Exploiting Conjugate Label Information for Multi-Instance Partial-Label Learning

Wei Tang[1,2] , **Weijia Zhang**[3] , **Min-Ling Zhang**[1,2*]

[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Lab. of Computer Network and Information Integration (Southeast University), MoE, China
[3]School of Information and Physical Sciences, The University of Newcastle, NSW 2308, Australia
tangw@seu.edu.cn, weijia.zhang@newcastle.edu.au, zhangml@seu.edu.cn

Abstract

Multi-instance partial-label learning (MIPL) addresses scenarios where each training sample is represented as a multi-instance bag associated with a candidate label set containing one true label and several false positives. Existing MIPL algorithms have primarily focused on mapping multi-instance bags to candidate label sets for disambiguation, disregarding the intrinsic properties of the label space and the supervised information provided by non-candidate label sets. In this paper, we propose an algorithm named ELIMIPL, i.e., *Exploiting conjugate Label Information for Multi-Instance Partial-Label learning*, which exploits the conjugate label information to improve the disambiguation performance. To achieve this, we extract the label information embedded in both candidate and non-candidate label sets, incorporating the intrinsic properties of the label space. Experimental results obtained from benchmark and real-world datasets demonstrate the superiority of the proposed ELIMIPL over existing MIPL algorithms and other well-established partial-label learning algorithms.

1 Introduction

Weakly supervised learning has emerged as a powerful strategy in scenarios with limited annotated data. Based on label quality and quantity, weak supervision can be broadly categorized into three types: inaccurate, inexact, and incomplete supervision [Zhou, 2018]. Inexact supervision refers to a coarse correspondence between instances and labels. To work with inexact supervision, these are two prevalent learning paradigms, i.e., *multi-instance learning* (MIL) [Amores, 2013; Carbonneau *et al.*, 2018; Ilse *et al.*, 2018; Zhang *et al.*, 2022c,b] and *partial-label learning* (PLL) [Cour *et al.*, 2011; Lyu *et al.*, 2020; Zhang *et al.*, 2022a; He *et al.*, 2022; Gong *et al.*, 2022; Li *et al.*, 2023]. In MIL, a sample is represented as a *bag of instances* and associated with a *single bag-level label*, while the instance-level labels are inaccessible to the learner. In PLL, a sample is represented as a *single instance* and linked to a *candidate label set*, including one true label and multiple false positives. Therefore, MIL and PLL can be perceived as two sides of the same coin: inexact supervision within MIL manifests in the instance space, whereas inexact supervision appears in the label space within PLL.

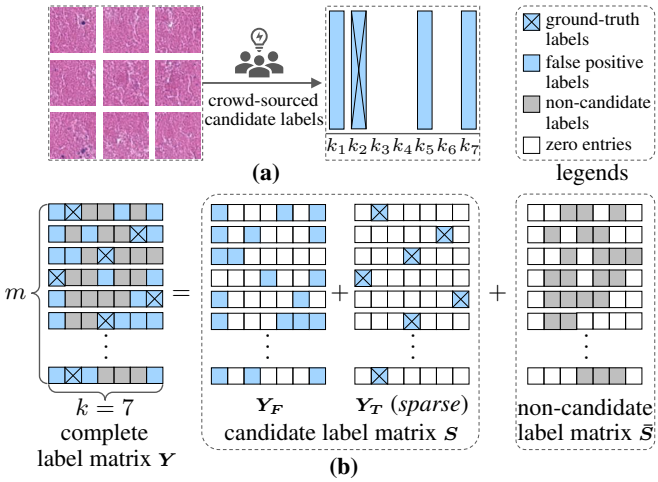


Figure 1: **(a)** A multi-instance bag is labeled with a candidate label set $\mathcal{S} = \{k_1, k_2, k_5, k_7\}$. **(b)** The decomposition of the complete label matrix, where m and k represent the number of multi-instance bags and categories, respectively.

However, many tasks exhibit a phenomenon of *dual inexact supervision*, where ambiguity arises in both instance and label spaces. To work with the dual inexact supervision, Tang *et al.* [2024] introduced a learning paradigm known as *multi-instance partial-label learning* (MIPL) and developed a Gaussian Processes-based algorithm (MIPLGP), which derives a bag-level predictor by aggregating predictions of all instances within the same bag. To capture global representations for multi-instance bags, an algorithm named DEMIPL equipped with an attention mechanism is introduced [Tang *et al.*, 2023]. The existing algorithms mainly operate in the instance space and only utilize the candidate label information.

The non-candidate label set holds crucial roles in MIPL. In histopathological image classification, images are commonly segmented into patches [Campanella *et al.*, 2019; Lu *et al.*, 2021], and their labels may come from crowd-sourced annotators rather than expert pathologists [Irshad *et al.*, 2017; Grote *et al.*, 2019]. Figure 1(a) illustrates that crowd-sourced annotators treat an image as a multi-instance bag $\boldsymbol{X}_i$ =

*Corresponding author

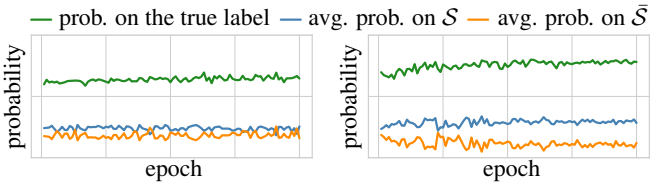


Figure 2: Predicted probabilities of DEMIPL (left) and ELIMIPL (right) on the sample in CRC-MIPL-Row dataset.

$\{\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \cdots, \boldsymbol{x}_{i,9}\}$ and provide a candidate label set $\mathcal{S}_i = \{k_1, k_2, k_5, k_7\}$, whose candidate label matrix can be written as $\boldsymbol{S}_i = [1, 1, 0, 0, 1, 0, 1]$. Similarly, the non-candidate label set $\bar{\mathcal{S}}_i = \{k_3, k_4, k_6\}$ corresponds to the non-candidate label matrix $\bar{\boldsymbol{S}}_i = [0, 0, 1, 1, 0, 1, 0]$, indicating that $\boldsymbol{X}_i$ must not belong to categories k_3, k_4, or k_6. Therefore, we can extract exact supervision from the non-candidate label set. As depicted in Figure 1(b), we decompose a complete label matrix $\boldsymbol{Y}$ into a candidate label matrix $\boldsymbol{S}$ and a non-candidate label matrix $\bar{\boldsymbol{S}}$. Subsequently, $\boldsymbol{S}$ may be further disintegrated into a false positive label matrix $\boldsymbol{Y_F}$ and a true label matrix $\boldsymbol{Y_T}$, i.e., $\boldsymbol{Y} = \boldsymbol{S} + \bar{\boldsymbol{S}} = \boldsymbol{Y_F} + \boldsymbol{Y_T} + \bar{\boldsymbol{S}}$. Notably, $\boldsymbol{Y_T}$ is sparse, as each row must have one and only one non-zero element. However, the current MIPL algorithms have predominantly concentrated on the mappings from multi-instance bags to $\boldsymbol{S}$, neglecting the sparsity of $\boldsymbol{Y_T}$ and the information from $\bar{\boldsymbol{S}}$.

Consequently, Figure 2 illustrates the predicted probabilities on the true label, along with the average predicted probabilities on each candidate label and non-candidate label. The left side depicts the probabilities of the DEMIPL, revealing proximity in the average predicted probabilities on candidate and non-candidate labels. This observation indicates that DEMIPL *encounters difficulty in effectively discerning between candidate and non-candidate labels*. To address this challenge, we introduce the concept of *conjugate label information* (CLI), encapsulating information from both candidate and non-candidate label sets, along with the sparsity of the true label matrix. The right side in Figure 2 shows the predicted probabilities when exploiting the CLI. It is evident that (a) the predicted probabilities on the true label exhibit a noticeable increase, (b) the average predicted probabilities on the non-candidate label are reduced, and (c) the average probabilities on each candidate label and non-candidate label are distinctly separated. This suggests that *the CLI conduce to train a more discriminative MIPL classifier*.

In this paper, we present an algorithm named ELIMIPL, i.e., *Exploiting conjugate Label Information for Multi-Instance Partial-Label learning*. Firstly, we introduce a scaled additive attention mechanism to aggregate each multi-instance bag into a bag-level feature representation. Secondly, to enhance the utilization of candidate label information, we leverage the mappings from the bag-level features to the candidate label sets, coupled with the sparsity of the candidate label matrix. Lastly, to incorporate the non-candidate label information, we propose an inhibition loss to diminish the model's predictions on the non-candidate labels. To the best of our knowledge, we are the first to introduce the scaled additive attention mechanism and the CLI in MIPL. Extensive experimental results demonstrate that ELIMIPL outperforms the state-of-the-art MIPL algorithms and the PLL algorithms.

The remainder is organized as follows. Firstly, we review related work in Section 2. Secondly, we present the proposed ELIMIPL in Section 3 and report the experimental results in Section 4. Lastly, we conclude this paper in Section 5.

2 Related Work

2.1 Multi-Instance Learning

Originating from drug activity prediction [Dietterich *et al.*, 1997], MIL has found extensive adoption in diverse applications, including text classification [Zhou *et al.*, 2009; Zhang, 2021] and image annotation [Wang *et al.*, 2018]. Contemporary deep MIL approaches predominantly rely on attention mechanisms [Wang *et al.*, 2022b; Chen *et al.*, 2022; Tan *et al.*, 2023]. Ilse *et al.* [2018] introduced attention mechanisms to aggregate each multi-instance bag into a feature vector. For multi-classification tasks, Shi *et al.* [2020] proposed a loss-based attention mechanism to learn instance-level weights, predictions, and bag-level predictions. Furthermore, researchers have explored the intrinsic attributes of attention mechanisms to improve performance [Cui *et al.*, 2023; Xiang *et al.*, 2023]. While these approaches achieve promising results in cases with exact bag-level labels, they face challenges in learning from ambiguous bag-level labels.

2.2 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 empirical risk minimization principle, Lv *et al.* [2020] devised a classifier-consistent risk estimator that progressively identifies true labels. Similarly, Feng *et al.* [2020] delved into the generation process of partial-labeled data, proposing both a risk-consistent approach and a classifier-consistent approach. Taking a more generalized stance, Wen *et al.* [2021] presented a weighted loss function capable of accommodating various methods through distinct weight assignments. Furthermore, Wu *et al.* [2022] proposed a supervised loss to constrain outputs on non-candidate labels, coupled with consistency regularization on candidate labels. While the supervised loss bears resemblance to our inhibition loss, our proposed CLI loss incorporates additional components, namely the mapping loss and the sparse loss. Although these methods effectively learn from partial-labeled data, they lack the capability to manage multi-instance bags.

