IRDA: Implicit Data Augmentation for Deep Imbalanced Regression

Weiyao Zhu[a], Ou Wu[a,*], Nan Yang[a]

[a]*Center of Mathematics, Tianjin University, Tianjin, China, Tianjin, 300072, China*

Abstract

Imbalanced data distributions are prevalent in real-world classification and regression tasks. Data augmentation is a commonly employed technique to mitigate this issue, with implicit methods gaining attention for their effectiveness and efficiency. However, implicit data augmentation methods have not been extensively explored in the context of regression tasks. To address this gap, we introduce IRDA, a novel learning method for regression that incorporates implicit data augmentation. Our approach includes developing a new augmentation strategy specifically tailored for deep imbalanced regression tasks, and a regression loss function that is suitable for real-world data with imbalanced label distributions and non-uniformly distributed features. We derive an easily computable surrogate loss and propose two implicit data augmentation algorithms, one incorporating meta-learning and one without. Additionally, we provide a regularization perspective to offer a deeper understanding of IRDA. We evaluate IRDA on five datasets, including a large-scale dataset, demonstrating its effectiveness in mitigating the adverse effects of imbalanced data distribution and its adaptability to various regression tasks.

Keywords: Deep imbalanced regression, Implicit data augmentation, Regularization, Regression loss

*Corresponding author.

Email addresses: weiyaozhu@tju.edu.cn (Weiyao Zhu), wuou@tju.edu.cn (Ou Wu), yny@tju.edu.cn (Nan Yang)

1. Introduction

Imbalanced data distribution is a common challenge in real-world machine learning applications, adversely affecting model generalizability [1]. This issue has recently garnered significant attention in deep regression tasks, as most real-world datasets used for deep regression are affected by imbalance. Several techniques have been proposed to address this issue.

In classification tasks, the issue of data imbalance has been extensively discussed and investigated. Various solutions have been proposed, including re-sampling [2], re-weighting [3], adjusting the loss function [4, 5], modifying the network [6, 7], and data augmentation [8, 9]. In this work, we specifically focus on data augmentation methods. Explicit data augmentation techniques are widely used to generate new samples for underrepresented or tail labels, i.e., targets with fewer observations, through methods such as cropping, rotation, and generative adversarial networks. The word "explicit" indicates that the samples are actually generated. Methods such as BalanceMix [9] and LeGAN [8] have been investigated to expand the dataset size, particularly for tail labels. A recent notable work by Ren et al. [10] proposes using generative adversarial networks to augment data and address the non-identical distributions between training and test sets. These explicit data augmentation methods have demonstrated significant effectiveness in handling imbalanced issues. However, as the number of augmented samples increases, the time required for augmentation also grows.

Previous work [11] has noted that allowing semantic transformations, such as changing the textures of an object, can enhance the effectiveness of data augmentation in improving generalizability. Traditional explicit methods, which apply transformations without considering semantic directions, are limited in their ability to improve generalizability, thereby reducing their effectiveness. The implicit data augmentation strategy emerges as a solution to the problem of time consumption, significantly enhancing the effectiveness of generalizability improvement. The term "implicit" indicates that the samples are virtually generated.

Wang et al. [12] introduced the first implicit data augmentation method ISDA, which implicitly generates samples along semantic directions by virtually sampling infinite samples from a constructed distribution. ISDA then optimizes the upper bound of the loss for these infinite generated samples. This method not only reduces time consumption but also proves to be more effective in improving generalizability than explicit augmentations. The co-

variance matrix is crucial for implicit data augmentation, as it stores semantic information. Recently, Li et al. [13] demonstrated that estimating a diversified covariance matrix using training samples is a challenging task for tail labels. To address this, Chen et al. [14] introduce semantic knowledge of other labels to enhance the diversity of the estimated covariance matrix, proposing a modified version of ISDA called RISDA, which has been shown to be more effective on long-tailed distribution data. Implicit data augmentations are also widely discussed and used for data imbalance in topics such as face recognition, active learning, and vehicle re-identification [15, 16, 17, 18], and are recognized as essential methods for image data augmentation [19].

Data imbalance is also a common issue in regression tasks and has received extensive attention in research over the years. To address this problem, the re-sampling strategy has been widely adopted. Torgo et al. [20] were the first to propose an adaptive re-sampling method for imbalanced regression called SMOTER, which over-samples instances from tail labels. Subsequently, Branco et al. [21] proposed SMOGN, which not only over-samples rare samples, but also adds Gaussian noise to frequent samples. Recently, with the rise of deep learning, Yang et al. [22] introduced a pioneering concept called Deep Imbalanced Regression (DIR), specifically designed to handle real-world tasks such as age inference and temperature prediction. Compared to regression tasks in shallow learning, DIR involves larger datasets and requires longer training times. Yang et al. proposed two loss adjustment solutions that introduce distribution smoothing strategies into DIR tasks. These strategies assign larger weights to tail labels and enhance the features of tail labels using the features of neighboring labels. Ren et al. [23] recently introduced a novel loss adjustment method, Balanced-MSE (BMSE), to reduce the bias caused by MSE, which is especially severe in DIR tasks.

In a recent work, Wu [24] highlighted the existence of intra-label imbalance in real-world data, describing the skewed distributions of features. This intra-label imbalance undermines the commonly adopted label-conditional invariance assumption, which presumes that the feature distribution is uniform across different labels. This assumption is presumed to hold during the inference of BMSE. Therefore, BMSE is inappropriate when the assumption of label-conditional invariance is violated. Explicit augmentation methods, such as RegMix [25] have been proposed to generate new samples by combining original samples. Although there are existing solutions for DIR tasks, they still have certain inadequacies. Explicit data augmentation methods

are time-consuming and inefficient. Despite the proven effectiveness and efficiency of implicit data augmentation in addressing imbalanced classification, no significant progress has been made in applying this approach to DIR tasks.

The contributions of our work are summarized as follows:

- We introduce the concept of implicit data augmentation, previously shown to be effective in classification tasks, to the domain of deep regression. To our knowledge, this is the first framework to apply implicit data augmentation to enhance regression tasks.
- We provide an in-depth theoretical analysis of related works from a regularization perspective, offering vivid geometric visualizations to compare different methods. A comprehensive comparison between our proposed method and several closely related methods is provided to further enhance the understanding of our approach.
- We propose two implicit learning algorithms, IRDA and Meta-IRDA. Our methods are compared with nine state-of-the-art techniques. Extensive experiments were conducted on five widely used benchmarks, including one large-scale dataset, and the results demonstrate the efficacy of our approach. Additionally, in-depth ablation studies and comprehensive sensitivity tests were performed.

2. Related Work

2.1. Imbalanced regression

Imbalanced regression has been a subject of investigation for years, though the concept of Deep Imbalanced Regression (DIR) has been introduced only recently, with limited related works. Various methods have been proposed to address imbalanced regression in both deep learning and shallow learning contexts.

Initially, data imbalance in regression was mitigated using re-sampling strategies. Torgo et al. [20] modified the SMOTE [2] method, originally designed for classification tasks, into a regression version called SMOTER, which over-samples rare samples and under-samples frequent samples. However, a common shortcoming of re-sampling strategies is their limited ability to enhance generalizability due to the insufficient semantic information contained in samples from tail labels.

Table 1: Summary of the specific concepts and abbreviations.

Concepts/Abbreviations	Meanings
MSE	Mean squared error.
DIR	Deep imbalanced regression.
LDS	Label distribution smoothing.
FDS	Feature distribution smoothing.
BMSE	Balanced MSE.
LA	Logit adjustment.
KDE	Kernel distribution smoothing.
NLL	Negative-log likelihood.
W-FDS	Weighted-feature distribution smoothing.
CE	Cross-entropy.
LCDI	Label-conditional distribution invariance.
PLCDI	Partition-projected label-conditional distribution invariance.
Tail label	Target with significant fewer observations.
Head label	Target with significant larger observations.
Explicit augmentation	Truly generating samples.
Implicit augmentation	Virtually generating samples.
Semantics	Information carried by features such as color, shape, and so on.
Skewness	The imbalanced ratio.
Compactness	The radius of the scatter of samples.

Recently, there has been increased focus on deep regression tasks with imbalanced distributions, primarily addressed through loss adjustment strategies. For example, techniques such as Logit Adjustment (LA)[26], initially developed for classification tasks, have been adapted into a regression context by Ren et al., resulting in BMSE [23]. This modification reformulates the commonly used MSE loss to emphasize samples from tail labels, thereby improving the model's ability to learn from them effectively. Furthermore, Yang et al.[22] introduced the distribution smoothing technique into DIR tasks, proposing two frameworks: Label Distribution Smoothing (LDS) and Feature Distribution Smoothing (FDS). LDS is a re-weighting method that smooths the label distribution using Kernel Density Estimation (KDE)[27] and reversely applies the smoothed distribution as a continuous weighting function to favor tail labels. Recognizing the similarity between neighboring labels in regression tasks, FDS leverages this similarity to compensate for the lack of information in tail label features. Similarly, RankSim [28] by Gong et al. and SupCR [29] by Zha et al. adjust the loss function based on feature similarity. Dubost et al. [30] introduced a novel regularized neural network

regressor called Hydranet, which regularizes the deep regressor according to an estimated minimum number for each label.

Additionally, explicit data augmentation has been extensively studied in imbalanced regression. Branco et al. proposed SMOGN [21], which builds on SMOTER by over-sampling rare samples and generating new frequent samples by adding Gaussian noise. Hwang et al.[25] introduced a data augmentation meta-learning framework named Mixrl for regression. Stocksieker et al.[31] developed DA-WR, combining traditional explicit data augmentations with a re-weighting strategy to enhance model performance.

2.2. Data augmentation

Data augmentation is one of the most investigated strategies in machine learning. It includes explicit augmentation methods such as BalanceMix [9] and LeGAN [8], as well as implicit augmentation methods like RISDA [14] and MetaSAug [13].

Typical explicit data augmentations such as BalanceMix apply linear combinations to generate new training samples. Ren et al.[10] proposed generating samples using a generative adversarial network and the Gaussian mixture model to reduce distribution differences between training and test sets. However, explicit augmentations often face limitations in improving generalizability as they randomly augment samples without considering semantic directions, which can make the augmented samples less valuable or even harmful to generalizability[14].

To address this, Chen et al.[32] introduced semantic transformations into one-shot learning with TriNet, generating samples along the co-linear semantic directions of a novel sample and existing training samples. Despite their benefits, explicit data augmentations that generate samples to enlarge the dataset are time-consuming and less efficient for deep learning.

Implicit data augmentation, on the other hand, augments towards semantic directions without generating actual samples. This concept was initially proposed by Li et al.[12] with ISDA, which extracts semantics from each label and augments original samples towards these semantic directions. Semantic information signifies features such as “red color”, “tail”, or “glasses”. Chen et al.[14] recognized the potential of implicit data augmentation for imbalanced classification and identified that an underestimation of the covariance matrix could lead to inaccurate semantic directions, resulting in insufficiently augmented samples. They proposed RISDA, which selects semantics from both the true label and easily confused labels for augmentation directions.

Li et al.[13] introduced MetaSAug, incorporating implicit augmentation into meta-learning to extract the most significant semantic information for optimal generalizability on validation sets.

Implicit data augmentations are also widely applied in various related topics. Low and Teoh [15] introduced implicit data augmentations into low-resolution face recognition tasks by training an IDEA-Net instance to augment the small-scale low-resolution face dataset implicitly in the representation space. Chen et al.[16] used ISDA in active learning with a diversity-aware semantic transformation framework named DAST-AL, which considers ISDA's effect in selecting unlabeled samples. Li et al.[17] applied ISDA to vehicle re-identification tasks, proposing bi-level implicit semantic data augmentation based on identity-level and superclass-level intra-class variations to generate more diverse semantic augmentations. Seo et al.[18] utilized implicit data augmentation in hand pose estimation, introducing metric learning and hand-dependent augmentation techniques. Yang et al.[19] provided a comprehensive survey on image data augmentations, discussing implicit data augmentation methods.

3. Methodology

This section begins by defining the problem setting, followed by a detailed description of the proposed method, IRDA. We then outline the two main components of IRDA: the augmentation strategy and the proposed surrogate loss, Rebalanced MSE. Furthermore, we provide a theoretical investigation of methods addressing the imbalance issue from a regularization perspective. Finally, we compare several closely related methods for addressing imbalanced data.

3.1. Problem Setting

Let $\boldsymbol{x}$ and y be the variable of the input and the target respectively. Denote the training set as $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, where $\boldsymbol{x}_i$ and y_i are realizations of $\boldsymbol{x}$ and y. In regression setting, the target y is continuous.

