Adaptive Momentum and EMA-weighted Modeling for Imbalanced Label Distribution Learning

Yongbiao Gao[1,2,3*], **Xiangcheng Sun**[1,2*], **Chao Tan**[4], **Chunyu Hu**[1,2], **Guohua Lv**[1,2†]

[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan),Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

[2]Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan, China

[3]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Application, (Southeast University), Ministry of Education, China

[4]School of Computer and Electronic Information / School of Artificial Intelligence, Nanjing Normal University, Nanjing, China

{gaoyb, 10431240010, hcy, guohualv}@qlu.edu.cn, 73022@njnu.edu.cn

Abstract

Label Distribution Learning (LDL) is a groundbreaking paradigm for addressing the task with label ambiguity. Subjectivity in annotating label description degrees often leads to imbalanced label distribution. Existing approaches either adopt representation alignment or decoupling strategies to solve the imbalanced label distribution learning (ILDL). However, representation alignment-based methods overlook the issue of gradient vanishing for non-dominant branches within imbalanced label distributions, while decoupling-based approaches fail to achieve adaptive weight optimization. To address these issues, we propose **A**daptive **M**omentum and **E**xponential **M**oving **A**verage weighted modeling (**AMEMA**). AMEMA combines EMA-based loss weighting with momentum allocation to mitigate gradient attenuation in non-dominant label learning and adaptively balance the optimization signals between dominant and non-dominant branches. It computes and updates Kullback-Leibler divergence losses for each branch using EMA, and applies different initial momenta to facilitate branch-specific optimization dynamics. Dynamic weighting coefficients, derived from EMA-smoothed losses, allow the model to adjust its learning direction adaptively and improve the learning of non-dominant labels. Extensive experiments on benchmark datasets show that AMEMA consistently outperforms state-of-the-art ILDL methods across various evaluation metrics.

Code — https://github.com/wenhuihji/AMEMA

Introduction

Developing reliable and generalizable AI systems is a fundamental challenge in machine learning. Learning with ambiguity is a central challenge in this context. Data ambiguity is pervasive in real-world scenarios, such as mobile crowdedge computing and satellite-ground collaborative AI platforms (Yao et al. 2025), often degrading the performance of traditional models (Kendall and Gal 2017). Label Distribution Learning (LDL) (Geng 2016) assigns a label distribution to each instance. Each value, known as a description degree, reflects the relevance of a label to an instance. By explicitly modeling ambiguity, LDL has been successfully applied to diverse fine-grained prediction tasks, such as age estimation (He et al. 2021; Yu et al. 2023; Lee et al. 2024), video parsing (Rachavarapu et al. 2023; Zhao et al. 2025a; Rachavarapu, Ramakrishnan et al. 2024), emotion analysis (Wang and Li 2025; Wu et al. 2025; Zhao et al. 2025b), head pose estimation (Algabri, Abdu, and Lee 2024; Jiang, Wang, and Chang 2024; Zou, Jia, and Tang 2025), facial beauty perception (Li et al. 2024; Cheng et al. 2024; Liu et al. 2025), etc. This capability positions LDL as a preferred approach in domains where precise modeling of label ambiguity is critical for advancing state-of-the-art performance.

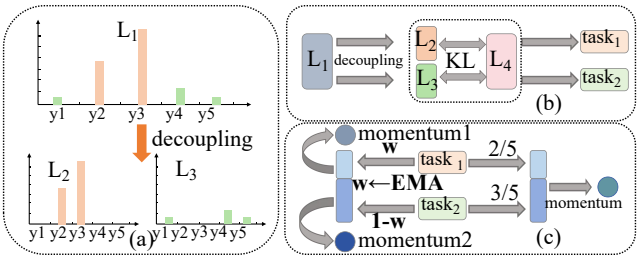


Figure 1: (a) In DILDL (Gao et al. 2025), the imbalanced label distribution is decoupled into dominant and non-dominant components. (b) Each component in DILDL is independently aligned, forming two separate optimization pathways. (c) We propose an AMEMA framework, employing adaptive EMA-based weighting and branch-specific momentum allocation.

LDL has seen rapid development, accompanied by the

*These authors contributed equally.

†Corresponding Author.

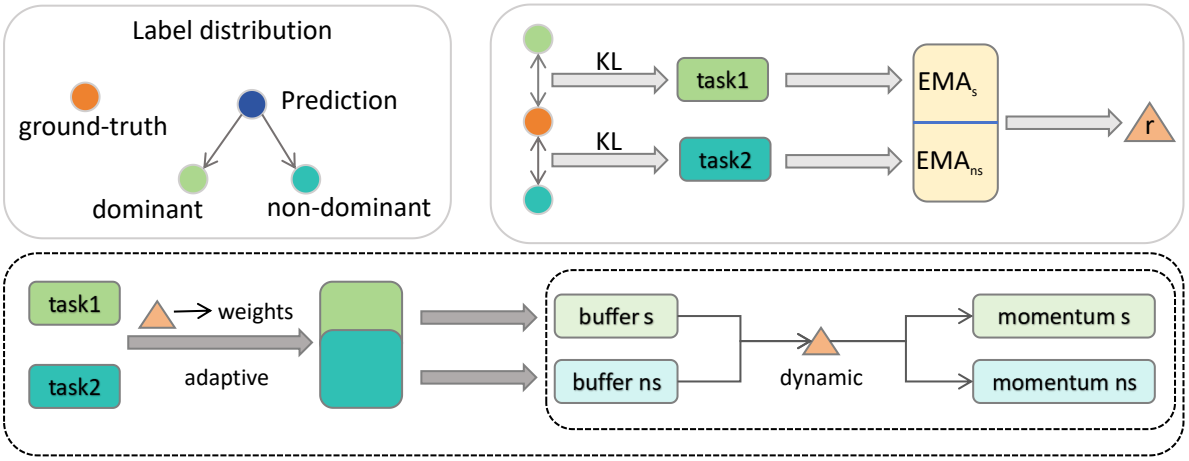


Figure 2: The EMA mechanism derives a dynamic weight ratio r to adaptively balance optimization signals between dominant/non-dominant branches, enabling label-aware coordinated optimization via distinct momentum coefficients.

proposal of several representative methods. The seminal work by Geng (Geng 2016) introduced the LDL framework, in which each instance is assigned a label distribution. Model training minimizes the Kullback-Leibler divergence between predicted and ground-truth label distributions, effectively addressing label ambiguity in real-world data. To further improve performance, LDL-HR (Wang and Geng 2019) explicitly distinguishes the label with the highest description degree from the others, leveraging this difference to enhance model accuracy. More recent studies focus on enhancing the model's robustness to label ambiguity, exploring alternative loss functions and incorporating label structure information to better capture complex label distributions (Kim, Lee, and Lee 2024; Sheng et al. 2024).

