Boosting Model Resilience via Implicit Adversarial Data Augmentation

Xiaoling Zhou[1] , **Wei Ye**[1*] , **Zhemg Lee**[2] , **Rui Xie**[1] and **Shikun Zhang**[1*]
[1]National Engineering Research Center for Software Engineering, Peking University, China
[2]Tianjin University, Tianjin, China
xiaolingzhou@stu.pku.edu.cn, wye@pku.edu.cn, zhemglee@tju.edu.cn, {ruixie,zhangsk}@pku.edu.cn

Abstract

Data augmentation plays a pivotal role in enhancing and diversifying training data. Nonetheless, consistently improving model performance in varied learning scenarios, especially those with inherent data biases, remains challenging. To address this, we propose to augment the deep features of samples by incorporating their adversarial and anti-adversarial perturbation distributions, enabling adaptive adjustment in the learning difficulty tailored to each sample's specific characteristics. We then theoretically reveal that our augmentation process approximates the optimization of a surrogate loss function as the number of augmented copies increases indefinitely. This insight leads us to develop a meta-learning-based framework for optimizing classifiers with this novel loss, introducing the effects of augmentation while bypassing the explicit augmentation process. We conduct extensive experiments across four common biased learning scenarios: long-tail learning, generalized long-tail learning, noisy label learning, and subpopulation shift learning. The empirical results demonstrate that our method consistently achieves state-of-the-art performance, highlighting its broad adaptability.

1 Introduction

Data augmentation techniques, designed to enrich the quantity and diversity of training samples, have demonstrated their effectiveness in improving the performance of deep neural networks (DNNs) [Maharana *et al.*, 2022]. Existing methods can be divided into two categories. The first, explicit data augmentation [Cubuk *et al.*, 2020; Xu and Zhao, 2023], applies geometric transformations to samples and directly integrates augmented instances into the training process, leading to reduced training efficiency. The second, a recent addition to this field, is the implicit data augmentation technique [Wang *et al.*, 2019], which is inspired by the existence of numerous semantic vectors within the deep feature space of DNNs. This method emphasizes enhancing the deep features of samples and is achieved by optimizing robust losses instead of explicitly conducting the augmentation process, resulting in a more efficient and effective approach. Subsequent studies in imbalanced learning have extended this approach. For example, MetaSAug [Li *et al.*, 2021] refines the accuracy of covariance matrices for minor classes by minimizing losses on a balanced validation set. RISDA [Chen *et al.*, 2022] generates diverse instances for minor classes by extracting semantic vectors from the deep feature space of both the current class and analogous classes.

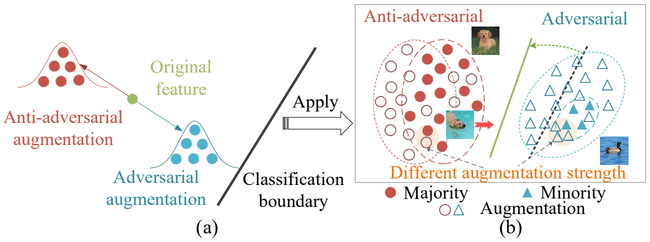


Figure 1: (a) Illustration for our augmentation strategy, which augments samples within their adversarial and anti-adversarial perturbation distributions. (b) Illustration of an imbalanced learning scenario. Our method employs adversarial and anti-adversarial augmentations for the minor and major classes, respectively.

Despite these promising efforts, current augmentation strategies still exhibit notable limitations. Firstly, these approaches mainly enhance samples within the original training data space [Xu and Zhao, 2023; Wang *et al.*, 2019], falling short in effectively mitigating the distributional discrepancies between training and test data, such as noise [Zheng *et al.*, 2021] and subpopulation shifts [Yao *et al.*, 2022]. Secondly, previous algorithms [Li *et al.*, 2021; Chen *et al.*, 2022] mostly operate at the category level, causing different samples within the same class to share identical augmentation distributions and strengths, which is potentially unreasonable and inaccurate. For instance, in noisy learning scenarios, it is advisable to treat noisy samples separately to mitigate their negative impact on model training [Zhou *et al.*, 2023].

This study introduces a novel **I**mplicit **A**dversarial **D**ata **A**ugmentation (IADA) approach, which conceptually embodies two main characteristics to address the two bottlenecks above. Firstly, as illustrated in Fig. 1(a), IADA enriches the deep features of samples by randomly sampling perturbation vectors from their adversarial and anti-adversarial perturbation distributions, surpassing the limitation of augmenting

[*]Corresponding authors.

within the original training distribution. With this strategy, the classifier is anticipated to be dynamically adjusted by modifying the learning difficulty of samples. Secondly, the augmentation distribution for each sample is tailored based on its unique training characteristics, granting it the ability to address data biases beyond the category level. Specifically, these distributions are modeled as multivariate normal distributions, characterized by sample-wise perturbations and class-specific covariance matrices. Fig. 1(b) demonstrates the two features of our method using imbalanced learning as an example scenario, where the classifier typically favors major classes while underperforming on the minor ones, as well as favors easy samples (e.g., dogs on grass) but commonly mispredicting hard samples (e.g., dogs in the water)[1]. In this scenario, our method adopts anti-adversarial augmentation for major classes and adversarial augmentation (with higher strength) for minor ones. Meanwhile, adversarial augmentation is also applied to hard samples within major classes, ultimately facilitating better learning of class boundaries.

By exploring an infinite number of augmentations, we theoretically derive a surrogate loss for our augmentation strategy, thereby eliminating the need for explicit augmentation. Subsequently, to determine the perturbation strategies for samples in this loss, we construct a meta-learning-based framework, in which a perturbation network is tasked with computing perturbation strategies by leveraging diverse training characteristics extracted from the classifier as inputs. The training of all parameters in the perturbation network is guided by a small, unbiased meta dataset, enabling the generation of well-founded perturbation strategies for samples. Extensive experiments show that our method adeptly addresses various data biases, such as noise, imbalance, and subpopulation shifts, consistently achieving state-of-the-art (SOTA) performance among all compared methods.

In summary, our contributions in this paper are threefold.

- We propose a novel perspective on data augmentation, wherein samples undergo augmentation within their adversarial and anti-adversarial perturbation distributions, to facilitate model training across diversified learning scenarios, particularly those with data biases.
- In accordance with our augmentation perspective, we derive a new logit-adjusted loss and incorporate it into a well-designed meta-learning framework to optimize the classifier, unlocking the potential of data augmentation without an explicit augmentation procedure.
- We conduct extensive experiments across four typical biased learning scenarios, encompassing long-tail (LT) learning, generalized long-tail (GLT) learning, noisy label learning, and subpopulation shift learning. The results conclusively demonstrate the effectiveness and broad applicability of our approach.

2 Related Work

Data Augmentation methods have showcased their capacity to improve DNNs' performance by expanding and diversifying training data [Maharana *et al.*, 2022]. Explicit augmentation directly incorporates augmented data into the training process, albeit at the expense of reduced training efficiency [Cubuk *et al.*, 2020; Taylor and Nitschke, 2018; Xu and Zhao, 2023]. Recently, Wang et al. [2019] introduced an implicit semantic data augmentation approach, named ISDA, which transforms the deep features of samples within the semantic space of DNNs and boils down to the optimization of a robust loss. Subsequent studies [Li *et al.*, 2021; Chen *et al.*, 2022] in image classification tasks have extended this approach. However, these methods still struggle with effectively improving model performance when dealing with data biases that go beyond the category level.

