Fast Multi-Instance Partial-Label Learning

Yin-Fang Yang[1,2], **Wei Tang**[1,2*], **Min-Ling Zhang**[1,2]

[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China
yangyf22@gmail.com, tangw@seu.edu.cn, zhangml@seu.edu.cn

Abstract

Multi-instance partial-label learning (MIPL) is a paradigm where each training example is encapsulated as a multi-instance bag associated with the candidate label set, which includes one true label and several false positives. Current MIPL algorithms typically assume that all instances are independent, thereby neglecting the dependencies and heterogeneity inherent in MIPL data. Moreover, these algorithms often prove to be excessively time-consuming when dealing with complex datasets, significantly limiting the practical application of MIPL. In this paper, we propose FASTMIPL, a framework that employs mixed-effects model to explicitly capture the dependencies and heterogeneity among instances and bags. FASTMIPL is able to learn from MIPL data both effectively and efficiently by utilizing the predefined dependencies modeling module and leveraging the posterior predictive probability disambiguation strategy. Experiments show that the performance of FASTMIPL is highly competitive to state-of-the-art methods, while significantly reducing computational time in benchmark and the real-world datasets.

Introduction

Weakly supervised learning has emerged as a potent strategy in scenarios characterized by a scarcity of annotated data. Based on label quality and quantity, weak supervision can be systematically classified into three primary categories, namely, inexact, inaccurate, and incomplete supervision (Zhou 2018). Furthermore, the inexact supervision indicates a coarse alignment between instances and labels, which is ubiquitous and challenging in real-world tasks. The two predominant learning paradigms for addressing issues related to the inexact supervision are multi-instance learning (MIL) (Ilse, Tomczak, and Welling 2018; Cui et al. 2023) and partial-label learning (PLL) (Tian, Yu, and Fu 2023; Hüllermeier and Beringer 2006).

In many real world applications, the inexact supervision can exist simultaneously in the instance space and the label space (Tang, Zhang, and Zhang 2024b). For example, in medical image analysis, each image (bag) may contain multiple regions of instances, with labels provided by doctors. To reduce labeling costs, each image can be assigned a candidate label set rather than an exact diagnosis, models can learn from partially labeled data, improving classification efficiency and accuracy. Multi-instance partial-label learning (MIPL) resolves scenarios where ambiguity exists in both instance and label spaces, making it a more natural and convenient approach to tasks involving such complex situations (Tang, Zhang, and Zhang 2024b).

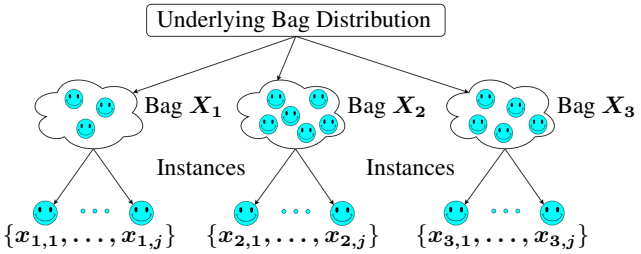

Figure 1: The hierarchical and nested data structure presents in the multi-instance bags indicate the presence of inherent dependencies and heterogeneous information embedded within the bag distribution.

However, existing predominantly MIPL address the problem under the assumption of independent samples, without considering the hierarchical structure between samples. This assumption simplifies the modeling process but often neglects the inherent dependencies and structural relationships in MIPL, as illustrated in Figure 1. Such oversight can hinder the model's ability to effectively capture the underlying data distribution, resulting in suboptimal performance. Furthermore, when extending these methods to more complex real-world applications, the computational burden becomes particularly pronounced. The independence assumption necessitates processing each bag with numerous instances individually, leading to substantial computational overhead. This challenge is exacerbated by the need to evaluate multiple potential labels for each bag under the setting partial-label learning, significantly increasing resource demands and limiting the scalability of MIPL in practical applications.

In this paper, we propose a new framework FASTMIPL to model dependent MIPL data efficiently. Designed for dependencies and efficiency in MIPL analyses, FASTMIPL explicitly models the dependencies and heterogeneity among multi-instance bags by using a hierarchical model extend-

*Corresponding author

ing the mixed-effects module. Building on the hierarchical model, FASTMIPL can effectively disentangle the complex interplay between instance-level characteristics and bag-level attributes, while simultaneously accounting for the latent structures that permeate the MIPL paradigm. To enhance computational efficiency, FASTMIPL jointly optimizes the evidence lower bound (ELBO) with respect to fixed effects covariates, posterior parameters, and prior hyperparameters using mini-batch gradient descent. Compared to prevailing MIPL methods, the proposed FASTMIPL consistently outperforms other implementations and improves the efficiency for more than 30 times on datasets. This work makes three contributions:

- We propose the FASTMIPL model, a synthetic model that integrates instance importance pooling function with statistical generalized linear hierarchical model to capture the dependencies and heterogeneity of MIPL data.
- We offer a transparent and interpretable framework for understanding the influence of the individual instance on the overall bag-level prediction.
- We significantly enhance computational efficiency by leveraging the computational advantages of linear models and employing a reparameterized variational inference framework to jointly optimize the objective function across benchmark and the real-world datasets.

Related Work

Multi-Instance Learning

Multi-instance learning (MIL) organizes data into bags of instances, where the bag label is known but the individual instance labels are unknown. Attention-based MIL algorithms focus on relevant instances within a bag, improving classification performance (Ilse, Tomczak, and Welling 2018). Loss-based attention mechanisms extend this paradigm to multi-class tasks (Shi et al. 2020). Despite their exceptional performance, attention-based MIL methods often suffer from high computational complexity, leading to significant training and inference times (Aminabadi et al. 2022). Multiple iterations for convergence and additional learnable parameters in the attention module also contribute to increased time complexity (Wibawa et al. 2022). More related to our setting, (Cui et al. 2023) used variational inference to estimate the posterior distribution of instance-level weights, enhancing model interpretability and uncertainty estimation. Accordingly, this time-consuming nature can hinder the practicality of attention-based MIL algorithms in real-world applications requiring efficiency and scalability, approaches with high efficiency are demanded.

Partial-Label Learning

Recent PLL approaches heavily rely on deep learning techniques. (Yao et al. 2020) employed deep convolutional neural networks for feature extraction and utilized the exponential moving average technique to uncover latent true labels. Building on the instance-dependent principle, (Xu et al. 2021) propose a novel PLL method that recovers the label distribution as a label enhancement process and trains the predictive model iteratively in every epoch. Following this line of thought, (Qiao, Xu, and Geng 2023) explicitly model the generation process of candidate labels in instance-dependent PLL. While these algorithms exhibit considerable efficacy in tackling instance-dependent partial-label learning problems, they encounter limitations in directly handling inexact supervision within the instance space. Consequently, they cannot be directly applied to multi-instance partial-label learning problems.