The above LDL methods typically assume that the description degrees are relatively balanced, which simplifies the training process by allowing the model to receive uniform supervisory signals (Geng 2016). However, in real-world scenarios, such balance is difficult to achieve due to the inherent subjectivity in label annotation (Xu et al. 2023; Zhao et al. 2023b). When label description degrees vary significantly, the performance of LDL can degrade substantially. This issue, known as *Imbalanced Label Distribution Learning* (ILDL) (Zhao et al. 2023b), arises from handling imbalanced label distributions.

Imbalanced classification and imbalanced regression are most similar to ILDL, both of which have been widely studied (Liu et al. 2019; Kim, Jeong, and Shin 2020; You et al. 2025; Wang and Wang 2023). For example, LDMLR (Han et al. 2024) enhances class discriminability by generating pseudo-features in the latent space, thereby improving recognition performance under long-tailed distributions. In regression, KNNOR-DeepReg (Wang and Wang 2023) leverages K-nearest neighbor oversampling in conjunction with deep autoencoders to synthesize both input features and target values, achieving competitive results on high-dimensional imbalanced datasets. However, these methods cannot be directly utilized for ILDL due to the fundamentally distinct task formulation. While traditional methods may focus on single labels, ILDL is designed to model label distributions. To address the imbalanced problem in LDL, several specialized methods have been proposed. For example, RDA (Zhao et al. 2023b) adopts a feature-label alignment strategy to mitigate distributional drift between label distributions and feature representations. However, its static optimization applies uniform alignment strength across all labels.

Decoupled Imbalanced Label Distribution Learning (DILDL) (Gao et al. 2025) addresses the gradient bias issue in non-dominant labels caused by large discrepancies in distribution degrees. As shown in Figure 1(a) and 1(b), DILDL structurally decouples dominant and non-dominant labels, but its modeling remains static and lacks adaptive flexibility during training. Furthermore, both branches share the same optimizer and optimization parameters, limiting the application of differentiated update strategies. This constraint restricts the independent optimization of each branch and reduces the method's effectiveness in fully mitigating the imbalance in LDL.

To address the limitations of static modeling in existing decoupled methods, we propose **A**daptive **M**omentum and **E**xponential **M**oving **A**verage weighted modeling (**AMEMA**). As depicted in Figure 1(c), our approach adopts adaptive weighting and branch-specific momentum allocation. Specifically, as shown in Figure 2, our framework retains the structural decoupling of dominant and non-dominant labels while introducing branch-specific momentum mechanisms and an EMA-based loss reweighting strategy to enhance the latent information of non-dominant labels. We initialize lower momentum values for dominant branches and higher values for non-dominant branches. We apply EMA to smooth KL divergence losses and dynamically assign weight coefficients proportional to the uncertainty of each branch based on these smoothed losses.

This combination of fixed branch-wise momentum initialization and dynamically generated EMA weights enables AMEMA to adapt optimization signals according to the training dynamics of each branch, thereby improving robustness to distributional shifts and enhancing the learning of non-dominant labels. Additionally, AMEMA dynamically adjusts branch learning via EMA-based loss weighting, prioritizing recent losses to emphasize unstable gradient branches. This adaptive scheme strengthens supervision of non-dominant labels, while branch-specific momentum optimizes decoupled optimization paths. By integrating momentum customization and EMA weighting, AMEMA addresses the imbalance optimization in ILDL by jointly adjusting gradient momentum and loss sensitivity. Extensive experiments demonstrate significant performance gains across benchmarks, validating the robustness of AMEMA in solving ILDL.

Our main contributions are summarized as follows:

- We propose AMEMA, a novel framework that extends existing decoupling-based methods by integrating structural label disentanglement with optimization-level differentiation for ILDL.
- We explicitly separate the model parameters for dominant and non-dominant label modeling. Each group has its own feature encoder and distribution predictor, and is assigned an adaptive momentum to achieve branch-specific optimization dynamics.
- We introduce an EMA-guided adaptive loss reweighting strategy to dynamically adjust the learning signals from the dominant and non-dominant branches based on their temporal stability.

Related Work

Label Distribution Learning

Label Distribution Learning (LDL) (Geng 2016) is a novel paradigm that assigns a label distribution to each instance, where each element represents the degree to which a label describes the instance. LDL has been widely applied to tackle tasks characterized by label ambiguity, such as age estimation (He et al. 2021; Yu et al. 2023; Lee et al. 2024), facial beauty perception (Li et al. 2024; Cheng et al. 2024; Liu et al. 2025), and emotion analysis (Wang and Li 2025; Wu et al. 2025; Zhao et al. 2025b). Traditional LDL methods are generally grouped into three categories. The first includes problem transformation methods, such as PT-SVM (Geng, Yin, and Zhou 2013) and PT-Bayes (Geng, Yin, and Zhou 2013), which recast LDL as conventional classification or regression problems. The second comprises algorithm adaptation methods, such as AA-KNN and AA-PT (Geng 2016), which extend classical machine learning algorithms to model label distributions. The third encompasses specialized models that directly learn from label distributions, often optimizing objectives such as maximum entropy (Geng 2016) or KL divergence (Fan et al. 2024; Li et al. 2025). Most existing methods assume that label distributions are balanced, but this assumption rarely holds in real-world scenarios, where dominant labels with high frequency often overshadow the training process, leaving non-dominant labels inadequately modeled.

Recent studies aim at improving the robustness of LDL under noisy or uncertain annotations. For example, graph-based approaches (Jin et al. 2024; Kim, Yun, and Song 2023) capture both local and global label dependencies to enhance distribution quality, while the directional label diffusion model (Hou et al. 2025) employs a generative diffusion-based strategy to recover more semantically coherent and robust label distributions under noisy label settings. Nevertheless, most current LDL methods employ static optimization and implicitly assume that the label distribution is relatively balanced across instances.

