
RankMatch: A Novel Approach to Semi-Supervised Label Distribution Learning Leveraging Rank Correlation between Labels

Zhiqiang Kou^{1,2}, Yucheng Xie^{1,2}, Hailin Wang⁵, Junyang Chen^{1,2}, Jing Wang^{1,2},
Ming-Kun Xie³, Shuo Chen³, Yuheng Jia^{1,2*}, Tongliang Liu⁴, Xin Geng^{1,2*}

¹School of Computer Science and Engineering, Southeast University, China

²Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China

³RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan

⁴Sydney AI Centre, The University of Sydney, Australia

⁵School of Mathematics and Statistics, Xian Jiaotong University, China

{zhiqiang_kou, xieyc, chenjunyang, wangjing91, yhjia, xgeng}@seu.edu.cn,
{ming-kun.xie, shuo.chen.ya}@riken.jp, wanghailin97@163.com,
tongliang.liu@sydney.edu.au

Abstract

Pseudo label based semi-supervised learning (SSL) for single-label and multi-label classification tasks has been extensively studied; however, semi-supervised label distribution learning (SSLDL) remains a largely unexplored area. Existing SSL methods fail in SSLDL because the pseudo-labels they generate only ensure overall similarity to the ground truth but do not preserve the ranking relationships between true labels, as they rely solely on KL divergence as the loss function during training. These skewed pseudo-labels lead the model to learn incorrect semantic relationships, resulting in reduced performance accuracy. To address these issues, we propose a novel SSLDL method called *RankMatch*. *RankMatch* fully considers the ranking relationships between different labels during the training phase with labeled data to generate higher-quality pseudo-labels. Furthermore, our key observation is that a flexible utilization of pseudo-labels can enhance SSLDL performance. Specifically, focusing solely on the ranking relationships between labels while disregarding their margins helps prevent model overfitting. Theoretically, we prove that incorporating ranking correlations enhances SSLDL performance and establish generalization error bounds for *RankMatch*. Finally, extensive real-world experiments validate its effectiveness.

1 Introduction

Label Distribution Learning (LDL) ² [10, 26] is a machine learning paradigm designed to address label ambiguity [11, 53]. Unlike Multi-label Learning [63, 30], which assigns a fixed number of labels to each instance [59], LDL extends this framework by quantifying the importance of each label through description degrees [20], thereby providing richer supervision information [19, 28].

*Corresponding authors.

²LDL is similar to learning from soft labels, but the soft-label formulation focuses on single-label problems (i.e., there is only one true label for each instance), while LDL considers multi-label problems (i.e., each instance can have multiple true labels).

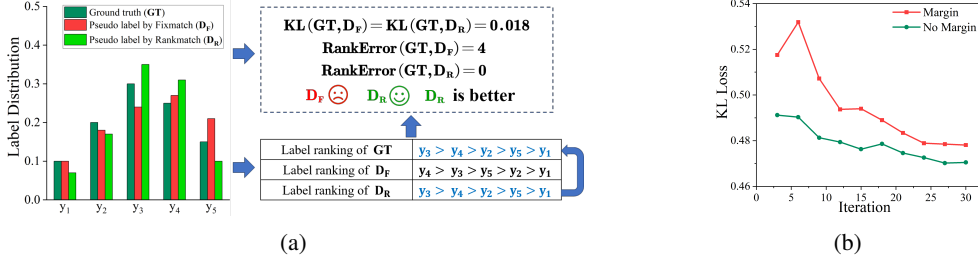


Figure 1: Illustration of the effectiveness of rank-aware pseudo-labeling. (a) Comparison of Pseudo-Labels by FixMatch (D_F) and RankMatch (D_R) on a sample from the Twitter-LDL dataset [61]. D_F fails to preserve the true label ranking (GT), despite a low KL divergence. In contrast, D_R maintains ranking relationships with higher Kendall tau (τ). (b) Performance on RAF-LDL dataset [24]. Flexible pseudo-label utilization improves performance by focusing on ranking relationships without strict margin alignment.

Deep learning has demonstrated remarkable success across various domains, primarily due to its ability to leverage large-scale and accurately labeled datasets [59, 39], which are essential for training deep neural networks (DNNs) with strong generalization. However, obtaining labeled data for LDL is particularly challenging and costly [31, 29]. For example, annotating the RAF-LDL dataset [34] required 315 trained annotators, with each image annotated multiple times to generate appropriate label distributions [34, 27]. This highlights the significant burden of creating labeled datasets for LDL. Given these challenges, the importance of semi-supervised LDL becomes evident.

Semi-supervised learning (SSL) [1, 9] has made significant progress, particularly in the areas of single-label [52, 23] and multi-label learning [59, 39] based on classical deep learning. However, Semi-Supervised Label Distribution Learning (SSLDL) remains relatively underexplored. One of the key techniques in SSL is leveraging trained models to generate pseudo-labels [56, 33] for unlabeled data, with the most well-known methods being FixMatch [52] and MixMatch [2]. These methods [52, 23] fail in SSLDL because, during training with labeled data, the model focuses solely on minimizing the overall similarity between predicted and ground truth label distributions (e.g., using KL divergence as loss function for training) without learning label ranking correlations, leading to biased pseudo-labels. As shown in Fig. 1(a), for a sample from the Twitter-LDL dataset [61], FixMatch generates pseudo-labels with a low KL divergence ($KL = 0.018$) but entirely incorrect label ranking correlations. This phenomenon is common across other datasets: ignoring the ranking relationships among labels during training produces distorted pseudo-label distributions (PLDs), and training with these PLDs causes the models performance to degrade. Moreover, we found that forcing the model to exactly match the pseudo-label distributions during training leads to overfitting. In contrast, using only the relative ranking among pseudo-labels preserves the underlying structure more robustly. Experiments on the RAF-LDL dataset (Fig. 1(b)) demonstrate that imposing only ranking constraints on pseudo-labels rather than enforcing strict numeric matchings significantly improves model performance.

In this paper, we first propose a pseudo-label-based SSLDL method called *RankMatch*, incorporates the ranking correlations between labels into the supervised training process. We introduce a novel loss function called the *Pairwise Ranking Relationship Loss (PRR Loss)* to enhance the ability of pseudo-labels to capture the ranking correlations between labels. Furthermore, we introduce a flexible pseudo-label training strategy that prioritizes ranking relationships between labels while disregarding margins, which prevents the model from overfitting to absolute label differences and enables a more robust utilization of pseudo-labels. In the theoretical aspect, we prove that incorporating ranking correlations between labels can enhance the performance of SSLDL and provide generalization error bounds for the *RankMatch*. Finally, extensive experiments on real-world datasets validate the effectiveness of our method. Our contributions can be summarized as follows:

- To the best of our knowledge, this is the first deep learning-based SSLDL algorithm utilizing pseudo-labels. Compared to existing SSL methods, our approach generates pseudo-labels that better align with the label distribution setting. Additionally, we propose a flexible pseudo-label utilization strategy for SSLDL.

- We theoretically demonstrate that incorporating label ranking correlations enhances model performance and provide a generalization bound for *RankMatch*.
- Extensive experiments on multiple datasets validate the effectiveness of *RankMatch*, consistently outperforming existing *SSLDL* methods.

2 RELATED WORK

Label Distribution Learning (LDL) [10, 54] assigns a distribution over labels to each instance, establishing a direct mapping between instances and their label distributions. Originally proposed for facial age estimation [12], LDL generates distributions across all possible age categories, offering richer supervisory signals than traditional single-label approaches. This paradigm has also demonstrated strong performance in facial emotion recognition, where it effectively models ambiguous emotional states by capturing uncertainty within the label space [49, 48, 38, 47].

Beyond facial analysis, LDL has shown broad applicability across diverse domains. For instance, NASA employed LDL to infer the chemical compositions of Martian meteorites [42], refining the method to predict elemental abundances from crystallographic data. In mental health, LDL has been used for depression detection via the Deep Joint Label Distribution and Metric Learning framework, which identifies subtle variations in facial expressions associated with different depression levels [66]. In crowd analysis, Ling [35] applied LDL to estimate indoor crowd densities by assigning label distributions that more accurately describe population levels in video frames.

Despite its success, LDL still faces challenges due to the scarcity of precisely annotated data [37, 60]. To mitigate this, several Semi-Supervised Label Distribution Learning (SSLDL) methods have been proposed. Hou [18] inferred the label distribution of unlabeled data by averaging the labels of its nearest neighbors, using both labeled and unlabeled samples for training. Jia [22] enhanced label distribution recovery by exploiting graph-structured relationships among instances. Liu [39] further developed a co-regularization-based SSLDL framework that leverages dual model structures to improve robustness and consistency.

However, these SSLDL methods are often not end-to-end and rely heavily on manual feature engineering, limiting their scalability to high-dimensional or large-scale data. They also underutilize unlabeled information. In contrast, deep learning provides a natural mechanism for automatic representation learning and has demonstrated remarkable success in data-rich environments. Consequently, integrating deep learning into SSLDL offers a promising direction to address existing limitations and unlock the full potential of label distribution learning under limited supervision.

3 Problem Statement and Notation

In SSLDL, the training data consists of a labeled dataset $\mathcal{D}_L = \{(\mathbf{x}_i, \mathbf{d}_i) | i = 1, 2, \dots, n\}$ and an unlabeled dataset $\mathcal{D}_U = \{\mathbf{x}_g | g = 1, 2, \dots, m\}$. Here, n and m represent the number of labeled and unlabeled samples, respectively. In the labeled dataset \mathcal{D}_L , \mathbf{x}_i is a labeled sample, and $\mathbf{d}_i = \{d_{\mathbf{x}_i}^{y_1}, d_{\mathbf{x}_i}^{y_2}, \dots, d_{\mathbf{x}_i}^{y_c}\}$ is the corresponding label distribution, where $d_{\mathbf{x}_i}^{y_j}$ represents the importance or relevance of label y_j to sample \mathbf{x}_i . The label distribution satisfies the normalization constraint $\sum_{j=1}^c d_{\mathbf{x}_i}^{y_j} = 1$. c denotes the number of labels in the label space $\mathcal{Y} = \{y_1, y_2, \dots, y_c\}$.

