Decoupled Imbalanced Label Distribution Learning

Yongbiao Gao[1,2,3,4], **Xiangcheng Sun**[1,2], **Miaogen Ling**[5], **Chao Tan**[6], **Yi Zhai**[1,2], **Guohua Lv**[1,2*]
[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[2]Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan, China
[3]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Application (Southeast University), Ministry of Education, China
[4]Shandong Key Laboratory of Ubiquitous Intelligent Computing, Jinan, China
[5]School of Computer and Software, Nanjing University of Information Science and Technology
[6]School of Computer and Electronic Information/School of Artificial Intelligence, Nanjing Normal University
{gaoyb, guohualv, yzhai, 10431240010}@qlu.edu.cn, 73022@njnu.edu.cn, mgling@nuist.edu.cn

Abstract

Label Distribution Learning (LDL) has been successfully implemented in numerous practical applications. However, the imbalance in label distributions presents a significant challenge due to the substantial variation in annotation information. To tackle this issue, we introduce *Decoupled Imbalance Label Distribution Learning* (DILDL), which decomposes the imbalanced label distribution into a dominant label distribution and a non-dominant label distribution. Our empirical findings reveal that an excessively high description degree of dominant labels can result in substantial gradient information attenuation for non-dominant labels during the learning process. Therefore, we employ the decoupling approach to balance the description degrees of both dominant and non-dominant labels independently. Furthermore, we align the feature representations with the representations of dominant and non-dominant labels separately, aiming to effectively mitigate the distribution shift problem. Experimental results demonstrate that our proposed DILDL outperforms other state-of-the-art methods for imbalance label distribution learning.

1 Introduction

Label Distribution Learning (LDL) [Geng, 2016] aims to establish a mapping from an instance to a label distribution, which encompasses a specific set of labels and indicates the degree to which each label describes the instance. LDL has been shown to be an effective approach to address the issue of label ambiguity [Wang and Geng, 2019; Kou *et al.*, 2024b; Wang *et al.*, 2024]. The underlying mechanism of label distribution mapping has also attracted significant research attention [Shen *et al.*, 2017; Lu and Jia, 2024; Li *et al.*, 2024]. LDL has been successfully applied to numerous practical scenarios and learning paradigms, including expression analysis [Le *et al.*, 2023], video parsing [Gao *et al.*, 2021; Zhang *et al.*, 2023], age estimation [Smith-Miles and Geng, 2020; Zhang *et al.*, 2021], image captioning [Yang *et al.*, 2023], beauty sense [Xie *et al.*, 2015; Ren and Geng, 2017], few-shot learning [An *et al.*, 2024], multi-label learning [Wang and Geng, 2024; Kou *et al.*, 2024b], and partial label learning [Xu *et al.*, 2023], big model [Peng *et al.*, 2025], etc.

LDL directly addresses the deeper and more ambiguous question: "How much does each label describe the instance?"

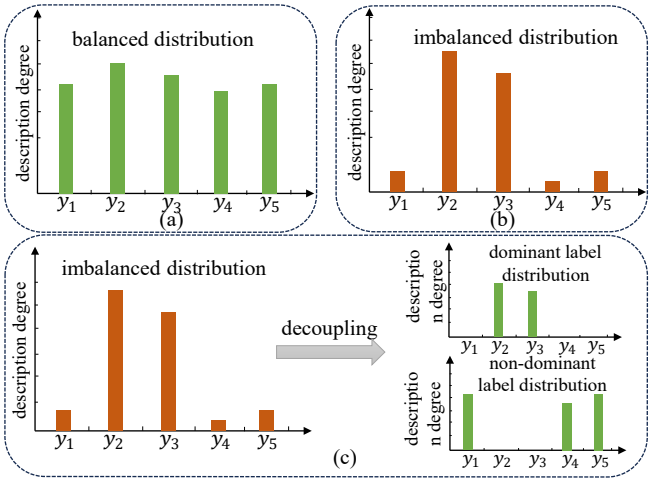

Figure 1: (a) indicates a balanced distribution, (b) indicates an imbalanced distribution, and (c) provides an example of decoupling the imbalanced label distribution. After decoupling, the description degrees of non-dominant labels can be enhanced in the LDL process.

*Corresponding Author

In other words, it considers the relative importance of each label in describing the instance. The effectiveness of LDL hinges on the relatively balanced nature of the description degrees, with generally small differences among them, enabling the use of more balanced supervision information to train the LDL model. However, achieving highly balanced description degrees is quite challenging in practical applications due to the significant subjectivity involved in annotating the label distribution. For example, when training a score distribution model for movies [Geng and Hou, 2015], obtaining an ideally balanced emotional distribution like Fig. 1(a) requires hundreds or even thousands of annotators to label the same movie. Yet, due to limitations such as insufficient numbers of annotators, varying backgrounds, ages, subjective opinions, annotation noise, etc., it is common for the score distribution to become imbalanced, as shown in Fig. 1(b). The dominant scores occupy an excessively large descriptive space, leaving very little space for the non-dominant labels. The excessive variance among these description degrees can severely impact the performance of label distribution learning in solving practical problems [He and Garcia, 2009; Wu *et al.*, 2020; Oh *et al.*, 2022; Fu *et al.*, 2024]. This newly emerging and challenging scenario is defined as *Imbalanced Label Distribution Learning (ILDL)* [Zhao *et al.*, 2023b]. RDA [Zhao *et al.*, 2023b] uncovers the underlying reason behind the performance degradation of imbalanced distribution learning from an alignment perspective. RDA asserts that existing LDL methods incorrectly assume that the consistency between the feature distributions of the training set and the test set is invalid. Consequently, RDA introduces a two-phase alignment approach.

Both LDL-HR [Wang and Geng, 2021b] and DKD [Zhao *et al.*, 2022] methods demonstrate that non-dominant labels can either enhance model generalization or transfer implicit knowledge. However, RDA solely focuses on enhancing the representation capabilities of the feature space and label space from a representation alignment perspective, without addressing the issue of excessive attenuation of non-dominant labels in LDL model learning due to the over-representation of dominant labels.

To solve the above issue, we propose a novel method named *Decoupled Imbalanced Label Distribution Learning* (DILDL). This method decouples the imbalanced label distribution into dominant and non-dominant label distributions. Specifically, we divide the label distribution learning process into two levels: (1) prediction of description degrees for dominant labels and (2) prediction of description degrees for non-dominant labels. Based on this division, we reformulate the LDL loss into two components, as illustrated in Figure 1(c). Despite the inherent imbalance in the initial label distribution, our decoupling methodology successfully establishes an equilibrium in the description degrees between dominant and non-dominant label distributions. After decoupling, the description degrees of non-dominant labels become independent of those of dominant labels, eliminating the need for their concurrent learning. From the perspective of gradient analysis, we demonstrate that decoupling can enhance the learning of gradient information related to non-dominant labels.

The overall framework of the proposed decoupled imbalanced label distribution learning (DILDL) is shown in Figure 2. We utilize the DILDL loss to learn the mapping from instances to label distributions in the first branch of the decoder $\mathcal{F}_\theta$. More importantly, DILDL also decouples the label distribution during the representation distribution alignment stage.