One of the most common methods for learning continuous targets in deep learning is to discretize the continuous target $\mathcal{Y}$ into B discrete bins. Following the approach in related works [22], we adopt the general settings for bin length in imbalanced regression, which represents the smallest granularity of the continuous labels. For example, the continuous target $\mathcal{Y}$ in the AgeDB-DIR dataset signifies human age, ranging from 0 to 101 years, with

a bin length of 1 year. Another dataset, NYUD2-DIR, contains images and depth maps for different indoor scenes, with the continuous target $\mathcal{Y}$ ranging from 0.7 to 10 meters, and a bin length of 0.1 meters. Although different datasets follow the same binning strategy, the bin length varies according to the specific domain and range of the targets.

Accordingly, we use the same setting of [22], where $\mathcal{Y}$ is divided into B groups $[y_0, y_1), ..., [y_b, y_{b+1}), ..., [y_{B-1}, y_B)$, and $b \in \mathcal{B} = \{0, ..., B-1\}$ signifies the the group index of the target value. The bin length is set as the smallest granularity of each dataset. $\boldsymbol{z}_i = f(\boldsymbol{x}_i, \theta)$ denotes the deep feature of $\boldsymbol{x}_i$ where $f(\boldsymbol{x}_i, \theta)$ is parameterized by a deep neural network model with parameter θ and $y_i \in [y_b, y_{b+1})$. The prediction $\hat{y}(\boldsymbol{x}_i)$ for $\boldsymbol{x}_i$ is given by a regressor $g(\cdot)$ that operates over $\boldsymbol{z}_i$, i.e., $\hat{y}(\boldsymbol{x}_i) = g(\boldsymbol{z}_i) = \boldsymbol{w}^\top \boldsymbol{z}_i + w_0$ with regressor's weight and bias $\boldsymbol{w}$ and w_0.

The most commonly used MSE loss can be written as

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}(\boldsymbol{x}_i))^2. \tag{1}$$

As is demonstrated in [33], minimizing MSE is equivalent to maximum likelihood estimation in regression with an underlying Gaussian error model [33]. Therefore, MSE loss equals to the Negative Log-Likelihood (NLL) loss of the prediction distribution $p_{train}(y|\boldsymbol{x}, \theta)$, i.e.,

$$\mathcal{L}_{NLL} = -\log p_{train}(y|\boldsymbol{x}, \theta). \tag{2}$$

In the classic probabilistic interpretation [34], the prediction distribution is considered as a noisy Gaussian distribution with the mean of $\hat{y}$, i.e., $p_{train}(y|\boldsymbol{x}, \theta) = \mathcal{N}(y; \hat{y}, \sigma^2_{noise})$ with σ_{noise} the scale of an i.i.d. noisy term $\epsilon \sim \mathcal{N}(0, \sigma^2_{noise})$.

3.2. The Proposed Method IRDA

The overall process of our proposed method, IRDA, is illustrated in Fig.1. IRDA consists of two primary components: a specifically designed augmentation strategy and a surrogate loss.

Given a long-tailed dataset used in DIR tasks, a feature subnet is employed to extract the deep feature $\boldsymbol{z}_i$ of each sample $(\boldsymbol{x}_i, y_i)$. The extracted features are clustered into k clusters using k-means. Sample numbers in each cluster are counted to obtain the posterior distribution. After extracting all

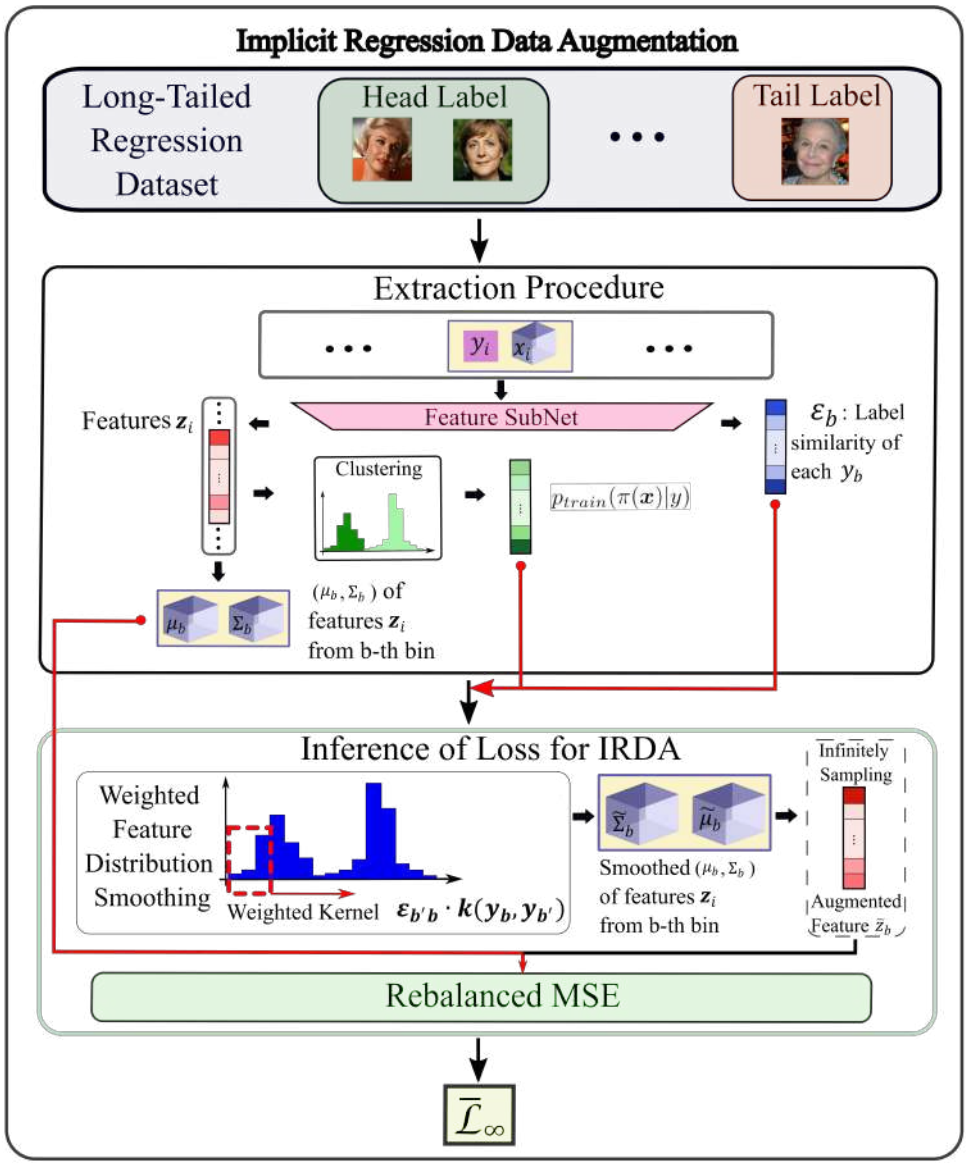


Figure 1: Overview of our proposed IRDA scheme. Features of each sample are extracted using a subnet. Label similarities and clustering distributions are calculated. The mean and covariance matrix are estimated and then smoothed using the given weights. Samples are augmented by infinitely sampling from the distribution defined by the smoothed mean and covariance matrix. These augmented samples are fed into the proposed loss function, Rebalanced-MSE, and an upper bound of the loss is calculated and used for optimization.

features, the mean $\boldsymbol{\mu}_b$ and the covariance matrix $\boldsymbol{\Sigma}_b$ of the features from the b-th bin are estimated. Additionally, the label similarities between the true target and other targets are estimated based on the probability of mislabeling. After estimation, $\boldsymbol{\mu}_b$ and $\boldsymbol{\Sigma}_b$ undergo a Weighted Feature Distribution Smoothing (W-FDS) process, where the mean and covariance are influenced by neighboring labels. The calculated label similarities are used as weights during the W-FDS process, ensuring that neighboring labels with greater similarity have a greater influence on the estimated mean and covariance of the target label.

Based on the smoothed feature statistics of the b-th bin, we simulate a Gaussian distribution of features. Assuming an infinite number of samples are drawn from the simulated distribution, we calculate the overall loss of infinite samples by deducing an upper bound, i.e., $\overline{\mathcal{L}}_\infty$.During this calculation, we assign weights to samples from each target, using the proportions of each label inversely as weights. This inferred loss $\overline{\mathcal{L}}_\infty$ is then used as the optimization objective function of IRDA.

3.3. Augmentation Strategy

Several implicit data augmentation methods have been proposed for classification tasks, such as ISDA and RISDA. ISDA randomly samples from $\mathcal{N}(0, \beta\boldsymbol{\Sigma}_b)$ to generate features with different semantic transformations according to their true targets, i.e.,

$$\tilde{\boldsymbol{z}}_i \sim \mathcal{N}(\boldsymbol{z}_i, \beta\boldsymbol{\Sigma}_b), \tag{3}$$

where β is a positive coefficient and $\boldsymbol{z}_i = f(\boldsymbol{x}_i, \theta)$ is the feature of sample $\boldsymbol{x}_i$ belonging to the category y_b. RISDA, on the other hand, enriches the covariance matrix of the true label by introducing the covariance matrix of similar labels, and then samples from this enriched covariance matrix.

Under the setting of regression, the statistics for features of samples belonging to the b-th bin $[y_b, y_{b+1})$, is denoted as $\{\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b\}$, where

$$\begin{aligned} \boldsymbol{\mu}_b &= \frac{1}{N_b}\sum_{b=1}^{B}\boldsymbol{z}_i, \\ \boldsymbol{\Sigma}_b &= \frac{1}{N_b - 1}\sum_{b=1}^{B}(\boldsymbol{z}_i - \boldsymbol{\mu}_b)(\boldsymbol{z}_i - \boldsymbol{\mu}_b)^T \end{aligned} \tag{4}$$

are the mean and the covariance matrix of features respectively. N_b is the number of samples in the b-th bin. As stated in previous work [13], the

statistics of features estimated using training samples are often underestimated and lack diversity, especially for tail labels. Therefore, we aim to enrich the statistics of tail labels using the following approach,

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_b &= \sum_{b' \in \mathcal{B}} s(y_b, y_{b'}) \boldsymbol{\mu}_{b'}, \\ \tilde{\boldsymbol{\Sigma}}_b &= \sum_{b' \in \mathcal{B}} s(y_b, y_{b'}) \boldsymbol{\Sigma}_{b'} \end{aligned} \tag{5}$$

where $s(y_b, y_{b'})$ a function over all target bins. Motivated by RISDA, we enrich the statistics of tail labels using easily confused labels. Previous work [22] also revealed that the features of tail labels can be effectively enhanced by neighboring labels. Based on these considerations, the function $s(y_b, y_{b'})$ is given by the following form:

$$s(y_b, y_{b'}) = \varepsilon_{bb'} k\left(y_b, y_{b'}\right),$$

with $k\left(y_b, y_{b'}\right)$ a symmetric kernel and

$$\varepsilon_{bb'} = \sum_{\boldsymbol{x}_i \text{ s.t. } y_i \in [y_b, y_{b+1})} \frac{\mathbb{I}\left(y_{b'} \leq \hat{y}(\boldsymbol{x}_i) < y_{b'+1}\right)}{N_b}. \tag{6}$$

Accordingly, $\mathbb{I}(\cdot)$ is an indicator function. $\varepsilon_{bb'}$ is the element in the b-th line, b'-th column of the matrix $\boldsymbol{\varepsilon}$ and denotes the similarity between the b-th bin and the b'-th bin.

This process degenerates to the FDS process when the matrix $\boldsymbol{\varepsilon}$ is an identity matrix. Accordingly, we name this process as Weighted-FDS (W-FDS) process.

For a sample $(\boldsymbol{x}_i, y_i)$ from the b-th bin with $\boldsymbol{z}_i = f(\boldsymbol{x}_i, \theta)$, the augmentation strategy of IRDA proposes to sample along $\mathcal{N}(\alpha\tilde{\mu}_b, \beta\tilde{\Sigma}_b)$ to generate features, i.e.,

$$\tilde{\boldsymbol{z}}_i \sim \mathcal{N}(\boldsymbol{z}_i + \alpha\tilde{\mu}_b, \beta\tilde{\Sigma}_b). \tag{7}$$

Assuming the sample $(\boldsymbol{x}_i, y_i)$ is augmented M times, the resulting dataset is $\{(\boldsymbol{z}_i^1, y_i), \cdots, (\boldsymbol{z}_i^m, y_i), \cdots, (\boldsymbol{z}_i^M, y_i)\}_{i=1}^{N_b}$, where $\boldsymbol{z}_i^m$ sampled from $\tilde{\boldsymbol{z}}_i$ is the m-th augmented feature of $\boldsymbol{z}_i$.