Imbalanced Label Distribution Learning

Imbalanced Label Distribution Learning (ILDL) (Zhao et al. 2023b) refers to scenarios where label description degrees vary significantly across labels, leading to highly skewed distributions. Unlike traditional imbalanced classification tasks that focus on discrete class occurrence, ILDL operates on continuous-valued label distributions (Jia et al. 2023; Han et al. 2024), in which each label is associated with a real-valued description degree reflecting its relevance to an instance. This imbalance introduces severe optimization challenges. Head labels often dominate the loss landscape and optimization dynamics, while tail labels experience gradient vanishing and insufficient representation learning (Gao et al. 2025; Zhao et al. 2023b; Jia et al. 2019).

To address the distributional mismatch, Zhao et al. (Zhao et al. 2023b) proposed the Representation Distribution Alignment (RDA) method, which aligns feature and label distributions in a shared latent space via a KL-divergence alignment to enhance the semantic coherence between encoded features and label distribution. Despite its effectiveness in enforcing distributional consistency, RDA has several critical limitations. Its static optimization process applies uniform alignment strength to all labels, which over-constrains dominant labels and provides insufficient gradient feedback to minority labels. Furthermore, the lack of an adaptive alignment modulation mechanism significantly limits its ability to model the evolving dynamic structure of highly imbalanced label distributions.

Although DILDL (Gao et al. 2025) alleviates some issues by decoupling imbalanced label distributions into dominant and non-dominant branches, it still relies on static weighting schemes and employs shared optimization hyperparameters across these branches, which fails to capture the evolving imbalance. Therefore, in this work, we propose an adaptive ILDL framework that dynamically balances optimization signals in a label-aware and training-aware manner.

Method

The proposed framework decouples the label distribution into dominant and non-dominant components and assigns distinct momentum to each branch. It also separates the KL divergence losses into dominant and non-dominant label components, applying separate momentum initializations for each. Moreover, an EMA-based loss-weighting mecha-

nism is integrated into the framework to dynamically balance optimization signals during training.

Dynamic Weight Allocation

Assume $f_\theta(\cdot)$ is the mapping function from the instance space $\mathcal{X}$ to the label distribution space $\mathcal{Y}$. The objective of LDL is to minimize the discrepancy between the ground truth and predicted label distributions. The KL divergence is commonly used as the loss function. Hence, the objective can be written as follows:

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

First, we decouple the distribution of the dominant label. For each instance $\boldsymbol{x}$, define the dominant-label distribution as $\bar{D}_{\boldsymbol{x}} = \left[\bar{d}_{\boldsymbol{x}}^{y_1}, \bar{d}_{\boldsymbol{x}}^{y_2}, \ldots, \bar{d}_{\boldsymbol{x}}^{y_c}\right]^\top$, where $\bar{d}_{\boldsymbol{x}}^{y_j}$ is defined by

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

and $y_{\boldsymbol{x}}$ denotes the label with the highest description degree. In this decoupled representation, the dominant label is assigned a description degree of 1, while all other labels receive 0.

Next, we normalize the description degrees of the non-dominant labels.

$$\hat{d}_{\boldsymbol{x}}^{y_j} = \begin{cases} 0, & \text{if } y_j = y_{\boldsymbol{x}}, \\ \dfrac{d_{\boldsymbol{x}}^{y_j}}{\sum_{j=1, y_j \neq y_{\boldsymbol{x}}}^{c} d_{\boldsymbol{x}}^{y_j}}, & \text{otherwise}. \end{cases} \tag{3}$$

According to Eq. (2) and Eq. (3), the KL divergence is divided into two components corresponding to the dominant and non-dominant branches, with the dominant branch defined as follows:

$$\begin{aligned} \mathcal{L}_s &= \mathrm{KL}\big(\bar{D}_{\boldsymbol{x}} \,\|\, \bar{f}_\theta(\boldsymbol{x})\big) \\ &= \sum_{j=1}^{c} \bar{d}_{\boldsymbol{x}}^{y_j} \left(\ln \bar{d}_{\boldsymbol{x}}^{y_j} - \ln \bar{f}_\theta^j(\boldsymbol{x})\right). \end{aligned} \tag{4}$$

The non-dominant branch is defined as follows:

$$\begin{aligned} \mathcal{L}_{ns} &= \mathrm{KL}\big(\hat{D}_{\boldsymbol{x}} \,\|\, \hat{f}_\theta(\boldsymbol{x})\big) \\ &= \sum_{j=1, y_j \neq y_{\boldsymbol{x}}}^{c} \hat{d}_{\boldsymbol{x}}^{y_j} \left(\ln \hat{d}_{\boldsymbol{x}}^{y_j} - \ln \hat{f}_\theta^j(\boldsymbol{x})\right). \end{aligned} \tag{5}$$

Here, $\bar{f}_\theta(\boldsymbol{x})$ and $\hat{f}_\theta(\boldsymbol{x})$ denote the model's predicted label distributions from the dominant and non-dominant branches.

Inspired by DWEMA (Lakkapragada et al. 2023) and EMAN (Cai et al. 2021), we denote the EMA update rules at step t as:

$$\mathrm{EMA}_s^{(t)} = m \cdot \mathrm{EMA}_s^{(t-1)} + (1-m) \cdot \mathcal{L}_s^{(t)}, \tag{6}$$

$$\mathrm{EMA}_{ns}^{(t)} = m \cdot \mathrm{EMA}_{ns}^{(t-1)} + (1-m) \cdot \mathcal{L}_{ns}^{(t)}, \tag{7}$$

where $m \in (0, 1)$ is the EMA smoothing factor. A larger value of m assigns more weight to historical losses, resulting in a more stable estimate over time.

Once the smoothed losses EMA_s and EMA_{ns} are computed, we calculate their ratio r to measure the relative learning difficulty between branches:

$$r^{(t)} = \frac{\mathrm{EMA}_{ns}^{(t)}}{\mathrm{EMA}_s^{(t)} + \varepsilon}, \tag{8}$$

where ε is a small positive constant to avoid division by zero or numerical instability. Intuitively, $r^{(t)}$ reflects how much larger the non-dominant branch's smoothed loss is compared to that of the dominant branch.

Since $r^{(t)} \in (0, \infty)$ is unbounded, we apply a sigmoid transformation to $(r^{(t)} - 1)$ and introduce a baseline $b \in (0, 1)$ to constrain the resulting weight within $[b, 1]$. Specifically,

$$w_{ns}^{(t)} = b + (1-b)\,\mathrm{sigmoid}\big(\alpha(r^{(t)} - 1)\big), \tag{9}$$

$$w_s^{(t)} = 1 - w_{ns}^{(t)}, \tag{10}$$

where $\alpha > 0$ controls the sharpness of the transition. When $r^{(t)} \gg 1$, the loss of the non-dominant branch becomes significantly larger than that of the dominant branch, causing $w_{ns}^{(t)}$ to approach 1. Consequently, most of the optimization weight is assigned to the non-dominant branch.

