Multi-Instance Multi-Label Classification from Crowdsourced Labels

Ziquan Wang[1,2], Mingxuan Xia[1,2], Xiangyu Ren[3], Jiaqing Zhou[4], Gengyu Lyu[5], Tianlei Hu[1,2], Haobo Wang[1,2]*

[1]School of Software Technology, Zhejiang University
[2] Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[3]College of Computer and Control Engineering, Northeast Forestry University
[4] ByteDance, Hangzhou
[5] College of Computer Science, Beijing University of Technology
zqwang@buaa.edu.cn, ren_xy@nefu.edu.cn, xiamingxuan@zju.edu.cn,
jiashu@bytedance.com, lyugengyu@gmail.com, htl@zju.edu.cn, wanghaobo@zju.edu.cn

Abstract

Multi-instance multi-label classification (MIML) is a fundamental task in machine learning, where each data sample comprises a bag containing several instances and multiple binary labels. Despite its wide applications, the data collection process involves matching multiple instances and labels, typically resulting in high annotation costs. In this paper, we study a novel yet practical crowdsourced multi-instance multi-label classification (CMIML) setup, where labels are collected from multiple crowd sources. To address this problem, we first propose a novel data generation process for CMIML, i.e., ***cross-label transition***, where cross-label annotation error is more likely to appear rather than previous single-label transition assumption, due to the inherent similarity of localized instances from different classes. Then, we formally define the cross-label transition by cross-label transition matrices which are dependent across classes. Subsequently, we establish the first unbiased risk estimator for CMIML and further improve it through aggregation techniques, along with a rigorous generalization error bound. We also provide a practical implementation of cross-label transition matrix estimation. Comprehensive experiments on six benchmark datasets under various scenarios demonstrate that our algorithm outperforms the baselines by a large margin, validating its effectiveness in handling the CMIML problem.

1 Introduction

Multi-instance multi-label classification (MIML) is a fundamental framework for modeling complex real-world data, where each sample is represented as a bag of instances and annotated with multiple labels (Feng and Zhou 2017; Zhou and Zhang 2006). Specifically, the bag-level label is positive if it contains at least one positive instance; otherwise, it is negative. MIML has diverse applications, including text classification (Zhou, Sun, and Li 2009), acoustic classification (Briggs et al. 2012), and gene function prediction (Wu, Huang, and Zhou 2014). Diverse methods tackle MIML challenges, including SVM (Zhou and Zhang 2006), convex optimization techniques (Li et al. 2012), neural networks (Feng and Zhou 2017), and attention-based approaches (Yang, Tang, and Min 2022). However, these methods rely on precise labels for optimal performance.

Since each bag may contain complex patterns among instances, acquiring precise annotation for MIML is both labor-intensive and costly (Nguyen and Raich 2021). When dealing with a large number of classes, annotators may overlook some labels due to limited energy or ability, leading to annotation errors. For instance, in MIML bird acoustic classification tasks, annotators must listen to all instances in each audio bag to detect the presence of specific birds (Briggs et al. 2012). The presence of numerous specific bird species requires repeated checks, increasing the likelihood of errors due to the labor-intensive work. To address this, researchers have explored weakly-supervised approaches for MIML, including semi-supervised MIML (Yang et al. 2019; Xu et al. 2012) and incomplete label MIML classification (Nguyen and Raich 2021). However, in complex and specialized tasks like protein function prediction (Wu, Huang, and Zhou 2014), these semi-supervised setups with missing labels or incomplete annotations may cause performance degradation.

In this work, we address that crowdsourcing (Bragg, Mausam, and Weld 2013; Zhang and Wu 2018), famous for its efficiency and cost-effectiveness, can be a feasible alternative for a full annotation on MIML instead of expensive expertise annotations. However, most existing crowdsourcing methods focus on scenarios where a single instance is associated with a single label (Rodrigues and Pereira 2018; Wei et al. 2023; Gao et al. 2022). Although some studies have addressed crowdsourcing with multiple labels (Rodrigues and Pereira 2018; Wei et al. 2023; Zhang and Wu 2018), the challenge of learning a robust classifier from an uncertain bag of instances remains unsolved, and these approaches also lack theoretical foundations.

To this end, we introduce a novel crowdsourced multi-instance multi-label classification (CMIML) framework and propose a theoretical-driven method that learns a robust MIML classifier end-to-end. Specifically, we excavate the pattern of annotation error in CMIML and propose a novel data generation process called ***cross-label transition***, where the confusion is more likely to appear between classes due to the inherent similarity of localized instances from different classes. By defining the cross-label transition matrix, we ***establish the first unbiased risk estimator (URE)*** for the

*Corresponding author.

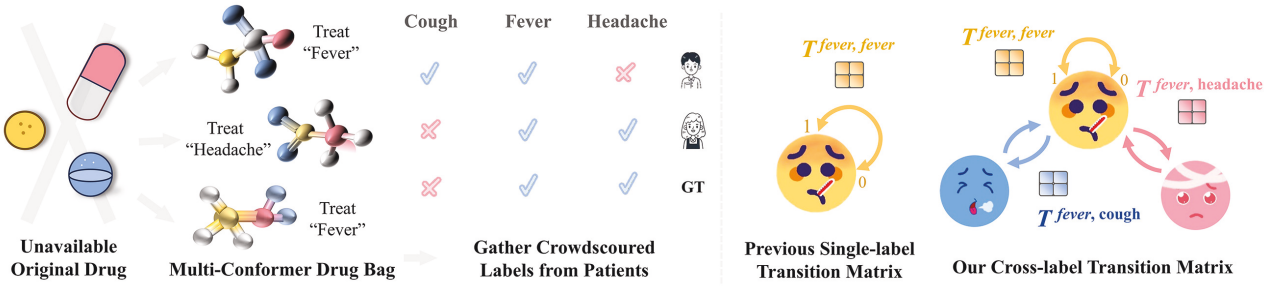

Figure 1: Drug activity prediction (left) and different transition matrix strategies in multi-label classification (right). The multiple conformers of drug molecules form a "multi-conformer bag", with each having different therapeutic effects. Then, crowdsourced labels from patients are collected. Due to individual differences, the efficacy of patient symptoms may be expressed in different correlation patterns, which can easily lead to cross-label errors.

CMIML problem and subsequently derive aggregated versions from the perspective of bag-level and instance-level, which ensures fast and realistic computation. Then, we devise a practical solution for cross-label transition matrix estimation. A generalization error bound for our UREs is also provided. Experimentally, we validate the performance of our proposed method on three datasets including image classification, text classification, and protein function prediction. Compared to other MIML and weakly-supervised multi-label classification methods, our method demonstrates excellent performance in all three scenarios.