Our contributions can be summarized as follows:

- We propose a decoupled method to address the issue of excessive emphasis on dominant labels, which leads to excessive attenuation of non-dominant labels in imbalanced label distribution learning.
- We prove that the decoupled approach in ILDL can further enhance the implicit knowledge of non-dominant labels from a gradient analysis perspective.
- We conduct extensive experiments to demonstrate the effectiveness of the proposed Decoupled Imbalanced Label Distribution Learning (DILDL), and our method achieves state-of-the-art performance.

2 Related Work

Label Distribution Learning. Label distribution learning (LDL) was first proposed by [Geng, 2016], and it has been successfully applied to ambiguous tasks. For example, [Geng and Hou, 2015] formulated movie scores from multiple annotators as a score distribution and simultaneously fit a sigmoid function to each component of the score distribution using a multi-output support vector machine. LDL-ALSG [Chen *et al.*, 2020] was proposed to address the facial expression recognition using the topological information of labels from related but more distinct tasks. The underlying assumption of LDL-ALSG is that facial images should have similar expression distributions to their neighbors in the label space of action unit recognition and facial landmark detection. More recently, [Le *et al.*, 2023] leveraged neighborhood information in the valence-arousal space to adaptively construct emotion distributions for training samples, taking into account the uncertainty of provided labels when incorporating them into the label distributions to solve the facial expression recognition problem. Additionally, LDL has been applied to age estimation [Zhang *et al.*, 2021], image sentiment analysis [Yang *et al.*, 2017a; Yang *et al.*, 2017b], text classification [Zhao *et al.*, 2023a], and other areas. Beyond its contributions across various application areas, LDL can further enhance the performance of other machine learning paradigms. For example, [Kou *et al.*, 2024b] introduced an auxiliary multi-label learning (MLL) process in LDL to capture low-rank label correlation on that MLL rather than LDL, justifying the advantages of exploiting low-rank label correlation in the auxiliary MLL through LDL. LDL-FCC [Wang *et al.*, 2024] was designed to explore fuzzy membership-induced correlation and to jointly realize fuzzy clustering and label correlation learning via LDL. However, as analyzed in [Zhao *et al.*, 2023b], previous LDL methods ignored the imbalance problem in label distribution and did not consider the distribution gap between the training set and test set.

Imbalance Label Distribution Learning. Imbalanced learning is a pressing topic that arises from the long-tail distribution of data. Under-sampling and over-sampling are two

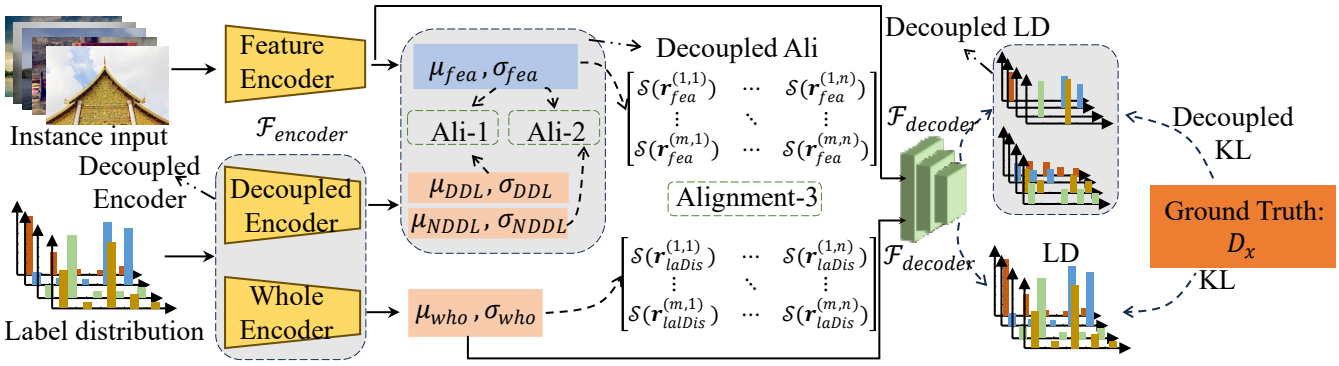

Figure 2: To achieve decoupled representation distribution alignment, we design the decoupled representation to conduct distinct representation learning for both dominant and non-dominant labels. Following this, we undertake the alignment of the Gaussian distribution of the feature representations with the Gaussian distributions corresponding to the dominant and non-dominant labels, respectively.

of the most popular approaches to tackle the long-tail recognition problem [Byrd and Lipton, 2019; Buda *et al.*, 2018; Ghosh *et al.*, 2024]. [Soltanzadeh *et al.*, 2023] addresses the imbalance issue by presenting an under-sampling approach based on a metaheuristic method, where the under-sampling problem is formulated as an optimization problem. The proposed method aims to select an optimal subset of majority samples to handle both the imbalance and class-overlap problems simultaneously, while avoiding excessive elimination of majority samples. Apart from imbalanced classification, imbalanced regression has also garnered significant attention in recent years [Yang *et al.*, 2021; Wang and Wang, 2024; Liu *et al.*, 2023]. VIR [Wang and Wang, 2024] borrows data with similar regression labels to compute the variational distribution of latent representations, predicts the entire normal-inverse-gamma distributions, and modulates the associated conjugate distributions to probabilistically re-weight the imbalanced data. However, both imbalanced classification and imbalanced regression focus on identifying imbalances at the label end. The uniqueness of imbalanced label distribution learning (ILDL) lies in the fact that each label is accompanied by a continuous description degree [Geng, 2016; Kou *et al.*, 2024a; Wang and Geng, 2019]. Consequently, traditional imbalanced classification and imbalanced regression methods cannot be directly applied to ILDL. To address this, RDA [Zhao *et al.*, 2023b] introduced the first specialized ILDL algorithm, which aligns the distributions of feature representations and label representations to bridge the gap between the training set and test set caused by imbalance. However, RDA does not effectively address the issue of excessive attenuation of non-dominant labels in ILDL model learning.

3 Approach

Inspired by LDL-HR [Wang and Geng, 2021b], AEKT [Park and Lee, 2024], and DKD [Zhao *et al.*, 2022], we reformulate the ILDL loss as a weighted sum of two components: one representing the distribution of dominant labels and the other representing the distribution of non-dominant labels. Additionally, we integrate our decoupled method into the RDA [Zhao *et al.*, 2023b] framework. Decoupling is implemented for dominant and non-dominant labels during both the label distribution prediction and representation alignment stages.