Multi-Instance Partial-Label Learning

MIPL extends both MIL and PLL. Only three MIPL algorithms have been proposed recently, all assuming instance independence. MIPLGP (Tang, Zhang, and Zhang 2024b) learns from MIPL data at the instance level using label augmentation and Dirichlet disambiguation. DEMIPL (Tang, Zhang, and Zhang 2023) identifies the true label from candidate labels, assuming instance independence and using disambiguation attention. ELIMIPL (Tang, Zhang, and Zhang 2024a) exploits candidate and non-candidate label set information by mapping bags to candidate label sets and learning the candidate label matrix sparsity. These methodologies have demonstrated commendable performance, corroborating the significance of the MIPL framework across diverse applications. Nevertheless, it is imperative to acknowledge that the multi-instance bags, corresponding to constituent parts of an object, inherit structural dependencies and heterogeneous information from bag distribution.

The FASTMIPL Approach

Formally, a MIPL training dataset is defined as $\mathcal{D} = \{(\boldsymbol{X}_i, \mathcal{S}_i) \mid 1 \leq i \leq m\}$, where $\mathcal{D}$ comprises m bags along with their associated candidate label sets. Crucially, each $\mathcal{S}_i$ includes a single ground-truth label y_i, i.e., $y_i \in \mathcal{S}_i$, alongside one or more false positives, thereby introducing inherent label ambiguity. Furthermore, we delineate the instance space as $\mathcal{X} = \mathbb{R}^d$ and the label space as $\mathcal{Y} = [k]$ (with k classes) where $[k] := \{1, 2, \ldots, k\}$. Both the candidate label set $\mathcal{S}_i$ and its complement $\overline{\mathcal{S}}_i$ are proper subsets of $\mathcal{Y}$, satisfying the condition $|\mathcal{S}_i| + \left|\overline{\mathcal{S}}_i\right| = |\mathcal{Y}|$, where $|\cdot|$ denotes the set cardinality. In the context of MIPL, each bag $\boldsymbol{X}_i = \{\boldsymbol{x}_{i,j}\}_{j=1}^{n_i}$ is constituted by n_i instances, each residing in a d-dimensional space. It is crucial to emphasize that the cardinality of instances n_i may exhibit variability across different bags, adding an additional layer of complexity to the problem. Given this intricate problem structure, the primary objective of MIPL is to accurately identify the ground-truth label from the candidate label set corresponding to each multi-instance bag. There are three main steps in framework of FASTMIPL illustrated in Figure 2:

- Leveraging predefined instance embeddings and formalizing the importance weights through an attention-inspired pooling function;
- Extending the generalized linear mixed model (GLMM) by incorporating fixed and random effect parameters to effectively capture the latent dependencies and heterogeneity presented in MIPL;

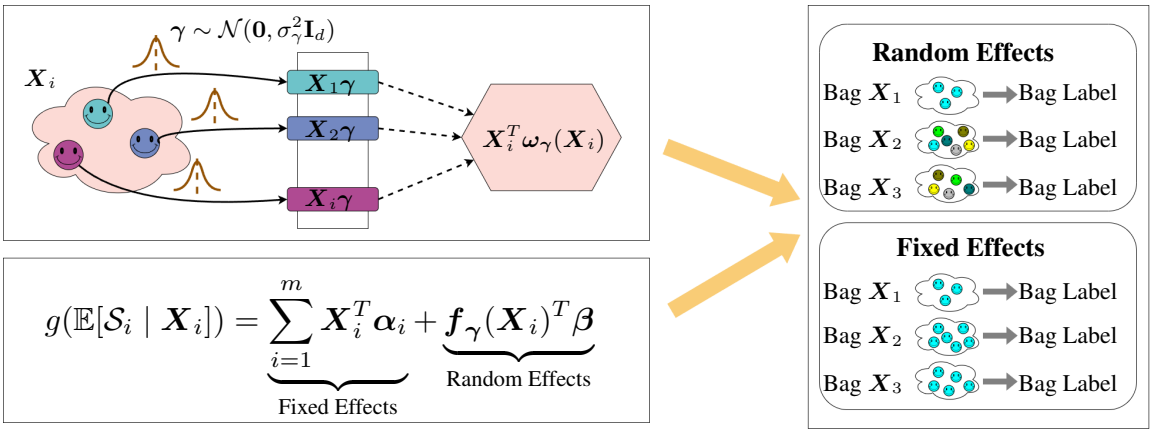

Figure 2: The framework of FASTMIPL, which comprises two key components: the Mixed-Effects Dependencies Modeling Module and the Bayesian Posterior Disambiguation Module.

- Introducing the posterior predictive probability disambiguation strategy to identify the true label from the candidate label set.

Instance Importance Weighting Function

We begin by utilizing predefined instance embeddings and proceed to model their importance weights via an attention-inspired pooling function. Furthermore, we quantify the instance importance weights by employing a single linear layer followed by a softmax link function across instances. Under these assumptions, the resulting bag embeddings can be mathematically represented as follows:

$$\boldsymbol{f}_{\boldsymbol{\gamma}}(\boldsymbol{X}_i) = \boldsymbol{X}_i^T \boldsymbol{\omega}_{\boldsymbol{\gamma}}(\boldsymbol{X}_i) \in \mathbb{R}^d, \tag{1}$$

where $\boldsymbol{\omega}_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)$ is defined as:

$$\begin{aligned} \boldsymbol{\omega}_{\boldsymbol{\gamma}}(\boldsymbol{X}_i) &= \operatorname{softmax}\left(\boldsymbol{X}_i \boldsymbol{\gamma}\right) \\ &= \operatorname{softmax}\left(\left[\boldsymbol{x}_{i,1}^T \boldsymbol{\gamma}, \boldsymbol{x}_{i,2}^T \boldsymbol{\gamma}, \ldots, \boldsymbol{x}_{i,n_i}^T \boldsymbol{\gamma}\right]\right) \in \mathbb{R}^{n_i}, \end{aligned} \tag{2}$$

where $\boldsymbol{X}_i \in \mathbb{R}^{n_i \times d}$ denotes the predefined bag embeddings across all instances, and it is crucial to note that both the bag embeddings $\boldsymbol{f}_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)$ and the weight function $\boldsymbol{\omega}_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)$ are parameterized by $\boldsymbol{\gamma}$, emphasizing their dependence on the parameters $\boldsymbol{\gamma}$. Building upon our previous discussion, the formulation provides a clear and interpretable representation of how instance-level information is aggregated to form bag-level embeddings (Ilse, Tomczak, and Welling 2018).