4 The Method

4.1 The Supervised Training Phase

In LDL, we transition from using the traditional binary cross-entropy loss, commonly employed in multi-label learning [17], to adopting Kullback-Leibler (KL) divergence as the loss function. This transition is essential because LDL predicts continuous real-valued label distributions instead of discrete binary outcomes. The KL divergence [17] is well-suited for measuring the difference between the ground-truth and predicted label distributions. The supervised loss is formulated as:

$$\mathcal{L}_s = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^c d_{\mathbf{x}_i}^{y_j} \ln \left(\frac{d_{\mathbf{x}_i}^{y_j}}{h(y_j | \text{Aug}_w(\mathbf{x}_i); \theta)} \right), \quad (1)$$

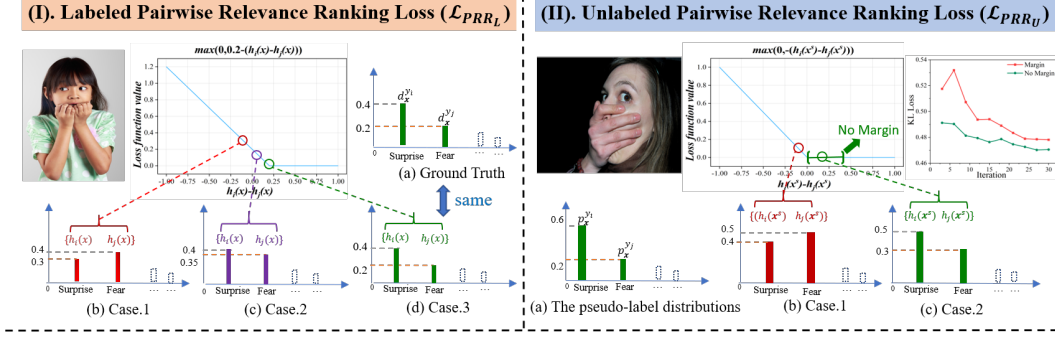


Figure 2: An example to illustrate the \mathcal{L}_{PRR} loss.

where $\text{Aug}_w(\mathbf{x}_i)$ represents a weakly augmented version of the i -th labeled sample [52]. The term $h(y_j | \text{Aug}_w(\mathbf{x}_i); \theta)$ denotes the predicted importance degree of label y_j for the augmented instance $\text{Aug}_w(\mathbf{x}_i)$, as determined by the model. This is computed as:

$$h(y_j | \mathbf{x}_i; \theta) = \frac{\exp(f_j(\mathbf{x}_i; \theta))}{\sum_{q=1}^c \exp(f_q(\mathbf{x}_i; \theta))}, \quad (2)$$

where $f_j(\mathbf{x}_i; \theta)$ represents the raw output of the DNN for label y_j with respect to instance \mathbf{x}_i . This formulation ensures that the predicted label distribution $h(y_j | \mathbf{x}_i; \theta)$ satisfies the normalization constraint $\sum_{j=1}^c h(y_j | \mathbf{x}_i; \theta) = 1$.

Existing SSL methods like FixMatch [52] and MixMatch [2] fail in SSLDL because they focus only on minimizing the KL divergence between predicted and ground truth distributions, neglecting label ranking relationships. This oversight leads to pseudo-labels that may have low KL divergence but incorrect label rankings, as shown in the example where FixMatch produces reversed label importance. In contrast, SSLDL requires preserving both the absolute importance and the relative ranking of labels to ensure semantic consistency and accurate predictions.

To produce more reliable pseudo-labels, we propose the Pairwise Relevance Ranking (PRR) loss \mathcal{L}_{PRR} , which aligns predictions with the inherent semantic structure of label distributions. For labeled data, \mathcal{L}_{PRR_L} strictly enforces alignment between the predicted and ground-truth label rankings while preserving meaningful margins. For example, when label description degrees $d_{\mathbf{x}_i}^{y_i} = 0.32$ and $d_{\mathbf{x}_i}^{y_k} = 0.33$, their negligible difference, likely caused by annotation noise, avoids unnecessary ranking adjustments. Let $h_j(\mathbf{x}_i)$ denote the predicted relevance for the j -th label after weak augmentation Aug_w . The \mathcal{L}_{PRR_L} loss is defined as:

$$\mathcal{L}_{PRR_L} = \sum_{1 < j < k < c} \left(s(j, k) \cdot g_\delta(j, k) + s(k, j) \cdot g_\delta(k, j) \right), \quad (3)$$

Here, $\delta = d_{\mathbf{x}_i}^{y_j} - d_{\mathbf{x}_i}^{y_k}$, and $f(j, k)$ and $g_\delta(j, k)$ are defined as follows:

$$s(j, k) = \begin{cases} 1, & \text{if } d_{\mathbf{x}_i}^{y_j} > d_{\mathbf{x}_i}^{y_k} \text{ and } d_{\mathbf{x}_i}^{y_j} - d_{\mathbf{x}_i}^{y_k} > t, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

$$g_\delta(j, k) = \begin{cases} 0, & \text{if } h_j(\mathbf{x}_i) - h_k(\mathbf{x}_i) \geq \delta, \\ \delta - (h_j(\mathbf{x}_i) - h_k(\mathbf{x}_i)), & \text{otherwise.} \end{cases} \quad (5)$$

Fig. 2(a) shows an example of the \mathcal{L}_{PRR_L} loss, using a sample from the RAF-LDL dataset. This loss penalizes two key scenarios (i) when the predicted ranking of labels deviates from the ground truth (Case 1); (ii) when the ranking is correct but the difference between label scores does not match the ground-truth margin (Case 2). Only in (iii), where both the ranking and the margin exactly match the ground truth (Case 3), does the loss drop to zero, indicating a perfectly correct prediction.

4.2 Self-Training Phase by Pseudo-label distribution

Pseudo-label distribution generation: To improve prediction stability and effectively utilize unlabeled data, we adopt an ensemble learning-based approach [67]. This method generates pseudo-label distributions (PLDs) for unlabeled instances by averaging the model’s outputs from multiple weak augmentations of the same image [52].

The pseudo-label generation process is as follows: given an unlabeled image \mathbf{x} , the model computes raw outputs (logits) for H weakly augmented versions $\text{Aug}_w(\mathbf{x})$. The PLD for \mathbf{x} , denoted as \mathbf{p}_i , is defined as:

$$\mathbf{p}_i(y_j) = \frac{\exp\left(\frac{1}{H} \sum_{k=1}^H f_j(\text{Aug}_w(\mathbf{x})_k; \theta)\right)}{\sum_{q=1}^c \exp\left(\frac{1}{H} \sum_{k=1}^H f_q(\text{Aug}_w(\mathbf{x})_k; \theta)\right)}, \quad (6)$$

where $f_j(\text{Aug}_w(\mathbf{x})_k; \theta)$ is the model’s raw output (logit) for label y_j on the k -th weak augmentation of \mathbf{x} .

Then, we define the unsupervised consistency loss, \mathcal{L}_{uc} , which aligns the PLD with the predictions on strongly augmented versions of the same instances [52]. It is expressed as:

$$\mathcal{L}_{uc} = \frac{1}{m} \sum_{u=1}^m \sum_{j=1}^c p_{\mathbf{x}_u}^{y_j} \ln \left(\frac{p_{\mathbf{x}_u}^{y_j}}{h(y_j | \text{Aug}_s(\mathbf{x}_u); \theta)} \right), \quad (7)$$

where $h(y_j | \text{Aug}_s(\mathbf{x}_u); \theta)$ is the predicted importance degree for label y_j after applying strong augmentation to \mathbf{x}_u . This loss encourages the model to exploit the underlying structure of the unlabeled data, improving learning from these instances.

In the unsupervised component, we adopt a more flexible strategy to utilize PLDs for training. Recognizing the potential inaccuracies in pseudo-labels, we focus on aligning the ranking relationships among labels rather than enforcing strict adherence to absolute values. To achieve this, we propose the unsupervised pairwise relevance ranking loss, \mathcal{L}_{PRR_u} , which prioritizes capturing inter-label ranking while ignoring margins. Let $h_j(\mathbf{x}_i^s)$ denote the predicted relevance of the j -th label after strong augmentation Aug_s . The loss is defined as:

$$\mathcal{L}_{PRR_u} = \sum_{1 \leq j < k \leq c} \left(s(j, k) \cdot g_0(j, k) + s(k, j) \cdot g_0(k, j) \right), \quad (8)$$

where:

$$s(j, k) = \begin{cases} 1, & \text{if } p_{\mathbf{x}_i}^{y_j} > p_{\mathbf{x}_i}^{y_k} \text{ and } p_{\mathbf{x}_i}^{y_j} - p_{\mathbf{x}_i}^{y_k} > t, \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

$$g_0(j, k) = \begin{cases} 0, & \text{if } h_j(\mathbf{x}_i^s) - h_k(\mathbf{x}_i^s) \geq 0, \\ h_k(\mathbf{x}_i^s) - h_j(\mathbf{x}_i^s), & \text{otherwise.} \end{cases} \quad (10)$$

As shown in Fig. 2(b), we illustrate the unlabeled Pairwise Relevance Ranking loss \mathcal{L}_{PRR_u} . In this example a sample from the RAF-LDL dataset the pseudolabel distribution is $(0.6, 0.2, \dots)$. The loss only penalizes cases where the predicted ranking conflicts with the order suggested by the pseudolabel distribution, i.e., when a label with higher pseudolabel score is ranked lower (Case.1), and it does not impose any margin constraints. Our experiments demonstrate that this flexible use of pseudolabels significantly improves SSLDL performance, as evidenced by the faster convergence and lower KL divergence shown in the top-right inset of Fig. 2(b).

Finally Loss Function: Overall, the RankMatch algorithm utilizes a dual-phase training strategy to effectively differentiate between labeled and unlabeled data. The combined application of supervised and unsupervised ranking losses under the PRR framework is modulated by a hyperparameter λ . The total loss is computed as follows:

$$\mathcal{L}_{total} = \mathcal{L}_s + \mathcal{L}_{uc} + \lambda(\mathcal{L}_{PRR_L} + \mathcal{L}_{PRR_u}), \quad (11)$$

5 Theoretical Analysis

In this section, we first investigate how the proposed PRR loss influences the generalization behavior of the SSLDL framework. Intuitively, incorporating PRR encourages the model to capture inter-label correlations and refine label ranking consistency, which helps the network generalize beyond

the labeled set. To formalize this intuition, we derive the following theorem, which shows that adding the PRR term leads to a tighter generalization bound compared with the KL-only objective.

Theorem 5.1. *Let \mathcal{F} be a hypothesis class of scoring functions, and define the empirical risks \hat{R}_{KL} , \hat{R}_{PRR} , and the combined risk \hat{R}_{tot} as above. Denote the corresponding minimizers f_{KL} and $f_{\text{KL+PRR}}$. Then, with probability at least $1 - \delta$,*

$$R_{\text{tot}}(f_{\text{KL+PRR}}) < R_{\text{tot}}(f_{\text{KL}}) + 2\mathfrak{R}_{n+m}(\ell \circ \mathcal{F}) + B\sqrt{\frac{\ln(1/\delta)}{2(n+m)}}.$$

This result theoretically confirms that when the PRR term effectively reduces the empirical total risk the overall population risk under PRR regularization becomes strictly smaller than that of the KL-only formulation, up to the standard complexity and confidence terms. In other words, the PRR loss not only improves empirical optimization but also strengthens the generalization guarantee of the model. This provides a theoretical foundation for the performance gains observed in our experiments.

Next we establish a theoretical foundation for our RankMatch by defining a generalization bound.