Adversarial and Anti-Adversarial Perturbations transform samples in directions that respectively move towards and away from the decision boundary, thereby modifying samples' learning difficulty [Lee *et al.*, 2023; Zhou *et al.*, 2023]. Consequently, models allocate varying levels of attention to samples subjected to their perturbations. Research has confirmed that incorporating adversarial and anti-adversarial samples during training assists models in achieving a better tradeoff between robustness and generalization [Zhou *et al.*, 2023; Zhu *et al.*, 2021]. However, existing adversarial training methods primarily focus on two specific types of perturbations that maximize and minimize losses [Xu *et al.*, 2021; Zhou *et al.*, 2023], posing limitations. Moreover, generating adversarial perturbations within the input space is time-consuming [Madry *et al.*, 2018]. Different from prior studies, our approach randomly selects perturbation vectors from both adversarial and anti-adversarial perturbation distributions, enabling the generation of multiple distinct adversarial and anti-adversarial samples. Furthermore, the perturbations are generated within the deep feature space, enhancing efficiency and ensuring universality across various data types.

3 Implicit Adversarial Data Augmentation

We initially introduce a sample-wise adversarial data augmentation strategy to facilitate model training across various learning scenarios. By considering infinite augmentations, we then derive a surrogate loss for our augmentation strategy.

3.1 Adversarial Data Augmentation

Consider training a deep classifier $\mathcal{F}$, with weights $\mathbf{\Phi}$ on a training set, denoted as $\mathcal{D}^{tr} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, where N refers to the number of training samples, and $y_i \in \{1, \cdots, \mathcal{C}\}$ represents the label of sample $\boldsymbol{x}_i$. The deep feature (before logit) learned by $\mathcal{F}$ for $\boldsymbol{x}_i$ is represented as a $\mathcal{H}$-dimensional vector $\boldsymbol{h}_i = \mathcal{F}_{\mathbf{\Phi}}(\boldsymbol{x}_i) \in \mathbb{R}^{\mathcal{H}}$.

Our augmentation strategy enhances samples within the deep feature space of DNNs. The perturbation vectors for the deep feature of each sample are randomly extracted from either its adversarial or anti-adversarial perturbation distributions. These distributions are modeled as multivariate normal distributions, $\mathcal{N}(\boldsymbol{\delta}_i, \mathbf{\Sigma}_{y_i})$, where $\boldsymbol{\delta}_i$ refers to the sample perturbation, and $\mathbf{\Sigma}_{y_i}$ represents the class-specific covariance matrix estimated from the features of all training samples in class y_i. As samples undergo augmentation within the deep feature space, perturbations should also be generated within

[1] This example stems from the observation that the majority of dogs in the training set are on grass and seldom in the water.

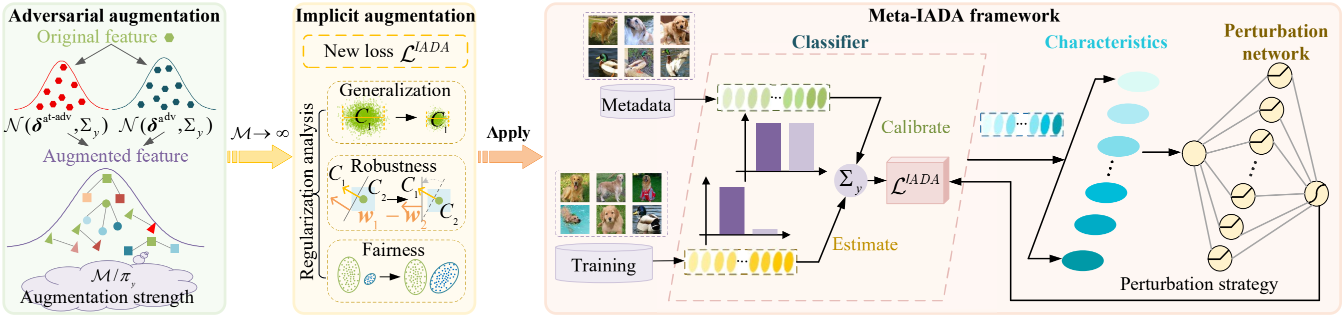


Figure 2: The overview of our method pipeline. We initiate with a sample-wise adversarial data augmentation strategy (Box 1), enriching the deep features of samples using perturbation vectors extracted from their adversarial and anti-adversarial perturbation distributions. Subsequently, by considering an infinite number of augmented instances, we derive a novel robust loss, termed IADA (Box 2). Regularization analysis reveals the efficacy of IADA in improving model generalization, robustness, and inter-class fairness. To facilitate optimization with IADA, we then establish a meta-learning-based framework called Meta-IADA (Box 3). Within it, a perturbation network is tasked with generating perturbation strategies for samples (denoted as $\epsilon_{\boldsymbol{x}}$) in the IADA loss, leveraging a set of $(K=15)$ training characteristics as inputs.

this space, facilitating semantic alterations for training samples. Consequently, the perturbation vector $\boldsymbol{\delta}_i$ for sample $\boldsymbol{x}_i$ is calculated as $\epsilon_i{\cdot}sign(\nabla_{\boldsymbol{h}_i}\ell_i^{CE})$, where $sign(\nabla_{\boldsymbol{h}_i}\ell_i^{CE})$ signifies the gradient sign of the CE loss ℓ_i^{CE} with respect to $\boldsymbol{h}_i$. The parameter ϵ_i plays a pivotal role in determining the perturbation strategy applied to $\boldsymbol{x}_i$, encompassing both the perturbation direction and bound. Its positive or negative sign signifies adversarial or anti-adversarial perturbations, respectively. Furthermore, the absolute value $|\epsilon_i|$ governs the perturbation bound. In practical applications, the value of ϵ_i is dynamically computed through a perturbation network based on the training characteristics of $\boldsymbol{x}_i$, which will be elaborated in Section 4. Additionally, the class-specific covariance matrix $\boldsymbol{\Sigma}_{y_i}$ within this distribution aids in preserving the covariance structure of each class. Its value is estimated in real-time by aggregating statistics from all mini-batches, as detailed in Section I of the Appendix. Regarding the augmentation strength quantified by the number of augmented instances $\mathcal{M}_i$ for $\boldsymbol{x}_i$, we define $\mathcal{M}_i$ as $\mathcal{M}/\pi_{y_i}$, where $\mathcal{M}$ is a constant and π_{y_i} represents the proportion of class y_i in the training data. Accordingly, a smaller proportion results in a larger number of augmented instances, ensuring class balance.