2.3 Multi-Instance Partial-Label Learning

In contrast to the inherent limitations of addressing only unilateral inexact supervision in MIL and PLL, MIPL possesses the capability to work with dual inexact supervision. To the best of our knowledge, there are only two viable MIPL algorithms. Tang *et al.* [2024] is the first to introduce the framework of MIPL along with a Gaussian processes-based algorithm (MIPLGP), which follows an instance-space paradigm. MIPLGP begins by augmenting a negative class for each candidate label set, subsequently treating the candidate label set of each multi-instance bag as that of each instance within the bag. Finally, it employs the Dirichlet disambiguation strategy and the Gaussian processes regression

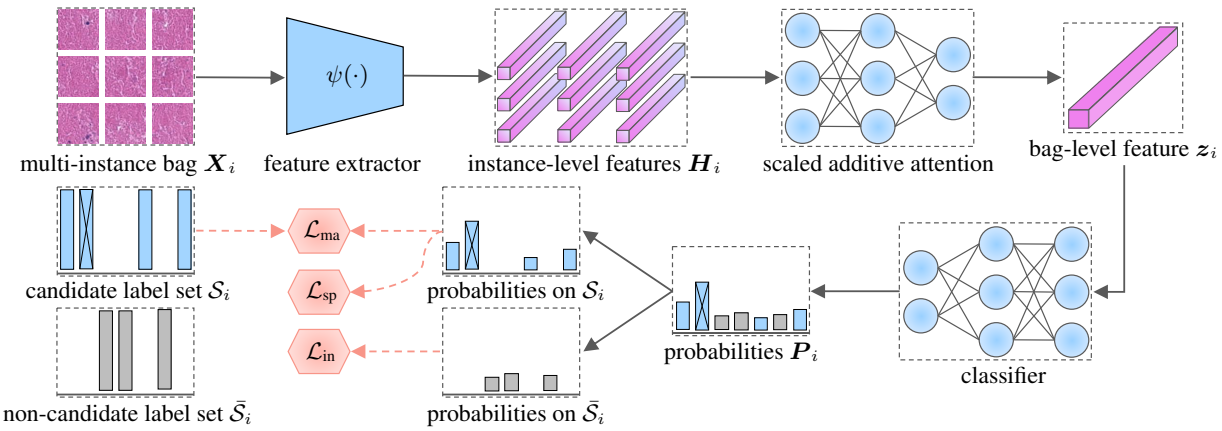


Figure 3: The pipeline of ELIMIPL, where $\mathcal{L}_{\text{ma}}$, $\mathcal{L}_{\text{sp}}$, and $\mathcal{L}_{\text{in}}$ refer to mapping loss, sparsity loss, and inhibition loss, respectively.

model for disambiguation. Differing from MIPLGP, DEMIPL follows the embedded-space paradigm and aggregates each multi-instance bag into a feature representation and employs a momentum-based disambiguation strategy to find true labels from candidate label sets [Tang *et al.*, 2023]. However, both methods primarily depend on mapping from instances or multi-instance bags to candidate label sets for disambiguation, without considering the proposed CLI in this paper.

3 Methodology

3.1 Preliminaries

In this study, we define a MIPL training dataset as $\mathcal{D} = \{(\boldsymbol{X}_i, \mathcal{S}_i) \mid 1 \leq i \leq m\}$, comprising m multi-instance bags and their corresponding candidate label sets. Specifically, a candidate label set $\mathcal{S}_i$ consists of one true label and multiple false positive labels, but the true label is unknown. It is crucial to note that a bag contains at least one instance pertaining to the true label, while excluding any instances corresponding to false positive labels. The instance space is denoted as $\mathcal{X} \in \mathbb{R}^d$, while the label space $\mathcal{Y} = \{1, 2, \cdots, k\}$ encompasses k class labels. The i-th bag $\boldsymbol{X}_i = \{\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \cdots, \boldsymbol{x}_{i,n_i}\}$ comprises n_i instances of dimension d. Both the candidate label set $\mathcal{S}_i$ and the non-candidate label set $\bar{\mathcal{S}}_i$ are proper subsets of the label space $\mathcal{Y}$, satisfying the conditions $|\mathcal{S}_i| + |\bar{\mathcal{S}}_i| = |\mathcal{Y}| = k$, where $|\cdot|$ denotes the cardinality of a set.

The pipeline of the proposed ELIMIPL is depicted in Figure 3, which contains three main components: an instance-level feature extractor, a scaled additive attention mechanism, and a classifier. When presented with a multi-instance bag $\boldsymbol{X}_i$ along with its associated candidate label set $\mathcal{S}_i$ and non-candidate label set $\bar{\mathcal{S}}_i$, we initially employ a feature extractor to procure instance-level feature representations. Subsequently, the scaled additive attention mechanism is applied to aggregate a bag of instances into a unified bag-level feature representation. Finally, the classifier is invoked to estimate the class probabilities based on the bag-level features. To utilize the CLI, we introduce a mapping loss $\mathcal{L}_{\text{ma}}$ and a sparsity loss $\mathcal{L}_{\text{sp}}$ to disambiguate the candidate label sets, along with an inhibition loss $\mathcal{L}_{\text{in}}$ to suppress the model's prediction over the non-candidate label sets.

3.2 Instance-Level Feature Extractor

For a given multi-instance bag $\boldsymbol{X}_i = \{\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \cdots, \boldsymbol{x}_{i,n_i}\}$ with n_i instances, instance-level feature representations $\boldsymbol{H}_i$ are learned using a feature extractor $\psi(\cdot)$ as follows:

$$\boldsymbol{H}_i = \psi(\boldsymbol{X}_i) = \{\boldsymbol{h}_{i,1}, \boldsymbol{h}_{i,2}, \cdots, \boldsymbol{h}_{i,n_i}\}, \tag{1}$$

where $\boldsymbol{h}_{i,j} \in \mathbb{R}^l$ indicates the feature representation of the j-th instance within the i-th multi-instance bag, and $\psi(\cdot)$ is a neural network comprised of two components, i.e., $\psi(\boldsymbol{X}_i) = \psi_2(\psi_1(\boldsymbol{X}_i))$. Here, $\psi_1(\cdot)$ is a feature extractor that can be tailored to the specific characteristics of the datasets, and $\psi_2(\cdot)$ is composed of fully connected layers that map instance-level features to an embedded space of dimension l.

3.3 Scaled Additive Attention Mechanism

To aggregate instance-level features into bag-level representations, we introduce a scaled additive attention mechanism specifically designed for MIPL. The existing attention mechanism for MIPL utilizes the sigmoid function for calculating attention scores, followed by normalization [Tang *et al.*, 2023]. The attention scores derived through the sigmoid function are constrained within the range $(0, 1)$, leading to a limited distinction between instances. Therefore, we introduce an additive attention mechanism calculating attention scores by the softmax function to distinguish instances, equipped with a scaling factor to prevent vanishing gradients [Vaswani *et al.*, 2017]. Specifically, we first denote the output of the additive attention mechanism as $\xi(h_{i,j})$, quantifying the impact of the j-th instance on the i-th bag as follows:

$$\xi(h_{i,j}) = \boldsymbol{W}^\top(\tanh(\boldsymbol{W}_t^\top \boldsymbol{h}_{i,j} + \boldsymbol{b}_t) \odot \text{sigm}(\boldsymbol{W}_s^\top \boldsymbol{h}_{i,j} + \boldsymbol{b}_s)), \tag{2}$$

where $\boldsymbol{W}^\top$, $\boldsymbol{W}_t^\top$, $\boldsymbol{W}_s^\top$, $\boldsymbol{b}_t$, and $\boldsymbol{b}_s$ are learnable parameters. $\tanh(\cdot)$ and $\text{sigm}(\cdot)$ are the hyperbolic tangent and sigmoid functions, respectively. The operator $\odot$ denotes element-wise multiplication. Then, we normalize $\xi(h_{i,j})$ using softmax with a scaling factor $1/\sqrt{l}$ to derive the attention score:

$$a_{i,j} = \frac{\exp\left(\xi\left(h_{i,j}\right)/\sqrt{l}\right)}{\sum_{j'=1}^{n_i} \exp\left(\xi\left(h_{i,j'}\right)/\sqrt{l}\right)}, \tag{3}$$

where $a_{i,j}$ represents the attention score of the j-th instance in the i-th bag. Finally, we consolidate the instance-level features into a bag-level representation, as demonstrated below:

$$\boldsymbol{z}_i = \sum_{j=1}^{n_i} a_{i,j} \boldsymbol{h}_{i,j}, \tag{4}$$

where $\boldsymbol{z}_i$ represents the bag-level representation of the i-th multi-instance bag. The bag-level representations of all multi-instance bags in the training dataset are denoted by $\mathcal{Z}$.

3.4 Conjugate Label Information

Candidate Label Information Once the bag-level feature representations are acquired, the subsequent task is to disambiguate the candidate label set. The disambiguation entails establishing the mapping relationship from the bag-level features to their corresponding candidate label set. The goal of precise mapping is to guide the classifier to assign higher class probabilities to true labels and lower probabilities to false positive labels. To attain this objective, we employ a weighted mapping loss function:

$$\mathcal{L}_{\text{ma}}(\mathcal{Z}, \mathcal{S}) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{c \in \mathcal{S}_i} w_{i,c}^{(t)} \log(f_c(\boldsymbol{z}_i)), \tag{5}$$

where f is the classifier, and $f_c(\cdot)$ represents the classifier's prediction probability for the candidate label c. $w_{i,c}^{(t)}$ denotes the weight assigned to the prediction of the c-th class at the t-th epoch, using the features of the i-th bag as input for the classifier. For candidate labels, we initialize $w_{i,c}^{(0)} = \frac{1}{|\mathcal{S}_i|}$ through an averaging approach. During training, we update $w_{i,c}^{(t)}$ by computing a weighted sum of the classifier's outputs at both the previous epoch and current epoch as follows:

$$w_{i,c}^{(t)} = \rho^{(t)} w_{i,c}^{(t-1)} + (1 - \rho^{(t)}) \frac{f_c(\boldsymbol{z}_i)}{\sum_{c' \in \mathcal{S}_i} f_{c'}(\boldsymbol{z}_i)}, \tag{6}$$

where $\rho^{(t)} = (T - t)/T$ is dynamically adjusted across epochs, and T is the maximum of the training epochs.