Combining the weighted losses yields the final optimization objective:

$$\begin{aligned} \mathcal{L}_{\mathrm{ILDL}}^{(t)} &= w_s^{(t)} \cdot \mathrm{KL}\big(\overline{D} \,\|\, \overline{f}_\theta\big) + w_{ns}^{(t)} \cdot \mathrm{KL}\big(\widehat{D} \,\|\, \widehat{f}_\theta\big) \\ &= w_s^{(t)} \cdot \mathcal{L}_s^{(t)} + w_{ns}^{(t)} \cdot \mathcal{L}_{ns}^{(t)}. \end{aligned} \tag{11}$$

Through EMA smoothing and Sigmoid-based reweighting, the proposed method adaptively emphasizes the branch with higher smoothed loss, while still retaining attention to the other branch. Thus, this mechanism enables more effective modeling under imbalanced label distributions.

Momentum Allocation Strategy

Previous decoupling-based method, DILDL (Gao et al. 2025), divides label distributions into dominant and non-dominant branches but relies on a shared optimizer using a fixed momentum parameter. This uniform approach fails to account for the distinct optimization dynamics of each branch. In particular, the non-dominant branch often suffers from gradient noise and slower convergence.

To address this challenge, we propose a branch-specific momentum allocation strategy. Each branch is assigned an independent momentum coefficient, which is adaptively updated based on the EMA-smoothed loss values to better reflect the learning difficulty of each branch.

Let μ_s and μ_{ns} denote the momentum coefficients for the dominant and non-dominant branches, respectively. Motivated by the observation that non-dominant labels typically exhibit higher noise and instability, we impose the constraint $\mu_{ns} > \mu_s$. This grants the non-dominant branch greater historical inertia.

Inspired by DOT (Zhao et al. 2023a), we employ two independent momentum-based SGD optimizers to enable

branch-specific optimization dynamics. At iteration t, the dominant branch parameters are updated as follows:

$$\boldsymbol{v}_s^{(t)} = \mu_s^{(t-1)}\, \boldsymbol{v}_s^{(t-1)} \; + \; \nabla_{\boldsymbol{\theta}_s} \mathcal{L}_s^{(t)}, \tag{12}$$

$$\boldsymbol{\theta}_s^{(t)} = \boldsymbol{\theta}_s^{(t-1)} \; - \; \eta\, \boldsymbol{v}_s^{(t)}. \tag{13}$$

The non-dominant branch is updated similarly:

$$\boldsymbol{v}_{ns}^{(t)} = \mu_{ns}^{(t-1)}\, \boldsymbol{v}_{ns}^{(t-1)} \; + \; \nabla_{\boldsymbol{\theta}_{ns}} \mathcal{L}_{ns}^{(t)}, \tag{14}$$

$$\boldsymbol{\theta}_{ns}^{(t)} = \boldsymbol{\theta}_{ns}^{(t-1)} \; - \; \eta\, \boldsymbol{v}_{ns}^{(t)}. \tag{15}$$

Here, $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_{ns}$ denote the learnable parameters of the dominant and non-dominant branches respectively. $\boldsymbol{v}_s$ and $\boldsymbol{v}_{ns}$ are the corresponding momentum buffers. $\mathcal{L}_s^{(t)}$ and $\mathcal{L}_{ns}^{(t)}$ represent the KL divergence losses for the dominant and non-dominant branches at step t. The η denotes the learning rate.

The EMA updates for the dominant and non-dominant branches are defined in Eq. (6) and Eq. (7), and the loss ratio $r^{(t)}$ is defined in Eq. (8).

We then apply a sigmoid transformation to derive an adaptive modulation factor:

$$\delta^{(t)} = \sigma\big(\beta{\cdot}(r^{(t)}{-}1)\big), \quad \sigma(u) = \frac{1}{1+e^{-u}}, \quad \beta > 0, \tag{16}$$

where β controls the steepness of the sigmoid curve, larger values make the update more sensitive to deviations from $r^{(t)} = 1$.

Finally, the momentum coefficients are updated as follows:

$$\mu_{ns}^{(t)} = \mu_{ns}^{(t-1)} + \gamma\, \delta^{(t)}, \tag{17}$$

$$\mu_s^{(t)} = \mu_s^{(t-1)} - \gamma\, \delta^{(t)}, \tag{18}$$

where $\gamma > 0$ is the adaptation rate that controls how fast the momentum coefficients respond to the loss dynamics.

By combining the static constraint $\mu_{ns} > \mu_s$ with adaptive updates from Eq. (17) and Eq. (18), our momentum allocation strategy dynamically adjusts the optimizer for each branch based on its training difficulty. Theoretically, the variance of the momentum buffer for the non-dominant branch satisfies,

$$\mathrm{Var}[\boldsymbol{v}_{ns}^{(t)}] \leq \frac{1}{1-\mu_{ns}}\, \mathrm{Var}[\nabla \mathcal{L}_{ns}], \tag{19}$$

where a larger momentum coefficient suppresses stochastic gradient noise and stabilizes optimization under imbalanced scenarios (Liu, Gao, and Yin 2020; Zhao et al. 2023a). Under standard assumptions (convexity and bounded variance), momentum-based SGD achieves improved convergence rates, with the expected suboptimality bounded by,

$$\mathbb{E}[f(\boldsymbol{\theta}_{ns}^{(t)})] - f^* \leq \mathcal{O}\left(\frac{1}{(1-\mu)t}\right) + \mathcal{O}(\eta\sigma^2), \tag{20}$$

where σ^2 is the variance of the stochastic gradient noise and f^* is the global minimum of the loss function. Thus, assigning a larger momentum to the non-dominant branch theoretically accelerates convergence and enhances stability. By dynamically adjusting the momentum of each branch based on EMA-smoothed losses, the optimization inertia is aligned with learning difficulty, thereby enhancing the robustness.

Experiment

In this section, we evaluate the effectiveness of the proposed AMEMA on six datasets. We compare AMEMA with state-of-the-art DILDL methods and report performance across multiple evaluation metrics.