Theorem 5.2. *Let f^* be the true risk minimizer and \hat{f} the empirical risk minimizer. Assume the loss function $\ell(\cdot)$ is bounded by B and that the pseudo-labeling error ϵ satisfies $\sum_{j=1}^m |\mathbb{I}(f_k(\mathbf{x}_j)) - \mathbb{I}(d_{\mathbf{x}_j}^{y_k})| / m \leq \epsilon$ for all $k \in [q]$. For a given Lipschitz constant L_E , Rademacher complexity $R_N(\mathcal{F})$ of the function class \mathcal{F} , and confidence parameter $\delta > 0$, the generalization gap is bounded as:*

$$R(\hat{f}) - R(f^*) \leq 2qB\epsilon + 4qL_ER_N(\mathcal{F}) + 2qB\sqrt{\frac{\log \frac{2}{\delta}}{2N}},$$

where $N = m + n$ is the total number of labeled and unlabeled samples. Theorem 3.2 provides a theoretical guarantee on the performance of the proposed *RankMatch* algorithm. Furthermore, it highlights key factors influencing the generalization error in SSLDL, including the pseudo-labeling error ϵ , the complexity of the hypothesis space captured by the Rademacher complexity $R_N(\mathcal{F})$, and the total number of training samples N . Moreover, increasing the training set size N further tightens the bound, reinforcing the benefits of leveraging large-scale unlabeled data in SSLDL. All the proof detail can be find in Appendix A and B.

6 Experiments

Experimental Datasets. We evaluate our approach on four real-world datasets: *Twitter-LDL* [61] (10,045 Twitter images labeled for eight emotions), *Flickr-LDL* [61] (10,700 Flickr images annotated for eight emotions by 11 annotators), *Emotion6* [43] (1,980 Flickr images labeled for six emotions), and *RAF-LDL* [34] (5,000 multi-label facial expression images).

Implementation Following [5, 58, 65], we employ ResNet-50 [16, 44, 55] pre-trained on ImageNet [32, 57] for training the classification model. For training images, we adopt standard flip-and-shift strategy [52, 51, 64] for weak data augmentation, and RandAugment [7, 46] and Cutout [8, 14, 3] for strong data augmentation. We employ AdamW [62, 45] optimizer and one-cycle policy scheduler [15] to train the model with maximal learning rate of 0.0001. For all datasets, the number of epochs is set as 30 and the batch size is set as 32. Furthermore, we perform exponential moving average (EMA) [25, 40] for the model parameter θ with a decay of 0.98. We adjust the parameter λ across a range of values, specifically $\{0.005, 0.01, 0.05, 0.1\}$. We perform all experiments on GeForce RTX 3090 GPUs. The random seed is set to 1 for all experiments. The datasets detail and the other implementation detail can be find in Appendix C.

Comparing Methods. To evaluate the effectiveness of our proposed RankMatch method, we benchmark it against four distinct groups of algorithms:

- *Semi-Supervised Multi-Label Learning (SSMLL) Algorithms:* We introduce two advanced algorithms, SSMLL-CAP(CAP) [59] and PCLP [36], developed to address the challenges of semi-supervised multi-label learning by improving the reliability of pseudo-labeling and leveraging label correlations within multi-label datasets.


Table 1: Comparison of testing results on the Emotion6, Flickr, RAF, and Twitter datasets using Canberra, Clark, Intersection, and Cosine metrics. The table reports performance under different labeled data proportions (10%, 20%, and 40%) used for training. The best performance in each metric is highlighted in bold.


	Method	Emotion6			Flickr-LDL			Twitter-LDL			RAF-LDL		
		10%	20%	40%	10%	20%	40%	10%	20%	40%	10%	20%	40%
Can. ↓	Rankmatch	3.3902	3.3176	3.2504	4.4060	3.9964	3.9013	3.7370	3.6962	3.2913	3.0178	2.9358	2.8341
	SSMLL-CAP	3.7951	3.7613	3.7248	5.3827	5.3235	5.2676	5.8983	5.7659	5.6366	3.4385	3.2808	3.1966
	PCLP	3.7011	3.6017	3.6030	5.2781	5.2292	5.1966	5.4909	5.3738	5.4133	3.3696	3.3383	3.3310
	Fixmatch-LDL	3.5080	3.5680	3.6050	5.5570	5.5310	5.4350	6.1750	6.0060	5.8340	3.1220	3.0920	3.0770
	Mixmatch-LDL	3.6080	3.4860	3.4880	5.6450	5.5026	5.5750	6.3530	6.2489	6.2960	3.1580	3.1111	3.0630
	GCT-LDL	3.5980	3.5490	3.6410	5.5860	5.5872	5.5260	6.3010	6.3078	6.2380	3.1920	3.1260	3.1470
	SALDL	3.4836	3.3737	3.1931	5.4612	4.7789	4.8199	5.0380	4.0868	4.0742	3.1947	3.1415	3.0527
	sLDLF	4.4164	4.3398	4.1322	6.2280	6.1238	6.2589	5.3084	6.0008	6.1910	4.0586	4.1705	4.1189
	DF-LDL	4.2427	4.0717	3.7221	5.5348	5.5549	5.5207	6.4184	6.3120	6.2588	3.3281	3.3865	3.3582
	LDL-LRR	4.6528	4.0496	3.7719	5.6325	5.4988	5.4319	6.4215	6.3295	6.2905	3.8677	4.0116	4.1890
Cla. ↓	Adam-LDL-SCL	4.0815	4.1128	4.1204	6.1634	5.9889	5.6508	6.5220	6.4081	6.3575	3.0891	3.0242	2.9912
	Rankmatch	1.5298	1.5050	1.4834	1.8189	1.7051	1.6737	1.6480	1.6190	1.5138	1.4506	1.4190	1.3843
	SSMLL-CAP	1.6705	1.6611	1.6502	2.1222	2.0988	2.0820	2.2590	2.2155	2.1733	1.5918	1.5332	1.5082
	PCLP	1.6397	1.6059	1.6083	2.0601	2.0478	2.0328	2.1002	2.0623	2.0728	1.5689	1.5636	1.5593
	Fixmatch-LDL	1.5950	1.6230	1.6390	2.2220	2.2110	2.1910	2.3830	2.3310	2.2820	1.5130	1.5060	1.5050
	Mixmatch-LDL	1.6240	1.5810	1.5840	2.2330	2.1996	2.2160	2.4280	2.4034	2.4150	1.5150	1.5020	1.4870
	GCT-LDL	1.6090	1.6050	1.6390	2.2200	2.2238	2.2080	2.4170	2.4216	2.4060	1.5350	1.5170	1.5290
	SALDL	1.6019	1.5751	1.5100	2.1967	2.0369	2.0446	2.1288	1.8938	1.8964	1.5445	1.5288	1.5035
	sLDLF	1.8922	1.8566	1.8049	2.3722	2.3436	2.3761	2.1480	2.3384	2.3746	1.9300	1.9645	1.9750
	DF-LDL	1.8217	1.7746	1.6781	2.2253	2.2072	2.1992	2.4313	2.4108	2.4033	1.6071	1.6229	1.6138
Int. ↑	LDL-LRR	1.9899	1.7745	1.6953	2.2285	2.2026	2.1919	2.4429	2.4223	2.4121	1.7907	1.8298	1.8919
	Adam-LDL-SCL	1.7851	1.7976	1.8014	2.3534	2.3093	2.2312	2.4639	2.4324	2.4160	1.5134	1.4980	1.4905
	Rankmatch	0.6735	0.6832	0.6940	0.6921	0.7073	0.7151	0.7036	0.7190	0.7316	0.6551	0.6813	0.7044
	SSMLL-CAP	0.5479	0.5587	0.5666	0.5815	0.6125	0.6377	0.6034	0.6324	0.6577	0.5264	0.5876	0.6092
	PCLP	0.6059	0.6370	0.6363	0.6392	0.6469	0.6490	0.6707	0.6784	0.6780	0.5471	0.5588	0.5590
	fixmatch-LDL	0.6638	0.6797	0.6916	0.6857	0.7042	0.7119	0.7009	0.7147	0.7283	0.6570	0.6760	0.6987
	Mixmatch-LDL	0.6372	0.6418	0.6496	0.6639	0.6686	0.6831	0.6819	0.6806	0.6986	0.6133	0.6381	0.6534
	GCT-LDL	0.6116	0.6602	0.6770	0.6639	0.6879	0.6863	0.6787	0.7018	0.7102	0.6321	0.6669	0.6910
	SALDL	0.6457	0.6612	0.6723	0.5559	0.5108	0.5091	0.6632	0.5724	0.5687	0.6298	0.6504	0.6708
	sLDLF	0.5935	0.5861	0.6162	0.4813	0.4750	0.4616	0.6487	0.5652	0.5336	0.2433	0.2315	0.2199
Cos. ↑	DF-LDL	0.5057	0.5461	0.6353	0.4173	0.4176	0.4169	0.3541	0.3536	0.3505	0.7022	0.7083	0.7085
	LDL-LRR	0.3721	0.6213	0.6626	0.5322	0.5519	0.5600	0.5746	0.5904	0.5979	0.5649	0.5389	0.4411
	Adam-LDL-SCL	0.3409	0.5627	0.6040	0.4724	0.3933	0.4628	0.5488	0.5828	0.5200	0.6177	0.5768	0.4843
	Rankmatch	0.8121	0.8257	0.8331	0.8489	0.8614	0.8679	0.8544	0.8698	0.8790	0.7901	0.8140	0.8375
	SSMLL-CAP	0.6850	0.6994	0.7185	0.7634	0.7885	0.8144	0.8109	0.8270	0.8442	0.6456	0.7119	0.7329
	PCLP	0.7421	0.7737	0.7778	0.8057	0.8146	0.8151	0.8391	0.8436	0.8448	0.6815	0.6962	0.6969
	Fixmatch-LDL	0.8079	0.8200	0.8312	0.8487	0.8573	0.8673	0.8517	0.8647	0.8758	0.7881	0.8123	0.8311
	Mixmatch-LDL	0.7585	0.7863	0.7901	0.7888	0.8381	0.8468	0.8463	0.8552	0.8602	0.7536	0.7680	0.7820
	GCT-LDL	0.7530	0.8017	0.8134	0.8313	0.8508	0.8531	0.8499	0.8587	0.8716	0.7660	0.7977	0.8181
	SALDL	0.7784	0.7874	0.7981	0.7361	0.6643	0.6624	0.8479	0.7612	0.7615	0.7711	0.7938	0.8135
	sLDLF	0.7037	0.6980	0.7350	0.6276	0.6066	0.5897	0.8002	0.7454	0.6988	0.3262	0.3506	0.3459
	DF-LDL	0.6035	0.6470	0.7689	0.5436	0.5539	0.5569	0.5069	0.5233	0.5209	0.8427	0.8492	0.8470
	LDL-LRR	0.4604	0.7362	0.7905	0.7020	0.7316	0.7399	0.7767	0.8027	0.8125	0.7253	0.6938	0.5757
	Adam-LDL-SCL	0.4311	0.6670	0.7144	0.6104	0.4888	0.6166	0.7163	0.7661	0.7403	0.7717	0.7337	0.6191


Table 2: Evaluation of Label Distribution Ranking Relationships for Test Samples. The table compares Kendall tau (τ_K) and Spearmans rank (ρ_S) correlation coefficients across datasets and different methods. Bold values indicate the best-performing method for each dataset and metric.