While the mapping loss can assess the relative labeling probabilities of candidate labels, it fails to capture the mutually exclusive relationships among the candidate labels. To address this issue in PLL, Feng and An [2019] introduced the maximum infinity norm on the predicted probabilities of all classes and alternately optimize the maximum infinity norm by solving k independent quadratic programming problems. However, as depicted in Figure 1(b), we observe that each row of the true label matrix exhibits sparsity. Although the true labels remain inaccessible during the training process, we encourage the classifier to generate sparse prediction probabilities for the candidate labels. Specifically, the goal is to push the prediction probability of the unknown true label toward 1 while simultaneously driving the prediction probabilities of other candidate labels toward 0. Therefore, we directly capture the mutually exclusive relationships among the candidate labels by implementing the sparsity loss, as detailed below:

$$\mathcal{L}_{\text{sp}}(\mathcal{S}) = \frac{1}{m} \sum_{i=1}^{m} \|\boldsymbol{P}_i \odot \boldsymbol{S}_i\|_0, \tag{7}$$

Algorithm 1 Training Procedure of ELIMIPL

Inputs:
$\mathcal{D}$: MIPL training set $\{(\boldsymbol{X}_i, \mathcal{S}_i) \mid 1 \leq i \leq m\}$
μ, γ : Weights for sparsity loss and inhibition loss
T: Maximum number of epochs
Process:
1: Initialize uniform weights $w_{i,c}^{(0)}$ $(c \in \mathcal{S}_i)$
2: **for** $t = 1$ to T **do**
3: Fetch a mini-batch $\mathcal{B}$ from $\mathcal{D}$
4: **for** $\boldsymbol{X} \in \mathcal{B}$ **do**
5: Extract instance-level features using Equation (1)
6: Calculate attention scores using Equations (2, 3)
7: Aggregate instance-level features into bag-level feature representations via Equation (4)
8: Update weights $w_{i,c}^{(t)}$ based on Equation (6)
9: Calculate $\mathcal{L}_{\text{ma}}$, $\mathcal{L}_{\text{sp}}$, and $\mathcal{L}_{\text{in}}$ via Equations (5, 7, 8)
10: Calculate total loss $\mathcal{L}$ as in Equation (9)
11: Set gradient $-\bigtriangledown_{\Phi} \mathcal{L}$
12: Update Φ using optimizer
13: **end for**
14: **end for**

where $\boldsymbol{P}_i$ and $\boldsymbol{S}_i$ is the prediction probabilities and the candidate label set matrix of the i-th bag, respectively. $\odot$ denotes element-wise multiplication. Since minimizing the ℓ_0 norm is NP-hard, we employ the ℓ_1 norm as a surrogate for the ℓ_0 norm, promoting sparsity while allowing for efficient optimization [Tibshirani, 1996; Wright and Ma, 2022].

Non-candidate Label Information For a multi-instance bag $\boldsymbol{X}_i$ linked to a candidate label set $\mathcal{S}_i$, the non-candidate label set $\bar{\mathcal{S}}_i$ complements the candidate label set $\mathcal{S}_i$ within the label space $\mathcal{Y}$. As the label space has a fixed size, an antagonistic relationship arises between the non-candidate and candidate label sets. To enhance the classifier's prediction probabilities for the candidate label set, a natural strategy is to diminish the classifier's prediction probabilities for the non-candidate label set. Motivated by this insight, we introduce an inhibition loss as follows:

$$\mathcal{L}_{\text{in}}(\mathcal{Z}, \bar{\mathcal{S}}) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{\bar{c} \in \bar{\mathcal{S}}_i} \log(1 - f_{\bar{c}}(\boldsymbol{z}_i)), \tag{8}$$

where $f_{\bar{c}}(\cdot)$ denotes the classifier's prediction probability over the non-candidate label $\bar{c}$.

CLI Loss During the training, CLI is formed by a loss function named CLI loss that is a weighted fusion of the mapping loss, sparsity loss, and inhibition loss, as shown below:

$$\mathcal{L} = \mathcal{L}_{\text{ma}}(\mathcal{Z}, \mathcal{S}) + \mu \mathcal{L}_{\text{sp}}(\mathcal{S}) + \gamma \mathcal{L}_{\text{in}}(\mathcal{Z}, \bar{\mathcal{S}}), \tag{9}$$

where μ and γ represent the weighting coefficients for the sparsity loss and the inhibition loss, respectively.

Algorithm 1 summarizes the training procedure of ELIMIPL. Firstly, the algorithm initializes the weights for the mapping loss uniformly (Step 1). Subsequently, instance-level features are extracted and aggregated into bag-level features within each mini-batch (Steps 5-8). The algorithm then updates the weights for the mapping loss and calculates the total

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

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

loss function (Steps 9-11). Finally, the model is optimized using gradient descent (Steps 12 and 13).

4 Experiments

In this section, we begin by introducing the experimental configurations, including the datasets, comparative algorithms, and the parameters used in the experiments. Subsequently, we present the experimental results on both benchmark and real-world datasets. Finally, we conduct further analysis to gain deeper insights into the impact of CLI.

4.1 Experimental Configurations

Datasets We employ four benchmark MIPL datasets [Tang *et al.*, 2024, 2023]: MNIST-MIPL, FMNIST-MIPL, Birdsong-MIPL, and SIVAL-MIPL, spanning diverse domains such as image analysis and biology [LeCun *et al.*, 1998; Xiao *et al.*, 2017; Briggs *et al.*, 2012; Settles *et al.*, 2007]. The characteristics of the datasets are presented in Table 1, where the abbreviations within parentheses in the first column represent the abbreviated names of the MIPL datasets. The dataset includes quantities of multi-instance bags and total instances, denoted as *#bag* and *#ins*, respectively. Additionally, we use *max. #ins*, *min. #ins*, and *avg. #ins* to indicate the maximum, minimum, and average instance count within all bags. The dimensionality of the instance-level feature is represented by *#dim*. Labeling details are elucidated using *#class* and *avg. #CLs*, signifying the length of the label space and the average length of candidate label sets, respectively. For a comprehensive performance assessment, we vary the count of false positive labels, denoted as r ($|\mathcal{S}_i| = r + 1$).

CRC-MIPL dataset is a real-world MIPL dataset for colorectal cancer classification. We utilize multi-instance features generated by four image bag generators [Wei and Zhou, 2016]: Row [Maron and Ratan, 1998], single blob with neighbors (SBN) [Maron and Ratan, 1998], k-means segmentation (KMeansSeg) [Zhang *et al.*, 2002], and scale-invariant feature transform (SIFT) [Lowe, 2004].

The appendix contains detailed information about the datasets and the four image bag generators.

Comparative Algorithms We conduct a comprehensive comparison involving ELIMIPL along with two established MIPL algorithms: MIPLGP [Tang *et al.*, 2024] and DEMIPL [Tang *et al.*, 2023]. These represent the entirety of available MIPL methods. Furthermore, we include four PLL algorithms: PRODEN [Lv *et al.*, 2020], RC [Feng *et al.*, 2020], LWS [Wen *et al.*, 2021], and PL-AGGD [Wang *et al.*, 2022a]. The first three algorithms can be equipped with diverse backbone networks, such as linear models and MLP. Due to spatial constraints, we present the results obtained from the linear models in the main body, while the results with MLP are shown in the appendix. Parameters for all algorithms are selected based on recommendations from original literature or refined through our search for enhanced outcomes.

Since PLL algorithms are not directly tailored for MIPL data, two common strategies, known as the Mean strategy and the MaxMin strategy, are employed to adapt MIPL data for PLL algorithms [Tang *et al.*, 2024]. The Mean strategy involves calculating average feature values across all instances within a bag, resulting in a bag-level feature representation. In contrast, the MaxMin strategy identifies both the maximum and minimum feature values for each dimension among instances within a bag, and then concatenates these values to form a bag-level feature representation.

Implementation We implement ELIMIPL using PyTorch and execute it on a single NVIDIA Tesla V100 GPU. We utilize the stochastic gradient descent (SGD) optimizer with a momentum value of 0.9 and a weight decay of 0.0001. The initial learning rate is selected from the set $\{0.01, 0.05\}$ and accompanied by a cosine annealing technique. We set the number of epochs uniformly to 100 for all datasets. For the MNIST-MIPL and FMNIST-MIPL datasets, μ is set to 1 or 0.1, γ is chosen from $\{0.1, 0.5\}$, and the feature extraction network $\psi_1(\cdot)$ is a two-layer convolutional neural network. For the remaining datasets, we set both μ and γ to 10, and $\psi_1(\cdot)$ is an identity transformation. The feature transformation network $\psi_2(\cdot)$ is implemented by a fully connected network, with the dimension l set to 512 for the CRC-MIPL dataset and 128 for the other datasets. The way of dataset partitioning is consistent with that of DEMIPL. We conduct ten random train/test splits with a ratio of $7:3$. We report the mean accuracies and standard deviations obtained from the ten runs, with the highest accuracy highlighted in bold. The code of ELIMIPL can be found at https://github.com/tangw-seu/ELIMIPL.

4.2 Results on the Benchmark Datasets

Table 2 presents the results of ELIMIPL and the comparative algorithms on benchmark datasets, considering varying numbers of false positive labels ($r \in \{1, 2, 3\}$). Compared to MIPL algorithms, ELIMIPL consistently achieves higher average accuracy than DEMIPL and MIPLGP. Furthermore, in contrast to PLL algorithms, ELIMIPL significantly outperforms them in all cases.

