Calibrated Disambiguation for Partial Multi-label Learning

Zhuoming Li [1,2], **Yuheng Jia** [1,2*], **Mi Yu** [2], **Zicong Miao** [2]

[1] School of Computer Science and Engineering, Southeast University
[2] China Telecom Cloud Computing Corporation, Beijing 100088, China
{leezhuoming, yhjia}@seu.edu.cn; {yumi2.cq, miaozc}@chinatelecom.cn

Abstract

Partial multi-label learning (PML) aims to train a classifier on dataset whose instances are over-annotated with not only relevant labels but also irrelevant labels, which is common when datasets are collected from crowd-sourcing platform. Existing works primarily approach it from a curriculum learning perspective, leveraging the memorization effect to disambiguate noisy labels and produce robust predictions. However, these methods are based on non-adaptive weighting functions and lack theoretical guidance for optimal weighting. To overcome these issues, a calibrated disambiguation model named PML-CD is proposed. We firstly formulate the optimal weighting function for curriculum-based disambiguation, which is equivalent to the calibration of the model's predicted confidences, thus provide a guidance for curriculum designing. To obtain the optimal weighting function from PML dataset during the training, a transferable calibrator is designed, which takes the histogram of positive samples' confidences as input, and outputs the optimal curriculum weighting for training. Prototype alignment regularization is also proposed to promote the model's performance. Experiments conducted on Pascal VOC, MS-COCO, NUS-WIDE and CUB have verified that our method outperforms existing state-of-the-art PML methods.

Code — https://github.com/lee-plus-plus/PML-CD

Introduction

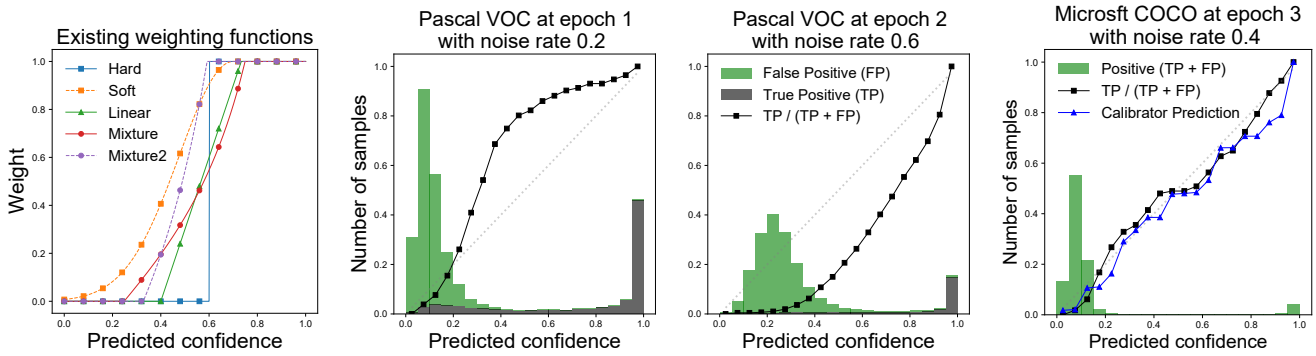
In recent years, multi-label learning has become a popular research topic, whose goal is to predict whether each category of label is relevant to a sample (Liu et al. 2021b). Given the fact that precisely annotated datasets are often difficult to be obtained due to high labor cost, crowd-sourcing platforms become common adopted for data collection, irrelevant labels are assigned by non-expert annotators, resulting the samples to be over-annotated, therefore leading to the problem of Partial Multi-label Learning (PML) (Yu et al. 2018). PML works with the problem that each instance is associated with a candidate label set, which contains not only all the relevant labels but also some irrelevant labels. Several PML methods have been proposed to disambiguate corrupted labels and learn robust model, achieving remarkable performance on PML table datasets (Xie and Huang 2021; Xu, Liu, and Geng 2020; Wang et al. 2023a). Inspired by the success of deep neural networks (DNNs), researchers begin to explore the potentials of designing PML methods with DNNs, particularly on PML image classification (Wang et al. 2023b), yet few methods have been proposed.

Curriculum learning, which let the model gradually learn from easy knowledge to hard knowledge by controlling the sample's weight during the training (Bengio et al. 2009), is learnt to be helpful to disambiguate noisy labels (Jiang et al. 2018; Kim et al. 2022). Therefore, it have been adopted by the majority of existing PML image classification methods (Wang et al. 2023b; Chen et al. 2024). In curriculum-based PML methods, the weighting function, also be named as the curriculum, is crucial for disambiguation. Among these proposed methods, however, although abundant weighting functions have been proposed (see Figure 1(a)) (Wang, Chen, and Zhu 2021), they are heuristically designed, e.g., hard threshold (Wang et al. 2023b), soft threshold (Chen et al. 2023), and temperature annealing (Xu et al. 2021), and therefore provide limited explanation for the criteria of designing curriculum, leading the evaluation or comparison between curriculum impossible.

Besides, it is noticed that the suitable curriculum should be vary on different dataset, different noise rate and different training epoch. Figure 1(b) provides such an example, showing that the best threshold to separate the true positive and false positive samples according to their predicted confidences vary. Therefore, a proper curriculum weighting function should be adaptive to the feedback of the training. However, existing curriculum-based PML approaches fails to achieve this (Wang et al. 2023b; Chen et al. 2024).

Motivated by the aforementioned limitations, we propose an approach named Calibrated Disambiguation for Partial Multi-label Learning (PML-CD), with a new perspective towards curriculum-based disambiguation by calibration. As it is shown in Figure 1(b), given the histogram of the confidences of candidate samples, the aim of curriculum weighting is to separate them into clean samples and corrupted samples, assigning high weights for the formal and low weights for the latter. Given the fact that samples with same confidence are indiscriminate during the disambiguation, we prove that the optimal mapping curve from

*Corresponding author



(a) Existing weighting functions.

(b) Histogram of the predict confidence.