This augmentation strategy, which perturbs the features' distribution along $\mathcal{N}(\alpha\tilde{\mu}_b, \beta\tilde{\Sigma}_b)$, differs from the typical Gaussian noise addition. Gaussian noise lacks semantic information, while our implicit augmentation module perturbs both the mean and variance towards specific semantic directions.

3.4. The Inference of Surrogate Loss

As mentioned earlier, implicit data augmentation was initially proposed for classification tasks. A crucial step in this approach estimating the upper bound of the total loss of M augmented samples as M tends to infinity. This estimation depends on the original choice of loss function, such as the commonly used Cross-Entropy (CE) loss for classification tasks. However, the CE loss cannot be directly applied to DIR and is incompatible with regression tasks. Additionally, existing loss functions may not be well-suited for imbalanced distributions or other real-world cases. In this regard, we introduce a new loss function for deep regression tasks, and then derive a surrogate loss that can be used in IRDA.

3.4.1. The introduction of Rebalanced-MSE

The MSE loss, which is widely used for regression tasks, has been found to be insufficient for DIR tasks dealing with imbalanced distributions in real-world datasets. It has been revealed that due to the disparate distributions between the training set and the test set, the MSE loss leads to low generalizability on the test set.

BMSE has recently been proposed as a surrogate loss for DIR tasks, replacing the typical MSE loss. First, we briefly introduce the main idea of BMSE. Considering the disparate distributions of the skewed training set and the balanced test set,

$$p_{train}(y) \neq p_{valid}(y). \tag{8}$$

BMSE aims to estimate $p_{valid}(y|\boldsymbol{x})$ accurately instead of training on $p_{train}(y|\boldsymbol{x})$ as in Eq.(2), because $p_{train}(y|\boldsymbol{x})$ is also skewed. By Bayes' Rule, following relations hold,

$$\begin{aligned} p_{train}(y|\boldsymbol{x}) &\propto p_{train}(\boldsymbol{x}|y)p_{train}(y) \\ p_{valid}(y|\boldsymbol{x}) &\propto p_{valid}(\boldsymbol{x}|y)p_{valid}(y) \end{aligned} \tag{9}$$

Eq. (10) infers to the following relation between skewed $p_{train}(y|\boldsymbol{x})$ and balanced $p_{valid}(y|\boldsymbol{x})$,

$$\frac{p_{train}(y|\boldsymbol{x})}{p_{valid}(y|\boldsymbol{x})} \propto \frac{p_{train}(y)}{p_{valid}(y)} \tag{10}$$

based on the following assumption.

Assumption 1. *The label-conditional distribution invariance (LCDI) holds if*

$$p_{train}(\boldsymbol{x}|y) = p_{valid}(\boldsymbol{x}|y). \tag{11}$$

Based on Eqs. (9) and Eq. (10), BMSE proposes to optimize the NLL loss of $p_{valid}(y|\boldsymbol{x},\theta)p_{train}(y)$. An implementation form of BMSE loss for the sample $(\boldsymbol{x}_i, y_i)$ with the deep feature $\boldsymbol{z}_i$ is given by

$$\ell_{BMSE}(\boldsymbol{z}_i) = -\log \frac{\exp(-(y_b - \hat{y}(\boldsymbol{x}_i))^2)}{\sum_{b' \in \mathcal{B}} \exp(-(y_{b'} - \hat{y}(\boldsymbol{x}_i))^2)}, \tag{12}$$

The aforementioned inference of BMSE holds if the LCDI assumption is satisfied. However, beyond imbalanced distributions, additional cases, such as the presence of clustering structures among features, have been revealed [35]. In such cases, the LCDI assumption is unsatisfied. We now present our proposed loss, namely Balanced MSE. Recent work [35] reveals that under a clustering structure, the label-conditional distribution can be refined as follows:

$$p(\boldsymbol{x}|y) = \sum_{k=1}^{K} p(\pi(\boldsymbol{x}) = k|y) p(\boldsymbol{x}|y, \pi(\boldsymbol{x}) = k), \tag{13}$$

where $\pi : \mathcal{X} \to \{1, 2, \cdots, K\}$ maps each $\boldsymbol{x} \in \mathcal{X}$ to the index of the cluster it belongs to. $p(\boldsymbol{x}|y, \pi(\boldsymbol{x}))$ is the label-conditional distribution projected onto the partition containing $\boldsymbol{x}$. Specifically, [35] states the following assumption.

Assumption 2. *The partition-projected label-conditional distribution invariance (PLCDI) holds if* $\forall k \in \{1, 2, \cdots, K\}$

$$p_{train}(\boldsymbol{x}|y, \pi(\boldsymbol{x}) = k) = p_{valid}(\boldsymbol{x}|y, \pi(\boldsymbol{x}) = k).$$

When $K = 1$, PLCDI is equivalent to LCDI assumption. Based on this assumption ensured by [35], we can further reconsider Eq. (10). Under PLCDI assumption and the assumption that the validation set is balanced, we rewrite Eq. (10),

$$\frac{p_{train}(y|\boldsymbol{x})}{p_{valid}(y|\boldsymbol{x})} \propto \frac{\sum_{k=1}^{K} p_{train}(\pi(\boldsymbol{x}) = k|y) p_{train}(y)}{\sum_{k=1}^{K} p_{valid}(\pi(\boldsymbol{x}) = k|y) p_{valid}(y)} \tag{14}$$

Accordingly, we propose a novel loss namely Rebalanced-MSE loss (ReMSE) which optimize the NLL loss of

$$\sum_{k=1}^{K} p_{valid}(y|\boldsymbol{x},\theta) p_{train}(\pi(\boldsymbol{x}) = k|y) p_{train}(y).$$

An implementation of ReMSE is given following the inference of BMSE in [23]. The ReMSE loss of the sample $(\boldsymbol{x}_i, y_i)$ with the deep feature $\boldsymbol{z}_i$ is given by

$$\ell_{ReMSE}(\boldsymbol{z}_i) \simeq -\sum_{k=1}^{K} \log \frac{\exp(-\eta_k^{y_i}(y_i - g(\boldsymbol{z}_i))^2)}{\sum_{k=1}^{K}\sum_{y_{b'}\in\mathcal{Y}} \exp(-\eta_k^{y_{b'}}(y_{b'} - g(\boldsymbol{z}_i)))^2)}, \tag{15}$$

where $\eta_k^{y_i} = \mathbb{P}(\pi(\boldsymbol{x}_i) = k | y = y_i)$. Denote $u_{y_{b'}}(\boldsymbol{z}_i) = -(y_{b'} - g(\boldsymbol{z}_i))^2$. By Bayes' Rule, $p(\pi|y) = p(y|\pi)p(\pi)/p(y)$. Therefore, $\mathbb{P}(\pi(\boldsymbol{x}_i) = k | y = y_b)$ can be obtained by estimating $\mathbb{P}(y = y_b | \pi(\boldsymbol{x}_i) = k)$ and $\mathbb{P}(\pi(\boldsymbol{x}_i) = k)$ given $\mathbb{P}(y = y_b)$. The estimation method follows [35], which is composed by firstly applying k-means clustering then counting samples in each cluster.

3.4.2. The deduction of surrogate loss

Thereafter, the entire loss of M augmented samples can be given,

$$\mathcal{L}_M = \sum_{b\in\mathcal{B}} \frac{1}{N_b} \sum_{i=1}^{N_b} \frac{1}{M} \sum_{m=1}^{M} \ell_{ReMSE}(\boldsymbol{z}_i^m), \tag{16}$$

Considering the needs of tail labels, M must be large enough since the effect of augmentation is limited with a small number of augmented samples. However, as M increases, the process becomes time-consuming. An implicit approach inspired by ISDA and RISDA is to let $M \to \infty$. The loss can be defined as

$$\mathcal{L}_{M\to\infty} = \sum_{b\in\mathcal{B}} \frac{1}{N_b} \sum_{i=1}^{N_b} \mathbb{E}_{\tilde{\boldsymbol{z}}_i}[\ell_{ReMSE}(\tilde{\boldsymbol{z}}_i)], \tag{17}$$

Though, Eq. (17) is still challenging to calculate. An upper bound for Eq. (17) can be obtained using the Jensen's inequality $\mathbb{E}[\log X] \leq \log \mathbb{E}[X]$,

$$\begin{aligned}
&\mathcal{L}_{M\to\infty} \\
&\leq \sum_{b\in\mathcal{B}} \frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{k=1}^{K} \log \left(\mathbb{E}_{\tilde{\boldsymbol{z}}_i} \left[\sum_{y_{b'}\in\mathcal{Y}} e^{\eta_k^{y_{b'}} u_{y_{b'}}(\tilde{\boldsymbol{z}}_i) - \eta_k^{y_i} u_{y_i}(\tilde{\boldsymbol{z}}_i)} \right] \right) \\
&\leq \sum_{b\in\mathcal{B}} \frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{k=1}^{K} \log \left(\sum_{y_{b'}\in\mathcal{Y}} \mathbb{E}_{\tilde{\boldsymbol{z}}_i} \left[e^{\eta_k^{y_{b'}} u_{y_{b'}}(\tilde{\boldsymbol{z}}_i) - \eta_k^{y_i} u_{y_i}(\tilde{\boldsymbol{z}}_i)} \right] \right)
\end{aligned} \tag{18}$$

Due to $\tilde{\boldsymbol{z}}_i \sim \mathcal{N}(\boldsymbol{z}_i + \alpha\tilde{\mu}_b, \beta\tilde{\Sigma}_b)$ and the moment-generating function

$$\mathbb{E}[e^{tX}] = e^{t\mu+\frac{1}{2}\sigma t^2},\ X \sim \mathcal{N}(\mu, \sigma),$$

we can deduce the following relation,

$$\begin{aligned}&\mathbb{E}_{\tilde{\boldsymbol{z}}_i}\left[e^{\eta_k^{y_{b'}} u_{y_{b'}}(\tilde{\boldsymbol{z}}_i) - \eta_k^{y_i} u_{y_i}(\tilde{\boldsymbol{z}}_i)}\right] \\ &= e^{\eta_k^{y_{b'}} u_{y_{b'}}(\boldsymbol{z}_i) - \eta_k^{y_i} u_{y_i}(\boldsymbol{z}_i) + r_1^{b'i}\alpha \boldsymbol{w}^\top \tilde{\mu}_b + r_2^{b'i}\beta \boldsymbol{w}^\top \tilde{\Sigma}_b \boldsymbol{w}},\end{aligned} \tag{19}$$

where $r_1^{b'ik} = (\eta_k^{y_i} - \eta_k^{y_{b'}}) + 2(\eta_k^{y_{b'}} y_{b'} - \eta_k^{y_i} y_i)$ and $r_2^{b'ik} = (\eta_k^{y_{b'}} y_{b'} - \eta_k^{y_i} y_i)^2 + (\eta_k^{y_i} - \eta_k^{y_{b'}})^2/4$. Finally, we can get the upper bound of the loss,

$$\overline{\mathcal{L}}_\infty = -\sum_{b\in\mathcal{B}} \frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{k=1}^{K} \log(\sigma_{i,k}), \tag{20}$$

where $\tilde{Z}_{i,k}^{y_{b'}} = r_1^{b'ik}\alpha|\boldsymbol{w}^\top \tilde{\mu}_b - y_i| + r_2^{b'ik}\beta \boldsymbol{w}^\top \tilde{\Sigma}_b \boldsymbol{w} + \eta_k^{y_{b'}} u_{y_{b'}}(\boldsymbol{z}_i)$ and

$$\sigma_{i,k} = \left(\frac{\exp(\tilde{Z}_{i,k}^{y_i})}{\sum_{k=1}^{K}\sum_{y_{b'}\in\mathcal{Y}} \exp(\tilde{Z}_{i,k}^{y_{b'}})}\right).$$

As pointed out by [14], if both head and tail labels are augmented infinitely, samples from the head labels will still dominate because the augmentation is based on the training samples, which are predominantly from head labels. To address this issue, a re-weighting strategy is employed to adjust the augmentation strength, ensuring that tail labels are given more weight during training.

$$\mathcal{L}_{IRDA} = -\sum_{b\in\mathcal{B}} \frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{k=1}^{K} \frac{1}{\pi_{y_i}} \log(\sigma_{i,k}), \tag{21}$$

where $\pi_{y_i} = \mathbb{P}(y = y_i)$. According to Logit Adjustment (LA) [26], π_{y_i} can be reformed as a logit perturbation, i.e.,

$$\begin{aligned}\mathcal{L}_{IRDA} = \sum_{b\in\mathcal{B}} \frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{k=1}^{K} \log(1 + \sum_{y_{b'}\neq y_i} \sum_{k=1}^{K} \exp(\Delta Z_{y_i k}^{y_{b'}} \\ + r_1^{b'ik}\alpha|\boldsymbol{w}^\top \tilde{\mu}_b - y_i| + r_2^{b'ik}\beta \boldsymbol{w}^\top \tilde{\Sigma}_b \boldsymbol{w} + \log(\pi_{y_{b'}}/\pi_{y_i}))),\end{aligned} \tag{22}$$

where $\Delta Z_{y_i k}^{y_{b'}} = \eta_k^{y_{b'}} u_{y_{b'}}(\boldsymbol{z}_i) - \eta_k^{y_i} u_{y_i}(\boldsymbol{z}_i)$. Eq. (22) provides the surrogate loss for IRDA.