3.1 Decoupled the Label Distribution

Assume $f_\theta(\cdot)$ is the mapping function from the instance space $\mathcal{X}$ to the label distribution space $\mathcal{Y}$. The objective function of LDL is to minimize the difference of the ground truth and predicted label distribution. The Kullback-Leibler (KL) divergence is the most common used loss function. Therefore, the objective function can be written as,

$$\mathcal{L}_{LDL} = \sum_{i=1}^{n}\sum_{j=1}^{c} d_{\boldsymbol{x}_i}^{y_j} \ln \frac{d_{\boldsymbol{x}_i}^{y_j}}{f_\theta^j(\boldsymbol{x}_i)}, \tag{1}$$

where $d_{\boldsymbol{x}_i}^{y_j}$ is the groundtruth of the description degree. To start, we decouple the distribution of dominant label. Specifically, for each instance $\boldsymbol{x}$, the distribution of dominant label is defined as $\bar{D}_{\boldsymbol{x}} = [\bar{d}_{\boldsymbol{x}}^{y_1}, \bar{d}_{\boldsymbol{x}}^{y_2}, ..., \bar{d}_{\boldsymbol{x}}^{y_c}]^T$, where $\bar{d}_{\boldsymbol{x}}^{y_j}$ is defined by

$$\bar{d}_{\boldsymbol{x}}^{y_j} = \begin{cases} 1 & if \quad y_j = y_{\boldsymbol{x}}, \\ 0 & otherwise, \end{cases} \tag{2}$$

where $y_{\boldsymbol{x}}$ is the label with the highest description degree, the decoupled distribution assigns a description degree of 1 to the dominant label and 0 to all other labels. According to the definition of LDL, the label distribution satisfies two constraints, *i.e.* $d_{\boldsymbol{x}_i}^{y_j} \in [0,1]$ and $\sum_{j=1}^{c} d_{\boldsymbol{x}_i}^{y_j} = 1$, the distribution of non-dominant labels is defined as $\hat{D}_{\boldsymbol{x}} = [\hat{d}_{\boldsymbol{x}}^{y_1}, \hat{d}_{\boldsymbol{x}}^{y_2}, ..., \hat{d}_{\boldsymbol{x}}^{y_c}]^T$, where $\hat{d}_{\boldsymbol{x}}^{y_j}$ is defined by

$$\hat{d}_{\boldsymbol{x}}^{y_j} = \begin{cases} 0 & if \quad y_j = y_{\boldsymbol{x}}, \\ exp(d_{\boldsymbol{x}_i}^{y_j}) / \sum\limits_{j=1, y_j \neq y_{\boldsymbol{x}}}^{c} exp(d_{\boldsymbol{x}_i}^{y_j}) & otherwise, \end{cases} \tag{3}$$

Therefore, Eq. (1) can be re-written as,

$$\mathcal{L}_{DILDL} = \alpha KL([\bar{d}_{\boldsymbol{x}_i}^{y_j}, \hat{d}_{\boldsymbol{x}_i}^{y_j}] || [\bar{d'}_{\boldsymbol{x}_i}^{y_j}, \hat{d'}_{\boldsymbol{x}_i}^{y_j}]) + (1-\alpha) KL(\hat{D} || \hat{f}_\theta), \tag{4}$$

where α is the trade-off parameter to balance the decoupled label distributions. $\hat{f}(\cdot)$ is the decoupled prediction of non-dominant labels. d' is the predicted description degree.

The first term of Eq.(4) for the distribution of the dominant label can be written as,

$$\mathcal{L}_{DDL} = \bar{d}_{\boldsymbol{x}_i}^{y_j}(\log \bar{d}_{\boldsymbol{x}_i}^{y_j} - \log \bar{d'}_{\boldsymbol{x}_i}^{y_j}) + \hat{d}_{\boldsymbol{x}_i}^{y_j}(\log \hat{d}_{\boldsymbol{x}_i}^{y_j} - \log \hat{d'}_{\boldsymbol{x}_i}^{y_j}), \tag{5}$$

Follow the gradient calculation method in AEKT [Park and Lee, 2024], the gradient of $\mathcal{L}_{DDL}$ with the logit z_i^k of the dominant label $y_{\boldsymbol{x}}$ can be calculated as follows,

$$\begin{aligned}\frac{\partial \mathcal{L}_{DDL}}{\partial z_i^k} &= \frac{\partial \mathcal{L}_{DDL}}{\partial \bar{d'}_{\boldsymbol{x}_i}^{y_k}} \frac{\partial \bar{d'}_{\boldsymbol{x}_i}^{y_k}}{z_i^k} + \frac{\partial \mathcal{L}_{DDL}}{\partial \hat{d'}_{\boldsymbol{x}_i}^{y_k}} \frac{\partial \hat{d'}_{\boldsymbol{x}_i}^{y_k}}{z_i^k} \\ &= \left(-\frac{\bar{d}_{\boldsymbol{x}_i}^{y_k}}{\bar{d'}_{\boldsymbol{x}_i}^{y_k}}\right)(\bar{d'}_{\boldsymbol{x}_i}^{y_k} - (\bar{d'}_{\boldsymbol{x}_i}^{y_k})^2) + \left(-\frac{\hat{d}_{\boldsymbol{x}_i}^{y_k}}{\hat{d'}_{\boldsymbol{x}_i}^{y_k}}\right)(-\bar{d'}_{\boldsymbol{x}_i}^{y_k} \cdot \hat{d'}_{\boldsymbol{x}_i}^{y_k}) \\ &= \bar{d'}_{\boldsymbol{x}_i}^{y_k} - \bar{d}_{\boldsymbol{x}_i}^{y_k},\end{aligned} \tag{6}$$

The gradient of $\mathcal{L}_{DDL}$ with the logit z_i^j of the non-dominant label y_j can be calculated as follows,

$$\begin{aligned}\frac{\partial \mathcal{L}_{DDL}}{\partial z_i^j} &= \frac{\partial \mathcal{L}_{DDL}}{\partial \bar{d'}_{\boldsymbol{x}_i}^{y_k}} \frac{\partial \bar{d'}_{\boldsymbol{x}_i}^{y_k}}{z_i^j} + \frac{\partial \mathcal{L}_{DDL}}{\partial \hat{d'}_{\boldsymbol{x}_i}^{y_k}} \frac{\partial \hat{d'}_{\boldsymbol{x}_i}^{y_k}}{z_i^j} \\ &= \left(-\frac{\bar{d}_{\boldsymbol{x}_i}^{y_k}}{\bar{d'}_{\boldsymbol{x}_i}^{y_k}}\right)(-\bar{d'}_{\boldsymbol{x}_i}^{y_k} \cdot \bar{d'}_{\boldsymbol{x}_i}^{y_j}) + \left(-\frac{\hat{d}_{\boldsymbol{x}_i}^{y_k}}{\hat{d'}_{\boldsymbol{x}_i}^{y_k}}\right)(\bar{d'}_{\boldsymbol{x}_i}^{y_j} - \hat{d'}_{\boldsymbol{x}_i}^{y_k}\bar{d'}_{\boldsymbol{x}_i}^{y_j}) \\ &= \left(1 - \frac{\hat{d}_{\boldsymbol{x}_i}^{y_k}}{\hat{d'}_{\boldsymbol{x}_i}^{y_k}}\right)\bar{d'}_{\boldsymbol{x}_i}^{y_j},\end{aligned} \tag{7}$$

The second term of Eq. (3) can be re-formulated as,

$$\mathcal{L}_{NDDL} = \sum_{j=1, y_j \neq y_{\boldsymbol{x}}}^{c} \hat{d}_{\boldsymbol{x}_i}^{y_j}(\log \hat{d}_{\boldsymbol{x}_i}^{y_j} - \log \hat{d'}_{\boldsymbol{x}_i}^{y_j}), \tag{8}$$