2 Methodology

MIML Classification. In MIML, let $\mathcal{X}$ denote the input space and $\mathcal{Y} = \{0,1\}^C$ the label space. The MIML dataset $\mathcal{D}$ consists of B bag-label pairs $(\mathbf{x}, \mathbf{y})$. Specifically, the i-th bag contains a set of instances $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_i}\}$, where n_i is the number of instances in the i-th bag, and each instance $\boldsymbol{x}_j \in \mathbb{R}^d$ has an unknown instance-level label $\boldsymbol{y}_j \in \{0,1\}^C$. The k-th class bag-level label y_k is set to 1 if any instance in the bag is relevant to the k-th label; otherwise, y_k is set to 0. We can express this rule in the formula $\mathsf{y}_k = 1 - \prod_{j=1}^{n_i}(1 - y_{j,k})$ (Maron and Lozano-Pérez 1997), where $y_{j,k}$ is the k-th class label for the j-th instance. The goal of MIML is to learn a bag-level classifier $f : \mathcal{X} \to \mathcal{Y}$ obtained by minimizing the following expected risk (Feng et al. 2021, 2023):

$$R(f) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P(\boldsymbol{X},\boldsymbol{Y})}[\mathcal{L}(f(\mathbf{x}), \mathbf{y})] \tag{1}$$

Where $\mathcal{L} : \mathbb{R}\times\mathcal{Y} \to \mathbb{R}^+$ is a MIML loss function, $\boldsymbol{X}$ ($\boldsymbol{Y}$) is the random variable on $\mathcal{X}$ ($\mathcal{Y}$), and $P(\boldsymbol{X}, \boldsymbol{Y})$ is the joint probability distribution. Note that a method guarantees ***risk consistency*** if an unbiased risk estimator is implemented, namely the risk estimator is equivalent to $R(f)$ given the same classifier f (Mohri, Rostamizadeh, and Talwalkar 2012).

Crowdsourced MIML Classification. In this paper, we address the crowdsourced MIML classification problem where the true label $\mathbf{y}$ is not accessible. Instead, weak labels are obtained from multiple crowdsourced annotators for each bag, resulting in a training dataset $\tilde{\mathcal{D}}$ of B bag-label pairs $(\mathbf{x}, \{\tilde{\mathbf{y}}^{(m)}\}_{m=1}^{M})$, where M is the number of annotators. Our goal is to establish an unbiased risk estimator $\tilde{R}(f)$ relative to $R(f)$ based on $\tilde{\mathcal{D}}$. For clarity, we list all the notations and their meanings in Table 1.

2.1 Cross-Label Transition

Annotation errors are common in crowdsourcing, impacting the robustness of learned models. In this subsection, we explore the pattern of annotation errors in CMIML, highlighting that beyond a lack of expertise in specific classes, confusion between instances of different classes is more prevalent in CMIML. To address this, we introduce a novel data generation process called cross-label transition, defined by cross-label transition matrices. This approach allows us to propose the first unbiased risk estimator for CMIML, which ensures risk consistency in addressing this challenge.

Cross-Label Annotation Error. Unlike previous single-label transition strategies (Song et al. 2022; Patrini et al. 2017; Xia et al. 2024), annotation errors in multi-label classification often manifest as cross-label patterns (Carbonneau et al. 2018). In Figure 1, we present an example of drug activity prediction. The input bags consist of multiple conformers of a drug molecule, with no additional information available beyond the conformers. After evaluating therapeutic effects from patients, the crowdsourced efficacy labels are collected. However, due to individual differences, a patient may experience multiple correlated diseases, leading to cross-label annotation errors. These cross-label correlations were neglected by previous work.

Cross-Label Transition Matrix. To deal with the frequent cross-label annotation error in CMIML, we define ***cross-label transition*** by cross-label transition matrices:

$$\boldsymbol{T}^{(p,q,m)} = (T_{i,j}^{(p,q,m)})_{2\times 2} \in [0,1]^{2\times 2} \tag{2}$$

where $m \in \{1,2,...,M\}$, $p, q \in \{1,2,...,C\}$ and $i, j \in \{0,1\}$. $T_{i,j}^{(p,q,m)}$ denotes the probability of the m-th annotator mistakenly labeling the p-th label i as the q-th label j. Assuming that the label transition process is independent of the samples (Tang, Zhang, and Zhang 2024; Cheng et al. 2022) , the transition probability can be derived as $T_{i,j}^{(p,q,m)} = P(\tilde{Y}_q^{(m)} = j | Y_p = i, \boldsymbol{X} = \mathbf{x}) = P(\tilde{Y}_q^{(m)} = j | Y_p = i)$.

The main advantage of our proposed cross-label transition lies in the following three aspects: *i)* Compared to the

previous single-label transition assumption (Li et al. 2022) in multi-label classification, which is dependent on every single class, cross-label transition matrices are defined with higher dimensions which are dependent across classes, enabling handling cross-label annotation errors in CMIML. *ii)* Armed with broader assumptions, the cross-label transition is more general and can also handle the single-label transition challenge since the latter is a subset of the former (see the right portion of Figure 1). *iii)* Modeling cross-label transition matrices can explore label relationships, which is also important in solving multi-label classification problems (Xu et al. 2022; Min et al. 2023; Liu et al. 2023).

Data Generation Process. With transition matrices, we express the data generation process in the following formula:

$$P(\tilde{Y}_k^{(m)} = j|\boldsymbol{X} = \mathbf{x}) = \sum_{t=1}^{C}\sum_{i=0}^{1} T_{i,j}^{(t,k,m)} P(Y_t = i|\boldsymbol{X} = \mathbf{x}) \quad (3)$$

Obviously, when given the original label, the transition probabilities satisfy the following normalization: $\sum_{q=1}^{C}\sum_{j=0}^{1} T_{i,j}^{(p,q,m)} = 1\ \forall p, m, i$.

Unbiased Risk Estimator. After we define the data generation process of the cross-label transition, we theoretically establish the first unbiased risk estimator for the CMIML problem. The theorem proposed below guarantees risk consistency when solving CMIML. All related proofs in this paper are provided in the supplementary materials.

Theorem 1 (The First URE)**.** *Let* $\boldsymbol{G}^{(k)} \in \mathbb{R}^{2^M\times 2^M}$ *or* $\boldsymbol{H}^{(k)}(\mathbf{x}) \in \mathbb{R}^{2^M\times 1}$ *be the matrix or vector all composed of the element* $\prod_{m=1}^{M}\sum_{t=1}^{C} T_{\cdot,\cdot,m}^{(t,k,m)}$*. Then,* $R(f)$ *can be equivalently expressed as* $\tilde{R}(f) = \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{y}})\sim P(\boldsymbol{X},\tilde{\boldsymbol{Y}})}(\tilde{L}(f(\mathbf{x}), \{\tilde{\mathbf{y}}^{(m)}\}_{m=1}^{M}))$*, where* $\tilde{L}(f(\mathbf{x}), \{\tilde{\mathbf{y}}^{(m)}\}_{m=1}^{M}) = \sum_{k=1}^{C}\tilde{l}(f_k(\mathbf{x}), \{\tilde{y}_k^{(m)}\}_{m=1}^{M})$ *and*

$$\forall\, k, i\ \ \tilde{l}(f_k(\mathbf{x}), \cdot) = det(\boldsymbol{G}^{(k)}|_{\boldsymbol{G}_{\cdot i}^{(k)} = \boldsymbol{H}^{(k)}(\mathbf{x})})[det(\boldsymbol{G}^{(k)})]^{-1}$$

$\boldsymbol{G}_{\cdot i}^{(k)} = \boldsymbol{H}^{(k)}(\mathbf{x})$ *means replacing the i-th column with* $\boldsymbol{H}^{(k)}(\mathbf{x})$ *and* f_k *is the* k*-th class prediction by classifier* f*.*

For readability and space, Theorem 1 simplifies complex expressions. The full one is in the supplementary materials.