3.5. Implementations of IRDA

Our proposed method, IRDA, includes three crucial parameters: $\tilde{\boldsymbol{\mu}}_b$, $\tilde{\boldsymbol{\Sigma}}_b$ and v_b, which vary according to different labels and are predefined. Accordingly, we propose two implementations of our proposed method, which are a direct estimation approach and a meta-learning-based approach.

3.5.1. Direct estimation approach (IRDA)

A straightforward approach to determining the values of the three crucial parameters is to directly calculate $\tilde{\boldsymbol{\mu}}_b$, $\tilde{\boldsymbol{\Sigma}}_b$ and choose a value for v_b based on ablation experiments. However, the covariance matrix is anisotropic during the initial iterations of the training process which can hinde the convergence of test error in high-dimensional regression [36].

To address this, we propose estimating an isotropic covariance matrix to accelerate the convergence of test error. The covariance matrix is estimated by adjusting the $\tilde{\boldsymbol{\Sigma}}_b$ with the identity matrix. Precisely,

$$\overline{\boldsymbol{\Sigma}}_b = (1 - v_b)\tilde{\boldsymbol{\Sigma}}_b + v_b\lambda_b\mathbf{I}, \tag{23}$$

where λ_b is the largest item of $\tilde{\boldsymbol{\Sigma}}_b$ and $\mathbf{I}$ is an identity matrix of the size of $\tilde{\boldsymbol{\Sigma}}_b$. Inspired by [37], the parameter can be calculated as follows,

$$v_b = \sum_{i=1}^{D} \left((\lambda_b - \sigma_{i,i}^b)^2 + \sum_{j \neq i} {\sigma_{i,j}^b}^2 \right), \tag{24}$$

where $\sigma_{i,j}^b$ is the item of $\tilde{\boldsymbol{\Sigma}}_b$ in the i-th line and j-th column. v_b for each label is normalized according to the whole label set. High values of v_b indicate severe correlation between features, implying high anisotropy. This adjustment coincides with whitening, a commonly used approach for decorrelation of features to avoid overfitting.

3.5.2. Meta-learning based approach (Meta-IRDA)

Inspired by previous work [13], meta-learning is also introduced to strengthen the effectiveness of IRDA. We tend to learn crucial parameters $\tilde{\boldsymbol{\mu}}_b$, $\tilde{\boldsymbol{\Sigma}}_b$ and v_b of each label by meta-learning using a series of label-concerning inputs, which are as follows,

- N_b: The normalized number of samples in each bin is employed to reflect the extent of imbalance.

Algorithm 1 IRDA

Input: Training data D^{train}, batch size n, number of iterations T, kernel k, α, β.

Output: Learned parameter $\boldsymbol{w}$.

Initialization: Initialize $\boldsymbol{w}^{(1)}$, $\boldsymbol{\varepsilon}^{(1)}$.
for $t = 1$ to T **do**
  Sample $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ from D^{train};
  Extract features of each sample $\{\boldsymbol{z}_i\}_{i=1}^{n}$;
  Calculate $\boldsymbol{\mu}_b^{(t)}$ and $\boldsymbol{\Sigma}_b^{(t)}$ for each label y_b;
  $\tilde{\boldsymbol{\mu}}^{(t)}, \tilde{\boldsymbol{\Sigma}}^{(t)} \leftarrow$ W-FDS $\left(\boldsymbol{\mu}_b^{(t)}, \boldsymbol{\Sigma}_b^{(t)}, \boldsymbol{\varepsilon}^{(t)}, k\right)$;
  Calculate v_b by Eq. (24);
  Formulate $\overline{\boldsymbol{\Sigma}}_b^{(t)}$ by Eq. (23) using v_b, λ_b and $\tilde{\boldsymbol{\Sigma}}_b$;
  $\boldsymbol{\eta} \leftarrow$ Cluster$(\boldsymbol{z}_i)$;
  Calculate $\mathcal{L}_{IRDA}$ by Eq. (22) using $\tilde{\boldsymbol{\mu}}^{(t)}, \overline{\boldsymbol{\Sigma}}_b^{(t)}, \boldsymbol{\eta}, \alpha, \beta$;
  Update $\boldsymbol{w}^{(t+1)}$ by back-propagation on $\mathcal{L}_{IRDA}$;
  Update $\boldsymbol{\varepsilon}^{t+1}$ by Eq. (6);
end for

- ρ_b: As mentioned in [22], two labels containing the same number of samples can exhibit different extents of imbalance due to the regions they belong to. A high-density region indicates a set of neighboring labels that are mostly head labels, whereas a low-density region indicates a set of tail labels. Therefore, it is important to consider region density, which is calculated as follows,

$$\rho_b = \frac{\sum_{b-2}^{b+2} N_b}{N},$$

  where N signifies the amount of samples of the entire dataset.

- $\mathcal{U}_b$: The uncertainty is a widely used characteristic of samples. Therefore, we compute the average uncertainty of samples from each label by calculating the average information entropy of the model predictions, i.e.,

$$\mathcal{U}_b = -\sum_{i=1}^{N_b} \hat{y}(\boldsymbol{x}_i) \log \hat{y}(\boldsymbol{x}_i).$$

- $\overline{L}_b$: The average loss of each label is also reported,

$$\overline{\mathcal{L}}_b = -\frac{1}{N_b}\sum_{i=1}^{N_b}\ell(\boldsymbol{x}_i),$$

where $\ell(\cdot)$ signifies the ReMSE loss of sample $(\boldsymbol{x}_i, y_i)$.

The four characteristics listed above are extracted as inputs for the meta-network, which is a two-layer MLP. Following MetaSAug [13] and Meta-Weight-Net [38], we optimize three label-wise parameters $\tilde{\boldsymbol{\mu}}_b$, $\tilde{\boldsymbol{\Sigma}}_b$ and v_b on metadata.

3.6. Regularization Analysis

Regularization is a common technique used in various machine learning tasks to enhance model generalizability. Ridge regression is a typical example of a regularization approach that mitigates overfitting by reducing the L_2 norm of a model's parameters and diminishing the variance of its predictions. BMSE introduces a modified MSE loss function that enables the derivation of a regularization term. This term in BMSE works to counteract the negative effects of imbalanced data by penalizing samples from tail labels.

The proposed IRDA approach can also be adapted into a regularization form, similar to ISDA-R and RISDA-R. From a regularization perspective, the rationale behind IRDA can be more clearly articulated by comparing it to existing algorithms. The proposed surrogate loss of IRDA in Eq. (22) can be rewritten in a regularization form using the first-order Taylor expansion. For a given loss $\ell(\boldsymbol{u}) = -\log\frac{\exp(u_i\eta_i)}{\sum_j\exp(u_j\eta_j)}$, we have

$$\ell(\boldsymbol{u}+\Delta\boldsymbol{u}) \approx \ell(\boldsymbol{u}) + \left(\frac{\partial l}{\partial \boldsymbol{u}}\right)^\top \Delta\boldsymbol{u}.$$

Accordingly, for a sample $(\boldsymbol{x}_i, y_i)$ falling in the b-th bin, i.e., $y_i \in [y_b, y_{b+1})$, denote

$$\boldsymbol{u}_i = \left(u_{y_1}(\boldsymbol{z}_i), \cdots, u_{y_b}(\boldsymbol{z}_i), \cdots, u_{y_B}(\boldsymbol{z}_i)\right)^\top$$

$$\Delta\boldsymbol{u}_i = \begin{pmatrix} r_1^{y_1}\alpha|\boldsymbol{w}^\top\tilde{\mu}_b - y_i| + r_2^{y_1}\beta\boldsymbol{w}^\top\tilde{\Sigma}_b\boldsymbol{w} + \delta_1 \\ \vdots \\ \delta_b \\ \vdots \\ r_1^{y_B}\alpha|\boldsymbol{w}^\top\tilde{\mu}_b - y_i| + r_2^{y_B}\beta\boldsymbol{w}^\top\tilde{\Sigma}_b\boldsymbol{w} + \delta_B \end{pmatrix},$$

Table 2: Regularization terms and generalization effect of existing methods.

Methods	Uncertainty	Skewness	Compactness	Precision
Ridge Regression	$\downarrow$	-	-	-
BMSE	-	$\downarrow$	-	-
ISDA-R	-	-	$\downarrow$	-
RISDA-R	-	-	$\downarrow$	$\downarrow$
IRDA	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$

as the logit and the logit perturbation successively and $\delta_{b'} = \log(\pi_{y_{b'}})$. Reformulate the continuous scalar target value $y_i = y_b$ into a one-hot vector $\boldsymbol{y}_i = (0, \cdots, 1, \cdots, 0)^\top$ where the b-th item of $\boldsymbol{y}_i$ equals to 1 and others equal to 0. Denote $\boldsymbol{q}_i = (q_{i,1}, \cdots, q_{i,B})^\top$ and $q_{i,b,k} = \exp(\eta_k^{y_i} u_{y_b}(\boldsymbol{z}_i)) / \sum_{y_{b'} \in \mathcal{Y}} \exp(\eta_k^{y'_b} u_{y_{b'}}(\boldsymbol{z}_i))$. Denote $\boldsymbol{\eta}_k = (\eta_k^{y_1}, \cdots, \eta_k^{y_B})$. The b-th item of $\Delta \boldsymbol{u}_i$ equals to 0. According to Eq. (22), we have

$$
\begin{aligned}
\ell(\boldsymbol{u}_i + \Delta \boldsymbol{u}_i) &\approx \ell(\boldsymbol{u}_i) + (\boldsymbol{\eta}_k \odot (\boldsymbol{q}_i - \boldsymbol{y}_i))^\top \Delta \boldsymbol{u}_i \\
&= \ell(\boldsymbol{u}_i) + (\eta_k^{y_1} q_{i,1,k}, \cdots, \eta_k^{y_b}(q_{i,b,k} - 1), \cdots, \eta_k^{y_B} q_{i,B,k})^\top \\
&\times \begin{pmatrix} r_1^{y_1 ik} \alpha |\boldsymbol{w}^\top \tilde{\mu}_b - y_i| + r_2^{y_1 ik} \beta \boldsymbol{w}^\top \tilde{\Sigma}_b \boldsymbol{w} + \delta_1 \\ \vdots \\ \delta_b \\ \vdots \\ r_1^{y_B ik} \alpha |\boldsymbol{w}^\top \tilde{\mu}_b - y_i| + r_2^{y_B ik} \beta \boldsymbol{w}^\top \tilde{\Sigma}_b \boldsymbol{w} + \delta_B \end{pmatrix} \\
&= \ell(\boldsymbol{u}_i) + \eta_k^{y_b} q_{i,b,k} \delta_b + R_1^{i,b,k} |\boldsymbol{w}^\top \tilde{\mu}_b - y_i| \\
&+ R_2^{i,b,k} \boldsymbol{w}^\top \tilde{\Sigma}_b \boldsymbol{w} + \sum_{y_{b'} \neq y_b} \eta_k^{y_{b'}} \delta_{b'},
\end{aligned}
\tag{25}
$$

where $R_1^{i,b,k} = \sum_{y_{b'} \neq y_b} q_{i,b',k} \alpha \eta_k^{y_{b'}} r_1^{y_{b'} ik}$ and $R_2^{i,b,k} = \sum_{y_{b'} \neq y_b} q_{i,b',k} \beta \eta_k^{y_{b'}} r_2^{y_{b'} ik}$. Since $\sum_{y_{b'} \neq y_b} \eta_k^{y_{b'}} \delta_{b'}$ is a constant item, we omit the term. Thereafter, the proposed IRDA loss in Eq. (22) can be rewritten in a regularization form as follow,

$$
\mathcal{L}_{IRDA} = \sum_{b \in \mathcal{B}} \frac{1}{N_b} \sum_{i=1}^{N_b} \mathcal{L}_{ReMSE} + R_{IRDA}, \tag{26}
$$

with

$$R_{IRDA} = \sum_{b\in\mathcal{B}} \frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{k=1}^{K} [\eta_k^{y_b} q_{i,b,k} \delta_b + R_1^{i,b,k} |\boldsymbol{w}^\top \tilde{\mu}_b - y_i| + R_2^{i,b,k} \boldsymbol{w}^\top \tilde{\Sigma}_b \boldsymbol{w}].$$

For a profound comprehension of the regularization analysis, we introduce the following two propositions concerning the averaged boundary distance and the mapping variance of features.