The gradient of $\mathcal{L}_{NDDL}$ with respect to the logits z_i^j of non-dominant labels y_j is calculated by,

$$\begin{aligned}\frac{\partial \mathcal{L}_{NDDL}}{\partial z_i^j} &= \sum_{m=1 y_m \neq y_{\boldsymbol{x}}}^{c} \frac{\partial \mathcal{L}_{NDDL}}{\partial \hat{d'}_{\boldsymbol{x}_i}^{y_m}} \frac{\partial \hat{d'}_{\boldsymbol{x}_i}^{y_m}}{\partial z_i^m} \\ &= \frac{\partial \mathcal{L}_{NDDL}}{\partial \hat{d'}_{\boldsymbol{x}_i}^{y_j}} \frac{\partial \hat{d'}_{\boldsymbol{x}_i}^{y_j}}{\partial z_i^j} + \sum_{m=1, m \neq j}^{c} \frac{\partial \mathcal{L}_{NDDL}}{\partial \hat{d'}_{\boldsymbol{x}_i}^{y_m}} \frac{\hat{d'}_{\boldsymbol{x}_i}^{y_m}}{\partial z_i^j} \\ &= \left(-\frac{\hat{d}_{\boldsymbol{x}_i}^{y_j}}{\hat{d'}_{\boldsymbol{x}_i}^{y_j}}\right)(\hat{d'}_{\boldsymbol{x}_i}^{y_j} - (\hat{d'}_{\boldsymbol{x}_i}^{y_j})^2) \\ &+ \sum_{m=1, m \neq j}^{c} \left(-\frac{\hat{d}_{\boldsymbol{x}_i}^{y_m}}{\hat{d'}_{\boldsymbol{x}_i}^{y_m}}\right)(-\hat{d'}_{\boldsymbol{x}_i}^{y_m} \cdot \hat{d'}_{\boldsymbol{x}_i}^{y_j}) \\ &= \hat{d'}_{\boldsymbol{x}_i}^{y_j} - \hat{d}_{\boldsymbol{x}_i}^{y_j} \\ &= \frac{1}{\hat{d'}_{\boldsymbol{x}_i}^{y_k}} \hat{d'}_{\boldsymbol{x}_i}^{y_j} - \frac{1}{\hat{d}_{\boldsymbol{x}_i}^{y_k}} \hat{d}_{\boldsymbol{x}_i}^{y_j},\end{aligned} \tag{9}$$

In summary, multiplying the decoupled gradients by the balanced parameter α and β can yield the gradients for dominant labels and non-dominant labels,

$$\frac{\partial \mathcal{L}_{LDL}}{\partial z_i^k} = \alpha(\bar{d'}_{\boldsymbol{x}_i}^{y_k} - \bar{d}_{\boldsymbol{x}_i}^{y_k}), \tag{10}$$

$$\frac{\partial \mathcal{L}_{LDL}}{\partial z_i^j} = \left\{\alpha\left(1 - \frac{\hat{d}_{\boldsymbol{x}_i}^{y_k}}{\hat{d'}_{\boldsymbol{x}_i}^{y_k}}\right) + \frac{\beta}{\hat{d'}_{\boldsymbol{x}_i}^{y_k}}\right\}\bar{d'}_{\boldsymbol{x}_i}^{y_j} - \frac{\beta}{\hat{d}_{\boldsymbol{x}_i}^{y_k}}\bar{d}_{\boldsymbol{x}_i}^{y_j}. \tag{11}$$

Based on the gradient analysis of the distributions of dominant and non-dominant labels, we can draw the follow the two conclusions:

- When the label distribution is very imbalance, *i.e.* the margin between $\bar{d}_{\boldsymbol{x}_i}^{y_k}$ and $\hat{d}_{\boldsymbol{x}_i}^{y_k}$ is very large, as shown in Eq. (11), the imbalance in label distribution can cause the LDL model to overly focus on dominant labels, leading to a small description degree $\bar{d}_{\boldsymbol{x}_i}^{y_j}$ for non-dominant labels (the second term in Eq. (11)), the gradient of the non-dominant labels $\partial \mathcal{L}_{LDL}/\partial z_i^j$ can be still delivered during the learning process since the β gives an independent balance for the distribution of non-dominant labels.
- As shown in the first term in Eq. (11), when the predicted label distribution $\hat{d'}_{\boldsymbol{x}_i}^{y_k}$ is large from the ground truth $\hat{d}_{\boldsymbol{x}_i}^{y_k}$, the weight for the predicted label distribution of the non-dominant labels $\hat{d'}_{\boldsymbol{x}_i}^{y_j}$ will be reduced, which will further address the issue of excessive losses resulting from inaccurate label distribution prediction during the learning process.

3.2 Decoupled Representation Distribution Alignment

RDA [Zhao *et al.*, 2023b] asserts that aligning the distributions of feature representations and label representations can narrow the distribution gap between the training set and test set, which is often exacerbated by the imbalance issue. Inspired by RDA [Zhao *et al.*, 2023b], we introduce the decoupled representation alignment approach to carry out distinct representation learning for both dominant and non-dominant labels.

As depicted in Figure 2, we design feature and label encoders to learn the representations of instances and label distributions, respectively. Specifically, the label distribution encoder consists of three branches, with two branches specifically dedicated to learning the representations of dominant and non-dominant labels.

The Central Limit Theorem as well as the information-theoretic state that the sum/mean of many independent random variables approximates a Gaussian distribution under suitable conditions and the Gaussian distribution has maximum entropy, making it the most "uncertain" (least assuming) distribution for data. Consequently, without specific distribution info, Gaussian is the default for label-feature alignment. Assuming that the distributions of features and labels adhere to Gaussian distributions, we utilize the KL divergence to align the decoupled label information with the feature representation of the model.

$$\mathcal{L}_{alig_1} = -\frac{1}{2}\sum_{i=1}^{I}\left[\log v_{DDL}^{(i)} - v_{DDL}^{(i)} - \tau_{DDL}^{(i)} + 1\right], \tag{12}$$

where $v_{DDL}^{(i)} = \frac{\sigma_{feature}^{(i)2}}{\sigma_{DDL}^{(i)2}}$, $\tau_{DDL}^{(i)} = \frac{\left(\mu_{feature}^{(i)} - \mu_{DDL}^{(i)}\right)^2}{\sigma_{DDL}^{(k)2}}$, i rep-

resents the i-th element of the latent space. Similarity,

$$\mathcal{L}_{alig_2} = -\frac{1}{2}\sum_{i=1}^{I}\left[\log v_{NDDL}^{(i)} - v_{NDDL}^{(i)} - \tau_{NDDL}^{(i)} + 1\right], \tag{13}$$

where $v_{NDDL}^{(i)} = \frac{\sigma_{feature}^{(i)^2}}{\sigma_{NDDL}^{(i)^2}}$, $\tau_{NDDL}^{(i)} = \frac{\left(\mu_{feature}^{(i)} - \mu_{NDDL}^{(i)}\right)^2}{\sigma_{NDDL}^{(k)^2}}$. To further align the distributions of feature representations and label representations, following RDA [Zhao *et al.*, 2023b], we use the reparameterization trick [Rezende *et al.*, 2014] to calculate the similarity of features and labels. For the labels: $r_{labelDis} = \mu_{whole} + \sigma_{whole}\delta_{whole}$, where μ_{whole} and σ_{whole} are calculated from the whole label encoder. $\delta_{whole} \sim \mathcal{N}(0, \boldsymbol{I})$. For the features: $r_{feature} = \mu_{feature} + \sigma_{feature}\delta_{feature}$, where $\mu_{feature}$ and $\sigma_{feature}$ are calculated from the feature encoder. $\delta_{feature} \sim \mathcal{N}(0, \boldsymbol{I})$. The alignment-3 is defined as,