Incorporating Fixed and Random Effects in MIPL

To capture the complex dependencies and heterogeneity inherent in MIPL problems, we leverage both fixed effects and random effects from GLMM to model the relationship between the bag candidate labels, denoted by $\mathcal{S}_i$ for the i-th bag, and the corresponding bag embeddings, $\boldsymbol{f}_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)$, derived from the bag's instance-level features $\boldsymbol{X}_i$. The model also incorporates shared bag-level covariates, denoted by $\sum_{i=1}^m \boldsymbol{X}_i^T \boldsymbol{\alpha}_i$. Specifically, let $g(\cdot)$ represent a suitable link function. The conditional expectation of the bag candidate labels, given the instance-level features and bag-level covariates, is modeled as:

$$g(\mathbb{E}[\mathcal{S}_i \mid \boldsymbol{X}_i]) = \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{\alpha}_i + \boldsymbol{f}_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)^T \boldsymbol{\beta}, \tag{3}$$

where $\boldsymbol{\alpha}_i \in \mathbb{R}^{|\mathcal{S}_i| \times k}$ denotes the fixed effect coefficients associated with the multi-instance bag on the k-dimensional shared bag-level covariates, while $\boldsymbol{\beta} \in \mathbb{R}^{d \times k}$ represents the random effect coefficients corresponding to the d-dimensional bag embeddings. Besides, $\mathbb{E}[\mathcal{S}_i \mid \boldsymbol{X}_i]$ characterizes the conditional mean of the bag's candidate labels for the i-th observation, contingent upon both instance-level descriptors and bag-specific covariates.

The first component fixed effects $\boldsymbol{\alpha}_i \in \mathbb{R}^{|\mathcal{S}_i| \times k}$ were introduced to model the impact of shared, invariant characteristics within multi-instance bags on their corresponding bag-level labels. It is grounded in the recognition that instances may share certain characteristics that exert a consistent influence on the overall bag label. Besides, these shared features are typically observable and measurable, aligning with the definition of fixed effects in GLMM (Cuomo et al. 2022). Consequently, they can be regarded as fixed effects for the entire MIPL dataset.

To address the heterogeneous influence of individual instance on bag-level labels, we introduce instance-specific variations as random effects $\boldsymbol{f}_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)^T \boldsymbol{\beta}$, a formulation grounded in several robust theoretical and empirical foundations. First, the contribution of each instance to the overall bag-level label can vary significantly, and random effects are particularly well-suited to capture this variability (Moore et al. 2019). Moreover, random effects adeptly model the correlations between instances nested within bags, which is crucial for improving the generalization performance of MIPL models. To further enhance the robustness of our regression framework, especially in scenarios characterized by limited sample sizes or high-dimensional instance embeddings, we impose Bayesian priors on the regression coefficients. Specifically, we assume that the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$

adhere to multivariate normal distributions:

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2 \mathbf{I}_d), \quad \boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}_d), \tag{4}$$

where $\mathbf{I}_d$ represents the $d \times d$ identity matrix, and σ_β^2 and σ_γ^2 denote the variances of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. These priors encode our epistemic uncertainty by imposing a zero-centered regularization structure on the coefficient space, where the magnitude of shrinkage is governed by the associated variance hyperparameters in a hierarchical framework.

Consequently, the marginal likelihood of observing the bag candidate labels $\mathcal{S}_i$ for the i-th bag, given the bag-level covariates $\sum_{i=1}^m \boldsymbol{X}_i^T \boldsymbol{\alpha}_i$, the fixed effect coefficients $\boldsymbol{\alpha}_i$ and the instance-level features $\boldsymbol{X}_i$, can be expressed as:

$$p(\mathcal{S}_i \,|\, \boldsymbol{\alpha}_i, \boldsymbol{X}_i) = \int\int p(\mathcal{S}_i \mid \sum_{i=1}^m \boldsymbol{X}_i^T \boldsymbol{\alpha}_i + \boldsymbol{f}_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)^\top \boldsymbol{\beta}) p(\boldsymbol{\beta}) p(\boldsymbol{\gamma})\, d\boldsymbol{\beta}\, d\boldsymbol{\gamma}. \tag{5}$$

Bayesian Inference and Optimization

The optimization objective is to characterize the posterior distribution of the random effect parameters, denoted by $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\gamma}\}$, given the observed data $\mathcal{D}$. However, due to the intractability of exact inference in our posterior distribution, we employ variational inference as a strategy that approximates the true posterior $p(\boldsymbol{\theta} \mid \mathcal{D})$ by introducing a variational family $q_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ parameterized by $\boldsymbol{\phi}$, and optimizing $\boldsymbol{\phi}$ to maximize the ELBO:

$$\begin{aligned} &\mathrm{ELBO}\left(\boldsymbol{\phi}, \sigma_\beta^2, \sigma_\gamma^2\right) = \\ &\quad \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\theta})}[\log p(\mathcal{D} \mid \boldsymbol{\theta})] - D_{KL}\left(q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta})\right). \end{aligned} \tag{6}$$

Here $D_{KL}\left(q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta})\right)$ denotes the Kullback-Leibler divergence between the variational approximation $q_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ and the prior distribution of the parameters $p(\boldsymbol{\theta})$. We here consider the variational family of multivariate Gaussian distributions with full rank covariance parameterized by mean parameters $\boldsymbol{\mu}_{\boldsymbol{\phi}}$ and covariance parameters $\boldsymbol{\Sigma}_{\boldsymbol{\phi}}$:

$$q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta} \mid \boldsymbol{\mu}_{\boldsymbol{\phi}}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}}\right). \tag{7}$$

We jointly optimize the ELBO with respect to fixed effects $\boldsymbol{\alpha}_i$, variational parameters $\boldsymbol{\phi}$, and prior hyperparameters σ_β^2 and σ_γ^2 using mini-batch gradient descent (Pirš and Štrumbelj 2019). To enable backpropagation through the expectation term in the ELBO, we refine the prior distribution and draw samples from the variational posterior distributions, conditioned on the observed data.

