ProMipl: A Probabilistic Generative Model for Multi-Instance Partial-Label Learning

Yin-Fang Yang[1, 2], Wei Tang[1, 2], Min-Ling Zhang*[1, 2]
[1] *School of Computer Science and Engineering, Southeast University, Nanjing, China*
[2] *Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China*
yangyf22@gmail.com, tangw@seu.edu.cn, zhangml@seu.edu.cn

***Abstract*—Multi-instance partial-label learning (MIPL) tackles scenarios where each training sample is represented as a multi-instance bag associated with a candidate label set. This set contains one true label and several false positives. Existing MIPL algorithms have predominantly focused on mapping multi-instance bags to candidate label sets for disambiguation. However, these algorithms may not be adequately generalizable in intricate real-world situations due to their reliance on heuristic methods for identifying true labels. In this paper, we propose ProMipl, i.e., a PRObabilistic generative model for Multi-instance partial-label learning, to address these challenges. ProMipl is the first attempt to explore the probabilistic generative model to infer latent ground-truth labeling information from the data generation process in multi-instance partial-label learning. Besides, the discovered underlying structures also provide improved explanations of the classification predictions. To circumvent the computationally intensive process of training the generative model, we formulate a unified variational lower bound within the stochastic gradient variational Bayesian framework for the model parameters. Experimental results from benchmark and real-world datasets show that our proposed ProMipl is competitive or superior to the state-of-the-art methods.**

***Index Terms*—Multi-Instance Partial-Label Learning, Generative Model, Probabilistic Disambiguation, Label Distribution, Variational Bayesian.**

I. Introduction

Weakly supervised learning is an effective strategy to train models in resource-constrained environments. Based on the quality and quantity of available labels, weak supervision can be classified into three categories: inexact, inaccurate, and incomplete supervision [1]. In particular, inexact supervision refers to a coarse correspondence between instances and labels, a common and challenging issue in real-world applications. Multi-instance learning (MIL) [1]–[5] and partial-label learning (PLL) [6]–[9] are two primary frameworks designed to handle inexact supervision in the instance space and label space, respectively. Recently, multi-instance partial-label learning (MIPL) [10] has been proposed to address dual inexact supervision, where inexact supervision co-occurs in both the instance and label spaces.

The occurrence of dual inexact supervision spans various domains. In histopathological image classification, the high resolution of these images requires dividing them into multi-instance bags, making the acquisition of ground truth labels from domain experts a costly and labor-intensive task [11]–[13]. A viable approach is to leverage crowd-sourced candidate label sets, which can substantially reduce the financial and resource burden associated with the labeling process [14].

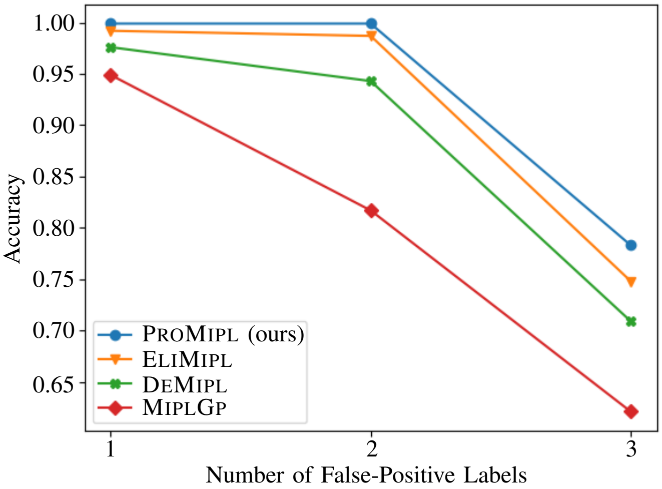

Fig. 1. Accuracy of our ProMipl and the existing MIPL approaches on the MNIST-MIPL dataset with the varying number of false-positive labels.

To address the challenge of dual inexact supervision in colorectal cancer classification, Tang et al. [15] introduced the MIPL algorithm DeMipl, which incorporates an attention mechanism to aggregate instances within a bag into a unified feature representation, alongside a disambiguation strategy to identify the true label. The attention mechanism mitigates inexactness in the instance space, while the disambiguation strategy resolves ambiguity in the label space. Building upon this, EliMipl algorithm [16] was proposed, extending DeMipl to utilize both candidate and non-candidate label information.

While existing approaches have demonstrated feasibility, their reliance on heuristic disambiguation rules limits their ability to generalize to more complex scenarios. As illustrated in Fig. 1, the performance of the existing MIPL methods degrades significantly under challenging conditions. To overcome this limitation, we introduce a novel framework called ProMipl, i.e., a PRObabilistic generative disambiguation model for MIPL. By treating the hidden ground-truth labels as latent variables, ProMipl builds a generative model that captures the underlying data generation process. Leveraging variational Bayesian principles, we derive a unified variational

[*] Corresponding author.

lower bound for the data log-likelihood via variational inference. This formulation facilitates simultaneous label disambiguation and model induction. By fitting a generative model, PROMIPL effectively extracts ground-truth label information, uncovering the inherent data structure for principled label clarification. The prediction model is subsequently optimized using a confidence-weighted cross-entropy loss. Comparative experimental results show that PROMIPL outperforms existing MIPL algorithms across various benchmarks.

Overall, our contributions are as follows: 1) We present the first generative framework specifically designed to address MIPL problems. 2) Our framework introduces a bag-wise Bayesian prior distribution, $p_{\theta}(z)$, enabling the model to effectively capture complex interactions between individual instances and collective bag-level information. 3) Extensive experiments demonstrate that PROMIPL consistently outperforms a wide range of existing MIPL and PLL algorithms.

The structure of the paper is as follows. Section II provides a brief review of related work. Section III details the proposed PROMIPL method. Section IV reports experimental results on both synthetic and real-world datasets. Finally, Section V concludes the paper and outlines potential future directions.

II. RELATED WORK

A. Multi-Instance Learning

Originating from drug activity prediction [17], MIL has gained significant attention in recent years due to its ability to handle complex data structures where labels are assigned to bags of instances rather than individual instances. Generative models have shown promising results in MIL tasks by capturing the underlying data distribution and generating informative representations. Contemporary deep MIL approaches predominantly leverage generative models for instance aggregation and label prediction [12], [13], [18]. Adel et al. [19] provide a methodological guide for modeling multiple-instance learning (MIL) tasks by introducing and analyzing generative models within a general framework and examining a variety of model structures and components. Pal et al. [20] employs a Bayesian graph neural network framework to jointly learn the parameters associated with the bag embedding, graph topology, and the weights of the graph neural network. Building on this concept, Zhang et al. [21] advanced the approach by utilizing a generative model with a shared bag-level latent factor and instance-level latent factors, effectively capturing both shared dependencies and individual variations. Additionally, the auxiliary classifier facilitates end-to-end prediction of instance and bag labels. Moreover, researchers have further explored the intrinsic capabilities of generative models to augment model performance [12], [14], [22], [23].

Nevertheless, although these methodologies have demonstrated encouraging outcomes with well-defined bag-level labels, they face considerable difficulties when confronted with the inherent ambiguity of bag-level annotations.