(c) Calibrator predict on unseen dataset with high precision.

Figure 1: Illustration of Calibrated Disambiguation. (a) shows the weighting function proposed in previous curriculum learning approaches, whose shape are heuristically designed; (b) is two histograms of predicted confidences on clean partial labels and corrupted partial samples. The optimal weighting curve, which is equivalent to the calibration curve, can be derived from the frequency statistics if the ground-truth information is given; (c) shows that the optimal weighting curve can be precisely predicted according to the histogram of predicted confidences on partial labels, perceive a **general relations between confidences histogram and calibration curve** according to the memorization effect.

samples' confidence to sample's weight should equals to the $TP/(TP + FP)$ curve, which is named as the optimal weighting function. By applying the optimal weighting function during the curriculum-based disambiguation, the potentiality of the memorization effect can be fully mined, providing a upper bound for curriculum-based disambiguation. Figure 2 shows that the optimal weighting function is equivalent to the calibration on sample's confidences.

In PML training, however, the statistics of whether a positive sample belongs to true positive or false positive, is unknown. To predict the optimal weighting function when training on PML dataset, we designed a transferable calibrator to learn the relationship between the distribution of positive samples' confidence, and the optimal weighting function. A data-driven approach is adopt to train the calibrator, which avoids making assumptions on the distribution of confidences. Figure 1(c) shows that our pre-trained calibrator can give fine-grained prediction for optimal weighting curve on unseen dataset.

To enhance the robustness of our model, prototype alignment regularization is also proposed, which helps the model learn better multi-label representation and perform better.

In summary, our contributions are listed as follows:

1. An analysis for the optimal weighting function is provided. which demonstrates the essence of weighting in curriculum-based disambiguation is confidence calibration.
2. An adaptive and transferable calibrator is developed, which is based on the common patterns observed between the confidence distribution histograms of candidate samples and the ratio of clean samples.
3. A prototype alignment regularization loss is proposed, which can help multi-label encoder produce not only separable features, but also discriminative features, therefore leading to robust performance.

Related Work

Multi-label Image Classification

In recent years, multi-label image classification has drawn increasing attention, since it is more reasonable for each image to contain multiple relevant labels(Chen and Yeh 2024). To fully exploit the label correlations, convolutional neural networks (CNNs) with new architectures have been proposed according to various perspectives, including recurrent neural networks (Wang et al. 2016), graph convolutional networks (Chen et al. 2019; Singh et al. 2024), graph propagation (Zhu et al. 2023), and Transformer architecture (Liu et al. 2021a; Lanchantin et al. 2021).

Partial Multi-label Learning

Partial Multi-label Learning (PML) considers a situation where each instance is associated with a set of candidate labels, while only part of them are the ground-truth labels. Some existing work utilizes global and local information to guide disambiguation, following the manifold assumption and the label correlation assumption (Xu, Liu, and Geng 2020; Wang et al. 2019; Zhao et al. 2022). Other existing works, following the low-rank decomposition assumption, prefer to separate ground truth and noise from corrupted data by applying low-rank constraints while optimizing (Sun et al. 2019; Xie and Huang 2021). Lately proposed PML methods usually incorporate both the manifold and low-rank assumptions in a delicate way (Xu et al. 2022; Lyu et al. 2021; Sun et al. 2021; Yang et al. 2024b).

In recent years, researchers began exploring the compatibility of PML with DNNs and investigated PML image classification. Current image PML methods are primarily based on memory effects (Arpit et al. 2017), and apply curriculum-guided disambiguation through small-loss criteria (Wang et al. 2023b; Chen et al. 2024).

Curriculum Learning

The concept of curriculum learning is inspired by the human learning process, where individuals first acquire simple and general knowledge, then gradually learns more complex and specialized knowledge and thereby achieving well understanding (Bengio et al. 2009). By leveraging the memorization effect (Arpit et al. 2017) and employing training sample loss as the difficulty measurer (Wang, Chen, and Zhu 2021), noisy label learning approaches can progressively disambiguate labels through curriculum learning strategies (Jiang et al. 2018).

Based on the small-loss criteria, a series of weighting functions that map samples' loss to their weights have been proposed in curriculum learning (Wang, Chen, and Zhu 2021), with various forms such as hard (Kumar, Packer, and Koller 2010), linear (Jiang et al. 2014), logarithmic (Jiang et al. 2014), Gaussian (Chen et al. 2023), and etc. Since the most appropriate curriculum varies on different training conditions, the weighting function should be adaptively adjusted by the feedback during the training. Existing approaches have proposed various curriculum adaptation strategies, by introducing the confidence histograms (Lee, Lim, and Chung 2021), the mean and variance of predicted confidence (Chen et al. 2023), quantized logical differences (Kim et al. 2024), loss rankings (Chen et al. 2024), and etc. Corresponding parameter updating schemes have also been designed in different approaches, including fitting two Gaussian distribution (Lee, Lim, and Chung 2021), temperature annealing (Xu et al. 2021), and etc.

Overall, **designing adaptive weighting strategies heuristically is challenging for three main reasons**: 1) a mathematical formulation for optimal curriculum lacks; 2) certain feature construction approaches, such as ranking, are inherently suboptimal for precise inference; 3) some predictors employ generative models that rely on assumptions, such as the confidence levels of noisy and clean samples following two separable normal distributions (Lee, Lim, and Chung 2021), which may not always be applicable.

In addition to re-weighting, some approaches explore how to re-label corrupted labels in a curriculum-based manner (Bengio et al. 2009; Kim et al. 2022), or combine re-weighting and re-labeling together (Zhou, Wang, and Bilmes 2020). It is also referred to as re-labeling or label correction. The main challenge with re-labeling lies in the possibility of incorrectly correcting difficult but accurate labels in the early stages of training, which could even damage further learning (Kim et al. 2022).