Posterior Predictive Probability Disambiguation

Considering the fundamental linear characteristics inherent to the instance importance weighting function, the cumulative impact of bag representations on the label manifold can be rigorously expressed through the following formalism:

$$\begin{aligned} \boldsymbol{f}_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)^\top \boldsymbol{\beta} &= \left(\boldsymbol{X}_i^T \boldsymbol{\omega}_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)\right)^\top \boldsymbol{\beta} = \boldsymbol{\omega}_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)^\top (\boldsymbol{X}_i \boldsymbol{\beta}) \\ &= \boldsymbol{\omega}_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)^\top \boldsymbol{h}_{\boldsymbol{\beta}}(\boldsymbol{X}_i), \end{aligned} \tag{8}$$

where $\boldsymbol{h}_{\boldsymbol{\beta}}(\boldsymbol{X}_i) = \boldsymbol{X}_i \boldsymbol{\beta} \in \mathbb{R}^{n_i}$ represents a vector-valued mapping that generates instance-specific latent representations for the constituent elements within bag $\boldsymbol{X}_i$. The attention-inspired pooling, parameterized by $\boldsymbol{\omega}_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)$, assigns probabilistic importance scores to individual instances, thereby facilitating a differentiable instance-level contribution aggregation scheme for the composite bag-level prediction.

To identify the true label from the candidate label set, we introduce the posterior predictive probability disambiguation strategy. After optimization, we employ the learned approximate posterior distribution $q_{\boldsymbol{\phi}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ to predict the true label of a new bag $\boldsymbol{X}_\star$ from its the candidate label set:

$$y_\star = \arg\max_{c \in \mathcal{Y}} \hat{p}_{\star,c} = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\beta},\boldsymbol{\gamma})}\left[\boldsymbol{\omega}_{\boldsymbol{\gamma}}\left(\boldsymbol{X}_\star\right)^T \boldsymbol{h}_{\boldsymbol{\beta}}(\boldsymbol{X}_\star)\right], \tag{9}$$

where $y_\star$ denotes the true label corresponding to the new bag embedding $\boldsymbol{X}_\star$, while $\hat{p}_{\star,c}$ represents the posterior predictive probability that the c-th class within the candidate label set is the true label for the new bag $\boldsymbol{X}_\star$.

Moreover, to retrieve important instances, we leverage the expected value of importance weights, namely $\mathbb{E}_{\mathrm{q}_{\boldsymbol{\phi}}(\boldsymbol{\beta},\boldsymbol{\gamma})}\left[\boldsymbol{\omega}_{\boldsymbol{\gamma}}\left(\boldsymbol{X}_\star\right)\right]$. This formulation offers a transparent and interpretable framework for understanding the influence of individual instances on the overall bag-level prediction. The FastMIPL pseudocode of the optimization procedure summarizes in the Appendix[1], where the candidate labels and the prediction model are updated simultaneously.

Experiments

We analyze the performance of effectiveness and efficiency between FastMIPL and comparative algorithms in MIPL tasks, and validate the potency of the model parameters introduced in FastMIPL through ablation studies using modified variants of FastMIPL.

Experimental Configuration

Datasets Table 1 provides an overview of the characteristics of all datasets. There are eight types of characteristics mentioned. The symbol $\#bag$ denotes the count of multi-instance bags, and $\#ins$ represents the number of total instances. $max.\ \#ins$, $min.\ \#ins$, and $avg.\ \#ins$ correspond to the maximum, minimum, and average instance count across all bags for describing the instance distribution. The symbol $\#dim$ signifies the number of dimensions associated with each instance-level feature. The length of the label space and the average length of the label space for candidate label sets are denoted by $\#class$ and $avg.\ \#CLs$, respectively. The number of false positive labels are identified as $r\ (|\mathcal{S}_i| = r + 1)$ on benchmark datasets for evaluating performance comprehensively, where $\mathcal{S}_i$ represents as the candidate label set for each bag.

Comparative Algorithms We compare FastMIPL with a broad range of baselines, covering MIPL, MIL, and PLL

[1]FastMIPL's code and appendix have been made publicly available on Github: https://github.com/yangyf22/FastMIPL

Dataset	$\#bag$	$\#ins$	$max.\#ins$	$min.\#ins$	$avg.\#ins$	$\#dim$	$\#class$	$avg.\#CLs$
MNIST-MIPL (MNIST)	500	20664	48	35	41.33	784	5	$2, 3, 4$
FMNIST-MIPL (FMNIST)	500	20810	48	36	41.62	784	5	$2, 3, 4$
Birdsong-MIPL (Birdsong)	1300	48425	76	25	37.25	38	13	$2, 3, 4$
SIVAL-MIPL (SIVAL)	1500	47414	32	31	31.61	30	25	$2, 3, 4$
CRC-MIPL-Row (C-Row)	7000	56000	8	8	8	9	7	2.08
CRC-MIPL-SBN (C-SBN)	7000	63000	9	9	9	15	7	2.08
CRC-MIPL-KMeansSeg (C-KMeans)	7000	30178	6	3	4.311	6	7	2.08
CRC-MIPL-SIFT (C-SIFT)	7000	175000	25	25	25	128	7	2.08

Table 1: Characteristics of benchmark and real-world MIPL datasets.

algorithms. Specifically, we reference three MIPL algorithms (Tang, Zhang, and Zhang 2024b, 2023, 2024a): MIPLGP , DEMIPL, and ELIMIPL. Furthermore, our comparison encompasses two types of PLL algorithms: the deep-learning-based approach with linear classifiers, including PRODEN (Lv et al. 2020), RC (Feng et al. 2020), LWS (Wen et al. 2021) and CAVL (Zhang et al. 2022), and a feature-aware disambiguation algorithm named PL-AGGD (Wang, Li, and Zhang 2019). For MIL algorithms, we incorporate two types of MIL algorithms that a variational autoencoder-based model MIVAE (**?**) and three attention-based models: ATTEN (Ilse, Tomczak, and Welling 2018), ATTEN-GATE (Ilse, Tomczak, and Welling 2018) and LOSS-ATTEN (Shi et al. 2020). Due to spatial limitations, results obtained from three MIPL algorithms are presented in the main body of the paper, while those with PLL and MIL algorithms are detailed in the Appendix. Parameters for all compared baselines have been meticulously tuned, drawing from recommendations in the original literature or refined through our pursuit of improved performance.