Metric	Dataset	10% Labeled Data						20% Labeled Data					
		RankMatch	GCT	CAP	PCLP	FixMatch	MixMatch	RankMatch	GCT	CAP	PCLP	FixMatch	MixMatch
τ_K	RAF	0.5696	0.4258	0.2079	0.2750	0.4643	0.4066	0.5463	0.4811	0.3598	0.2949	0.4946	0.4524
	Emotion6	0.5535	0.3718	0.1237	0.3325	0.4899	0.4562	0.5620	0.4594	0.1536	0.4394	0.4985	0.4528
	Flickr	0.5618	0.5005	0.4030	0.4904	0.5215	0.5119	0.5627	0.5215	0.4475	0.5005	0.5416	0.5265
	Twitter	0.4927	0.4806	0.4407	0.4887	0.4744	0.5016	0.5452	0.5121	0.4828	0.5037	0.5039	0.5177
ρ_S	RAF	0.6726	0.5122	0.2474	0.3365	0.5564	0.4900	0.6649	0.5754	0.4297	0.3602	0.5917	0.5497
	Emotion6	0.6545	0.4495	0.1529	0.4037	0.5933	0.5528	0.6512	0.5624	0.1923	0.5332	0.5989	0.5559
	Flickr	0.6537	0.5904	0.4831	0.5793	0.6097	0.6019	0.6551	0.6112	0.5335	0.5903	0.6304	0.6171
	Twitter	0.5740	0.5637	0.5193	0.5735	0.5517	0.5864	0.6315	0.5966	0.5658	0.5898	0.5853	0.6024

- *Dual-Network SSLDL Algorithm:* We present and evaluate our own GCT-LDL(GCT), a dual-network [4] SSLDL approach that we developed, which leverages mutual supervision of unlabeled data between two independent networks for enhanced learning.
- *Deep Learning SSLDL Algorithms:* We introduce two novel algorithms, FixMatch-LDL [52] and MixMatch-LDL [2], designed to bridge the gap in open-source semi-supervised LDL (SSLDL) approaches within deep learning frameworks.

		anger	disgust	fear	joy	sad	surprise	neutral
	Ground Truth	0.07	0.17	0.06	0.03	0.36	0.03	0.28
	Rankmatch-LDL	0.02	0.04	0.05	0.08	0.50	0.08	0.23
	Fixmatch-LDL	0.03	0.08	0.13	0.17	0.16	0.16	0.27
	Mixmatch-LDL	0.04	0.06	0.28	0.12	0.23	0.09	0.18
	GCT-LDL	0.04	0.06	0.12	0.24	0.15	0.10	0.29

		surprise	fear	disgust	happy	sad	anger
	Ground Truth	0.45	0.52	0.00	0.00	0.00	0.03
	Rankmatch-LDL	0.37	0.52	0.02	0.00	0.08	0.01
	Fixmatch-LDL	0.45	0.32	0.04	0.03	0.14	0.02
	Mixmatch-LDL	0.39	0.10	0.19	0.09	0.13	0.10
	GCT-LDL	0.30	0.19	0.05	0.03	0.42	0.01

		anger	disgust	fear	joy	sad	surprise	neutral
	Ground Truth	0.00	0.00	0.00	0.63	0.00	0.20	0.17
	Rankmatch-LDL	0.01	0.04	0.11	0.34	0.07	0.20	0.23
	Fixmatch-LDL	0.02	0.12	0.06	0.20	0.06	0.09	0.45
	Mixmatch-LDL	0.04	0.17	0.07	0.18	0.10	0.09	0.35
	GCT-LDL	0.04	0.14	0.10	0.23	0.13	0.11	0.25


		surprise	fear	disgust	happy	sad	anger
	Ground Truth	0.27	0.00	0.03	0.67	0.03	0.00
	Rankmatch-LDL	0.35	0.02	0.02	0.59	0.00	0.02
	Fixmatch-LDL	0.64	0.10	0.02	0.22	0.01	0.01
	Mixmatch-LDL	0.55	0.13	0.07	0.06	0.15	0.03
	GCT-LDL	0.44	0.17	0.08	0.12	0.15	0.04

Figure 3: Examples illustrating *RankMatch-LDL*’s ability to generate superior pseudo-label distributions compared to existing semi-supervised methods. The first two images are from the Emotion6 [43] dataset, and the last two are from the RAF dataset [34]. *RankMatch-LDL* more accurately aligns with the ground truth label ranking.

- *Traditional SSLDL Algorithm:* The traditional SA-LDL [18] algorithm, originally for tabular data, is adapted for image datasets through necessary feature engineering, detailed in Appendix D.
- *SOTA LDL Algorithms:* Comparisons are also made with state-of-the-art LDL algorithms including Adam-LDL-SCL [20], sLDLF [50], DF-LDL [13], and LDL-LRR [21], highlighting their potential limitations in SSLDL contexts.

Evaluation Metrics. We evaluate LDL methods using eight metrics [10]: Chebyshev, Clark, Canberra distances, and Kullback-Leibler divergence (lower is better), along with Intersection and Cosine similarities, Spearmans rank correlation (ρ_S), and Kendall tau correlation (τ_K) [21] (higher is better).

6.1 Comparative Experiment Analysis

We employed a range of labeled data proportions (10%, 20%, and 40%) to simulate varying levels of label availability, a critical factor in semi-supervised learning scenarios. The experiments are presented in Table 1 and Table 2. from that we can draw the following conclusions

- RankMatch consistently achieves top performance across all datasets and metrics. Compared to SSMLL-CAP and PCLP, RankMatch shows stronger label relationship modeling, leveraging PRR losses to refine pseudo-label rankings and outperforming traditional SSLDL methods like FixMatch-LDL and MixMatch-LDL.
- The results in Table 2 confirm the consistent superiority of *RankMatch* in capturing label ranking relationships, achieving the highest Kendall tau (τ_K) and Spearmans rank (ρ_S) correlation coefficients across almost all datasets and metrics. Notably, *RankMatch* maintains robust performance as the proportion of labeled data increases, demonstrating strong generalization and adaptability under varying supervision levels.
- With increasing labeled data, RankMatch scales effectively, achieving significant performance improvements. On Twitter, the Canberra distance improves from 3.7370 (10% labeled data) to 3.2913 (40% labeled data).

6.2 Analysis of Pseudo-Label Performance

To evaluate the performance of pseudo-labeling, we assess the pseudo-labeling quality of different algorithms and visualize two sample images from the Emotion6 and RAF datasets as examples. The experimental results are presented in Table 3 and Fig. 3. Based on these results, we derive the following conclusions:

- Our method effectively captures the true label distribution, achieving the best performance in overall distance metrics, as evidenced by the lowest KL divergence across all datasets.
- By incorporating the PRR loss, our approach generates pseudo-labels with ranking structures that more closely align with the ground truth, leading to higher-quality pseudo-labels and improved model performance.

Table 3: Comparison of pseudo-labeling performance with different methods trained on 10% and 20% labeled data. Kendalls Tau (τ_K) and Spearmans rank correlation (ρ_S) measure ranking quality (higher is better), while KL divergence quantifies distribution alignment (lower is better).

Dataset	10% Labeled Data						20% Labeled Data					
	RankMatch	FixMatch	MixMatch	GCT	CAP	PCLP	RankMatch	FixMatch	MixMatch	GCT	CAP	PCLP
ρ_S												
Emotion6	0.6266	0.5372	0.4763	0.3470	0.1838	0.2921	0.6287	0.5685	0.4760	0.4929	0.4449	0.3325
RAF	0.6617	0.5348	0.5024	0.4774	0.2593	0.3471	0.6730	0.5772	0.5480	0.5730	0.2354	0.4313
Flickr	0.6646	0.6262	0.6153	0.6024	0.4902	0.5933	0.6618	0.6490	0.6377	0.6328	0.5528	0.5988
Twitter	0.6446	0.5551	0.6006	0.5729	0.5197	0.5681	0.6440	0.5990	0.6147	0.6108	0.5602	0.5830
KL												
Emotion6	2.5458	3.4885	3.9461	5.1156	5.6353	5.0965	2.4344	3.0090	4.1522	3.7812	3.5549	4.8452
RAF	2.0564	2.9951	3.9584	3.5445	4.9070	4.5420	1.9380	2.5717	3.1612	2.7052	5.4632	3.9629
Flickr	2.6495	3.7454	3.6983	4.2128	6.0978	3.7667	2.6802	2.9860	4.0924	3.7720	5.2461	3.7492
Twitter	2.5243	3.1827	3.4360	4.1143	6.5147	3.2599	2.5900	3.1581	3.9771	3.2475	5.7258	3.0865

Table 4: Ablation Results on Flickr and RAF Datasets.

		Che.↓	Cla.↓	Can.↓	KL↓	Cos.↑	Int.↑
Flickr	pretrain	0.2411	2.2594	5.6885	0.5371	0.8427	0.6873
	pretrain + consistency	0.2262(6.2%↑)	2.1131(6.5%↑)	5.1536(9.4%↑)	0.5293(1.5%↑)	0.8633(2.4%↑)	0.7188(4.6%↑)
	pretrain + consistency+PRR loss	0.2184(3.4%↑)	2.0158(4.6%↑)	4.9008(4.9%↑)	0.5227(1.2%↑)	0.8714(0.9%↑)	0.7208(0.3%↑)
RAF	pretrain	0.2938	1.5412	3.206	0.5146	0.7687	0.6411
	pretrain + consistency	0.255(13.2%↑)	1.5021(2.5%↑)	3.1345(2.2%↑)	0.3699(28.1%↑)	0.8189(28.1%↑)	0.7073(10.3%↑)
	pretrain + consistency+PRR loss	0.2341(8.2%↑)	1.4914(0.7%↑)	3.0459(2.8%↑)	0.3464(6.4%↑)	0.8476(3.5%↑)	0.7194(1.7%↑)

6.3 Further Analysis

Ablation Study Our ablation study analyzed the impact of PRR loss and unsupervised consistency loss on the performance of RankMatch. Initially, the model was pre-trained with only 10% of labeled data to establish a baseline. This phase highlighted the model’s ability to utilize minimal data effectively.

Next, unsupervised consistency loss was applied to enhance learning from unlabeled data. In the final phase, PRR loss was introduced, leveraging the same 10% labeled data to refine the model further with supervised ranking loss. Ablation experiment results are shown in Table 4. From this, we can draw the following conclusions

- The integration of unsupervised consistency loss markedly improves RankMatch’s performance across datasets, as observed in the ablation results. This confirms the effectiveness of using unsupervised data to enhance model accuracy.
- The incorporation of pairwise relevance ranking (PRR) loss significantly boosts performance, particularly in scenarios where it surpasses the baseline. This improvement demonstrates the PRR loss’s critical role in refining label discrimination within the semi-supervised learning framework.