For the MNIST-MIPL and FMNIST-MIPL datasets, each with

Algorithm	r	MNIST	FMNIST	Birdsong	SIVAL
ELIMIPL	1	**.992±.007**	**.903±.018**	**.771±.018**	**.675±.022**
	2	**.987±.010**	**.845±.026**	**.745±.015**	**.616±.025**
	3	**.748±.144**	**.702±.055**	**.717±.017**	**.600±.029**
DEMIPL	1	.976±.008	.881±.021	.744±.016	.635±.041
	2	.943±.027	.823±.028	.701±.024	.554±.051
	3	.709±.088	.657±.025	.696±.024	.503±.018
MIPLGP	1	.949±.016	.847±.030	.716±.026	.669±.019
	2	.817±.030	.791±.027	.672±.015	.613±.026
	3	.621±.064	.670±.052	.625±.015	.569±.032
Mean					
PRODEN	1	.605±.023	.697±.042	.296±.014	.219±.014
	2	.481±.036	.573±.026	.272±.019	.184±.014
	3	.283±.028	.345±.027	.211±.013	.166±.017
RC	1	.658±.031	.753±.042	.362±.015	.279±.011
	2	.598±.033	.649±.028	.335±.011	.258±.017
	3	.392±.033	.401±.063	.298±.009	.237±.020
LWS	1	.463±.048	.726±.031	.265±.010	.240±.014
	2	.209±.028	.720±.025	.254±.010	.223±.008
	3	.205±.013	.579±.041	.237±.005	.194±.026
PL-AGGD	1	.671±.027	.743±.026	.353±.019	.355±.015
	2	.595±.036	.677±.028	.314±.018	.315±.019
	3	.380±.032	.474±.057	.296±.015	.286±.018
MaxMin					
PRODEN	1	.508±.024	.424±.045	.387±.014	.316±.019
	2	.400±.037	.377±.040	.357±.012	.287±.024
	3	.345±.048	.309±.058	.336±.012	.250±.018
RC	1	.519±.028	.731±.027	.390±.014	.306±.023
	2	.469±.035	.666±.027	.371±.013	.288±.021
	3	.380±.048	.524±.034	.363±.010	.267±.020
LWS	1	.242±.042	.435±.049	.225±.038	.289±.017
	2	.239±.048	.406±.040	.207±.034	.271±.014
	3	.218±.017	.318±.064	.216±.029	.244±.023
PL-AGGD	1	.527±.035	.391±.040	.383±.014	.397±.028
	2	.439±.020	.371±.037	.372±.020	.360±.029
	3	.321±.043	.327±.028	.344±.011	.328±.023

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

5 class labels, ELIMIPL achieves an average accuracy at least 0.016 higher than DEMIPL and between 0.032 to 0.17 higher than MIPLGP. In the case of the Birdsong-MIPL dataset that comprises 13 class labels, ELIMIPL's average accuracy surpasses DEMIPL by at least 0.021 and MIPLGP by at least 0.055. The SIVAL-MIPL dataset spans 25 class labels, encompassing diverse categories such as fruits and commodities. ELIMIPL's average accuracy surpasses DEMIPL by 0.04 to 0.097 and MIPLGP by an average of 0.013. Notably, DEMIPL demonstrates relatively superior performance with fewer class labels, while MIPLGP excels in scenarios with more class labels. In contrast, ELIMIPL consistently maintains the highest average accuracy in both fewer and more class labels. This indicates that ELIMIPL exhibits superior capabilities compared to existing MIPL algorithms. PLL algorithms exhibit decent results on the MNIST-MIPL and FMNIST-MIPL datasets when $r = 1$ or $r = 2$. However, their performance significantly deteriorates when $r = 3$ or on the Birdsong-MIPL and SIVAL-MIPL datasets. This observation underscores the intrinsic complexity of MIPL problems, highlighting that they cannot be reduced to PLL problems.

The above analysis not only highlights the robustness of ELIMIPL across diverse label space but also emphasizes the limitations of addressing MIPL problems using PLL algorithms. The results underscore the importance of algorithmic designs specifically tailored to MIPL tasks.

Algorithm	C-Row	C-SBN	C-KMeans	C-SIFT
ELIMIPL	.433±.008	.509±.007	**.546±.012**	**.540±.010**
DEMIPL	.408±.010	.486±.014	.521±.012	.532±.013
MIPLGP	.432±.005	.335±.006	.329±.012	–
Mean				
PRODEN	.365±.009	.392±.008	.233±.018	.334±.029
RC	.214±.011	.242±.012	.226±.009	.209±.007
LWS	.291±.010	.310±.006	.237±.008	.270±.007
PL-AGGD	.412±.008	.480±.005	.358±.008	.363±.012
MaxMin				
PRODEN	.401±.007	.447±.011	.265±.027	.291±.011
RC	.227±.012	.338±.010	.208±.007	.246±.008
LWS	.299±.008	.382±.009	.247±.005	.230±.007
PL-AGGD	**.460±.008**	**.524±.008**	.434±.009	.285±.009

Table 3: The classification accuracies (mean±std) of ELIMIPL and comparative algorithms on the real-world datasets.

Dataset	r	ELIMIPL	MA+SP	MA+IN	MA
Birdsong	1	.771±.018	.742±.014	.746±.015	.733±.011
	2	.745±.015	.665±.024	.689±.020	.677±.017
	3	.717±.017	.592±.031	.674±.023	.652±.016
SIVAL	1	.675±.022	.618±.021	.626±.019	.620±.022
	2	.616±.025	.532±.041	.550±.040	.540±.038
	3	.600±.029	.545±.027	.521±.025	.521±.032

Table 4: The classification accuracies of the variants on the Birdsong-MIPL and SIVAL-MIPL datasets.

4.3 Results on the Real-World Datasets

Table 3 provides the results of ELIMIPL and the comparative algorithms on the CRC-MIPL dataset. The symbol – denotes cases where results could not be obtained due to memory overflow on our server. Compared to MIPL algorithms, ELIMIPL consistently achieves the highest average accuracies. Additionally, in comparison to PLL algorithms, ELIMIPL significantly outperforms them in 30 out of 32 cases.

For the CRC-MIPL dataset, both ELIMIPL and DEMIPL exhibit improved performance as the complexity of the image bag generator increases. This observation aligns with the intuition that, while avoiding overfitting, intricate feature extractors tend to produce higher classification accuracies. However, this phenomenon is not consistently observed for MIPLGP and PLL algorithms. For example, these algorithms do not consistently achieve superior results on the C-KMeans and C-SIFT datasets compared to the results on the CRC-Row or C-SBN dataset. We posit that the intricate features exceed the capability limits of these algorithms. Thus, the development of effective MIPL algorithms becomes imperative.

In most cases, the MaxMin strategy tends to yield superior outcomes than the Mean strategy. We postulate that this difference arises from the significant distinction between tissue cells and the background in the CRC-MIPL dataset. Applying the Mean strategy to features generated by simple image bag generators (i.e., Row and SBN) diminishes the distinction between tissue cells and the background, making it challenging to learn discriminative features. Conversely, for features generated by more complex image bag generators (i.e., KMeansSeg and SIFT), both the Mean and MaxMin strategies demonstrate their respective merits. Therefore, both strategies are worthy of consideration and application.

4.4 Further Analyses

Effectiveness of CLI To evaluate the impact of CLI, we modify the loss function in Equation (9) and propose three

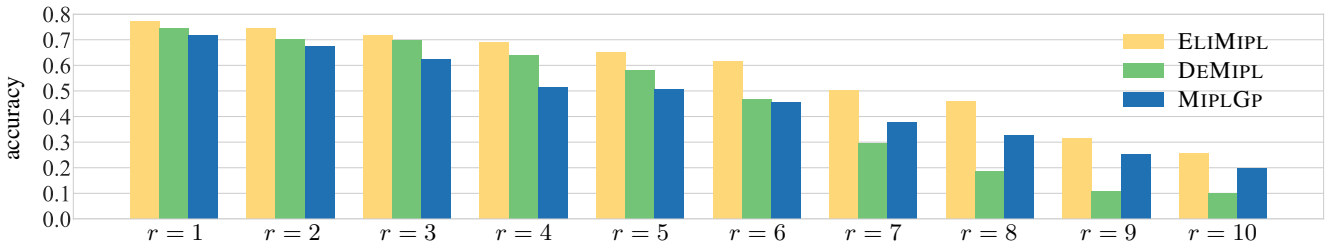


Figure 4: The classification accuracies of ELIMIPL, DEMIPL, and MIPLGP on the Birdsong-MIPL dataset with varying r.

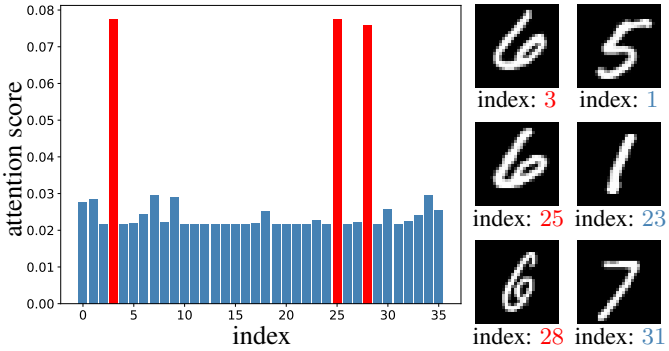


Figure 5: Attention scores for a test bag. Red and blue are the attention scores of positive and negative instances, respectively.

variants: MA+SP, MA+IN, and MA. These variants respectively represent the removal of inhibition loss, sparsity loss, and the simultaneous elimination of both inhibition and sparsity losses. Table 4 presents the experimental results conducted on the Birdsong-MIPL and SIVAL-MIPL datasets. With MA as the baseline, the introduction of individual sparse loss or inhibition loss tends to yield marginal performance improvements in most cases, while in some cases, performance degradation may occur. In contrast, ELIMIPL, using the CLI demonstrates a substantial boost in classification accuracy.

Challenging Disambiguation Scenarios We select different quantities of false positive labels from 1 to 10 on the Birdsong-MIPL dataset. Figure 4 presents the experimental results of ELIMIPL, DEMIPL, and MIPLGP with $r \in \{1, 2, \cdots, 10\}$. Particularly, ELIMIPL and DEMIPL adhere to the embedded-space paradigm, while MIPLGP follows the instance-space paradigm. Three distinct phenomena are observed: (a) ELIMIPL consistently exhibits higher average accuracy compared to DEMIPL and MIPLGP. (b) For $r < 7$, DEMIPL outperforms MIPLGP. However, when $r \geq 7$, MIPLGP surpasses DEMIPL. (c) The gaps between ELIMIPL and DEMIPL are greater when $r \in \{6, 7, 8, 9, 10\}$ than when $r \in \{1, 2, 3, 4, 5\}$. The widening gaps signify the growing significance of the scaled additive attention mechanism and CLI. Therefore, Figure 4 clearly demonstrates that ELIMIPL outperforms both MIPLGP and DEMIPL in disambiguation, even when confronted with challenging scenarios.

Comparison of CLI and Cross-Entropy Loss For a comparative analysis between the CLI loss and the cross-entropy (CE) loss, we substitute the mapping loss and the CLI loss with the CE loss, resulting in variants CE-SP-IN (which utilizes CE loss, sparsity loss, and inhibition loss) and CE (which only utilizes CE loss), respectively. Table 5 illustrates that, in all cases, accuracies obtained with the CLI loss surpass those achieved with CE-SP-IN and CE. Notably, the incorporation of inhibition loss and sparsity loss enhances the performance of the CE loss, underscoring the importance of considering the intrinsic properties of the label space and the information from non-candidate label sets.