Implementation FASTMIPL is implemented using PyTorch and trained on a single NVIDIA GeForce RTX 4090 GPU. The optimization process employs stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 0.0001. For instance-level feature extraction, a two-layer convolutional neural network and a fully connected network are applied to the MNIST-MIPL and FMNIST-MIPL datasets, while a fully connected network is employed on the Birdsong-MIPL and SIVAL-MIPL datasets with preprocessed features. For the CRC-MIPL dataset, a fully connected network follows one of four image bag generators or ResNet-34 as the feature extractor. The learning rate is selected from the predefined set $\{0.0005, 0.001, 0.002, 0.005\}$, the training batch size equals to the count of bags in the training set, and the value of posterior samples to approximate the expectation is chosen from the set $\{10, 20, 30, 40, 50\}$. The number of epochs is set to 200 for the MNIST-MIPL and FMNIST-MIPL datasets and 500 for the remaining three datasets. The data partition follows the strategies of DEMIPL and ELIMIPL, dividing the data into training and testing sets with a ratio of 7:3. The average and the standard deviation of accuracy are recorded by conducting the experiments with random train/test splits ten times, and the highest accuracy is highlighted in bold. We report the time consumption of experiments obtained from running the experiments at a time, with the least time consumption highlighted in bold.

Algorithm	r	MNIST	FMNIST	Birdsong	SIVAL
FASTMIPL	1	**.999±.002**	**.911±.022**	**.797±.024**	**.779±.030**
	2	**.998±.004**	**.901±.027**	**.792±.021**	**.708±.026**
	3	**.975±.074**	**.816±.071**	**.772±.022**	**.615±.031**
ELIMIPL	1	.991±.005	.904±.016	.770±.019	.676±.025
	2	.989±.013	.843±.026	.745±.017	.615±.023
	3	.749±.148	.701±.053	.717±.019	.599±.025
DEMIPL	1	.977±.008	.883±.019	.741±.015	.631±.042
	2	.944±.027	.822±.026	.702±.026	.551±.056
	3	.711±.088	.656±.027	.694±.024	.502±.017
MIPLGP	1	.951±.019	.846±.031	.714±.026	.669±.020
	2	.818±.033	.792±.027	.671±.015	.614±.023
	3	.623±.062	.669±.052	.626±.015	.570±.031

Table 2: The classification accuracy (mean±std) of algorithms on benchmark datasets with the varying numbers of false positive labels ($r \in \{1, 2, 3\}$).

Effectiveness Comparison

In MIPL tasks, the effectiveness evaluation typically focus solely on the metric of classification accuracy. The effectiveness evaluation is considered among four MIPL algorithms, five PLL algorithms and four MIL algorithms.

Results on Benchmark Datasets Table 2 presents a comprehensive comparison of FASTMIPL's effectiveness against three MIPL algorithms (MIPLGP, DEMIPL, ELIMIPL). The corresponding effectiveness results of PLL and MIL algorithms are recorded in the Appendix. Benchmark datasets used for the algorithm evaluation differentiate among different levels of false positive labels. FASTMIPL demonstrates statistically significant superiority in predictive performance relative to all baseline methodologies, as evidenced by mean accuracy metrics across four standardized benchmark datasets.

FASTMIPL exhibits a lower accuracy loss in the complex scenario of datasets characterized by a higher proportion of false positive labels. Specifically, FASTMIPL demonstrates a smaller difference in average accuracy compared to other MIPL algorithms when transitioning from a scenario with a false positive labels ($r = 1$) to one with two false positive labels ($r = 2$) on the Birdsong-MIPL dataset.

As shown in Tabel 3, FASTMIPL achieves statistically better performance against other approaches. The superior performance of FASTMIPL is consistent across almost all synthetic data sets and real-world data under the challenging

Algorithm	Prediction Performance (Kendalltau)	FastMIPL Improvement (t-test P-value)
FastMIPL	**.975 ± .074**	–
EliMIPL	.749 ± 148	**win**[< 1e-3]
DeMIPL	.711 ± .088	**win**[< 7e-9]
MIPLGP	.623 ± .062	**win**[< 6e-6]

Table 3: Summary of the Kendalltau correlation paired t-test across 5 labels for FastMIPL against other comparing approaches on the MNIST-MIPL ($r = 3$).

Algorithm	C-Row	C-SBN	C-KMeans	C-SIFT
FastMIPL	**.487±.038**	**.573±.031**	**.573±.013**	.526±.029
EliMIPL	.434±.008	.510±.008	.545±.013	**.539±.010**
DeMIPL	.410±.011	.484±.013	.523±.012	.531±.013
MIPLGP	.435±.006	.335±.008	.331±.014	–

Table 4: The classification accuracy (mean±std) of algorithms on the real-world dataset.

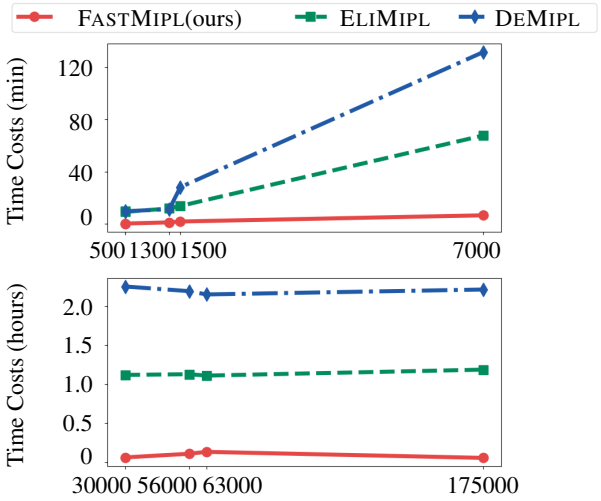

Figure 3: Time consumption trending for MIPL algorithms on data with different number of bags and instances, which subplots' horizontal axis represent the number of bags and the number of instances respectively.

number of false positive labels, which provides a strong evidence for the effectiveness of FastMIPL to facilitate MIPL.

Results on the Real-World Dataset Table 4 illustrates the accuracy comparison on the CRC-MIPL dataset for FastMIPL and three MIPL algorithms. The corresponding effectiveness results of PLL algorithms are demonstrated in the Appendix. The symbol "−" denotes the algorithm cannot be applied to C-SIFT dataset, because the computational limitation of memory overflow in our server. FastMIPL consistently outperforms in 9 out of 11 cases against MIPL algorithms. FastMIPL achieves better effectiveness than PLL and MIL algorithms in all cases.