Preliminaries

We first provide a detailed explanation of the notations used in the paper. The multi-label training dataset is denoted as $\mathcal{D} = \{(x_i, y_i) \mid i = 1, \dots, N\}$, where N is the number of samples of the training set, $x_i \in \mathbb{R}^{W \times H \times 3}$ is the image of the i-th sample, and $y_i = [y_i^1, \dots, y_i^C]^T \in \{0, 1\}^C$ denotes the ground truth labels for each category of the i-th sample, with a total of C categories. In PML, since labels are not precisely annotated, only a corrupted dataset $\tilde{\mathcal{D}} = \{(x_i, \tilde{y}_i) \mid i = 1, \dots, n\}$ is available, whose candidate

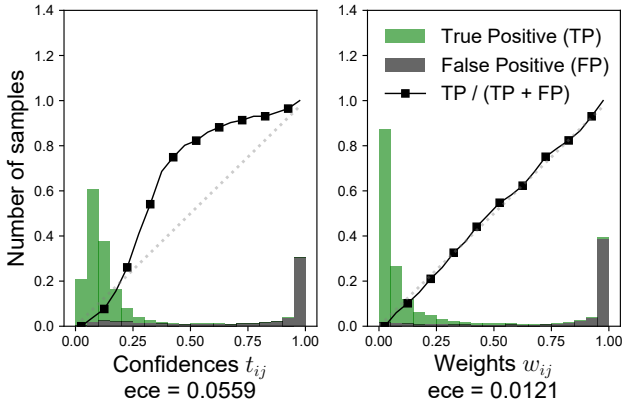


Figure 2: An example of optimal curriculum weighting with histograms of model's predicted confidences (left) and weights (right) by assigning $w_{ij} = P(y_{ij} = \tilde{y}_{ij} \mid t_{ij})$. It is shown that weighting is equivalent to calibrating model's predicted confidences. The expected calibration error (ECE) of confidences and weights are presented.

labels $\tilde{y}_i = [y_{i1}, \dots, y_{ic}]^T \in \{0, 1\}^C$ for the i-th sample is contains not only relevant labels corresponding to y_i, but also additional noise. For convenience, we denote $\{\cdot\}$ in this paper to succinctly represent a collection of certain variable of every sample on the training set, e.g., $\{x_i\}$.

Given a model $f : \mathbb{R}^{W \times H \times 3} \to [0, 1]^C$, the confidence predictions $t_i = f(x_i) \in [0, 1]^C$ for each sample on training set can be collected. According to the memorization effect, when a model learns on a noisy dataset, it tends to first memorize the clean samples. It suggests that the small-loss criteria to filter reliable samples and assign higher weights to samples with lower loss during the training. Such disambiguation procedure is also known as curriculum learning, which can be formulated as

$$\ell_{\text{curr}}(\hat{y}_{ij}, \tilde{y}_{ij}) = w_{ij} \ell_{\text{BCE}}(\hat{y}_{ij}, \tilde{y}_{ij}), \tag{1}$$

where $\hat{y}_{ij} = f_j(x_i)$ is the model's predicted confidence for i-th sample and j-th class, and $\ell_{\text{BCE}}(\hat{y}, \tilde{y}) = -(\tilde{y} \log(\hat{y}) + (1 - \tilde{y}) \log(1 - \hat{y}))$ is the binary cross entropy loss. Given a predicted score t, we hope to assign weight $w(t)$ to samples to better train the model. It is expected to assign as much weight as possible to clean samples, and as little weight as possible to corrupted samples. which can be evaluated by

$$\begin{aligned} \ell_{\text{BCE}}\left(w(t), \mathbb{I}[y = \tilde{y}]\right) = & \underbrace{-w(t) \log(\mathbb{I}[y = \tilde{y}])}_{\text{maximize clean samples' weights}} + \\ & \underbrace{-(1 - w(t)) \log(\mathbb{I}[y \neq \tilde{y}])}_{\text{minimize corrunpted samples' weights}}, \end{aligned} \tag{2}$$

and the *optimal* weighting function $w^*(t)$ can be induced as

$$\begin{aligned} w^*(t) &= \arg\min_{w(t)} \mathbb{E}_{y|t} \left[\ell_{\text{BCE}}(w(t), \mathbb{I}[y = \tilde{y}])\right] \\ &= P(y = \tilde{y} \mid t). \end{aligned} \tag{3}$$

Since other baseline methods try to model $w^*(t)$ in a binarization way, e.g., $w(t) = \mathbb{I}[t \geq \eta]$, our method pro-

vides theoretical upper bound. Therefore, the optimal curriculum weighting function is to model the posterior probability $P(y_{ij} = \tilde{y}_{ij} \mid t_{ij})$ as

$$\ell^*_{\text{curr}}(\hat{y}_{ij}, \tilde{y}_{ij}, y_{ij}) = P(y_{ij} = \tilde{y}_{ij} \mid t_{ij}) \ell_{\text{BCE}}(\hat{y}_{ij}, \tilde{y}_{ij}). \quad (4)$$

To better understand the meaning of Equation 4, an example of optimal curriculum weighting is given in Figure 2. In summary, our analysis below provides three insights:

1. In curriculum learning based disambiguation schemes, **the essence of weighting is the calibration of model's predicted confidence**.
2. The optimal weighting function $P(y_{ij} = \tilde{y}_{ij} \mid t_{ij})$ provides a **theoretically upper bounds for small-loss guided disambiguation**.
3. The **quality of weighing function can be evaluated according to the quality of calibration**, via metrics like expected calibration error (ECE).

Methods

In this section, we first introduce the overall architecture and training process of our proposed model PML-CD, which is also presented in Figure 3. Design proposals and implementation details of each module will be introduced later.