Algorithm	r	MNIST	FMNIST	Birdsong	SIVAL
CLI loss	1	.992±.007	.903±.018	.771±.018	.675±.022
	2	.987±.010	.845±.026	.745±.015	.616±.025
	3	.748±.144	.702±.055	.717±.017	.600±.029
CE-SP-IN	1	.899±.037	.825±.035	.740±.013	.639±.030
	2	.847±.027	.679±.037	.687±.024	.587±.022
	3	.636±.112	.610±.037	.592±.036	.578±.022
CE	1	.919±.017	.709±.257	.704±.019	.587±.028
	2	.833±.016	.645±.044	.616±.032	.534±.025
	3	.628±.096	.551±.032	.459±.045	.514±.025

Table 5: The comparison between the CLI loss and the CE loss.

Interpretability of the Attention Mechanism Figure 5 displays the attention scores of a test multi-instance bag in the MNIST-MIPL dataset ($r = 1$). The bag contains positive instances represented by the digit 6, while negative instances are drawn from digits $\{1, 3, 5, 7, 9\}$. Additionally, we visualize the attention scores of all three positive instances and the negative instances. Figure 5 illustrates that ELIMIPL can accurately identify all positive instances by assigning significantly higher attention scores to them, and the attention scores can be directly utilized for interpretability.

Due to space constraints, additional experimental results and analyses are presented in the appendix.

5 Conclusion

This paper investigates a multi-instance partial-label learning algorithm that introduces a scaled additive attention mechanism and exploits conjugate label information. This information includes both candidate label information and non-candidate label information, along with the sparsity of the true label matrix. Experimental results demonstrate the superiority of our proposed ELIMIPL algorithm. The utilization of conjugate label information proves significantly more effective than relying on incomplete label information, especially in challenging disambiguation scenarios. In the future, we will explore the instance-depend multi-instance partial-label learning algorithm and conduct theoretical analyses to develop more effective algorithms.

References

Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013.

Forrest Briggs, Xiaoli Z. Fern, and Raviv Raich. Rank-loss support instance machines for MIML instance annotation. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China*, pages 534–542, 2012.

Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018.

Wei-Chen Chen, Xin-Yi Yu, and Linlin Ou. Pedestrian attribute recognition in video surveillance scenarios based on view-attribute attention localization. *Machine Intelligence Research*, 19(2):153–168, 2022.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *The Journal of Machine Learning Research*, 12:1501–1536, 2011.

Yufei Cui, Ziquan Liu, Xiangyu Liu, Xue Liu, Cong Wang, Tei-Wei Kuo, Chun Jason Xue, and Antoni B. Chan. Bayes-MIL: A new probabilistic perspective on attention-based multiple instance learning for whole slide images. In *Proceedings of the 11th International Conference on Learning Representations, Kigali, Rwanda*, pages 1–17, 2023.

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

Lei Feng and Bo An. Partial label learning with self-guided retraining. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, USA*, pages 3542–3549, 2019.

Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, and Masashi Sugiyama. Provably consistent partial-label learning. In *Advances in Neural Information Processing Systems 33, Virtual Event*, pages 10948–10960, 2020.

Xiuwen Gong, Dong Yuan, and Wei Bao. Partial label learning via label influence function. In *Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA*, pages 7665–7678, 2022.

Anne Grote, Nadine S. Schaadt, Germain Forestier, Cédric Wemmert, and Friedrich Feuerhake. Crowdsourcing of histological image labeling and object delineation by medical students. *IEEE Transactions Medical Imaging*, 38(5):1284–1294, 2019.

Shuo He, Lei Feng, Fengmao Lv, Wen Li, and Guowu Yang. Partial label learning with semantic label representations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA*, pages 545–553, 2022.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning, Stockholmsmässan, Stockholm, Sweden*, pages 2132–2141, 2018.

Humayun Irshad, Eun-Yeong Oh, Daniel Schmolze, Liza M Quintana, Laura Collins, Rulla M Tamimi, and Andrew H Beck. Crowdsourcing scoring of immunohistochemistry images: Evaluating performance of the crowd and an automated computational method. *Scientific reports*, 7(1):43286, 2017.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Ximing Li, Yuanzhi Jiang, Changchun Li, Yiyuan Wang, and Jihong Ouyang. Learning with partial labels from semi-supervised perspective. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington, DC, USA*, pages 8666–8674, 2023.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In *Proceedings of the 37th International Conference on Machine Learning, Virtual Event*, pages 6500–6510, 2020.

Gengyu Lyu, Songhe Feng, Yidong Li, Yi Jin, Guojun Dai, and Congyan Lang. HERA: Partial label learning by combining heterogeneous loss with sparse and low-rank regularization. *ACM Transactions on Intelligent Systems and Technology*, 11(3):1–19, 2020.

Oded Maron and Aparna Lakshmi Ratan. Multiple-instance learning for natural scene classification. In *Proceedings of the 15th International Conference on Machine Learning, Madison, Wisconsin, USA*, pages 341–349, 1998.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in Neural Information Processing Systems 20, Vancouver, British Columbia, Canada*, pages 1289–1296, 2007.

Xiaoshuang Shi, Fuyong Xing, Yuanpu Xie, Zizhao Zhang, Lei Cui, and Lin Yang. Loss-based attention for deep multiple instance learning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, NY, USA*, pages 5742–5749, 2020.

Shuo Tan, Lei Zhang, Xin Shu, and Zizhou Wang. A feature-wise attention module based on the difference with surrounding features for convolutional neural networks. *Frontiers of Computer Science*, 17(6):176338: 1–10, 2023.

Wei Tang, Weijia Zhang, and Min-Ling Zhang. Disambiguated attention embedding for multi-instance partial-label learning. In *Advances in Neural Information Processing Systems 36, New Orleans, LA, USA*, pages 56756–56771, 2023.

Wei Tang, Weijia Zhang, and Min-Ling Zhang. Multi-instance partial-label learning: Towards exploiting dual inexact supervision. *Science China Information Sciences*, 67(3):Article 132103: 1–14, 2024.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30, Long Beach, CA, USA*, pages 5998–6008, 2017.

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

Deng-Bao Wang, Min-Ling Zhang, and Li Li. Adaptive graph guided disambiguation for partial label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8796–8811, 2022.

Yang Wang, Jinjia Peng, Huibing Wang, and Meng Wang. Progressive learning with multi-scale attention network for cross-domain vehicle re-identification. *Science China Information Sciences*, 65(6):160103:1–15, 2022.

Xiu-Shen Wei and Zhi-Hua Zhou. An empirical study on image bag generators for multi-instance learning. *Machine Learning*, 105:155–198, 2016.

Hongwei Wen, Jingyi Cui, Hanyuan Hang, Jiabin Liu, Yisen Wang, and Zhouchen Lin. Leveraged weighted loss for partial label learning. In *Proceedings of the 38th International Conference on Machine Learning, Virtual Event*, pages 11091–11100, 2021.

John Wright and Yi Ma. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press, 2022.

Dong-Dong Wu, Deng-Bao Wang, and Min-Ling Zhang. Revisiting consistency regularization for deep partial label learning. In *Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA*, pages 24212–24225, 2022.

Jinxi Xiang, Xiyue Wang, Jun Zhang, Sen Yang, Xiao Han, and Wei Yang. Exploring low-rank property in multiple instance learning for whole slide image classification. In *Proceedings of the 11th International Conference on Learning Representations, Kigali, Rwanda*, pages 1–18, 2023.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

Yao Yao, Jiehui Deng, Xiuhua Chen, Chen Gong, Jianxin Wu, and Jian Yang. Deep discriminative CNN with temporal ensembling for ambiguously-labeled image classification. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence, New York, NY, USA*, pages 12669–12676, 2020.

Qi Zhang, Sally A. Goldman, Wei Yu, and Jason E. Fritts. Content-based image retrieval using multiple-instance learning. In *Proceedings of the 19th International Conference on Machine Learning, Sydney, Australia*, pages 682–689, 2002.

Fei Zhang, Lei Feng, Bo Han, Tongliang Liu, Gang Niu, Tao Qin, and Masashi Sugiyama. Exploiting class activation value for partial-label learning. In *Proceedings of the 10th International Conference on Learning Representations, Virtual Event*, pages 1–17, 2022.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the 35th IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA*, pages 18802–18812, 2022.

Weijia Zhang, Xuanhui Zhang, Han-Wen Deng, and Min-Ling Zhang. Multi-instance causal representation learning for instance label prediction and out-of-distribution generalization. In *Advances in Neural Information Processing Systems 35, New Orleans, LA, USA*, pages 34940–34953, 2022.

Weijia Zhang. Non-i.i.d. multi-instance learning for predicting instance and bag labels with variational auto-encoder. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence, Virtual Event / Montreal, Canada*, pages 3377–3383, 2021.

Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-i.i.d. samples. In *Proceedings of the 26th International Conference on Machine Learning, Montreal, Quebec, Canada*, pages 1249–1256, 2009.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018.

Exploiting Conjugate Label Information for Multi-Instance Partial-Label Learning (Appendix)

Wei Tang[1,2], **Weijia Zhang**[3], **Min-Ling Zhang**[1,2*]
[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Lab. of Computer Network and Information Integration (Southeast University), MoE, China
[3]School of Information and Physical Sciences, The University of Newcastle, NSW 2308, Australia
tangw@seu.edu.cn, weijia.zhang@newcastle.edu.au, zhangml@seu.edu.cn

1 Additional Experiment Results

1.1 Results of PLL algorithms with MLP

We compared with four PLL algorithms, i.e., PRODEN [Lv *et al.*, 2020], RC [Feng *et al.*, 2020], LWS [Wen *et al.*, 2021], and PL-AGGD [Wang *et al.*, 2022]. Among these, the first three algorithms can be used with either linear classifiers or multi-layer perceptrons (MLP). Due to space limitations, we presented only the results obtained using the linear classifiers in the main body of the paper. Table A1 and A2 display the results of ELIMIPL and the comparative PLL algorithms with MLP on the benchmark and CRC-MIPL datasets, respectively.