B. Partial-Label Learning

Partial-label learning deals with the challenge of disambiguating the true label from a set of candidate labels provided for each instance. Traditional supervised learning methods are ill-suited for this setting due to the inherent label noise and ambiguity [24]–[26]. PLL methods can be broadly categorized into two types: disambiguation-based and identification-based [27]. Generative models, which can model the underlying data distribution and the process generating the labels, have been increasingly employed to enhance PLL by leveraging their ability to represent complex data distributions and infer missing information [28]. Probabilistic models provide a natural way to handle the uncertainty in PLL. The seminal work by [6] pioneered a probabilistic approach treating the true label as a latent variable, aiming to maximize the likelihood of observed data. Feng et al. [29] dissected the generation process of partially labeled data using contrastive learning within an Expectation-Maximization (EM) framework. They iteratively identify true labels and refine model parameters accordingly. Yao et al. [30] employed deep convolutional neural networks for feature extraction and utilized an exponential moving average technique to infer latent true labels. Zhang et al. [31] introduced a GAN-based approach for partially labeled learning, where the generator models the data distribution and the discriminator distinguishes true labels among candidates. Lv et al. [32] proposed a classifier-consistent risk estimator grounded in empirical risk minimization principles for progressive true label discovery. Similarly, Wen et al. [33] introduced a generalized weighted loss function adaptable across different methods through customized weight assignments.

Despite their efficacy in tackling PLL challenges, these algorithms encounter difficulties in accommodating imprecise supervision across instances, thereby constraining their direct applicability to MIPL scenarios.

C. Multi-Instance Partial-Label Learning

Multi-instance partial-label learning (MIPL) is a novel learning framework extending both multi-instance learning (MIL) and partial-label learning (PLL). It addresses the challenge of inexact supervision present in both instance and label spaces. There are three existing primary MIPL algorithms, named MIPLGP [10], DEMIPL [15], and ELIMIPL [16]. Tang et al. [10] pioneered the MIPL framework with MIPLGP, operating within the instance-space paradigm. Conversely, DEMIPL adopts an embedded-space paradigm, utilizing a two-step process. It first aggregates each multi-instance bag into a unified feature representation via a disambiguated attention mechanism. Subsequently, it employs a momentum-based disambiguation strategy to identify true labels within the candidate set. Building on this, ELIMIPL leverages information from both candidate and non-candidate label sets through three distinct loss functions [16], learning mappings from multi-instance bags to candidate label sets while considering candidate label matrix sparsity.

Nevertheless, existing MIPL algorithms fall short in harnessing the potential to reconstruct ground-truth label, thereby

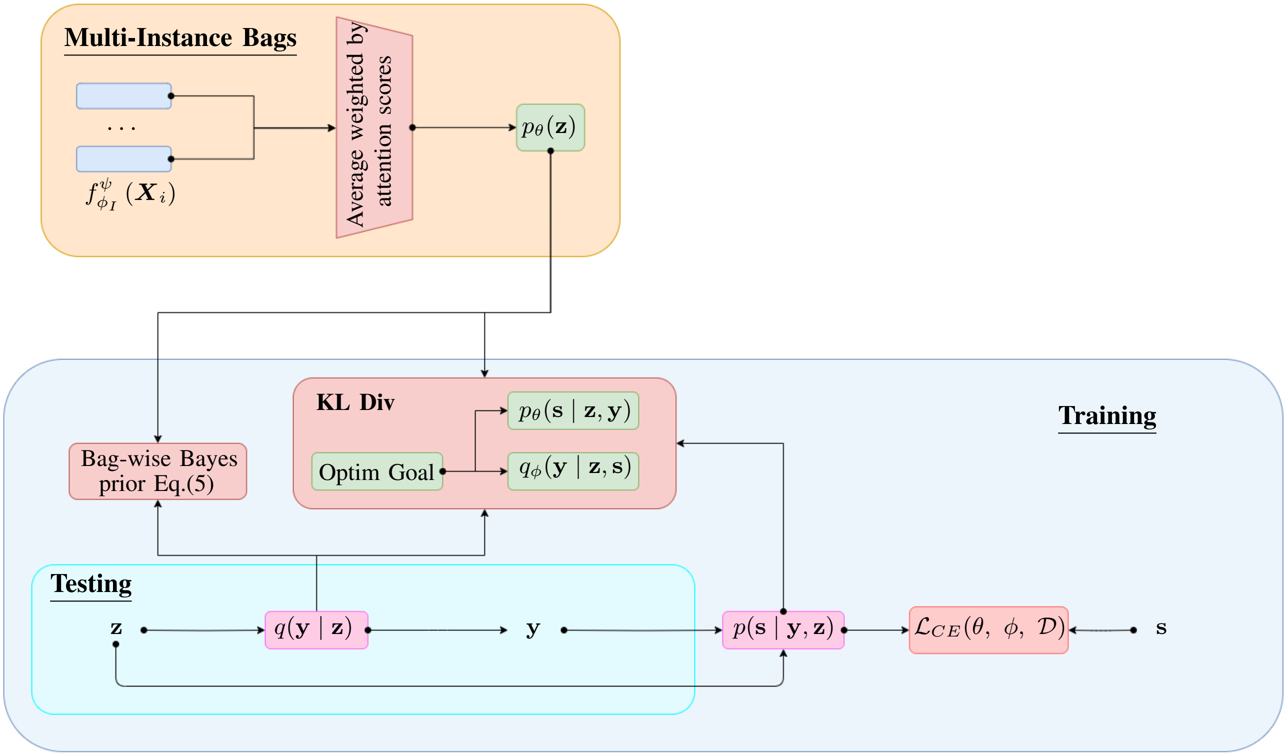
Multi-Instance Bags
. . .
$f^{\psi}_{\phi_I}(\boldsymbol{X}_i)$
Average weighted by attention scores
$p_\theta(\mathbf{z})$
KL Div
$p_\theta(\mathbf{s} \mid \mathbf{z}, \mathbf{y})$
Optim Goal
$q_\phi(\mathbf{y} \mid \mathbf{z}, \mathbf{s})$
Training
Bag-wise Bayes prior Eq.(5)
Testing
$\mathbf{z}$
$q(\mathbf{y} \mid \mathbf{z})$
$\mathbf{y}$
$p(\mathbf{s} \mid \mathbf{y}, \mathbf{z})$
$\mathcal{L}_{CE}(\theta,\ \phi,\ \mathcal{D})$
$\mathbf{s}$

Fig. 2. Description of our proposed PROMIPL framework, which follows Eq. (9) and Eq. (10) for model training and Eq. (5) for multi-instance bags wise-prior distribution. The term $q_\phi(y \mid z,\ s)$ represents the variational posterior distribution. The unshaded y represents the unobserved true label variable and $q_\phi(\mathbf{y} \mid \mathbf{z},\ \mathbf{s})$ is constructed to disambiguate the candidate label set by inferring the most probable ground-truth label from which candidate labels.

imposing constraints on the predictive performance of the models. This paper endeavors to bridge this lacuna by propounding an exploratory methodology that encapsulates the generative process of the MIPL data through the utilization of a bespoke probabilistic disambiguation model.

III. The Proposed Approach

Formally, we define a MIPL training dataset as $\mathcal{D} = \{(\boldsymbol{X}_i,\ \mathcal{S}_i) \mid 1 \leqslant i \leqslant m\}$, which consists of m bags and their corresponding candidate label sets. Each $\mathcal{S}_i$ contains a single true label and one or more false positives. The instance space is denoted as $\mathcal{X} = \mathbb{R}^d$, and the label space as $\mathcal{Y} = \{1,\ 2,\ \ldots,\ k\}$, encompassing k distinct classes. Each bag $\boldsymbol{X}_i$ is composed of n_i instances in a d-dimensional space. Both the candidate label set $\mathcal{S}_i$ and its complement $\overline{\mathcal{S}}_i$ are proper subsets of $\mathcal{Y}$, satisfying the condition $|\mathcal{S}_i| + |\overline{\mathcal{S}}_i| = |\mathcal{Y}| = k$, where $|\cdot|$ denotes the set cardinality.

The PROMIPL workflow is depicted in Fig. 2. The procedure initiates by leveraging a feature extractor, denoted as $f^{\psi}_{\phi_I}$, to produce instance-level feature vectors $\boldsymbol{Z}_i$ from each multi-instance bag $\boldsymbol{X}_i$. Subsequently, an adaptive weighted attention mechanism is deployed to construct a meticulously calibrated prior distribution $p_\theta(z)$ for the aggregated bag-level factors. To elucidate the underlying causes of label ambiguity in the MIPL data, we conceptualize the hidden ground-truth label as a latent variable and devise a generative model to encapsulate the generation process of the MIPL data.

