Imbalanced Label Distribution Learning

Xingyu Zhao[1,2], Yuexuan An[1,2], Ning Xu[1,2], Jing Wang[1,2], Xin Geng[1,2]*

[1]School of Computer Science and Engineering, Southeast University, Nanjing 211189, China
[2]Key Laboratory of Computer Network and Information Integration (Ministry of Education), Southeast University, Nanjing 211189, China
{xyzhao, yx_an, xning, wangjing91, xgeng}@seu.edu.cn

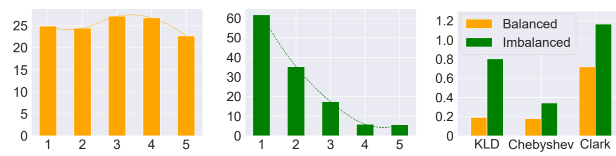
Abstract

Label distribution covers a certain number of labels, representing the degree to which each label describes an instance. The learning process on the instances labeled by label distributions is called Label Distribution Learning (LDL). Although LDL has been applied successfully to many practical applications, one problem with existing LDL methods is that they are limited to data with balanced label information. However, annotation information in real-world data often exhibits imbalanced distributions, which significantly degrades the performance of existing methods. In this paper, we investigate the *Imbalanced Label Distribution Learning* (ILDL) problem. To handle this challenging problem, we delve into the characteristics of ILDL and empirically find that the representation distribution shift is the underlying reason for the performance degradation of existing methods. Inspired by this finding, we present a novel method named Representation Distribution Alignment (RDA). RDA aligns the distributions of feature representations and label representations to alleviate the impact of the distribution gap between the training set and the test set caused by the imbalance issue. Extensive experiments verify the superior performance of RDA. Our work fills the gap in benchmarks and techniques for practical ILDL problems.

Introduction

Learning with ambiguity is one of the most important machine learning topics since data ambiguity is ubiquitous in the real world (Geng 2016; Gao et al. 2017). *Label distribution learning* (LDL) is a novel paradigm for dealing with data ambiguity. LDL assigns each instance a label distribution and learns the mapping from instances to label distributions. Each element of a label distribution is called the label description degree that explicitly indicates the relative importance of the corresponding label to an instance. As the utility of dealing with ambiguity explicitly, LDL has been successfully applied to many real applications, such as facial landmark detection (Su and Geng 2019), age estimation (Gao et al. 2018), head poses estimation (Geng and Xia 2014), zero-shot learning (Huo and Geng 2017), emotion analysis (Yang et al. 2021a) and autism spectrum disorder classification (Wang et al. 2022).

*Corresponding author.

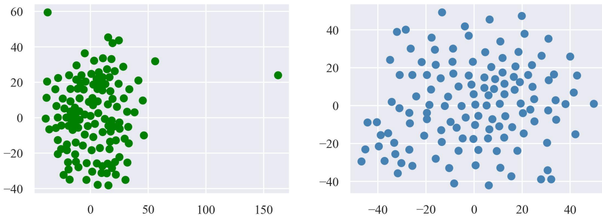


(a) Balanced (b) Imbalanced (c) Results

Figure 1: Performance comparison of a standard LDL model (i.e., SA-BFGS) respectively trained on the balanced *Movie* dataset and imbalanced *Movie* dataset. In (a) and (b), the x-axes indicate the rating score, the y-axes indicate the sum of the description degree. In (c), the x-axis indicates the LDL evaluation criterion, the y-axis indicates the result value (the smaller the better).

Despite the fact that LDL achieved success in many applications, one limitation with existing LDL methods is that they are designed for data with balanced supervision information in different labels. That is, the distribution of label annotation information is balanced, which means the sum of the label description degree for each label is approximately equal. However, annotation information in real-world data often exhibits imbalanced distributions, which significantly degrades the performance of existing methods (He and Garcia 2009; Wu et al. 2020). For example, when training a movie rating distribution model for some types of movies, the description degree distribution corresponding to a certain rating may be much higher than other ratings due to the possible deviation of the data collection means. Therefore, the ideal dataset shown in Figure 1(a) may be difficult to gather, and it is possible to obtain the imbalanced dataset shown in Figure 1(b). We use a standard LDL method (i.e., SA-BFGS) to train two models from these training sets separately and test them on the given balanced test set (Geng 2016). From the results given in Figure 1(c), we can find that the performance of the model trained on the imbalanced dataset is significantly worse than that of the model trained on the balanced dataset on each evaluation criterion. Therefore, how to learn an LDL model resilient to the imbalanced label distribution is challenging and meaningful for the practicality of LDL.

We refer to this new and challenging scenario as *Imbalanced Label Distribution Learning* (ILDL) and systemati-



(a) Training feature (b) Test feature

Figure 2: The T-SNE (Van der Maaten and Hinton 2008) visualizations of the distribution of feature representations in the imbalanced *Movie* dataset. In the distribution of training feature representations, most of the points are clustered, while the points in the test set are relatively uniform.

cally investigate the ILDL problem. We delve into the characteristics of ILDL and reveal the underlying reason hidden from the performance degradation of existing methods in ILDL. As we can see from Figure 2, it exists an obvious shift in feature representation distribution due to the significant differences in the distribution of label annotation information, where most points in Figure 2(a) are obviously clustered, while the distribution in Figure 2(b) is almost uniform. However, existing LDL methods are based on the assumption that the distribution of feature representations in the training set and test set is consistent and directly generalizes the mapping from feature to label distribution on a training set to a test set. Obviously, this assumption is violated in ILDL. Therefore, the performance degradation of these methods is serious in the ILDL scenarios. In fact, feature and label distribution are two kinds of description for a same instance and a natural potential consistency exists in their representation space. If we leverage this potential consistency to establish a reasonable relationship between feature representations and label representations, we can effectively infer label representations from feature representations and obtain high-performance predictions no longer subject to the effect of the feature representation distribution shifts.