Table A1 clearly illustrates that ELIMIPL consistently outperforms the classification accuracies of the comparative PLL algorithms with MLP. ELIMIPL consistently outperforms the comparative PLL algorithms in all cases. However, when employing MLP, the comparative PLL algorithms did not consistently achieve superior outcomes when compared to themselves using linear classifiers. This is particularly evident when dealing with datasets containing relatively simple features, where MLP results in lower performance than linear classifiers. This phenomenon suggests that for the comparative PLL algorithms, linear classifiers possess sufficient capacity to handle relatively simple features, while MLP might lead to overfitting on the benchmark datasets.

Table A2 reveals that ELIMIPL significantly outperforms the comparative PLL algorithms in 20 out of 24 cases while showing inferior performance in 3 cases out of 24. Notably, PLL algorithms utilizing MLP consistently outperform those using linear classifiers across almost all cases. When replacing linear classifiers with MLP, results obtained from the KMeansSeg image bag generator exhibit a substantial improvement compared to those generated by simpler image bag generators (i.e., Row and SBN), while the improvements are less pronounced with the SIFT image bag generator. In conclusion, although the PLL algorithms can attain satisfactory results using MLP in certain scenarios, the development of dedicated MIPL algorithms is essential.

1.2 Win/tie/loss counts of Experimental Results

To ensure the reliability of the results, we perform the pairwise t-test at a significance level of 0.05. We present the win/tie/loss counts between ELIMIPL and the comparative algorithms on the benchmark datasets for varying numbers of false positive labels ($r \in \{1, 2, 3\}$), as well as the CRC-MIPL dataset, in Table A3. Several key observations emerge: (a) ELIMIPL demonstrates statistical superiority over MIPL and PLL algorithms in 67.7% and 96.9% of cases, respectively. (b) Across the benchmark datasets, ELIMIPL exhibits statistical superiority over comparative algorithms in 95.3% of cases. (c) Specifically, for the CRC-MIPL dataset, ELIMIPL

Algorithm	r	MNIST	FMNIST	Birdsong	SIVAL
ELIMIPL	1	**.992±.007**	**.903±.018**	**.771±.018**	**.675±.022**
	2	**.987±.010**	**.845±.026**	**.745±.015**	**.616±.025**
	3	**.748±.144**	**.702±.055**	**.717±.017**	**.600±.029**
Mean					
PRODEN	1	.555±.033	.652±.033	.303±.016	.303±.020
	2	.372±.038	.463±.067	.287±.017	.274±.022
	3	.285±.032	.288±.039	.278±.006	.242±.009
RC	1	.660±.031	.697±.166	.329±.014	.344±.014
	2	.577±.039	.684±.029	.301±.014	.299±.015
	3	.362±.029	.414±.050	.288±.019	.256±.013
LWS	1	.605±.030	.702±.033	.344±.018	.346±.014
	2	.431±.024	.547±.040	.310±.014	.312±.015
	3	.335±.029	.411±.033	.289±.021	.286±.018
MaxMin					
PRODEN	1	.465±.023	.358±.019	.339±.010	.322±.018
	2	.338±.031	.315±.023	.329±.016	.295±.021
	3	.260±.037	.265±.031	.305±.015	.244±.018
RC	1	.518±.022	.421±.016	.379±.014	.304±.015
	2	.462±.028	.363±.018	.359±.015	.268±.023
	3	.366±.039	.294±.053	.332±.024	.244±.014
LWS	1	.457±.028	.346±.033	.349±.013	.345±.013
	2	.351±.043	.323±.031	.336±.013	.314±.019
	3	.274±.037	.267±.034	.307±.016	.268±.019

Table A1: The classification accuracies (mean±std) of ELIMIPL and comparative PLL algorithms on the benchmark datasets with varying numbers of false positive candidate labels ($r \in \{1, 2, 3\}$).

Algorithm	Row	SBN	KMeans	SIFT
ELIMIPL	.433±.008	.509±.007	.546±.012	**.540±.010**
Mean				
PRODEN	.405±.012	.515±.010	.512±.014	.352±.015
RC	.290±.010	.394±.010	.304±.017	.248±.008
LWS	.360±.008	.440±.009	.422±.035	.338±.009
MaxMin				
PRODEN	**.453±.009**	**.529±.010**	**.563±.011**	.294±.008
RC	.347±.013	.432±.008	.366±.010	.204±.008
LWS	.381±.011	.442±.009	.335±.049	.287±.009

Table A2: The classification accuracies (mean±std) of ELIMIPL and comparative PLL algorithms on the real-world datasets.

*Corresponding author

	ELIMIPL against						In total
	DEMIPL	MIPLGP	PRODEN	RC	LWS	PL-AGGD	
$r = 1$	2/2/0	3/1/0	16/0/0	16/0/0	16/0/0	8/0/0	61/3/0
$r = 2$	3/1/0	3/1/0	16/0/0	16/0/0	16/0/0	8/0/0	62/2/0
$r = 3$	2/2/0	2/2/0	16/0/0	16/0/0	16/0/0	8/0/0	60/4/0
CRC-MIPL	4/0/0	2/1/0	11/2/3	16/0/0	16/0/0	6/0/2	55/3/5
In total	11/5/0	10/5/0	59/2/3	64/0/0	64/0/0	30/0/2	**238/12/5**

Table A3: Win/tie/loss counts on the classification performance of ELIMIPL against the comparing algorithms.

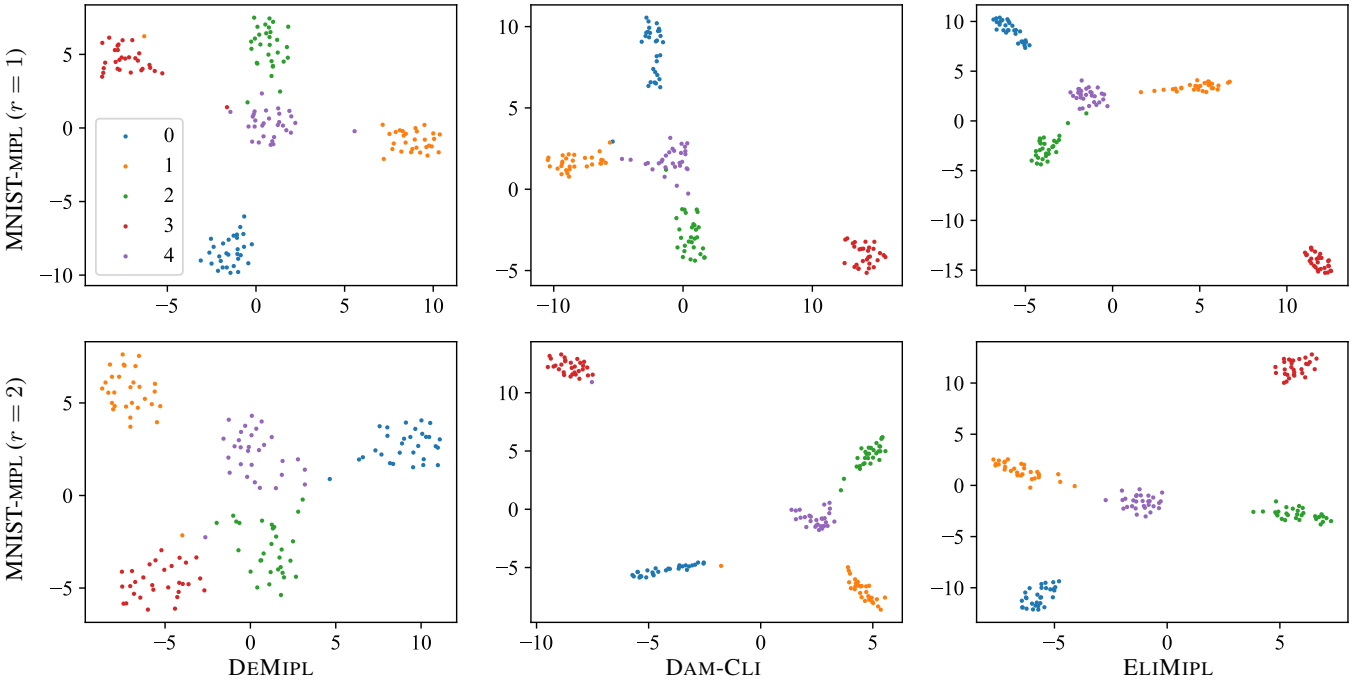


Figure A1: t-SNE visualization of aggregated bag-level feature representations produced by the attention mechanisms in DEMIPL, DAM-CLI, and ELIMIPL on the test set of the MNIST-MIPL dataset ($r \in \{1,\ 2\}$).

Algorithm	Row	SBN	KMeans	SIFT
ELIMIPL	.433±.008	.509±.007	.546±.012	.540±.010
DAM-CLI	.424±.007	.501±.008	.534±.012	.531±.010

Table A4: The classification accuracies of ELIMIPL and DAM-CLI.

shows statistical superiority over the comparative algorithms in 87.3% of cases. In summary, ELIMIPL achieves either superior or competitive performance comparative to the MIPL and PLL algorithms.