A. Bayesian Prior in the Instance Space

Specifically, each multi-instance bag is denoted as $\boldsymbol{X}_i = \{\boldsymbol{x}_{i,1},\ \boldsymbol{x}_{i,2},\ \ldots,\ \boldsymbol{x}_{i,n_i}\} \in \mathbb{R}^{d \times n_i}$, encompassing n_i individual instances. To derive instance-level features, we utilize a neural network-based function $f^{\psi}_{\phi_I}$, parameterized by ϕ_I and ψ, which processes the multi-instance bag as input. This function is formally articulated as follows:

$$\boldsymbol{Z}_i = f^{\psi}_{\phi_I}(\boldsymbol{X}_i) = \{\boldsymbol{z}_{i,1},\ \boldsymbol{z}_{i,2},\ \cdots,\ \boldsymbol{z}_{i,n_i}\}, \tag{1}$$

where the instance-level features for the multi-instance bag $\boldsymbol{X}_i$ are encoded in $\boldsymbol{Z}_i \in \mathbb{R}^{l \times n_i}$, where each row $\boldsymbol{z}_{i,\ j}$ corresponds to the feature representation of the j-th instance within the bag. This matrix structure allows for the extraction of distinct characteristics for each constituent instance.

In practical settings, our observations are constrained to multi-instance bags and their associated candidate labels, with instances within each bag showcasing diverse characteristics and traits. To address this diversity, we devise a weighting mechanism that assigns attention scores to individual instances, reflecting their relevance and importance. This strategy enables the model to adeptly manage the inherent bag heterogeneity and effectively leverage the information from diverse instances. Initially, during the early stages of training, attention scores distribute evenly across all instances in multi-instance partial-label learning. However, as training progresses, the scores for positive instances consistently outpace those of their negative counterparts.

The dual nature of inexact supervision poses challenges for the attention mechanism in distinguishing between positive and negative samples during the early training phase, leading to imprecise attention score computations. As training advances, the mechanism gradually assigns distinct attention weights to these instances. To enhance alignment with the model's performance, we propose a dynamic approach that adjusts the attention scores. The calculation of attention scores is mathematically represented by the following equation:

$$a_{ij} = \operatorname{softmax}\left(\frac{\boldsymbol{W}^{\top}\left(\tanh\left(\boldsymbol{W}_1^{\top}\boldsymbol{z}_{i,j}\right) \odot \operatorname{sigm}\left(\boldsymbol{W}_2^{\top}\boldsymbol{z}_{i,j}\right)\right)}{\tau^{(t)}}\right), \tag{2}$$

where $\boldsymbol{W}^{\top}$, $\boldsymbol{W}_1^{\top}$, and $\boldsymbol{W}_2^{\top}$ are learnable parameters. $\tanh(\cdot)$ and $\operatorname{sigm}(\cdot)$ are the hyperbolic tangent and sigmoid functions, respectively. The operator $\odot$ denotes element-wise multiplication, and $\tau^{(t)}$denotes the temperature parameter of the margin-aware attention mechanism. More precisely, during the initial training stages, a higher temperature parameter is strategically employed to flatten the attention score distribution, preventing the mechanism from assigning excessive scores to indeterminate instances. As training progresses, a lower temperature parameter is employed to refine the distribution, amplifying the distinction between attention scores assigned to positive and negative samples. The dynamic adjustment of the temperature parameter at the t-th epoch is mathematically formulated as:

$$\tau^{(t)} = \max\left\{\tau_m,\ \tau^{(t-1)} * (1 - \Delta\tau)\right\}, \tag{3}$$

where τ_m and $\tau^{(t-1)}$ represent the minimum temperature and the temperature at the $(t-1)$-th epoch, respectively. $\Delta\tau$ represents the decremental rate of the temperature parameter. Consequently, Eq. (3) encapsulates an adaptive mechanism that adjusts the temperature parameter for the margin-sensitive attention mechanism, thereby facilitating a progressive refinement of the attention score distribution commensurate with the model's evolving ability for discriminating between positive and negative instances as training advances.

In multi-instance bags with varying quantities of positive instances, the attention score distribution exhibits heterogeneity, necessitating diverse temperature parameters for optimal accuracy across different bags. To address this issue, we propose a normalization strategy for attention scores, which ensures a balanced and efficient allocation of focus, regardless of the number of positive elements per bag:

$$\bar{a}_{ij} = \frac{a_{ij} - \bar{a}_i}{n_i - 1}, \tag{4}$$

where $\bar{a}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} a_{ij}$ is the mean value of the attention scores in the i-th multi-instance bag. Subsequent to obtaining normalized attention scores, we aggregate the instance-level feature representations to compose the bag-wise prior distribution $p_\theta\left(\mathbf{z}\right)$ in the following manner:

$$\begin{aligned} p_\theta(\mathbf{z}) = \prod_{k=1}^{D_{z_B}} \mathcal{N}(&\frac{1}{\sum_{j=1}^{n_i}\bar{a}_{ij}}\sum_{j=1}^{n_i}\bar{a}_{ij} f_{\phi_I}^{\psi}(\boldsymbol{X}_j), \\ &\frac{1}{\sum_{j=1}^{n_i}\bar{a}_{ij}}\sum_{j=1}^{n_i}\bar{a}_{ij} f_{\phi_I}^{\pi}(\boldsymbol{X}_j)). \end{aligned} \tag{5}$$

To elucidate further, the parameter D_{z_B} defines the dimensionality of the latent factors at the bag level in the context of multi-instance learning. The functions $f_{\phi_I}^{\psi}$ and $f_{\phi_I}^{\pi}$ represent the means and variances of Gaussian distributions respectively, and these distributions are parameterized by neural networks. This process then calculates the prior distribution for the entire bag by aggregating these factors through attention mechanisms. The attention-weighted aggregation results in a more nuanced and comprehensive representation of the latent variables. This refined approach enables the model to effectively capture the complex interactions between individual instances and the collective bag-level information, while also taking into account the importance of each instance within the broader multi-instance framework. This strategy allows the model to adeptly capture the intricate interplay between instance-specific and aggregate information, while also considering the relative significance of each instance within the broader context of the multi-instance bags setup.

B. Probabilistic Disambiguation in the Label Space

In real-world applications involving multi-instance partial-label datasets, our method develops a probabilistic model tailored to capture the underlying generative mechanism. This is achieved by introducing an unobserved latent variable $\mathbf{y}$ to embody the true class labels. The generative process is structured into three distinct sequential steps:

1) Sampling a multi-instance bag $\mathbf{z}$ from the bag-wise Bayesian prior probabilistic distribution $p_\theta(\mathbf{z})$;
2) Extracting a latent variable $\mathbf{y}$ as the true label of the bag $\mathbf{z}$ by means of stochastic sampling from the authentic class posterior distribution $p_\theta(\mathbf{y} \mid \mathbf{z})$, which encapsulates the underlying probabilistic relationship between the latent true label and the observed bag;
3) Distorting the ground-truth label of the bag to derive its set of candidate labels $\mathbf{s}$ via $p_\theta(\mathbf{s} \mid \mathbf{z},\ \mathbf{y})$ which is parameterized by θ.