Datasets and Evaluation Metrics

We conduct experiments on 6 standard datasets: *SCUT-FBP* (Xie et al. 2015), *Flickr_LDL* (Yang, Sun, and Sun 2017), *Movie* (Geng and Hou 2015), *Emotion6* (Peng et al. 2015), *Natural Scene* (Geng 2016), and *RAF-ML* (Li and Deng 2019). These datasets exhibit varying degrees of label distribution skew, making them suitable for a comprehensive evaluation of AMEMA under ILDL scenarios.

To ensure robustness and generalization, we split each dataset into training, validation, and test subsets in a ratio of 80%, 10%, and 10%, respectively. Following the ILDL evaluation protocol (Zhao et al. 2023b; Gao et al. 2025), we adopt four distance-based metrics—Chebyshev↓, Clark↓, Canberra↓, and KL↓, where the arrows ↓ indicate that lower values are preferred. Additionally, we use two similarity-based metrics—Cosine↑ and Intersection↑, where ↑ indicates that higher values are preferred.

We compare AMEMA with 10 representative methods, which fall into three categories: (1) standard LDL algorithms including SA-BFGS (Geng 2016), EDL-LRL (Jia et al. 2019), LDLSF (Ren et al. 2019), and LDL-LCLR (Ren et al. 2019); (2) adaptive ILDL methods including Adam-LDL-SCL (Jia et al. 2019) and LDL-LDM (Wang and Geng 2021); and (3) specially designed ILDL approaches including OFR-FL (Zhao et al. 2025b), OFR-CB (Zhao et al. 2025b), OFR-DB (Zhao et al. 2025b), RDA (Zhao et al. 2023b), and DILDL (Gao et al. 2025).

Finally, we conduct sensitivity analyses on two key components of AMEMA, momentum allocation and dynamic branch weighting. Both components are crucial for regulating branch-specific optimization dynamics and ensuring optimal performance.

Implementation Details

All experiments are conducted under the PyTorch framework on an NVIDIA GeForce RTX 4060 GPU. The training setup includes a batch size of 64, an initial learning rate of 0.001, and a maximum of 50 epochs. We employ SGD with gradient clipping to improve training stability and prevent gradient explosion. Specifically, the dominant and non-dominant branches are trained with different initial momenta (0.75 and 0.95), which are dynamically adjusted based on EMA-smoothed KL divergence losses during training. Typically, $b = 0.2$ and $\alpha = 5$ are selected based on empirical observations. Each experiment is conducted using 10-fold cross-validation, and the results are reported as mean ± standard deviation to ensure statistical robustness. To assess statistical significance, two-tailed t-tests at the 0.05 level are used. In the result tables, solid circles (•) indicate our method is significantly better than the comparison, while hollow circles (○) indicate the results are tied.

Method	Movie	SCUT-FBP	Emotion6	Flickr_LDL	RAF-ML	Natural Scene
SA-BFGS	0.3415±0.0070•	0.7266±0.0326•	0.8292±0.0179•	0.8948±0.0149•	0.7575±0.0149•	0.6621±0.0198•
EDL-LRL	0.3638±0.0118•	0.3522±0.0236•	0.4175±0.0074•	0.5811±0.0060•	0.4784±0.0137•	0.4341±0.0233•
LDLSF	0.3624±0.0107•	0.4701±0.0307•	0.4355±0.0106•	0.5697±0.0092•	0.4177±0.0174•	0.4440±0.0249•
LDL-LCLR	0.3346±0.0072•	0.3332±0.0246•	0.5239±0.0136•	0.7033±0.0126•	0.3849±0.0107•	0.5680±0.0225•
Adam-LDL-SCL	0.7175±0.0487•	0.4460±0.0218•	0.4711±0.0333•	0.6711±0.0547•	0.5848±0.0300•	0.4773±0.0344•
LDL-LDM	0.4858±0.0285•	0.4030±0.0441•	0.4739±0.0159•	0.5816±0.0085•	0.5348±0.0275•	0.4769±0.0234•
OFR-FL	0.3416±0.0151•	0.3364±0.0357•	0.3910±0.0102•	0.5636±0.0054•	0.5081±0.0236•	0.4323±0.0201•
OFR-CB	0.3337±0.0177•	0.3447±0.0289•	0.3922±0.0091•	0.5658±0.0059•	0.5057±0.0161•	0.4329±0.0209•
OFR-DB	0.2548±0.0080•	0.3199±0.0384•	0.3772±0.0072•	0.5252±0.0205•	0.4638±0.0196•	0.3872±0.0254•
RDA	0.1962±0.0068•	0.2849±0.0157○	0.3598±0.0079•	0.5208±0.0075•	0.3756±0.0068•	0.3768±0.0208○
DILDL	0.1752±0.0091•	0.2684±0.0182•	0.3485±0.0061•	0.5025±0.0073•	0.3484±0.0102•	0.3624±0.0212•
AMEMA	**0.1703±0.0043**	**0.2115±0.0110**	**0.3358±0.0169**	**0.4531±0.0054**	**0.3433±0.0064**	**0.3243±0.0223**

Table 1: Experimental results on ILDL datasets measured by Chebyshev Distance (↓).

Method	Movie	SCUT-FBP	Emotion6	Flickr_LDL	RAF-ML	Natural Scene
SA-BFGS	0.8007±0.0539•	13.04±4.1007•	21.8514±1.0523•	27.1262±1.5508•	18.2051±1.2023•	4.7976±0.3734•
EDL-LRL	0.7797±0.0472•	0.8111±0.1085•	1.4348±0.1160•	9.9140±4.5756•	1.2838±0.0994•	2.5862±1.5835•
LDLSF	3.1338±0.3786•	8.4136±1.6575•	9.4371±0.5063•	12.8509±1.0510•	7.0684±1.1409•	8.8454±0.5594•
LDL-LCLR	0.6803±0.0314•	0.6034±0.0788•	2.2820±0.1581•	6.2168±0.2896•	1.0106±0.0704•	2.9449±0.2527•
Adam-LDL-SCL	19.1715±1.6303•	2.3768±1.1735•	8.1116±4.8903•	17.1944±8.5188•	6.1170±4.2557•	9.6209±4.8989•
LDL-LDM	1.8123±0.2788•	1.0253±0.2190•	1.7890±0.1369•	2.7424±0.2096•	1.9157±0.2248•	1.7753±0.2056•
OFR-FL	0.6459±0.0567•	0.6415±0.1438•	1.1829±0.0959•	2.5998±0.1650•	1.3672±0.1676•	1.3364±0.0981•
OFR-CB	0.6288±0.0604•	0.6581±0.1171•	1.1904±0.0776•	2.6285±0.3774•	1.3264±0.1110•	1.3280±0.0932•
OFR-DB	0.3883±0.0160•	0.5577±0.1317•	0.9238±0.0238•	1.7751±0.2858•	1.1481±0.0823•	1.1746±0.0898•
RDA	0.2491±0.0149•	0.4331±0.0328•	0.7677±0.0218•	1.6071±0.1107•	0.7058±0.0203•	1.1188±0.0591•
DILDL	0.2211±0.0121•	0.4131±0.0315•	0.7393±0.0254•	1.5011±0.0925•	0.6025±0.0320•	1.0734±0.0437•
AMEMA	**0.1924±0.0075**	**0.3253±0.0257**	**0.6611±0.0296**	**1.1771±0.0100**	**0.5455±0.0149**	**0.8344±0.0337**