Inspired by the above insights, we propose a novel method named *Representation Distribution Alignment* (RDA), which establishes a relationship between feature representations and label representations. RDA first maps the features and labels of instances into different latent spaces. Then, it constructs continuous distributions of feature representations and label representations which are transformed from the mapped values in a common space. By aligning the distributions of feature representations and label representations of instances in the common space, RDA enhances the joint representation ability of the model in both feature space and label space. As shown in Figures 3(a) to 3(d), the distributions of feature representations and label representations of instances can be effectively aligned by RDA, which allows the model to effectively mitigate the representation distribution shift problem to tackle the imbalance problem.

To support the practical evaluation of ILDL, we reshape several objective functions of existing imbalanced learning approaches as strong baselines for the ILDL problem. Moreover, we curate and benchmark ILDL datasets for common real-world tasks in movie rating, facial beauty perception, and visual sentiment distribution perception. We further set up benchmarks for proper ILDL performance evaluation.

Our contributions are as follows:

- We identify Imbalance Label Distribution Learning (ILDL) as a new challenging topic and formally define the setting of the ILDL problem.
- We delve into the characteristics of the ILDL problem and empirically find that the representation distribution shift is the underlying reason for the performance degradation of existing methods.
- We propose a novel method named Representation Distribution Alignment (RDA) for the ILDL problem based on our findings.
- We set up three strong baseline methods for the ILDL problem by reshaping the objective functions of existing imbalanced learning approaches.
- We curate several benchmark datasets for proper ILDL performance evaluation.

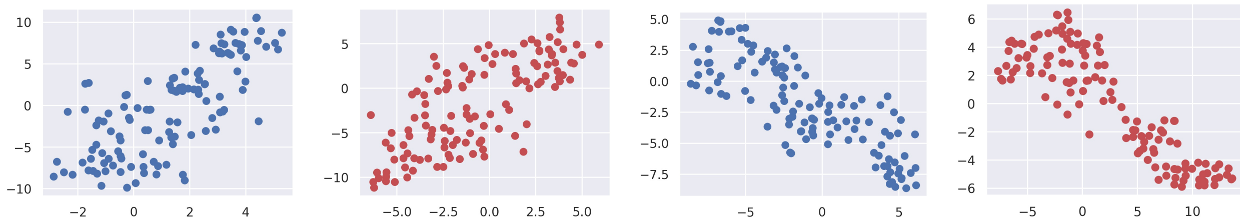
Related Work

Label Distribution Learning

Label distribution learning (LDL) is a novel learning paradigm, which assigns an instance a label distribution and learns a mapping from instances to label distributions straightly (Geng 2016). In recent years, LDL has been widely studied. (Geng, Yin, and Zhou 2013) proposes the first specialized LDL algorithm, whose objective function consists of the maximum entropy model (Berger, Pietra, and Pietra 1996) and KL divergence. (Zhao and Zhou 2018) casts the label correlations exploration as a ground metric learning problem and adopts optimal transport distance to measure the quality of prediction. (Ren et al. 2019a) exploits the label correlations and learns the common features for all labels and specific features for each label simultaneously. (Jia et al. 2019) exploits local label correlation by capturing low-rank structure on clusters of samples with trace-norm regularization. (Zheng, Jia, and Li 2018) and (Jia et al. 2021) consider label correlation to be local and learn optimal encoding vector and label distribution simultaneously. (Ren et al. 2019b) captures global label correlation with a low-rank matrix and updates the matrix on different clusters to explore local label correlation, which exploits both global and local label correlations. However, these methods are not resilient to the ILDL problem since they do not consider the distribution gap between the training set and the test set.

Imbalanced Learning

Arising from long-tail distributions of natural data, imbalanced learning has been extensively studied. Imbalanced classification (also referred to as long-tailed recognition) (Liu et al. 2019) is one popular topic and numerous methods have been proposed. These works mainly follow two directions. One line of these approaches is re-sampling, which uses under-sampling (Buda, Maki, and Mazurowski 2018)



(a) Training feature (RDA) (b) Training labels (RDA) (c) Test feature (RDA) (d) Test labels (RDA)

Figure 3: The T-SNE (Van der Maaten and Hinton 2008) visualizations of the feature representations and label representations encoded by RDA.

or over-sampling (Byrd and Lipton 2019) to achieve a relatively balanced dataset. However, the former might weaken feature learning capacity due to omitting a number of valuable samples, and the latter might lead to over-fitting minority classes with duplicated samples. In the meantime, (Wu et al. 2020) indicates that the adaption of this technology into the multi-label setting will not cause a significant change in the label frequency. The other line of these approaches is cost-sensitive, which assigns a weight to each sample according to cost metrics. (Lin et al. 2020) uses the output of the predictive model as the weights. (Cui et al. 2019) proposes a novel class-balanced loss that re-weighting the loss of different labels by the inverse of the effective number of samples. (Wu et al. 2020) applies re-weighting based on the class frequency and modifies the loss gradient with a regularization as well for better optimization.

There are also several works that focus on imbalanced regression. (Torgo et al. 2013) is the first work to address this problem by adopting the SMOTE algorithm (Chawla et al. 2002). (Branco, Torgo, and Ribeiro 2017) presents a Gaussian noise-based synthetic case generation method. (Branco, Torgo, and Ribeiro 2018) introduces a bagging-based ensemble method. Recently, (Yang et al. 2021b) further delves into this problem. It exploits the similarity between nearby targets in target space and feature space, and proposed two smoothing methods for targets and features.