1.3 Effectiveness of the Attention Mechanism

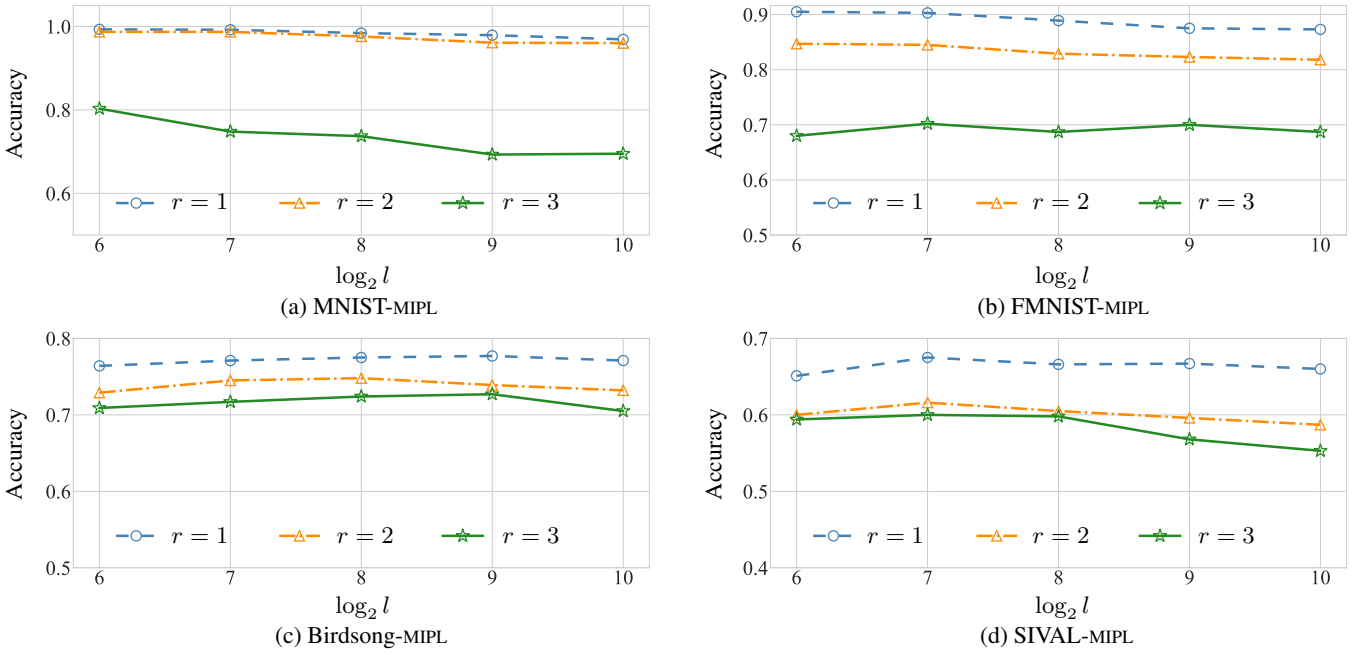
We now show the results to demonstrate the effectiveness of our scaled additive attention mechanism by contrasting it with the disambiguation attention mechanism proposed previously [Tang *et al.*, 2023]. DAM-CLI is derived by substituting the scaled additive attention mechanism in ELIMIPL with the disambiguation attention mechanism from DEMIPL. Consequently, the sole distinction between ELIMIPL and DAM-CLI lies in the utilization of different attention mechanisms.

Table A4 illustrates that ELIMIPL consistently attains higher average accuracies compared to DAM-CLI, indicating the effectiveness of the scaled additive attention mechanism.

1.4 Visualization of the Feature Representations

To delve deeper into the scaled additive attention mechanism, we employ t-SNE [Van der Maaten and Hinton, 2008] to visualize the aggregated bag-level feature representations, i.e., $\boldsymbol{z}_i$ in Eq. (4), on the test set of the MNIST-MIPL dataset when $r \in \{1, 2\}$. The t-SNE algorithm is implemented by the `sklearn.manifold` package with default parameters.

Figure A1 illustrates the feature representations generated by the attention mechanisms in DEMIPL, DAM-CLI, and ELIMIPL on the test set of the MNIST-MIPL dataset when $r \in \{1, 2\}$. Here, DAM-CLI signifies the use of the disambiguation attention mechanism in DEMIPL to replace the scaled additive attention mechanism in ELIMIPL. In Figure A1, the feature representations produced by the disambiguation attention mechanism in DEMIPL exhibit more intersections between different categories, suggesting the reduced discriminations of the representations. In contrast, the feature representations generated by the disambiguation attention mechanisms in DAM-CLI form more compact clusters than those produced by the disambiguation attention mechanisms in DEMIPL. Additionally, the feature representations generated by the scaled additive attention mechanisms in ELIMIPL exhibit increased accuracy and separability compared to those produced by the disambiguation attention mechanisms



(a) MNIST-MIPL (b) FMNIST-MIPL (c) Birdsong-MIPL (d) SIVAL-MIPL

Figure A2: The classification accuracies of ELIMIPL with varying l on the benchmark datasets.

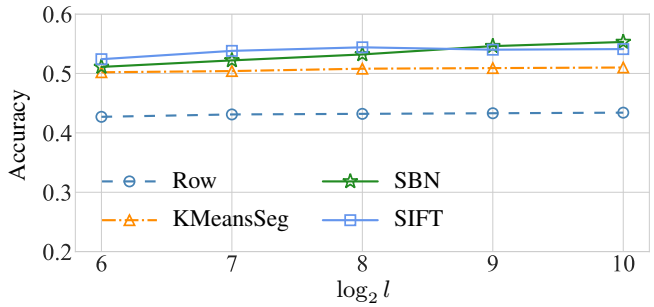


Figure A3: The classification accuracies of ELIMIPL with varying l on the CRC-MIPL dataset.

in DEMIPL and DAM-CLI.

Consequently, the proposed scaled additive attention mechanism is more effective for aggregating bag-level feature representations than the disambiguation attention mechanism proposed in DEMIPL [Tang *et al.*, 2023]. Moreover, the CLI enhances the attention mechanism's ability to aggregate more compact bag-level feature representations.

1.5 Robustness to the Embedded Space Dimension

In the instance-level feature extractor, the feature transformation network $\psi_2(\cdot)$ maps instance-level features to an embedded space of dimension l. Furthermore, the scaling factor in the scaled additive attention is $1/\sqrt{l}$. To examine the impact of parameter l on disambiguation outcomes, we vary l within the set $\{64, 128, 256, 512, 1024\}$. Especially, for each dataset, all experiments maintain consistent data partitioning and other parameters except for varying l.

Figures A2 and A3 depict the classification accuracies of ELIMIPL on the benchmark and CRC-MIPL datasets with varying l, respectively. On the benchmark dataset, ELIMIPL's performance demonstrates insensitivity to dimension l for $r = 1$ or 2. However, when $r = 3$, some variations in classification accuracy emerge for the MNIST-MIPL and SIVAL-MIPL datasets. Specifically, an increase in l correlates with a decrease in classification accuracy. This phenomenon can be attributed to the relatively straightforward features of the benchmark datasets, resulting in an undue emphasis on their feature representation when projected into a higher-dimensional embedding space. This effect becomes particularly pronounced under challenging disambiguation conditions, i.e., $r = 3$. Overall, ELIMIPL achieves improved classification accuracies on the benchmark datasets when employing smaller values of dimension l. On the CRC-MIPL dataset, ELIMIPL's classification accuracies remain stable across variations in l. Notably, when utilizing the Row and KMeansSeg image bag generators, ELIMIPL demonstrates strong robustness to the varying dimension l.

Based on the insights derived from the above analysis, we opt for uniform settings of $l = 128$ for the benchmark datasets and $l = 512$ for the CRC-MIPL dataset, corresponding to $\log_2 l = 7$ and $\log_2 l = 9$, respectively. From Figures A2 and A3, it is evident that such parameter configurations of dimension l can yield commendable results.

1.6 Effectiveness of the Scaling Factor

The scaling factor prevents the softmax function from entering regions with small gradients, thereby mitigating the issue of gradient vanishing. To assess the impact of the scaling factor, we introduce a variant called ELIMIPL *wo* $\frac{1}{\sqrt{l}}$, differing from ELIMIPL only by excluding the scaling factor.

Table A5 displays the classification accuracies of ELIMIPL *wo* $\frac{1}{\sqrt{l}}$ and ELIMIPL on the benchmark datasets. The results

Algorithm	r	MNIST	FMNIST	Birdsong	SIVAL
ELIMIPL	1	.992±.007	.903±.018	.771±.018	.675±.022
	2	.987±.010	.845±.026	.745±.015	.616±.025
	3	.748±.144	.702±.055	.717±.017	.600±.029
ELIMIPL *wo* $\frac{1}{\sqrt{l}}$	1	.200±.000	.275±.080	.138±.019	.143±.030
	2	.211±.032	.229±.049	.138±.020	.139±.029
	3	–	.200±.000	.129±.024	.131±.022

Table A5: The classification accuracies (mean±std) of ELIMIPL and comparative algorithms on the benchmark datasets with varying numbers of false positive candidate labels ($r \in \{1, 2, 3\}$). The symbol "–" indicates that ELIMIPL *wo* $\frac{1}{\sqrt{l}}$ fails to achieve accuracy due to gradient vanishing.

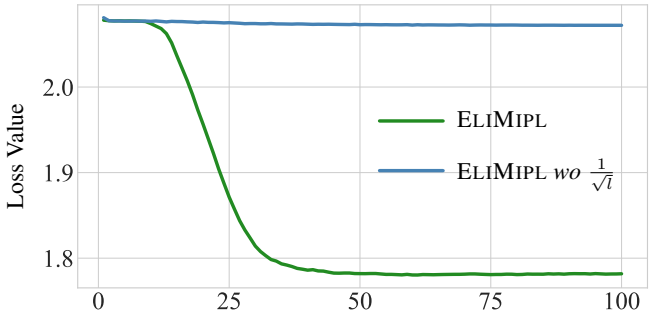


Figure A4: The loss values of ELIMIPL and ELIMIPL *wo* $\frac{1}{\sqrt{l}}$ on the MNIST-MIPL dataset ($r = 1$).

highlight that without the scaling factor, the model's performance resembles random classification, demonstrating ineffective learning and susceptibility to gradient vanishing. Furthermore, Figure A4 illustrates that without the scaling factor, the loss value fails to converge. Therefore, in ELIMIPL, the scaling factor plays a pivotal role in achieving convergence.

1.7 Parameter Sensitivity

Figure A5 illustrates the classification accuracies of ELIMIPL on the FMNIST-MIPL dataset across varying parameters μ and γ. Specifically, μ and γ are chosen from the sets $\{0.7, 0.8, 0.9, 1.0, 1.1\}$ and $\{0.4, 0.5, 0.6, 0.7, 0.8\}$, respectively. ELIMIPL exhibits robustness to various combinations of parameters μ and γ. The accuracies remain stable in most cases, even when $r = 3$. In our experiments, we set $\mu = 1$ and $\gamma = 0.5$ for ELIMIPL on the FMNIST-MIPL dataset. The results of these experiments validate the efficacy of such parameter configurations.

1.8 Computational Complexity

Table A6 presents the floating-point operations (FLOPs), number of parameters (Params), peak GPU memory usage (PM), average time per test multi-instance bag (Times), and average accuracy (Acc) over 10 trials, providing comprehensive metrics to assess model complexity. The complexity of MIPLGP is denoted as $\mathcal{O}((k+1)n^2)$, where k and n represent the number of classes and instances, respectively.

As observed in Table A6, the computational complexities of ELIMIPL and DEMIPL are comparable, yet ELIMIPL exhibits higher accuracy than DEMIPL. Moreover, ELIMIPL achieves this superior accuracy with a lighter computational burden compared to MIPLGP. This suggests that, while keeping the computational cost comparable to that of DEMIPL, ELIMIPL attains superior accuracy, outperforming MIPLGP in terms of both accuracy and computational complexity.

