Asymmetric Beta Loss for Evidence-Based Safe Semi-Supervised Multi-Label Learning

Hao-Zhe Liu*
Nanjing University of Aeronautics and Astronautics
Nanjing, China
haozheliu@nuaa.edu.cn

Ming-Kun Xie*
Nanjing University of Aeronautics and Astronautics
Nanjing, China
mkxie@nuaa.edu.cn

Chen-Chen Zong
Nanjing University of Aeronautics and Astronautics
Nanjing, China
chencz@nuaa.edu.cn

Sheng-Jun Huang†
Nanjing University of Aeronautics and Astronautics
Nanjing, China
huangsj@nuaa.edu.cn

Abstract

The goal of semi-supervised multi-label learning (SSMLL) is to improve model performance by leveraging the information of unlabeled data. Recent studies usually adopt the pseudo-labeling strategy to tackle unlabeled data based on the assumption that labeled and unlabeled data share the same distribution. However, in realistic scenarios, unlabeled examples are often collected through cost-effective methods, inevitably introducing out-of-distribution (OOD) data, leading to a significant decline in model performance. In this paper, we propose a safe semi-supervised multi-label learning framework based on the theory of evidential deep learning (EDL), with the goal of achieving robust and effective unlabeled data exploitation. On one hand, we propose the asymmetric beta loss to not only compensate for the lack of robustness in common MLL losses, but also to solve the inherent positive-negative imbalance problem faced by the EDL losses in MLL. On the other hand, to construct a robust SSMLL framework, we adopt a dual-head structure to generate class probabilities and instance uncertainties. The former are used to generate pseudo-labels, while the latter are utilized to filter OOD examples. To avoid the need for threshold estimation, we develop a dual-measurement weighted loss function to safely perform unlabeled training. Extensive experiments on multiple benchmark datasets verify the effectiveness of the proposed method in both OOD detection and SSMLL tasks. Implementation is available at: https://github.com/hz681/AsymmetricBetaLoss.

CCS Concepts

• **Computing methodologies** → *Semi-supervised learning settings.*

Keywords

Semi-Supervised Multi-Label Learning, Evidential Learning

ACM Reference Format:
Hao-Zhe Liu, Ming-Kun Xie, Chen-Chen Zong, and Sheng-Jun Huang. 2024. Asymmetric Beta Loss for Evidence-Based Safe Semi-Supervised Multi-Label Learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671756

1 Introduction

Multi-label learning (MLL) stands as a pivotal machine learning paradigm designed to tackle situations where each instance can be associated with multiple class labels, as opposed to traditional single-label learning where each instance is assigned with a single label. The objective of MLL is to develop a classifier capable of predicting all relevant labels for unseen examples.

Due to the exponentially larger output space compared to single-label learning, training an effective MLL classifier necessitates a substantial number of precisely labeled examples. Unfortunately, in realistic tasks, acquiring a large scale of precise annotations proves to be challenging and costly. In order to handle such a problem, the semi-supervised multi-label learning (SSMLL) framework has been proposed to leverage the information of enormous unlabeled ones, and in consequence, several advanced methods have emerged to enhance the performance of SSMLL [14, 33, 38].

Typical SSMLL methods assume that labeled and unlabeled data share the same distribution. However, in many real-world scenarios, this assumption hardly holds since unlabeled examples are often obtained through cost-effective methods, *e.g.*, web crawling, inevitably introducing out-of-distribution (OOD) data. An intuitive strategy to handle OOD-corrupted unlabeled examples is combining a multi-label OOD detection method, which filters out OOD examples, and a SSMLL method, which exploits rest in-distribution (ID) ones. Unfortunately, due to the limited number of labeled examples, it struggles to obtain an effective OOD detector, resulting in a large number of ID examples being misclassified as OOD. This leads to a subsequent decline in the performance of SSMLL, which can be validated by Figure 1, which shows the performance comparison between our proposed method and CAP+JE (composed of a recent SSMLL method CAP and a multi-label OOD detection method Joint Energy) when unlabeled data is involved with OOD examples. It can be observed that our proposed method outperforms CAP+JE

*Both authors contributed equally to this research.
†Corresponding author.

KDD '24, August 25–29, 2024, Barcelona, Spain.

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671756

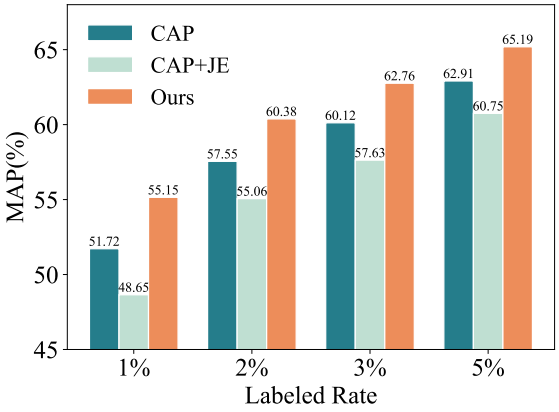


Figure 1: Performance of different semi-supervised multi-label learning method when unlabeled data is involved with OOD samples. The results demonstrate that simple combination methods, *e.g.*, CAP+JE, can hardly work and result in unfavorable performance.

with a significant margin under different labeled rates. Even only CAP can achieve better performance than CAP+JE. These results demonstrate that simple combination methods can hardly work and result in unfavorable performance. As the proportion of labeled data decreases, the gap between the two becomes larger. This means that the SOTA SSMLL method does not work in a safe SSMLL scenario due to the introduction of OOD examples.

To address this problem, we propose the evidence-based safe SSMLL framework to perform OOD detection and unlabeled data exploitation simultaneously. Considering that the commonly used BCE loss lacks robustness to OOD, we develop the asymmetric beta loss that not only produces class probabilities but also provides an uncertainty measurement. This allows us to utilize the former to generate pseudo-labels, while using the latter to measure the likelihood of an unlabeled example being an OOD. On one hand, to prevent the model from potential corruption by OOD examples, we adopt a dual-head model architecture. This comprises a clean head exclusively trained on labeled data for OOD detection and a noisy head trained on additional unlabeled data for multi-label classification. On the other hand, to avoid threshold estimation, we utilize the soft pseudo-label and uncertainty to weight the contributions of unlabeled examples. Extensive experimental results verify that, in comparison to various methods, our method achieves superior performance in both OOD detection and SSMLL tasks.

2 PRELIMINARIES

2.1 Problem Setting

We first formulate the problem of SSMLL with OOD data as follows. Let $\boldsymbol{x} \in \mathcal{X}$ represent a feature vector, and $\boldsymbol{y} \in \mathcal{Y}$ denote its corresponding label vector. Here, $\mathcal{X} = \mathbb{R}^d$ is the feature space, and $\mathcal{Y} = \{0, 1\}^K$ is the label space with K class labels. The training data can be divided into three subsets, $\mathcal{D}_{lb} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid i \in [N_{lb}]\}$ for labeled data, $\mathcal{D}_{id} = \{\boldsymbol{x}_i \mid i \in [N_{id}]\}$ for ID unlabeled data, and $\mathcal{D}_{od} = \{\boldsymbol{x}_i \mid i \in [N_{od}]\}$ for OOD unlabeled data, where N_{lb}, N_{id}, and N_{od} represent their respective example counts. The unlabeled set can be represented as a combination $\mathcal{D}_{ub} = \mathcal{D}_{id} \cup \mathcal{D}_{od}$. An instance is considered as an OOD if it does not contain any label in the label space $\mathcal{Y}$ [43]. Notably, we do not know which examples in the unlabeled data are ID and which ones are OOD during training. Our goal is to train a model based on both $\mathcal{D}_{lb}$ and $\mathcal{D}_{ub}$, which aims to leverage the usefulness of ID unlabeled examples, while alleviate the harmfulness of OOD unlabeled ones.

2.2 Evidential Deep Learning

Evidential Deep Learning (EDL) [27], which considers evidence as a measure of support for classifying an instance to a specific class, has emerged as a widely-used method for quantifying the uncertainty in model predictions. This idea can also be extended to determine whether an input instance is in- or out-of-distribution. In our safe semi-supervised multi-label learning scenario, it is essential to have a metric to distinguish OOD examples. It is significantly different from traditional MLL studies, which only require predicted probabilities. This motivates us to incorporate evidential deep learning theory, which provides an uncertainty measurement for distinguishing OOD examples.