Accordingly, the joint probability distribution $p_\theta(\mathbf{z},\ \mathbf{y},\ \mathbf{s})$ can be factorized as:

$$p_\theta(\mathbf{z},\ \mathbf{y},\ \mathbf{s}) = p_\theta(\mathbf{z})p_\theta(\mathbf{y} \mid \mathbf{z})p_\theta(\mathbf{s} \mid \mathbf{y},\ \mathbf{z}). \tag{6}$$

Considering the multi-instance partial-label training set $\mathcal{D}$, the aforementioned latent generative process can be acquired through the maximization of the log-likelihood function on the observed data. Nevertheless, the direct estimation of the generative model's parameters via likelihood maximization often presents a formidable challenge due to the inherent computational intractability. To circumvent this obstacle and render the optimization process more manageable, the generative and inference models can be simultaneously learned by maximizing the marginal likelihood function over the bags, which entails integrating out the latent variables and inferring the posterior distribution of the labels given the instances and their corresponding partial-label sets:

$$
\begin{aligned}
\log p_\theta(\mathbf{s} \mid \mathbf{z}) &= \log \int p_\theta(\mathbf{s},\ \mathbf{y} \mid \mathbf{z}) d\mathbf{y} \\
&= \log \int p_\theta(\mathbf{s} \mid \mathbf{z},\ \mathbf{y}) p_\theta(\mathbf{y} \mid \mathbf{z}) d\mathbf{y} \\
&= \log \int q_\phi(\mathbf{y} \mid \mathbf{z},\ \mathbf{s}) \frac{p_\theta(\mathbf{s} \mid \mathbf{z},\ \mathbf{y}) p_\theta(\mathbf{y} \mid \mathbf{z})}{q_\phi(\mathbf{y} \mid \mathbf{z},\ \mathbf{s})} d\mathbf{y} \\
&\geqslant \mathbb{E}_{q_\phi(\mathbf{y}\mid\mathbf{z},\ \mathbf{s})} \left[\log \frac{p_\theta(\mathbf{s} \mid \mathbf{z},\ \mathbf{y}) p_\theta(\mathbf{y} \mid \mathbf{z})}{q_\phi(\mathbf{y} \mid \mathbf{z},\ \mathbf{s})} \right] \\
&= \mathcal{L}(\mathbf{z},\ \mathbf{s};\ \theta,\ \phi),
\end{aligned}
\tag{7}
$$

in this formulation, $q_\phi(\mathbf{y} \mid \mathbf{z},\ \mathbf{s})$ represents the introduced variational posterior distribution, serving as an approximation to the true label posterior distribution $p_\theta(\mathbf{y} \mid \mathbf{z},\ \mathbf{s})$. The term $\mathcal{L}(\mathbf{z},\ \mathbf{s};\ \theta,\ \phi)$ denotes the derived variational lower bound, which functions as a surrogate loss function for the log-likelihood. This lower bound can be reformulated and expressed in the following manner:

$$
\begin{aligned}
\mathcal{L}(\mathbf{z},\ \mathbf{s};\ \theta,\ \phi) =& \mathbb{E}_{q_\phi(\mathbf{y}\mid\mathbf{z},\mathbf{s})} \left[\log p_\theta(\mathbf{s} \mid \mathbf{z},\ \mathbf{y}) \right] \\
&- KL \left[q_\phi(\mathbf{y} \mid \mathbf{z},\ \mathbf{s}) \mid p_\theta(\mathbf{y} \mid \mathbf{z}) \right],
\end{aligned}
\tag{8}
$$

the symbol $KL[\cdot \mid \cdot]$ represents the Kullback-Leibler divergence between two distributions. In general cases, the KL-divergence term in Eq. (8) cannot be analytically integrated. To render it computationally feasible, we employ the mean-field approximation technique and derive a closed-form solution for the KL-divergence term as follows [34]:

$$
\begin{aligned}
& KL \left[q_\phi(\mathbf{y} \mid \mathbf{z},\ \mathbf{s}) \mid p_\theta(\mathbf{y} \mid \mathbf{z}) \right] \\
&= \sum_{k=1}^{t} \mathbb{E}_{q_\phi(y_k\mid\mathbf{z},\ \mathbf{s})} \left[\log \frac{q_\phi\left(y_k \mid \mathbf{z},\ \mathbf{s}\right)}{p_\theta\left(y_k \mid \mathbf{z}\right)} \right] \\
&= \sum_{k=1}^{t} p_\phi^{y_k} \log \frac{p_\phi^{y_k}}{p_\theta^{y_k}} + \left(1 - p_\phi^{y_k}\right) \log \frac{1 - p_\phi^{y_k}}{1 - p_\theta^{y_k}},
\end{aligned}
\tag{9}
$$

in this context, the semantics of the ground-truth class posterior distribution $p_\theta(\mathbf{y} \mid \mathbf{z})$ are perspicuous, constituting the quintessential prediction model. Concurrently, the variational posterior $q_\phi(\mathbf{y} \mid \mathbf{z},\ \mathbf{s})$ (colloquially referred to as the inference model) endeavors to elucidate the ambiguity inherent in the candidate label set by inferring the labels with the highest probability of being the ground-truth label, from which the candidate labels could have been obfuscated, given the multi-instance bag $\mathbf{z}$. Concomitantly, $p_\theta(\mathbf{s} \mid \mathbf{z},\ \mathbf{y})$ (colloquially termed the generative model) obfuscates the ascertained ground-truth label to reconstruct the observed candidate labels. All distributions encompassed within the framework of Eq. (9) are instantiated as Gaussian distributions, the parameters of which are determined by neural networks.

To further elucidate the learning dynamics and behaviors exhibited by the aforementioned trio of model, we undertake the reformulation of the variational lower bound through the meticulous unfolding of the KL-divergence term, thereby yielding the following expression:

$$
\begin{aligned}
\mathcal{L}(\mathbf{z},\ \mathbf{s};\ \theta,\ \phi) =& \underbrace{\mathbb{E}_{q_\phi(\mathbf{y}\mid\mathbf{z},\ \mathbf{s})} \left[\log p_\theta(\mathbf{y} \mid \mathbf{z}) \right]}_{\mathcal{L}_{PRI}(\theta,\ \mathcal{D})} + \underbrace{H \left[q_\phi(\mathbf{y} \mid \mathbf{z},\ \mathbf{s}) \right]}_{\mathcal{L}_{CE}(\theta,\ \phi,\ \mathcal{D})} \\
&+ \underbrace{\mathbb{E}_{q_\phi(\mathbf{y}\mid\mathbf{z},\ \mathbf{s})} \left[\log p_\theta(\mathbf{s} \mid \mathbf{z},\ \mathbf{y}) \right]}_{\mathcal{L}_{KL}(\theta,\ \phi,\ \mathcal{D})}.
\end{aligned}
\tag{10}
$$

where $H[\cdot]$ signifies the entropy of a distribution. The initial two components of the objective function describe an entropy-regularized autoencoder mechanism with respect to s. This mechanism is designed to capture and leverage the intrinsic structural information of the data, thereby facilitating the disambiguation process. Conversely, the final component represents a cross-entropy loss, which serves to train the predictive model by incorporating the identified ground-truth labeling information $q_\phi(\mathbf{y} \mid \mathbf{z},\ \mathbf{s})$. Through the optimization of these terms within a unified variational lower bound framework, the model incrementally learns the underlying generative process inherent to the MIPL data.

Hence the complete loss function to minimise is

$$
\mathcal{L}(\mathcal{D},\ \theta,\ \phi) = \mathcal{L}_{PRI}(\theta,\ \mathcal{D}) + \mathcal{L}_{CE}(\theta,\ \phi,\ \mathcal{D}) + \mathcal{L}_{KL}(\theta,\ \phi,\ \mathcal{D}).
\tag{11}
$$

Algorithm 1 is the complete procedure of ProMipl. Consequently, this enables the disambiguation of the candidate label set and the concurrent induction of the target predictive model.