Intuitively, ILDL is similar to the existing imbalanced classification and regression problems in that specific target values have significantly fewer observations (Liu et al. 2019; Cao et al. 2019; Zhou et al. 2020; Yang et al. 2021b). However, it brings greater challenges distinct from imbalanced classification and regression. Compared with imbalanced classification, its target values for each label become continuous, which causes ambiguity when directly applying existing approaches such as re-sampling and re-weighting. Compared with imbalanced regression, ILDL considers not only the continuous target values but also the distribution of these values brought by multiple related label dimensions.

Problem Setting

First, the main notations used in this paper are listed as follows. The instance variable is denoted by $\boldsymbol{x}$, the particular i-th instance is denoted by $\boldsymbol{x}_i$, the label variable is denoted by y, the particular j-th label value is denoted by y_j, the description degree of y to $\boldsymbol{x}$ is denoted by $d_{\boldsymbol{x}}^{y}$, the label distribution of $\boldsymbol{x}_i$ is denoted by $\boldsymbol{d}_i = \{d_{\boldsymbol{x}_i}^{y_1}, d_{\boldsymbol{x}_i}^{y_2}, \cdots, d_{\boldsymbol{x}_i}^{y_c}\}$, where c is the number of possible label values, $d_{\boldsymbol{x}_i}^{y_j} \in [0,1]$ and $\sum_{j=1}^{c} d_{\boldsymbol{x}_i}^{y_j} = 1$. In this paper, we consider the imbalanced label distribution setting where the sum of description degrees of different labels are significantly different. Formally, we define the imbalanced label distribution learning problem as follows.

Problem 1 (Imbalanced Label Distribution Learning, ILDL)**.** *Let $\mathcal{X} = \mathbb{R}^q$ denote the input space and $\mathcal{Y} = \{y_1, y_2, \cdots, y_c\}$ denote the complete set of labels. We consider an imbalance training set $S = \{(\boldsymbol{x}_i, \boldsymbol{d}_i)\}_{i=1}^{N}$, where $\boldsymbol{x}_i \in \mathcal{X}$ is a q-dimensional real value vector, $\boldsymbol{d}_i = \left(d_{\boldsymbol{x}_i}^{y_1}, d_{\boldsymbol{x}_i}^{y_2}, \cdots, d_{\boldsymbol{x}_i}^{y_c}\right)^T \in [0,1]^c$ is the corresponding c-dimensional label distribution. The imbalanced factor of the training set $\gamma = \max\left\{\left(\sum_{i=1}^{N} d_{\boldsymbol{x}_i}^{y_j}\right)\right\}_{j=1}^{c} / \min\left\{\left(\sum_{i=1}^{N} d_{\boldsymbol{x}_i}^{y_j}\right)\right\}_{j=1}^{c}$ is greater than a large threshold (e.g. $\gamma > 10$). The goal of ILDL is to learn a mapping from an instance $\boldsymbol{x}$ to its corresponding label distribution $\mathbf{d}$, which can achieve high performance on a balanced test set $S^* = \{(\boldsymbol{x}_i^*, \boldsymbol{d}_i^*)\}_{i=1}^{N^*}$.*

Our purpose is to train an LDL model on the imbalanced training set, which can achieve better performance in the relatively balanced test set.

Methods

In this section, we first reshape several objective functions of existing imbalanced learning approaches as strong baselines for the ILDL problem. Furthermore, we propose a novel method, *Representation Distribution Alignment*, which alleviates the impact of the distribution gap between the training set and the test set by aligning the distributions of feature representations and label representations of instances.

Objective Function Reshaping

In LDL, Kullback-Leibler (KL) divergence between the ground truth and the predicted label distribution is a commonly used loss function. Assume $f_\theta(\cdot)$ is a mapping from $\mathcal{X}$ to $\mathcal{Y}$. The objective function of LDL can be formulated

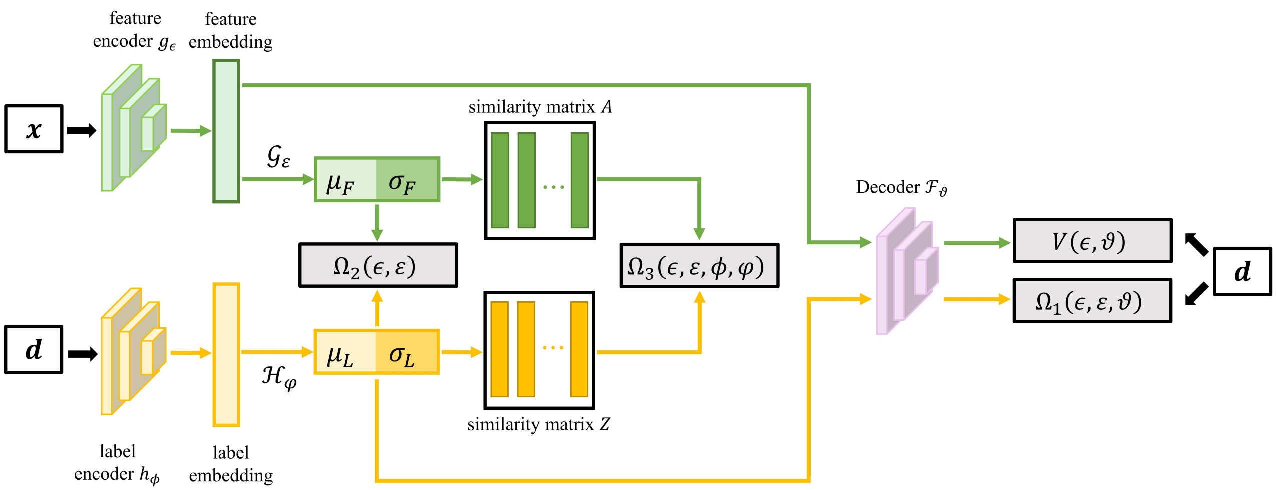


Figure 4: Overview of the proposed Representation Distribution Alignment method.

by

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

where $f_\theta^{(j)}(\boldsymbol{x}_i)$ is the output for the j-th label of $f_\theta(\boldsymbol{x}_i)$. The plain KL divergence loss function may be vulnerable to label imbalance due to the observations of different labels are significantly different. Therefore, we reshape the objective functions from imbalanced classification to make it be resilient to ILDL.