Table 2: Experimental results on ILDL datasets measured by Kullback–Leibler divergence (↓).

Results

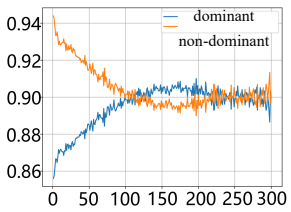
We evaluate the AMEMA method across six ILDL datasets: *Movie*, *SCUT-FBP*, *Emotion6*, *Flickr_LDL*, *RAF-ML*, and *Natural Scene*. As shown in Tables 1 and 2, AMEMA consistently outperforms both DILDL and RDA across all datasets in terms of Chebyshev distance and KL divergence. For example, on the *SCUT-FBP* dataset, AMEMA achieves a Chebyshev distance of 0.2115 ± 0.0110, which is significantly lower than DILDL's 0.2684 ± 0.0182. On the *Movie* dataset, AMEMA achieves a Clark distance of 0.6892 ± 0.0128, which is significantly lower than DILDL's 0.7683 ± 0.0315. On the *Flickr_LDL* dataset, AMEMA achieves a Canberra distance of 5.8434 ± 0.1102, which is significantly lower than DILDL's 6.1463 ± 0.0615. On the *RAF-ML* dataset, AMEMA achieves an Intersection similarity of 0.5488 ± 0.0084, which is significantly higher than DILDL's 0.5372 ± 0.0392. On the *Natural Scene* dataset, AMEMA achieves a Cosine coefficient of 0.6207 ± 0.0256, which is significantly higher than DILDL's 0.6103 ± 0.0207. On the *Emotion6* dataset, AMEMA achieves a KL divergence of 0.6611 ± 0.0296, clearly outperforming DILDL's 0.7393 ± 0.0254.

These results demonstrate the effectiveness of our optimization strategy in enhancing convergence and modeling of label distributions, leading to more accurate and reliable performance across diverse datasets.

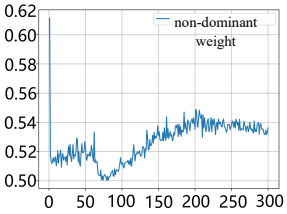
Analysis of Adaptive Loss Weighting Mechanism

We analyze the adaptive loss weighting mechanism in AMEMA from two perspectives, the loss trajectories of both the dominant and non-dominant branches, and the evolution of the non-dominant branch's weight during training. As shown in Figure 3(a), both branches exhibit a consistent downward trend in KL divergence, with the non-dominant branch showing a sharper decline. This suggests that the adaptive weighting mechanism effectively stabilizes learning dynamics under imbalanced label distributions, preventing the model from over-relying on the dominant branch.

In Figure 3(b), the weight assigned to the non-dominant branch follows a non-monotonic trajectory. At the early stage of training, the weight increases rapidly as the model focuses on reducing errors associated with the non-dominant branch. As the dominant branch stabilizes, the weight gradually decreases, reflecting a more balanced allocation of optimization effort. Toward the later stages, the weight increases again to alleviate potential overfitting to the dominant branch and sustain learning pressure on the non-dominant branch. These dynamics highlight the model's ability to adaptively modulate its optimization attention in response to evolving loss patterns across branches.

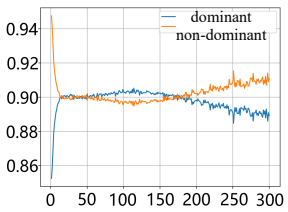


(a) Loss Evolution

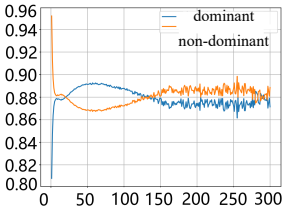


(b) Weight Variation

Figure 3: The curves of loss evolution for dominant and non-dominant branches during training (a), and weight variation for the non-dominant branch over time (b), on the imbalanced *Emotion6* dataset.



(a) Momentum on *SCUT-FBP*



(b) Momentum on *Movie*

Figure 4: The curves of momentum evolution for the dominant and non-dominant branches during training on the *SCUT-FBP* (a) and *Movie* (b) datasets.

WEIG	MOME	Cheb ↓	Clark ↓	Can ↓	KL ↓	Cos ↑	Inter ↑
		0.175	0.768	1.529	0.221	0.834	0.729
✓		0.172	0.701	1.312	0.204	0.844	0.733
	✓	0.173	0.712	1.311	0.199	0.848	0.741
✓	✓	**0.170**	**0.689**	**1.289**	**0.192**	**0.855**	**0.748**

(a) Ablation study on *Movie* dataset.

WEIG	MOME	Cheb ↓	Clark ↓	Can ↓	KL ↓	Cos ↑	Inter ↑
		0.362	2.479	6.907	1.073	0.610	0.409
✓		0.336	2.241	6.900	0.945	0.619	0.420
	✓	0.339	2.113	6.871	0.897	0.620	0.419
✓	✓	**0.324**	**2.041**	**6.801**	**0.834**	**0.621**	**0.424**

(b) Ablation study on *Natural Scene* dataset.

Table 3: Ablation results of WEIG and MOME on the *Movie* (a) and *Natural Scene* (b) datasets.

Analysis of Dynamic Momentum Allocation

To evaluate the effectiveness of the proposed momentum allocation strategy, we examine the evolution of branch-specific momentum coefficients, μ_s for the dominant branch and μ_{ns} for the non-dominant branch, throughout training. As shown in Figure 4, results on the *SCUT-FBP* and *Movie* datasets reveal a consistent trend that μ_s gradually decreases, while μ_{ns} steadily increases. This indicates that the optimizer progressively increases the historical inertia assigned to the non-dominant branch, which typically faces noisy gradients and unstable supervision, thereby enhancing its convergence and stability.