Algorithm	FLOPs (M)	Params (M)	PM (MiB)	Times (s)	Acc
ELIMIPL	109.86	0.43	1824	1.554	.992
DEMIPL	109.86	0.43	1822	1.426	.976
MIPLGP	–	–	12938	1.187	.949

Table A6: The outcomes on the MNIST-MIPL dataset ($r = 1$).

2 Why ELIMIPL Works?

The experimental results presented in the main body of the paper and the supplementary material demonstrate that ELIMIPL outperforms the comparative MIPL and PLL algorithms across the majority of scenarios. Furthermore, we conduct a thorough validation of the efficacy of each component within ELIMIPL. In this section, we provide insights into the key factors contributing to the success of ELIMIPL.

Scaled Additive Attention Mechanism The t-SNE visualization in Figure A1 reveals that our proposed scaled additive attention mechanism generates feature representations that are not only more compact but also more accurate compared to the disambiguation attention mechanism in DEMIPL [Tang *et al.*, 2023]. The results presented in Table A4 further affirm that the utilization of the scaled additive attention mechanism leads to higher classification accuracy than employing the disambiguation attention mechanism. Additionally, as illustrated in Table A5 and Figure A4, the inclusion of the scaling factor in the scaled additive attention mechanism ensures the model's convergence with satisfactory accuracy. Without the scaling factor, the model fails to converge, resulting in classification outputs resembling random guesses.

CLI loss During training, ELIMIPL learns conjugate label information by minimizing the CLI loss, comprising mapping loss, sparse loss, and inhibition loss. Figure 2 in the main body of the paper illustrates that using CLI loss enhances the predicted probabilities of the classifier on true labels while suppressing probabilities on non-candidate labels. Tables 4 and 5 in the main body of the paper demonstrate that the CLI loss significantly improves the model's performance compared to using mapping loss, mapping loss with sparse loss, mapping loss with inhibition loss, and cross-entropy loss. Thus, CLI loss effectively exploits the information from the label sets and the intrinsic properties of the label space, enhancing the model's disambiguation performance.

In summary, the effectiveness of ELIMIPL can be attributed to two pivotal components: (a) the scaled additive attention mechanism, which is responsible for generating discriminative bag-level feature representations, and (b) CLI loss, which is proficient in exploiting the information from the label sets and the intrinsic properties of the label space.

3 MIPL Datasets

We employ four benchmark datasets and one real-world dataset, all of which are publicly accessible [Tang *et al.*, 2024, 2023]. Next, we will provide detailed descriptions of the data preparation procedures for each of these datasets.

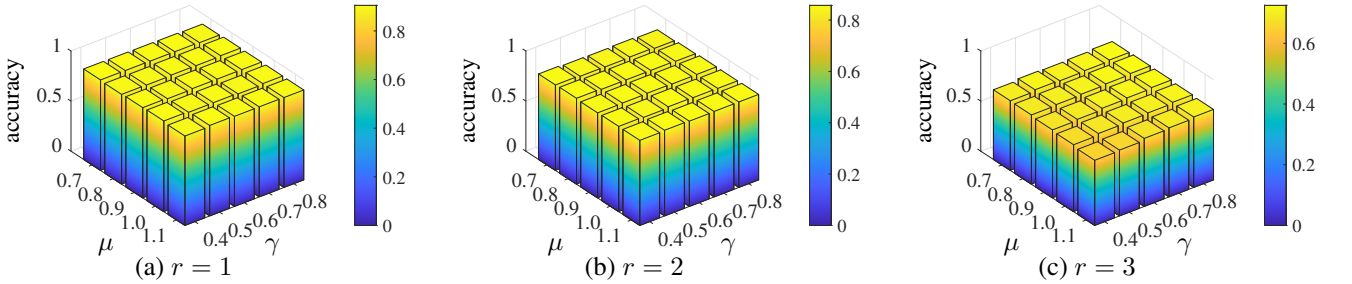


Figure A5: The performance of ELIMIPL with varying μ and γ on the FMNIST-MIPL dataset ($r \in \{1, 2, 3\}$).

3.1 Benchmark MIPL Datasets

The MNIST-MIPL and FMNIST-MIPL datasets are adaptations of the original MNIST and Fashion-MNIST datasets [LeCun *et al.*, 1998; Xiao *et al.*, 2017], respectively. To construct the MNIST-MIPL and FMNIST-MIPL datasets, positive and negative instances within each multi-instance bag are selected from targeted and reserved class labels, respectively. For the MNIST-MIPL dataset, the classes $\{0, 2, 4, 6, 8\}$ are designated as the targeted classes, ensuring the presence of positive instances corresponding to these classes. Conversely, negative instances are drawn randomly from the reserved classes $\{1, 3, 5, 7, 9\}$. Similarly, the FMNIST-MIPL dataset is constructed with the targeted class labels {*T-shirt, Trouser, Coat, Sneaker, Bag*} and the reserved class labels {*Pullover, Dress, Sandal, Shirt, Ankle boot*}.

The Birdsong dataset is widely employed in both multi-instance multi-label learning [Briggs *et al.*, 2012] and PLL [Lv *et al.*, 2020]. This dataset comprises 548 multi-instance bags, collectively containing 10232 instances. Each instance is represented by a 38-dimensional feature vector and corresponds to a single label, which is either one of the 13 specific target classes or a singular negative class. In the Birdsong-MIPL dataset, the 13 targeted classes are utilized to select positive instances, while the negative class serves as the reserved label encompassing the negative instances.

The SIVAL is a MIL benchmark dataset for content-based image retrieval with 1500 images [Settles *et al.*, 2007]. Each image serves as a multi-instance bag containing either 31 or 32 instances, linked to one of 25 distinct class labels. Each instance is characterized by a feature vector in a 30-dimensional space. To create the SIVAL-MIPL dataset from the SIVAL dataset, the arrangement of multi-instance bags remains unchanged. Every candidate label set is generated by retaining the true label and randomly choosing r false positive labels from the remaining 24 classes.

3.2 Real-World MIPL Datasets

The CRC-MIPL dataset consists of 7000 images used for classifying colorectal cancer in the absence of exact labels. These images are uniformly selected from the 7 classes of the NCT-CRC-HE-100K dataset [Kather *et al.*, 2019]. To form a candidate label set for each image, an expert trains three crowdsourced workers before annotation. The final candidate label set is obtained by aggregating the candidate labels from all three workers. The methodology is elaborated as follows: Firstly, workers assign candidate labels with non-zero probabilities, thereby creating a label set per image. Higher probabilities indicate a greater likelihood of being the true label, whereas zero probabilities indicate non-candidate labels. Secondly, the aggregated candidate label set is derived from the three label sets, which includes labels present in two or three sets. If the aggregated set contains only one or no label, the labels with the highest probabilities in each set are selected. In contrast to requiring expert annotation of true labels for each image, this annotation approach effectively reduces the expert workload while achieving satisfactory outcomes.

4 The Image Bag Generators

To learn multi-instance features on the CRC-MIPL dataset, we employ four image bag generators [Wei and Zhou, 2016]:

Row Generator [Maron and Ratan, 1998]: This approach treats each row within the image as an independent instance. For feature extraction, it calculates the average RGB color value of each row and analyzes the color differences with adjacent rows. The resulting instance feature encompasses the RGB values of the current instance, along with the disparities in RGB values between the current instance and the preceding one, as well as the subsequent one. This procedure yields a 9-dimensional feature representation for each instance.

SBN Generator [Maron and Ratan, 1998]: This approach utilizes five 2×2 blobs within the image to generate an instance-level feature. This feature includes RGB color values of the central blob and its four neighboring blobs. Instances are generated by iteratively shifting one pixel at a time, while the SBN generator omits feature information at the image's four corners. This results in a 15-dimensional feature vector for each instance.

KMeansSeg Generator [Zhang *et al.*, 2002]: This generator partitions the image into k segments, producing 6-dimensional features for each segment. The initial three dimensions represent color values within the YCbCr color space, while the subsequent three dimensions derived through wavelet transformation of the luminance Y component.

SIFT Generator [Lowe, 2004]: Using the scale-invariant feature transform (SIFT) algorithm, the SIFT generator divides instances into multiple 4×4 subregions and maps gradients of pixels within these subregions to 8 bins. As a result, SIFT generates a 128-dimensional feature for each instance.

References

Forrest Briggs, Xiaoli Z. Fern, and Raviv Raich. Rank-loss support instance machines for MIML instance annotation.

In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China*, pages 534–542, 2012.

Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, and Masashi Sugiyama. Provably consistent partial-label learning. In *Advances in Neural Information Processing Systems 33, Virtual Event*, pages 10948–10960, 2020.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Medicine*, 16(1):e1002730:1–22, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In *Proceedings of the 37th International Conference on Machine Learning, Virtual Event*, pages 6500–6510, 2020.

Oded Maron and Aparna Lakshmi Ratan. Multiple-instance learning for natural scene classification. In *Proceedings of the 15th International Conference on Machine Learning, Madison, Wisconsin, USA*, pages 341–349, 1998.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In *Advances in Neural Information Processing Systems 20, Vancouver, British Columbia, Canada*, pages 1289–1296, 2007.

Wei Tang, Weijia Zhang, and Min-Ling Zhang. Disambiguated attention embedding for multi-instance partial-label learning. In *Advances in Neural Information Processing Systems 36, New Orleans, LA, USA*, pages 56756–56771, 2023.

Wei Tang, Weijia Zhang, and Min-Ling Zhang. Multi-instance partial-label learning: Towards exploiting dual inexact supervision. *Science China Information Sciences*, 67(3):Article 132103: 1–14, 2024.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

Deng-Bao Wang, Min-Ling Zhang, and Li Li. Adaptive graph guided disambiguation for partial label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8796–8811, 2022.

Xiu-Shen Wei and Zhi-Hua Zhou. An empirical study on image bag generators for multi-instance learning. *Machine Learning*, 105:155–198, 2016.

Hongwei Wen, Jingyi Cui, Hanyuan Hang, Jiabin Liu, Yisen Wang, and Zhouchen Lin. Leveraged weighted loss for partial label learning. In *Proceedings of the 38th International Conference on Machine Learning, Virtual Event*, pages 11091–11100, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

Qi Zhang, Sally A. Goldman, Wei Yu, and Jason E. Fritts. Content-based image retrieval using multiple-instance learning. In *Proceedings of the 19th International Conference on Machine Learning, Sydney, Australia*, pages 682–689, 2002.