To compute the augmented features $\tilde{\boldsymbol{h}}_i$ from $\boldsymbol{h}_i$, we transform $\boldsymbol{h}_i$ along random directions sampled from $\mathcal{N}(\boldsymbol{\delta}_i,\boldsymbol{\Sigma}_{y_i})$. This transformation yields $\tilde{\boldsymbol{h}}_i\sim\mathcal{N}(\boldsymbol{h}_i+\boldsymbol{\delta}_i,\alpha\boldsymbol{\Sigma}_{y_i})$, where the parameter α controls the extent of dispersion for augmented samples. In summary, our adversarial data augmentation strategy offers the following advantages:

- Instead of augmenting samples within the original data space, our approach enhances them within their adversarial and anti-adversarial perturbation distributions. This method effectively adjusts the learning difficulty distribution of training samples, fostering improved generalization and robustness in DNNs.
- Our sample-wise augmentation distribution customizes the mean vector based on the unique training characteristics of each sample. This personalized strategy significantly enhances models' ability to address data biases, encompassing those beyond the category level.

3.2 IADA Loss

With our augmentation strategy, a straightforward way to train a classifier involves augmenting each $\boldsymbol{h}_i$ for $\mathcal{M}_i$ times. This procedure generates an augmented feature set for each sample, $\{\tilde{\boldsymbol{h}}_i^1,\cdots,\tilde{\boldsymbol{h}}_i^{\mathcal{M}_i}\}$. Subsequently, the CE loss for all augmented samples is as follows:

$$\mathcal{L}^{\mathcal{M}}(\boldsymbol{W},\boldsymbol{b})=\frac{1}{\hat{\mathcal{M}}}\sum_{i=1}^{N}\sum_{k=1}^{\mathcal{M}_i}-\log\frac{e^{\boldsymbol{w}_{y_i}^T\tilde{\boldsymbol{h}}_i^k+b_{y_i}}}{\sum_{j=1}^{\mathcal{C}}e^{\boldsymbol{w}_j^T\tilde{\boldsymbol{h}}_i^k+b_j}},\tag{1}$$

where $\hat{\mathcal{M}}=\sum_{i=1}^{N}\mathcal{M}_i$. Additionally, $\boldsymbol{W}=[\boldsymbol{w}_1,\cdots,\boldsymbol{w}_{\mathcal{C}}]^T\in\mathbb{R}^{\mathcal{C}\times\mathcal{H}}$ and $\boldsymbol{b}=[b_1,\cdots,b_{\mathcal{C}}]^T\in\mathbb{R}^{\mathcal{C}}$, in which $\boldsymbol{w}_j$ and b_j refer to the weight vector and bias corresponding to the last fully connected layer for class j. Considering augmenting more data while enhancing training efficiency, we explore augmenting an infinite number of times for the deep feature of each training sample. As $\mathcal{M}$ in $\mathcal{M}_i$ approaches infinity, the expected CE loss is expressed as:

$$\mathcal{L}^{\mathcal{M}\to\infty}(\boldsymbol{W},\boldsymbol{b})=\sum_{i=1}^{N}\frac{1}{\pi_{y_i}}\mathrm{E}_{\tilde{\boldsymbol{h}}_i}\big[-\log\frac{e^{\boldsymbol{w}_{y_i}^T\tilde{\boldsymbol{h}}_i+b_{y_i}}}{\sum_{j=1}^{\mathcal{C}}e^{\boldsymbol{w}_j^T\tilde{\boldsymbol{h}}_i+b_j}}\big].\tag{2}$$

However, accurately calculating Eq. (2) poses a challenge. Hence, we proceed to derive a computationally efficient surrogate loss for it. Given the concavity of the logarithmic function $\log(\cdot)$, using Jensen's inequality, $\mathrm{E}[\log X]\leq\log\mathrm{E}[X]$, we derive an upper bound of Eq. (2) as follows:

$$\mathcal{L}^{\mathcal{M}\to\infty}(\boldsymbol{W},\boldsymbol{b})\leq\sum_{i=1}^{N}\frac{1}{\pi_{y_i}}\log\Bigg(\sum_{j=1}^{\mathcal{C}}\mathrm{E}_{\tilde{\boldsymbol{h}}_i}\big[e^{\Delta\boldsymbol{w}_{j,y_i}\tilde{\boldsymbol{h}}_i+\Delta b_{j,y_i}}\big]\Bigg),\tag{3}$$

where $\Delta\boldsymbol{w}_{j,y_i}=\boldsymbol{w}_j^T-\boldsymbol{w}_{y_i}^T$ and $\Delta b_{j,y_i}=b_j-b_{y_i}$. As $\tilde{\boldsymbol{h}}_i$ is a Gaussian random variable adhering to $\mathcal{N}(\boldsymbol{h}_i+\boldsymbol{\delta}_i,\alpha\boldsymbol{\Sigma}_{y_i})$, we know that $\Delta\boldsymbol{w}_{j,y_i}\tilde{\boldsymbol{h}}_i+\Delta b_{j,y_i}\sim\mathcal{N}\big(\Delta\boldsymbol{w}_{j,y_i}(\boldsymbol{h}_i+\boldsymbol{\delta}_i)+\Delta b_{j,y_i},\alpha\Delta\boldsymbol{w}_{j,y_i}\boldsymbol{\Sigma}_{y_i}\Delta\boldsymbol{w}_{j,y_i}^T\big)$. Subsequently, utilizing the moment-generating function,

$$\mathrm{E}\left[e^{tX}\right]=e^{t\mu+\frac{1}{2}\sigma^2t^2},\quad X\sim\mathcal{N}(\mu,\sigma^2),\tag{4}$$

Name	Term	Function
$\mathcal{G}$	$\sum_{i=1}^{N}\sum_{j\neq y_i} q_{i,j}\rho_i^j$	Generalization
$\mathcal{R}$	$\sum_{i=1}^{N}\sum_{j\neq y_i} q_{i,j}\Delta \boldsymbol{w}_{j,y_i}\boldsymbol{\delta}_i$	Robustness
$\mathscr{F}$	$\sum_{i=1}^{N}\sum_{j\neq y_i} q_{i,j}\log(\pi_j/\pi_{y_i})$	Inter-class fairness

Table 1: The regularization terms incorporated in the IADA loss and their corresponding functions. $\boldsymbol{q}_i$ denotes the Softmax output linked to sample $\boldsymbol{x}_i$. The derivation process is presented in the Appendix.

the upper bound of Eq. (3) can be represented as

$$\mathcal{L}^{\mathcal{M}\to\infty}(\boldsymbol{W},\boldsymbol{b}) \leq \sum_{i=1}^{N} -\frac{1}{\pi_{y_i}}\log\frac{e^{\mathcal{Z}_i^{y_i}}}{\sum_{j=1}^{\mathcal{C}} e^{\mathcal{Z}_i^{j}}}, \tag{5}$$

where $\mathcal{Z}_i^j = \boldsymbol{w}_j(\boldsymbol{h}_i + \boldsymbol{\delta}_i) + b_j + \alpha\rho_i^j$ and $\rho_i^j = \frac{1}{2}\Delta\boldsymbol{w}_{j,y_i}\boldsymbol{\Sigma}_{y_i}\Delta\boldsymbol{w}_{j,y_i}^T$.