Remark 1. While instance similarity and cross-label annotation errors in CMIML present significant challenges, Theorem 1 shows that determinant calculations using the cross-label transition matrix can implicitly restore normal learning. Despite its merits, Theorem 1 may suffer from high time costs and accumulated errors in implementation. On the one hand, for large M, calculating the determinant of the 2^M-order matrix $G^{(k)}$ is extremely time-consuming. On the other hand, the product $\prod_{m=1}^{M}\sum_{t=1}^{C} T_{\cdot,\cdot,m}^{(t,k,m)}$ across M transition matrices can lead to unreliable numerical results. Both issues stem from the number of annotators M, and we focus on reducing its influence in what follows.

2.2 Improved Version of URE

To make our cross-label transition-based unbiased learning practically applicable, in this subsection, we proposed two improved versions of URE for CMIML, namely, the bag-level URE and the instance-level URE, that greatly reduce the computation cost overhead by aggregation techniques. Inspired by crowdsourcing methods (Ma et al. 2015; Li, Rubinstein, and Cohn 2019) that aggregate labels from different annotators, we aggregate both labels and transition matrices below, which eliminate the influence of M.

Notation	Meaning
C / B / M	the number of classes / bags / annotators
n_i	the number of instances in the i-th bag
$\mathcal{X}$ / $\mathcal{Y}$	feature space / label space
$\boldsymbol{X}$ / $\boldsymbol{Y}$ **(capital)**	**random variable** on $\mathcal{X}$ / $\mathcal{Y}$
$\boldsymbol{x}_j$ **(italic)**	the j-th **instance** in the bag $\mathbf{x}$
$\boldsymbol{y}_j$ **(italic)**	the label for the j-th **instance** in the bag $\mathbf{x}$
$\mathbf{x}$ **(normal)**	the multi-instance **bag** $\{\boldsymbol{x}_j\}_{j=1}^{n_i}$
$\mathbf{y}$ **(normal)**	the true label for the **bag** $\mathbf{x}$
$\tilde{\mathbf{y}}^{(m)}$ **(tilde)**	the **crowdsourced** label for the m-th annotation of the bag $\mathbf{x}$
$\tilde{\mathcal{D}}$ **(tilde)**	**crowdsourced** MIML training dataset

Table 1: Notations in this paper.

Definition 1 (Aggregation)**.** *Let* A *denote the random variable representing the annotator index. The allocation of annotators is generally independent of the true label, i.e.,* $P(A = m|\boldsymbol{X} = \mathbf{x}, \boldsymbol{Y} = \mathbf{y}) = P(A = m|\boldsymbol{X} = \mathbf{x})$*. Then,*

The ***Aggregated Label*** $\bar{\mathbf{y}} = (\bar{y}_1, \bar{y}_2, ..., \bar{y}_C)$ *is defined as* $\bar{y}_i = \sum_{m=1}^{M} P(A = m|\boldsymbol{X} = \mathbf{x})\tilde{y}_i^{(m)}$ *for* $i = 1, 2, \ldots, C$*.*

The ***Aggregated Transition Matrix*** $\boldsymbol{T}^{(p,q)} = (T_{i,j}^{(p,q)})_{2\times 2}$ *is defined as* $T_{i,j}^{(p,q)} = P(\tilde{Y}_q = j|Y_p = i, \boldsymbol{X} = \mathbf{x}) = \sum_{m=1}^{M} T_{i,j}^{(p,q,m)} P(A = m|\boldsymbol{X} = \mathbf{x})$*, where* $\tilde{Y}_k$ *is the random variable of the aggregated label* $\bar{y}_k$*.*

Note that even with the elimination of M, the information from multi-source annotation is still preserved. Moreover, this aggregation method, similar to ensemble learning, mitigates high variance from cross-label errors through averaging and enhances overall reliability. With the aggregation defined above, we derive an improved version of the unbiased risk estimator, denoted as the bag level URE.

Theorem 2 (Bag-level URE)**.** *Let* $\bar{\mathbf{y}} = (\bar{y}_1, \bar{y}_2, ..., \bar{y}_C)$ *be the aggregated label for a bag* $\mathbf{x}$*,* l *be the base loss* $\mathcal{L}(f(\mathbf{x}), \mathbf{y}) = \sum_{k=1}^{C} l(f_k(\mathbf{x}), \mathbf{y})$ *and* $\phi = \mathcal{A}^{-1}\cdot\mathcal{B} = (\phi_1, \phi_2, ..., \phi_C)^T$*, where* $\mathcal{A} = (T_{1,1}^{(i,j)} - T_{0,1}^{(i,j)})_{C\times C}$ *and* $\mathcal{B} = (P(\tilde{Y}_j = 1|\boldsymbol{X} = \mathbf{x}) - \sum_{t=1}^{C} T_{0,1}^{(t,j)})_{C\times 1}$*. Then, the unbiased risk estimator with respect to* $R(f)$ *is* $\tilde{R}(f) = \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{y}})\sim P(\boldsymbol{X},\tilde{\boldsymbol{Y}})}(\tilde{L}_b(f(\mathbf{x}), \bar{\mathbf{y}}))$*, where* $\tilde{L}_b(f(\mathbf{x}), \bar{\mathbf{y}}) = \sum_{k=1}^{C}\tilde{l}_b(f_k(\mathbf{x}), \bar{y}_k)$*,*

$$\begin{aligned}\tilde{l}_b(f_k(\mathbf{x}), \bar{y}_k) = \frac{1}{2P(\tilde{Y}_k = \bar{y}_k|\boldsymbol{X} = \mathbf{x})}[&\phi_k l(f_k(\mathbf{x}), 1)\\ &+ (1 - \phi_k) l(f_k(\mathbf{x}), 0)]\end{aligned} \quad (4)$$

Remark 2. Theorem 2, the key URE in this paper, is more practical to implement than Theorem 1. While Theorem 1 requires computing a 2^M-order matrix with time complexity of $O(2^{3M})$ for each sample, Theorem 2 reduces the time complexity to that of standard cross-entropy loss by pre-

computing a vector $\boldsymbol{\phi}$ of only C dimensions using the cross-label transition matrices and the noisy posterior probabilities, both of which can be effectively estimated (discussed later). Note that $\boldsymbol{\phi}$ is essentially a label confidence vector representing the estimated true label posterior probability.

Practical Implementation. We approximate the noise posterior probability by averaging the crowdsourced labels of M annotators $P(\tilde{Y}_k = i|\boldsymbol{X} = \mathbf{x}) \approx \frac{1}{M}\sum_{m=1}^{M} \tilde{\mathbf{y}}_k^{(m)} = \bar{\mathbf{y}}_k$, where we **assume each annotator tags each bag with uniform contribution**. Therefore, we can obtain the matrix $\mathcal{A}$ and vector $\mathcal{B}$ before training, adding minimal overhead to training time.

Additionally, with the improved version of bag-level URE proposed above, we explore the instance-level URE (Lin et al. 2022; Andrews, Tsochantaridis, and Hofmann 2002) since learning an instance-level classifier could further capture instance-label relationships which is more desirable in fine-grained applications. Note that the bag-level predictions can be obtained by instance-level classifiers, i.e., $\hat{\mathbf{y}} = 1 - \prod_{j=1}^{n_i}(1 - g(\boldsymbol{x}_j))$, where g is the instance-level classifier and $\hat{\mathbf{y}}$ is the bag-level prediction.