Two phase training. The training progress of our PML-CD model is divided into two phases. The first phase is the warm-up phase, where the model is trained via Binary Cross-Entropy (BCE) loss on raw training set until early stopping (Bai et al. 2021). This phase aims to initialize the model's random parameters to a reasonable state and to produce reasonable multi-label representations. The second phase is the disambiguation phase. In this phase, we collect the model's predicted confidence on the training set, and apply the calibrator to estimate the reliability of candidate labels, then apply the reliability score as the weights in a curriculum learning. The model is robustly trained via weighted BCE loss along with the prototype alignment regularization until early stopping.

Multi-label encoder. Current researches in multi-label learning suggests that it is more appropriate to learn a representation for each categories in an image. In multi-label images, objects of different classes are often situated in distinct regions of the image and can sometimes be quite small. Therefore, compressing all regional features into one feature vector through adaptive pooling is suboptimal (Liu et al. 2021a; Chen and Yeh 2024). Applying attention mechanisms on regional features enables the model to learn which regional features are relevant to each class, thus helping the model to learn more consistent feature representations (Liu et al. 2021a; Ridnik et al. 2023). Following the architecture of ML-Decoder (Ridnik et al. 2023), we first extract regional features from the image using pre-trained CNN backbone, and then pass them into a Transformer with C query vectors, therefore outputs the multi-label representations of an image $z \in \mathbb{R}^{C \times D}$.

Group linear projection. To produce confidence predictions, the Group Linear Projection (Ridnik et al. 2023) is adopted, where the representation in each class is passed through its corresponding linear layer.

Calibrator

Following the above analysis that **the essence of weighting is calibration**, unlike previous methods that heuristically formulating weighting function, we formulate our calibrator from the analysis of the optimal calibrator. Given the information of the ground-truth labels $\{y_i\}$, the optimal calibrator can be derived from the confidences on true partial labels and false partial labels in a histogram way. Given the predicted confidence on partial labels $\{t_{ij} \mid \tilde{y}_{ij} = 1\}$, we split the partial confidences $\{t_{ij}\}^{\text{P}} = \{t_{ij} \mid \tilde{y}_{ij} = 1\}$ into two group, the True Positive (TP) group $\{t_{ij}\}^{\text{TP}} = \{t_{ij} \mid \tilde{y}_{ij} = 1, y_{ij} = 1\}$ and the False Positive (FP) group $\{t_{ij}\}^{\text{FP}} = \{t_{ij} \mid \tilde{y}_{ij} = 1, y_{ij} = 0\}$. We denote the number of bins $K = 20$, confidences in each bin $\{t_{ij}\}^{[k]} = \{(k-1)/K \le t_{ij} < k/K\}$, and then the frequency distribution of TP and FP confidences are collected as

$$\begin{aligned} q_k^{\text{P}} &= \frac{|\{t_{ij}\}^{[k]} \cap \{t_{ij}\}^{\text{P}}|}{|\{t_{ij}\}^{\text{P}}|}, \\ q_k^{\text{FP}} &= \frac{|\{t_{ij}\}^{[k]} \cap \{t_{ij}\}^{\text{TP}}|}{|\{t_{ij}\}^{\text{P}}|}, \\ q_k^{\text{TP}} &= \frac{|\{t_{ij}\}^{[k]} \cap \{t_{ij}\}^{\text{FP}}|}{|\{t_{ij}\}^{\text{P}}|}. \end{aligned} \quad (5)$$

Therefore, the reliability histogram can be derived as

$$r_k = q_k^{\text{TP}} / (q_k^{\text{TP}} + q_k^{\text{FP}}). \quad (6)$$

In the end, we can export the confidence-reliability relationship between t_i and $P(y_{ij} = \tilde{y}_{ij} \mid t_{ij})$ via applying interpolation on reliability histogram

$$P(y_{ij} = \tilde{y}_{ij} \mid t_{ij}) = (\frac{k+1}{K} - t_{ij}) r_k + (t_{ij} - \frac{k}{K}) r_{k+1}, \quad s.t. \ \frac{k}{K} \le t_{ij} < \frac{k+1}{K}. \quad (7)$$

To produce precise calibration, the confidence-reliability relationship can be individually accomplished on each class.

It has been witnessed that the histogram of predicted confidence is highly relevant to the ground-truth confidence-reliability relationship during the training (Lee, Lim, and Chung 2021; Chen et al. 2024). Reasonable attempts that fit the TP and FP confidences via two Gaussian distribution has been made (Lee, Lim, and Chung 2021). However, according to our observation, it can hardly be adopted to general situations where the histogram of TP and FP confidences are skewed or are highly overlapping.

To overcome this issue, a *data-driven approach* is adopted. We model our calibrator $g : [0,1]^K \to [0,1]^K$ with the frequency of confidences in each bin $\left(q^{\text{P}}\right)_{k \in K}$ as input, and the confidence-reliability relationship $\left(q_k^{\text{TP}} / (q_k^{\text{TP}} + q_k^{\text{FP}})\right)_{k \in K}$ as output. The structure of the calibrator is designed as a 2-layer Multi-label Perception with the activation of LeakyReLU. To enable the calibrator to learn the relationship, we collect information of $\{q^{\text{P}}, q^{\text{TP}}, q^{\text{FP}}\}$ during the training of baseline model on every epoch at the disambiguation phase. The baseline model is implemented with