Drawing inspiration from the Logit Adjustment (LA) approach [Menon *et al.*, 2021], we introduce its logit adjustment term in place of the class-wise weight $1/\pi_{y_i}$, providing a more effective solution for imbalanced class distributions. Consequently, the final IADA loss is formulated as follows:

$$\mathcal{L}^{IADA}(\boldsymbol{W},\boldsymbol{b}) := \sum_{i=1}^{N} -\log\frac{e^{\tilde{\mathcal{Z}}_i^{y_i}}}{\sum_{j=1}^{\mathcal{C}} e^{\tilde{\mathcal{Z}}_i^{j}}}, \tag{6}$$

where $\tilde{\mathcal{Z}}_i^j = \boldsymbol{w}_j(\boldsymbol{h}_i + \boldsymbol{\delta}_i) + b_j + \alpha\rho_i^j + \beta\log\pi_j$. The symbols α and β serve as two hyperparameters in the IADA loss, frequently set to 0.5 and 1, respectively, in practical applications. The IADA loss, essentially a logit-adjusted variant of the CE loss, operates as a surrogacy for our proposed adversarial data augmentation strategy. Therefore, instead of explicitly executing the augmentation process, we can directly optimize this loss, thereby improving efficiency.

We further explain the IADA loss, defined in Eq. (6), from a regularization perspective using the Taylor expansion. This process reveals three regularization terms stemming from our adversarial data augmentation strategy, as outlined in Table 1. A detailed analysis of these terms is presented in Section III of the Appendix. Through our analysis, the first term $\mathcal{G}$ diminishes the mapped variances of deep features within each class, thereby enhancing intra-class compactness and improving the generalization ability of models. The second term $\mathcal{R}$ reinforces model robustness by increasing the cosine similarity between the classification boundary and the gradient vectors of adversarial samples. Furthermore, the third term $\mathscr{F}$ promotes fairness among classes by favoring less-represented categories. Collectively, the regularization terms incorporated in the IADA loss significantly contribute to strengthening the generalization, robustness, and inter-class fairness of DNNs, as depicted in Fig. 2 (Box 2).

4 Optimization Using IADA Loss

Due to the necessity of pre-determining the value of ϵ_i in $\boldsymbol{\delta}_i$, which governs the perturbation strategies, when utilizing the IADA loss, we construct a meta-learning-based framework named Meta-IADA. This framework is designed to optimize classifiers using the IADA loss. As illustrated in Fig. 2 (Box 3), Meta-IADA comprises four main components: the classifier, the characteristics extraction module, the perturbation network, and the meta-learning-based learning strategy.

Algorithm 1: Algorithm of Meta-IADA

```
Input: Training data $\mathcal{D}^{tr}$, metadata $\mathcal{D}^{meta}$, batch
       sizes $n$ and $m$, ending iterations $\mathcal{T}_1$ and $\mathcal{T}_2$.
Output: Learned classifier parameter $\boldsymbol{\Phi}$.
Initialize classifier parameter $\boldsymbol{\Phi}^1$ and perturbation
  network parameter $\boldsymbol{\Omega}^1$;
for $t \leq \mathcal{T}_1$ do
    Sample $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $\mathcal{D}^{tr}$;
    Update $\boldsymbol{\Phi}^{t+1} \leftarrow \boldsymbol{\Phi}^t - \eta_1\frac{1}{n}\sum_{i=1}^n \nabla_{\boldsymbol{\Phi}}\ell_i^{CE}$;
end
for $\mathcal{T}_1 < t \leq \mathcal{T}_2$ do
    Sample $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $\mathcal{D}^{tr}$;
    Sample $\{(\boldsymbol{x}_i^{meta}, y_i^{meta})\}_{i=1}^m$ form $\mathcal{D}^{meta}$;
    Obtain current covariance matrices $\boldsymbol{\Sigma}$;
    Formulate $\overline{\boldsymbol{\Phi}}^{t+1}$ by Eq. (7);
    Update $\boldsymbol{\Omega}^{t+1}$ by Eq. (8);
    Update $\boldsymbol{\Sigma}^{t+1}$ by Eq. (9);
    Update $\boldsymbol{\Phi}^{t+1}$ by Eq. (10);
end
```

Considering that determining perturbation strategies for samples involves factors including their learning difficulty, class distribution, and noise levels [Zhou *et al.*, 2023; Xu *et al.*, 2021], we extract fifteen training characteristics (e.g., sample loss and margin) from the classifier to encompass these aspects. All characteristics are comprehensively introduced in Section IV of the Appendix. These extracted characteristics then serve as inputs to the perturbation network, assisting in computing the values of ϵ_i. Within our framework, the perturbation network employs a two-layer MLP[2], known theoretically as a universal approximator for nearly any continuous function. Its output passes through a Tanh function to constrain the range of ϵ_i within $(-1, 1)$.

Meta-IADA employs a meta-learning-based strategy to iteratively update the classifier and the perturbation network. This involves leveraging a high-quality (clean and balanced) yet small meta dataset, $\mathcal{D}^{meta} = \{(\boldsymbol{x}_i^{meta}, y_i^{meta})\}_{i=1}^M$. Additionally, to mitigate the accuracy compromise within $\boldsymbol{\Sigma}_{y_i}$ for minor classes due to the constraints of limited training data, their values are further updated on the metadata. Let $\boldsymbol{\Omega}$ represent the parameters of the perturbation network. The optimization process within Meta-IADA is detailed as follows:

Initially, the parameters of the classifier, $\boldsymbol{\Phi}$, are updated using stochastic gradient descent (SGD) on a mini-batch of training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with the following objective:

$$\overline{\boldsymbol{\Phi}}^{t+1} \leftarrow \boldsymbol{\Phi}^t - \eta_1\frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\Phi}}\ell^{IADA}(\mathcal{F}(\boldsymbol{x}_i;\boldsymbol{\Phi}^t), y_i; \boldsymbol{\Sigma}_{y_i}^t, \epsilon(\boldsymbol{f}_i^t, \boldsymbol{\Omega}^t)), \tag{7}$$

where η_1 is the step size. Additionally, $\boldsymbol{f}_i^t$ denotes the concatenated vector containing the extracted training character-

[2] A comparative analysis of alternative architectures for the perturbation network is detailed in Section V of the Appendix.