Corollary 3 (Instance-level URE)**.** *The notation is the same as that in Theorem 2. Assume the base loss l is symmetric about labels and predictions $l(\hat{p}, y) = l(1 - \hat{p}, 1 - y)$ and satisfies the rule of product addition about the prediction $l(\prod_{j=1}^{n_i}\hat{p}_j, \cdot) = \sum_{j=1}^{n_i} l(\hat{p}_j, \cdot)$. Then the instance-level URE with respect to $R(f)$ is $\tilde{R}(f) = \mathbb{E}_{(\mathbf{x},\bar{\mathbf{y}})\sim P(\boldsymbol{X},\tilde{\boldsymbol{Y}})}(\tilde{L}_t(f(\mathbf{x}), \bar{\mathbf{y}}))$, where $\tilde{L}_t(f(\mathbf{x}), \bar{\mathbf{y}}) = \sum_{k=1}^{C} \tilde{l}_t(f_k(\mathbf{x}), \bar{\mathbf{y}}_k)$,*

$$\tilde{l}_t(f_k(\mathbf{x}), \bar{\mathbf{y}}_k) = \frac{1}{2P(\tilde{Y}_k = \bar{\mathbf{y}}_k|\boldsymbol{X} = \mathbf{x})}[\phi_k \sum_{j=1}^{n_i} l(g_k(\boldsymbol{x}_j), 1) + (1 - \phi_k)\sum_{j=1}^{n_i} l(g_k(\boldsymbol{x}_j), 0)] \tag{5}$$

Remark 3. The cross-entropy loss satisfies the assumptions. The formula $\tilde{l}_t$ allows direct accumulation of instance-level loss, simplifying calculations with an instance-level classifier. The instance prediction vector in Corollary 3 can also be used to calculate instance similarity via methods like inner product or cosine similarity.

In experiments, we implement the beg-level URE as the training objective which turns out to be reliable for CMIML on various datasets, and we leave more studies on instance-level URE to future works.

2.3 Estimation of Transition Matrix

Another challenge for implementing our cross-label transition-based unbiased learning is the estimation of the transition matrices. In this subsection, we detail the estimation process of the aggregated cross-label transition matrices. Specifically, we extend previous estimation methods tailored for noisy label learning on single label (Patrini et al. 2017; Liu and Tao 2015) to the scenario of cross-label transition for multi-label classification.

Theorem 4 (Estimation of Transition Matrix)**.** *Let $\mathbf{x}^*_{\{p\to i,q\to j\}}$ and $\mathbf{x}^*_{\{p\to i\}}$ be two or single -class anchor point, respectively, i.e. $P(Y_p = i, Y_q = j|\boldsymbol{X} = \mathbf{x}^*_{\{p\to i,q\to j\}}) = 1$ and $P(Y_p = i|\boldsymbol{X} = \mathbf{x}^*_{\{p\to i\}}) = 1$. Then these anchor points can be used to estimate the probability of cross-label transition.*

$$T_{i,j}^{(p,q)} = \begin{cases} P(\tilde{Y}_q = j|\boldsymbol{X} = \mathbf{x}^*_{\{p\to i,q\to 1-j\}}) & if\ p \neq q \\ P(\tilde{Y}_q = j|\boldsymbol{X} = \mathbf{x}^*_{\{p\to 1-i\}}) & else \end{cases}$$

Theorem 4 shows that the transition matrix can be estimated using single or two-class anchor points. However, since the probabilities for true labels $P(Y_p = i|\boldsymbol{X} = \mathbf{x}^*)$ are unavailable, the anchor points must also be estimated. Proposition 5 suggests that the sample point farthest from the classification boundary can serve as the anchor point, as expanded by methods (Xia et al. 2019; Liu and Tao 2015).

Proposition 5 (Estimation of Anchor Point)**.** *Let f be the MIML classifier trained with the aggregated labels $\bar{\mathbf{y}}$. Then the sample point with the highest probability of the classifier f is chosen as the corresponding anchor point: $\hat{\mathbf{x}}^*_{\bigcup_{k\in\mathcal{I}}\{k\to j_k\}} = \arg\max_{\mathbf{x}} \sum_{k\in\mathcal{I}} [I_{[j_k=1]} f_k(\mathbf{x}) + I_{[j_k=0]}(1 - f_k(\mathbf{x}))]$, where $\mathcal{I}$ is a subset of the class set $\{1, 2, ..., C\}$ with one or two elements.*

With the estimated anchor points, the transition matrix can be readily calculated using Theorem 4. Note that the noise probability can be estimated by $P(\tilde{Y}_q = j|\boldsymbol{X} = \hat{\mathbf{x}}^*) = \bar{\mathbf{y}}_q$. Nevertheless, experiments show that selecting the top K samples and approximating the transition probability by averaging their noise rates $\frac{1}{K}\sum_{k=1}^{K} \bar{y}_{k,q}$ is more efficient than choosing the top one, where $\bar{y}_{k,q} = P(\tilde{Y}_q = 1|\boldsymbol{X} = \hat{\mathbf{x}}^*_k)$ is the aggregated label for $\hat{\mathbf{x}}^*_k$, the k-th anchor point.

In summary, the training process is outlined in appendix. Before formal training, we estimate the cross-label transition matrix and calculate the confidence $\boldsymbol{\phi}$, which is computationally efficient compared to the training phase. Specifically, we use neural networks to fit aggregated labels $\bar{\mathbf{y}}$, sort predictions to select the top k anchor points, and then calculate $\boldsymbol{\phi}$ by solving the linear equation system in Theorem 2.

2.4 Generalization Bound

Let $\mathcal{F}$ denote the hypothesis set and $H_k = \{h|h : \mathbf{x} \mapsto f_k(\mathbf{x}), f \in \mathcal{F}\}$ denote the function space for the k-th class. With the expected Rademacher complexity $\mathfrak{R}_n(\cdot)$ of the composite space $\cdot$ (Mohri, Rostamizadeh, and Talwalkar 2012), a generalization error bound of our proposed unbiased risk estimator can be justified as the following Theorem.

Theorem 6 (Generalization Error Bound)**.** *Assume that the unbiased loss function $\tilde{l}(f(\mathbf{x}), \mathbf{y})$ is L_p-Lipschtz continuous with respect to $f(\mathbf{x})$, and the base loss function l is upper-bounded by constant $\mathcal{J}$, let $\mu = max\{L_p, \mathcal{J}\}$ and $\lambda = \min_{1\le k\le C}\ \min_{\tilde{\mathbf{y}}_k\in\{0,1\}} P(\tilde{Y}_k = \tilde{\mathbf{y}}_k)$. Then, $\forall\delta > 0, f \in \mathcal{F}$, with probability $\geq 1 - \delta$ over a sample $S = \{(\mathbf{x}_i, \bar{\mathbf{y}}_i)\}_{i=1}^n$, we have*

$$R(f) \le \sum_{i=1}^{n} \frac{\tilde{L}_b(f(\mathbf{x}_i), \bar{\mathbf{y}}_i)}{n} + 2\sum_{k=1}^{C} \mathfrak{R}_n(\tilde{l} \circ H_k) + \frac{C\mu}{2\lambda}\sqrt{\frac{1}{2n} ln \frac{1}{\delta}}$$

Theorem 6 indicates optimizing the empirical risk can reduce the population level error on the true risk estimator, which is an important theoretical guarantee for training CMIML with the unbiased risk estimator.