IV. Experiments

A. Experimental Configurations

1) Datasets: In the experimental design, we adhere to the methodology of DeMipl by employing four benchmark datasets for multi-instance partial-label learning, complemented by a real-world dataset. These benchmark datasets include MNIST-MIPL, FMNIST-MIPL, Birdsong-MIPL, and SIVAL-MIPL, which span a diverse range of applications such as image analysis and bioinformatics. The CRC-MIPL dataset, a real-world example, is further divided into four sub-datasets: CRC-MIPL-Row (C-Row), CRC-MIPL-SBN (C-SBN), CRC-MIPL-KMeansSeg (C-KMeans), and CRC-MIPL-SIFT (C-SIFT), each featuring distinct multi-instance characteristics. These features are derived from four different image bag generators: Row, single blob with neighbors (SBN), k-means segmentation (KMeansSeg), and scale-invariant feature transform (SIFT).

Table I presents a detailed overview of the dataset's characteristics. It includes the count of multi-instance bags, denoted as $\#bag$, and the total instances, represented by $\#ins$. To describe the instance distribution, we use $max.\ \#ins$, $min.\ \#ins$, and $avg.\ \#ins$, which correspond to the maximum, minimum, and average instance count across all bags. The dimension of each instance-level feature representation is denoted by $\#dim$, while $\#class$ and $avg.\ \#CLs$ signify the

TABLE I
CHARACTERISTICS OF THE BENCHMARK AND REAL-WORLD MIPL DATASETS.

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

```
Algorithm 1 $Y_* = \text{ProMipl}(\mathcal{D},\ T,\ X_*)$
Inputs:
  $\mathcal{D}$: the MIPL training set $\{(X_i,\ S_i) \mid 1 \leqslant i \leqslant m\}$, where $X_i = \{x_{i,1},\ x_{i,2},\ \ldots,\ x_{i,n_i}\}$, $x_{i,j} \in \mathcal{X}$, $\mathcal{X} = \mathbb{R}^d$, $S_i \subseteq \mathcal{Y}$, $\mathcal{Y} = \{1,\ 2,\ \ldots,\ k\}$
  $T$: the maximum number of training epochs
  $X_*$: the unseen multi-instance bag with $n_*$ instances
Outputs:
  $Y_*$: the predicted label for $X_*$
Initialize model parameters $\theta$, $\pi$
for $t = 1$ to $T$ do
    Shuffle training set $\mathcal{D}$ into $B$ mini-batches
    for $b = 1$ to $B$ do
        Extract the instance-level features based on Eq. (1)
        Calculate the attention scores with the temperature $\tau^{(t)}$ as stated by Eqs. (2) and (3), normalize the attention scores as stated by Eq. (4)
        Calculate the attention scores and instance-level features into bag-wise prior distribution based on Eq. (5)
        Compute unbiased estimator of the variational lower bound on $\mathcal{D}$ by Eq. (10);
    end for
    Update model parameters $\theta$, $\phi$ via gradient ascent.
end for
Extract the instance-level features of $X_*$ based on Eq. (1)
Calculate the attention scores and instance-level features into bag-wise prior distribution $p_\theta(z_*)$ according to Eqs. (2), (3), (4), and (5)
return $Y_* = \arg\max q_\phi(y \mid s,\ *)$
```

label space length and the average length of candidate label sets, respectively. To evaluate performance comprehensively, we manipulate the number of false-positive labels on the benchmark datasets, represented as r ($|\mathcal{S}_i| = r + 1$), where $\mathcal{S}_i$ denotes the set of candidate labels for each instance.

2) Comparative Algorithms: We extensively compare ProMipl with a diverse array of baselines, encompassing MIPL and PLL techniques. For MIPL algorithms, we consider MiplGp, DeMipl, and EliMipl. To address PLL algorithms, which are not specifically designed for the MIPL data, we employ two adaptation strategies, named the Mean strategy and the MaxMin strategy. The Mean strategy calculates the average feature values across all instances within a bag, resulting in a bag-level feature representation, and the MaxMin strategy identifies the maximum and minimum feature values for each dimension among instances within a bag and concatenates these values to form a bag-level feature representation. There are five involved PLL algorithms, including four deep-learning-based approaches (Proden, Rc, Lws and Cavl), and one feature-aware disambiguation algorithm (Pl-aggd).

3) Implementation: We employed PyTorch to implement ProMipl and trained the model using a single NVIDIA GeForce RTX 4090 GPU, and the code has been made publicly available on Github[1]. Utilizing Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a weight decay of 0.0001, we designed the optimization process. For feature extraction, a two-layer convolutional neural network and a fully connected network were employed for MNIST-mipl and FMNIST-mipl, while preprocessed features in Birdsong-mipl and SIVAL-mipl datasets necessitated only a fully connected network. In the CRC-mipl dataset, the feature extractor varied between four image bag generators or ResNet-34, followed by a fully connected network. The initial learning rate was selected from the set of $\{0.005,\ 0.01,\ 0.015,\ 0.02\}$, and a cosine annealing technique was applied to Birdsong-mipl and SIVAL-mipl. We set the number of epochs to 100 for MNIST-mipl, FMNIST-mipl datasets, 200 for Birdsong-mipl, SIVAL-mipl, and 300 for CRC-mipl. The initial configuration of the temperature parameter's annealing schedule was set as follows: $\{\tau(0) = 5,\ \Delta\tau = 0.0\}$ for MNIST-mipl and FMNIST-mipl, $\{\tau(0) = 5,\ \Delta\tau = 0.05,\ \tau_m = 0.05\}$ for Birdsong-mipl and SIVAL-mipl, and $\{\tau_m = 0.1,\ \Delta\tau = 0.05,\ \tau_m = 0.5\}$ for CRC-mipl. The selection process for the weights of these components involved evaluating a range of values, specifically $\{0.1, 0.5, 1.0, 1.5\}$. Following the same dataset partitioning strategy as DeMipl and EliMipl, we conducted ten random train/test splits with a 7:3 ratio and reported mean accuracy and standard deviations across these runs.

B. Comparison with MIPL and PLL Algorithms

Given the incompatibility of partial-label learning algorithms with the multi-instance structure of the MIPL data,

[1] ProMipl: https://github.com/yangyf22/ProMIPL

TABLE II
THE CLASSIFICATION ACCURACY (MEAN±STD) OF PROMIPL AND COMPARATIVE ALGORITHMS ON THE BENCHMARK DATASETS WITH THE VARYING NUMBERS OF FALSE-POSITIVE LABELS ($r \in \{1, 2, 3\}$).

Algorithm	r	MNIST-MIPL		FMNIST-MIPL		Birdsong-MIPL		SIVAL-MIPL	
ProMipl	1	**.999±.003**		**.922±.024**		**.776±.015**		**.682±.032**	
	2	**.999±.003**		**.889±.022**		.719±.018		**.633±.023**	
	3	**.783±.116**		.659±.041		.694±.021		.539±.024	
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	MaxMin	Mean	MaxMin	Mean	MaxMin	Mean	MaxMin
Proden	1	.605±.023	.508±.024	.697±.042	.424±.045	.296±.014	.387±.014	.219±.014	.316±.019
	2	.481±.036	.400±.037	.573±.026	.377±.040	.272±.019	.357±.012	.184±.014	.287±.024
	3	.283±.028	.345±.048	.345±.027	.309±.058	.211±.013	.336±.012	.166±.017	.250±.018
Rc	1	.658±.031	.519±.028	.753±.042	.731±.027	.362±.015	.390±.014	.279±.011	.306±.023
	2	.598±.033	.469±.035	.649±.028	.666±.027	.335±.011	.371±.013	.258±.017	.288±.021
	3	.392±.033	.380±.048	.408±.044	.390±.058	.298±.016	.363±.010	.237±.020	.267±.020
Lws	1	.463±.048	.242±.042	.726±.031	.535±.049	.265±.010	.225±.038	.240±.014	.289±.017
	2	.209±.028	.239±.048	.720±.025	.406±.040	.254±.012	.207±.034	.223±.008	.271±.014
	3	.205±.013	.218±.017	.579±.218	.318±.064	.205±.016	.216±.029	.194±.026	.244±.023
Cavl	1	.596±.074	.481±.030	.728±.047	.544±.012	.370±.013	.354±.015	.260±.013	.251±.023
	2	.412±.039	.389±.027	.586±.035	.265±.037	.335±.008	.237±.001	.216±.011	.216±.011
	3	.315±.020	.292±.032	.352±.035	.285±.022	.313±.017	.197±.014	.175±.020	.175±.020
Pl-aggd	1	.671±.027	.527±.035	.743±.026	.394±.012	.353±.019	.383±.014	.355±.015	.397±.028
	2	.595±.036	.439±.020	.678±.020	.371±.037	.314±.012	.372±.020	.315±.019	.360±.029
	3	.380±.032	.321±.043	.474±.057	.327±.028	.296±.015	.344±.011	.286±.018	.328±.023