The Impact of \mathcal{L}_{PRR} Loss on Pseudo-Labeling: We evaluated the effect of \mathcal{L}_{PRR} loss on pseudo-labeling by comparing models trained with and without it on the RAF and Emotion6 datasets. Pseudo-labels were generated for part of the validation set during training, and their ranking performance was analyzed. As shown in Fig. 4, incorporating PRR loss consistently improved the ranking quality of pseudo-labels, aligning them more closely with the ground truth. This demonstrates the ability of \mathcal{L}_{PRR} loss to effectively capture inter-label ranking relationships, thereby enhancing both the training process and the overall model performance.

Parameter Sensitivity Analysis Fig. 5 illustrates the impact of the parameter λ on RankMatch’s performance across the Emotion6, Flickr-LDL, RAF-LDL, and Twitter-LDL datasets, focusing on KL divergence and Cosine similarity metrics. Fig. 5 shows that RankMatch performs consistently well when λ ranges from 0.01 to 0.05, with minimal variations in performance metrics such as KL

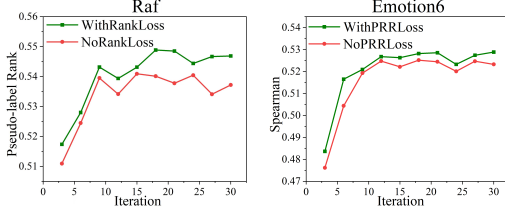


Figure 4: The effect of incorporating *PRR Loss* on the ranking performance of pseudo-labels during training.

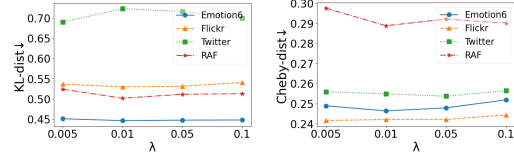


Figure 5: Parameter sensitivity analysis on Emotion6, Flickr-LDL, RAF-LDL, and Twitter-LDL datasets.

divergence, Intersection, and Cosine similarity across all datasets. However, at $\lambda = 0.005$, a noticeable performance drop, particularly in the Emotion6 dataset, highlights the reduced effectiveness of regularization. Conversely, higher values, such as $\lambda = 0.1$, slightly hinder performance, suggesting over-regularization. This analysis indicates that λ values between 0.01 and 0.05 strike the best balance for effective learning. Detailed results for additional metrics can be found in Appendix D.

7 Conclusion and Limitations

Conclusion. In this paper, we introduce RankMatch, the first deep-learningbased semi-supervised label distribution learning (SSLDL) method that explicitly models inter-label ranking relationships via pseudo-labels. By combining the standard KL-divergence loss with our novel Pairwise Ranking Relationship (PRR) loss within a single training framework, RankMatch produces higher-quality pseudo-label distributions and flexibly leverages unlabeled data without overfitting to noisy absolute values. In our theoretical analysis, we prove that adding the PRR loss tightens the models generalization bound and provide an explicit generalization error bound for RankMatch. Finally, we empirically validate on four real-world LDL benchmarks (Twitter-LDL, Flickr-LDL, Emotion6, and RAF-LDL) that RankMatch consistently outperforms all baselines.

Limitations. While SSLDL significantly lowers the demand for fully annotated data, it still depends on a nontrivial amount of manual labeling. Going forward, we plan to investigate how to harness large language models to generate cost-effective, high-quality pseudo-labels and to integrate LLMs directly into the SSLDL training pipeline.

Acknowledgments

This work was supported in part by the Jiangsu Science Foundation (BG2024036, BK20243012), the National Natural Science Foundation of China (62125602, U24A20324, 92464301, 62306073), the China Postdoctoral Science Foundation (2022M720028), the Xplorer Prize, and the National Natural Science Foundation of China under Grant U24A20322. Additional support was provided by the Big Data Computing Center of Southeast University.

References

- [1] Hritam Basak and Zhaozheng Yin. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19786–19797. IEEE, 2023.
- [2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.
- [3] Linbo Cao and Jinman Zhao. Pretraining on the test set is no longer all you need: A debate-driven approach to QA benchmarks. In *Second Conference on Language Modeling*, 2025.
- [4] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *IEEE Conference on Computer Vision and Pattern*

- Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2613–2622. Computer Vision Foundation / IEEE, 2021.
- [5] Elijah Cole, Oisín Mac Aodha, Titouan Lorieu, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 933–942, 2021.
 - [6] Richard Combes. An extension of mediarimid’s inequality. *arXiv preprint arXiv:1511.05240*, 2015.
 - [7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.
 - [8] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.
 - [9] Enrico Fini, Pietro Astolfi, Karteek Alahari, Xavier Alameda-Pineda, Julien Mairal, Moin Nabi, and Elisa Ricci. Semi-supervised learning made simple with self-supervised clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 3187–3197. IEEE, 2023.
 - [10] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016.
 - [11] Xin Geng, Xin Qian, Zengwei Huo, and Yu Zhang. Head pose estimation based on multivariate label distribution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
 - [12] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE transactions on pattern analysis and machine intelligence*, 35(10):2401–2412, 2013.
 - [13] Manuel González, Germán González-Almagro, Isaac Triguero, José-Ramón Cano, and Salvador García. Decomposition-fusion for label distribution learning. *Information Fusion*, 66:64–75, 2021.
 - [14] Shunxin Guo, Hongsong Wang, Shuxia Lin, Zhiqiang Kou, and Xin Geng. Addressing skewed heterogeneity via federated prototype rectification with personalization. *IEEE Transactions on Neural Networks and Learning Systems*, 36(5):8442–8454, 2025.
 - [15] Mahammad A Hannan, Dickson NT How, Muhamad Bin Mansor, Md S Hossain Lipu, Pin Jern Ker, and Kashem M Muttaqi. State-of-charge estimation of li-ion battery using gated recurrent unit with one-cycle learning rate policy. *IEEE Transactions on Industry Applications*, 57(3):2964–2971, 2021.
 - [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
 - [17] John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP’07*, volume 4, pages IV–317. IEEE, 2007.
 - [18] Peng Hou, Xin Geng, Zeng-Wei Huo, and Jia-Qi Lv. Semi-supervised adaptive label distribution learning for facial age estimation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
 - [19] Xiuyi Jia, Weiwei Li, Junyu Liu, and Yu Zhang. Label distribution learning by exploiting label correlations. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, volume 32, 2018.
 - [20] Xiuyi Jia, Zechao Li, Xiang Zheng, Weiwei Li, and Sheng-Jun Huang. Label distribution learning with label correlations on local samples. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1619–1631, 2019.

- [21] Xiuyi Jia, Xiaoxia Shen, Weiwei Li, Yunan Lu, and Jihua Zhu. Label distribution learning by maintaining label ranking relation. *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- [22] Xiuyi Jia, Tao Wen, Weiping Ding, Huaxiong Li, and Weiwei Li. Semi-supervised label distribution learning via projection graph embedding. *Information Sciences*, 581:840–855, 2021.
- [23] Yangbangyan Jiang, Xiaodan Li, Yuefeng Chen, Yuan He, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. Maxmatch: Semi-supervised learning with worst-case consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5970–5987, 2022.
- [24] Jae-Yong Kim, Eun-Young Lee, Jin Kyun Park, Yeong Wook Song, Jae-Ryong Kim, and Kyung-Hyun Cho. Patients with rheumatoid arthritis show altered lipoprotein profiles with dysfunctional high-density lipoproteins that can exacerbate inflammatory and atherogenic process. *PLoS One*, 11(10):e0164564, 2016.
- [25] Frank Klinker. Exponential moving average versus moving exponential average. *Mathematische Semesterberichte*, 58:97–107, 2011.
- [26] Zhiqiang Kou, Si Qin, Hailin Wang, Jing Wang, Mingkun Xie, Shuo Chen, Yuheng Jia, Tongliang Liu, Masashi Sugiyama, and Xin Geng. Label distribution learning with biased annotations assisted by multi-label learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 8 2025.
- [27] Zhiqiang Kou, Jing Wang, Yuheng Jia, and Xin Geng. Inaccurate label distribution learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):10237–10249, 2024.
- [28] Zhiqiang Kou, Jing Wang, Yuheng Jia, and Xin Geng. Progressive label enhancement. *Pattern Recognition*, 160:111172, 2025.
- [29] Zhiqiang Kou, Jing Wang, Yuheng Jia, Biao Liu, and Xin Geng. Instance-dependent inaccurate label distribution learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):1425–1437, 2025.
- [30] Zhiqiang Kou, Jing Wang, Jiawei Tang, Yuheng Jia, Boyu Shi, and Xin Geng. Exploiting multi-label correlation in label distribution learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 4326–4334, 8 2024.
- [31] Zhiqiang Kou, Haoyuan Xuan, Jingyu Zhu, Hailin Wang, Ming-kun Xie, Changwei Wang, Jing Wang, Yuheng Jia, and Xin Geng. Tail-aware reconstruction of incomplete label distributions with low-rank and sparse modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2025.
- [32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [33] Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry S Davis. Rethinking pseudo labels for semi-supervised object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 1314–1322, 2022.
- [34] Shan Li and Weihong Deng. Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. *International Journal of Computer Vision*, 127(6-7):884–906, 2019.
- [35] Miaogen Ling and Xin Geng. Indoor crowd counting by mixture of gaussians label distribution learning. *IEEE Transactions on Image Processing*, 28(11):5691–5701, 2019.
- [36] Biao Liu, Ning Xu, Xiangyu Fang, and Xin Geng. Correlation-induced label prior for semi-supervised multi-label learning. In *Forty-first International Conference on Machine Learning*.
- [37] Biao Liu, Ning Xu, Xiangyu Fang, and Xin Geng. Correlation-induced label prior for semi-supervised multi-label learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 32224–32238, 21–27 Jul 2024.

