D2ACE: Multi-Label Batch Selection Guided by Dual Dynamics and Adaptive Correlation Enhancement

Bin Liu[1*], **Haoyu Peng**[1,2], **Zhijia Wei**[1,2], **Jiajing Zhang**[1], **Grigorios Tsoumakas**[3]
[1] Key Laboratory of Data Engineering and Visual Computing, Chongqing University of Posts and Telecommunications, China
[2] School of Computer Science and Technology, Chongqing University of Posts and Telecommunications
[3] School of Informatics, Aristotle University of Thessaloniki, Greece
liubin@cqupt.edu.cn, {s240231185,s250201123,s250331042}@stu.cqupt.edu.cn, greg@csd.auth.gr

Abstract

Batch selection is crucial for improving both training efficiency and predictive performance in deep multi-label classification (MLC). Existing batch selection methods typically rely on a single metric to assess instance importance and use static label weights to distinguish label significance, neglecting the dynamic evolution of metric utility and label significance during training. In addition, the method that explicitly exploits label correlations is largely affected by abundant irrelevant labels and insensitive to local label distributions. To address these issues, we propose D2ACE, a novel multi-label batch selection method guided by Dual Dynamics and Adaptive Correlation Enhancement. D2ACE explicitly captures metric and label-level training dynamics by combining stage-wise Bernoulli mixture sampling, which balances uncertainty and noise-resistant hardness, with dynamic label weighting to recalibrate label priorities at each epoch based on current metric statistics. Furthermore, D2ACE introduces a local context-aware correlation enhancement to focus on relevant labels with instance-adaptive dependencies. Extensive experiments on tabular and image benchmarks demonstrate that D2ACE outperforms existing batch selection approaches across various deep MLC models, achieving stronger predictive performance and more efficient correlation modeling.

1 Introduction

By modeling multiple outputs simultaneously, multi-label classification (MLC) methods are able to capture complex real-world phenomena where multiple concepts often co-occur and label dependencies are intrinsic [Shou *et al.*, 2023]. This flexibility has led to the extensive application of MLC methods in domains such as text classification [Chai *et al.*, 2024; Lin *et al.*, 2023], image annotation [Guo *et al.*, 2023], and protein function prediction [Bai *et al.*, 2024].

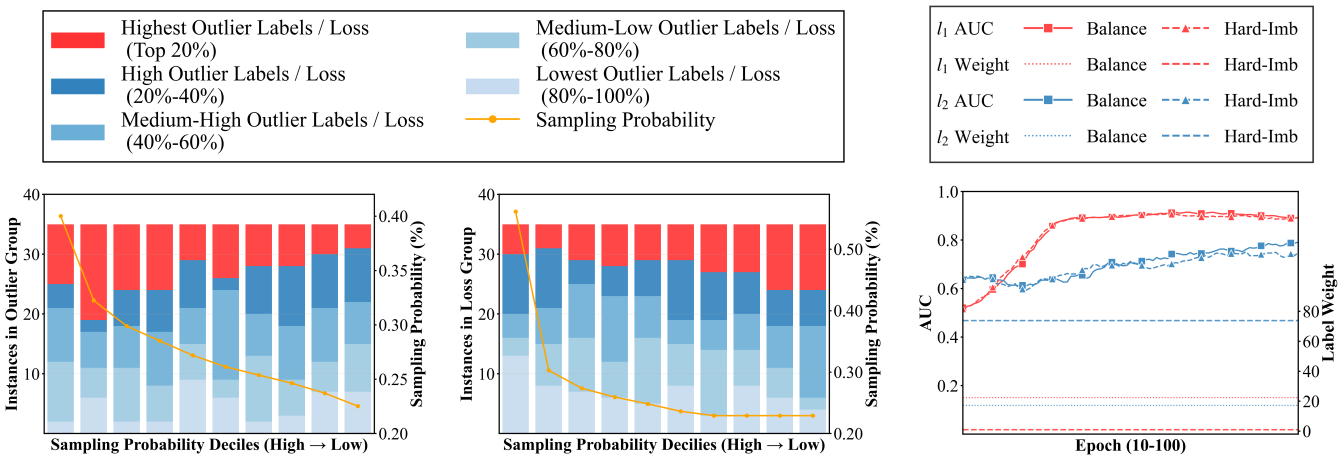
Recently, deep learning architectures for MLC have achieved significant progress. Deep MLC models automatically extract semantically rich features from input data and predict multiple labels for each instance simultaneously [Liu *et al.*, 2021]. Using techniques such as prompt-tuning [Guo *et al.*, 2023], attention mechanisms [Lanchantin *et al.*, 2021], and graph neural networks [Hang and Zhang, 2021], these models capture complex feature–label and label–label dependencies, improving both accuracy and generalization.