Metrics	Datasets	MIML-LLMC	MIMLSVM	KiSar	DeepMIML	PLAIN	Ours-BCE	Ours
Macro-F1↑ [a]	Scene	0.3952	0.0056	0.5874	0.2582	0.5588	0.3967	**0.6204**
	Reuters	**0.8742**	0.0838	0.6484	0.7504	0.7259	0.7481	0.7823
	AzoVin	0.0295	0.0022	0.0013	0.0008	0.0007	0.0542	**0.0731**
	GeoSul	0.0452	0.0037	0.0014	0.0013	0.0050	0.0305	**0.0746**
	HalMar	0.1035	0.0099	0.0061	0.0019	0.0099	**0.1125**	0.0799
	PyrFur	0.0291	0.0008	0.0037	0.0006	0.0009	0.0691	**0.0766**
Coverage↓ [a]	Scene	1.97	2.07	**0.95**	3.05	2.18	2.48	1.47
	Reuters	1.30	1.66	0.47	1.42	1.53	1.34	**0.31**
	AzoVin	50.98	59.77	62.65	50.87	49.28	59.18	**44.56**
	GeoSul	40.34	41.11	46.02	43.24	39.52	46.78	**37.01**
	HalMar	30.76	32.89	36.54	33.90	40.01	40.01	**27.03**
	PyrFur	49.61	57.69	69.11	58.22	51.62	62.25	**46.80**
AvgRec↑ [a]	Scene	0.3655	0.2366	0.5130	0.2269	0.6429	0.5477	**0.6492**
	Reuters	0.9198	0.4339	0.6460	0.8646	0.7971	0.9105	**0.9723**
	AzoVin	0.1042	0.0904	0.0019	0.0002	0.0019	0.0815	**0.4759**
	GeoSul	0.1489	0.1234	0.0030	0.0011	0.0271	0.1108	**0.4959**
	HalMar	0.1898	0.1565	0.0188	0.0010	0.0546	0.1910	**0.4294**
	PyrFur	0.1200	0.1047	0.0114	0.0003	0.0066	0.1262	**0.4737**

[a] ↑ (↓) represents the higher the better (the lower the better). The same applies to other tables.

Table 2: Experimental results when $T_{1,1}^{(q,q,m)} = 0.5$ and $T_{0,0}^{(q,q,m)} = 0.5$.

3 Experiment

3.1 Setup

Datasets. We evaluate our method on six datasets from various fields: 1) Scene[1], an image dataset for MIML (Zhou and Zhang 2006); 2) Reuters[2], a text dataset for MIML (Zhou, Sun, and Li 2009); 3) Geobacter sulfurreducens (GeoSul), Azotobacter vinelandii (AzoVin), Haloarcula marismortui (HalMar), and Pyrococcus furiosus (PyrFur)[3], datasets used for genome-wide protein function prediction through MIML (Wu, Huang, and Zhou 2014). We set the number of annotators M to 30, calculate the noise rate using Equation 3, and generate uniformly distributed random numbers between 0 and 1 to determine noise labels. For the transition matrix, we consider three CMIML scenarios across all classes: a) $T_{1,1}^{(q,q,m)} = T_{0,0}^{(q,q,m)}$, b) $T_{1,1}^{(q,q,m)} < T_{0,0}^{(q,q,m)}$ and c) $T_{1,1}^{(q,q,m)} > T_{0,0}^{(q,q,m)}$. Specifically, the true rates $(T_{1,1}^{(q,q,m)}, T_{0,0}^{(q,q,m)})$ are set to a) $(0.5, 0.5)$, b) $(0.3, 0.5)$ and c) $(0.5, 0.3)$. To satisfy the normalization conditions in Section 2.1, the remaining elements $T_{i,\cdot}^{(p,q,m)}$ are sampled from a normal distribution $\mathcal{N}(\psi_i, \left(\frac{\psi_i}{3}\right)^2)$, where $\psi_i = \frac{2(1-T_{i,i}^{(p,p,m)})}{C \cdot M}$ and $i \in \{0, 1\}$. The original transition matrix is then normalized using the equation in Section 2.1, and the aggregated matrix is computed as described in Section 2.1.

Baseline. To demonstrate the excellence of our method, we compare it against different types of baselines: 1) Four MIML methods: MIMLSVM (Zhou and Zhang 2006), which employs Hausdorff distance among bags for SVM-based classification; KiSar (Li et al. 2012), which uses convex optimization to discover relations between input patterns and output labels; DeepMIML (Feng and Zhou 2017), which models label-instance relations and establishes instance-level classifiers; and MIML-LLMC (Yang, Tang, and Min 2022), which improves MIML classification by extracting local label correlations. 2) One partial multi-label learning (Xu et al. 2023) (PMLL) method: PLAIN (Wang et al. 2023), which leverages instance- and label-level similarities to generate pseudo-labels via label propagation and trains a deep model to fit the disambiguated labels. 3) a variant of our method that uses binary cross entropy (BCE) loss instead of unbiased loss. Note that we use mean squared error (MSE) loss for MIML-LLMC and PLAIN, Hinge loss for MIMLSVM and KiSar, and BCE loss for DeepMIML. For all the baselines, the aggregated crowdsourced label $\bar{\mathbf{y}}$ is used. Moreover, the experimental results using the majority voting labels are provided in the appendix.

Inplementation Details. We use a neural network with hidden layers of 256, 512, and 256 units, which is applied to all models for consistency. Our model is trained with the Adam optimizer, with a learning rate $1e^{-4}$ and a weight decay $1e^{-5}$. The hyperparameter K, used to estimate the transition matrix in Section 2.3, is set to 5. Other parameters are set to their default values. The base loss l in Theorem 2 is BCE loss. Following *Ockham's Razor* and given that bag-level classifiers outperform instance-level ones (Zhou 2004), we use bag-level loss in Theorem 2 instead of instance-level loss.

For the performance evaluations, three metrics widely used in MIML are selected to analyze experimental results, including macro-averaged F1 (Macro-F1), average Recall(AvgRec) and coverage (Aggarwal et al. 2017). For the first two metrics, a higher value represents a better performance, while for coverage, the lower the better. We perform ten-fold cross-validation and report the mean as well as the

[1]https://www.lamda.nju.edu.cn/data_MIMLimage.ashx
[2]https://www.lamda.nju.edu.cn/data_MIMLtext.ashx
[3]https://www.lamda.nju.edu.cn/data_MIMLprotein.ashx