- [38] Biao Liu, Ning Xu, Jiaqi Lv, and Xin Geng. Revisiting pseudo-label for single-positive multi-label learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 22249–22265, 23–29 Jul 2023.
- [39] Xinyuan Liu, Jihua Zhu, Qinghai Zheng, Zhiqiang Tian, and Zhongyu Li. Semi-supervised label distribution learning with co-regularization. *Neurocomputing*, 491:353–364, 2022.
- [40] Lei Meng, Zhuang Qi, Lei Wu, Xiaoyu Du, Zhaochuan Li, Lizhen Cui, and Xiangxu Meng. Improving global generalization and local personalization for federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36, 2024.
- [41] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.
- [42] Shaunna M Morrison, Feifei Pan, Olivier C Gagné, Anirudh Prabhu, Ahmed Eleish, Peter Arthur Fox, Robert T Downs, Thomas Bristow, Elizabeth B Rampe, David Frederick Blake, et al. Predicting multi-component mineral compositions in gale crater, mars with label distribution learning. In *AGU Fall Meeting Abstracts*, volume 2018, pages P21I–3438, 2018.
- [43] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadvnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 860–868, 2015.
- [44] Zhuang Qi, Lei Meng, Zitan Chen, Han Hu, Hui Lin, and Xiangxu Meng. Cross-silo prototypical calibration for federated learning with non-iid data. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3099–3107, 2023.
- [45] Zhuang Qi, Lei Meng, Zhaochuan Li, Han Hu, and Xiangxu Meng. Cross-silo feature space alignment for federated learning on clients with imbalanced data. In *The 39th Annual AAAI Conference on Artificial Intelligence (AAAI-25)*, pages 19986–19994, 2025.
- [46] Zhuang Qi, Sijin Zhou, Lei Meng, Han Hu, Han Yu, and Xiangxu Meng. Federated deconfounding and debiasing learning for out-of-distribution generalization. *arXiv preprint arXiv:2505.04979*, 2025.
- [47] Congyu Qiao, Ning Xu, and Xin Geng. Decompositional generation process for instance-dependent partial label learning. In *International Conference on Learning Representations*, 2023.
- [48] Congyu Qiao, Ning Xu, Yihao Hu, and Xin Geng. Ulares: A unified label refinement framework for learning with inaccurate supervision. In *International Conference on Machine Learning*, 2024.
- [49] Congyu Qiao, Ning Xu, Jiaqi Lv, Yi Ren, and Xin Geng. Fredis: A fusion framework of refinement and disambiguation for unreliable partial label learning. In *International Conference on Machine Learning*, pages 28321–28336. PMLR, 2023.
- [50] Wei Shen, Kai Zhao, Yilu Guo, and Alan L Yuille. Label distribution learning forests. *Advances in neural information processing systems*, 30, 2017.
- [51] Boyu Shi, Shiyu Xia, Xu Yang, Haokun Chen, Zhiqiang Kou, and Xin Geng. Building variable-sized models via learngene pool. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13):14946–14954, Mar. 2024.
- [52] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.
- [53] Jing Wang and Xin Geng. Label distribution learning machine. In *Proc. Int. Conf. Mach. Learn.*, pages 10749–10759. PMLR, 2021.
- [54] Jing Wang, Zhiqiang Kou, Yuheng Jia, Jianhui Lv, and Xin Geng. Label enhancement by manifold fusion of feature and label spaces. *Pattern Recognition*, 168:111854, 2025.

- [55] Yining Wang, Jinman Zhao, Chuangxin Zhao, Shuhao Guan, Gerald Penn, and Shinan Liu. λ -grpo: Unifying the grpo frameworks with learnable token preferences, 2025.
- [56] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4248–4257, 2022.
- [57] Chunlin Wen, Yu Zhang, Jie Fan, Hongyuan Zhu, Xiu-Shen Wei, Yijun Wang, Zhiqiang Kou, and Shuzhou Sun. Object-level correlation for few-shot segmentation, 2025.
- [58] Ming-Kun Xie, Jia-Hao Xiao, Gang Niu, Lei Feng, Zhiqiang Kou, Min-Ling Zhang, and Masashi Sugiyama. What makes "good" distractors for object hallucination evaluation in large vision-language models?, 2025.
- [59] Ming-Kun Xie, Jiahao Xiao, Hao-Zhe Liu, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- [60] Miao Xu and Zhi-Hua Zhou. Incomplete label distribution learning. In *Proc. Int. Joint Conf. Artif. Intell.*, pages 3175–3181, 2017.
- [61] Jufeng Yang, Ming Sun, and Xiaoxiao Sun. Learning visual sentiment distributions via augmented conditional probability neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
- [62] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.
- [63] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.*, 26(8):1819–1837, 2014.
- [64] Xueyan Zhang, Jinman Zhao, Zhifei Yang, Yibo Zhong, Shuhao Guan, Linbo Cao, and Yining Wang. UORA: Uniform orthogonal reinitialization adaptation in parameter efficient fine-tuning of large models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11709–11728, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [65] Yibo Zhong, Jinman Zhao, and Yao Zhou. Low-rank interconnected adaptation across layers. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17005–17029, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [66] Xiuzhuang Zhou, Zeqiang Wei, Min Xu, Shan Qu, and Guodong Guo. Facial depression recognition by deep joint label distribution and metric learning. *IEEE Transactions on Affective Computing*, 13(3):1605–1618, 2020.
- [67] Zhi-Hua Zhou and Zhi-Hua Zhou. *Ensemble learning*. Springer, 2021.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the papers contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction clearly state our three major contributions(i) the RankMatch SSLDL framework, (ii) the novel PRR loss, and (iii) the theoretical generalization boundand these are fully supported by the experiments (Sec. 1).

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We include a dedicated Limitations paragraph at the end of Sec. 7, noting that SSLDL still requires some manual labels and outlining future directions for LLMbased low-cost annotation and integration of LLMs into the SSLDL loop.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: All assumptions are stated with each theorem (Sec. 5), and full proofs appear in Appendix A and B.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results?

Answer: [\[Yes\]](#)

Justification: We provide dataset splits, pseudocode, hyperparameters, training details in Sec. 6 and Appendix C.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to reproduce the main experimental results?

Answer: [\[Yes\]](#)

Justification: The benchmark datasets used in this work are publicly available, and the download links are provided in Appendix C. The implementation code and pretrained models will be released on our GitHub repository upon publication to facilitate full reproducibility.

6. Experimental setting/details

Question: Does the paper specify all training and test details necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: We describe data splits, augmentations, optimizers, and hyperparameter sweeps in Appendix C.

7. Experiment statistical significance

Question: Does the paper report error bars or other appropriate information about statistical significance?

Answer: [\[No\]](#)

Justification: For all experiments, we used a random seed of 1.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on compute resources needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix C reports GPU type (NVIDIA 3090), run time per epoch, and total training hours.

9. Code of ethics

Question: Does the research conform to the NeurIPS Code of Ethics?

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics and found no ethical issues.

10. Broader impacts

Question: Does the paper discuss both potential positive and negative societal impacts?

Answer: [NA]

Justification: We found no societal impact of this work.

11. Safeguards

Question: Does the paper describe safeguards for responsible release of data or models with high misuse risk?

Answer: [NA]

Justification: Our contributions do not pose highrisk dualuse scenarios requiring special release controls.

12. Licenses for existing assets

Question: Are existing assets properly credited with license information?

Answer: [Yes]

Justification: We cite all datasets (Twitter-LDL, RAF-LDL, etc.) and reference their original licenses (Sec. 6).

13. New assets

Question: Are new assets introduced in the paper well documented alongside the assets?

Answer: [NA]

Justification: his paper does not release new assets.

14. Crowdsourcing and human subjects

Question: For crowdsourcing experiments, does the paper include full instructions and compensation details?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

15. IRB approvals

Question: Does the paper describe IRB approvals or equivalent for research with human subjects?

Answer: [NA]

Justification: We did not conduct new humansubjects research.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if they are an important component of the core methods?

Answer: [NA]

Justification: Our core SSLDL method does not rely on external LLMs; future work will explore LLM-based labeling but that is out of scope here.

A The Proof of Theorem 3.1

We now show that adding the pairwise relevance ranking (PRR) loss to the usual KLdivergence objective strictly tightens the generalization bound in semisupervised label distribution learning.

Theorem A.1. *Let \mathcal{F} be a hypothesis class of scoring functions $f: \mathcal{X} \rightarrow \mathbb{R}^C$. Define the empirical risks on the labeled set $\{(x_i, d_i)\}_{i=1}^n$ and unlabeled set $\{x_j\}_{j=1}^m$ by*

$$\hat{R}_{\text{KL}}(f) = \frac{1}{n} \sum_{i=1}^n D_{\text{KL}}(d_i \| f(x_i)), \hat{R}_{\text{PRR}}(f) = \frac{1}{m} \sum_{j=1}^m L_{\text{PRR}}(f(x_j), \hat{d}_j),$$

where $D_{\text{KL}}(\cdot \| \cdot)$ is the KullbackLeibler divergence and L_{PRR} is the pairwise ranking loss computed against pseudo-labels \hat{d}_j . Let the combined empirical risk be

$$\hat{R}_{\text{tot}}(f) = \hat{R}_{\text{KL}}(f) + \lambda \hat{R}_{\text{PRR}}(f).$$

Denote

$$f_{\text{KL}} = \arg \min_{f \in \mathcal{F}} \hat{R}_{\text{KL}}(f), \quad f_{\text{KL+PRR}} = \arg \min_{f \in \mathcal{F}} \hat{R}_{\text{tot}}(f).$$

Assume all losses are bounded by B and Lipschitz continuous. Define the population risks

$$R_{\text{KL}}(f) = \mathbb{E}[D_{\text{KL}}(d \| f(x))], \quad R_{\text{PRR}}(f) = \mathbb{E}[L_{\text{PRR}}(f(x), d)], \quad R_{\text{tot}}(f) = R_{\text{KL}}(f) + \lambda R_{\text{PRR}}(f).$$

Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the training draw,

$$R_{\text{tot}}(f_{\text{KL+PRR}}) < R_{\text{tot}}(f_{\text{KL}}) + 2\mathfrak{R}_{n+m}(\ell \circ \mathcal{F}) + B\sqrt{\frac{\ln(1/\delta)}{2(n+m)}},$$

where $\mathfrak{R}_{n+m}(\ell \circ \mathcal{F})$ is the Rademacher complexity of the combined loss class $\ell(x, y) = D_{\text{KL}}(y \| x) + \lambda L_{\text{PRR}}(x, y)$. In particular, if adding L_{PRR} strictly lowers the empirical total risk, then the resulting population total risk is strictly smaller than that of the KL-only solution.