Batch selection plays a key role in improving both training efficiency and prediction accuracy in deep learning. Instance importance is typically measured using hardness and uncertainty. Hardness-based methods assess the learning difficulty of instances, guiding the model to focus on easy, hard, or near-boundary samples [Loshchilov and Hutter, 2015; Zhou *et al.*, 2020; Song *et al.*, 2020b], whereas uncertainty-based methods prioritize instances with unstable or ambiguous predictions [Chang *et al.*, 2017; Song *et al.*, 2020a]. While existing multi-label batch selection methods rely on either hardness [Zhou *et al.*, 2024a] or uncertainty [Zhou *et al.*, 2025], both metrics have clear limitations when applied in isolation (see Figure 1(a) and 1(b)). Hardness-based selection (Hard-Imb) can overemphasize noisy or outlier samples during early training, increasing the risk of overfitting. In contrast, uncertainty-based selection (ML-Unc) ignores genuinely hard (high-loss) instances that exhibit low uncertainty but remain misclassified in later training stages. Current methods lack a systematic approach to harmonize the complementary strengths of these two metrics.

In MLC, different labels can exhibit distinct and evolving importance throughout training, as evidenced by their AUC trajectories. As shown in Figure 1(c), although label l_1 has a lower AUC than l_2 in the early stages, its AUC increases more rapidly as training progresses. However, existing multi-label batch selection methods, such as Balance [Hand *et al.*, 2018] and Hard-Imb [Zhou *et al.*, 2024a], rely on fixed label priorities and predefined static weights, failing to adapt to the evolving label learning dynamics.

Moreover, label correlations play a crucial role in capturing complex output structures and improving performance in multi-label models. ML-Unc [Zhou *et al.*, 2025] pioneers the incorporation of label correlations into batch selection by prioritizing instances that exhibit synergistic uncertain behavior across multiple labels. However, it propagates correlation-enhanced uncertainty uniformly over all

*Corresponding author.



(a) Hard-Imb (epoch 20) (b) ML-Unc (epoch 80) (c) Label Weight

Figure 1: (a) and (b): Distribution of instance properties across sampling probability deciles for Hard-Imb and ML-Unc using CLIF base model on the CAL500 dataset, where the x-axis shows deciles of instances sorted by descending sampling probability, and color bars indicate counts from each outlier/loss group within each decile. (c): The validation AUC curves and label weight changes of different labels during training for Balance and Hard-Imb using static label weights.

labels, implicitly assuming equal contributions from relevant and irrelevant labels as well as the same label dependencies across instances. The inherent sparsity of multi-label datasets, where each instance is associated with only a few relevant labels [Ning *et al.*, 2025; Liu *et al.*, 2025; Zhou *et al.*, 2024b], causes ML-Unc to disproportionately benefit irrelevant labels over more informative relevant ones. Furthermore, the globally estimated correlation-based metric score overlooks local variations in co-occurrence, as globally correlated labels may not co-exist in every local region.

To address the above issues, we propose D2ACE, a novel multi-label batch selection strategy that considers training dynamics from both metric and label perspectives while effectively leveraging label correlations. First, D2ACE jointly integrates uncertainty and noise-resistant-hardness through a stage-wise Bernoulli mixture sampling scheme, which dynamically shifts the sampling focus from uncertain to hard instances as training progresses. This stage-adaptive design exploits the complementary strengths of the two metrics, enabling the selection of informative instances at different learning stages. Second, D2ACE introduces a dynamic label weighting scheme that recalibrates label priority at each epoch based on current metric statistics to capture the evolving importance of labels during training. Third, D2ACE develops a local context-aware label correlation enhancement that explicitly focuses on relevant labels to prevent the dominance of irrelevant ones and incorporates local label context to exploit instance-adaptive dependencies, thereby adaptively amplifying informative metric scores and improving the efficiency of correlation mining.

Our main contributions are summarized as follows:

- **Training Dynamics**: We propose a unified multi-label batch selection framework D2ACE that concurrently addresses metric dynamics by adaptively balancing uncertainty and hardness through stage-wise Bernoulli mixture sampling while tackling label dynamics via epoch-wise recalibration of label priorities.
- **Correlation Mining**: We develop a novel label correlation enhancement strategy that estimates relevance-filtered label correlations and propagates correlations based on local label context, facilitating efficient metric enhancement in an instance-adaptive manner.
- **Empirical Validation**: Our method achieves the largest performance improvement compared with state-of-the-art single- and multi-label batch selection methods, demonstrating consistent superiority across diverse deep multi-label models on tabular and image datasets.