Definition 1. *For the samples $\{\boldsymbol{x}_i, y_i\}_{i=1,\cdots,N_b}$ falling in b-th bin of the label space $\mathcal{Y}$, the averaged residual distance of features according to the true model is defined by $|\boldsymbol{w}^\top \tilde{\boldsymbol{\mu}}_b - y_b|$.*

As shown in Fig. 2, the average prediction of features $\{\boldsymbol{x}_i, y_i\}_{i=1,\cdots,N_b}$ has a residual distance to the true center of the feature. According to the problem setting, which divides the label space into B bins, the true center of the features should be located at y_b. If the residual distance is large, the true model is poorly approximated by the learned one. Therefore, the learned model should be adjusted to approach the true model, which can be achieved by drawing the average prediction closer to the true center of the features.

Proposition 1. *For the samples $\{\boldsymbol{x}_i, y_i\}_{i=1,\cdots,N_b}$ falling in b-th bin of the label space $\mathcal{Y}$, the mapping variance of features $\{\boldsymbol{z}_i\}_{i=1,\cdots,N_b}$ is $\boldsymbol{w}^\top \tilde{\Sigma}_b \boldsymbol{w}$.*

Proof. The boundary distance d describes the vertical distance between the features and the true model. It also illustrate the position of projected feature $\boldsymbol{z}_i$ according to the model. The mapping value of each feature can be expressed by

$$\phi(\boldsymbol{z}_i) = |\boldsymbol{w}^\top \boldsymbol{z}_i - y_b|.$$

Expand $\phi(\boldsymbol{z}_i)$ as $\phi(\boldsymbol{z}_i) = \boldsymbol{w}^\top \boldsymbol{z}_i - y_b$ which gives the possibility that the features could be projected on the both side of the true model. The mapping variance can be deduced as follows,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{z}_i}[\phi(\boldsymbol{z}_i)] &= \mathbb{E}_{\boldsymbol{z}_i}\left[(\boldsymbol{w}^\top \boldsymbol{z}_i - y_b - \mathbb{E}_{\boldsymbol{z}_i}[\boldsymbol{w}^\top \boldsymbol{z}_i - y_b])^2\right] \\
&= \mathbb{E}_{\boldsymbol{z}_i}\left[(\boldsymbol{w}^\top \boldsymbol{z}_i - y_b - \boldsymbol{w}^\top \mathbb{E}_{\boldsymbol{z}_i}[\boldsymbol{z}_i] + y_b)^2\right] \\
&= \mathbb{E}_{\boldsymbol{z}_i}\left[(\boldsymbol{w}^\top \boldsymbol{z}_i - \boldsymbol{w}^\top \mathbb{E}_{\boldsymbol{z}_i}[\boldsymbol{z}_i])^2\right] \\
&= \mathbb{E}_{\boldsymbol{z}_i}\left[\boldsymbol{w}^\top (\boldsymbol{z}_i - \mathbb{E}_{\boldsymbol{z}_i}[\boldsymbol{z}_i])(\boldsymbol{z}_i - \mathbb{E}_{\boldsymbol{z}_i}[\boldsymbol{z}_i])^\top \boldsymbol{w}\right] \\
&= \boldsymbol{w}^\top \mathbb{E}_{\boldsymbol{z}_i}\left[(\boldsymbol{z}_i - \tilde{\boldsymbol{\mu}}_b)(\boldsymbol{z}_i - \tilde{\boldsymbol{\mu}}_b)^\top\right] \boldsymbol{w} \\
&= \boldsymbol{w}^\top \tilde{\Sigma}_b \boldsymbol{w}
\end{aligned}$$

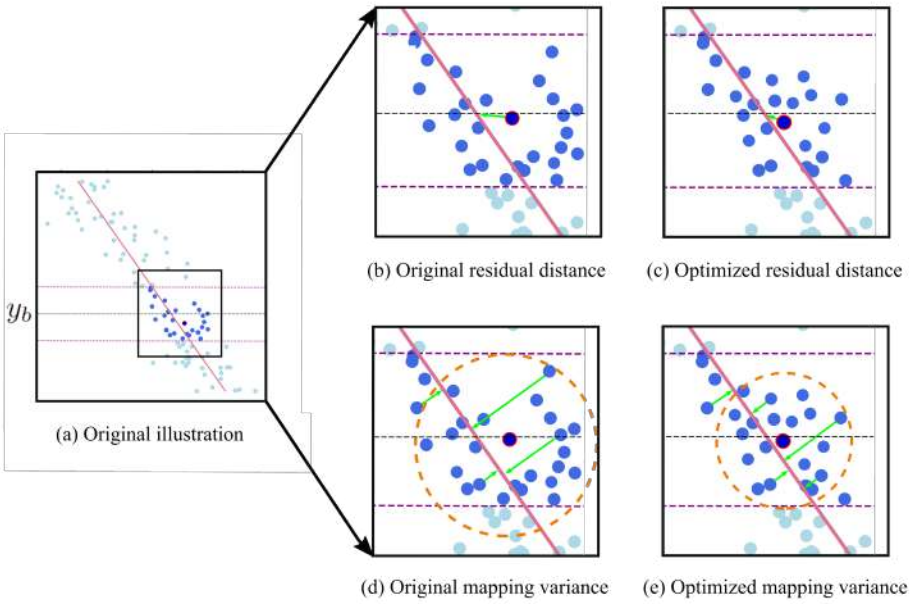


Figure 2: For regularization analysis, an example of linear regression is given. The purple dotted line restricts the b-th bin of the label space and the black dotted line is the mean value of b-th bin. The learned model is shown as the pink line. The black squared area is zoomed in and illustrations before and after optimization of the focused area are shown in (b)-(e).

□

Apart from our proposed IRDA, the regularization terms of Ridge Regression, BMSE, ISDA-R, and RISDA-R are formulated as follows,

- Ridge Regression: $R_{ridge} = \lambda \boldsymbol{w}^\top \boldsymbol{w}$
- BMSE : $R_{BMSE} = \sum_{b\in\mathcal{B}} \frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{y_{b'}\neq y_b} q_{i,b'} \delta_{b'}$
- ISDA-R : $R_{ISDA-R} = \sum_{b\in\mathcal{B}} \frac{1}{N_b} \sum_{i=1}^{N_b} \beta q_{i,b} \boldsymbol{w}^\top \Sigma_b \boldsymbol{w}$
- RISDA-R : $R_{RISDA-R} = \sum_{b\in\mathcal{B}} \frac{1}{N_b} \sum_{i=1}^{N_b} q_{i,b} [\beta \boldsymbol{w}^\top \Sigma_b \boldsymbol{w} + \sum_{y_{b'}\neq y_b} (\alpha |\boldsymbol{w}^\top \tilde{\boldsymbol{\mu}}_{b'} - y_b| + \beta \boldsymbol{w}^\top \tilde{\Sigma}_{b'} \boldsymbol{w})]$

The effects of the regularization terms of the aforementioned methods are summarized in Table 2. The analysis reveals that the typical regularization term used in Ridge Regression improves generalizability by reducing overall

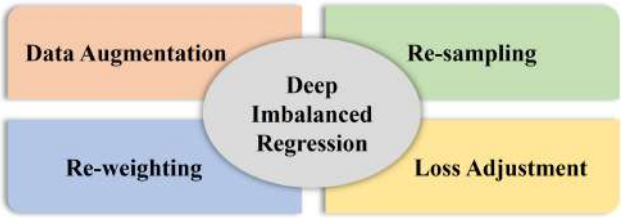


Figure 3: An overview of existing strategies for DIR tasks. According to data augmentation, only explicit data augmentation methods have been investigated in the previous literature.

prediction variance, thereby significantly decreasing prediction uncertainty. The regularization term of BMSE addresses imbalance by increasing the loss for tail samples, thus reducing the imbalance extent. ISDA-R aims to reduce the variance of features within each label, making samples from different labels more distinct and increasing sample compactness. RISDA-R reduces the variance of features for tail labels and decreases the average residual distance, bringing each sample closer to the true model.

IRDA encourages greater compactness of samples within each label. Additionally, as previously discussed, each cluster comprises samples from different labels, and the distribution within each cluster is also imbalanced. Under this clustering structure, IRDA encourages tail samples in each cluster to have larger losses and to be more focused.

3.7. Comparison between related works

In this section, we compare the differences between several related methods from two perspectives: motivation and implementation.

- **Motivation:** Our proposed IRDA is an implicit data augmentation approach. Implicit data augmentation has recently been proposed for imbalanced classification and has demonstrated significant effectiveness. Our work introduces the implicit data augmentation into deep imbalanced regression for the first time. The motivation behind implicit data augmentation is to redesign the loss by virtually generating samples and enriching the features with more semantic information.

  Several methods proposed for addressing imbalanced issues fall into the following categories: re-sampling, re-weighting, loss adjustment, and explicit data augmentation. Fig. 3 illustrates the existing types of methods proposed for DIR tasks. For example, LDS proposed by [22]

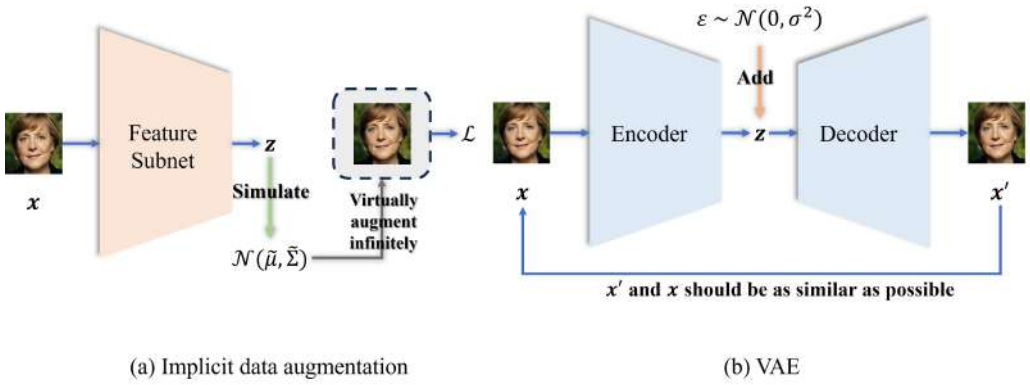


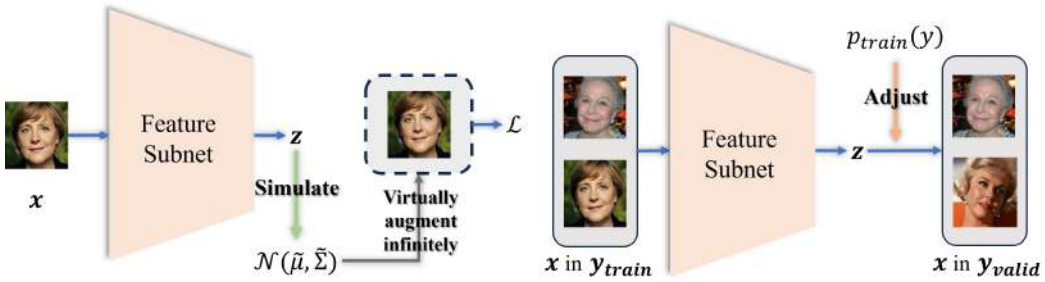
Figure 4: The comparison between IRDA and VAE.

is a re-weighting approach. The motivation behind re-weighting approaches is to rebalance the skewed distribution of labels. BMSE proposed by [23] is a loss adjustment approach. The motivation behind loss adjustment approaches is to modify the loss according to various heuristic discoveries. Variational Auto Encoder (VAE) is a typical explicit data augmentation approach. The motivation of explicit data augmentation is to actually generate samples and increase the data size. Accordingly, different approaches have distinct motivation.

- **Implementation:** The implementation of each method differs. We take the following four methods as example and compare with our proposed IRDA.

  - **LDS:** The differences between LDS and IRDA can be summarized as follows:

    * The information considered in LDS and IRDA is different. LDS uses the proportion of each label in the training set as the weight. The prior of the label in the training set, i.e., $p_{train}(y)$, , guides the redesign of the loss. IRDA uses the semantic information extracted from features of easily-confused labels to enrich the features of tail labels.
    * The application targets of LDS and IRDA differ. LDS is a label-wise approach, assigning the same weight to all samples from the same target. IRDA is a sample-wise approach, processing different samples differently.