Focal loss. Focal loss places a higher weight of loss on instances predicted with low probability on ground truth to emphasize the importance of "hard-to-classify" instances. (Lin et al. 2020). In ILDL, we modify the original focal loss to the following form:

$$\mathcal{L}_{OFR-FL} = \sum_{i=1}^{n} \sum_{j=1}^{c} \left(1 - f_\theta^{(j)}(\boldsymbol{x}_i)\right)^{\gamma} d_{\boldsymbol{x}_i}^{y_j} \ln \frac{d_{\boldsymbol{x}_i}^{y_j}}{f_\theta^{(j)}(\boldsymbol{x}_i)}, \tag{2}$$

where $\gamma \geq 0$ is a tunable focusing parameter. The idea of Eq.(2) is consistent with the original focal loss, which utilizes the predicted values of different labels to weight the original loss, so as to better deal with label imbalance.

Class-balanced focal loss. Class-balanced focal loss estimates the effective number of samples of each class and uses them to further reweight focal loss (Cui et al. 2019). Compared with focal loss, class-balanced focal loss integrates the class-level information into the loss function, which can capture the diminishing marginal benefits of data and reduce redundant information of head classes. In ILDL, we modify the number of samples of each class to the sum of the description degree of each label, i.e., $\hat{N}_j = \sum_{i=1}^{N} d_{\boldsymbol{x}_i}^{y_j}$, and the "effective number" of samples of each class is redefined as the "effective description degree" of each label, i.e., $r_{CB}^{j} = 1 - \beta/(1 - \beta^{\hat{N}_j})$. The modified class-balanced focal loss is defined as

$$\mathcal{L}_{OFR-CB} = \sum_{i=1}^{n} \sum_{j=1}^{c} r_{CB}^{j} \left(1 - f_\theta^{(j)}\right)^{\gamma} d_{\boldsymbol{x}_i}^{y_j} \ln \frac{d_{\boldsymbol{x}_i}^{y_j}}{f_\theta^{(j)}(\boldsymbol{x}_i)}. \tag{3}$$

Distribution-balanced focal loss. Distribution-balanced loss (Wu et al. 2020) is first proposed for multi-label classification. It consists of re-balanced weighting and negative-tolerant regularization. Re-balanced weighting assigns different weights to each label for each sample based on the re-sampling strategy. In ILDL, we modify the re-balancing weight to $r_i^j = P_C^j(\boldsymbol{x}_i)/P_I(\boldsymbol{x}_i)$, where $P_C^j(\boldsymbol{x}_i) = \frac{1}{c}\frac{1}{\hat{N}_j}$ is the expectation of label-level sampling frequency and $P_I(x_i) = \frac{1}{c}\sum_{j=1}^{c} d_{\boldsymbol{x}_i}^{y_j}/\hat{N}_j$ is the expectation of instance-level sampling frequency. We also use the smoothing function (Wu et al. 2020) to map r into a smoothed value $\hat{r}$. Negative-tolerant regularization (NTR) tries to deal with the issue that the gradients of the positive classes and the negative classes are significantly different. In ILDL, we modify the predicted value after negative-tolerant regularization to $q_{\boldsymbol{x}_i}^{y_j} = \exp\left(f_\theta^{(j)}(\boldsymbol{x}_i) - v_j\right)/\sum_{k=1}^{c} \exp\left(f_\theta^{(k)}(\boldsymbol{x}_i) - v_k\right)$ where ν is a class-specific bias. Combine re-balanced weighting with negative-tolerant regularization, we have the modified distribution-balanced focal loss:

$$\begin{aligned} \mathcal{L}_{OFR-DB} = \sum_{i=1}^{n} \sum_{j=1}^{c} \frac{1}{c} \Bigg[\hat{r}_i^j \left(1 - f_\theta^{(j)}(\boldsymbol{x}_i)\right)^{\gamma} \cdot \\ \left(\left(1 - \frac{1}{\lambda}\right) d_{\boldsymbol{x}_i}^{y_j} + \frac{1}{\lambda}\right) \ln \left(\frac{d_{\boldsymbol{x}_i}^{y_j}}{q_{\boldsymbol{x}_i}^{y_j}}\right) \Bigg]. \end{aligned} \tag{4}$$

where λ is a balanced hyperparameter.

Representation Distribution Alignment for ILDL

In the former subsection, we reshape several objective functions commonly used in imbalanced classification. These approaches, however, are all independently inside each label, and there is no insight into the distribution of the whole label set and no interaction among different labels. At the same time, these approaches only pay attention to the processing of the label space, and do not effectively utilize the information of the feature space to improve the performance of the predictive model. As a result, these methods still cannot fully avoid the impact of the representation distribution shift problem.