Dataset	CIFAR10		CIFAR100	
Imbalance ratio	100:1	10:1	100:1	10:1
Class-Balanced CE[†] [Cui *et al.*, 2019]	72.68	86.90	38.77	57.57
Class-Balanced Focal[†] [Cui *et al.*, 2019]	74.57	87.48	39.60	57.99
LDAM-DRW[†] [Cao *et al.*, 2019]	78.12	88.37	42.89	58.78
Meta-Weight-Net[†] [Shu *et al.*, 2019]	73.57	87.55	41.61	58.91
De-confound-TDE* [Tang *et al.*, 2020]	80.60	88.50	44.10	59.60
LA [Menon *et al.*, 2021]	77.67	88.93	43.89	58.34
MiSLAS* [Zhong *et al.*, 2021]	82.10	90.00	47.00	63.20
LADE [Hong *et al.*, 2021]	81.17	89.15	45.42	61.69
MetaSAug[†] [Li *et al.*, 2021]	80.54	89.44	46.87	61.73
LPL* [Li *et al.*, 2022]	77.95	89.41	44.25	60.97
RISDA* [Chen *et al.*, 2022]	79.89	89.36	50.16	62.38
LDAM-DRW-SAFA* [Hong *et al.*, 2022]	80.48	88.94	46.04	59.11
BKD* [Zhang *et al.*, 2023]	82.50	89.50	46.50	62.00
Meta-IADA (Ours)	**84.01**	**91.73**	**52.18**	**64.72**

Table 2: Accuracy (%) comparison on CIFAR-LT benchmark. The best results are highlighted in bold. "$*$" represents the results in the original paper, while "†" denotes the results in [Li *et al.*, 2021].

istics for $\boldsymbol{x}_i$ at the tth iteration. Subsequently, leveraging the optimized $\overline{\boldsymbol{\Phi}}^{t+1}$, the parameter update in the perturbation network, $\boldsymbol{\Omega}$, entails utilizing a mini-batch of metadata $\{(\boldsymbol{x}_i^{meta}, y_i^{meta})\}_{i=1}^{m}$:

$$\boldsymbol{\Omega}^{t+1} \leftarrow \boldsymbol{\Omega}^t - \eta_2 \frac{1}{m} \sum\nolimits_{i=1}^{m} \nabla_{\boldsymbol{\Omega}} \ell^{CE}\big(\mathcal{F}(\boldsymbol{x}_i^{meta}; \overline{\boldsymbol{\Phi}}^{t+1}), y_i^{meta}\big), \tag{8}$$

where η_2 denotes the step size. Simultaneously, the optimization of covariance matrices is conducted using the metadata, outlined as follows:

$$\boldsymbol{\Sigma}^{t+1} \leftarrow \boldsymbol{\Sigma}^t - \eta_2 \frac{1}{m} \sum\nolimits_{i=1}^{m} \nabla_{\boldsymbol{\Sigma}} \ell^{CE}\big(\mathcal{F}(\boldsymbol{x}_i^{meta}; \overline{\boldsymbol{\Phi}}^{t+1}), y_i^{meta}\big). \tag{9}$$

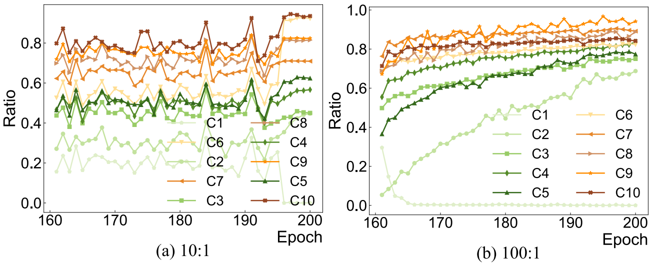
Finally, leveraging the computed perturbations and updated covariance matrices, we proceed to update the parameters $\boldsymbol{\Phi}$ of the classifier in the following manner:

$$\boldsymbol{\Phi}^{t+1} \leftarrow \boldsymbol{\Phi}^t - \eta_1 \frac{1}{n} \sum\nolimits_{i=1}^{n} \nabla_{\boldsymbol{\Phi}} \ell^{IADA}(\mathcal{F}(\boldsymbol{x}_i; \boldsymbol{\Phi}^t), y_i; \boldsymbol{\Sigma}_{y_i}^{t+1}, \epsilon(\boldsymbol{f}_i^t, \boldsymbol{\Omega}^{t+1})). \tag{10}$$

Similar to MetaAug [Li *et al.*, 2021], to acquire better-generalized representations, the classifier is initially trained using vanilla CE loss, followed by trained with Meta-IADA. The algorithm for Meta-IADA is delineated in Algorithm 1.

5 Experiments

We experiment across four typical biased learning scenarios, including LT learning, GLT learning, noisy label learning, and subpopulation shift learning, involving image and text datasets. The excluded settings and results (including those on standard datasets) are detailed in the Appendix. The code for Meta-IADA is available in the supplementary materials.



(a) 10:1 (b) 100:1

Figure 3: Ratio of adversarial samples in each class during the last forty epochs on CIFAR10 under imbalance ratios of 10:1 and 100:1. From "C1" to "C10", the class proportions progressively decrease.

Method	Accuracy↑
Class-Balanced CE[†] [Cui *et al.*, 2019]	66.43
Class-Balanced Focal[†] [Cui *et al.*, 2019]	61.12
LDAM-DRW[†] [Cao *et al.*, 2019]	68.00
BBN[†] [Zhou *et al.*, 2020]	66.29
Meta-Class-Weight[†] [Jamal *et al.*, 2020]	67.55
MiSLAS* [Zhong *et al.*, 2021]	71.60
LADE* [Hong *et al.*, 2021]	70.00
MetaSAug[†] [Li *et al.*, 2021]	68.75
LDAM-DRS-SAFA* [Hong *et al.*, 2022]	69.78
RISDA* [Chen *et al.*, 2022]	69.15
BKD* [Zhang *et al.*, 2023]	71.20
Meta-IADA (Ours)	**72.55**

Table 3: Accuracy (%) comparison on iNat 2018 dataset.

5.1 Long-Tail Learning

Four LT image classification benchmarks, CIFAR-LT [Cui *et al.*, 2019], ImageNet-LT [Liu *et al.*, 2019], Places-LT [Liu *et al.*, 2019], and iNaturalist (iNat) 2018 [Jamal *et al.*, 2020], are evaluated. Additionally, two imbalanced text classification datasets are included. Due to space limitations, we only present experiments for CIFAR-LT and iNat in the main text.

Experiments on CIFAR-LT Datasets. We employ the ResNet-32 model [He *et al.*, 2016] with an initial learning rate of 0.1. The training employs the SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4} on a single GPU, spanning 200 epochs. The learning rate is decayed by 0.01 at the 160th and 180th epochs. Additionally, the perturbation network is optimized using Adam, with an initial learning rate of $1{\times}10^{-3}$. To construct metadata, we randomly select ten images per class from the validation data. For the hyperparameters in the IADA loss, α is selected from $\{0.1, 0.25, 0.5, 0.75, 1\}$, while keeping β fixed at 1.

From the results reported in Table 2, Meta-IADA displays remarkable superiority over other LT baselines, underscoring its effectiveness in managing imbalanced class distributions. Additionally, it outperforms previous implicit augmentation methods, providing evidence for the efficacy of augmenting samples within their adversarial and anti-adversarial perturbation distributions. Furthermore, as depicted in Fig. 3, Meta-IADA induces a lower proportion of adversarial samples in major classes ("C1" to "C5"), whereas minor classes ("C6" to "C10") exhibit a higher proportion. This manifests the model's increased attention on samples in tail classes. Comparative analyses of confusion matrices involving CE loss, MetaSAug, and Meta-IADA are presented in Section V of the Appendix. The results unveil that Meta-IADA remarkably increases the accuracy of both major and minor classes. Conversely, while MetaSAug improves the performance of minor categories, it compromises the accuracy of major classes.