The effectiveness of both EliMIPL and DeMIPL was enhanced by employing more sophisticated image bag generators on the CRC-MIPL dataset. Conversely, MIPLGP did not exhibit improved model effectiveness with complex features, which can be attributed to their inability to effectively utilize such features. Notably, FastMIPL incorporates a random effect component to capture the intricate heterogeneity between instance bags in complex real-world datasets. Additionally, features such as bag-level covariates do not directly influence label prediction but may introduce confounding effects; thus, we treat them as fixed effects. This mixed-effects modeling approach aligns with our intuitive understanding of multi-instance partial labeling and effectively facilitates regression in realistic applications.

Efficiency Comparison

Evaluating the efficiency of the compared MIPL methods is essential, as our primary objective is to devise a method capable of efficiently processing complicated MIPL datasets. Therefore, we also incorporate time consumption as a metric for all competing algorithms under identical conditions.

In the efficiency analysis, only MIPL algorithms are considered because PLL and MIL algorithms generally demonstrate poor effectiveness in complex MIPL problem scenarios, as evident in the aforementioned effectiveness comparison. All experiments are performed on a machine with an Intel Core i7-13700K CPU, 64 GB main memory, and a single NVIDIA GeForce RTX 4090 GPU. Three MIPL algorithms are considered, including FastMIPL, EliMIPL, and DeMIPL, since MIPLGP could not be deployed with a single RTX 4090 GPU. Figure 3 illustrates the time consumption of each algorithm with varying data sizes, which records the time consumption to conduct an experiment on benchmark datasets and a real-world dataset. Figure 3 shows the least time consumption for different numbers of bags for each algorithm, and illustrates the time consumption for different counts of instances with the same number of bags. FastMIPL, represented by the red line, achieves lower time consumption compared to other MIPL algorithms. FastMIPL demonstrates superior computational efficiency, particularly in high-dimensional scenarios characterized by increasing multiplicities of bags or instances.

We combine the accuracy-runtime performance curves for FastMIPL and comparative algorithms on five datasets in Figure 4. Points represent different methods with various shapes of markers, and lines denote the accuracy-runtime performance for FastMIPL and comparative algorithms on the same dataset. Clearly, the closer the point of the MIPL algorithm are to the point $(0, 1)$, the better the performance obtained. FastMIPL, which the marker is denoted by the diamond symbol, consistently outperforms the other algorithm on all datasets. It is worth mentioning that FastMIPL can reduce the time cost of model training by up to nearly 20 times compared to EliMIPL and up to about 30 times compared to DeMIPL while maintaining the prediction accuracy on the CRC-MIPL dataset.

Variant Comparison

To further elucidate the inner workings of FastMIPL, we investigated two variants, FastMIPL-v1 and FastMIPL-

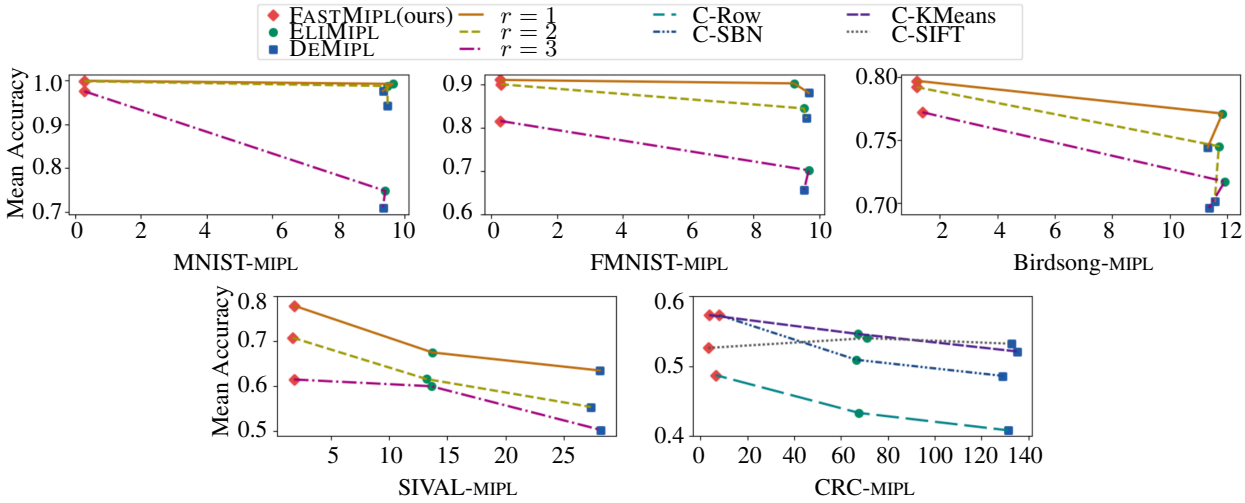

Figure 4: The performance of FastMIPL and comparative algorithms based on mean accuracy and time consumption, which subplots' horizontal axis represents the time consumption (minutes) and their vertical axis shows values of average accuracy.

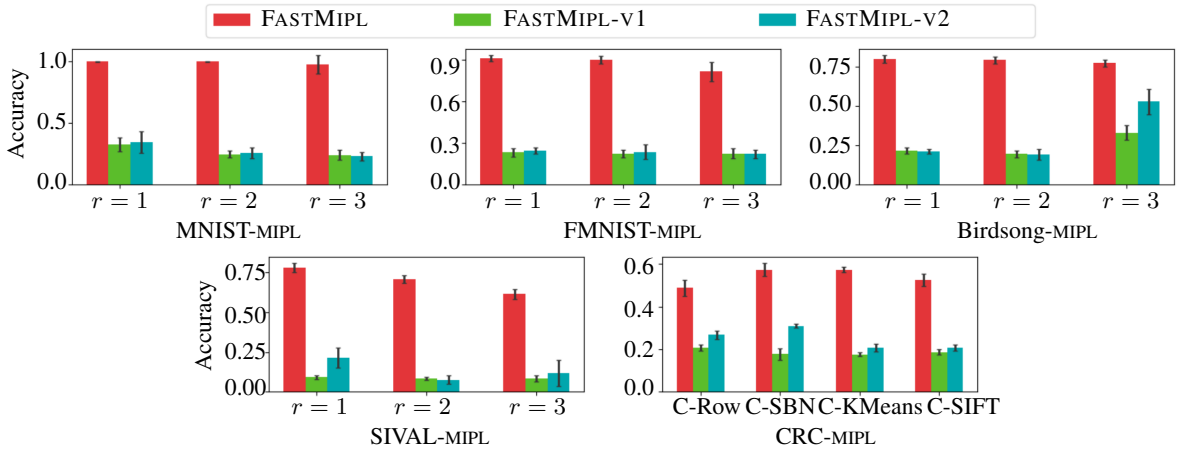

Figure 5: The effectiveness comparison among FastMIPL and its variant algorithms, subplots' horizontal axis means the number of false positives or names of sub-datasets, and their vertical axis denotes average and standard deviation of accuracy.

v2. FastMIPL-v1 removes the random effect coefficient β from Equation 3 and learns a linear model for each bag directly. This variant is designed to examine the efficacy of incorporating random effects in addressing the complexities inherent in MIPL, which exhibits a hierarchical nested structure. FastMIPL-v2, on the other hand, removes the parameter γ from Equation 3, aiming to assess the effectiveness of constructing an embeddings space for feature extraction and instance-level information aggregation.