Proof. By standard Rademacher complexity bounds (see [41]), for every $f \in \mathcal{F}$, with probability at least $1 - \delta$,

$$R_{\text{tot}}(f) \leq \hat{R}_{\text{tot}}(f) + 2\mathfrak{R}_{n+m}(\ell \circ \mathcal{F}) + B\sqrt{\frac{\ln(1/\delta)}{2(n+m)}}.$$

By definition of the minimizers,

$$\hat{R}_{\text{KL}}(f_{\text{KL}}) \leq \hat{R}_{\text{KL}}(f_{\text{KL+PRR}}), \quad \hat{R}_{\text{tot}}(f_{\text{KL+PRR}}) \leq \hat{R}_{\text{tot}}(f_{\text{KL}}).$$

Subtracting these two inequalities shows that

$$\hat{R}_{\text{tot}}(f_{\text{KL+PRR}}) - \hat{R}_{\text{tot}}(f_{\text{KL}}) < 0 \implies \hat{R}_{\text{tot}}(f_{\text{KL+PRR}}) < \hat{R}_{\text{tot}}(f_{\text{KL}}).$$

Apply the uniform convergence bound from Step 1 to both $f_{\text{KL+PRR}}$ and f_{KL} . The difference in their population risks is upperbounded by the difference in empirical risks (which is negative by Step 2) plus the same complexity term. Hence

$$R_{\text{tot}}(f_{\text{KL+PRR}}) < R_{\text{tot}}(f_{\text{KL}}) + 2\mathfrak{R}_{n+m}(\ell \circ \mathcal{F}) + B\sqrt{\frac{\ln(1/\delta)}{2(n+m)}},$$

as claimed. \square

B The Proof of the Theorem 3.2

We study the generalization performance of Rankmatch. Before providing the main results, we first define the true risk with respect to the classification model $f(x; \theta)$:

$$R(f) = \mathbb{E}_{(x, y)}[L(f(\mathbf{x}), \mathbf{d})].$$

Our goal is to learn a good classification model by minimizing the empirical risk $\hat{R}(f) = \hat{R}_L(f) + \hat{R}_U(f)$, where $\hat{R}_L(f)$ and $\hat{R}_U(f)$ are respectively the empirical risk of the labeled loss $L_L(f(\mathbf{x}), \mathbf{d})$ and unlabeled loss $L_U(f(\mathbf{x}), \mathbf{d})$:

$$\hat{R}_L(f) = \frac{1}{n} \sum_{i=1}^n L(f(\mathbf{x}_i), \mathbf{d}_i), \quad \hat{R}_U(f) = \frac{1}{m} \sum_{j=1}^m L_U(f(\mathbf{x}_j), \mathbf{d}_j).$$

Note that during the training, we cannot train a model directly by optimizing $\hat{R}_U(f)$, since the labels of unlabeled data are inaccessible. Instead, we train the model with $\hat{R}'_U(f) = \frac{1}{m} \sum_{j=1}^m L_U(f(\mathbf{x}_j), \hat{\mathbf{d}}_j)$, where $\hat{\mathbf{d}}_j$ represents the pseudo-label vector of the instance \mathbf{x}_j .

Let $L_k(f(\mathbf{x})) = d_{\mathbf{x}}^{y_k} \ln \left(\frac{d_{\mathbf{x}}^{y_k}}{h(y_k | \text{Aug}_w(\mathbf{x}))} \right)$ be the loss for the label k , and L_E be any (not necessarily the best) Lipschitz constant of L . Let $R_N(\mathcal{F})$ be the expected Rademacher complexity of \mathcal{F} with $N = m + n$ training points. Let \hat{f} be the empirical risk minimizer, where \mathcal{F} is a function class, and f^* be the true minimizer. We derive the following theorem, which provides a generalization error bound for the proposed method.

Theorem B.1. *Suppose that $\ell(\cdot)$ is bounded by B . For some $\epsilon > 0$, if $\sum_{j=1}^m |\mathbb{I}(f_k(\mathbf{x}_j)) - \mathbb{I}(d_{\mathbf{x}_j}^{y_k})| / m \leq \epsilon$ for any $k \in [q]$, for any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$R(\hat{f}) - R(f^*) \leq 2qB\epsilon + 4qL_ER_N(\mathcal{F}) + 2qB\sqrt{\frac{\log \frac{2}{\delta}}{2N}}.$$

From Theorem 2, it can be observed that the generalization performance of \hat{f} mainly depends on two factors, i.e., the pseudo-labeling error ϵ and the number of training examples N . Apparently, a smaller pseudo-labeling error ϵ often leads to better generalization performance. Thanks to its robustness and the empirical evidence supporting the model, we anticipate strong performance in practical applications.

Theorem B.2. *Suppose that $\ell(\cdot)$ is bounded by B . For some $\epsilon > 0$, if $\sum_{j=1}^m |\mathbb{I}(f_k(\mathbf{x}_j)) - \mathbb{I}(d_{\mathbf{x}_j}^{y_k})| / m \leq \epsilon$ for any $k \in [q]$ for any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$R(\hat{f}) - R(f^*) \leq 2qB\epsilon + 4qL_ER_N(\mathcal{F}) + 2qB\sqrt{\frac{\log \frac{2}{\delta}}{2N}}.$$

Proof. Before proving the theorem, we first provide two useful lemmas as follows. We primarily derive the uniform deviation bound between $R(\hat{f})$ and $R(f)$.

Lemma B.3. *Suppose that the loss function ℓ is L_E -Lipschitz continuous with respect to θ . For any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$|R(\hat{f}) - \hat{R}(f)| \leq 2qL_ER_{n+m}(\mathcal{F}) + qB\sqrt{\frac{\log \frac{2}{\delta}}{2(n+m)}} \quad (12)$$

Proof. In order to prove this lemma, we define the Rademacher complexity of L and \mathcal{F} with $m + n$ training examples as follows:

$$R_{n+m}(L \circ \mathcal{F}) = \mathbb{E}_{\mathbf{x}, \mathbf{d}, \sigma} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma_i \ell(f(\mathbf{x}_i), \mathbf{d}_i) + \sum_{j=1}^m \sigma_j \ell(f(\mathbf{x}_j), \mathbf{d}_j) \right]$$

where σ_i and σ_j are Rademacher variables.

Considering that $C(f(\mathbf{x}), \mathbf{d}) = \sum_{i=1}^m \ell(f_k, \mathbf{d}_k)$, we have

$$R_{n+m}(L \circ \mathcal{F}) \leq qR_{n+m}(\ell \circ \mathcal{F}) \leq qL_ER_{n+m}(\mathcal{F})$$

where the second line is due to the Lipschitz continuity of the loss function ℓ .

Then, we proceed the proof by showing that one direction $\sup_{f \in \mathcal{F}} R(f) - R(\hat{f})$ is bounded with probability at least $1 - \delta/2$, and the other direction can be proved similarly. According to *McDiarmid's inequality* [6], for any $\delta > 0$, with probability at least $1 - \delta/2$, we have

$$\sup_{f \in \mathcal{F}} R(\hat{f}) - R(f) \leq \sup_{f \in \mathcal{F}} R(\hat{f}) - R(f) + qB \sqrt{\frac{\log \frac{2}{\delta}}{2(n+m)}}$$

According to the result in [41] (Theorem 3.3) that shows $\mathbb{E} \sup_{f \in \mathcal{F}} R(\hat{f}) - R(f) \leq 2R_m(\mathcal{F})$, by further considering the other direction $\sup_{f \in \mathcal{F}} R(f) - R(\hat{f})$, with probability at least $1 - \delta$, we have

$$\sup_{f \in \mathcal{F}} |R(\hat{f}) - R(f)| \leq 2qL_E R_m(\mathcal{F}) + qB \sqrt{\frac{\log \frac{2}{\delta}}{2n+m}}$$

which completes the proof. \square

Then, we can bound the difference between $R(\hat{f})$ and $R(f)$ as follows:

Lemma B.4. *Suppose that $\ell(\cdot)$ is bounded by B . For some $\epsilon > 0$, if $\sum_{j=1}^m |\mathbb{I}(f_k(\mathbf{x}_j)) - \mathbb{I}(d_{\mathbf{x}_j}^{y_k})|/m \leq \epsilon$ for any $k \in [q]$ for any $\delta > 0$, we have:*

$$|\hat{R}_U(f) - R_U(f)| \leq qB\epsilon$$

Proof. Without loss of generality, assume that ϵ is the largest pseudo-labeling error among q classes, i.e., $\epsilon = \max_{k=1}^q \sum_{j=1}^m |\mathbb{I}(f_k(\mathbf{x}_j)) - \mathbb{I}(d_{\mathbf{x}_j}^{y_k})|/m \leq \epsilon$ for any $k \in [q]$. Obviously, ϵ consists below pseudo-labeling error:

$$\epsilon = \frac{\sum_{j=1}^m \mathbb{I}(f_k(\mathbf{x}_j), d_{\mathbf{x}_j}^{y_k})}{m} \quad (13)$$

Then, we prove the following side, which provide the bounds for $R_U(f)$. Firstly, we prove its upper bound:

$$\begin{aligned} \hat{R}'_u(f) &= \frac{1}{m} \sum_{j=1}^m \sum_{k=1}^q \mathbb{I}(f_k(\mathbf{x}_j)) \ell(f_k(\mathbf{x}_j)) \\ &\leq \frac{1}{m} \sum_{j=1}^m \sum_{k=1}^q \mathbb{I}(d_{\mathbf{x}_j}^{y_k}) \ell(f_k(\mathbf{x}_j)) + \mathbb{I}(d_{\mathbf{x}_j}^{y_k}, f_k(\mathbf{x}_j)) \ell(f_k(\mathbf{x}_j)) \\ &\leq \frac{1}{m} \sum_{j=1}^m \mathcal{L}(f(\mathbf{x}_j), d_{\mathbf{x}_j}^{y_k}) + \epsilon \sum_{k=1}^q \ell(f_k(\mathbf{x}_j)) \\ &\leq \hat{R}_u(f) + qB\epsilon \end{aligned} \quad (14)$$

where the second line holds based on Eq.(13). Then, we prove its low bound:

$$\begin{aligned} \hat{R}'_u(f) &= \frac{1}{m} \sum_{j=1}^m \sum_{k=1}^q \mathbb{I}(f_k(\mathbf{x}_j)) \ell(f_k(\mathbf{x}_j)) \\ &\geq \frac{1}{m} \sum_{j=1}^m \sum_{k=1}^q \mathbb{I}(d_{\mathbf{x}_j}^{y_k}) \ell(f_k(\mathbf{x}_j)) - \mathbb{I}(d_{\mathbf{x}_j}^{y_k}, f_k(\mathbf{x}_j)) \ell(f_k(\mathbf{x}_j)) \\ &\geq \frac{1}{m} \sum_{j=1}^m \mathcal{L}(f(\mathbf{x}_j), d_{\mathbf{x}_j}^{y_k}) + \epsilon \sum_{k=1}^q \ell(f_k(\mathbf{x}_j)) \\ &\geq \hat{R}_u(f) + qB\epsilon \end{aligned} \quad (15)$$

By combining these two sides, we can obtain the following result:

$$|\hat{R}_U(f) - R_U(f)| \leq qB\epsilon$$

which concludes the proof.