$$\mathcal{L}_{alig_3} = \frac{1}{M^2}\sum_{m=1}^{M}\sum_{n=1}^{N}\left(A_{mn} - Z_{mn}\right)^2, \tag{14}$$

where A_{mn} and Z_{mn} are cosine similarity matrix of feature representations and label representations,

$$\begin{aligned} A_{mn} &= \mathcal{S}\left(r_{feature}^{(m)}, r_{feature}^{(n)}\right), \\ Z_{mn} &= \mathcal{S}\left(r_{labelDis}^{(m)}, r_{labelDis}^{(n)}\right). \end{aligned} \tag{15}$$

where m and n are m-th and n-th instances. In Figure 2, each element of the similarity matrix is abbreviated as $\mathcal{S}(r_{feature}^{mn})$ or $\mathcal{S}(r_{labelDis}^{mn})$

To reduce the likelihood of significant disparities in the label distribution representation, $r_{labelDis}$ derived using the reparameterization technique from the Gaussian distribution of label representation are fed into the decoder $\mathcal{F}_{de}$, enabling an alignment between the predicted label distribution and the ground truth distribution.

$$\mathcal{L}_{alig_4} = KL(D||\mathcal{F}_{de}(r_{labelDis})). \tag{16}$$

where D indicates the ground truth, $r_{labelDis}$ is the reparameterization result from the Gaussian distribution of the label representation. $\mathcal{F}_{de}$ is the decoder network, which is represented by three green cubes in Figure 2.

3.3 Objective

During the training phase, the optimization objective of our proposed DILDL consists of six components, namely, two decoupled imbalance label distribution learning losses and four alignment losses.

$$\begin{aligned} \mathcal{L}_{total} = &\underbrace{\alpha\mathcal{L}_{DDL} + (1-\alpha)\mathcal{L}_{NDDL}}_{DILDL} \\ &\underbrace{+\underbrace{\lambda(\alpha\mathcal{L}_{alig_1} + (1-\alpha)\mathcal{L}_{alig_2})}_{Decoupled\ Alignment} + \beta\mathcal{L}_{alig_3} + \gamma\mathcal{L}_{alig_4}}_{Alignment}. \end{aligned} \tag{17}$$

where α, β, γ and λ are balance parameters to balance the weight during the training process in the total loss $\mathcal{L}_{total}$. In the inference stage, the predicted label distribution can be obtained from $\mathcal{F}_{decoder}(\mathcal{F}_{encoder}(\boldsymbol{x}^*))$.

4 Experiment

In this section, we conduct extensive experiments on six ILDL datasets, which are sampled from standard LDL datasets, to assess the effectiveness of our proposed decoupled imbalance label distribution learning approach. In the following subsections, we will report on the datasets used, the evaluation metrics, the experiment setup, the results, and further analysis. All experiments were implemented using the PyTorch framework and executed on one NVIDIA GeForce RTX 4060 GPU. The code of the paper has been open-sourced.

4.1 Datasets and Evaluation

The datasets encompass a diverse range of sources, including SCUT-FBP [Xie *et al.*, 2015], Flicker-LDL [Yang *et al.*, 2017a; Yang *et al.*, 2017b], Movie [Geng and Hou, 2015], Emotion6 [Peng *et al.*, 2015], Natural Scene [Geng, 2016], and RAF-ML [Li and Deng, 2019], each providing unique insights and challenges for our research.

To ensure the robustness and generalizability of our model, we adopt a rigorous experimental design. Specifically, we randomly split each dataset 10 times, allocating a substantial portion of 90% of the data to the combined training and validation sets. Within this 90%, we typically further subdivide the data into separate training and validation subsets to fine-tune our models and prevent overfitting. The remaining 10% of the data is reserved for the test set, serving as an unbiased evaluation of our model's performance on unseen data.

To better verify the proposed DILDL, following ILDL [Zhao *et al.*, 2023b], four distance metrics (Chebyshev ↓, Clark↓, Canberra↓, Kullback-Leibler↓) and two similarity metrics (Cosine↑, Intersection↑) are adopted to evaluate the performance of all the methods. The "↓" after the distance metrics indicates "the smaller the better", and the "↑" after the similarity metrics indicates "the larger the better".

4.2 Implementation Details

We select 10 comparison methods, which fall into three major categories: LDL algorithms, adaptation ILDL algorithms, and specially designed ILDL algorithm. SA-BFGS [Geng, 2016], EDL-LRL [Jia *et al.*, 2019b], LDLSF [Ren *et al.*, 2019a], LDL-LCLR [Ren *et al.*, 2019b], Adam-LDL-SCL [Jia *et al.*, 2019a] and LDL-LDM [Wang and Geng, 2021a] are six state-of-the-art LDL algorithms. Following the objective function reshaping method [Zhao *et al.*, 2023b], OFR-FL, OFR-CB, OFR-DB techniques are used to reshape the LDL algorithms into three adaptation ILDL approaches. In addition, the latest specially designed ILDL algorithm RDA [Zhao *et al.*, 2023b] is also selected as the comparison method.

The learning rate is set 0.001. The batch size is 50. The trade-off parameter α in Eq. (17) is 0.6, which is selected from parameter sensitivity analysis. The trade-off parameters λ, β and γ for alignment are set 0.1. The maximum epoch is 300. During the inference stage, the predicted label distribution is obtained from the decoder network after the feature encoder.

Algorithm	Movie	SCUT - FBP	Emotion6	Flickr_LDL	RAF - ML	Natural Scene
SA - BFGS	0.3415±0.0070•	0.7266±0.0326•	0.8292±0.0179•	0.8948±0.0149•	0.7575±0.0149•	0.6621±0.0198•
EDL - LRL	0.3638±0.0118•	0.3522±0.0236•	0.4175±0.0074•	0.5811±0.0060•	0.4784±0.0137•	0.4341±0.0233•
LDLSF	0.3624±0.0107•	0.4701±0.0307•	0.4355±0.0106•	0.5697±0.0092•	0.4177±0.0174•	0.4440±0.0249•
LDL - LCLR	0.3346±0.0072•	0.3332±0.0246•	0.5239±0.0136•	0.7033±0.0126•	0.3849±0.0107•	0.5680±0.0225•
Adam - LDL - SCL	0.7175±0.0487•	0.4460±0.0218•	0.4711±0.0333•	0.6711±0.0547•	0.5848±0.0300•	0.4773±0.0344•
LDL - LDM	0.4858±0.0285•	0.4030±0.0441•	0.4739±0.0159•	0.5816±0.0085•	0.5348±0.0275•	0.4769±0.0234•
OFR - FL	0.3416±0.0151•	0.3364±0.0357•	0.3910±0.0102•	0.5636±0.0054•	0.5081±0.0236•	0.4323±0.0201•
OFR - CB	0.3337±0.0177•	0.3447±0.0289•	0.3922±0.0091•	0.5658±0.0059•	0.5057±0.0161•	0.4329±0.0209•
OFR - DB	0.2548±0.0080•	0.3199±0.0384•	0.3772±0.0072•	0.5252±0.0205•	0.4638±0.0196•	0.3872±0.0254•
RAD	0.1962±0.0068•	0.2849±0.0157	0.3598±0.0079•	0.5208±0.0075•	0.3756±0.0068•	0.3768±0.0208
DILDL (Ours)	**0.1752±0.0091**	**0.2684±0.0182**	**0.3485±0.0061**	**0.5025±0.0073**	**0.3484±0.0102**	**0.3624±0.0212**

Table 1: Experimental results on ILDL datasets measured by Chebyshev distance ↓.