Figure 5 presents the effectiveness comparisons on real-world and benchmark datasets, respectively. The results demonstrate that FastMIPL-v1 exhibits significantly lower accuracy across all datasets compared to FastMIPL. This underscores the crucial role of leveraging random effects to capture the individual contributions of instances within each bag towards the bag-level label, thereby effectively addressing the inherent heterogeneity in MIPL. FastMIPL-v2 consistently underperforms FastMIPL across all datasets, which highlights the efficacy of retaining the computational advantages of a linear model while effectively representing the instance contributions in a probabilistic framework.

Conclusion

Existing MIPL approaches are typically too time-consuming to handle complicated data. We propose FastMIPL to learn from MIPL data both effectively and efficiently. On one hand, effectiveness is achieved by leveraging mixed effects to capture the complex dependencies and heterogeneity inherent in MIPL. On the other hand, efficiency is significantly enhanced by maintaining the computational advantages of a linear model and jointly optimizing the ELBO using mini-batch SGD. Future studies could explore advanced statistical conjectures to develop more effective models.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Science Foundation of China (62225602) and the Big Data Computing Center of Southeast University.

References

Aminabadi, R. Y.; Rajbhandari, S.; Awan, A. A.; Li, C.; Li, D.; Zheng, E.; Ruwase, O.; Smith, S.; Zhang, M.-J.; Rasley, J.; et al. 2022. Deepspeed-inference: enabling efficient inference of transformer models at unprecedented scale. In *Proceedings of the 34th International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–15. Dallas, TX.

Breslow, N. E.; and Clayton, D. G. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421): 9–25.

Cui, Y.-F.; Liu, Z.-Q.; Liu, X.-Y.; Liu, X.; Wang, C.; Kuo, T.-W.; Xue, C. J.; and Chan, A. B. 2023. Bayes-MIL: A New Probabilistic Perspective on Attention-based Multiple Instance Learning for Whole Slide Images. In *Proceedings of the 11st International Conference on Learning Representations*. Kigali, Rwanda.

Cuomo, A. S.; Heinen, T.; Vagiaki, D.; Horta, D.; Marioni, J. C.; and Stegle, O. 2022. CellRegMap: a statistical framework for mapping context-specific regulatory variants using scRNA-seq. *Molecular Systems Biology*, 18(8).

Engelmann, J. P.; Palma, A.; Tomczak, J. M.; Theis, F.; and Casale, F. P. 2024. Mixed Models with Multiple Instance Learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, 3664–3672. Palau de Congressos, Valencia.

Feng, L.; Lv, J.-Q.; Han, B.; Xu, M.; Niu, G.; Geng, X.; An, B.; and Sugiyama, M. 2020. Provably Consistent Partial-Label Learning. In *Advances in Neural Information Processing Systems 33*, 10948–10960. Virtual Event.

Gao, Y.; Xu, M.; and Zhang, M.-L. 2024. Complementary to Multiple Labels: A Correlation-Aware Correction Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9179–9191.

Hang, J.-Y.; and Zhang, M.-L. 2023. Partial Multi-Label Learning with Probabilistic Graphical Disambiguation. In *Advances in Neural Information Processing Systems 36*. New Orleans, LA.

Haußmann, M.; Hamprecht, F. A.; and Kandemir, M. 2017. Variational bayesian multiple instance learning with gaussian processes. In *Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6570–6579. Honolulu, HI.

Hüllermeier, E.; and Beringer, J. 2006. Learning from Ambiguously Labeled Examples. *Intelligent Data Analysis*, 10(5): 419–439.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based Deep Multiple Instance Learning. In *Proceedings of the 35th International Conference on Machine Learning*, 2127–2136. Stockholmsmässan, Stockholm.

Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R. J.; and Wasserman, L. 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111.

Liu, L.-P.; and Dietterich, T. 2012. A Conditional Multinomial Mixture Model for Superset Label Learning. In *Advances in Neural Information Processing Systems 25*, 557–565. Lake Tahoe, Nevada.

Liu, L.-P.; and Dietterich, T. 2014. Learnability of the Superset Label Learning Problem. In *Proceedings of the 31th International Conference on Machine Learning*, 1629–1637. Beijing, China.

Liu, Y.; Wu, Y.-H.; Sun, G.; Zhang, L.; Chhatkuli, A.; and Van Gool, L. 2024. Vision Transformers with Hierarchical Attention. *Machine Intelligence Research*, 21(4): 670–683.

Lowe, D. G. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2): 91–110.

Lv, J.-Q.; Liu, Y.-F.; Xia, S.-Y.; Xu, N.; Xu, M.; Niu, G.; Zhang, M.-L.; Sugiyama, M.; and Geng, X. 2024. What Makes Partial-Label Learning Algorithms Effective? In *Advances in Neural Information Processing Systems 37*. Vancouver, Canada.

Lv, J.-Q.; Xu, M.; Feng, L.; Niu, G.; Geng, X.; and Sugiyama, M. 2020. Progressive Identification of True Labels for Partial-Label Learning. In *Proceedings of the 37th International Conference on Machine Learning*, 6500–6510. Virtual Event.

Mao, J.-X.; Wang, W.; and Zhang, M.-L. 2023. Label Specific Multi-Semantics Metric Learning for Multi-Label Classification: Global Consideration Helps. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, 4055–4063. Macao, China.

Maron, O.; and Ratan, A. L. 1998. Multiple-Instance Learning for Natural Scene Classification. In *Proceedings of the 15th International Conference on Machine Learning*, 341–349. Madison, Wisconsin.

Moore, R.; Casale, F. P.; Jan Bonder, M.; Horta, D.; Franke, L.; Barroso, I.; and Stegle, O. 2019. A linear mixed-model approach to study multivariate gene–environment interactions. *Nature genetics*, 51(1): 180–186.

Oh, C.-Y.; Tomczak, J.; Gavves, E.; and Welling, M. 2019. Combinatorial Bayesian Optimization using the Graph Cartesian Product. In *Advances in Neural Information Processing Systems 32*, 2910–2920. Vancouver, BC.