Experiments on iNat 2018 Dataset. The ResNet-50 model serves as the backbone classifier, pre-trained on Ima-

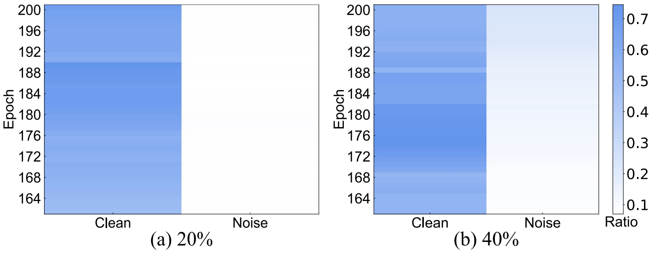
Protocol	**CLT**		**GLT**		**ALT**	
Metric	**Acc.↑**	**Prec.↑**	**Acc.↑**	**Prec.↑**	**Acc.↑**	**Prec.↑**
CE loss‡	42.52	47.92	34.75	40.65	41.73	41.74
MixUp‡ [Zhang *et al.*, 2018]	38.81	45.41	31.55	37.44	42.11	42.42
LDAM‡ [Cao *et al.*, 2019]	46.74	46.86	38.54	39.08	42.66	41.80
ISDA [Wang *et al.*, 2019]	42.66	44.98	36.44	37.26	43.34	43.56
cRT‡ [Kang *et al.*, 2020]	45.92	45.34	37.57	37.51	41.59	41.43
LWS‡ [Kang *et al.*, 2020]	46.43	45.90	37.94	38.01	41.70	41.71
De-confound-TDE‡ [Tang *et al.*, 2020]	45.70	44.48	37.56	37.00	41.40	42.36
BLSoftmax‡ [Ren *et al.*, 2020]	45.79	46.27	37.09	38.08	41.32	41.37
BBN‡ [Zhou *et al.*, 2020]	46.46	49.86	37.91	41.77	43.26	43.86
RandAug‡ [Cubuk *et al.*, 2020]	46.40	52.13	38.24	44.74	46.29	46.32
LA‡ [Menon *et al.*, 2021]	46.53	45.56	37.80	37.56	-	-
MetaSAug [Li *et al.*, 2021]	48.53	54.21	40.27	44.38	47.62	48.26
IFL‡ [Tang *et al.*, 2022]	45.97	52.06	37.96	44.47	45.89	46.42
RISDA [Chen *et al.*, 2022]	46.31	51.24	38.45	42.77	43.65	43.23
BKD [Zhang *et al.*, 2023]	46.51	50.15	37.93	41.50	42.17	41.83
Meta-IADA (Ours)	**53.45**	**58.05**	**44.36**	**50.07**	**52.54**	**53.23**

Table 4: Accuracy and precision (%) of CLT, GLT, and ALT protocols on the ImageNet-GLT benchmark. ‡ indicates the results reported in [Tang *et al.*, 2022].

geNet [Russakovsky *et al.*, 2015] and iNat 2017 [Horn *et al.*, 2018] datasets. The perturbation network is optimized using Adam, with an initial learning rate of 1×10^{-4}. Other settings follow those outlined in the MetaSAug [Li *et al.*, 2021] paper. As reported in Table 3, Meta-IADA outperforms other comparative approaches tailored for LT learning. This suggests that, even in situations with imbalanced class distributions, employing a sample-level strategy proves more effective, as finer-grained imbalances within each class may also exist.

5.2 Generalized Long-Tail Learning

GLT learning considers both long-tailed class and attribute distributions in the training data. We employ two GLT benchmarks, ImageNet-GLT, and MSCOCO-GLT [Tang *et al.*, 2022]. Each benchmark comprises three protocols, CLT, ALT, and GLT, showcasing variations in class distribution, attribute distribution, and combinations of both between training and testing datasets. The ResNeXt-50 [Xie *et al.*, 2017] model is utilized as the backbone network. We train models with a batch size of 256 and an initial learning rate of 0.1, using SGD with a weight decay of 5×10^{-4} and a momentum of 0.9. The perturbation network is optimized using the Adam optimizer, initialized with a learning rate of 1×10^{-3}. The metadata utilized in this experiment is balanced in both classes and attributes. We adopt the construction method of Tang et al. [2022], which involves clustering images in each class into six groups using KMeans with a pre-trained ResNet-50 model. Two images are then randomly selected from each group and class within the validation data.



(a) 20%

(b) 40%

Figure 4: Ratio of adversarial samples for noisy and clean samples in the final forty epochs on CIFAR10 with 20% and 40% flip noise.

Dataset	**CIFAR10**		**CIFAR100**	
Noise ratio	**20%**	**40%**	**20%**	**40%**
CE loss	76.85	70.78	50.90	43.02
D2L [Ma *et al.*, 2018]	87.64	83.90	63.39	51.85
Co-teaching [Han *et al.*, 2018]	82.85	75.43	54.19	44.92
GLC [Hendrycks *et al.*, 2018]	89.77	88.93	63.15	62.24
MentorNet [Jiang *et al.*, 2018]	86.41	81.78	62.00	52.71
L2RW [Ren *et al.*, 2018]	87.88	85.70	57.51	51.00
DMI [Xu *et al.*, 2019]	88.43	84.00	58.87	42.95
Meta-Weight-Net [Shu *et al.*, 2019]	90.35	87.65	64.31	58.67
APL [Ma *et al.*, 2020]	87.45	81.08	59.86	53.31
JoCoR [Wei *et al.*, 2020]	90.78	83.67	65.21	46.44
MLC [Zheng *et al.*, 2021]	91.55	89.53	66.32	62.29
MFRW-MES [Ricci *et al.*, 2023]	91.45	90.72	65.27	62.35
Meta-IADA (Ours)	**93.44**	**91.99**	**69.16**	**64.37**

Table 5: Accuracy (%) comparison on CIFAR datasets with 20% and 40% flip noise.

The comparative results for ImageNet-GLT are outlined in Table 4, while those for MSCOCO-GLT are presented in the Appendix. Meta-IADA demonstrates substantial enhancements in accuracy and precision across all three protocols, emphasizing its ability to address distributional skewness in three scenarios, including attribute imbalance, class imbalance, and their combination. The efficacy of Meta-IADA stems from its ability to apply adversarial augmentation to samples from tail classes and those with rare attributes. This significantly magnifies the impact of these samples on model training, resulting in improved prediction accuracy. However, methods specifically designed for LT learning demonstrate subpar performance on the ALT protocol, primarily due to their class-level characteristics.

5.3 Noisy Label Learning

We examine two types of label corruptions: uniform and pair-flip noise [Shu *et al.*, 2019], using CIFAR [Krizhevsky and Hinton, 2009] datasets. The Wide ResNet-28-10 (WRN-28-10) [Zagoruyko and Komodakis, 2016] and ResNet-32 models are utilized for uniform and flip noises, respectively. 1,000 images with clean labels are selected from the validation set to compile the metadata. The ResNet settings match those used for CIFAR-LT. WRN-28-10 is trained using SGD with an initial learning rate of 0.1, a momentum of 0.9, and a weight decay of 5×10^{-4}; the learning rate is decayed by 0.1 at the 50th and 55th epochs during the total 60 epochs. The Adam optimizer, initialized with a learning rate of 1×10^{-3}, is utilized for optimizing the perturbation network.