To tackle these issues, we propose *Representation Distribution Alignment* (RDA) for ILDL. RDA aligns the distributions of feature representations and label representations of instances to effectively leverage the information hidden in both feature space and label space and alleviate the impact of the distribution gap between the training set and the test set. Specifically, as shown in Figure 4, it first utilizes two mapping functions g_ϵ and h_ϕ to map the feature $\boldsymbol{x}$ and label distribution vector $\boldsymbol{d}$ into latent spaces. Then, it maps the latent vectors $g_\epsilon(\boldsymbol{x})$ and $h_\phi(\boldsymbol{d})$ into a common space and aligns the distributions of feature representations and label representations of instances in this space. Assuming the distributions of features and labels are Gaussian, i.e., there are two mapping functions $\mathcal{G}_\varepsilon$ and $\mathcal{H}_\varphi$ which maps $g_\epsilon(\boldsymbol{x})$ to $\mathcal{N}(\boldsymbol{\mu}_F, \boldsymbol{\sigma}_F^2)$ and maps $h_\phi(\boldsymbol{d})$ to $\mathcal{N}(\boldsymbol{\mu}_L, \boldsymbol{\sigma}_L^2)$, respectively. Then RDA minimizes the differences between $\mathcal{N}(\boldsymbol{\mu}_F, \boldsymbol{\sigma}_F^2)$ and $\mathcal{N}(\boldsymbol{\mu}_L, \boldsymbol{\sigma}_L^2)$. In order to better utilize the learned knowledge for description degree prediction, RDA also adopts a decoding function $\mathcal{F}_\vartheta$ to decode the feature encoding value to the description degree. Formally, RDA optimizes the following objective function:

$$\begin{aligned}\min_{\epsilon,\varepsilon,\phi,\varphi,\vartheta} & V(\epsilon,\vartheta)+\lambda_1\Omega_1(\phi,\varphi,\vartheta)\\ & +\lambda_2\Omega_2(\epsilon,\varepsilon)+\lambda_3\Omega_3(\epsilon,\varepsilon,\phi,\varphi),\end{aligned} \tag{5}$$

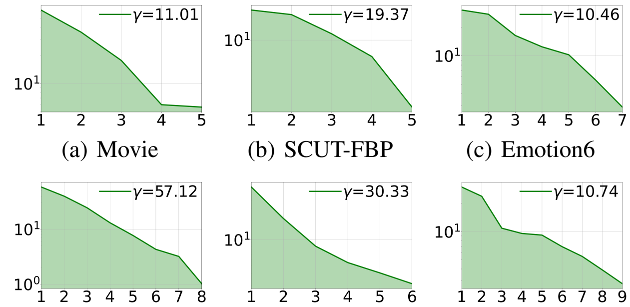
where $V(\epsilon,\vartheta)$ is the loss function for description degree prediction, Ω_1 and Ω_2 are used to learn reasonable label representation mapping and feature representation mapping, respectively, Ω_3 is used for aligning the distributions of feature representations and label representations, λ_1, λ_2 and λ_3 are balanced parameters.

The purpose of V is to reduce the divergence between the real distribution $\boldsymbol{d}$ and predicted distribution $\mathcal{F}_\vartheta(g_\epsilon(\boldsymbol{x}))$. Any reshaped objective function can be leveraged as the loss function V. The goal of Ω_1 is to make h_ϕ, $\mathcal{H}_\varphi$ and $\mathcal{F}_\vartheta$ have better label representation ability. To achieve that, we first sample values for label representations $\boldsymbol{r}_L$ using the reparameterization trick (Rezende, Mohamed, and Wierstra 2014): $\boldsymbol{r}_L=\boldsymbol{\mu}_L+\boldsymbol{\sigma}_L\boldsymbol{\delta}_L$, where $\boldsymbol{\mu}_L$ and $\boldsymbol{\sigma}_L$ are computed from $\mathcal{H}_\varphi(h_\phi(\boldsymbol{d}))$ and $\boldsymbol{\delta}_L\sim\mathcal{N}(0,\boldsymbol{I})$. Then we define Ω_1 as the following form:

$$\Omega_1(\phi,\varphi,\vartheta)=\sum_{i=1}^{n}\sum_{j=1}^{c}d_{\boldsymbol{x}_i}^{y_j}\ln\frac{d_{\boldsymbol{x}_i}^{y_j}}{\mathcal{F}_\vartheta(\boldsymbol{r}_L)}. \tag{6}$$

For Ω_2, we leverage the information of label space to promote the feature representation ability of the model. Specifically, we optimize the KL divergence between $\mathcal{N}(\boldsymbol{\mu}_F, \boldsymbol{\sigma}_F^2)$ and $\mathcal{N}(\boldsymbol{\mu}_L, \boldsymbol{\sigma}_L^2)$:

$$\Omega_2(\epsilon,\varepsilon)=-\frac{1}{2}\sum_{k=1}^{K}\left[\log\boldsymbol{\upsilon}^{(k)}-\boldsymbol{\upsilon}^{(k)}-\boldsymbol{\tau}^{(k)}+1\right], \tag{7}$$



(a) Movie (b) SCUT-FBP (c) Emotion6
(d) Flickr_LDL (e) RAF-ML (f) Natural Scene

Figure 5: Overview of the distribution of label annotation information in the training sets on six ILDL datasets. In each subfigure, the x-axis denotes the label ID (sorted by frequency), the y-axis denotes the sum description degree of each label, "γ" denotes the imbalance factor.

where K is the dimension of the latent space, $(\cdot)^{(k)}$ denotes the k-th element, $\boldsymbol{\upsilon}^{(k)}=\frac{\boldsymbol{\sigma}_F^{(k)2}}{\boldsymbol{\sigma}_L^{(k)2}}$, $\boldsymbol{\tau}^{(k)}=\frac{\left(\boldsymbol{\mu}_F^{(k)}-\boldsymbol{\mu}_L^{(k)}\right)^2}{\boldsymbol{\sigma}_L^{(k)2}}$, $\boldsymbol{\mu}_F$ and $\boldsymbol{\sigma}_F$ are computed from $\mathcal{G}_\varepsilon(g_\epsilon(\boldsymbol{x}))$.