Metrics	Datasets	MIML-LLMC	MIMLSVM	KiSar	DeepMIML	PLAIN	Ours-BCE	Ours
Macro-F1↑	Scene	0.4010	0.0057	0.5860	0.2335	0.5183	0.3648	**0.6203**
	Reuters	**0.8686**	0.0796	0.6328	0.7343	0.7142	0.7281	0.7823
	AzoVin	0.1987	0.0021	0.0019	0.0005	0.0012	**0.2156**	0.0634
	GeoSul	0.0696	0.0076	0.0026	0.0002	0.0061	0.0462	**0.0749**
	HalMar	0.0463	0.0095	0.0113	0.0008	0.0096	**0.1191**	0.0957
	PyrFur	0.0288	0.0011	0.0036	0.0002	0.0007	0.0305	**0.0737**
Coverage↓	Scene	2.11	2.08	**0.94**	2.21	2.88	2.62	1.04
	Reuters	1.29	1.67	0.48	1.36	2.54	1.83	**0.39**
	AzoVin	56.00	56.86	63.63	55.27	51.53	52.66	**48.65**
	GeoSul	45.98	42.37	45.54	38.78	42.53	41.85	**37.42**
	HalMar	52.08	32.41	37.14	29.46	28.48	34.52	**24.83**
	PyrFur	48.01	55.03	67.36	54.28	57.43	56.26	**45.60**
AvgRec↑	Scene	0.2543	0.2367	0.5130	0.1403	0.6254	0.5652	**0.6686**
	Reuters	0.9132	0.4327	0.6264	0.7830	0.7479	0.8913	**0.9557**
	AzoVin	0.2751	0.0862	0.0034	0.0003	0.0082	0.0946	**0.4670**
	GeoSul	0.2350	0.1353	0.0082	0.0002	0.0921	0.1052	**0.4653**
	HalMar	0.1127	0.1574	0.0328	0.0005	0.0620	0.1262	**0.4341**
	PyrFur	0.1099	0.1119	0.0112	0.0006	0.0272	0.1516	**0.5048**

Table 3: Experimental results when $T_{1,1}^{(q,q,m)} = 0.3$ and $T_{0,0}^{(q,q,m)} = 0.5$.

Metrics	Datasets	MIML-LLMC	MIMLSVM	KiSar	DeepMIML	PLAIN	Ours-BCE	Ours
Macro-F1↑	Scene	0.6002	0.0060	0.5942	0.2344	0.5931	0.3649	**0.6248**
	Reuters	**0.8833**	0.0787	0.6378	0.7362	0.6847	0.7263	0.7729
	AzoVin	0.0400	0.0020	0.0040	0.0005	0.0008	0.0661	**0.0720**
	GeoSul	0.0287	0.0029	0.0021	0.0010	0.0046	0.0463	**0.0727**
	HalMar	**0.1572**	0.0077	0.0083	0.0007	0.0106	0.0854	0.0957
	PyrFur	0.0379	0.0018	0.0022	0.0005	0.0011	0.0613	**0.0782**
Coverage↓	Scene	1.97	2.07	**0.95**	3.05	2.18	2.48	1.47
	Reuters	1.30	1.66	0.47	1.42	1.51	1.34	**0.32**
	AzoVin	50.98	59.78	62.65	50.87	49.28	59.17	**44.56**
	GeoSul	40.34	41.11	46.03	43.24	39.54	46.78	**37.01**
	HalMar	30.76	32.89	36.54	33.90	32.06	40.02	**27.03**
	PyrFur	49.64	57.69	69.11	58.22	51.62	62.25	**46.80**
AvgRec↑	Scene	0.5240	0.2367	0.5200	0.1418	0.6747	0.5973	**0.6582**
	Reuters	0.9291	0.4322	0.6282	0.7851	0.7819	0.8684	**0.9585**
	AzoVin	0.1156	0.0880	0.0036	0.0006	0.0084	0.0975	**0.4503**
	GeoSul	0.0989	0.1233	0.0066	0.0007	0.0316	0.0893	**0.5087**
	HalMar	0.2751	0.1313	0.0248	0.0009	0.06185	0.0786	**0.4342**
	PyrFur	0.2267	0.1147	0.0089	0.0009	0.1812	0.1467	**0.4415**

Table 4: Experimental results when $T_{1,1}^{(q,q,m)} = 0.5$ and $T_{0,0}^{(q,q,m)} = 0.3$.

standard deviation for metric results.

3.2 Comparison Results

We report experimental results across three CMIML scenarios, as shown in Tables 2, 3, and 4. Overall, our model outperforms the baseline on multiple datasets and metrics. In terms of avgRec, our method consistently exceeds others; for instance, on the PyrFur dataset in Table 3, our recall rate is **4.5 times** higher than the second-best model, MIMLSVM. This is because most baselines fail to correct errors from incorrect labeling, but predicting mislabeled annotations leads to a low true positive rate and poor recall. In contrast, our CMIML unbiased loss generates negative gradients, helping the model correct mislabels and avoid incorrect predictions, thus significantly improving recall. For macro-F1, our method shows at least a 4.1% improvement over the baseline, with a maximum increase of **1.65 times**. In terms of coverage, our model, using CMIML unbiased loss, improves by at least 3.6%, with up to **1.5 times** the coverage of the second-best model. The strong performance across different scenarios demonstrates that our approach effectively and accurately handles complex CMIML tasks.

Error between Confidence $\boldsymbol{\phi}$ and True Label y. Please note that the label confidence ϕ_k in Theorem 2 is consistent in form with $I_{[\mathrm{y}_k=1]}$ in the basic loss function $l(f_k(\mathbf{x}), \mathrm{y}_k) = I_{[\mathrm{y}_k=1]} l(f_k(\mathbf{x}), 1) + (1 - I_{[\mathrm{y}_k=1]}) l(f_k(\mathbf{x}), 0)$. In fact, $\boldsymbol{\phi}$ indeed

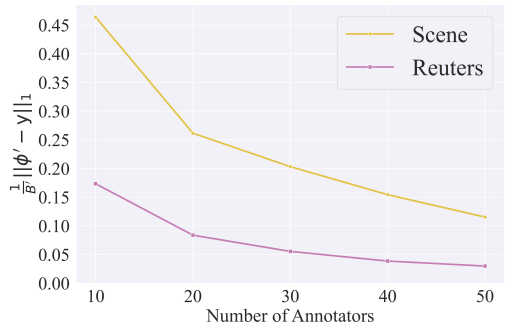

Figure 2: Average absolute error per element between the clamped confidence $\boldsymbol{\phi}'$ of $\boldsymbol{\phi}$ in Theorem 2 and true label $\mathbf{y}$.

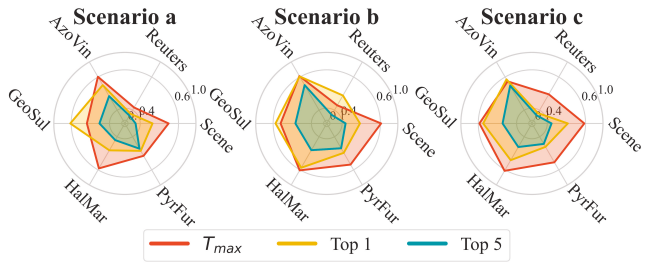

Figure 3: The relative error of transition matrix estimation under different strategies.

reflects the confidence in the true label. Figure 2 illustrates the error between the clamped $\boldsymbol{\phi}$ (restricted to the range 0 to 1) and the true labels $\mathbf{y}$ across different numbers of annotators. The error used is the average absolute error per element, i.e. $\frac{1}{B'}||\boldsymbol{\phi}' - \mathbf{y}||_1$, where B' is the number of training samples and $\boldsymbol{\phi}'$ is the clamped $\boldsymbol{\phi}$. Both $\boldsymbol{\phi}$ and $\mathbf{y}$ are two-dimensional matrices, with the first dimension representing samples and the second representing classes. Figure 2 clearly shows that as the number of annotators increases, the error between $\boldsymbol{\phi}$ and $\mathbf{y}$ decreases. Notably, when the number of annotators reaches 50, the error on the Scene dataset drops to 0.11, and on the Reuters dataset, it falls below 0.04. This suggests that our method can nearly restore the true label $\mathbf{y}$ using the label confidence $\boldsymbol{\phi}$ with a large number of annotators. This finding supports the superiority of our method.