For any $\delta > 0$, with probability at least $1 - \delta$, we have:

$$\begin{aligned} R(f) &\leq \hat{R}(f) + R_U(f) + 2qL_ER_{n+m}(\mathcal{F}) + qB\sqrt{\frac{\log \frac{2}{\delta}}{2N}} \\ &\leq \hat{R}(f) + R_U(f) + qB\epsilon + 2qL_ER_{n+m}(\mathcal{F}) + qB\sqrt{\frac{\log \frac{2}{\delta}}{2N}} \\ &\leq \hat{R}(f) + R_U(f) + 2qB\epsilon + 2qL_ER_{n+m}(\mathcal{F}) + qB\sqrt{\frac{\log \frac{2}{\delta}}{2N}} \\ &\leq \hat{R}(f) + R_U(f) + 2qB\epsilon + 4qL_ER_{n+m}(\mathcal{F}) + 2qB\sqrt{\frac{\log \frac{2}{\delta}}{2N}} \\ &\leq R(f) + 2qB\epsilon + 4qL_ER_{n+m}(\mathcal{F}) + 2qB\sqrt{\frac{\log \frac{2}{\delta}}{2N}} \end{aligned}$$

where the first and fifth lines are based on Eq. 6, and second and fourth lines are due to Lemma B.3. The third line is by the definition of f . Putting all these together, the proof is then finished. \square

C Others

Experimental Datasets :In this paper, we validate our approach using four distinct real-world datasets³. The details of these datasets are as follows:

Twitter-LDL : A large-scale Visual Sentiment Distribution dataset was constructed from Twitter, encompassing eight distinct emotions Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear, Sadness. Approximately 30,000 images were collected by searching various emotional keywords, such as "sadness," "heartbreak," and "grief." Subsequently, eight annotators were hired to label this dataset. The resulting Twitter LDL dataset comprises 10,045 images.

Flickr-LDL : A subset of the Flickr dataset, unlike other datasets that searched for images using emotional terms, the Flickr dataset collected 1,200 pairs of adjective-noun pairs, resulting in 500,000 images. We employed 11 annotators to label this subset with tags for eight common emotions. In the end, the Flickr LDL was created, containing 10,700 images, with roughly equal quantities for each class.

Emotion6 : Emotion6: We collected 1,980 images from Flickr using six category keywords and synonyms as search terms for Emotion6. A total of 330 images were collected for each category, and each image was assigned to only one category (dominant emotion). Emotion6 represents the emotions related to each image in the form of a probability distribution, consisting of 7 bins, including Ekman's 6 basic emotions and neutral.

RAF-LDL : RAF-LDL is a multi-label distribution facial expression dataset, comprising approximately 5,000 diverse facial images downloaded from the internet. These images exhibit variations in emotion, subject identity, head pose, lighting conditions, and occlusions. During annotation, 315 well-trained annotators are employed to ensure each image can be annotated enough independent times. And images with multi-peak label distribution are selected out to constitute the RAF-LDL.

Comparing methods In order to assess the effectiveness of the proposed approach, we benchmark it against four sets of methods:

1) For the semi-supervised multi-label learning (SSMLL) algorithms, we follow the hyperparameter configurations provided in their original papers. Specifically, for SSMLL-CAP (CAP) [59] and

³The dataset's author has made the dataset publicly available at the following link: <http://cv.nankai.edu.cn/projects/SentiLDL>.

PCLP [36], we adopt the same backbone networks, learning rates, batch sizes, and training schedules as specified in their respective works. Additionally, to adapt these methods to our semi-supervised label distribution learning (SSLDL) setting, we modify their final activation layer: the original sigmoid function is replaced with a softmax activation. This transformation enables the models to generate pseudo-label distributions rather than independent multi-label probabilities, ensuring better

2) The second group consists of two deep learning SSLDL algorithms that we introduced, named FixMatch-LDL and MixMatch-LDL. Since there are currently no open-source semi-supervised LDL works in deep learning, these two algorithms were developed by us, based on the current most effective two deep learning SSL algorithms.

(i) FixMatch-LDL :Fixmatch-LDL is an adaptation we made based on the classic semi-supervised algorithm fixmatch [52]. Specifically, we pre-trained on images using ResNet50, then trained the model with labeled data. Subsequently, we assigned pseudo-label distributions to the unlabeled data, and finally, we aligned the model’s strongly augmented output with the pseudo-label distribution. For all datasets, the number of epochs is set as 30 and the batch size is set as 32. We perform all experiments on GeForce RTX 3090 GPUs. The random seed is set to 1 for all experiments.

(ii) MixMatch-LDL: Mixmatch is a semi-supervised LDL algorithm designed by us. Specifically, we first use linear interpolation to blend images, creating new samples. Similarly, we generate the label distributions for these new samples. Following this, we train the data using the same training strategy as Mixmatch. It’s worth mentioning that producing new samples enhances the model’s ability to prevent overfitting. For all datasets, the number of epochs is set as 30 and the batch size is set as 32. We perform all experiments on GeForce RTX 3090 GPUs. The random seed is set to 1 for all experiments.

3) The third group of algorithms is a deep learning SSLDL algorithm based on the dual-network concept, which we named GCT-LDL. The core idea involves mutual supervision of the outputs from two independent networks using unlabeled data. GCT-LDL : Two models utilized two different pretrained initializations of ResNet50 provided by PyTorch (ResNet50-Weights.IMAGENET1K-V1 and ResNet50-Weights.IMAGENET1K-V2). During training, labeled and unlabeled data were mixed. The loss used is the cross-entropy loss, divided into two parts: for labeled data, the loss is calculated directly between the prediction results and the ground truth. For unlabeled data, the loss is calculated between the prediction results of each model and the results of the other model. Hyperparameter settings are the same as those used in other methods.

4) The fourth group consists of traditional SSLDL algorithms, referred to as SA-LDL [18]. Since SA-LDL is an SSLDL algorithm designed for tabular data, we needed to perform feature engineering on image data, first, we use ResNet-50 for feature extraction from all datasets, followed by dimensionality reduction to 128 dimensions using PCA. For the remaining settings, we adhere to the defaults as specified in the paper.

5) The finally category consists of existing LDL algorithms. As there is currently only one open-source SSLDL algorithm, which is SA-LDL [18], we compared it with some state-of-the-art LDL algorithms. In this regard, we selected four state-of-the-art LDL algorithms: Adam-LDL-SCL [20], sLDLF [50], DF-LDL [13], and LDL-LRR [21]. These algorithm settings are defaulted to be consistent with those specified in the paper. Additionally, for these algorithms, we directly use labeled data to train the classifier. Then, we use the trained model to assign pseudo-labels to the unlabeled samples. Finally, we use the pseudo-labels to update the model.

Evaluation Metrics: We evaluate LDL algorithms using six metrics: five distance-based (Chebyshev, Clark, Kullback-Leibler, and Canberra) and two similarity-based (Cosine and Intersection). Lower values indicate better performance for distance-based metrics (\downarrow), while higher values indicate better performance for similarity-based metrics (\uparrow).

Table 5: The distribution distance/similarity measures and ranking correlation metrics

Measure / Metric	Formula
Chebyshev ↓	$\text{Dis}_1(\mathbf{d}, \hat{\mathbf{d}}) = \max_j d_j - \hat{d}_j $
Clark ↓	$\text{Dis}_2(\mathbf{d}, \hat{\mathbf{d}}) = \sqrt{\sum_{j=1}^c \frac{(d_j - \hat{d}_j)^2}{(d_j + \hat{d}_j)^2}}$
Canberra ↓	$\text{Dis}_3(\mathbf{d}, \hat{\mathbf{d}}) = \sum_{j=1}^c \frac{ d_j - \hat{d}_j }{d_j + \hat{d}_j}$
Kullback-Leibler ↓	$\text{Dis}_4(\mathbf{d}, \hat{\mathbf{d}}) = \sum_{j=1}^c d_j \ln \frac{d_j}{\hat{d}_j}$
Cosine ↑	$\text{Sim}_1(\mathbf{d}, \hat{\mathbf{d}}) = \frac{\sum_{j=1}^c d_j \hat{d}_j}{\sqrt{\sum_{j=1}^c d_j^2} \sqrt{\sum_{j=1}^c \hat{d}_j^2}}$
Intersection ↑	$\text{Sim}_2 = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^c \min(d_{x_i}^{y_j}, \hat{d}_{x_i}^{y_j})$
Spearman's rank ρ_S ↑	$\rho_S = 1 - \frac{6 \sum_i (\bar{D} - D)^2}{k(k^2 - 1)}$
Kendall tau correlation τ_K ↑	$\tau_K = \frac{n_c - n_d}{\frac{1}{2}k(k-1)}$

C.1 The Rest Experimental Results

Convergence Analysis Fig. 6 illustrates the convergence curves of the RankMatch algorithm on the Flickr-LDL and Twitter-LDL datasets. The rapid decline in the initial loss for Flickr-LDL indicates quick adaptation and efficient optimization during early epochs, stabilizing as the model converges. On the other hand, Twitter-LDL demonstrates a more gradual decline, reflecting a steadier learning process. These results confirm the robust optimization capability of RankMatch across diverse datasets.

C.2 Parameter Sensitivity Analysis

To investigate the robustness of our method with respect to the trade-off parameter λ in the PRR regularization term, we conduct a sensitivity analysis on four datasets: Emotion6, Flickr-LDL, RAF-LDL, and Twitter-LDL. As shown in Fig. 7, the performance remains stable across a wide range of λ values from 0.005 to 0.1, demonstrating that the proposed framework is not overly sensitive to this hyperparameter. A small λ (e.g., 0.01) generally achieves the best or near-best results across most metrics, indicating that a moderate contribution from the PRR loss is sufficient to capture label-ranking consistency without dominating the primary KL-divergence objective. These results validate the robustness and general applicability of the proposed PRR-regularized SSLDL framework across diverse datasets.

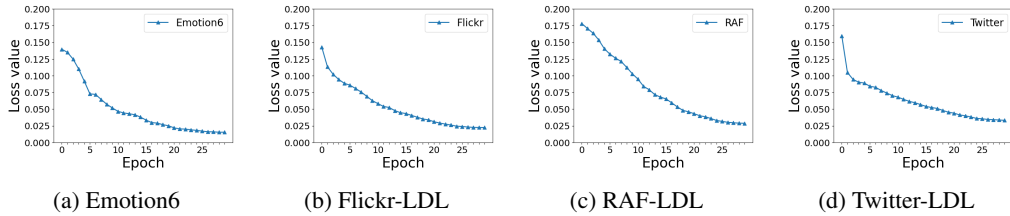


Figure 6: The convergence curve on Emotion6, Flickr-LDL, RAF-ML, and Twitter-LDL.

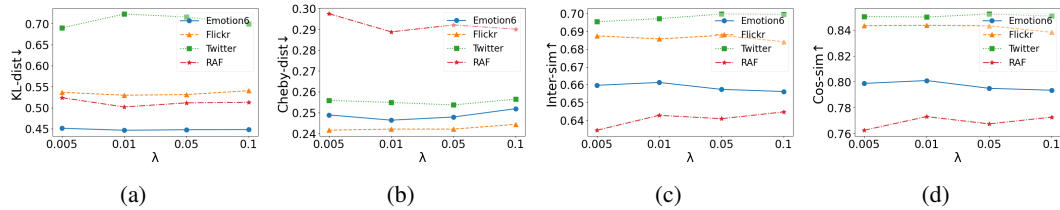


Figure 7: Parameter Sensitivity Analysis on 4 datasets.