Algorithm	Movie	SCUT-FBP	Emotion6	Flickr_LDL	RAF-ML	Natural Scene
SA-BFGS	0.8007±0.0539•	13.0419±4.1007•	21.8514±1.0523•	27.1262±1.5508•	18.2051±1.2023•	4.7976±0.3734•
EDL-LRL	0.7797±0.0472•	0.8111±0.1085•	1.4348±0.1160•	9.9140±4.5756•	1.2838±0.0994•	2.5862±1.5835•
LDLSF	3.1338±0.3786•	8.4136±1.6575•	9.4371±0.5063•	12.8509±1.0510•	7.0684±1.1409•	8.8454±0.5594•
LDL-LCLR	0.6803±0.0314•	0.6034±0.0788•	2.2820±0.1581•	6.2168±0.2896•	1.0106±0.0704•	2.9449±0.2527•
Adam-LDL-SCL	19.1715±1.6303•	2.3768±1.1735•	8.1116±4.8903•	17.1944±8.5188•	6.1170±4.2557•	9.6209±4.8989•
LDL-LDM	1.8123±0.2788•	1.0253±0.2190•	1.7890±0.1369•	2.7424±0.2096•	1.9157±0.2248•	1.7753±0.2056•
OFR-FL	0.6459±0.0567•	0.6415±0.1438•	1.1829±0.0959•	2.5989±0.1650•	1.3672±0.1676•	1.3364±0.0981•
OFR-CB	0.6288±0.0604•	0.6581±0.1171•	1.1904±0.0776•	2.6285±0.3774•	1.3264±0.1110•	1.3280±0.0932•
OFR-DB	0.3883±0.0160•	0.5577±0.1317•	0.9238±0.0238•	1.7751±0.2858•	1.1481±0.0823•	1.1746±0.0898•
RAD	0.2491±0.0149•	0.4313±0.0328	0.7677±0.0218•	1.6071±0.1107•	0.7058±0.0203•	1.1188±0.0591
DILDL (Ours)	**0.2211±0.0121**	**0.4131±0.0315**	**0.7393±0.0254**	**1.5011±0.0925**	**0.6025±0.0320**	**1.0734±0.0437**

Table 2: Experimental results on ILDL datasets measured by Kullback-Leibler divergence ↓.

DRDA	DLD	Cheb↓	Clark↓	Can↓	KL↓	Cos↑	Inter↑
		0.198	0.802	1.561	0.254	0.808	0.703
✓		0.182	0.781	1.544	0.241	0.821	0.716
	✓	0.187	0.792	1.552	0.246	0.815	0.710
✓	✓	**0.175**	**0.768**	**1.529**	**0.221**	**0.834**	**0.729**

Table 3: The results of the ablation study on Movie dataset.

DRDA	DLD	Cheb↓	Clark↓	Can↓	KL↓	Cos↑	Inter↑
		0.375	2.485	6.930	1.118	0.597	0.404
✓		0.369	2.481	6.916	1.071	0.606	0.406
	✓	0.370	2.481	6.919	1.096	0.604	0.405
✓	✓	**0.362**	**2.479**	**6.907**	**1.073**	**0.610**	**0.409**

Table 4: The results of the ablation study on Natural Scene dataset.

4.3 Results

Table 1 and 2 show the main results of different methods on six dataset measured by Chebyshev and KL divergence. The best performances are highlighted in bold. And •/◦ represents whether our proposed method is statistically superior/inferior to the comparing methods, which is calculated by two-tailed t-test under 0.05 significance level. From the results, our proposed algorithm consistently achieves the lowest Chebyshev distance across all datasets, indicating superior performance compared to the other algorithms. Specifically, our proposed method outperforms all others with a distance of 0.1752±0.0091, significantly lower than the next best (RAD with 0.1962±0.0068) on Movie dataset. Similarly, our method achieves 0.2684±0.0182, outperforming RAD (0.2849±0.0157) and other competitors on SCUT-FBP dataset. Our proposed algorithm also demonstrates the best performance in terms of KL divergence across all datasets, consistent with the Chebyshev distance results. For example, on Emotion6 dataset, our method gains the best performance with a divergence of 0.7393±0.0254. Our method's divergence of 1.5011±0.1165 is notably lower than RAD's 1.6071±0.1107 on Flickr_LDL, underlining its efficiency. Table 5 also verifies the effectiveness of our method.

In summary, the experimental results clearly demonstrate that our proposed algorithm significantly outperforms the other algorithms in terms of both Chebyshev distance, KL divergence and Euclidean distance across multiple ILDL datasets. These findings underscore the robustness and effectiveness of our method in tackling imbalance label distribution learning tasks. The consistent lower distances and divergences achieved by our algorithm across diverse datasets suggest its generalizability and potential for practical applications where accurate imbalance label distribution learning is paramount importance.

4.4 Ablation Study

The decoupled section of our proposed method seamlessly integrates both the decoupled label distribution and the decoupled representation distribution alignment. To evaluate the impact of each component within our proposed method, we conducted a rigorous ablation study by defining variations of DILDL. Specifically, "DRDA" denotes the decoupled representation distribution alignment component, while "DLD" signifies the decoupled label distribution component.

Tables 3 and 4 present the results of the ablation studies conducted on both the Movie dataset and the Natural Scene

Algorithm	Movie			SCUT-FBP			Emotion6			Flickr_LDL			RAF-ML			Nature Scene		
	all	tail	head	all	tail	head	all	tail	head	all	tail	head	all	tail	head	all	tail	head
SA-BFGS	.473•	.346•	.303•	.914•	.760•	.355•	.992•	.888•	.269•	1.108•	.617•	.864•	.936•	.849•	.221•	.834•	.736•	.257•
EDL-LRL	.495•	.360•	.315•	.473•	.323•	.326•	.572•	.508•	.214•	.779•	.567•	.501•	.652•	.490•	.416•	.582•	.511•	.242•
LDLSF	.502•	.359•	.337•	.643•	.442•	.447•	.602•	.540•	.198•	.771•	.552•	.492•	.565•	.490•	.218•	.611•	.565•	.181•
LDL-LCLR	.466•	.337•	.309•	.449•	.318•	.298•	.692•	.625•	.201•	.919•	.599•	.632•	.502•	**.438**◦	.177•	.736•	.672•	.211•
Adam-LDL-SCL	.856•	.676•	.378•	.639•	.461•	.435•	.672•	.618•	.225•	.914•	.579•	.679•	.806•	.587•	.527•	.649•	.563•	.297•
LDL-LDM	.618•	.484•	.297•	.539•	.420•	.286•	.643•	.583•	.203•	.779•	.563•	.498•	.721•	.555•	.416•	.634•	.578•	.197•
OFR-FL	.497•	.336•	.338•	.469•	.337•	.315•	.540•	.495•	.190•	.755•	.556•	.480•	.691•	.496•	.473•	.574•	.499•	.253•
OFR-CB	.472•	.334•	.329•	.481•	.342•	.330•	.541•	.497•	.190•	.760•	.555•	.488•	.687•	.491•	.473•	.575•	.499•	.256•
OFR-DB	.377•	.285•	.243•	.443•	.332•	.284•	.503•	.465•	.163•	.666•	.524	.374•	.636•	.498•	.384•	.526•	.492•	.162•
RAD	.295•	.245•	.157•	.386•	.299•	.234•	.464•	.429•	.133•	.645•	.526•	.333•	.484•	.454•	.132•	.515•	.486•	.151•
DILDL (Ours)	**.251**	**.227**	**.133**	**.350**	**.273**	**.198**	**.452**	**.427**	**.112**	**.631**	**.524**	**.313**	**.451**	.452	**.121**	**.497**	**.473**	**.132**