Error of Different Transition Matrix Estimation Strategies. Figure 3 shows the error of different transition matrix estimation strategies. Based on prior research (Li et al. 2022; Patrini et al. 2017; Xia et al. 2019), $T_{\max}$ estimates the transition matrix using the prediction probability of the most reliable sample, while Top k averages the crowdsourced labels of the top k anchor points. The error is measured as the relative error $\frac{||\hat{\boldsymbol{T}}' - \boldsymbol{T}'||_2}{||\boldsymbol{T}||_2}$, where $\hat{\boldsymbol{T}}'$ and $\boldsymbol{T}'$ are the flattened estimated and true transition matrices, respectively. The results show the Top 5 method achieves the lowest estimation error.

4 Related Work

4.1 Multi-Instance Multi-Label Classification

In past decades, multi-instance multi-label classification (MIMLC) has attracted much research attention. (Zhou and Zhang 2006; Zhou, Sun, and Li 2009) approach scene classification and text categorization by decomposing the MIML problem into multi-label classification (Li, Ouyang, and Zhou 2015; Wei, Shi, and Li 2021; Wei et al. 2022) and multi-instance classification (Zhou 2004; Eberts and Ulges 2021; Lin et al. 2023) tasks. However, this decomposition can lead to the loss of critical information. To address this, methods have been developed that directly tackle the MIML problem (Li et al. 2012). With the rapid advancement of deep learning, recent approaches leverage deep models to generate features directly (Chen et al. 2013; Wu, Huang, and Zhou 2014; Feng and Zhou 2017; Yang et al. 2017; Yu et al. 2019), eliminating the need for additional instance generators. Traditional MIML approaches consider the training data to be fully supervised. However, obtaining precise data annotations is expensive and time-consuming. To mitigate this problem, weakly-supervised MIML has been proposed (Xu et al. 2012; Yang, Jiang, and Zhou 2013; Yang et al. 2019; Nguyen and Raich 2021). In this paper, we explore a more practical setup and propose the CMIML framework, which reduces the cost of data annotation and achieves robust classification without precise annotations.

4.2 Learning from Crowdsourced Labels

Crowdsourcing (Dalvi et al. 2013; Patrini et al. 2017) is widely used in practice and increasingly studied in academia, as it collects low-cost but unreliable labels to reduce the budget of large-scale data annotation. Traditional crowdsourcing methods using expectation-maximization(EM) algorithms identify accurate labels by modeling crowdsourced labels (Dalvi et al. 2013; Patrini et al. 2017). Subsequently, advanced techniques rooted in deep learning have been introduced, showcasing remarkable efficacy. These methodologies address the challenge of crowdsourced label noise through the acquisition of label transition matrices (Gao et al. 2022; Chen et al. 2020). Going beyond the single-instance single-label scenario, (Li et al. 2018; Zhang and Wu 2018) studies the crowdsourcing problem in the context of learning with multiple labels and (Del Amor et al. 2023) with multiple instances. Some works study crowdsourced multi-label learning (Xia et al. 2024), but overlook cross-label transitions in multi-instance learning.

5 Conclusion

In order to reduce annotation costs, this article focuses on the crowdsourced MIML problem labeled by multi-source annotation. Firstly, the first URE was proposed, which was further simplified into bag-level and instance-level versions by combining aggregated labels and transition matrices. A generalization upper bound is proposed to provide theoretical assurance. In terms of understanding, this method can be seen as using negative gradients to help the model forget incorrect annotations. Extensive experiments have demonstrated the superior performance of our method.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (Grant No. 62402424). Haobo Wang is also supported by the Pioneer R&D Program of Zhejiang (No. 2024C01035). This work is also partially supported by the Fundamental Research Funds for the Central Universities (226-2024-00049) and (226-2024-00145).

References

Aggarwal, A.; Ghoshal, S.; Shetty, A. M. S.; Sinha, S.; Ramakrishnan, G.; Kar, P.; and Jain, P. 2017. Scalable Optimization of Multivariate Performance Measures in Multi-instance Multi-label Learning. In *AAAI*, 1698–1704. AAAI Press.

Andrews, S.; Tsochantaridis, I.; and Hofmann, T. 2002. Support Vector Machines for Multiple-Instance Learning. In *NeurIPS*, 561–568. MIT Press.

Bragg, J.; Mausam; and Weld, D. S. 2013. Crowdsourcing Multi-Label Classification for Taxonomy Creation. In *AAAI*, 25–33. AAAI Press.

Briggs, F.; Lakshminarayanan, B.; Neal, L.; Fern, X. Z.; Raich, R.; Hadley, S. J.; Hadley, A. S.; and Betts, M. G. 2012. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *The Journal of the Acoustical Society of America*, 131(6): 4640–4650.

Carbonneau, M.-A.; Cheplygina, V.; Granger, E.; and Gagnon, G. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77: 329–353.

Chen, Z.; Chi, Z.; Fu, H.; and Feng, D. 2013. Multi-instance multi-label image classification: A neural approach. *Neurocomputing*, 99: 298–306.

Chen, Z.; Wang, H.; Sun, H.; Chen, P.; Han, T.; Liu, X.; and Yang, J. 2020. Structured Probabilistic End-to-End Learning from Crowds. In *IJCAI*, 1512–1518. ijcai.org.

Cheng, D.; Liu, T.; Ning, Y.; Wang, N.; Han, B.; Niu, G.; Gao, X.; and Sugiyama, M. 2022. Instance-Dependent Label-Noise Learning with Manifold-Regularized Transition Matrix Estimation. In *CVPR*, 16609–16618. IEEE.

Dalvi, N. N.; Dasgupta, A.; Kumar, R.; and Rastogi, V. 2013. Aggregating crowdsourced binary ratings. In *WWW*, 285–294. International World Wide Web Conferences Steering Committee / ACM.

Del Amor, R.; Pérez-Cano, J.; López-Pérez, M.; Terradez, L.; Aneiros-Fernandez, J.; Morales, S.; Mateos, J.; Molina, R.; and Naranjo, V. 2023. Annotation protocol and crowdsourcing multiple instance learning classification of skin histological images: The CR-AI4SkIN dataset. *Artif. Intell. Medicine*, 145: 102686.

Eberts, M.; and Ulges, A. 2021. An End-to-end Model for Entity-level Relation Extraction using Multi-instance Learning. In *EACL*, 3650–3660. Association for Computational Linguistics.

Feng, J.; and Zhou, Z. 2017. Deep MIML Network. In *AAAI*, 1884–1890. AAAI Press.

Feng, L.; Shu, S.; Cao, Y.; Tao, L.; Wei, H.; Xiang, T.; An, B.; and Niu, G. 2021. Multiple-Instance Learning from Similar and Dissimilar Bags. In *SIGKDD*, 374–382. ACM.

Feng, L.; Shu, S.; Cao, Y.; Tao, L.; Wei, H.; Xiang, T.; An, B.; and Niu, G. 2023. Multiple-Instance Learning From Unlabeled Bags With Pairwise Similarity. *TKDE*, 35(11): 11599–11609.

Gao, Z.; Sun, F.-K.; Yang, M.; Ren, S.; Xiong, Z.; Engeler, M.; Burazer, A.; Wildling, L.; Daniel, L.; and Boning, D. S. 2022. Learning from multiple annotator noisy labels via sample-wise label fusion. In *ECCV*, 407–422. Springer.