The aim of Ω_3 is to align the distributions of feature representations and label representations. Specifically, we align the similarities of the distributions of both representations. For the feature representations r_F, we again use the reparameterization trick : $\boldsymbol{r}_F=\boldsymbol{\mu}_F+\boldsymbol{\sigma}_F\boldsymbol{\delta}_F$, where $\boldsymbol{\delta}_F\sim\mathcal{N}(0,\boldsymbol{I})$. Then the similarity matrix A of the distribution of feature representations can be obtained by:

$$A_{mn}=\mathcal{S}\left(\boldsymbol{r}_F^{(m)},\boldsymbol{r}_F^{(n)}\right), \tag{8}$$

where $\mathcal{S}$ is cosine similarity, $\boldsymbol{r}_F^{(m)}$ and $\boldsymbol{r}_F^{(n)}$ are the m-th and n-th instances. Meanwhile, the similarity matrix Z of the distribution of label representations can be obtained by:

$$Z_{mn}=\mathcal{S}\left(\boldsymbol{r}_L^{(m)},\boldsymbol{r}_L^{(n)}\right). \tag{9}$$

For Ω_3, the distance between A and Z is minimized:

$$\Omega_3(\epsilon,\varepsilon,\phi,\varphi)=\frac{1}{M^2}\sum_{m=1}^{M}\sum_{n=1}^{M}(A_{mn}-Z_{mn})^2, \tag{10}$$

In the training stage, the gradient-based method is used to optimize (5). In the prediction stage, given an instance $\boldsymbol{x}^*$, the prediction of RDA can be obtained by $\mathcal{F}_\vartheta(g_\epsilon(\boldsymbol{x}^*))$.

Experiments

Datasets

We curate six ILDL benchmarks that span movie rating, facial beauty perception and visual sentiment distribution perception. These datasets are sampled from six standard LDL datasets, including *Movie* (Geng 2016), *SCUT-FBP* (Xie et al. 2015), *Emotion6* (Peng et al. 2015), *Flickr_LDL* (Yang, Sun, and Sun 2017), *RAF-ML* (Shang and Deng 2019) and *Natural Scene* (Geng 2016). The sampling process is performed 10 times, for each time, we sample the training set

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±0.0101•	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
RDA (Ours)	**0.1962±0.0068**	**0.2849±0.0157**	**0.3598±0.0079**	**0.5208±0.0075**	**0.3756±0.0068**	**0.3768±0.0208**

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
RDA (Ours)	**0.2491±0.0149**	**0.4313±0.0328**	**0.7677±0.0218**	**1.6071±0.1107**	**0.7058±0.0203**	**1.1188±0.0591**

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

and validation set from the original training set which occupies 90% of the examples, while the test set remains unchanged. Figure 5 illustrates the distribution of label annotation information of these datasets and their level of imbalance.

Evaluation Criteria

Six standard LDL measures (Chebyshev Distance, Clark Distance, Canberra Metric, Kullback-Leibler Divergence, Cosine Coefficient, and Intersection Similarity between ground-truth label distributions and predicted label distributions) are selected to evaluate different methods for the prediction of label distributions. Besides, Euclidean Distance is also adopted to evaluate the performance of different methods on *tail*, *head* and *all* labels. It is worth noting that the evaluation criteria of ILDL are different from imbalanced classification and regression problems. In ILDL, the description degrees of the *tail* labels in the training set tend to increase in the test set, while the description degrees of the *head* labels tend to decrease. Therefore, *head* labels and *tail* labels are equally important in ILDL.

Methodology

Several existing state-of-the-art LDL algorithms, i.e., SA-BFGS (Geng 2016), EDL-LRL (Jia et al. 2019), LDLSF (Ren et al. 2019a), LDL-LCLR (Ren et al. 2019b), Adam-LDL-SCL (Jia et al. 2021) and LDL-LDM (Wang and Geng 2021) are set as baselines. Three objective function reshaping approaches, i.e., OFR-FL, OFR-CB and OFR-DB, are also performed in the experiments. Moreover, Our RDA is compared with these methods. In RDA, g_ϵ, h_ϕ and $\mathcal{F}_\vartheta$ are set as linear projections, $\mathcal{G}_\varepsilon$ and $\mathcal{H}_\varphi$ are set as single-layer neural network with two outputs including mean and variance of Gaussian, and the modified distribution-balanced focal loss is adopted as the loss function V. Hyperparameters λ_1, λ_2 and λ_3 are selected by grid search from the set $\{0.01, 0.05, 0.1, 0.2, 0.5\}$.

Main Results

Comparisons on Distribution Criteria Tables 1 and 2 tabulate the experimental results of different methods on Chebyshev Distance and Kullback-Leibler Divergence. For each evaluation criterion, "↓" indicates the smaller the better. In Tables 1 and 2, the two-tailed t-test at 0.05 significance level is conducted, and the best performances are highlighted in bold. •/◦ indicates whether RDA is statistically superior/inferior to the comparing methods. From Tables 1 and 2, it can be observed that: 1) The existing LDL methods show poor performances in solving ILDL tasks. 2) The performances of the objective function reshaping approaches are superior to existing methods. 3) Compared with these baseline methods, Our RDA has achieved better results on Chebyshev Distance and Kullback-Leibler Divergence and significantly outperforms other algorithms in most cases. These observations indicate that the RDA can effectively alleviate the impact of the distribution gap between the training set and the test set.