Table 5: Experimental results (*tail, head and all labels*) on ILDL datasets measured by Euclidean Distance ↓.

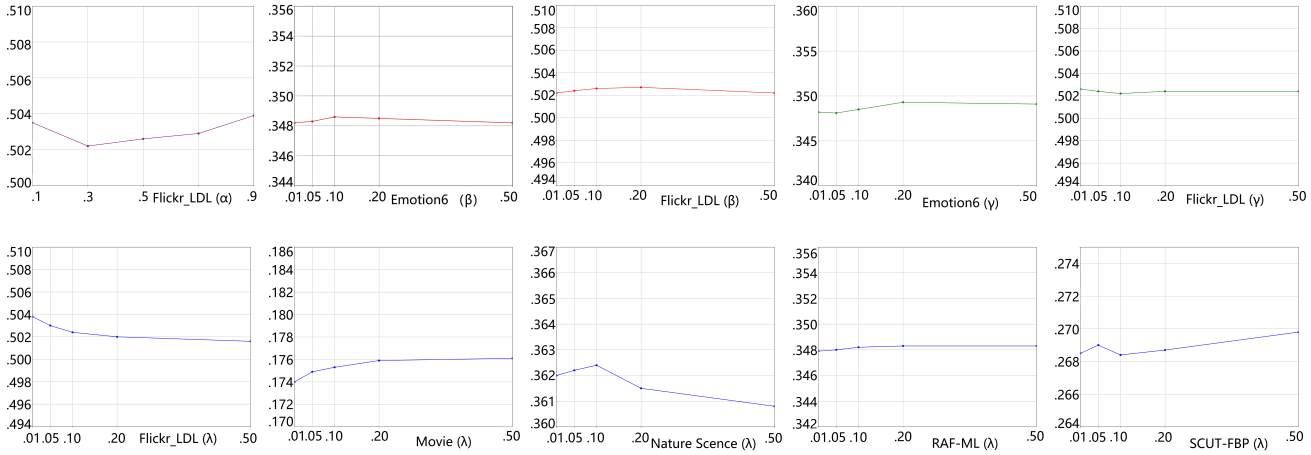

Figure 3: Effects of the values of α, β, γ and λ settings on ILDL datasets on Chebyshev Distance ↓. The first row shows the results of parameters α, β, γ on Emotion6 and Flickr_LDL datasets. The second row shows the results of parameter λ on six datasets.

dataset, offering insights into the influence of the DRDA and DLD components across various evaluation metrics. The table for the Movie dataset (Table 4) illustrates the performance of different configurations with and without DRDA and DLD. When both DRDA and DLD are active (indicated by checkmarks), we observe significant improvements across all metrics compared to when either or both are deactivated. Similarly, the results for the Natural Scene dataset (Table 5) demonstrate the advantages of combining DRDA and DLD. For instance, the result decreases from 1.118 to 1.063 when both DRDA and DLD are enabled, indicating better alignment of distributions. Without DRDA, the result increases from 1.063 to 1.096 under the KL divergence metric, which underscores the effectiveness of the proposed DRDA. The combined results from both datasets consistently reveal that enabling both DRDA and DLD leads to enhanced performance across all metrics. This underscores the positive contribution of both components in improving the quality of imbalance label distribution learning.

4.5 Parameter Sensitivity Analysis

To select the optimal balance parameters, we compare the performances of our proposed DILDL with various values of hyperparameters across six datasets, evaluating the performance using the Chebyshev distance. Figure 3 illustrates the results of the parameter sensitivity analysis. From the figure, we can conclude that the parameter α exhibits a higher degree of sensitivity compared to the other three parameters, suggesting that meticulous tuning of this parameter can further enhance the model's performance. When $\alpha = 0.6$, the model's performance is optimal in most cases. Additionally, DILDL exhibits better performances with 0.1 for the other three hyperparameters.

5 Conclusion

LDL has been proven to be an efficient learning paradigm for solving the label ambiguity problems. However, it may encounter the challenge posed by ILDL. In this paper, we propose a novel ILDL method named Decoupled Imbalanced Label Distribution Learning (DILDL). DILDL decomposes imbalanced label distributions into dominant and non-dominant components. By employing the decoupling approach, we independently balance the description degrees of both dominant and non-dominant labels. Furthermore, we separately align feature representations with those of dominant and non-dominant labels, thereby significantly mitigating the issue of gradient information attenuation for non-dominant labels. Experimental results demonstrate that our proposed method outperforms other methods. In the future, we aim to explore the application of feature decoupling methods to further enhance the performance of the DILDL approach.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62406155, the project No. ZR2024QF115 supported by Shandong Provincial Natural Science Foundation, the National Natural Science Foundation of China under No. 62471202 and No. 62476135, the Innovation Capability Enhancement Project for Technology-based Small and Medium-sized Enterprises of Shandong Province under Grant No. 2024TSGC0777, the Opening Fund of Shandong Provincial Key Laboratory of Ubiquitous Intelligent Computing, the Shandong Province Youth Innovation Team Project under Grant No. 2024KJH032), and the Development Program Project of Youth Innovation Team of Institutions of Higher Learning in Shandong Province.

References

[An *et al.*, 2024] Yuexuan An, Hui Xue, Xingyu Zhao, Ning Xu, Pengfei Fang, and Xin Geng. Leveraging bilateral correlations for multi-label few-shot learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Buda *et al.*, 2018] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

[Byrd and Lipton, 2019] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, 2019.

[Chen *et al.*, 2020] Shikai Chen, Jianfeng Wang, Yuedong Chen, Zhongchao Shi, Xin Geng, and Yong Rui. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[Fu *et al.*, 2024] Huiqiao Fu, Kaiqiang Tang, Yuanyang Lu, Yiming Qi, Guizhou Deng, Flood Sung, and Chunlin Chen. Ess-infogail: Semi-supervised imitation learning from imbalanced demonstrations. In *Advances in Neural Information Processing Systems*, 2024.

[Gao *et al.*, 2021] Yongbiao Gao, Ning Xu, and Xin Geng. Video summarization via label distributions dual-reward. In *International Joint Conference on Artificial Intelligence*, 2021.