EDL is derived from the theory of Subjective Logic (SL) [16]. In SL, there exists a belief mass b_j associated with each exclusive class $j = 1, ..., K$, along with a global uncertainty mass u. These masses are all non-negative and adhere to the constraint $u + \sum_{j=1}^{K} b_j = 1$. The masses can be derived from the evidence with respect to every class j.

$$b_j = \frac{e_j}{S}, \; u = \frac{K}{S}, \tag{1}$$

where $S = \sum_{j=1}^{K}(e_j + 1)$ and $e_j \geq 0$ is the evidence with respect to the j-th class. According to the results in [27], a belief mass assignment corresponds to a Dirichlet distribution $Dir(p|\boldsymbol{\alpha})$ with parameters $\alpha_j = e_j + 1$, *i.e.*, the probability density function of the prediction. Accordingly, the expected probability with respect to the j-th class can be computed as $\mathbb{E}[p_j] = \frac{\alpha_j}{S}$.

Consequently, we regard $\boldsymbol{\alpha}$ as the output of the neural network, which is employed to model the Dirichlet distribution. This approach enables the simultaneous estimation of probability and uncertainty for an instance. In the context of single-label scenarios, we often take cross-entropy loss as the base loss function and compute its Bayes risk as

$$\begin{aligned} \mathcal{L}_{ECE}(\alpha_{ij}, y_{ij}) &= \int_0^1 y_{ij} log p_{ij} Dir(p_{ij}|\boldsymbol{\alpha}_i) dp_{ij} \\ &= y_{ij}\big(\psi(S_i) - \psi(\alpha_{ij})\big), \end{aligned} \tag{2}$$

where $\psi(\cdot)$ represents digamma function.

3 The Proposed Method

3.1 Overview

Figure 2 provides an illustration of our proposed framework. The clean head is trained using the labeled loss $\mathcal{L}_{lb}$ over the labeled examples, while the noisy head is trained using both $\mathcal{L}_{lb}$ and the unlabeled loss $\mathcal{L}_{ub}$ over the unlabeled examples. Before talking about these losses, we first introduce the asymmetric beta loss (ABL) designed to handle OOD-corrupted multi-label data. Subsequently,

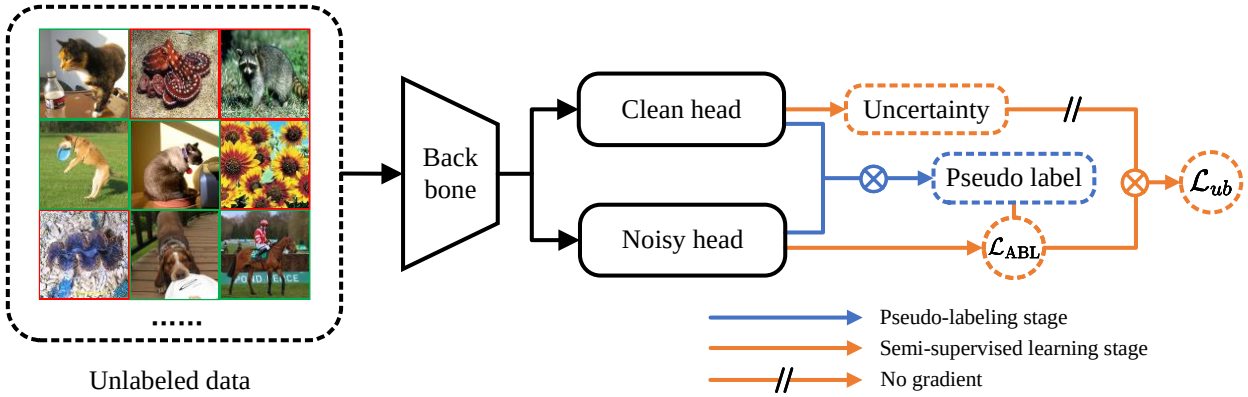


Figure 2: Overview of the proposed method. The clean head is updated solely based on labeled data, and is responsible for predicting the uncertainty of unlabeled data, which is omitted in the figure for simplicity. In contrast, the noisy head is exposed to both OOD samples (red box) and ID samples (green box) and is trained using the dual-measurement weighted loss function. Pseudo labels are generated through the combination of the outputs from both heads.

we will provide the detailed explanation of the entire learning framework. The pseudocode is illustrated in Algorithm 1.

3.2 Asymmetric Beta Loss

In MLL, the learning task can be regarded as a combination of K independent binary classification problems. From the evidence-based perspective, we consider each binary classification task as a binary evidential learning problem using a Beta distribution, *i.e.*, a Dirichlet distribution with two parameters. Specifically, each class j consists of two exclusive singletons (belongs to the class or not) and has masses b_j^+, b_j^-, u_j, where b_j^+, b_j^- are the belief masses of positive and negative labels respectively and u_j is the uncertainty mass for the j-th class. The three masses sum up to one, *i.e.*, $b_j^+ + b_j^- + u_j = 1$, where satisfy $b_j^+ \geq 0$, $b_j^- \geq 0$ and $u_j \geq 0$.

Given the prediction function $f(\cdot)$, for an instance $\boldsymbol{x}$, we obtain evidences $(e_j^+, e_j^-) = f_j(\boldsymbol{x})$ for positive and negative classes respectively, where $f_j(\boldsymbol{x})$ denotes the j-th component of $f(\boldsymbol{x})$. Based on the evidences, the masses can be derived by

$$b_j^+ = \frac{e_j^+}{S_j}, \; b_j^- = \frac{e_j^-}{S_j}, \; u_j = \frac{2}{S_j}, \tag{3}$$

where $S_j = e_j^+ + e_j^- + 2$ is referred to as the Beta strength. We use a Beta distribution $\text{Beta}(p_j|\alpha_j, \beta_j)$ to model the predicted probability $p_j \in [0, 1]$ for the j-th class, where $\alpha_j = e_j^+ + 1$ and $\beta_j = e_j^- + 1$ are two parameters to characterize the Beta distribution. The expected probability for the j-th class is the mean of the corresponding Beta distribution and computed as

$$\bar{p}_j = \mathbb{E}[p_j] = \frac{\alpha_j}{\alpha_j + \beta_j} \tag{4}$$

In MLL, the most common loss function is the binary cross entropy (BCE) loss. Similar to the ECE loss in Eq.(2), the evidential BCE can be defined in the same way for every class j. Note that in the following definition, we omit the index j when the context is clear.

$$\mathcal{L}_{\text{EBCE}}(\alpha, \beta, y) = \begin{cases} \psi(\alpha + \beta) - \psi(\alpha), & \text{if } y = 1, \\ \psi(\alpha + \beta) - \psi(\beta), & \text{if } y = 0. \end{cases} \tag{5}$$

where ψ is the digamma function. Although EBCE loss enjoys advantageous theoretical properties of evidential learning, it suffers from the issue of inherent positive-negative imbalance in MLL, *i.e.*, negative labels dominate the majority while positive ones constitute a smaller portion for every class, resulting in a degradation of model performance. To deal with this problem, ASL [23] loss is an improved version of BCE loss, which down-weights easy negative examples and enforces models to focus on positive ones. Formally, ASL loss can be defined as

$$\mathcal{L}_{\text{ASL}}(p, y) = \begin{cases} -(1-p)^{\gamma^+} log(p), & \text{if } y = 1, \\ -(p-m)^{\gamma^-} log(1-p), & \text{if } y = 0. \end{cases} \tag{6}$$

where $m = \min(p, c)$ and c is a constant probability shift parameter used for neglecting very easy negative examples. γ^+ and γ^- stand for focusing parameter. In practice, we often set $\gamma^- > \gamma^+$ to focus on positive examples.

Similar to EBCE loss, by taking ASL as the base loss function, we obtain the asymmetric beta loss (ABL) by calculating its Bayesian risk. When $y = 1$, it is easy to derive the analytical solution for the ABL loss through the integral operation. When $y = 0$, the probability shifting technique adopted by ASL loss cannot be directly used to ABL loss due to the fact that our model outputs the probability distribution rather than the certain probability. We define a shifted random variable p_c following the distribution $Beta(p_c|\alpha_c, \beta)$, where $\alpha_c = \max(\alpha - \frac{c}{1-c}\beta, 0)$. This modification guarantees that $\mathbb{E}(p_c) = 0$ when $\mathbb{E}(p) \leq c$ to discard negative samples when their probability is very low.