Li, S.; Xia, X.; Zhang, H.; Zhan, Y.; Ge, S.; and Liu, T. 2022. Estimating Noise Transition Matrix with Label Correlations for Noisy Multi-Label Learning. In *NeurIPS*, 24184–24198. Curran Associates, Inc.

Li, S.-Y.; Jiang, Y.; Chawla, N. V.; and Zhou, Z.-H. 2018. Multi-label learning from crowds. *TKDE*, 31(7): 1369–1382.

Li, X.; Ouyang, J.; and Zhou, X. 2015. Centroid prior topic model for multi-label classification. *Pattern Recognit. Lett.*, 62: 8–13.

Li, Y.; Hu, J.; Jiang, Y.; and Zhou, Z. 2012. Towards Discovering What Patterns Trigger What Labels. In *AAAI*, 1012–1018. AAAI Press.

Li, Y.; Rubinstein, B.; and Cohn, T. 2019. Exploiting worker correlation for label aggregation in crowdsourcing. In *ICML*, 3886–3895. PMLR.

Lin, T.; Xu, H.; Yang, C.; and Xu, Y. 2022. Interventional Multi-Instance Learning with Deconfounded Instance-Level Prediction. In *AAAI*, 1601–1609. AAAI Press.

Lin, T.; Yu, Z.; Hu, H.; Xu, Y.; and Chen, C. W. 2023. Interventional Bag Multi-Instance Learning On Whole-Slide Pathological Images. In *CVPR*, 19830–19839. IEEE.

Liu, B.; Xu, N.; Lv, J.; and Geng, X. 2023. Revisiting Pseudo-Label for Single-Positive Multi-Label Learning. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, 22249–22265. PMLR.

Liu, T.; and Tao, D. 2015. Classification with noisy labels by importance reweighting. *TPAMI*, 38(3): 447–461.

Ma, F.; Li, Y.; Li, Q.; Qiu, M.; Gao, J.; Zhi, S.; Su, L.; Zhao, B.; Ji, H.; and Han, J. 2015. FaitCrowd: Fine Grained Truth Discovery for Crowdsourced Data Aggregation. In *SIGKDD*, 745–754. ACM.

Maron, O.; and Lozano-Pérez, T. 1997. A Framework for Multiple-Instance Learning. In *NeurIPS*, 570–576. The MIT Press.

Min, C.; Lin, H.; Li, X.; Zhao, H.; Lu, J.; Yang, L.; and Xu, B. 2023. Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective. *Inf. Fusion*, 96: 214–223.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2012. *Foundations of Machine Learning*. MIT Press.

Nguyen, T.; and Raich, R. 2021. Active Learning in Incomplete Label Multiple Instance Multiple Label Learning. *CoRR*, abs/2107.10804.

Patrini, G.; Rozza, A.; Menon, A. K.; Nock, R.; and Qu, L. 2017. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *CVPR*, 2233–2241. IEEE Computer Society.

Rodrigues, F.; and Pereira, F. C. 2018. Deep Learning from Crowds. In *AAAI*, 1611–1618. AAAI Press.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2022. Learning from noisy labels with deep neural networks: A survey. *TNNLS*, 34(11): 8135–8153.

Tang, W.; Zhang, W.; and Zhang, M.-L. 2024. Multi-instance partial-label learning: Towards exploiting dual inexact supervision. *Science China Information Sciences*, 67(3): 132103.

Wang, H.; Yang, S.; Lyu, G.; Liu, W.; Hu, T.; Chen, K.; Feng, S.; and Chen, G. 2023. Deep Partial Multi-Label Learning with Graph Disambiguation. In *IJCAI*, 4308–4316. ijcai.org.

Wei, H.; Xie, R.; Feng, L.; Han, B.; and An, B. 2023. Deep Learning From Multiple Noisy Annotators as A Union. *TNNLS*, 34(12): 10552–10562.

Wei, T.; Mao, Z.; Shi, J.; Li, Y.; and Zhang, M. 2022. A Survey on Extreme Multi-label Learning. *CoRR*, abs/2210.03968.

Wei, T.; Shi, J.; and Li, Y. 2021. Probabilistic Label Tree for Streaming Multi-Label Learning. In *SIGKDD*, 1801–1811. ACM.

Wu, J.; Huang, S.; and Zhou, Z. 2014. Genome-Wide Protein Function Prediction through Multi-Instance Multi-Label Learning. *IEEE ACM Trans. Comput. Biol. Bioinform.*, 11(5): 891–902.

Xia, M.; Huang, Z.; Wu, R.; Lyu, G.; Zhao, J.; Chen, G.; and Wang, H. 2024. Unbiased Multi-Label Learning from Crowdsourced Annotations. In *ICML*, 54064–54081. PMLR.

Xia, X.; Liu, T.; Wang, N.; Han, B.; Gong, C.; Niu, G.; and Sugiyama, M. 2019. Are Anchor Points Really Indispensable in Label-Noise Learning? In *NeurIPS*, 6835–6846. Curran Associates, Inc.

Xu, N.; Liu, Y.; Zhang, Y.; and Geng, X. 2023. Progressive Enhancement of Label Distributions for Partial Multilabel Learning. *TNNLS*, 34(8): 4856–4867.

Xu, N.; Qiao, C.; Lv, J.; Geng, X.; and Zhang, M.-L. 2022. One Positive Label is Sufficient: Single-Positive Multi-Label Learning with Label Enhancement. In *NeurIPS*, 21765–21776. Curran Associates, Inc.

Xu, X.; Jiang, Y.; Xue, X.; and Zhou, Z. 2012. Semi-supervised multi-instance multi-label learning for video annotation task. In *MM*, 737–740. ACM.

Yang, H.; Zhou, J. T.; Cai, J.; and Ong, Y. 2017. MIML-FCN+: Multi-Instance Multi-Label Learning via Fully Convolutional Networks with Privileged Information. In *CVPR*, 5996–6004. IEEE Computer Society.

Yang, M.; Tang, W.; and Min, F. 2022. Multi-instance Multi-label Learning Based on Parallel Attention and Local Label Manifold Correlation. In *DSAA*, 1–10. IEEE.

Yang, S.; Jiang, Y.; and Zhou, Z. 2013. Multi-Instance Multi-Label Learning with Weak Label. In *IJCAI*, 1862–1868. IJCAI/AAAI.

Yang, Y.; Fu, Z.-Y.; Zhan, D.-C.; Liu, Z.-B.; and Jiang, Y. 2019. Semi-supervised multi-modal multi-instance multi-label deep network with optimal transport. *TKDE*, 33(2): 696–709.

Yu, W.-J.; Chen, Z.-D.; Luo, X.; Liu, W.; and Xu, X.-S. 2019. DELTA: A deep dual-stream network for multi-label image classification. *Pattern Recognition*, 91: 322–331.

Zhang, J.; and Wu, X. 2018. Multi-Label Inference for Crowdsourcing. In *SIGKDD*, 2738–2747. ACM.

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

Zhou, Z.; and Zhang, M. 2006. Multi-Instance Multi-Label Learning with Application to Scene Classification. In *NeurIPS*, 1609–1616. MIT Press.

Zhou, Z.-H. 2004. Multi-instance learning: A survey. *Department of Computer Science & Technology, Nanjing University, Tech. Rep*, 1.