(a) Implicit data augmentation (b) BMSE

Figure 5: The comparison between IRDA and BMSE.

– **FDS:** The differences between FDS and IRDA can be summarized as follows:

  * The focus of FDS and IRDA differs. FDS uses the information extracted from neighbor labels to enrich the features of tails labels. IRDA uses the semantic information extracted from features of easily-confused labels to enrich the features of tail labels.
  * The attentions on which FDS and IRDA pay attention are different. FDS treats samples from each label equally, whereas IRDA applies weights to samples, giving more focus to those from tail labels.

– **VAE:** The differences between VAE and IRDA can be summarized as follows:

  * Their applied scenarios are different. VAE is an unsupervised method, whereas IRDA is a supervised method.
  * The goals of VAE and IRDA are different. VAE aims to generate samples that are similar to the input samples. IRDA aims to extract semantic information by virtually generating samples to enrich deficient features.
  * The perturbations used in VAE and IRDA are different. VAE perturbs features with a zero-mean error and random scale. IRDA perturbs features using semantic information extracted from easily-confused labels.

– **BMSE:** The differences between BMSE and IRDA can be summarized as follows:
  * BMSE and IRDA are based on different assumptions. BMSE relies on the LCDI assumption, i.e., $p_{train}(x|y) = p_{valid}(x|y)$. IRDA, on the other hand, uses the PLCDI assumption, which posits that samples from each label have a clustering structure that differs from label to label.
  * The goals of BMSE and IRDA are different. BMSE aims to learn a balanced posterior by redesigning the loss based on the distribution of labels in the training set.
  * The application targets of BMSE and IRDA are different. BMSE is a label-wise approach, meaning that it applies identical adjustments to all samples from the same target. In contrast, IRDA is a sample-wise approach, meaning that it differentially influences distinct samples.

Fig. 4 illustrates the implementation differences between VAE and IRDA. Fig. 5 shows the implementation differences between BMSE and IRDA. As demonstrated, the aforementioned methods have clear distinctions from our proposed IRDA.

4. Experimental results

In this section, we present the validation of the proposed method described in Section III through extensive experiments. We compare our proposed methodology with nine recent methods from four strategies: re-sampling, explicit data augmentation, loss adjustment, and re-weighting. Additionally, we use five benchmarks that have been employed in three recent works for comparison. Section IV-A describes the DIR benchmarks used, which span computer vision and healthcare. The implementation details are provided in Section IV-B, while Section IV-C details the various state-of-the-art methods chosen for comparison, along with the evaluation metrics used. The main results are presented in Section IV-D, followed by further analysis in Section IV-E.

4.1. Datasets

Five DIR benchmarks spanning computer vision and healthcare are employed. The label density distribution and the level of imbalance are detailed

in [22].

- *AgeDB-DIR*: Constructed from the AgeDB dataset [39], which contains face images with corresponding ages. Ages range from 0 to 101 with a bin length of 1. The maximum bin density is 353 images, and the minimum bin density is 1 in the 12K training samples. 2.1K images are split as the balanced validation and test sets.

- *IMDB-WIKI-DIR*: IMDB-WIKI-DIR [40] is a large-scale face image dataset for age estimation from single input images. The original dataset contains 523.0K face images and corresponding ages, with 460.7K images from the IMDB website and 62.3K images from Wikipedia. The IMDB-WIKI-DIR constructed by [22] used in this work contains 191.5K images for training and 11.0K images for validation and testing.

- *NYUD2-DIR*: Constructed from the NYU Depth Dataset V2 [41], containing images and depth maps for different indoor scenes. Depth ranges from 0.7 to 10 meters with a bin length of 0.1. The training set contains 50,688 images, and the balanced test set contains 654 images.

- *TUAB*: TUAB [42] is used for brain-age estimation from EEG resting-state signals. The dataset comes from EEG exams at the Temple University Hospital in Philadelphia. Following the setting in [42], the dataset is split into a 1,246-subject training set and a 139-subject test set.

- *SkyFinder*: SkyFinder [43, 44] is used for temperature prediction from outdoor webcam images. It contains 35,417 images captured by 44 cameras around 11am on each day under a wide range of weather and illumination conditions. It is split into a 28,373-image training set, a 3,522-image validation set and a 3,522-image test set.

4.2. Implementation Details

4.2.1. Network architectures

The proposed method is implemented using PyTorch. We adopt ResNet-50 [45] as our backbone network for AgeDB-DIR, IMDB-WIKI-DIR and SkyFinder. Besides, ResNet-50-based encoder-decoder architecture [46] is employed for NYUD2-DIR and a 24-layer 1D ResNet used in [47] is employed for TUAB.

Table 3: Benchmarking results on AgeDB-DIR.

Metrics	MAE ↓				GM ↓			
Shot	All	Many	Med.	Few	All	Many	Med.	Few
VANILLA	7.77	6.62	9.55	13.67	5.05	4.23	7.01	10.75
SMOTER [20]	8.16	7.39	8.65	12.28	5.21	4.65	5.69	8.49
SMOGN [21]	8.26	7.64	9.01	12.09	5.36	4.90	6.19	8.44
FOCAL-R [22]	7.64	6.68	9.22	13.00	4.90	4.26	6.39	9.52
SMOGN + LDS + FDS [22]	7.90	7.32	8.51	11.19	4.98	4.64	5.41	7.35
FOCAL-R + LDS + FDS [22]	7.47	6.69	8.30	12.55	4.71	4.25	5.36	8.59
BMC [23]	7.96	6.94	8.92	12.04	5.11	4.87	5.84	7.62
GAI [23]	7.67	6.75	9.04	11.32	4.86	4.24	5.91	8.89
SupCR(MSE) [47]	7.36	6.53	8.37	11.00	4.78	4.35	5.66	7.17
RankSim [28]	7.23	6.51	8.77	10.92	4.48	4.01	5.87	7.79
ISDA-R	7.11	**6.01**	8.14	9.95	4.55	3.99	5.40	7.09
RISDA-R	7.19	6.33	8.09	9.98	4.61	4.09	5.55	6.87
IRDA	6.62	6.11	7.89	8.99	4.17	3.35	**4.75**	6.64
Meta-IRDA	**6.58**	6.02	**7.60**	**8.47**	**3.94**	**3.18**	4.80	**6.30**
SOTA (Best) **vs.** VANILLA	+0.54	+0.11	+1.18	+2.75	+0.57	+0.22	+1.60	+3.58
OURS (Best) **vs.** VANILLA	+1.19	+0.61	+1.95	+5.20	+1.11	+1.05	+2.26	+4.45

4.2.2. Training details

The standard model without any technique for dealing with imbalanced data is used as the backbone. For fair comparison, the same settings for SMOTER and SMOGN described in [22] are used, as well as the settings for Focal-R, LDS, and FDS.

All models are trained for 90 epochs using the Adam optimizer [48] with an initial learning rate of 10^{-3}, which is decayed by 0.1 at the 60th and 80th epochs. The batch size is fixed at 256. For LDS, FDS, and our W-FDS, we use a Gaussian kernel with a kernel size of 5 and a standard deviation of 2. The momentum of W-FDS and FDS is fixed at 0.9 following [22] for fair comparison. α and β in Eq. (7) are set to 0.5 and 0.75 respectively. The choices of α and β are detailed in Section IV.E 2) through sensitive tests.

4.3. Compared Methods and Evaluation Metrics

4.3.1. Compared methods

Few methods have been proposed for DIR, and we describe here both the compared methods and adapted imbalanced classification methods here.

- *Re-sampling*: The SMOTER algorithm first separates samples into rare and frequent categories, then performs oversampling on the rare samples and undersampling on the frequent samples.

- *Explicit data augmentation*: SMOGN generates new rare samples by using linear combinations of existing rare samples, and generates new frequent samples by adding Gaussian noise to the existing frequent samples.

- *Weighting strategy*: FOCAL-R is proposed as a sample-wise error-aware weighting strategy for DIR. It is initially proposed as a regression version of Focal Loss (FL) [3]. The loss is mapped to a continuous weighting function, with samples having larger loss being assigned higher weights.

- *Distribution Smoothing*: LDS reversely uses the smoothed label distribution as a continuous label-wise weighting function, assigning larger weights to labels with low density. FDS smooths the feature distribution by KDE to mitigate the limitation of information contained in tail labels.

- *Loss adjustment*: To address the shortcomings of MSE in DIR, balanced-MSE is proposed to fit imbalanced situations. GMM-based Analytical Integrarion (GAI) and Batch-based Monte-Carlo (BMC) are two implementations of balanced-MSE. GAI tends to express the label distribution as a Gaussian Mixture Model (GMM), while BMC treats all labels in a training batch as random samples from the entire training set and adds a loss term for the label-sampling process. SupCR combines supervised contrastive learning with DIR by applying data augmentation to the batch to obtain a two-view batch, encouraging samples that are close in label space to be similar in feature space. Similarly, RankSim encourages features of samples from neighboring labels to have high similarity, ordering the features' similarity and labels' distance and enforcing the two orders to coincide.

- *ISDA-R*: The proposed regression version of ISDA, namely ISDA-R is

compared. The final optimization function for ISDA-R is as follows,

$$\begin{aligned}\mathcal{L}_{ISDA-R} = \sum_{b\in\mathcal{B}} \frac{1}{N_b} \sum_{i=1}^{N_b} \log(1 + \sum_{y_{b'}\neq y_i} \exp(\Delta Z_{y_i k}^{y_{b'}} \\ + \beta \boldsymbol{w}^\top \boldsymbol{\Sigma}_b \boldsymbol{w})),\end{aligned} \tag{27}$$

where $\Delta Z_{y_i}^{y_{b'}} = u_{y_{b'}}(\boldsymbol{z}_i) - u_{y_i}(\boldsymbol{z}_i)$. The ISDA-R uses the covariance matrix before W-FDS. ISDA-R doesn't consider the augmentation strength. The hyper-parameter β in ISDA-R follows the settings in [12].

- *RISDA-R*: The regression version of RISDA is also compared by remaining the original augmentation strategy and changing the CE loss into BMSE. RISDA separates the label space into two sets, a $\mathcal{H}$ indicating head labels containing majority of samples and a $\mathcal{T}$ indicating tail labels possessing few samples, and $\mathcal{B} = \mathcal{H} \cup \mathcal{T}$.

$$\begin{aligned}\mathcal{L}_{RISDA-R} = \sum_{b\in\mathcal{H}} \frac{1}{N_b} \sum_{i=1}^{N_b} \log(1 + \sum_{y_{b'}\neq y_i} \exp(\Delta Z_{y_i}^{y_{b'}})) \\ + \sum_{b\in\mathcal{T}} \frac{\gamma_b}{N_b} \sum_{i=1}^{N_b} \log(1 + \sum_{y_{b'}\neq y_i} \exp(\Delta Z_{y_i}^{y_{b'}} \\ + \alpha|\boldsymbol{w}^\top \hat{\boldsymbol{\mu}}_b - y_i| + \beta \boldsymbol{w}^\top \hat{\boldsymbol{\Sigma}}_b \boldsymbol{w})),\end{aligned} \tag{28}$$

where $\hat{\boldsymbol{\mu}}_b = \sum_{b'} \varepsilon_{bb'} \boldsymbol{\mu}_b$, $\hat{\boldsymbol{\Sigma}}_b = \boldsymbol{\Sigma}_b + \sum_{b'} \varepsilon_{bb'} \boldsymbol{\Sigma}_b$, and $\Delta Z_{y_i}^{y_{b'}} = u_{y_{b'}}(\boldsymbol{z}_i) - u_{y_i}(\boldsymbol{z}_i)$. RISDA-R has an augmentation strength $\gamma_b = (1-\gamma)/(1-\gamma^{N_b})$ on tail labels in $\mathcal{T}$ with γ a hyper-parameter. The choice of γ is set to 0.5 following the settings and the results of ablation experiments in [14].

4.3.2. Evaluation metrics

Following [22], the whole test set is separated into *many-shot region* (bins with > 100 training samples), *medium-shot region* (bins with 20 to 100 training samples), and *few-shot region* (bins with < 20 training samples).

Aprat from the widely used Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), we also use following metrics for specific cases.

Table 4: Benchmarking results on IMDB-WIKI-DIR.