Algorithm 1 The Main Procedures of the Proposed Method

Input: Labeled dataset $\mathcal{D}_{lb} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N_l}$; unlabeled dtaset $\mathcal{D}_{ub} = \{\boldsymbol{x}_i\}_{i=1}^{N_u}$; feature extractor $g(\cdot)$; clean head $h^c(\cdot)$; noisy head $h^n(\cdot)$.

for $i = 1$ to *WarmupEpoch* **do**
  $\boldsymbol{z}_{lb} = g(\boldsymbol{x}_{lb})$, $(\boldsymbol{\alpha}^c_{lb}, \boldsymbol{\beta}^c_{lb}) = h^c(\boldsymbol{z}_{lb})$, $(\boldsymbol{\alpha}^n_{lb}, \boldsymbol{\beta}^n_{lb}) = h^n(\boldsymbol{z}_{lb})$. ▷ *calculate the output Beta Distribution*
  Update $g(\cdot)$ and $h^c(\cdot)$ with $\mathcal{L}_{ABL}(\boldsymbol{\alpha}^c_{lb}, \boldsymbol{\beta}^c_{lb}, \boldsymbol{y})$.
  Update $g(\cdot)$ and $h^n(\cdot)$ with $\mathcal{L}_{ABL}(\boldsymbol{\alpha}^n_{lb}, \boldsymbol{\beta}^n_{lb}, \boldsymbol{y})$.
end for
for i = *WarmupEpoch*+1 to *MaxEpoch* **do**
  $\boldsymbol{z}_{lb} = g(\boldsymbol{x}_{lb})$, $(\boldsymbol{\alpha}^c_{lb}, \boldsymbol{\beta}^c_{lb}) = h^c(\boldsymbol{z}_{lb})$, $(\boldsymbol{\alpha}^n_{lb}, \boldsymbol{\beta}^n_{lb}) = h^n(\boldsymbol{z}_{lb})$.
  $\boldsymbol{z}_{ub} = g(\boldsymbol{x}_{ub})$, $(\boldsymbol{\alpha}^c_{ub}, \boldsymbol{\beta}^c_{ub}) = h^c(\boldsymbol{z}_{ub})$, $(\boldsymbol{\alpha}^n_{ub}, \boldsymbol{\beta}^n_{ub}) = h^n(\boldsymbol{z}_{ub})$.
  Assign the pseudo labels $\hat{\boldsymbol{p}}_c = \frac{\boldsymbol{\alpha}^c_{ub}}{\boldsymbol{\alpha}^c_{ub}+\boldsymbol{\beta}^c_{ub}}$, $\hat{\boldsymbol{p}}_n = \frac{\boldsymbol{\alpha}^n_{ub}}{\boldsymbol{\alpha}^n_{ub}+\boldsymbol{\beta}^n_{ub}}$, $\tilde{\boldsymbol{y}} = \hat{\boldsymbol{p}}_c \otimes \hat{\boldsymbol{p}}_n$. ▷ *calculate pseudo labels by element-wise multiplication*
  Obtain the enhanced pseudo labels $\hat{\boldsymbol{y}}$ by $\hat{y}^+ = \frac{(\tilde{y})^m}{\tilde{y}^m+(1-\tilde{y})^m}$ and $\hat{y}^- = \frac{(1-\tilde{y})^m}{(\tilde{y})^m+(1-\tilde{y})^m}$.
  Calculate the uncertainties $\boldsymbol{u}$ and its normalized version $\hat{\boldsymbol{u}}$, where $u_i = \frac{2K}{\sum_{j=1}^{K} \boldsymbol{\alpha}^c_{ij}}$, $\hat{u}_i = \frac{u_i}{\max_{i' \in [N_u]} u_{i'}}$. ▷ *the shape of u is $(B, 1)$*
  Update $g(\cdot)$ and $h^c(\cdot)$ with $\mathcal{L}_{ABL}(\boldsymbol{\alpha}^c_{lb}, \boldsymbol{\beta}^c_{lb}, \boldsymbol{y})$.
  Update $g(\cdot)$ and $h^n(\cdot)$ with $\mathcal{L}_{ub}(\hat{\boldsymbol{u}}, \boldsymbol{\alpha}^n_{ub}, \boldsymbol{\beta}^n_{ub}, \hat{\boldsymbol{y}}) + \mathcal{L}_{ABL}(\boldsymbol{\alpha}^u_{lb}, \boldsymbol{\beta}^u_{lb}, \boldsymbol{y})$.
end for

Output: The trained neural network with $g(\cdot)$, $h^c(\cdot)$ and $h^n(\cdot)$.

Finally, the Asymmetric Beta Loss is defined as follows[1]

$$\mathcal{L}_{\text{ABL}} = \begin{cases} -\int_0^1 (1-p)^{\gamma^+} log(p) Beta(p|\alpha,\beta)dp, & \text{if } y = 1, \\ -\int_0^1 p_c^{\gamma^-} log(1-p_c) Beta(p_c|\alpha_c,\beta)dp_c, & \text{if } y = 0, \end{cases}$$
$$= \begin{cases} w^+[\psi(\alpha+\beta+\gamma^+) - \psi(\alpha)], & \text{if } y = 1, \\ w^-[\psi(\alpha_c+\beta+\gamma^-) - \psi(\beta)], & \text{if } y = 0, \end{cases} \tag{7}$$

where

$$\begin{cases} w^+ = \prod_{r=0}^{\gamma^+-1} \frac{\beta+r}{\alpha+\beta+r}, \\ w^- = \prod_{r=0}^{\gamma^--1} \frac{\alpha_c+r}{\alpha_c+\beta+r}. \end{cases} \tag{8}$$

Here, different from ASL loss, γ^+ and γ^- should be non-negative integers to calculate the weighting coefficients.

3.3 Evidence-Based Safe SSMLL Framework

Based on favorable theoretical properties of ABL loss, we can obtain two reliable measurements, the probability for generating pseudo-labels, and the uncertainty for detecting OOD examples. Below, we will introduce these two key components of our framework.

In SSMLL, the key of generating pseudo-labels lies in estimating a threshold to separate positive and negative labels for each instance or class. The recent method [38] has developed the class-distribution-aware thresholding strategy to separate positive and negative labels for each class according to the class proportions of labeled examples. Although this method can generate pseudo-labels with the proportions that approximates the true ones, it suffers from the issue of introducing many false positive labels. This is attributed to the alteration in the class proportions of unlabeled data caused by the presence of OOD data. Even if a OOD detection method is used, since detecting all OOD examples is challenging, it still results in the introduction of false positive labels. To solve this problem, we propose to generate soft pseudo-labels, which avoid the estimation of thresholds.

Specifically, to enhance the quality of pseudo-labels, we employ two classification head, *i.e.*, clean head trained only on labeled data and noisy head trained on additional unlabeled data. For notational simplicity, we decompose the classifier f into the backbone $g(\cdot)$, the clean head $h^c(\cdot)$ and the noisy head $h^n(\cdot)$. Similarly, we can obtain two groups of parameters $(\boldsymbol{\alpha}^c, \boldsymbol{\beta}^c) = h^c \circ g(\boldsymbol{x})$ and $(\boldsymbol{\alpha}^n, \boldsymbol{\beta}^n) = h^n \circ g(\boldsymbol{x})$. Then we generate pseudo-labels for an unlabeled instance $\boldsymbol{x}_i$ as

$$\tilde{\boldsymbol{y}}_i = \hat{\boldsymbol{p}}^c_i \otimes \hat{\boldsymbol{p}}^n_i, \tag{9}$$

where $\hat{\boldsymbol{p}}^c_i$ and $\hat{\boldsymbol{p}}^n_i$ are expected probabilities over K classes, and $\otimes$ means element-wise multiplication. From the equation, it is evident that only when both classification heads provide high predicted probabilities can we obtain a relatively confident positive pseudo-label. This reduces the risk of introducing false positives into model training, which contributes to enhancing pseudo-labeling performance.