The divergence between the two momentum trajectories becomes most pronounced during the middle training phase, where overfitting to the dominant branch is likely to occur. Being guided by EMA-smoothed KL divergence losses, the optimizer shifts focus toward the non-dominant branch to maintain training balance. Both momentum values stabilize during the later epochs, demonstrating that the proposed Sigmoid-based modulation combined with EMA smoothing enables smooth and reliable adaptation. These dynamics confirm the strategy's capacity to improve convergence stability and learning robustness under imbalanced label distributions.

Ablation Study

To evaluate the contribution of each component in AMEMA, we conduct ablation experiments by independently disabling adaptive loss weighting (WEIG) and dynamic momentum allocation (MOME). Table 3 reports the results on the *Movie* and *Natural Scene* datasets under six evaluation metrics.

Disabling either component results in significant performance degradation. Removing adaptive weighting by assigning equal weights to both branches reduces the model's ability to balance attention between dominant and non-dominant labels, leading to suboptimal learning. Fixing momentum values across branches weakens the stability of the non-dominant branch optimization due to its noisy gradient signals. Disabling both mechanisms reduces AMEMA to a standard ILDL baseline, causing further performance decline. In contrast, the full AMEMA configuration consistently achieves the lowest KL divergence, demonstrating that the combination of dynamic weighting and momentum allocation is crucial for improving convergence and enhancing model robustness under imbalanced label distributions. In summary, the ablation experiments demonstrate that removing WEIG and MOME from the model leads to deteriorated experimental results. Therefore, the WEIG and MOME methods we employ are effective in enhancing the model's performance.

Conclusion

In this study, we address the challenge of imbalanced label distribution learning by introducing AMEMA, a dual-strategy framework that decouples the learning process into dominant and non-dominant branches. AMEMA integrates dynamic momentum allocation and EMA-based adaptive weighting to effectively balance optimization across branches. By jointly regulating optimization strength and learning direction, the proposed framework achieves robust convergence and generalization, providing a principled solution for real-world ILDL scenarios. Extensive experiments on six benchmark datasets show that AMEMA consistently outperforms state-of-the-art methods across multiple evaluation metrics. In future work, we plan to extend AMEMA to more complex learning paradigms such as multi-modal and multi-task ILDL, aiming to further enhance its adaptability. We also intend to investigate the theoretical convergence properties of its adaptive momentum mechanism to establish stronger guarantees for stable optimization.

Acknowledgments

This work was supported by the project No. ZR2024QF115 supported by Shandong Provincial Natural Science Foundation, the National Natural Science Foundation of China under No. 62406155 and No. 62476135, the Innovation Capability Enhancement Project for Technology-based Small and Medium-sized Enterprises of Shandong Province under Grant No. 2025TSGCCZZB0077 and No. 2024TSGC0777, the National Science Foundation of China (Nos. 62002187), the Shandong Provincial Natural Science Foundation (No. ZR2024QF306), the Youth Innovation Team of Colleges and Universities in Shandong Province (2023KJ331, 2024KJH032), the approved research project of the Shandong Provincial Health Commission for the breakthrough initiative on large models (including DeepSeek) in the healthcare sector, the Open Funding of Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Application (Southeast University), Ministry of Education, China.

References

Algabri, R.; Abdu, A.; and Lee, S. 2024. Deep Learning and Machine Learning Techniques for Head Pose Estimation: A Survey. *Artificial Intelligence Review*, 57(10): 1–288.

Cai, Z.; Ravichandran, A.; Maji, S.; Fowlkes, C.; Tu, Z.; and Soatto, S. 2021. Exponential Moving Average Normalization for Self-Supervised and Semi-Supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 194–203.

Cheng, Z.; Cheng, Z.-Q.; He, J.-Y.; Wang, K.; Lin, Y.; Lian, Z.; Peng, X.; and Hauptmann, A. 2024. Emotion-llama: Multimodal Emotion Recognition and Reasoning with Instruction Tuning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 37, 110805–110853.

Fan, Y.; Liu, J.; Tang, J.; Liu, P.; Lin, Y.; and Du, Y. 2024. Learning Correlation Information for Multi-Label Feature Selection. *Pattern Recognition*, 145: 109899.

Gao, Y.; Sun, X.; Ling, M.; Tan, C.; Zhai, Y.; and Lv, G. 2025. Decoupled Imbalanced Label Distribution Learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Geng, X. 2016. Label Distribution Learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7): 1734–1748.

Geng, X.; and Hou, P. 2015. Pre-Release Prediction of Crowd Opinion on Movies by Label Distribution Learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 3511–3517.

Geng, X.; Yin, C.; and Zhou, Z.-H. 2013. Facial Age Estimation by Learning from Label Distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10): 2401–2412.

Han, P.; Ye, C.; Zhou, J.; Zhang, J.; Hong, J.; and Li, X. 2024. Latent-Based Diffusion Model for Long-Tailed Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2639–2648.

He, H.; Zhao, S.; Xi, Y.; and Ho, J. 2021. AGE: Enhancing the Convergence on GANs Using Alternating Extra-Gradient with Gradient Extrapolation. In *Proceedings of the NeurIPS Workshop on Deep Generative Models and Downstream Applications*. Virtual.

Hou, S.; Jiang, G.; Zhang, J.; Yang, S.; Guo, H.; Guo, Y.; and Wang, W. 2025. Directional Label Diffusion Model for Learning from Noisy Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25738–25748.

Jia, X.; Li, Z.; Zheng, X.; Li, W.; and Huang, S.-J. 2019. Label Distribution Learning with Label Correlations on Local Samples. *IEEE Transactions on Knowledge and Data Engineering*, 33(4): 1619–1631.

Jia, X.; Qin, T.; Lu, Y.; and Li, W. 2023. Adaptive Weighted Ranking-Oriented Label Distribution Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35: 6041–6052.

Jiang, Y.; Wang, Z.; and Chang, H. 2024. Head Pose Estimation via mmWave Radar. In *Proceedings of the IEEE Global Communications Conference*, 259–264. IEEE.

Jin, Y.; Gao, R.; He, Y.; and Zhu, X. 2024. GLDL: Graph Label Distribution Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12965–12974.