Dataset	**CelebA**		**CMNIST**		**Waterbirds**		**CivilComments**	
Metric	**Avg.↑**	**Worst↑**	**Avg.↑**	**Worst↑**	**Avg.↑**	**Worst↑**	**Avg.↑**	**Worst↑**
CORAL+ [Li *et al.*, 2018]	93.8	76.9	71.8	69.5	90.3	79.8	88.7	65.6
IRM+ [Arjovsky *et al.*, 2019]	94.0	77.8	72.1	70.3	87.5	75.6	88.8	66.3
GroupDRO+ [Sagawa *et al.*, 2020]	92.1	87.2	72.3	68.6	91.8	90.6	89.9	70.0
DomainMix+ [Xu *et al.*, 2020]	93.4	65.6	51.4	48.0	76.4	53.0	90.9	63.6
IB-IRM+ [Ahuja *et al.*, 2021]	93.6	85.0	72.2	70.7	88.5	76.5	89.1	65.3
V-REx+ [Krueger *et al.*, 2021]	92.2	86.7	71.7	70.2	88.0	73.6	90.2	64.9
UW+ [Yao *et al.*, 2022]	92.9	83.3	72.2	66.0	**95.1**	88.0	89.8	69.2
Fish+ [Shi *et al.*, 2022]	93.1	61.2	46.9	35.6	85.6	64.0	89.8	71.1
LISA+ [Yao *et al.*, 2022]	92.4	89.3	74.0	73.3	91.8	89.2	89.2	72.6
Meta-IADA (Ours)	**94.5**	**91.2**	**78.0**	**75.9**	94.5	**92.5**	**92.1**	**74.8**

Table 6: Comparison of the average and worst-group accuracy (%) on four subpopulation shift datasets. + indicates the results reported in [Yao *et al.*, 2022].

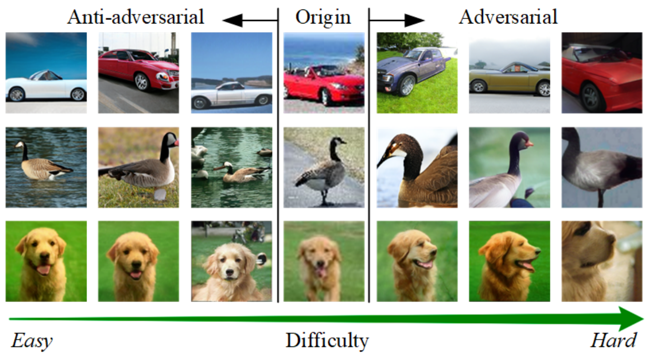


Figure 5: Visualization of instances corresponding to deep features augmented by Meta-IADA.

Table 5 presents the comparison results for flip noise, while those for uniform noise are detailed in the Appendix. Meta-IADA consistently achieves SOTA performance when compared to all other methods, surpassing the highest accuracy among comparative methods by an average of 2%. These outcomes underscore its capacity to bolster DNNs' robustness against noise. An analysis of Fig. 4 indicates that nearly all noisy samples undergo anti-adversarial augmentations in Meta-IADA, effectively mitigating the negative impact of noisy samples on the overall performance of DNNs.

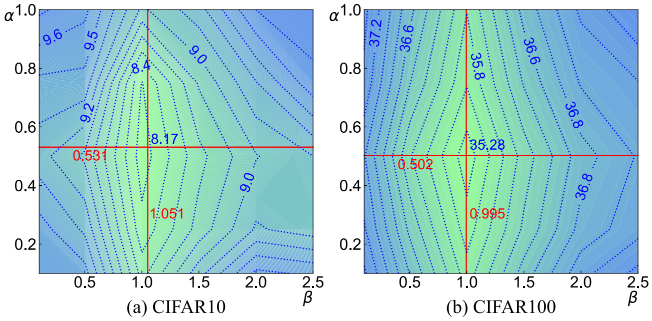
5.4 Subpopulation Shift Learning

We conduct evaluations on four binary classification datasets characterized by subpopulation shifts: Colored MNIST (CMNIST) [Yao *et al.*, 2022], Waterbirds [Sagawa *et al.*, 2020], CelebA [Liu *et al.*, 2016], and CivilComments [Borkan *et al.*, 2019]. For three image datasets (i.e., CMNIST, Waterbirds, and CelebA), we utilize ResNet-50 [He *et al.*, 2016] as the backbone network, while for the text dataset, CivilComments, we employ DistilBert [Sanh *et al.*, 2019]. To provide a comprehensive assessment, we report both average and worst-group accuracy. Detailed experimental settings and dataset introductions are available in Section V of the Appendix.

Based on the results in Table 6, Meta-IADA achieves the highest worst-group accuracy across the four datasets. This emphasizes its effectiveness in enhancing performance for underrepresented groups, such as samples in the landbird class with a water background. Typically, these samples benefit from adversarial augmentation, enhancing their influence on model training. Furthermore, except for the Waterbirds dataset, Meta-IADA surpasses other methods in terms of average accuracy. These findings illustrate Meta-IADA's capability to bolster model resilience against subpopulation shifts.

Regularization terms			CIFAR10		CIFAR100	
$\mathcal{G}$	$\mathcal{R}$	$\mathscr{F}$	10:1	100:1	10:1	100:1
✓	✓	✓	**8.17**	**15.99**	**35.28**	**47.82**
✗	✓	✓	9.48	18.20	36.95	51.01
✓	✗	✓	10.06	19.11	37.62	52.47
✓	✓	✗	9.77	17.36	37.18	49.87

Table 7: Results of ablation studies using the ResNet-32 model on CIFAR-LT benchmark. Top-1 error rates are reported.



(a) CIFAR10

(b) CIFAR100

Figure 6: Results of sensitivity tests on CIFAR datasets with an imbalance ratio of 10:1, using the ResNet-32 model.

5.5 Visualization Results

We utilize the visualization method introduced in ISDA [Wang *et al.*, 2019], projecting the features augmented by Meta-IADA back into the pixel space. The results, depicted in Fig. 5, highlight the diversity encapsulated within the generated adversarial and anti-adversarial samples. Additionally, samples created through adversarial and anti-adversarial augmentation commonly showcase varying levels of difficulty compared to the original samples. These modifications are attributed to transformations within the deep feature space, influenced by semantic vectors associated with attributes such as angles, colors, and backgrounds.

5.6 Ablation and Sensitivity Studies

Ablation studies are conducted to analyze the impact of the three regularization terms within the IADA loss. The results reported in Table 7, emphasize the crucial roles played by both the generalization term $\mathcal{G}$ and the robustness term $\mathcal{R}$. Additionally, in the context of an imbalanced scenario, the fairness term $\mathscr{F}$ also demonstrates its significance. Furthermore, sensitivity tests are performed on the hyperparameters α and β, which respectively control the effects of $\mathcal{G}$ and $\mathscr{F}$. As shown in Fig. 6, Meta-IADA achieves optimal performance when α is around 0.5 and β approaches 1. Besides, the stable ranges for α and β lie within $[0.25, 0.75]$ and $[0.75, 1.25]$, respectively. Based on these findings, we recommend setting $\alpha = 0.5$ and $\beta = 1$ for practical applications.