While generating soft pseudo-labels helps avoid threshold estimation, the mod el may suffer from the under-fitting issue due to unconfident pseudo-labels. We enhance the confidence of pseudo-labels by introducing a trick, which involves taking the power of pseudo-labels and normalizing the result. Specifically, for every pseudo-label label $\tilde{y}$ (the index j is omitted), by taking power operation and normalization, we obtain its enhanced versions $\hat{y}^+ = \frac{(\tilde{y})^m}{\tilde{y}^m+(1-\tilde{y})^m}$ and $\hat{y}^- = \frac{(1-\tilde{y})^m}{(\tilde{y})^m+(1-\tilde{y})^m}$. To show why this trick can effectively enhance the confidence of pseudo-labels, Figure 3 illustrates the value of $\hat{y}$ with the increase of value of $\tilde{y}$ as m changes. From the figure, we can observe that as the pseudo-label values approach one or zero, the enhanced pseudo-labels become increasingly confident.

[1]The detailed derivation can be found in Appendix.

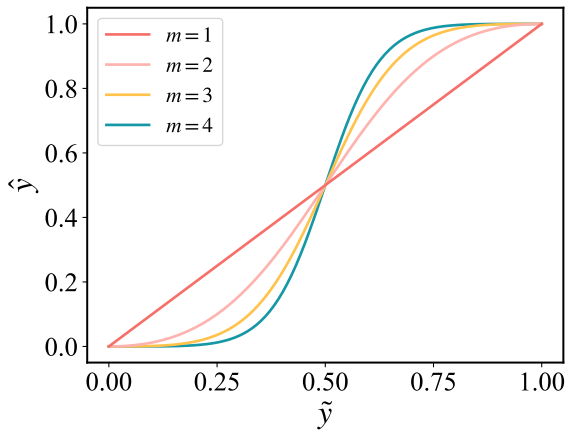


Figure 3: The relationship between $\hat{y}$ and $\tilde{y}$ with different m. As the pseudo-label value $\tilde{y}$ approach one or zero, the enhanced pseudo-label $\hat{y}$ become increasingly confident.

Given the pseudo-labels $\hat{\boldsymbol{y}}_i$ and evidences $(\boldsymbol{\alpha}_i^n, \boldsymbol{\beta}_i^n)$ predicted by the noisy head, we define the unlabeled loss for the i-th unlabeled instance as

$$\mathcal{L}(\boldsymbol{\alpha}^n, \boldsymbol{\beta}^n, \hat{\boldsymbol{y}}) = \sum_{j=1}^{K} \hat{y}_j \mathcal{L}^+_{\mathrm{ABL}}(\alpha_j^n, \beta_j^n) + (1-\hat{y}_j)\mathcal{L}^-_{\mathrm{ABL}}(\alpha_j^n, \beta_j^n). \quad (10)$$

To alleviate the harmfulness of OOD examples, we down-weight their losses based on the uncertainty measurement, with the goal of implicitly highlight the contributions of ID examples. Specifically, for each unlabeled instance $\boldsymbol{x}_i$, given that its model outputs following a Beta distribution, we compute its uncertainty as $u_i = \frac{2K}{\sum_{j=1}^{K} \alpha_j}$. A higher uncertainty, a large probability of an unlabeled instance to be an OOD. We use the certainty $1 - u_i$ as the weight to alleviate the harmfulness of OOD examples. Formally, we define the dual-measurement weighted unlabeled loss as

$$\mathcal{L}_{ub} = \sum_{i=1}^{N_{ub}} (1 - \hat{u}_i)\mathcal{L}(\boldsymbol{\alpha}_i^n, \boldsymbol{\beta}_i^n, \hat{\boldsymbol{y}}_i), \quad (11)$$

where $\hat{u}_i = \frac{u_i}{\max_{i' \in [N_{ub}]} u_{i'}}$ is the normalized version of uncertainty u_i.

Given the label vector $\boldsymbol{y}_i$ and evidences $(\boldsymbol{\alpha}_i^c, \boldsymbol{\beta}_i^c)$, $(\boldsymbol{\alpha}_i^n, \boldsymbol{\beta}_i^n)$ predicted by clean and noisy head respectively, we define the labeled loss as

$$\mathcal{L}_{lb} = \sum_{i=1}^{N_{lb}} \mathcal{L}_{\mathrm{ABL}}(\boldsymbol{\alpha}_i^c, \boldsymbol{\beta}_i^c, \boldsymbol{y}_i) + \mathcal{L}_{\mathrm{ABL}}(\boldsymbol{\alpha}_i^n, \boldsymbol{\beta}_i^n, \boldsymbol{y}_i). \quad (12)$$

Finally, we define the overall loss function as

$$\mathcal{L} = \lambda \mathcal{L}_{lb} + (1 - \lambda)\mathcal{L}_{ub}. \quad (13)$$

4 Experiments

In this section, we conduct experiments to validate the effectiveness of the proposed method. Subsequently, we perform ablation studies to assess the contribution of each component within our method.

4.1 Experimental Settings

Datasets. To evaluate the proposed method, we perform experiments on multiple multi-label benchmark datasets, including Pascal VOC-2012 (VOC for short) [2] [7], MS-COCO-2014 (COCO for short) [3] [20], and NUS-WIDE (NUS for short) [4] [3]. There are 5,717 and 82,081 images in VOC and COCO respectively, while NUS consists of 150,000 examples. We adopt two methods to generate labeled examples: 1) sampling a fixed proportion p of examples from the ID data as labeled data; 2) taking a fixed number of examples as labeled data.

To construct OOD-corrupted datasets, following the previous work [13], we manually select a subset of ImageNet[5, 22] as OOD data. Specifically, for VOC and COCO, we use 20 OOD classes from ImageNet-21K identical to [13]. These classes have no overlap with ImageNet-1K, VOC, nor COCO. For NUS, we select another 20 classes from ImageNet-21K based on the same principle as the previous work. Notably, these 20 classes exclude high-level concepts such as animals, plants, and flowers in NUS. The OOD dataset for VOC and COCO comprises 17,635 samples while the dataset for NUS contains 30,212 samples. The remaining ID examples and OOD examples consist of the unlabeled dataset.

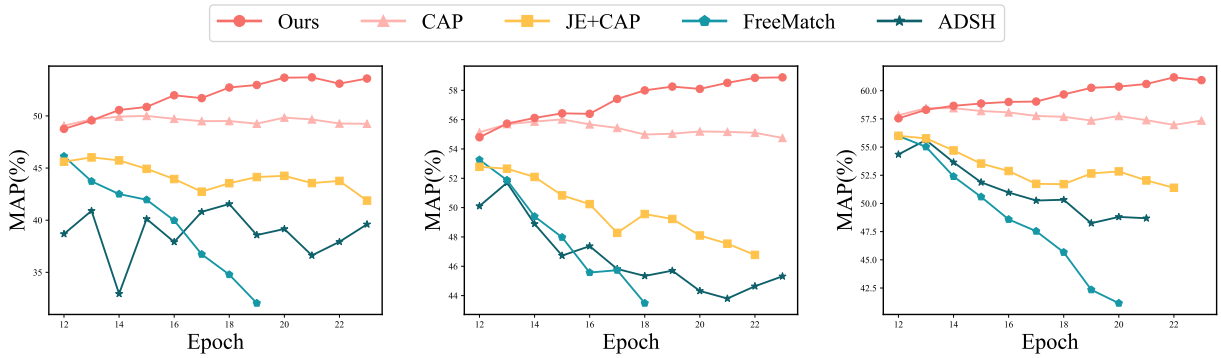
In contrast to conventional semi-supervised learning, where datasets are typically divided into two segments, *i.e.*, labeled data and unlabeled data, our OOD-corrupted datasets are partitioned into labeled data, unlabeled ID data, and unlabeled OOD data. The proportion of these three partitions will affect the experimental outcomes. Therefore, we conduct experiments with two different settings. Firstly, we make use of the full datasets and adjust the labeled rate. The labeled rate here means the number of labeled instance dividing the number of unlabeled ID instances. The evaluated labeled rate in COCO and NUS is $p \in \{0.01, 0.02, 0.03, 0.05, 0.1, 0.15, 0.2\}$, while $p \in \{0.05, 0.1, 0.15, 0.2\}$ in VOC. Secondly, considering the former experiments can not reflect the influence of the scale of OOD datasets, we perform another experimental setting by altering the ratio of unlabeled ID samples to unlabeled OOD samples (ID rate). In implementation, we alter ID rate by increasing or decreasing the number of unlabeled samples and keep the OOD datasets unchanged. The ID rates tested are $q \in \{0.5, 1, 2, 4\}$. Each ID rate is combined with 4 fixed numbers of labeled samples (1000, 2000, 3000, 4000 for COCO and 2000, 4000, 6000, 8000 for NUS). More details about our dataset settings can be found in the Appendix.