Comparisons on Imbalance Criteria Table 3 gives the experimental evaluations of different algorithms on *tail*, *head* and *all* labels. From Table 3, we can find that most objective function reshaping approaches have better performance than existing LDL methods. In the meantime, RDA achieves the best performance in almost all cases. In particular, RDA achieved the best performance on *head* and *all* labels in all

Algorithm	Movie			SCUT-FBP			Emotion6			Flickr_LDL			RAF-ML			Natural 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	.479	.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
RDA (Ours)	**.295**	**.245**	**.157**	**.386**	**.299**	**.234**	**.464**	**.429**	**.133**	**.645**	.526	**.333**	**.484**	.454	**.132**	**.515**	**.486**	**.151**

Table 3: Experimental results on ILDL datasets measured by Euclidean Distance ↓.

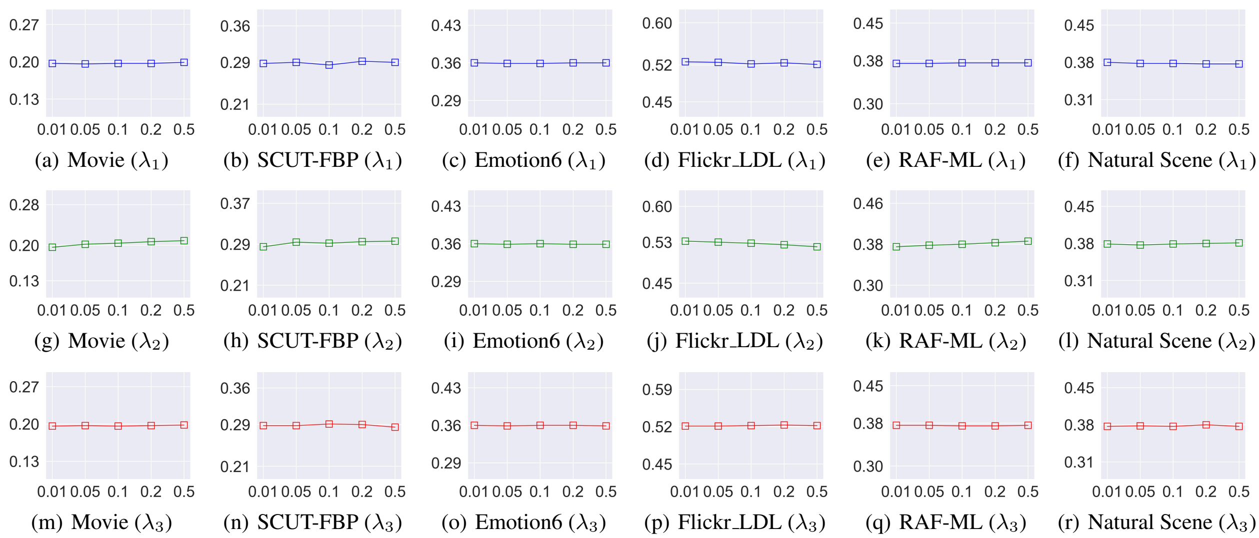


Figure 6: Effects of λ_1, λ_2 and λ_3 on Chebyshev Distance ↓.

cases. These observations demonstrate the effectiveness of the proposed methods in tackling ILDL tasks.

Effect of Hyperparameters λ_1, λ_2 and λ_3

In this subsection, we explore the effect of hyperparameters λ_1, λ_2 and λ_3. We compare the performances of RDA with different values of λ_1, λ_2 and λ_3 on the six datasets measured by Chebyshev Distance. Figure 6 illustrates the performances of RDA with different values of λ_1, λ_2 and λ_3. From these curves, we can find that: 1) Overall, RDA has stable performances with a wide range of hyperparameter values on all six datasets; 2) Appropriate values of λ_2 can bring slight performance gains on some datasets; 3) The performance of the model hardly changes with changes in λ_1 and λ_3. These findings further demonstrate the robustness of the proposed RDA.

Further Analyses

We compare the average ranks of different algorithms over all the six ILDL datasets and find that RDA surpasses the compared methods by a significant margin across all the evaluation criteria, which further indicates the effectiveness of the proposed RDA. Details of the further analyses are provided in the appendix, which is available at: https://github.com/ailearn-ml/RDA.

Conclusion

We study a challenging and meaningful problem, i.e., Imbalanced Label Distribution Learning (ILDL), in this paper. We curate several benchmark ILDL datasets and offer three strong baselines. Moreover, we delve into the characteristics of the ILDL problem and find that the representation distribution shift is the underlying reason for the performance degradation of existing methods. Based on this finding, we propose a novel method named Representation Distribution Alignment, which can align the distributions of feature representations and label representations to effectively alleviate the impact of the distribution gap between the training set and the test set caused by the imbalance issue. Extensive experiments confirm the superior performance of our proposed method. Our work fills the gap in benchmarks and techniques for practical ILDL problems.

Acknowledgements

This research was supported by the National Key Research & Development Plan of China (No. 2018AAA0100104), the National Science Foundation of China (62125602, 62076063, 62206050), China Postdoctoral Science Foundation (2021M700023, 2022M720028), Jiangsu Province Science Foundation for Youths (BK20210220), Young Elite Scientists Sponsorship Program of Jiangsu Association for Science and Technology (TJ-2022-078).

References

Berger, A. L.; Pietra, S. D.; and Pietra, V. J. D. 1996. A Maximum Entropy Approach to Natural Language Processing. *Comput. Linguistics*, 22(1): 39–71.