6 Conclusion

This paper presents a novel adversarial data augmentation strategy to facilitate model training across diverse learning scenarios, particularly those with data biases. This strategy enriches the deep features of samples by incorporating their adversarial and anti-adversarial perturbation distributions, dynamically adjusting the learning difficulty of training samples. Subsequently, we formulate a surrogate loss for our augmentation strategy and establish a meta-learning framework to optimize classifiers using this loss. Extensive experiments are conducted across various biased learning scenarios involving different networks and datasets, showcasing the effectiveness and broad applicability of our approach.

References

[Ahuja *et al.*, 2021] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. In *NeurIPS*, pages 3438–3450, 2021.

[Arjovsky *et al.*, 2019] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[Borkan *et al.*, 2019] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *WWW*, pages 491–500, 2019.

[Cao *et al.*, 2019] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, pages 1567–1578, 2019.

[Chen *et al.*, 2022] Xiaohua Chen, Yucan Zhou, Dayan Wu, Wanqian Zhang, Yu Zhou, Bo Li, and Weiping Wang. Imagine by reasoning: A reasoning-based implicit semantic data augmentation for long-tailed classification. In *AAAI*, pages 356–364, 2022.

[Cubuk *et al.*, 2020] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, pages 3008–3017, 2020.

[Cui *et al.*, 2019] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pages 9268–9277, 2019.

[Han *et al.*, 2018] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, pages 8536–8546, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hendrycks *et al.*, 2018] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*, pages 10477–10486, 2018.

[Hong *et al.*, 2021] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *CVPR*, pages 6626–6636, 2021.

[Hong *et al.*, 2022] Yan Hong, Jianfu Zhang, Zhongyi Sun, and Ke Yan. Safa: Sample-adaptive feature augmentation for long-tailed image classification. In *ECCV*, pages 587–603, 2022.

[Horn *et al.*, 2018] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778, 2018.

[Jamal *et al.*, 2020] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *CVPR*, pages 7610–7619, 2020.

[Jiang *et al.*, 2018] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Fei-Fei Li. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, pages 2304–2313, 2018.

[Kang *et al.*, 2020] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020.

[Krizhevsky and Hinton, 2009] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, 2009.

[Krueger *et al.*, 2021] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, pages 5815–5826, 2021.

[Lee *et al.*, 2023] Sungyoon Lee, Hoki Kim, and Jaewook Lee. Graddiv: Adversarial robustness of randomized neural networks via gradient diversity regularization. *IEEE TPAMI*, 45(2):2645–2651, 2023.

[Li *et al.*, 2018] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409, 2018.

[Li *et al.*, 2021] Shuang Li, Kaixiong Gong, Chi-Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *CVPR*, pages 5208–5217, 2021.

[Li *et al.*, 2022] Mengyang Li, Fengguang Su, Ou Wu, et al. Logit perturbation. In *AAAI*, pages 1359–1366, 2022.

[Liu *et al.*, 2016] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2016.

[Liu *et al.*, 2019] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, pages 2537–2546, 2019.

[Ma *et al.*, 2018] Xingjun Ma, Yisen Wang, Michael E. Houle, Shuo Zhou, Sarah Erfani, Shutao Xia, Sudanthi Wijewickrema, et al. Dimensionality-driven learning with noisy labels. In *ICML*, pages 3355–3364, 2018.

[Ma *et al.*, 2020] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *ICML*, pages 6543–6553, 2020.

[Madry *et al.*, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[Maharana *et al.*, 2022] Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. A review: Data preprocessing and data augmentation techniques. *Global Transitions Proceedings*, 3(1):91–99, 2022.

[Menon *et al.*, 2021] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021.

[Ren *et al.*, 2018] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, pages 4334–4343, 2018.

[Ren *et al.*, 2020] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *NeurIPS*, pages 4175–4186, 2020.

[Ricci *et al.*, 2023] Simone Ricci, Tiberio Uricchio, and Alberto Del Bimbo. Meta-learning advisor networks for long-tail and noisy labels in social image classification. *ACM TOMM*, 19(5s):1–23, 2023.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[Sagawa *et al.*, 2020] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, et al. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

[Sanh *et al.*, 2019] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[Shi *et al.*, 2022] Yuge Shi, Jeffrey Seely, Philip H.S. Torr, N. Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *ICLR*, 2022.

[Shu *et al.*, 2019] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, pages 1919–1930, 2019.

[Tang *et al.*, 2020] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, pages 1513–1524, 2020.

[Tang *et al.*, 2022] Kaihua Tang, Mingyuan Tao, Jiaxin Qi, Zhenguang Liu, and Hanwang Zhang. Invariant feature learning for generalized long-tailed classification. In *ECCV*, pages 709–726, 2022.

[Taylor and Nitschke, 2018] Luke Taylor and Geoff Nitschke. Improving deep learning with generic data augmentation. In *SSCI*, pages 1542–1547, 2018.

[Wang *et al.*, 2019] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Cheng Wu, and Gao Huang. Implicit semantic data augmentation for deep networks. In *NeurIPS*, pages 12635–12644, 2019.

[Wei *et al.*, 2020] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, pages 13723–13732, 2020.

[Xie *et al.*, 2017] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 5987–5995, 2017.

[Xu and Zhao, 2023] Xiaogang Xu and Hengshuang Zhao. Universal adaptive data augmentation. In *IJCAI*, pages 1596–1603, 2023.

[Xu *et al.*, 2019] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *NeurIPS*, pages 6225–6236, 2019.

[Xu *et al.*, 2020] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI*, pages 6502–6509, 2020.

[Xu *et al.*, 2021] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *ICML*, pages 11492–11501, 2021.

[Yao *et al.*, 2022] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *ICML*, pages 25407–25437, 2022.

[Zagoruyko and Komodakis, 2016] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[Zhang *et al.*, 2018] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[Zhang *et al.*, 2023] Shaoyu Zhang, Chen Chen, Xiyuan Hu, and Silong Peng. Balanced knowledge distillation for long-tailed learning. *Neurocomputing*, 527:36–46, 2023.

[Zheng *et al.*, 2021] Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. Meta label correction for noisy label learning. In *AAAI*, pages 11053–11061, 2021.

[Zhong *et al.*, 2021] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *CVPR*, pages 16489–16498, 2021.

[Zhou *et al.*, 2020] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, pages 9719–9728, 2020.

[Zhou *et al.*, 2023] Xiaoling Zhou, Nan Yang, and Ou Wu. Combining adversaries with anti-adversaries in training. In *AAAI*, pages 11435–11442, 2023.

[Zhu *et al.*, 2021] Jianing Zhu, Jingfeng Zhang, Bo Han, Tongliang Liu, Gang Niu, Hongxia Yang, Mohan Kankanhalli, and Masashi Sugiyama. Understanding the interaction of adversarial training with noisy labels. *arXiv preprint arXiv:2102.03482*, 2021.