Comparing Methods. To validate the effectiveness of the proposed method, we compare our method with two advanced semi-supervised methods, a SSMLL method and an intuitive method for OOD-corrupted SSMLL setting. FreeMatch [34] and ADSH [8] are two state-of-the-art SSL methods. CAP [38] is a state-of-the-art SSMLL method. As there is no end-to-end framework available for our targeted problem setting to the best of our knowledge, we construct a method by combining the basic SSMLL algorithm with OOD detection methods. Joint Energy [32] is an advanced OOD detection method for multi-label classification by aggregating

[2] http://host.robots.ox.ac.uk/pascal/VOC/
[3] https://cocodataset.org
[4] https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html

Table 1: Comparison results on COCO and NUS in terms of mAP (%) with different labeled rate p.

Dataset	COCO					NUS				
Method	FreeMatch	ADSH	CAP+JE	CAP	Ours	FreeMatch	ADSH	CAP+JE	CAP	Ours
$p = 0.01$	47.46	48.93	48.65	51.72	**55.15**	32.62	31.87	34.89	36.23	**37.56**
$p = 0.02$	54.44	55.29	55.06	57.55	**60.38**	37.47	37.44	38.74	40.83	**41.72**
$p = 0.03$	57.38	58.06	57.63	60.12	**62.76**	40.02	40.53	41.60	43.24	**43.72**
$p = 0.05$	60.56	61.49	60.75	62.91	**65.19**	42.49	43.40	43.81	45.77	**46.14**
$p = 0.10$	63.97	64.90	65.67	66.96	**68.28**	45.07	45.85	46.47	48.10	**48.25**
$p = 0.15$	65.71	66.70	67.87	68.90	**69.72**	46.30	47.42	47.91	**49.53**	49.20
$p = 0.20$	67.00	67.73	69.11	70.02	**70.79**	47.14	48.37	48.72	**50.46**	50.15



(a) $p = 0.01$ **(b)** $p = 0.02$ **(c)** $p = 0.03$

Figure 4: Illustration of model's precision throughout semi-supervised training process on COCO. Only our method exhibits a consistent improvement, while other methods either continues to decrease or fluctuates during semi-supervised learning stage.

Table 2: Comparison results on VOC in terms of mAP (%) with different labeled rate p.

Method	FreeMatch	ADSH	CAP+JE	CAP	Ours
$p = 0.05$	75.22	75.01	74.72	73.23	**77.08**
$p = 0.10$	81.03	81.03	80.52	79.92	**81.30**
$p = 0.15$	83.01	82.42	82.72	82.11	**82.98**
$p = 0.20$	83.82	83.15	83.48	83.06	**84.06**

label-wise energy scores from multiple labels. We use the Gaussian Mixture Model to decide whether the unlabeled data is OOD according to the joint energy score, and then apply CAP algorithm on the predicted ID data.

Implementation. For each method, we employ ResNet-50 [11] pre-trained on ImageNet as the backbone. We adopt RandAugment [4] and Cutout [6] for data augmentation. We employ the AdamW optimizer [21] and the one-cycle policy scheduler [28] to train the model and the maximum learning rate is 0.0001. The warm-up epoch is set to 12, and the maximum training epoch is 40. The batch size is set to 64 for all datasets. In our method, hyper-parameters are set as $\gamma_+ = 0$, $\gamma_- = 4$, $c_{lb} = 0.2$, $c_{ub} = 0.05$, $m = 2$, $\lambda = 0.5$. For CAP, we employ Asymmetric Loss [23] as loss function. We also perform an Exponential Moving Average (EMA) for the model parameter with a decay of 0.9997. The random seed is set to 1 for all experiments.

Evaluation. We evaluate our model in two aspects, multi-label classification ability among ID class space and OOD recognition ability among unlabeled dataset. For the multi-label classification problem, we take mAP (mean average precision) of the ID label space as the metric. For OOD recognition, we adopt AUROC (area under receiver operating characteristic curve) and report the best result during the training process.

4.2 Empirical Results

Table 1 and Table 2 show the experiment results of the metric mAP on three datasets COCO, NUS and VOC. From the table we can find that our method has achieved best performance in most cases and has an obvious advantage especially when the labeled rate is low. Comparing CAP with its modified version (CAP+JE), the modification that excludes OOD instances decreases its performance unexpectedly, although Joint Energy is a remarkable OOD detection method. It reveals the fact that sometimes no detector is better than bad detectors. Moreover, our method outperforms CAP in almost every test case. The reason lies in CAP assigning pseudo labels based on the assumption that labeled and unlabeled sets share the same distribution; however, this assumption is violated by OOD data. NUS is the largest dataset and its OOD proportion is relatively

Table 3: Comparison results on COCO and NUS in terms of mAP (%) with different ID rate q.

ID Rate	Labeled Counts	COCO					Labeled Counts	NUS				
		FreeMatch	ADSH	CAP+JE	CAP	Ours		FreeMatch	ADSH	CAP+JE	CAP	Ours
$q = 0.5$	1000	50.62	49.93	50.70	50.59	54.31	2000	35.76	34.67	35.32	36.22	36.40
	2000	56.58	56.59	56.74	56.42	58.83	4000	38.45	40.60	39.23	41.07	41.53
	3000	59.27	59.40	59.67	59.64	60.88	6000	41.17	42.42	42.35	42.84	44.00
	4000	60.78	61.17	61.10	59.87	62.72	8000	43.13	44.09	44.03	44.49	45.21
$q = 1$	1000	50.62	49.85	50.53	51.82	55.45	2000	35.61	34.54	36.41	36.96	37.73
	2000	56.90	57.04	56.80	57.37	60.07	4000	39.67	40.54	41.22	41.88	41.95
	3000	59.37	59.74	59.44	60.26	61.95	6000	41.14	42.42	42.31	43.32	44.13
	4000	60.84	61.21	60.96	61.84	63.31	8000	43.07	44.17	43.81	44.94	45.28
$q = 2$	1000	50.08	50.51	50.20	52.47	56.45	2000	35.03	35.25	36.14	37.88	38.04
	2000	56.64	56.95	56.26	58.30	60.80	4000	39.78	40.50	40.77	42.40	42.60
	3000	59.34	59.65	59.22	60.72	62.70	6000	41.02	42.43	42.11	43.87	44.53
	4000	60.80	61.31	60.86	62.21	63.92	8000	42.97	44.14	43.91	45.53	45.77
$q = 4$	1000	49.49	50.51	49.63	53.35	57.03	2000	34.90	34.65	36.16	38.26	39.39
	2000	56.20	56.71	56.10	58.90	61.39	4000	39.56	40.16	40.70	42.55	43.27
	3000	58.86	59.52	58.87	61.44	63.35	6000	40.84	42.21	42.18	44.21	45.10
	4000	60.65	61.26	60.58	62.68	64.81	8000	42.83	43.94	43.50	45.99	46.16

small, which may cause our method to have the worst comparison results on it.

We visualize the trend of the model's precision change during the semi-supervised learning phase on COCO in Figure 4. Only our method exhibits a consistent improvement, while other methods either continue to decrease or fluctuate during the semi-supervised learning phase.

The experimental results in Table 3 show comparison results with different ID rate. As the ID rate increases, *i.e.*, as more ID samples are included in the unlabeled dataset, our method maintains a substantial advantage over other approaches. As the amount of unlabeled data increases, our method demonstrates improved performance, whereas some methods (FreeMatch and CAP+JE) remain unchanged or even exhibit a decline in their performance. This observation indicates that our method effectively leverages the unlabeled ID samples to a greater extent. From Table **??** we obtain that our method shows more superiority than other methods when OOD data constitutes the main part of the unlabeled dataset. Besides, when the OOD rate is low, our method still outperforms other methods. In fact, further experiments illustrate that our method can even be applied to semi-supervised learning without OOD data.

In Table 4, we report the OOD detection performance with the metric AUROC. We can see that our model discriminates OOD samples better than Joint Energy. Comparing the OOD detection results with the semi-supervised learning results, we find that the OOD detection ability is independent of the classification ability and the scale of labeled data. This suggests that our method is a suitable choice for OOD detection when labeled data is scarce.

4.3 Ablation Study and Discussion

To further explore the effectiveness of the proposed method, we perform extensive ablation studies to validate the effectiveness of each component in our method.