Branco, P.; Torgo, L.; and Ribeiro, R. P. 2017. SMOGN: a Pre-processing Approach for Imbalanced Regression. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 36–50.

Branco, P.; Torgo, L.; and Ribeiro, R. P. 2018. REBAGG: REsampled BAGGing for Imbalanced Regression. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 67–81.

Buda, M.; Maki, A.; and Mazurowski, M. A. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106: 249–259.

Byrd, J.; and Lipton, Z. C. 2019. What is the Effect of Importance Weighting in Deep Learning? In *Proceedings of the 36th International Conference on Machine Learning*, 872–881.

Cao, K.; Wei, C.; Gaidon, A.; Aréchiga, N.; and Ma, T. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *Advances in Neural Information Processing Systems 32*, 1565–1576.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.*, 16: 321–357.

Cui, Y.; Jia, M.; Lin, T.; Song, Y.; and Belongie, S. J. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9268–9277.

Gao, B.; Xing, C.; Xie, C.; Wu, J.; and Geng, X. 2017. Deep Label Distribution Learning With Label Ambiguity. *IEEE Trans. Image Process.*, 2825–2838.

Gao, B.; Zhou, H.; Wu, J.; and Geng, X. 2018. Age Estimation Using Expectation of Label Distribution Learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 712–718.

Geng, X. 2016. Label Distribution Learning. *IEEE Trans. Knowl. Data Eng.*, 28(7): 1734–1748.

Geng, X.; and Xia, Y. 2014. Head Pose Estimation Based on Multivariate Label Distribution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1837–1842.

Geng, X.; Yin, C.; and Zhou, Z. 2013. Facial Age Estimation by Learning from Label Distributions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(10): 2401–2412.

He, H.; and Garcia, E. A. 2009. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.*, 21(9): 1263–1284.

Huo, Z.; and Geng, X. 2017. Ordinal Zero-Shot Learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1916–1922.

Jia, X.; Li, Z.; Zheng, X.; Li, W.; and Huang, S. 2021. Label Distribution Learning with Label Correlations on Local Samples. *IEEE Trans. Knowl. Data Eng.*, 33(4): 1619–1631.

Jia, X.; Zheng, X.; Li, W.; Zhang, C.; and Li, Z. 2019. Facial Emotion Distribution Learning by Exploiting Low-Rank Label Correlations Locally. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9841–9850.

Lin, T.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2020. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2): 318–327.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-Scale Long-Tailed Recognition in an Open World. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2537–2546.

Peng, K.; Chen, T.; Sadovnik, A.; and Gallagher, A. C. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 860–868.

Ren, T.; Jia, X.; Li, W.; Chen, L.; and Li, Z. 2019a. Label distribution learning with label-specific features. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 3318–3324.

Ren, T.; Jia, X.; Li, W.; and Zhao, S. 2019b. Label Distribution Learning with Label Correlations via Low-Rank Approximation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 3325–3331.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31st International Conference on Machine Learning*, 1278–1286.

Shang, L.; and Deng, W. 2019. Blended Emotion in-the-Wild: Multi-label Facial Expression Recognition Using Crowdsourced Annotations and Deep Locality Feature Learning. *Int. J. Comput. Vis.*, 127(6-7): 884–906.

Su, K.; and Geng, X. 2019. Soft Facial Landmark Detection by Label Distribution Learning. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 5008–5015.

Torgo, L.; Ribeiro, R. P.; Pfahringer, B.; and Branco, P. 2013. SMOTE for Regression. In *Proceedings of the 16th Portuguese Conference on Artificial Intelligence*, 378–389.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, J.; and Geng, X. 2021. Label distribution learning by exploiting label distribution manifold. *IEEE transactions on neural networks and learning systems*.

Wang, J.; Zhang, F.; Jia, X.; Wang, X.; Zhang, H.; Ying, S.; Wang, Q.; Shi, J.; and Shen, D. 2022. Multi-Class ASD Classification via Label Distribution Learning with Class-Shared and Class-Specific Decomposition. *Medical Image Anal.*, 102294.

Wu, T.; Huang, Q.; Liu, Z.; Wang, Y.; and Lin, D. 2020. Distribution-Balanced Loss for Multi-label Classification in Long-Tailed Datasets. In *Proceedings of the European Conference on Computer Vision*, 162–178.

Xie, D.; Liang, L.; Jin, L.; Xu, J.; and Li, M. 2015. SCUT-FBP: A Benchmark Dataset for Facial Beauty Perception. In *Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics*, 1821–1826.

Yang, J.; Li, J.; Li, L.; Wang, X.; and Gao, X. 2021a. A Circular-Structured Representation for Visual Emotion Distribution Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4237–4246.

Yang, J.; Sun, M.; and Sun, X. 2017. Learning Visual Sentiment Distributions via Augmented Conditional Probability Neural Network. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 224–230.

Yang, Y.; Zha, K.; Chen, Y.; Wang, H.; and Katabi, D. 2021b. Delving into Deep Imbalanced Regression. In *Proceedings of the 38th International Conference on Machine Learning*, 11842–11851.

Zhao, P.; and Zhou, Z. 2018. Label Distribution Learning by Optimal Transport. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 4506–4513.

Zheng, X.; Jia, X.; and Li, W. 2018. Label Distribution Learning by Exploiting Sample Correlations Locally. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 4556–4563.

Zhou, B.; Cui, Q.; Wei, X.; and Chen, Z. 2020. BBN: Bilateral-Branch Network With Cumulative Learning for Long-Tailed Visual Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9716–9725.