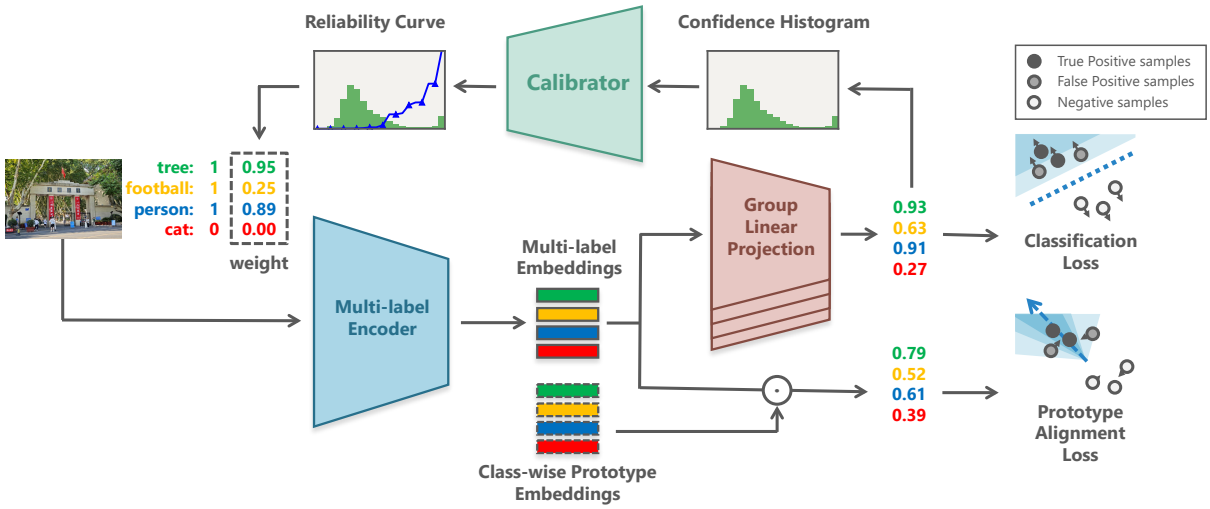


Figure 3: Illustration of the proposed model PML-CD. During the training, the model is updated by both classification loss and prototype alignment loss. Predicted confidences on training set are collected and are fed into the calibrator to generate reliability of candidate labels, which guides the curriculum disambiguation training in the next epoch.

the same architecture as PML-CD but adopts a trivial disambiguation strategy $P(y_{ij} = \tilde{y}_{ij} \mid t_{ij}) = t_{ij}$ adopted. In order to enable our calibrator be adaptive to different situation, the training data is collected from multiple training procedures on a dataset with different noise rate. After pre-training, the calibrator can be run on any dataset with any noise rate by loading parameters, and therefore serving as a plugin. Since our calibrator avoids making assumptions on the distribution of TP and FP confidences, and learns from data with both low noise and high noise, it can adapt to various scenario.

Prototype Alignment Regularization

It is common for multi-label DNN methods to first encode the multi-label representation, then produce the confidence through a *Linear Projection* layer for each class. Given the multi-label representation of i-th sample on j-th class $z_{ij} \in \mathbb{R}^D$, and the parameter of j-th class linear projection layer $\theta_j \in \mathbb{R}^D, b_j \in \mathbb{R}$, the confidence $\hat{y}_{ij}^{\text{clf}}$ is predicted as

$$\hat{y}_{ij}^{\text{clf}} = \theta_j^T z_{ij} + b_j. \tag{8}$$

Therefore, the classification loss

$$\mathcal{L}_{\text{clf}} = \sum_{i \in N} \sum_{j \in C} \ell_{\text{curr}}(\hat{y}_{ij}^{\text{clf}}, \tilde{y}_{ij}) \tag{9}$$

can be viewed as seeking an optimal hyperplane to separate the representations of positive and negative examples for each class. However, the decision boundary learned through linear projection can be easily disrupted by noisy labels, therefore lacking robustness (Yi et al. 2022). On the other hand, another widely applied way to produce confidence via multi-label representations is *Prototype Alignment*, which is reckoned to be robust in noisy label learning and is less prone to be over-fitting. Given the prototype representation of j-th class $\bar{z}_j \in \mathbb{R}^D$, the confidence $\hat{y}_{ij}^{\text{proto}}$ is predicted as

$$\hat{y}_{ij}^{\text{proto}} = \frac{\langle z_{ij}, \bar{z}_j \rangle}{\|z_{ij}\| \cdot \|\bar{z}_j\|}. \tag{10}$$

Given the prototype alignment loss

$$\mathcal{L}_{\text{proto}} = \sum_{i \in N} \sum_{j \in C} \ell_{\text{curr}}(\hat{y}_{ij}^{\text{proto}}, \tilde{y}_{ij}), \tag{11}$$

unlike maximizing the distance of representations from the decision hyperplane, the prototype alignment method compute the cosine similarity between sample's representation and the corresponding class prototype. It has been widely applied in partial label learning to guide disambiguation and provide more reliable label guesses for uncertain samples (Wang et al. 2022a,b; Yang et al. 2024a). However, this projection method struggles to achieve fine-grained fitting, resulting in suboptimal prediction performance.

Taking of impact on encoder into consideration, optimization on linear projection score $\ell_{\text{curr}}(\hat{y}_{ij}^{\text{clf}}, \tilde{y}_{ij})$ lets the embeddings of positive and negative samples linear separable, while optimization on prototype alignment score $\ell_{\text{curr}}(\hat{y}_{ij}^{\text{proto}}, \tilde{y}_{ij})$ lets the embeddings of positive and negative be clustered around their prototype embeddings. By applying both of them, a representation which is not only separable, but also discriminative, is encouraged (Wen et al. 2016).

Above all, in warm-up phase, the training loss is

$$\mathcal{L}_{\text{warmup}} = \mathcal{L}_{\text{clf}}, \tag{12}$$

and training loss in disambiguation phase is defined as

$$\mathcal{L}_{\text{disam}} = \mathcal{L}_{\text{clf}} + \mathcal{L}_{\text{proto}}. \tag{13}$$

During the training, we allow the prototype representations be updated via gradient, therefore no explicit updating strategy is required. Since the multi-label representation is not well initialized during the warm-up training, making their prototypes meaningless, the prototype alignment regularization is only applied during the disambiguation phase.