we employed two data transformation techniques: the Mean-based approach and the MaxMin strategy, as described in [15]. The Mean strategy computes a bag-level feature representation by averaging the feature values of all instances within a bag. In contrast, the MaxMin strategy involves extracting the maximum and minimum feature values for each dimension within a multi-instance bag, and subsequently concatenating these values to create a unified bag-level feature representation.

1) Results on the Benchmark Datasets: Table II presents a comprehensive comparison of ProMipl's performance against three MIPL algorithms (MiplGp, DeMipl, EliMipl), four deep-learning-based PLL algorithms utilizing linear classifiers (Proden, Rc, Lws and Cavl), and the feature-aware disambiguation PLL algorithm (Pl-aggd). The evaluation process makes use of benchmark datasets that exhibit a range of false-positive label frequencies, allowing for a comprehensive analysis of the system's performance across different levels of label accuracy.

ProMipl consistently outperforms MiplGp in terms of average accuracy across almost all instances within the instance space paradigm. When focusing on methods operating within the embedding space paradigm, ProMipl demonstrates superior performance compared to DeMipl and EliMipl in 17 out of 24 cases. Notably, ProMipl surpasses MiplGp, DeMipl, and EliMipl in 29 out of 36 cases. This advantage is particularly evident on the MNIST-MIPL, FMNIST-MIPL and SIVAL-MIPL datasets, where ProMipl consistently achieves higher average accuracy than DeMipl. Furthermore, in scenarios with two candidate labels per bag ($r = 1$), ProMipl's average accuracy surpasses all other multi-instance partial-label learning algorithms, highlighting its effectiveness.

Interestingly, while partial-label learning algorithms perform adequately on simpler datasets like MNIST-MIPL and FMNIST-MIPL, their effectiveness diminishes with increasing dataset complexity, as observed with Birdsong-MIPL and SIVAL-MIPL. This trend underscores the limitations of these algorithms in handling complex data. Regarding data degradation strategies, the Mean strategy generally outperforms the MaxMin strategy on MNIST-MIPL and FMNIST-MIPL. Conversely, the MaxMin strategy tends to yield superior results on Birdsong-MIPL and SIVAL-MIPL. The findings indicate that the selection of the most appropriate data degradation approach is contingent upon the inherent characteristics and complexities of the dataset under consideration. Datasets exhibiting a relatively simplistic structure and composition may derive optimal benefits from the implementation of the straightforward Mean strategy, which involves the calculation and substitution of the arithmetic mean value. Conversely, datasets characterized by a higher degree of intricacy and multifaceted nature might necessitate the adoption of the more sophisticated MaxMin strategy, which entails the identification and substitution of the extreme upper and lower bounds within datasets, consequently maintaining a more comprehensive representation of the inherent dispersion and heterogeneity present in the original dataset.

2) Results on the Real-World Datasets: Table III provides a comprehensive comparative analysis of the performance

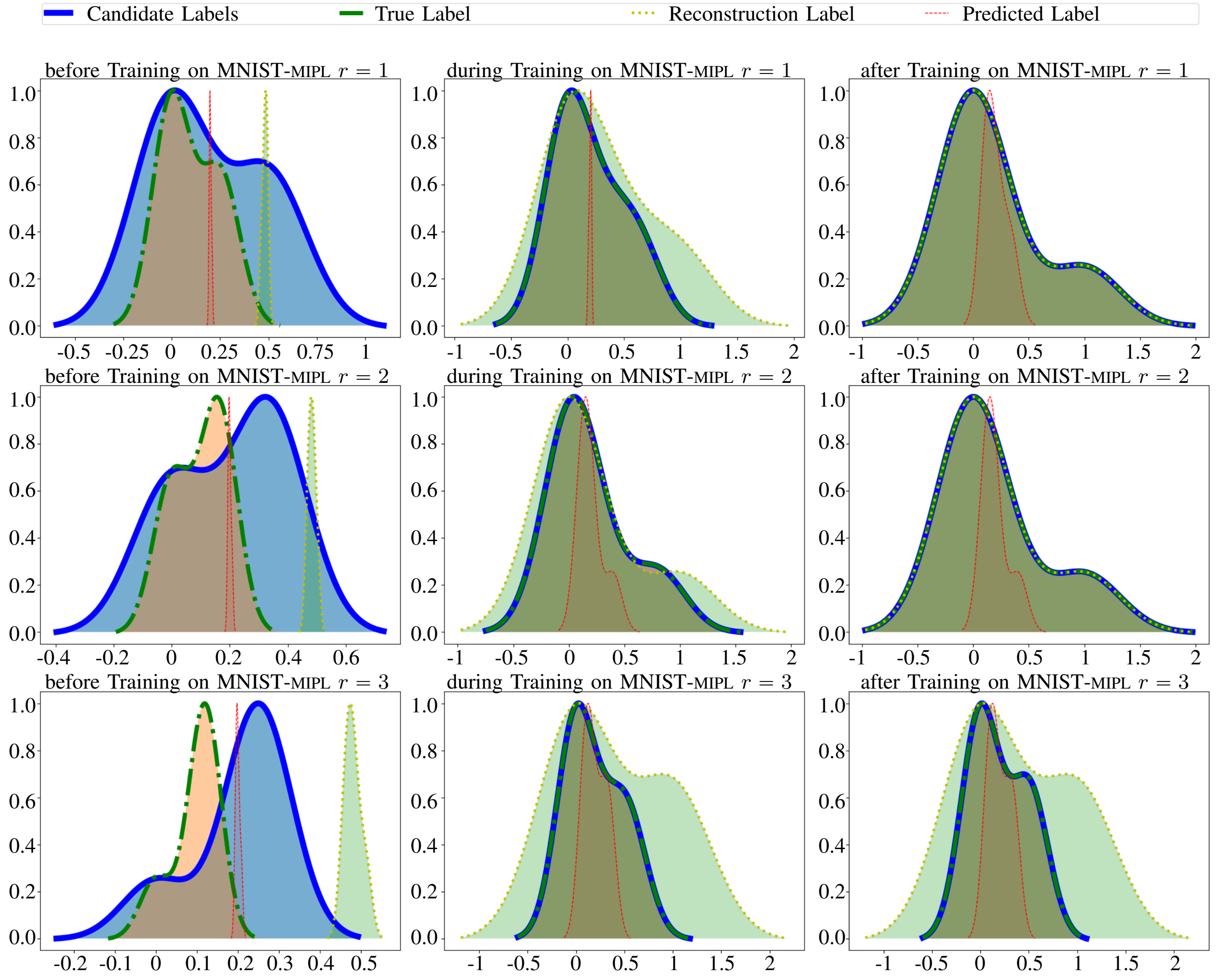

Fig. 3. Evolution of label distribution during training on the MNIST-MIPL dataset, which curves represent probability density functions. Subplots' horizontal coordinates express density and vertical coordinates denote values.

metrics across the CRC-MIPL dataset, juxtaposing our proposed methodology, henceforth denoted as ProMipl, with three MIPL algorithms, namely MiplGp, DeMipl, and EliMipl, four deep-learning-based partial-label learning algorithms that employ linear classifiers, specifically Proden, Rc, Lws, and Cavl, as well as the feature-aware disambiguation partial-label learning algorithm, designated as Pl-aggd. Notably, results for MiplGp on the C-SIFT dataset are unavailable due to computational limitations that indicated by the symbol ”–”.