[Geng and Hou, 2015] Xin Geng and Peng Hou. Pre-release prediction of crowd opinion on movies by label distribution learning. In *International Joint Conference on Artificial Intelligence*, 2015.

[Geng, 2016] Xin Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016.

[Ghosh *et al.*, 2024] Kushankur Ghosh, Colin Bellinger, Roberto Corizzo, Paula Branco, Bartosz Krawczyk, and Nathalie Japkowicz. The class imbalance problem in deep learning. *Machine Learning*, 113(7):4845–4901, 2024.

[He and Garcia, 2009] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[Jia *et al.*, 2019a] 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.

[Jia *et al.*, 2019b] Xiuyi Jia, Xiang Zheng, Weiwei Li, Changqing Zhang, and Zechao Li. Facial emotion distribution learning by exploiting low-rank label correlations locally. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[Kou *et al.*, 2024a] Zhiqiang Kou, Jing Wang, Yuheng Jia, and Xin Geng. Inaccurate label distribution learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[Kou *et al.*, 2024b] Zhiqiang Kou, Jing Wang, Jiawei Tang, Yuheng Jia, and Xin Geng. Exploiting multi-label correlation in label distribution learning. In *International Joint Conference on Artificial Intelligence*, 2024.

[Le *et al.*, 2023] Nhat Le, Khanh Nguyen, Quang Tran, Erman Tjiputra, Bac Le, and Anh Nguyen. Uncertainty-aware label distribution learning for facial expression recognition. In *IEEE Winter Conference on Applications of Computer Vision*, 2023.

[Li and Deng, 2019] 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):884–906, 2019.

[Li *et al.*, 2024] Weiwei Li, Wei Qian, Lei Chen, and Xiuyi Jia. Sample diversity selection strategy based on label distribution morphology for active label distribution learning. *Pattern Recognition*, 150:110322, 2024.

[Liu *et al.*, 2023] Gang Liu, Tong Zhao, Eric Inae, Tengfei Luo, and Meng Jiang. Semi-supervised graph imbalanced regression. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

[Lu and Jia, 2024] Yunan Lu and Xiuyi Jia. Predicting label distribution from ternary labels. In *Advances in Neural Information Processing Systems*, 2024.

[Oh *et al.*, 2022] Youngtaek Oh, Dong-Jin Kim, and In So Kweon. Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[Park and Lee, 2024] Hyungkeun Park and Jong-Seok Lee. Adaptive explicit knowledge transfer for knowledge distillation. *arXiv preprint arXiv:2409.01679*, 2024.

[Peng *et al.*, 2015] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: model, predict, and transfer emotion distributions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[Peng *et al.*, 2025] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

[Ren and Geng, 2017] Yi Ren and Xin Geng. Sense beauty by label distribution learning. In *International Joint Conference on Artificial Intelligence*, 2017.

[Ren *et al.*, 2019a] Tingting Ren, Xiuyi Jia, Weiwei Li, Lei Chen, and Zechao Li. Label distribution learning with label-specific features. In *International Joint Conference on Artificial Intelligence*, 2019.

[Ren *et al.*, 2019b] Tingting Ren, Xiuyi Jia, Weiwei Li, and Shu Zhao. Label distribution learning with label correlations via low-rank approximation. In *International Joint Conference on Artificial Intelligence*, 2019.

[Rezende *et al.*, 2014] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

[Shen *et al.*, 2017] Wei Shen, Kai Zhao, Yilu Guo, and Alan L Yuille. Label distribution learning forests. *Advances in Neural Information Processing Systems*, 2017.

[Smith-Miles and Geng, 2020] Kate Smith-Miles and Xin Geng. Revisiting facial age estimation with new insights from instance space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2689–2697, 2020.

[Soltanzadeh *et al.*, 2023] Paria Soltanzadeh, M Reza Feizi-Derakhshi, and Mahdi Hashemzadeh. Addressing the class-imbalance and class-overlap problems by a metaheuristic-based under-sampling approach. *Pattern Recognition*, 143:109721, 2023.

[Wang and Geng, 2019] Jing Wang and Xin Geng. Theoretical analysis of label distribution learning. In *AAAI Conference on Artificial Intelligence*, 2019.

[Wang and Geng, 2021a] Jing Wang and Xin Geng. Label distribution learning by exploiting label distribution manifold. *IEEE Transactions on Neural Networks and Learning Systems*, 34(2):839–852, 2021.

[Wang and Geng, 2021b] Jing Wang and Xin Geng. Learn the highest label and rest label description degrees. In *International Joint Conference on Artificial Intelligence*, 2021.

[Wang and Geng, 2024] Jing Wang and Xin Geng. Large margin weighted k-nearest neighbors label distribution learning for classification. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):116720–116732, 2024.

[Wang and Wang, 2024] Ziyan Wang and Hao Wang. Variational imbalanced regression: Fair uncertainty quantification via probabilistic smoothing. *Advances in Neural Information Processing Systems*, 2024.

[Wang *et al.*, 2024] Jing Wang, Zhiqiang Kou, Yuheng Jia, Jianhui Lv, and Xin Geng. Label distribution learning by exploiting fuzzy label correlation. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Wu *et al.*, 2020] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision*, 2020.

[Xie *et al.*, 2015] Duorui Xie, Lingyu Liang, Lianwen Jin, Jie Xu, and Mengru Li. Scut-fbp: A benchmark dataset for facial beauty perception. In *IEEE International Conference on Systems, Man, and Cybernetics*, 2015.

[Xu *et al.*, 2023] Ning Xu, Biao Liu, Jiaqi Lv, Congyu Qiao, and Xin Geng. Progressive purification for instance-dependent partial label learning. In *International Conference on Machine Learning*, 2023.

[Yang *et al.*, 2017a] Jufeng Yang, Ming Sun, and Xiaoxiao Sun. Learning visual sentiment distributions via augmented conditional probability neural network. In *AAAI Conference on Artificial Intelligence*, 2017.

[Yang *et al.*, 2017b] Jufeng Yang, Ming Sun, and Xiaoxiao Sun. Learning visual sentiment distributions via augmented conditional probability neural network. In *AAAI Conference on Artificial Intelligence*, 2017.

[Yang *et al.*, 2021] Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *International Conference on Machine Learning*, 2021.

[Yang *et al.*, 2023] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36:40924–40943, 2023.

[Zhang *et al.*, 2021] Huiying Zhang, Yu Zhang, and Xin Geng. Practical age estimation using deep label distribution learning. *Frontiers of Computer Science*, 15:1–6, 2021.

[Zhang *et al.*, 2023] Yu Zhang, Junjie Zhao, Zhengjie Chen, Siya Mi, Hongyuan Zhu, and Xin Geng. A closer look at video sampling for sequential action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7503–7514, 2023.

[Zhao *et al.*, 2022] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[Zhao *et al.*, 2023a] Xingyu Zhao, Yuexuan An, Ning Xu, and Xin Geng. Variational continuous label distribution learning for multi-label text classification. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[Zhao *et al.*, 2023b] Xingyu Zhao, Yuexuan An, Ning Xu, Jing Wang, and Xin Geng. Imbalanced label distribution learning. In *AAAI Conference on Artificial Intelligence*, 2023.