Soft Pseudo Label. Our method adopts soft pseudo labels rather than hard pseudo labels. To verify its effectiveness, we compare it with the hard pseudo-labeling method. In experiments, we take $\mathbb{E}(p) = 0.5$ as the threshold to distinguish OOD data from ID data. From Table 6 we find this modification leads to a decrease in accuracy. This phenomenon occurs because thresholds amplify errors when OOD or negative instances are predicted with high confidence labels. There are two main advantages to using soft pseudo-labels: 1) We do not need to determine thresholds. In practice, this is a challenging task, as optimal thresholds vary across classes and datasets. 2) We can use all examples for training. Existing SSL work often uses a threshold to filter out a subset of unreliable pseudo-labels, resulting in underutilization of unlabeled examples. By using soft pseudo-labels, we can not only fully utilize all unlabeled examples, but also adaptively adjust the contribution of different examples to the learning process.

Dual-head Classifier. In our method, the dual-head classifier has two effects. On the one hand, it separates clean data from noisy data, making the clean head more accurate on the OOD detection task. On the other hand, it helps to generate pseudo labels by multiplication, reducing the probability of misclassifying negative instances as positive ones, which is dangerous in our setting. For comparison, we try to remove one classifier and the experimental results is in Table 6. The result indicates that the dual-head structure is the most important component of our framework compared to other ablation experiments.

Uncertainty Weighting. Instead of definitively separating OOD data from unlabeled data, the proposed method uses a flexible way to exclude the influence of OOD data. In the compared method, we take the instances with the highest $k\%$ uncertainty score as OOD data and the rest instances as ID data, where $k\%$ is the true OOD rate in the unlabeled dataset. Figure 5 shows the comparison result

Table 4: Comparison results of OOD detection performance in terms of AUROC (%).

Dataset	Method	$p = 0.01$	$p = 0.02$	$p = 0.03$	$p = 0.05$	$p = 0.1$	$p = 0.15$	$p = 0.2$
VOC	CAP+JE	51.61	59.57	80.39	79.66	80.15	81.26	83.31
	Ours	**79.84**	**81.73**	**89.09**	**90.24**	**89.96**	**91.57**	**90.77**
COCO	CAP+JE	81.70	80.54	80.95	83.62	82.52	83.48	85.29
	Ours	**89.48**	**87.66**	**88.21**	**88.79**	**90.01**	**89.41**	**90.88**
NUS	CAP+JE	77.66	83.15	84.18	86.30	**87.14**	**87.88**	**87.71**
	Ours	**86.05**	**86.29**	**86.03**	**88.06**	86.14	84.94	86.19

Table 5: Comparison results of SSMLL performance on COCO and NUS in terms of mAP(%).

Dataset	Method	$p = 0.01$	$p = 0.02$	$p = 0.03$	$p = 0.05$	$p = 0.1$	$p = 0.15$	$p = 0.2$
COCO	CAP	52.61	58.19	60.75	63.52	67.48	69.30	70.41
	Ours	**55.64**	**60.83**	**63.10**	**65.68**	**68.59**	**70.08**	**71.05**
NUS	CAP	32.24	37.88	40.76	44.30	**48.63**	**50.14**	**50.85**
	Ours	**36.36**	**41.52**	**44.15**	**45.83**	48.19	49.25	50.38

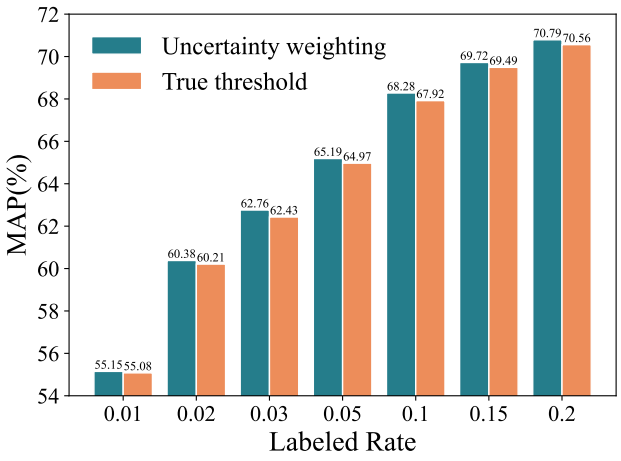


Figure 5: Comparison results of the proposed method on COCO with certainty-based weighting and true threshold. Despite the unfairness in the comparison due to the impracticality of accessing true thresholds in practice, our method still achieves similar or even superior performance.

of the proposed method under uncertainty-based weighting and true thresholds. Although it is an unfair comparison, our method still achieves similar or even better performance. This demonstrates the effectiveness of our uncertainty weighting method.

Asymmetric Beta Loss. We compare Asymmetric Beta Loss with the basic EBCE loss function in Eq. (5). Results are shown in Table 6 and it proves that Asymmetric Beta Loss is an advanced choice for multi-label evidential learning since it leverages the

Application in SSMLL. Although the proposed method is designed for OOD-corrupted SSMLL scenarios, it can also be a general framework for other settings such as SSMLL. We compare the

Table 6: Ablation results of different components in the proposed method on COCO.

Method	$p = 0.05$	$p = 0.1$	$p = 0.15$	$p = 0.2$
w/o soft pseudo label	64.54	67.67	69.25	70.44
w/o dual-head classifier	61.59	65.52	67.38	68.91
w/o Asymmetric Beta Loss	64.00	67.25	68.76	69.78
Ours	**65.19**	**68.28**	**69.72**	**70.79**

SSMLL performance of our model with the state-of-the-art SSMLL method CAP. Results in Table 5 show that our method also achieves good performance in the SSMLL task and is significantly better than CAP when labeled samples are scarce. Exploring the application of our method in other settings and its underlying mechanism is crucial in our future research.

5 Related Work

5.1 Semi-Supervised Learning

Recent years have witnessed the great development of semi-supervised learning [1]. Pseudo-labeling [18] and consistency regularization [17, 26, 31] are the most popular methods to utilize unlabeled data. Pseudo-labeling assigns model's predictions as pseudo labels to unlabeled data and this augmented dataset is used for further training to improve model performance. Consistency regularization suggests that a neural network should be invariant when confronted with different perturbations of the same instance. FixMatch [29] is a famous work that combines the two techniques in a concise framework. Many subsequent works [8, 34, 40, 42, 44] are based on its paradigm with locality improvement. Despite the significant successes of pseudo-labeling, it is inherently tied to the closed-set assumption and cannot be directly applied to the open world. This phenomenon occurs because noisy pseudo-labeled outliers can diminish the performance of self-supervised training.

5.2 Open-Set Semi-Supervised Learning

Open-set semi-supervised learning (OSSL) [9] is a special semi-supervised multi-class classification problem where the unlabeled set contains other classes different from the labeled set. Sometimes it should recognize the unknown class during the inference stage, which is the task of open-set recognition. Early studies on OSSL follow the basic strategy of utilizing only sufficiently confident ID samples within traditional SSL schemes [2, 12, 41]. For example, MTC [41] employs a binary classifier to predict the likelihood of a given instance belonging to the outlier category and updates the network parameters and the outlier score alternately. OpenMatch [24] uses an one-vs-all (OVA) [25] classifier as its outlier detector and introduces the application of soft consistency regularization to the outlier detector. However, this detect-and-exclude strategy will encounter challenges when labeled data is limited or the detector is not accurate enough. Many experiments have found that an unreliable outlier detector is more harmful than the outliers themselves. Aware of this phenomenon, some works start to make use of OOD data rather than considering them as merely negative noise. IOMatch [19] employs a multi-binary classifier and a standard closed-set classifier to generate unified open-set classification targets, which regard all outliers as a single new class. Taking these targets as open-set pseudo-labels, it can utilize both inliers and outliers. HOOD [15] divides OOD data into benign ones and malign ones and identify them through content and style from each image. Benign OOD data helps to train the closed-set classifier while malign OOD data helps to deceive anomalies.

5.3 Semi-Supervised Multi-label Learning

Semi-supervised multi-label learning (SSMLL) is studying about the semi-supervised method for multi-label classification problem [10]. Although there are many different settings in MLL area, *e.g.*, partial multi-label learning [35–37], multi-label learning with missing label [30, 39], there are few works studying on the original SSMLL setting in recent years. DRML [33] takes feature-label and label-label relations into account simultaneously with dual-classifier domain adaptation strategy. PercentMatch [14] proposes a dynamic threshold adjusting method and unlabeled loss weights as an extension of FixMatch to SSMLL. CAP [38] employs a class-aware approach to determine the threshold in different classes. Since these are all traditional SSL methods and rely heavily on the same distribution assumption, these SSMLL works are unable to deal with the SSMLL setting that involves unlabeled OOD samples.