Algorithm 1: Algorithm of PML-CD

Input: partial multi-label learning dataset $\{(x_i, \tilde{y}_i)\}$,
Parameter: pre-trained calibrator g, learning rate η
Output: network $f(\cdot \mid \theta)$ for PML

repeat
  Update network parameters $\theta \leftarrow \theta - \eta\Delta\mathcal{L}_{\text{warmup}}$
until early stopping
Initialize $w_{ij} \leftarrow 1$
repeat
  Collect predicted confidences $\{t_{ij}\} \leftarrow f(\{x_i\})$
  Compute q^{P} via Equation. 5
  Predict reliability histogram r via Equation. 6
  Update curriculum weights $w_{ij} = P(y_{ij} = \tilde{y}_{ij} \mid t_{ij})$ via Equation. 7, if $\tilde{y}_{ij} = 1$
  Update network parameters $\theta \leftarrow \theta - \eta\Delta\mathcal{L}_{\text{disam}}$
until early stopping

Experiments

Settings

Datasets. To validate our proposed approach, we conducted experiments on several commonly used real-world multi-label image datasets, including MS-COCO 2014 (Lin et al. 2014), Pascal VOC 2007 (Everingham et al. 2010), NUS-WIDE (Chua et al. 2009), and CUB 200 (Wah et al. 2011). Specifically, we took the 312 binary attributes instead of the bird category as predicted labels, therefore making it a multi-label dataset. To create PML datasets from multi-label datasets, we injected random noise into the label matrices by randomly flipping the zero values in the original label matrix with a certain probability, denoted as noise rate ρ.

Comparison Methods. We compare our PML-CD model with three groups of methods.

- **Multi-label learning methods**. **BCE** is a standard multi-label classification model trained with binary cross entropy loss. **ASL** (Ridnik et al. 2021) is a multi-label learning method that addresses the issue of imbalance between positive and negative labels. **ML-Decoder** (Ridnik et al. 2023) re-designs the decoder architecture and introduces a group-decoding scheme to provide efficiency. These methods do not specifically address noise disambiguation, and have achieved state-of-the-art results on many multi-label image classification datasets.
- **Traditional PML methods**. **PML-NI** (Xie and Huang 2021) applies low-rank constraints to decompose the PML label matrix, while **PML-LRS** (Sun et al. 2019) decomposes the projection matrix of linear regression model. **PLAIN** (Wang et al. 2023a) disambiguates gradually with instance similarity and label correlations. These methods are shallow methods that are not specifically designed for image classification. Therefore, we use ResNet-101 (He et al. 2016) pre-trained on ImageNet to extract feature embeddings to make them appliable.
- **Curriculum-based PML methods**. **LL-R** (Kim et al. 2022) ignores training samples with high loss during the training. **PML-CDCR** (Wang et al. 2023b) trains on strongly-augmented images and selecting reliable samples on weakly-augmented images with a loss threshold. **UNM** (Chen et al. 2024) applies cyclical learning rate to transform the model from over-fitting status back to under-fitting status, therefore providing more reliable disambiguation during the training. These methods adopt a curriculum disambiguation framework similar to ours.

Setting and evaluation. Our experiments are run on GeForce RTX 4090 with PyTorch 1.13.1. We employ ResNet-101 (He et al. 2016) pre-trained on ImageNet as the backbone, and the detailed setting for all experiments are described as follow.

During the training, all images are resized to 224×224, and strong data augmentation are applied for training set, including horizontal flip, RandAugment (Cubuk et al. 2020) and Cutout. Same data processing are adopted for almost all comparison methods. We train the model with the following optimization setting: an Adam Optimizer with fixed learning rate 1×10^{-4} for VOC and CUB, 1×10^{-5} for COCO and NUS-WIDE; a weight decay of 5×10^{-5}. The size of mini-batch is set to be 32. Early stopping is applied for all comparison methods, since noisy labels learning is highly dependent on it (Bai et al. 2021). **Mean Average Precision** (mAP) is applied to evaluate models' performance.

Calibrator pre-training. Our calibrator is pre-trained with the following procedure. First, the baseline model without prototype alignment regularization loss and with an identity weighting function is trained on VOC dataset, with noise rate ranging from $10\%, 20\%, \ldots, 90\%$. The partial label matrix, ground-truth label matrix, and the confidence prediction matrix at every epoch are collected during the training. On every epoch, the confidence histogram and the reliability histogram for each class are calculated independently, therefore making C feature-label pairs (q^{P}, r). After the dataset is constructed, the calibrator is trained by minimizing **Mean Squared Error** (MSE) via an Adam optimizer with a fixed learning rate 1×10^{-2}.

Comparison Results

Results are reported in Table 1, which shows that our model outperforms other baseline models on $15/16$ conditions, thus achieving state-of-the-art results. PML-CD performs slightly inferior to other baseline methods under low noise condition on VOC, but performs significantly superior to others under high noise conditions, which can be explained as PML-CDCR and LL-R is more suitable on low noise condition since they adopt simpler network architecture, i.e., ResNet only. It is noticed that our model performs well on CUB, which has over 300 fine-grained visual attributes, making it challenging for PML. Our model outperforms other curriculum-based PML baseline models with significant advantages, for the reason that our model adopts the ML-Decoder's multi-layer transformer architecture as the encoder, therefore advantageous in extracting fine-grained visual features.

To summarize, our model outperforms other baseline models for $15/16$ times under different training conditions, which shows that our model, together with the pre-trained