Kendall, A.; and Gal, Y. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Proceedings of the Advances in Neural Information Processing Systems*, volume 30.

Kim, J.; Jeong, J.; and Shin, J. 2020. M2M: Imbalanced Classification via Major-to-Minor Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13896–13905.

Kim, J.-H.; Yun, S.; and Song, H. O. 2023. Neural Relation Graph: A Unified Framework for Identifying Label Noise and Outlier Data. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, 43754–43779.

Kim, N.-r.; Lee, J.-S.; and Lee, J.-H. 2024. Learning with Structural Labels for Learning with Noisy Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27610–27620.

Lakkapragada, A.; Sleiman, E.; Surabhi, S.; and Wall, D. P. 2023. Mitigating Negative Transfer in Multi-Task Learning with Exponential Moving Average Loss Weighting Strategies (Student Abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 16246–16247.

Lee, T.; Shin, W.; Lee, J.-H.; Lee, S.; Yeom, H.-G.; and Yun, J. P. 2024. Resolving the Non-Uniformity in the Feature Space of Age Estimation: A Deep Learning Model Based on Feature Clusters of Panoramic Images. *Computerized Medical Imaging and Graphics*, 112: 102329.

Li, D.; Wang, S.; Zhao, W.; Kang, L.; Dong, L.; Wang, J.; and Wang, X. 2025. ADGaze: Anisotropic Gaussian Label Distribution Learning for Fine-grained Gaze Estimation. *Pattern Recognition*, 164: 111536.

Li, S.; and Deng, W. 2019. Blended Emotion in-the-Wild: Multi-Label Facial Expression Recognition Using Crowdsourced Annotations and Deep Locality Feature Learning. *International Journal of Computer Vision*, 127(6): 884–906.

Li, W.; Qian, W.; Chen, L.; and Jia, X. 2024. Sample Diversity Selection Strategy Based on Label Distribution Morphology for Active Label Distribution Learning. *Pattern Recognition*, 150: 110322.

Liu, S.; Huang, E.; Zhou, Z.; Xu, Y.; Kui, X.; Lei, T.; and Meng, H. 2025. Lightweight Facial Attractiveness Prediction Using Dual Label Distribution. *IEEE Transactions on Cognitive and Developmental Systems*, 17. Early Access.

Liu, Y.; Gao, Y.; and Yin, W. 2020. An Improved Analysis of Stochastic Gradient Descent with Momentum. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, 18261–18271.

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

Peng, K.-C.; Chen, T.; Sadovnik, A.; and Gallagher, A. C. 2015. A Mixed Bag of Emotions: Model, Predict, and Transfer Emotion Distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 860–868.

Rachavarapu, K. K.; Ramakrishnan, K.; et al. 2024. Weakly-Supervised Audio-Visual Video Parsing with Prototype-Based Pseudo-Labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18952–18962.

Rachavarapu, K. K.; et al. 2023. Boosting Positive Segments for Weakly-Supervised Audio-Visual Video Parsing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10192–10202.

Ren, T.; Jia, X.; Li, W.; Chen, L.; and Li, Z. 2019. Label Distribution Learning with Label-Specific Features. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 3–10.

Sheng, M.; Sun, Z.; Pei, G.; Chen, T.; Luo, H.; and Yao, Y. 2024. Enhancing Robustness in Learning with Noisy Labels: An Asymmetric Co-Training Approach. In *Proceedings of the ACM International Conference on Multimedia*, 4406–4415.

Wang, J.; and Geng, X. 2019. Theoretical Analysis of Label Distribution Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5256–5263.

Wang, J.; and Geng, X. 2021. Learn the Highest Label and Rest Label Description Degrees. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 3097–3103.

Wang, X.; and Li, J. 2025. SBPM Model for Analyzing Students' Learning Behavior Based on Fine Grained Emotion Analysis and Emotion Assessment. *Informatica (Slovenia)*, 49(7): 187–200.

Wang, Z.; and Wang, H. 2023. Variational Imbalanced Regression: Fair Uncertainty Quantification via Probabilistic Smoothing. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, 30429–30452.

Wu, S.; He, D.; Wang, X.; Wang, L.; and Dang, J. 2025. Enriching Multimodal Sentiment Analysis through Textual Emotional Descriptions of Visual-Audio Content. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1601–1609. Vancouver, Canada.

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

Xu, H.; Liu, X.; Zhao, Q.; Ma, Y.; Yan, C.; and Dai, F. 2023. Gaussian Label Distribution Learning for Spherical Image Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1033–1042.

Yang, J.; Sun, M.; and Sun, X. 2017. Learning Visual Sentiment Distributions via Augmented Conditional Probability Neural Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 3801–3808.

Yao, S.; Wang, M.; Ren, J.; Xia, T.; Wang, W.; Xu, K.; Xu, M.; and Zhang, H. 2025. Multi-Agent Reinforcement Learning for Task Offloading in Crowd-Edge Computing. *IEEE Transactions on Mobile Computing*. Early Access.

You, N.; Zhao, X.; Gao, Z.; and Song, X. 2025. Explainable Label Distribution Learning by Exploiting Neighborhood. *International Journal of Machine Learning and Cybernetics*, 16(2): 182–199.

Yu, Z.; Chen, R.; Gui, P.; Ju, L.; Shang, X.; Zhu, Z.; He, M.; and Ge, Z. 2023. Retinal Age Estimation with Temporal Fundus Images Enhanced Progressive Label Distribution Learning. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 629–638. Springer.

Zhao, B.; Cui, Q.; Song, R.; and Liang, J. 2023a. Dot: A Distillation-Oriented Trainer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6189–6198.

Zhao, P.; Zhou, J.; Zhao, Y.; Guo, D.; and Chen, Y. 2025a. Multimodal Class-Aware Semantic Enhancement Network for Audio-Visual Video Parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10448–10456. Vancouver, Canada.

Zhao, Q.; Xia, Y.; Long, Y.; Xu, G.; and Wang, J. 2025b. Leveraging Sensory Knowledge into Text-to-Text Transfer Transformer for Enhanced Emotion Analysis. *Information Processing & Management*, 62(1): 103876.

Zhao, X.; An, Y.; Xu, N.; Wang, J.; and Geng, X. 2023b. Imbalanced Label Distribution Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 11336–11344.

Zou, Z.; Jia, D.; and Tang, W. 2025. Towards Unsupervised Learning of Joint Facial Landmark Detection and Head Pose Estimation. *Pattern Recognition*, 162: 111393.