ProMipl consistently demonstrates superior performance compared to MIPL algorithms across all evaluated cases. This dominance is particularly evident against DeMipl and EliMipl, where ProMipl consistently outperforms them across all four datasets. Moreover, ProMipl achieves statistically superior results compared to all partial-label learning algorithms. Interestingly, it exhibits commendable performance on simpler datasets like C-Row and C-SBN, while its performance significantly improves on more intricate datasets like C-KMeans and C-SIFT. This performance disparity is also observed between simpler and more complex datasets when comparing ProMipl, DeMipl and EliMipl.

Conversely, MiplGp and the partial-label learning algorithms exhibit the opposite trend, highlighting their limitations in effectively modeling complex features inherent in datasets like C-KMeans and C-SIFT. This deficiency underscores the urgent need for more effective MIPL algorithms specifically designed to handle the complexities of such data.

3) Effectiveness of the probabilistic generation process: To evaluate the efficacy of the probabilistic generation process, Fig. 3 illustrates the progression of probability density functions for candidate labels, the true label, reconstructed labels, and predicted labels during training on the MNIST-MIPL dataset. These label distributions are influenced by the generation process and disambiguation mechanism. As training advances, the distributions of candidate and reconstructed

TABLE III
THE CLASSIFICATION ACCURACY (MEAN±STD) OF PROMIPL AND COMPARATIVE ALGORITHMS ON THE REAL-WORLD DATASETS.

Algorithm	C-Row	C-SBN	C-KMeans	C-SIFT
PROMIPL	**.435±.009**	**.516±.012**	**.565±.013**	**.562±.011**
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
CAVL	.312±.043	.364±.066	.286±.062	.329±.033
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
CAVL	.368±.054	.503±.025	.311±.038	.274±.018
PL-AGGD	.460±.008	.524±.008	.434±.009	.285±.009

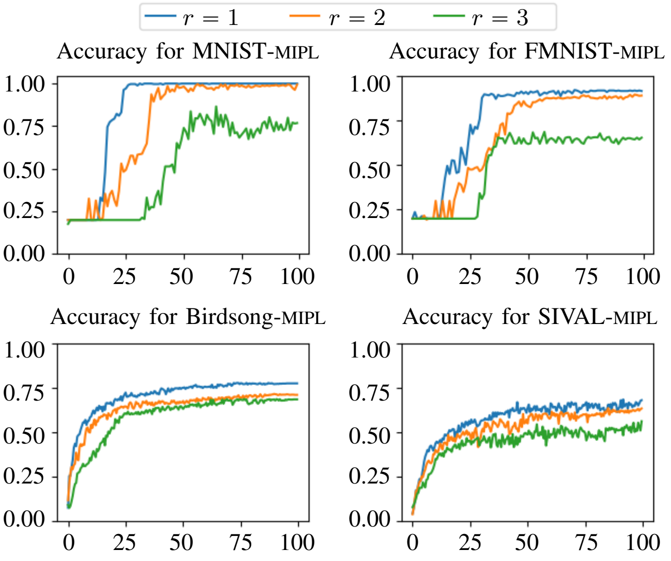

Fig. 4. Trend of accuracy in iteration on the testing set, which horizontal coordinates denote number of epochs and vertical coordinates represent values of accuracy.

labels converge towards the true label's distribution. The curve representing the true label's probability density closely resembles the shapes of other label distributions. On datasets with lower complexity, such as each bag with fewer false-positive labels, the reconstructed label's density function better aligns with the candidate labels, indicating a more challenging reconstruction task on high-complexity datasets.

Fig. 4 substantiates the earlier assertion regarding the influence of dataset complexity on the probabilistic generation process, as it presents the test set accuracy for each epoch on the first fold of all datasets. The curve representing the test accuracy on the higher complexity dataset consistently falls below the one for the dataset with fewer false-positive labels. Notably, PROMIPL converges to similar levels of test set accuracy across datasets of varying complexity, indicating its consistent predictive efficiency even in complex data.

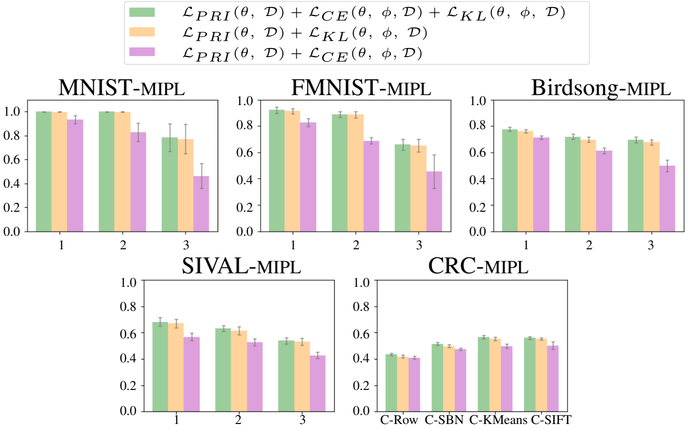

Fig. 5. The accuracy of PROMIPL with the variant on all datasets, which horizontal coordinates denote number of false-positive labels r or names of sub-dataset and vertical coordinates represent values of mean accuracy.

To delve deeper into the probabilistic generation process, we present two variants of PROMIPL. Eq. (11) is composed of three components, denoted as $\mathcal{L}_{PRI}(\theta,\ \mathcal{D})$, $\mathcal{L}_{CE}(\theta,\ \phi,\ \mathcal{D})$ and $\mathcal{L}_{KL}(\theta,\ \phi,\ \mathcal{D})$, each contributing to the model's objective. We devise two variants to evaluate their contributions: one where we exclude $\mathcal{L}_{CE}(\theta,\ \phi,\ \mathcal{D})$ (the entropy term), and another where $\mathcal{L}_{KL}(\theta,\ \phi,\ \mathcal{D})$ (the reconstruction term) is omitted. These modifications serve as a means to analyze their respective effects on the overall performance of the algorithm.

Fig. 5 illustrates the performance degradation observed across various datasets. The variant algorithm $\mathcal{L}_{PRI}(\theta,\ \mathcal{D}) + \mathcal{L}_{KL}(\theta,\ \phi,\ \mathcal{D})$ maintains performance metrics closely aligned with the original PROMIPL. In contrast, the variant $\mathcal{L}_{PRI}(\theta,\ \mathcal{D}) + \mathcal{L}_{CE}(\theta,\ \phi,\ \mathcal{D})$ displays a significant drop in mean accuracy, accompanied by an increase in standard deviation, when compared to the baseline. This disparity in data underscores the impact of the omitted reconstruction loss on the decline of the variant's performance, particularly in terms of prediction accuracy and model stability.

V. CONCLUSION

In this paper, we propose a probabilistic generative framework for multi-instance partial-label learning, termed PROMIPL. To the best of our knowledge, this is the first work to reformulate the MIPL problem using a probabilistic generative model. By modeling the correlations between instances and their bag-level label assignments, the proposed PROMIPL algorithm effectively disambiguates the candidate label set and identifies the most credible label for each training bag. Extensive experiments demonstrate that PROMIPL achieves superior or comparable performance to state-of-the-art methods. By developing PROMIPL, we endeavor to delve into a myriad of Bayesian MIPL paradigms to tackle the multifaceted challenges inherent in real-world MIPL scenarios.