Metrics	MAE ↓				GM ↓			
Shot	All	Many	Med.	Few	All	Many	Med.	Few
VANILLA	8.06	7.23	15.12	26.33	4.57	4.17	10.59	20.46
SMOTER [20]	8.14	7.42	14.15	25.28	4.64	4.30	9.05	19.46
SMOGN [21]	8.03	7.30	14.02	25.93	4.63	4.30	8.74	20.12
FOCAL-R [22]	7.97	7.12	15.14	26.96	4.49	4.10	10.37	21.20
SMOGN + LDS + FDS [22]	7.97	7.38	13.22	22.95	4.59	4.39	7.84	14.94
FOCAL-R + LDS + FDS [22]	7.88	7.10	14.08	25.75	4.47	4.11	9.32	18.67
BMC [23]	8.08	7.52	12.47	23.29	4.61	4.21	8.86	16.33
GAI [23]	8.12	7.58	12.27	23.05	4.59	4.27	9.11	16.42
SupCR(MSE) [47]	8.07	7.46	13.02	23.41	4.56	4.09	7.91	15.22
RankSim [28]	7.72	6.93	14.48	25.38	4.27	3.90	10.02	15.84
ISDA-R	7.32	7.04	12.13	22.87	4.16	3.85	7.63	14.29
RISDA-R	7.21	6.76	12.47	21.75	4.09	3.89	8.21	13.76
IRDA	7.09	**5.90**	12.87	20.99	4.06	3.99	7.95	13.06
Meta-IRDA	**6.94**	5.97	**11.76**	**20.73**	**3.70**	**3.55**	**7.42**	**12.94**
SOTA (Best) **vs.** VANILLA	+0.34	+0.30	+2.85	+3.38	+0.30	+0.27	+2.75	+5.52
OURS (Best) **vs.** VANILLA	+1.12	+1.33	+3.36	+5.60	+0.87	+0.62	+3.17	+7.52

- GM: The geometric mean proposed by [23] is defined as $\sqrt[N]{\Pi_{i=1}^{N}|y_i - \hat{y}_i|}$ characterizing the uniformity of model predictions.
- δ_1: The threshold accuracy is defined as the percentage of δ_1 such that $\max(d_i/g_i, g_i/d_i) = \delta_i < 1.25^i$ where g_i denotes the value of a pixel in the ground truth depth image and d_i represents the value of its corresponding pixel in the predicted depth image. δ_1 is used as a standard depth estimation evaluation metric.

4.4. Main Results

For five DIR benchmarks employed, the main results obtained are reported in this section.

4.4.1. Age inference

We evaluated the performance of our proposed method, IRDA, and Meta-IRDA, against existing SOTA methods using two benchmarks for age inference: AgeDB-DIR and IMDB-WIKI-DIR. The results for AgeDB-DIR and IMDB-WIKI-DIR are shown in Table 3 and Table 4 respectively. SMOTER and SMOGN exhibited the worst performance overall, even worse than the

Table 5: Benchmarking results on NYUD2-DIR.

Metrics	ReMSE ↓				δ_1 ↑			
Shot	All	Many	Med.	Few	All	Many	Med.	Few
VANILLA	1.477	0.591	0.952	2.123	0.677	0.777	0.693	0.570
VANILLA + LDS + FDS [22]	1.338	0.670	0.851	1.880	0.705	0.730	0.764	0.655
BMC [23]	1.283	0.787	0.870	1.736	0.694	0.622	0.806	0.723
GAI [23]	1.251	0.692	0.959	1.851	0.702	0.676	0.734	0.715
ISDA-R	1.179	0.586	0.881	1.749	0.772	**0.790**	0.739	0.781
RISDA-R	1.192	0.610	0.877	1.732	0.769	0.788	0.712	0.793
IRDA	1.096	0.514	**0.787**	1.700	0.797	0.777	**0.820**	0.809
Meta-IRDA	**1.045**	**0.499**	0.791	**1.697**	**0.805**	0.774	0.817	**0.839**
SOTA (Best) **vs.** VANILLA	+0.226	-0.079	+0.101	+0.387	+0.028	-0.047	+0.113	+0.153
OURS (Best) **vs.** VANILLA	+0.432	+0.092	+0.165	+0.423	+0.128	+0.013	+0.127	+0.269

vanilla model, though they slightly improved generalizability in few-shot regions. Focal-R showed improved overall performance, with a more distinct improvement in many-shot regions than in few-shot regions. Applying distribution smoothing strategies slightly enhance the effectiveness of Focal-R and SMOGN. Two implementations of Balanced-MSE demonstrated similar performance to Focal-R but were inferior to the distribution-smoothed Focal-R. SupCR and RankSim showed better performance among all SOTA methods.

Among the four implicit data augmentation methods, there was a significant promotion in few-shot regions compared to the vanilla model, with an increase of 38% and a 20% improvement over the best SOTA algorithm. Notably, Meta-IRDA achieved the best performance among all implicit data augmentation methods, with IRDA being the second best.

4.4.2. Depth inference

The NYUD2-DIR was used for the depth inference and the results are presented in Table 5. Due to the dataset's specific characteristics, which involve learning pixel-wise information for each image, methods like SMOTER and SMOGN cannot be easily employed. Hence, only the vanilla model, the distribution-smoothed vanilla model and two implements of Balanced-MSE are compared with our proposed methods. Among the existing SOTA methods, BMC showed the best performance in few-shot regions, whereas GAI demonstrated overall impressive performance. Nevertheless, the improvement brought by Meta-IRDA was even more significant, with an elevation of 29.25%, compared to the SOTA, which only showed a 15.3% overall improve-

Table 6: Benchmarking results on TUAB.

Metrics	MAE ↓				GM ↓			
Shot	All	Many	Med.	Few	All	Many	Med.	Few
VANILLA	8.96	7.45	10.05	14.77	6.62	5.38	8.24	11.92
SMOTER [20]	8.85	7.59	9.65	14.33	6.15	5.92	7.99	11.73
SMOGN [21]	8.79	7.51	9.97	14.11	6.09	5.43	8.06	11.81
FOCAL-R [22]	8.77	7.43	9.76	14.08	6.33	5.56	8.45	11.03
SMOGN + LDS + FDS [22]	8.64	7.53	9.72	13.81	6.03	5.37	8.04	10.35
FOCAL-R + LDS + FDS [22]	8.61	7.33	9.61	14.07	6.11	5.41	8.39	10.98
BMC [23]	8.72	7.29	9.50	13.74	6.26	5.47	8.12	10.37
GAI [23]	8.76	7.31	9.42	13.34	6.07	5.36	8.09	10.72
SupCR(MSE) [47]	8.58	7.22	9.58	13.42	6.03	5.29	8.17	10.67
RankSim [28]	8.41	7.18	9.39	13.86	5.93	4.97	8.76	11.02
ISDA-R	8.11	6.54	9.32	13.41	5.06	4.28	7.87	10.13
RISDA-R	8.09	6.90	9.19	13.04	5.07	4.31	7.71	9.98
IRDA	7.78	6.12	8.98	12.57	4.96	**4.09**	7.34	9.53
Meta-IRDA	**7.73**	**6.11**	**8.88**	**12.41**	**4.98**	4.11	**7.30**	**9.18**
SOTA (Best) **vs.** VANILLA	+0.55	+0.27	+0.63	+1.43	+0.69	+0.41	+0.20	+1.57
OURS (Best) **vs.** VANILLA	+1.23	+1.34	+1.17	+2.36	+1.66	+1.29	+0.94	+2.74

ment and performed worse than the vanilla model in many-shot regions.

4.4.3. Brain-age estimation

The TUAB benchmark was used for the brain-age estimation and the results are presented in Table 6. All of the compared methods were tested on TUAB, and among all existing state-of-the-art techniques, RankSim achieved the best overall performance. However, the implicit data augmentation strategies generally outperformed other strategies for DIR. In particular, the proposed IRDA and Meta-IRDA showed significant improvements in both metrics. Meta-IRDA achieved an overall improvement 13.73% compared to the vanilla model, whereas the existing SOTA only achieved a 6.28% improvement. In the few-shot regions, IRDA and Meta-IRDA showed improvements of 14.9% and 15.98%, respectively, compared to the existing SOTA's 9.68% improvement.

4.4.4. Temperature prediction

The SkyFinder dataset was used for temperature prediction and the results are presented in Table 7. Compared to SupCR, which is the optimal SOTA on SkyFinder, all implicit data augmentation methods outperformed

SupCR. Among all implicit augmentation methods, IRDA and Meta-IRDA further improved both the overall performance and performance in few-shot regions. Specifically, Meta-IRDA achieved a 27.92% improvement on the entire dataset and a 22.92% improvement in the few-shot regions, whereas the SOTA showed only a 6.17% improvement on the entire dataset and a 13.62% improvement in the few-shot regions.

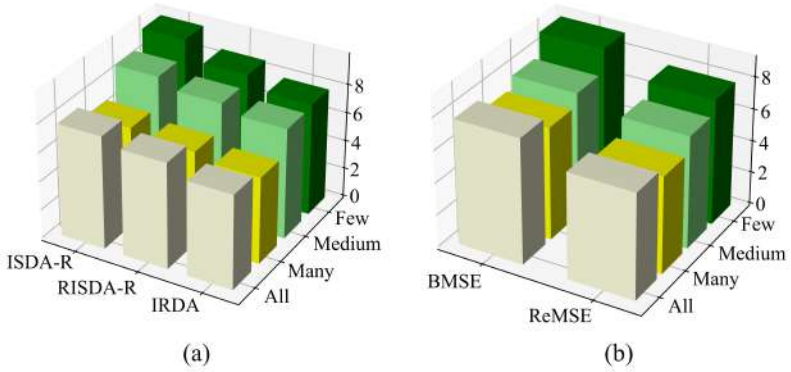


Figure 6: Results of ablation experiments comparing proposed augmentation strategy of IRDA with existing augmentation strategies shown in (a) and comparing ReMSE with existing BMSE and the typical MSE shown in (b).

4.5. Further Analysis

4.5.1. Ablation study

We conducted an ablation study to evaluate the effectiveness of the augmentation strategy and ReMSE separately. The study consisted of two parts: assessing the effectiveness of the augmentation strategies and ReMSE.

To evaluate the effectiveness of the augmentation strategy, we compared IRDA with ISDA-R and RISDA-R without using ReMSE on AgeDB-DIR. The results, shown in Fig. 6 (a), indicate that the augmentation strategy of IRDA achieves the best performance.

We further demonstrated the effectiveness of ReMSE by evaluating IRDA with and without ReMSE on five benchmarks. The results, shown in Fig. 6 (b), indicate that IRDA with ReMSE significantly outperforms IRDA without ReMSE.

Table 7: Benchmarking results on SkyFinder.

Metrics	MAE ↓				GM ↓			
Shot	All	Many	Med.	Few	All	Many	Med.	Few
VANILLA	3.08	2.11	4.15	6.24	2.55	1.90	3.76	5.02
SMOTER [20]	3.21	2.57	4.09	6.10	2.78	2.00	3.55	4.91
SMOGN [21]	3.29	2.66	4.02	6.02	2.79	2.03	3.47	4.88
FOCAL-R [22]	3.01	2.23	4.07	6.07	2.47	1.95	3.41	4.09
SMOGN + LDS + FDS [22]	3.05	2.58	3.91	5.39	2.39	1.87	3.33	4.23
FOCAL-R + LDS + FDS [22]	3.16	2.13	3.99	5.97	2.39	2.03	3.09	4.14
BMC [23]	3.17	2.64	3.86	5.86	2.52	1.99	3.05	3.89
GAI [23]	2.97	2.01	3.97	5.67	2.12	1.85	3.06	3.89
SupCR(MSE) [47]	2.89	1.91	3.81	5.72	2.03	1.79	2.95	3.71
RankSim [28]	2.99	2.05	3.88	5.69	2.15	1.89	3.00	3.80
ISDA-R	2.81	1.89	3.79	5.61	1.99	1.75	2.91	3.65
RISDA-R	2.87	1.92	3.85	5.24	2.02	1.81	2.93	3.41
IRDA	2.41	1.56	**3.21**	4.96	1.56	1.24	**2.31**	3.05
Meta-IRDA	**2.22**	**1.47**	3.29	**4.81**	**1.47**	**1.19**	2.32	**2.98**
SOTA (Best) **vs.** VANILLA	+0.19	+0.20	+0.34	+0.85	+0.52	+0.11	+0.81	+1.31
OURS (Best) **vs.** VANILLA	+0.86	+0.64	+0.94	+1.43	+1.08	+0.71	+1.45	+2.04

4.5.2. Sensitive test

There are three hyper-parameters in IRDA: α, β and v_b. v_b is a hyper-parameter that can be directly calculated or learned through mete-learning. Therefore, conducted a sensitivity test on α and β and discussed v_b.

A sensitivity test was conducted for α and β. The values of α and β were selected from $\{0.25, 0.50, 0.75, 1.00, 1.25, 1.50\}$ for our proposed IRDA and Meta-IRDA on AgeDB-DIR. The values of MAE for many-shot, medium-shot, and few-shot categories were compared. For many shots, MAE achieved the minimum value at $\alpha = 0.50$ and $\beta = 0.50$. For medium shots, MAE achieved the minimum value at $\alpha = 0.50$ and $\beta = 0.75$. For few shots, MAE achieved the minimum value at $\alpha = 0.50$ and $\beta = 1.00$. Overall, MAE achieved the minimum value at $\beta = 0.75$.