6 Conclusion

In this paper, we focus on the difficulties of the SSMLL problem when OOD samples are involved in unlabeled data and propose a unified framework to solve this problem. Our framework utilizes the theory of evidential deep learning to detect OOD data and an adaptive weight to implicitly weaken them. A dual-head structure enables the model to perform both OOD detection and multi-label classification simultaneously. Considering the property of multi-label data, we improve the conventional loss by introducing the asymmetric beta loss. Experiments have demonstrated the effectiveness of each component of our contribution and underscore the strong potential for its application to other problems.

7 ACKNOWLEDGMENTS

This work was supported by the National Science and Technology Major Project (2020AAA0107000), the Natural Science Foundation of Jiangsu Province of China (BK20211517, BK20222012), and NSFC (62222605).

References

[1] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. 2006. *Semi-Supervised Learning*. The MIT Press. https://doi.org/10.7551/mitpress/9780262033589.001.0001

[2] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. 2020. Semi-Supervised Learning under Class Distribution Mismatch. In *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:211526565

[3] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval* (Santorini, Fira, Greece) *(CIVR '09)*. Association for Computing Machinery, New York, NY, USA, Article 48, 9 pages. https://doi.org/10.1145/1646396.1646452

[4] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. 2019. RandAugment: Practical automated data augmentation with a reduced search space. arXiv:1909.13719 [cs.CV]

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[6] Terrance DeVries and Graham W. Taylor. 2017. Improved Regularization of Convolutional Neural Networks with Cutout. arXiv:1708.04552 [cs.CV]

[7] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2015. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision* 111, 1 (01 Jan 2015), 98–136. https://doi.org/10.1007/s11263-014-0733-5

[8] Lan-Zhe Guo and Yu-Feng Li. 2022. Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 8082–8094. https://proceedings.mlr.press/v162/guo22e.html

[9] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. 2020. Safe Deep Semi-Supervised Learning for Unseen-Class Unlabeled Data. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 3897–3906. https://proceedings.mlr.press/v119/guo20i.html

[10] Meng Han, Hongxin Wu, Zhiqiang Chen, Muhang Li, and Xilong Zhang. 2023. A survey of multi-label classification based on supervised and semi-supervised learning. *International Journal of Machine Learning and Cybernetics* 14, 3 (01 Mar 2023), 697–724. https://doi.org/10.1007/s13042-022-01658-9

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[12] Rundong He, Zhongyi Han, Xiankai Lu, and Yilong Yin. 2022. Safe-Student for Safe Deep Semi-Supervised Learning with Unseen-Class Unlabeled Data. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14565–14574. https://doi.org/10.1109/CVPR52688.2022.01418

[13] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. 2022. Scaling Out-of-Distribution Detection for Real-World Settings. arXiv:1911.11132 [cs.CV]

[14] Junxiang Huang, Alexander Huang, Beatriz C. Guerra, and Yen-Yun Yu. 2022. PercentMatch: Percentile-based Dynamic Thresholding for Multi-Label Semi-Supervised Classification. arXiv:2208.13946 [cs.CV]

[15] Zhuo Huang, Xiaobo Xia, Li Shen, Bo Han, Mingming Gong, Chen Gong, and Tongliang Liu. 2023. Harnessing Out-Of-Distribution Examples via Augmenting Content and Style. arXiv:2207.03162 [cs.LG]

[16] Audun Jøsang. 2016. *Subjective Logic: A Formalism for Reasoning Under Uncertainty* (1st ed.). Springer Publishing Company, Incorporated.

[17] Samuli Laine and Timo Aila. 2017. Temporal Ensembling for Semi-Supervised Learning. arXiv:1610.02242 [cs.NE]

[18] Dong-Hyun Lee. 2013. Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. https://api.semanticscholar.org/CorpusID:18507866

[19] Zekun Li, Lei Qi, Yinghuan Shi, and Yang Gao. 2023. IOMatch: Simplifying Open-Set Semi-Supervised Learning with Joint Inliers and Outliers Utilization. arXiv:2308.13168 [cs.CV]

[20] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *CoRR* abs/1405.0312 (2014). arXiv:1405.0312 http://arxiv.org/abs/1405.0312

[21] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs.LG]
[22] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. 2021. ImageNet-21K Pretraining for the Masses. arXiv:2104.10972 [cs.CV]
[23] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric Loss For Multi-Label Classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 82–91. https://doi.org/10.1109/ICCV48922.2021.00015
[24] Kuniaki Saito, Donghyun Kim, and Kate Saenko. 2021. OpenMatch: Open-set Consistency Regularization for Semi-supervised Learning with Outliers. arXiv:2105.14148 [cs.CV]
[25] Kuniaki Saito and Kate Saenko. 2021. OVANet: One-vs-All Network for Universal Domain Adaptation. arXiv:2104.03344 [cs.CV]
[26] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/30ef30b64204a3088a26bc2e6ecf7602-Paper.pdf
[27] Murat Sensoy, Melih Kandemir, and Lance M. Kaplan. 2018. Evidential Deep Learning to Quantify Classification Uncertainty. *CoRR* abs/1806.01768 (2018). arXiv:1806.01768 http://arxiv.org/abs/1806.01768
[28] Leslie N. Smith and Nicholay Topin. 2018. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. arXiv:1708.07120 [cs.LG]
[29] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. arXiv:2001.07685 [cs.LG]
[30] Anhui Tan, Jiye Liang, Wei-Zhi Wu, and Jia Zhang. 2022. Semi-supervised partial multi-label classification via consistency learning. *Pattern Recognition* 131 (2022), 108839. https://doi.org/10.1016/j.patcog.2022.108839
[31] Antti Tarvainen and Harri Valpola. 2018. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv:1703.01780 [cs.NE]
[32] Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. 2021. Can multi-label classification networks know what they don't know? arXiv:2109.14162 [cs.LG]
[33] Lichen Wang, Yunyu Liu, Can Qin, Gan Sun, and Yun Fu. 2020. Dual Relation Semi-Supervised Multi-Label Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 04 (Apr. 2020), 6227–6234. https://doi.org/10.1609/aaai.v34i04.6089
[34] Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. 2023. FreeMatch: Self-adaptive Thresholding for Semi-supervised Learning. arXiv:2205.07246 [cs.LG]
[35] Ming-Kun Xie and Sheng-Jun Huang. 2018. Partial Multi-Label Learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 4302–4309. https://doi.org/10.1609/AAAI.V32I1.11644
[36] Ming-Kun Xie and Sheng-Jun Huang. 2022. Partial Multi-Label Learning With Noisy Label Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 7 (2022), 3676–3687. https://doi.org/10.1109/TPAMI.2021.3059290
[37] Ming-Kun Xie and Sheng-Jun Huang. 2024. Multi-label learning with pairwise relevance ordering. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*. Curran Associates Inc., Red Hook, NY, USA, Article 1803, 12 pages.
[38] Ming-Kun Xie, Jia-Hao Xiao, Hao-Zhe Liu, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. 2023. Class-Distribution-Aware Pseudo Labeling for Semi-Supervised Multi-Label Learning. arXiv:2305.02795 [cs.LG]
[39] Zexian Xie, Peipei Li, Jinling Jiang, and Xindong Wu. 2023. Semi-supervised Multi-Label Learning with Missing Labels via Correlation Information. In *2023 International Joint Conference on Neural Networks (IJCNN)*. 01–08. https://doi.org/10.1109/IJCNN54540.2023.10191722
[40] Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. 2021. Dash: Semi-Supervised Learning with Dynamic Thresholding. arXiv:2109.00650 [cs.LG]
[41] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. 2020. Multi-Task Curriculum Framework for Open-Set Semi-Supervised Learning. arXiv:2007.11330 [cs.CV]
[42] Bowen Zhang, Yidong Wang, Wenxin Hou, HAO WU, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. 2021. FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 18408–18419. https://proceedings.neurips.cc/paper_files/paper/2021/file/995693c15f439e3d189b06e89d145dd5-Paper.pdf
[43] Dell Zhang and Bilyana Taneva-Popova. 2023. A Theoretical Analysis of Out-of-Distribution Detection in Multi-Label Classification. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval* (Taipei, Taiwan) *(ICTIR '23)*. Association for Computing Machinery, New York, NY, USA, 275–282. https://doi.org/10.1145/3578337.3605116
[44] Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. 2022. SimMatch: Semi-Supervised Learning With Similarity Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14471–14481.