Method	VOC				COCO				NUS-WIDE				CUB			
	20%	40%	60%	80%	20%	40%	60%	80%	20%	40%	60%	80%	20%	40%	60%	80%
PML-NI	62.57	35.37	20.97	13.84	60.14	53.41	43.02	24.52	21.26	16.39	12.02	7.89	12.92	12.04	11.42	10.96
PML-LRS	79.96	63.45	32.86	16.66	61.56	53.80	42.85	24.23	21.52	16.23	11.85	7.76	14.56	12.91	11.83	11.19
PLAIN	80.76	72.45	54.59	27.24	60.13	53.71	44.99	21.56	19.22	17.85	13.72	7.44	14.98	11.38	10.55	10.37
BCE	87.73	84.26	75.27	48.86	73.66	69.62	63.52	50.47	50.81	45.78	37.90	27.06	29.66	27.01	22.19	17.39
ASL	88.63	84.43	77.23	52.95	73.19	68.53	62.56	50.46	46.27	37.86	31.67	21.97	28.58	25.75	22.66	17.35
MLDecoder	87.94	85.49	80.41	64.46	74.29	70.35	68.55	60.55	50.92	46.11	40.01	30.83	29.72	27.63	24.99	21.13
UNM	87.99	86.58	79.94	44.07	65.56	63.51	61.68	27.87	37.83	34.57	29.62	12.04	23.93	19.73	18.42	11.66
PML-CDCR	88.56	86.87	69.26	38.31	72.44	69.90	66.88	42.17	42.91	39.96	31.99	22.37	26.67	24.63	23.05	16.38
LL-R	88.99	82.87	72.36	40.32	74.35	67.79	61.47	41.28	43.75	42.09	30.56	24.76	29.03	26.47	22.54	18.62
PML-CD	88.83	87.12	83.87	71.70	75.26	74.65	69.88	62.10	51.30	46.11	40.41	32.10	30.52	27.87	25.42	21.69

Table 1: Comparison of our model PML-CD with baseline methods, with noise rate 20%, 40%, 60% and 80%. mAP(%) is adopted as metric. The best performances on each column are highlighted with background color.

calibrator, can not only perform well on VOC dataset, but also on unseen dataset including COCO, NUS-WIDE, and CUB. Given the complexity of multi-label encoder can sometimes lead to sub-optimal performance on on easy PML dataset, e.g., VOC with low noise rate, it enables our model to encode fine-grained visual features on images, which is of necessity when training on CUB.

Ablation Study

Effectiveness of calibrator. To evaluate whether the pre-trained calibrator can accurately predict the optimal calibration curve on unseen datasets, and therefore calibrate their predicted confidences, we evaluate the expected calibration error (ECE) of predicted confidences and the curriculum weighting, during the disambiguation phase of the training on VOC, COCO datasets. Also, to compare with other weighting policy, we adopt the pre-trained calibrator, the optimal calibrator, and other commonly used curriculum weighting functions for comparison. We denote the **hard** weighting function $v^{\text{hard}}(t_{ij}) = \mathbb{I}[t_{ij} \geq 0.6]$, the **soft** weighting function $v^{\text{soft}}(t_{ij}) = (t_{ij})^2$, the **identity** weighting function $v^{\text{identity}}(t_{ij}) = t_{ij}$. Comparison result are shown in Table 2.

weighting policy	ECE of confidences (epoch 1)	ECE of weights (epoch 1)	mAP per class (overall)
hard	76.78	92.96	59.79
soft	76.78	61.68	60.08
identity	76.78	76.78	59.93
calibrator	76.78	3.74	61.48
optimal	76.78	3.74	66.42

Table 2: Effectiveness of weighting policies on MS-COCO with noise rate 80%. ECE(%) and mAP(%) are presented.

Effectiveness of prototype alignment regularization. To verify whether the prototype alignment regularization (PAR) enhances the performance of PML-CD, we train our model on the VOC and COCO datasets with corresponding settings: 1) applying no PAR; 2) applying PAR at the disambiguation phase; 3) applying PAR at both the warmup phase and disambiguation phase. Comparison results are presented in Table 3. It is shown that prototype alignment regularization helps the model perform better especially in high noise condition.

dataset	$\mathcal{L}_{\text{proto}}$	noise rate 20%	40%	60%	80%
VOC	neither	89.28	87.93	85.05	66.91
	2nd-phase only	88.77	88.12	85.07	71.70
	1st & 2nd-phase	87.89	86.11	83.90	71.40
COCO	neither	75.54	73.86	69.78	60.13
	2nd-phase only	75.26	74.65	69.88	62.10
	1st & 2nd-phase	74.79	72.58	70.27	62.16

Table 3: Ablation study of the prototype alignment regularization. mAP(%) is adopted as metric.

Conclusion

In this work, we present a novel model PML-CD to apply curriculum-based disambiguation on PML in a calibration way. We derive the optimal weighting function from existing curriculum-based PML methods, and provide an insight that assigning curriculum weights for samples is equivalent to the calibration of the model's predicted confidences. Guided by the optimal weighting function, we design a calibrator based on the general relations between confidences histogram and calibration curve, which can provide high quality calibration on unseen dataset after pre-training, therefore serving as an adaptive weighting function on different training condition. We also propose the prototype alignment regularization to improve model's robustness during the training. Comparison experiments have been conducted to verify the effectiveness of PML-CD.

Acknowledgements

This research work is supported in part by National Natural Science Foundation of China (Grant 62106044 and Grant U24A20322), in part by the Big Data Computing Center of Southeast University.

References

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; et al. 2017. A closer look at memorization in deep networks. In *International conference on machine learning*, 233–242. PMLR.

Bai, Y.; Yang, E.; Han, B.; Yang, Y.; Li, J.; Mao, Y.; Niu, G.; and Liu, T. 2021. Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34: 24392–24403.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Chen, C.-Y.; and Yeh, M.-C. 2024. Self-Supervised Multi-Label Classification with Global Context and Local Attention. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 934–942.

Chen, H.; Tao, R.; Fan, Y.; Wang, Y.; Wang, J.; Schiele, B.; Xie, X.; Raj, B.; and Savvides, M. 2023. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*.

Chen, J.-Y.; Li, S.-Y.; Huang, S.-J.; Chen, S.; Wang, L.; and Xie, M.-K. 2024. UNM: A Universal Approach for Noisy Multi-label Learning. *IEEE Transactions on Knowledge and Data Engineering*.

Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5177–5186.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 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*, 1–9.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 702–703.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jiang, L.; Meng, D.; Mitamura, T.; and Hauptmann, A. G. 2014. Easy samples first: Self-paced reranking for zero-example multimedia search. In *Proceedings of the 22nd ACM international conference on Multimedia*, 547–556.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.-J.; and Fei-Fei, L. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, 2304–2313. PMLR.