Pal, S.; Valkanas, A.; Regol, F.; and Coates, M. 2022. Bag graph: Multiple instance learning using bayesian graph neural networks. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 7922–7930. Virtual Event.

Pirš, G.; and Štrumbelj, E. 2019. Bayesian combination of probabilistic classifiers using multivariate normal mixtures. *Journal of Machine Learning Research*, 20(51): 1–18.

Qiao, C.-Y.; Xu, N.; and Geng, X. 2023. Decompositional generation process for instance-dependent partial label learning. In *Proceedings of the 11st International Conference on Learning Representations*. Kigali, Rwanda.

Ranganath, R.; Gerrish, S.; and Blei, D. 2014. Black box variational inference. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, 814–822. Reykjavik, Iceland.

Ren, L.-J.; Jiang, L.-X.; Zhang, W.-J.; and Li, C.-Q. 2024. Label distribution similarity-based noise correction for crowdsourcing. *Frontiers of Computer Science*, 18(5): 185323.

Shi, X.-S.; Xing, F.-Y.; Xie, Y.-P.; Zhang, Z.-Z.; Cui, L.; and Yang, L. 2020. Loss-Based Attention for Deep Multiple Instance Learning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 5742–5749. New York, NY.

Tang, W.; Yang, Y.-F.; Wang, Z.; Zhang, W.; and Zhang, M.-L. 2024. Multi-Instance Partial-Label Learning with Margin Adjustment. In *Advances in Neural Information Processing Systems 37*. Vancouver, Canada.

Tang, W.; Zhang, W.; and Zhang, M.-L. 2023. Disambiguated Attention Embedding for Multi-Instance Partial-Label Learning. In *Advances in Neural Information Processing Systems 36*, 56756–56771. New Orleans, LA.

Tang, W.; Zhang, W.-J.; and Zhang, M.-L. 2024a. Exploiting Conjugate Label Information for Multi-Instance Partial-Label Learning. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, 4973–4981. Jeju, South Korea.

Tang, W.; Zhang, W.-J.; and Zhang, M.-L. 2024b. Multi-instance partial-label learning: Towards exploiting dual inexact supervision. *Science China Information Sciences*, 67(3): 1–14.

Tian, Y.-J.; Yu, X.-T.; and Fu, S.-J. 2023. Partial label learning: Taxonomy, analysis and outlook. *Neural Networks*, 161: 708–734.

Tomczak, J.; and Welling, M. 2018. VAE with a VampPrior. In *Proceedings of the 21st International Conference on International Conference on Artificial Intelligence and Statistics*, volume 84, 1214–1223. Lanzarote, Canary Islands.

van Krieken, E.; Thanapalasingam, T.; Tomczak, J.; Van Harmelen, F.; and Ten Teije, A. 2023. A-NeSI: A Scalable Approximate Method for Probabilistic Neurosymbolic Inference. In *Advances in Neural Information Processing Systems 36*. New Orleans, LA.

Wang, D.-B.; Li, L.; and Zhang, M.-L. 2019. Adaptive Graph Guided Disambiguation for Partial Label Learning. In *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 83–91. Anchorage, AK.

Wei, X.-S.; and Zhou, Z.-H. 2016. An Empirical Study on Image Bag Generators for Multi-Instance Learning. *Machine Learning*, 105: 155–198.

Wen, H.-W.; Cui, J.-Y.; Hang, H.-Y.; Liu, J.-B.; Wang, Y.-S.; and Lin, Z.-C. 2021. Leveraged Weighted Loss for Partial Label Learning. In *Proceedings of the 38th International Conference on Machine Learning*, 11091–11100. Virtual Event.

Wibawa, M. S.; Lo, K.-W.; Young, L.; and Rajpoot, N. 2022. Multi-Scale Attention-based Multiple Instance Learning for Classification of Multi-Gigapixel Histology Images. In *Proceedings of the 17th European Conference on Computer Vision*, 635–647. Tel Aviv, Israel.

Wu, D.-D.; Wang, D.-B.; and Zhang, M.-L. 2024. Distilling Reliable Knowledge for Instance-Dependent Partial Label Learning. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 15888–15896.

Xia, S.-Y.; Lv, J.-Q.; Xu, N.; and Geng, X. 2022. Ambiguity-Induced Contrastive Learning for Instance-Dependent Partial Label Learning. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, 3615–3621. Vienna, Austria.

Xia, S.-Y.; Lv, J.-Q.; Xu, N.; Niu, G.; and Geng, X. 2023. Towards effective visual representations for partial-label learning. In *Proceedings of the 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15589–15598. Vancouver, Canada.

Xu, N.; Qiao, C.-Y.; Geng, X.; and Zhang, M.-L. 2021. Instance-dependent partial label learning. In *Advances in Neural Information Processing Systems 34*, 27119–27130. Virtual Event.

Yang, Y.-F.; Tang, W.; and Zhang, M.-L. 2024. ProMIPL: A probabilistic generative model for multi-instance partial-label learning. In *Proceedings of the 24th IEEE International Conference on Data Mining*. Abu Dhabi, UAE.

Yao, Y.; Deng, J.; Chen, X.; Gong, C.; Wu, J.; and Yang, J. 2020. Deep Discriminative CNN with Temporal Ensembling for Ambiguously-Labeled Image Classification. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 12669–12676. New York, NY.

Yildizer, E.; Balci, A. M.; Hassan, M.; and Alhajj, R. 2002. Content-Based Image Retrieval Using Multiple-Instance Learning. In *Proceedings of the 19th International Conference on Machine Learning*, 682–689. Sydney, Australia.

Zhang, F.; Feng, L.; Han, B.; Liu, T.-L.; Niu, G.; Qin, T.; and Sugiyama, M. 2022. Exploiting Class Activation Value for Partial-Label Learning. In *Proceedings of the 10th International Conference on Learning Representations*, 1–17. Virtual Event.

Zhang, K.; and Hyvarinen, A. 2009. On the Identifiability of the Post-Nonlinear Causal Model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 647–655. Montreal, QC.

Zhang, K.; Peters, J.; Janzing, D.; and Schölkopf, B. 2011. Kernel-based Conditional Independence Test and Application in Causal Discovery. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, 804–813. Barcelona, Spain.

Zhang, M.-L.; and Zhang, K. 2010. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 999–1008. Washington, DC.

Zhou, Z.-H. 2018. A brief introduction to weakly supervised learning. *National science review*, 5(1): 44–53.