A Further Ablation Studies

Ablation of Hyper-parameters. $\gamma^+, \gamma^-, c_{lb}, c_{ub}$ are hyper-parameters inherited from Asymmetric Loss. We set them as default values given in the original paper. Table 7 and Table 8 report the ablation studies on the parameters m and λ. These experiments were conducted on COCO with a labeled ratio of $p = 0.05$. From Table 7, we observe a significant improvement when m changes from 1 to 2. However, there is no significant change in the results when m exceeds 2. From Table 8, we find that the method achieves its best performance with $\lambda = 0.5$, indicating an equal treatment of labeled and unlabeled loss.

Table 7: Ablation results of different m.

m	1	2	3	4
mAP	64.29	65.19	65.16	65.25

Table 8: Ablation results of different λ.

lambda	0.1	0.3	0.5	0.7
mAP	62.79	64.32	65.19	63.42

Other baselines. The comparison methods shown in the main body are the standard methods in the close domain of our studying problems. There are also other potential comparison methods, and we report their performance based on our experimental setup on COCO in Table 9. DRML and PercentMatch are SSMLL methods, as is CAP. We observe that CAP outperforms DRML and PercentMatch with large margin so that we select CAP as the default SSMLL method. We also compare the combination of CAP with different OOD detectors. We observe that CAP with maximum energy outperforms CAP with Joint Energy when the labeled rate is low. However, the situation is reversed when the labeled rate is high. We choose CAP with Joint Energy in the main content because Joint Energy is the OOD detecting metric tailored for multi-label scenarios.

B Dataset details

The experiments presented in Table **??** and Table **??** investigate the impact of different ratios of ID and OOD examples on model performance. Considering that the number of OOD examples is fixed (see Section 4.1), we adjust the ratio between OOD and ID by changing the number of unlabeled ID examples. To make it clearer, we report the numbers of labeled ID, unlabeled ID, and OOD examples under different ID ratios in Table 10 and Table 11.

Table 9: Comparing results on COCO in terms of mAP(%).

Method	p=0.05	p=0.1	p=0.15	p=0.2
DRML	41.98	54.18	57.63	59.13
PercentMatch	57.86	62.91	65.54	67.38
CAP	62.91	66.96	68.90	70.02
CAP+maximum softmax	60.92	64.80	66.53	68.12
CAP+maximum energy	61.39	65.58	67.27	68.76
CAP+Joint Energy	60.75	65.67	67.87	69.11

Table 10: The numbers of labeled ID, unlabeled ID and OOD examples under different ID ratios on COCO.

ID rate	Labeled Counts	labeled ID : unlabeled ID : OOD
$q = 0.5$	1000	1000:8817:17635
$q = 0.5$	2000	2000:8817:17635
$q = 0.5$	3000	3000:8817:17635
$q = 0.5$	4000	4000:8817:17635
$q = 1$	1000	1000:17635:17635
$q = 1$	2000	2000:17635:17635
$q = 1$	3000	3000:17635:17635
$q = 1$	4000	4000:17635:17635
$q = 2$	1000	1000:35270:17635
$q = 2$	2000	2000:35270:17635
$q = 2$	3000	3000:35270:17635
$q = 2$	4000	4000:35270:17635
$q = 4$	1000	1000:70540:17635
$q = 4$	2000	2000:70540:17635
$q = 4$	3000	3000:70540:17635
$q = 4$	4000	4000:70540:17635

Table 11: The numbers of labeled ID, unlabeled ID and OOD examples under different ID ratios on NUS.

ID rate	Labeled Counts	labeled ID : unlabeled ID : OOD
$q = 0.5$	2000	2000:15106:30212
$q = 0.5$	4000	4000:15106:30212
$q = 0.5$	6000	6000:15106:30212
$q = 0.5$	8000	8000:15106:30212
$q = 1$	2000	2000:30212:30212
$q = 1$	4000	4000:30212:30212
$q = 1$	6000	6000:30212:30212
$q = 1$	8000	8000:30212:30212
$q = 2$	2000	2000:60424:30212
$q = 2$	4000	4000:60424:30212
$q = 2$	6000	6000:60424:30212
$q = 2$	8000	8000:60424:30212
$q = 4$	2000	2000:120848:30212
$q = 4$	4000	4000:120848:30212
$q = 4$	6000	6000:120848:30212
$q = 4$	8000	8000:120848:30212

C Derivation of Eq. (7)

Since p is a random variable follows $Beta(p|\alpha, \beta)$, its probability dense function is

$$Beta(p|\alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)},$$

where

$$B(\alpha, \beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}dp = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

and Γ represents gamma function and it has property

$$\Gamma(z+1) = z\Gamma(z).$$

Hence,

$$B(\alpha, \beta+\gamma) = \prod_{r=0}^{\gamma-1}\left(\frac{\beta+r}{\alpha+\beta+r}\right) B(\alpha, \beta).$$

Besides,

$$\psi(z) = \frac{\Gamma'(z)}{\Gamma(z)} = \frac{d\ ln\Gamma(z)}{dz}.$$

As a result, when $y = 1$,

$$\begin{aligned}
\mathcal{L}_{\text{ABL}} &= -\mathbb{E}\left[(1-p)^{\gamma} lnp\right] \\
&= -\int_0^1 (1-p)^{\gamma}\ lnp\ \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}dp \\
&= -\frac{1}{B(\alpha, \beta)}\int_0^1 \frac{\partial[p^{\alpha-1}(1-p)^{\beta+\gamma-1}]}{\partial\alpha}dp \\
&= -\frac{1}{B(\alpha, \beta)}\frac{\partial\int_0^1 p^{\alpha-1}(1-p)^{\beta+\gamma-1}dp}{\partial\alpha} \\
&= -\frac{1}{B(\alpha, \beta)}\frac{\partial B(\alpha, \beta+\gamma)}{\partial\alpha} \\
&= -\prod_{r=0}^{\gamma-1}\left(\frac{\beta+r}{\alpha+\beta+r}\right)\frac{1}{B(\alpha, \beta+k)}\frac{\partial B(\alpha, \beta+\gamma)}{\partial\alpha} \\
&= -\prod_{r=0}^{\gamma-1}\left(\frac{\beta+r}{\alpha+\beta+r}\right)\frac{\partial\ lnB(\alpha, \beta+\gamma)}{\partial\alpha} \\
&= -\prod_{r=0}^{\gamma-1}\left(\frac{\beta+r}{\alpha+\beta+r}\right)\left(\frac{\partial\ ln\Gamma(\alpha)}{\partial\alpha} - \frac{\partial\ ln\Gamma(\alpha+\beta+\gamma)}{\partial\alpha}\right) \\
&= \prod_{r=0}^{\gamma-1}\left(\frac{\beta+r}{\alpha+\beta+r}\right)\left(\psi(\alpha+\beta+\gamma) - \psi(\alpha)\right).
\end{aligned}$$

It is noted that here we use γ to represent γ^+ for simplicity.

Similarly, when $y = 0$,

$$\begin{aligned}
\mathcal{L}_{\mathrm{ABL}} &= -\mathbb{E}\left[p^{\gamma} ln(1-p)\right] \\
&= -\int_0^1 p^{\gamma}\ ln(1-p)\ \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} dp \\
&= -\frac{1}{B(\alpha,\beta)} \frac{\partial \int_0^1 p^{\alpha-1}(1-p)^{\beta+\gamma-1} dp}{\partial \alpha} \\
&= -\frac{1}{B(\alpha,\beta)} \frac{\partial B(\alpha+\gamma,\beta)}{\partial \alpha} \\
&= -\prod_{r=0}^{\gamma-1} \left(\frac{\alpha+r}{\alpha+\beta+r}\right) \frac{\partial\, lnB(\alpha+\gamma,\beta)}{\partial \alpha} \\
&= \prod_{r=0}^{\gamma-1} \left(\frac{\alpha+r}{\alpha+\beta+r}\right) \big(\psi(\alpha+\beta+\gamma) - \psi(\beta)\big).
\end{aligned}$$

Here, α denotes α_c, and λ represents λ^- for simplicity.