Kim, S.; Lee, D.; Kang, S.; Chae, S.; Jang, S.; and Yu, H. 2024. Learning Discriminative Dynamics with Label Corruption for Noisy Label Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22477–22487.

Kim, Y.; Kim, J. M.; Akata, Z.; and Lee, J. 2022. Large loss matters in weakly supervised multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14156–14165.

Kumar, M.; Packer, B.; and Koller, D. 2010. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23.

Lanchantin, J.; Wang, T.; Ordonez, V.; and Qi, Y. 2021. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16478–16488.

Lee, J.; Lim, H.; and Chung, K.-S. 2021. CLC: Noisy label correction via curriculum learning. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–7. IEEE.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, S.; Zhang, L.; Yang, X.; Su, H.; and Zhu, J. 2021a. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*.

Liu, W.; Wang, H.; Shen, X.; and Tsang, I. W. 2021b. The emerging trends of multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 7955–7974.

Lyu, G.; Feng, S.; Jin, Y.; Wang, T.; Lang, C.; and Li, Y. 2021. Prior knowledge regularized self-representation model for partial multilabel learning. *IEEE Transactions on Cybernetics*, 53(3): 1618–1628.

Ridnik, T.; Ben-Baruch, E.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; and Zelnik-Manor, L. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 82–91.

Ridnik, T.; Sharir, G.; Ben-Cohen, A.; Ben-Baruch, E.; and Noy, A. 2023. Ml-decoder: Scalable and versatile classification head. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 32–41.

Singh, I. P.; Ghorbel, E.; Oyedotun, O.; and Aouada, D. 2024. Multi-label image classification using adaptive graph convolutional networks: from a single domain to multiple domains. *Computer Vision and Image Understanding*, 247: 104062.

Sun, L.; Feng, S.; Liu, J.; Lyu, G.; and Lang, C. 2021. Global-local label correlation for partial multi-label learning. *IEEE Transactions on Multimedia*, 24: 581–593.

Sun, L.; Feng, S.; Wang, T.; Lang, C.; and Jin, Y. 2019. Partial multi-label learning by low-rank and sparse decomposition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 5016–5023.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, H.; Liu, W.; Zhao, Y.; Zhang, C.; Hu, T.; and Chen, G. 2019. Discriminative and Correlative Partial Multi-Label Learning. In *IJCAI*, 3691–3697.

Wang, H.; Xiao, R.; Li, Y.; Feng, L.; Niu, G.; Chen, G.; and Zhao, J. 2022a. Pico: Contrastive label disambiguation for partial label learning. In *International conference on learning representations*.

Wang, H.; Xiao, R.; Li, Y.; Feng, L.; Niu, G.; Chen, G.; and Zhao, J. 2022b. Pico+: Contrastive label disambiguation for robust partial label learning. *arXiv preprint arXiv:2201.08984*.

Wang, H.; Yang, S.; Lyu, G.; Liu, W.; Hu, T.; Chen, K.; Feng, S.; and Chen, G. 2023a. Deep partial multi-label learning with graph disambiguation. *arXiv preprint arXiv:2305.05882*.

Wang, H.; Yang, S.; Lyu, G.; Liu, W.; Hu, T.; Chen, K.; Feng, S.; and Chen, G. 2023b. Deep partial multi-label learning with graph disambiguation. *arXiv preprint arXiv:2305.05882*.

Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; and Xu, W. 2016. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2285–2294.

Wang, X.; Chen, Y.; and Zhu, W. 2021. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 4555–4576.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11–14, 2016, proceedings, part VII 14*, 499–515. Springer.

Xie, M.-K.; and Huang, S.-J. 2021. Partial multi-label learning with noisy label identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3676–3687.

Xu, N.; Liu, Y.-P.; and Geng, X. 2020. Partial multi-label learning with label distribution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 6510–6517.

Xu, N.; Wu, Y.-D.; Qiao, C.; Ren, Y.; Zhang, M.; and Geng, X. 2022. Multi-view partial multi-label learning via graph-fusion-based label enhancement. *IEEE Transactions on Knowledge and Data Engineering*, 35(11): 11656–11667.

Xu, Y.; Shang, L.; Ye, J.; Qian, Q.; Li, Y.-F.; Sun, B.; Li, H.; and Jin, R. 2021. Dash: Semi-supervised learning with dynamic thresholding. In *International conference on machine learning*, 11525–11536. PMLR.

Yang, F.; Cheng, J.; Liu, H.; Dong, Y.; Jia, Y.; and Hou, J. 2024a. Mixed Blessing: Class-Wise Embedding guided Instance-Dependent Partial Label Learning. *arXiv preprint arXiv:2412.05029*.

Yang, F.; Jia, Y.; Liu, H.; Dong, Y.; and Hou, J. 2024b. Noisy Label Removal for Partial Multi-Label Learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3724–3735.

Yi, L.; Liu, S.; She, Q.; McLeod, A. I.; and Wang, B. 2022. On learning contrastive representations for learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16682–16691.

Yu, G.; Chen, X.; Domeniconi, C.; Wang, J.; Li, Z.; Zhang, Z.; and Wu, X. 2018. Feature-induced partial multi-label learning. In *2018 IEEE international conference on data mining (ICDM)*, 1398–1403. IEEE.

Zhao, P.; Zhao, S.; Zhao, X.; Liu, H.; and Ji, X. 2022. Partial multi-label learning based on sparse asymmetric label correlations. *Knowledge-Based Systems*, 245: 108601.

Zhou, T.; Wang, S.; and Bilmes, J. 2020. Robust curriculum learning: from clean label detection to noisy label self-correction. In *International Conference on Learning Representations*.

Zhu, X.; Liu, J.; Liu, W.; Ge, J.; Liu, B.; and Cao, J. 2023. Scene-aware label graph learning for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1473–1482.