol. 6, no. 4, pp. 259–272, 1989.

[33] C. Li, P. Huang, J. Qin, and X. Luo, "Knowledge-driven multiple instance learning with hierarchical cluster-incorporated aware filtering for larynx pathological grading," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–13, 2025.