References

1 Zhou, Z.-H., “A brief introduction to weakly supervised learning,” *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

2 Zhou, Z.-H., Sun, Y.-Y., and Li, Y.-F., “Multi-instance learning by treating instances as non-i.i.d. samples,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, Quebec, 2009, pp. 1249–1256.

3 Amores, J., “Multiple instance classification: Review, taxonomy and comparative study,” *Artificial intelligence*, vol. 201, pp. 81–105, 2013.

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

5 Ilse, M., Tomczak, J., and Welling, M., “Attention-based deep multiple instance learning,” in *Proceedings of the 35th International Conference on Machine Learning*, Stockholmsmässan, Stockholm, 2018, pp. 2127–2136.

6 Cour, T., Sapp, B., and Taskar, B., “Learning from partial labels,” *The Journal of Machine Learning Research*, vol. 12, pp. 1501–1536, 2011.

7 He, S., Feng, L., Lv, F., Li, W., and Yang, G., “Partial label learning with semantic label representations,” in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Washington D. C., 2022, pp. 545–553.

8 Li, X., Jiang, Y., Li, C., Wang, Y., and Ouyang, J., “Learning with partial labels from semi-supervised perspective,” in *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, Washington D.C., 2023, pp. 8666–8674.

9 Liu, L. and Dietterich, T., “A conditional multinomial mixture model for superset label learning,” in *Advances in Neural Information Processing Systems 25 (NIPS'12)*, Lake Tahoe, Nevada, 2012, pp. 557–565.

10 Tang, W., Zhang, W., and Zhang, M.-L., “Multi-instance partial-label learning: Towards exploiting dual inexact supervision,” *Science China Information Sciences*, vol. 67, no. 3, pp. 1–14, 2024.

11 Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., and Huang, J., “Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks,” *Medical Image Analysis*, vol. 65, p. 101789, 2020.

12 Doran, G. and Ray, S., “Multiple-instance learning from distributions,” *Journal of Machine Learning Research*, vol. 17, no. 128, pp. 1–50, 2016.

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

14 Li, X.-C., Zhan, D.-C., Yang, J.-Q., and Shi, Y., “Deep multiple instance selection,” *Science China Information Sciences*, vol. 64, pp. 1–15, 2021.

15 Tang, W., Zhang, W., and Zhang, M.-L., “Disambiguated attention embedding for multi-instance partial-label learning,” in *Advances in Neural Information Processing Systems 36*, New Orleans, LA, 2023, pp. 56 756–56 771.

16 Tang, W., Zhang, W., and Zhang, M.-L., “Exploiting conjugate label information for multi-instance partial-label learning,” in *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, Jeju, South Korea, 2024, pp. 4973–4981.

17 Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T., “Solving the multiple instance problem with axis-parallel rectangles,” *Artificial intelligence*, vol. 89, pp. 31–71, 1997.

18 Wehrmann, J., Cerri, R., and Barros, R., “Hierarchical multi-label classification networks,” in *Proceedings of the 35th International Conference on Machine Learning*, Stockholmsmässan, Stockholm, 2018, pp. 5075–5084.

19 Adel, T., Smith, B., Urner, R., Stashuk, D., and Lizotte, D. J., “Generative multiple-instance learning models for quantitative electromyography,” *arXiv preprint arXiv:1309.6811*, 2013.

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

21 Zhang, W., “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 Conference, 2021, pp. 3377–3383.

22 Cui, Y., Liu, Z., Liu, X., Liu, X., Wang, C., Kuo, T.-W., Xue, C., and Chan, A., “Bayes-mil: A new probabilistic perspective on attention-based multiple instance learning for whole slide images,” in *Proceedings of the 11st International Conference on Learning Representations*, Kigali, Rwanda, 2023.

23 Kandemir, M. and Hamprecht, F. A., “Instance label prediction by dirichlet process multiple instance learning,” in *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, Quebec City, Quebec, 2014, pp. 380–389.

24 Zhang, M.-L., Zhou, B.-B., and Liu, X.-Y., “Partial label learning via feature-aware disambiguation,” in *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, 2016, pp. 1335–1344.

25 Nguyen, N. and Caruana, R., “Classification with partial labels,” in *Proceedings of the 14th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Las Vegas, Nevada, 2008, pp. 551–559.

26 Ishida, T., Niu, G., Menon, A., and Sugiyama, M., “Complementary-label learning for arbitrary losses and models,” in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, CA, 2019, pp. 2971–2980.

27 Wang, H., Xiao, R., Li, Y., Feng, L., Niu, G., Chen, G., and Zhao, J., “Pico+: Contrastive label disambiguation for robust partial label learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3183–3198, 2024.

28 Zhang, M.-L. and Yu, F., “Solving the partial label learning problem: an instance-based approach,” in *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, 2015, pp. 4048–4054.

29 Feng, L., Lv, J., Han, B., Xu, M., Niu, G., Geng, X., An, B., and Sugiyama, M., “Provably consistent partial-label learning,” in *Advances in Neural Information Processing Systems 33*, Vancouver, Canada, 2020, pp. 10 948–10 960.

30 Yao, Y., Deng, J., Chen, X., Gong, C., Wu, J., and Yang, J., “Deep discriminative CNN with temporal ensembling for ambiguously-labeled image classification,” in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020, pp. 12 669–12 676.

31 Zhang, Y., Yang, G., Zhao, S., Ni, P., Lian, H., Chen, H., and Li, C., “Partial label learning via generative adversarial nets,” in *Proceedings of the 24th European Conference on Artificial Intelligence*, Santiago de Compostela, Spain, 2020, pp. 1674–1681.

32 Lv, J., Xu, M., Feng, L., Niu, G., Geng, X., and Sugiyama, M., “Progressive identification of true labels for partial-label learning,” in *Proceedings of the 37th International Conference on Machine Learning*, Virtual Conference, 2020, pp. 6500–6510.

33 Wen, H., Cui, J., Hang, H., Liu, J., Wang, Y., and Lin, Z., “Leveraged weighted loss for partial label learning,” in *Proceedings of the 38th International Conference on Machine Learning*, Virtual Conference, 2021, pp. 11 091–11 100.

34 Kingma, D. P. and Welling, M., “Auto-encoding variational bayes,” in *Proceedings of the 2nd International Conference on Learning Representations*, Banff, AB, 2014.

35 Wang, X., Yan, Y., Tang, P., Bai, X., and Liu, W., “Revisiting multiple instance neural networks,” *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

36 Struski, L., Tabor, J., and Zieli'nski, B., “Propall: Probabilistic partial label learning,” *ArXiv*, vol. abs/2208.09931, 2022.

37 Xu, Y., Gong, M., Chen, J., Liu, T., Zhang, K., and Batmanghelich, K., “Generative-discriminative complementary learning,” in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020, pp. 6526–6533.

38 Xu, N., Qiao, C., Geng, X., and Zhang, M.-L., “Instance-dependent partial label learning,” in *Advances in Neural Information Processing Systems 34*, Virtual Conference, 2021, pp. 27 119–27 130.

39 Tang, C.-Z. and Zhang, M.-L., “Confidence-rated discriminative partial label learning,” in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, San Francisco, CA, 2017, pp. 2611–2617.

40 Yeh, C.-K., Wu, W.-C., Ko, W.-J., and Wang, Y.-C. F., “Learning deep latent space for multi-label classification,” in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, San Francisco, CA, 2017, pp. 2838–2844.

41 Wang, D.-B., Li, L., and Zhang, M.-L., “Adaptive graph guided disambiguation for partial label learning,” in *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Anchorage, AK, 2019, pp. 83–91.

42 Yan, Y. and Guo, Y., “Multi-level generative models for partial label learning with non-random label noise,” in *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, Virtual Conference, 2021, pp. 3264–3270.