As mentioned in Section IV.B, the estimations of feature statistics can be inaccurate. We use a class-wise parameter v_b to adjust the covariance matrix. To investigate the effect of v_b, we plotted the variation of v_b learned by the meta network during the entire training process for different regions. As shown in Fig. 7(b), v_b sharply decreases at the beginning of the training process for a head label and rapidly converges to 0. For a tail label, v_b

gently decreases and hardly reaches 0 before the end of training, as shown in Fig. 7(a). Our investigation indicates that samples from tail labels are more likely to have correlations between features.

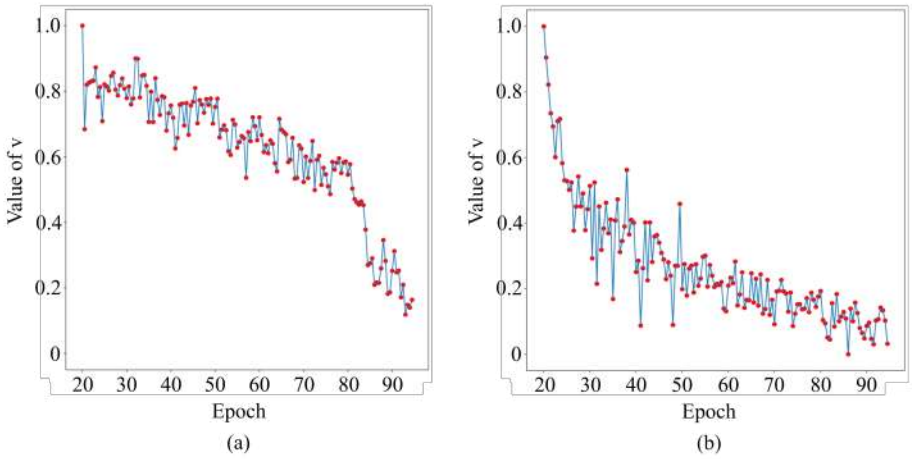


Figure 7: The variation of hyper-parameter v_b during a whole training process. Two labels are discussed, including a tail label and a head label.

4.6. Future Work

Learning continuous targets differs significantly from learning discrete target. In deep learning, a common approach is to convert the continuous target $\mathcal{Y}$ into B discrete bins. According to related works [22, 23, 39, 40, 42, 43, 44], we have adhered to the general settings of the length of bins for imbalanced regression, which represents the smallest granularity of the continuous labels. We are also interested in how the length of bins could influence the optimization of deep regression tasks. We plan to explore the discretization of continuous labels as a future direction.

5. Conclusion

Our study proposes a novel approach to investigate DIR tasks using implicit data augmentation strategies. First, we suggest a new augmentation strategy that incorporates information from neighboring labels during augmentation. Then, we introduce a novel regression loss function, ReMSE, which accounts for both imbalanced label distribution and skewed feature distribution. Our proposed method, IRDA, integrates the augmentation

strategy and the loss function into an easy-to-compute objective function. We provide two implementations of IRDA: a direct estimation approach and a meta-learning-based approach, Meta-IRDA. We offer a regularization analysis and a theoretical examination of the rationality of IRDA. Through extensive experiments, we compare existing SOTA methods with our proposed IRDA and Meta-IRDA. The results demonstrate the superiority of IRDA and Meta-IRDA. Our work presents a new approach to solving DIR tasks.

References

[1] N. A. Azhar, M. S. M. Pozi, A. M. Din, A. Jatowt, An investigation of smote based methods for imbalanced datasets with data complexity analysis, IEEE Transactions on Knowledge and Data Engineering 35 (7) (2023) 6651–6672.

[2] N. Chawla, K. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (1) (2002) 321–357.

[3] T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 99 (2017) 2999–3007.

[4] Z. Linbin, L. Xiangguang, M. Xiaojie, J. Kefeng, K. Gangyao, L. Li, Data distribution loss for imbalanced SAR vehicle target recognition, IEEE Geoscience and Remote Sensing Letters 21 (2024) 1–5.

[5] S. Qiu, X. Cheng, H. Lu, H. Zhang, R. Wan, X. Xue, J. Pu, Subclassified loss: Rethinking data imbalance from subclass perspective for semantic segmentation, IEEE Transactions on Intelligent Vehicles 9 (2024) 1547–1558.

[6] J. Cui, S. Liu, Z. Tian, Z. Zhong, J. Jia, Reslt: Residual learning for long-tailed recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 3695–3706.

[7] L. Wei, C. Jinlin, C. Jiannong, M. Chao, W. Jia, C. Xiaohui, C. Ping, Eid-gan: Generative adversarial nets for extremely imbalanced data augmentation, IEEE Transactions on Industrial Informatics 19 (3) (2023) 3208–3218.

[8] D. Hongwei, H. Nana, W. Yaoxin, C. Xiaohui, Legan: Addressing intra-class imbalance in gan-based medical image augmentation for improved imbalanced data classification, IEEE Transactions on Instrumentation and Measurement 73 (2024) 1–14.

[9] H. Song, M. Kim, J. Lee, Toward robustness in multi-label classification: A data augmentation strategy against imbalance and noise, in: Association for the Advancement of Artificial Intelligence, 2024, pp. 21592–21601.

[10] X. Ren, W. Lin, X. Yang, X. Yu, H. Gao, Data augmentation in defect detection of sanitary ceramics in small and non-i.i.d datasets, IEEE Transactions on Neural Networks and Learning Systems 33 (1) (2022) 1–10.

[11] Z. Hao, C. Ying, Y. Dong, H. Su, J. Song, J. Zhu, GSmooth: Certified robustness against semantic transformations via generalized randomized smoothing, in: Proceedings of the 39th International Conference on Machine Learning, Vol. 162, 2022, pp. 8465–8483.

[12] Y. Wang, X. Pan, S. Song, H. Zhang, G. Huang, C. Wu, Implicit semantic data augmentation for deep networks, in: Advances in Neural Information Processing Systems, Vol. 32, 2019, pp. 3320–3329.

[13] S. Li, K. Gong, C. H. Liu, Y. Wang, F. Qiao, X. Cheng, Metasaug: Meta semantic augmentation for long-tailed visual recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5208–5217.

[14] X. Chen, Y. Zhou, D. Wu, W. Zhang, Y. Zhou, B. Li, W. Wang, Imagine by reasoning: A reasoning-based implicit semantic data augmentation for long-tailed classification, in: Association for the Advancement of Artificial Intelligence, 2022, pp. 356–364.

[15] C.-Y. Low, A. Beng-Jin Teoh, An implicit identity-extended data augmentation for low-resolution face representation learning, IEEE Transactions on Information Forensics and Security 17 (2022) 3062–3076.

[16] Z. Chen, J. Zhang, P. Wang, J. Chen, J. Li, When active learning meets implicit semantic data augmentation, in: European Conference on Computer Vision, Cham, 2022, pp. 56–72.

[17] W. Li, H. Guo, H. Dong, M. Tang, Y. Zhou, J. Wang, Bi-level implicit semantic data augmentation for vehicle re-identification, IEEE Transactions on Intelligent Transportation Systems 24 (4) (2023) 4364–4376.

[18] K. Seo, H. Cho, D. Choi, J.-D. Park, Implicit semantic data augmentation for hand pose estimation, IEEE Access 10 (2022) 84680–84688.

[19] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, F. Shen, Image data augmentation for deep learning: A survey, ArXiv abs/2204.08610 (2022).

[20] L. Torgo, R. P. Ribeiro, B. Pfahringer, P. Branco, Smote for regression, in: Portuguese Conference on Artificial Intelligence, 2013, pp. 1–12.

[21] P. Branco, L. Torgo, R. P. Ribeiro, Smogn: a pre-processing approach for imbalanced regression, in: First international workshop on learning with imbalanced domains: Theory and applications, 2017, pp. 36–50.

[22] Y. Yang, K. Zha, Y. Chen, H. Wang, D. Katabi, Delving into deep imbalanced regression, in: Proceedings of the 38th International Conference on Machine Learning, Vol. 139, 2021, pp. 11842–11851.

[23] J. Ren, M. Zhang, C. Yu, Z. Liu, Balanced mse for imbalanced visual regression, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7926–7935.

[24] O. Wu, Rethinking class imbalance in machine learning, arxiv:2305.03900 (2023).

[25] S. Hwang, S. E. Whang, Mixrl: Data mixing augmentation for regression using reinforcement learning, arxiv:2106.03374 (2021).

[26] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, S. Kumar, Long-tail learning via logit adjustment, in: 9th International Conference on Learning Representations, 2022, pp. 1–27.

[27] E. Parzen, On estimation of a probability density function and mode, The annals of mathmatical statistics 33 (3) (1962) 1065–1076.

[28] Y. Gong, G. Mori, F. Tung, RankSim: Ranking similarity regularization for deep imbalanced regression, in: Proceedings of the 39th International Conference on Machine Learning, Vol. 162, 2022, pp. 7634–7649.

[29] K. Zha, P. Cao, Y. Yang, D. Katabi, Supervised contrastive regression, arxiv:2210.01189 (2022).

[30] F. Dubost, G. Bortsova, H. Adams, M. A. Ikram, W. Niessen, M. Vernooij, M. de Bruijne, Hydranet: Data augmentation for regression neural networks, in: Medical Image Computing and Computer Assisted Intervention, 2019, pp. 438–446.

[31] S. Stocksieker, D. Pommeret, A. Charpentier, Data augmentation for imbalanced regression, in: Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, Vol. 206, 2023, pp. 7774–7799.

[32] Z. Chen, Y. Fu, Y. Zhang, Y.-G. Jiang, X. Xue, L. Sigal, Multi-level semantic feature augmentation for one-shot learning, IEEE Transactions on Image Processing 28 (9) (2019) 4594–4605.

[33] D. Nix, A. Weigend, Estimating the mean and variance of the target probability distribution, in: Proceedings of IEEE International Conference on Neural Networks, 1994, pp. 55–60.

[34] P. McCullagn, J. A. Nelder, Generalized linear models, European Journal of Operational Research 16 (13) (1984) 285–292.

[35] D. Zeiberg, S. Jain, P. Radivojac, Leveraging structure for improved classification of grouped biased data, in: Association for the Advancement of Artificial Intelligence, 2023, pp. 1–14.

[36] G. Mel, S. Ganguli, A theory of high dimensional regression with arbitrary correlations between input features and target functions: sample complexity, multiple descent curves and a hierarchy of phase transitions, in: Proceedings of the 38th International Conference on Machine Learning, 2021, pp. 7578–7587.

[37] A. Bardes, J. Ponce, Y. LeCun, VICReg: Variance-invariance-covariance regularization for self-supervised learning, in: International Conference on Learning Representations, 2022, pp. 1–23.

[38] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, D. Meng, Meta-weight-net: Learning an explicit mapping for sample weighting, in: Advances in Neural Information Processing Systems, Vol. 32, 2019, pp. 3388–3398.

[39] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, S. Zafeiriou, Agedb: The first manually collected, in-the-wild age database, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 5977–5896.

[40] R. Rothe, R. Timofte, L. V. Gool, Deep expectation of real and apparent age from a single image without facial landmarks, International Journal of Computer Vision 126 (2018) 144–157.

[41] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: European Conference on Computer Vision, 2012, pp. 1–14.

[42] D. A. Engemann, A. Mellot, R. Höchenberger, H. Banville, D. Sabbagh, L. Gemein, T. Ball, A. Gramfort, A reusable benchmark of brain-age prediction from m/eeg resting-state signals, NeuroImage 262 (2022) 119521.

[43] R. P. Mihail, S. Workman, Z. Bessinger, N. Jacobs, Sky segmentation in the wild: An empirical study, in: IEEE Winter Conference on Applications of Computer Vision, 2016, pp. 1–6.

[44] W.-T. Chu, K.-C. Ho, A. Borji, Visual weather temperature prediction, in: IEEE Winter Conference on Applications of Computer Vision, 2018, pp. 234–241.

[45] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, IEEE Conference on Computer Vision and Pattern Recognition (2016) 770–778.

[46] J. Hu, M. Ozay, Y. Zhang, T. Okatani, Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries, in: IEEE Winter Conference on Applications of Computer Vision, 2019, pp. 1043–1051.

[47] K. Zha, P. Cao, Y. Yang, D. Katabi, Supervised contrastive regression, arxiv:2210.01189 (2022).

[48] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arxiv:1412.6980 (2014).