2 Related Work

2.1 Multi-Label Classification

Traditional methods for classifying multi-label tabular data fall into two main categories: algorithm adaptation, which extends conventional machine learning models to directly handle multi-label outputs [Zhang and Zhou, 2007; Yang *et al.*, 2020], and problem transformation, which converts the MLC task into multiple single-label classification problems via various label space transformations [Zhang *et al.*, 2018; Tsoumakas *et al.*, 2010; Li *et al.*, 2023]. Recently, deep learning models have advanced MLC by effectively capturing feature–label and label–label dependencies. CLIF [Hang and Zhang, 2021] integrates a graph autoencoder to encode label semantics to guide informative label-specific feature extraction. HOT-VAE [Zhao *et al.*, 2021] employs attention mechanisms to adaptively exploit high-order label dependencies. DELA [Hang and Zhang, 2024] identifies label-specific non-informative features and learns robust classifiers through stochastic feature perturbation. PACA [Hang *et al.*, 2022] formulates an end-to-end probabilistic framework based on prototype-based latent metric spaces with label correlation

	Method	Metric	Metric dynamics	Label dynamics	Label correlation	Local context
single -label	Active [Chang *et al.*, 2017]	uncertainty	-	-	-	-
	Recent [Song *et al.*, 2020a]	uncertainty	-	-	-	-
	DIHCL [Zhou *et al.*, 2020]	hardness	-	-	-	-
multi -label	Balance [Hand *et al.*, 2018]	imbalance	-	✗	✗	✗
	Hard-Imb [Zhou *et al.*, 2024a]	hardness & imbalance	✗	✗	✗	✓
	ML-Unc [Zhou *et al.*, 2025]	uncertainty	-	✗	✓	✗
	D2ACE (Ours)	hardness & uncertainty	✓	✓	✓	✓

Table 1: The summary of single-label and multi-label batch selection approaches.

regularization. FLEM [Zhao *et al.*, 2025] integrates label enhancement into model training to jointly optimize label importance recovery and predictive learning.

Multi-label image classification methods typically employ pre-trained ResNet [He *et al.*, 2016] backbones to extract visual features from raw images, upon which multi-label prediction is performed with label dependency modeling. ML-GCN [Chen *et al.*, 2019] captures global label correlations by generating inter-dependent classifiers from semantic label representations using graph convolutional networks, enabling shared parameter learning across labels. C-Tran [Lanchantin *et al.*, 2021] is a transformer-based method that models label dependencies via self-attention and reconstruction of masked label embeddings conditioned on visual features, without relying on predefined label graphs. HST [Chen *et al.*, 2024] exploits both intra-image co-occurrence and cross-image semantic similarity through heterogeneous semantic transfer. TAI++ [Wu *et al.*, 2024] leverages pseudo-visual prompts and a dual-adapter co-learning strategy to transfer visual knowledge from text to image representations.

2.2 Batch Selection

Single-label batch selection methods typically prioritize instances based on either uncertainty or learning difficulty. To measure uncertainty, Active [Chang *et al.*, 2017] estimates prediction variance over the training process, while Recent [Song *et al.*, 2020a] focuses on the entropy of recent predictions. In contrast, DIHCL [Zhou *et al.*, 2020] characterizes sample hardness by tracking loss variation and prediction flips between consecutive epochs, and employs an exponential moving average to progressively reduce the influence of earlier training stages. Multi-label batch selection methods extend these metrics by operating on sample–label pairs and explicitly accounting for label-specific characteristics. The Balance method [Hand *et al.*, 2018] constructs mini-batches to match desired label distributions based on global label imbalance. Hard-Imb [Zhou *et al.*, 2024a] integrates static local label imbalance with label-wise loss values to identify hard instances. ML-Unc [Zhou *et al.*, 2025] proposes an uncertainty metric that combines current prediction reliability with fine-grained changes in recent outputs, and further exploits global label correlations to emphasize instances exhibiting synergistic uncertainty across more labels.

As summarized in Table 1, D2ACE is distinguished from existing methods by the joint consideration of hardness and uncertainty in a stage-aware manner, dynamic label weighting, and local context-aware label correlation enhancement.

3 Method

3.1 Preliminary

Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ be a multi-label dataset consisting of n instances, where the i-th instance consists of a feature vector $\mathbf{x}_i \in \mathbb{R}^d$ and a corresponding binary label vector $\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{iq}] \in \{0,1\}^q$. Given a label set $\mathcal{L} = \{l_1, l_2, \ldots, l_q\}$, $y_{ij} = 1$ indicates that $\mathbf{x}_i$ is relevant to label l_j, and $y_{ij} = 0$ otherwise. The objective of multi-label classification is to learn a mapping function $\mathbb{R}^d \to \{0,1\}^q$ that predicts the label subset associated with unseen instances. To accommodate memory and computational limitations during training, deep multi-label model parameters are updated iteratively using mini-batches $\mathcal{B} = \{(\mathbf{x}_k, \mathbf{y}_k)\}_{k=1}^b \subseteq \mathcal{D}$.

3.2 Importance Metrics

Focusing exclusively on either uncertain or hard samples can be suboptimal, as it may overlook persistently misclassified instances in later training stages or lead to overfitting noisy outliers in early epochs. To address this issue, D2ACE employs the uncertainty metric from [Zhou *et al.*, 2025] and further develops a hardness metric that integrates prediction stability with loss to mitigate the effect of noise.

Uncertainty Metric

We adopt the uncertainty metric of [Zhou *et al.*, 2025] that jointly accounts for both the ambiguity of the current model prediction and the temporal instability of recent predictions. Let $\hat{y}_{ij}^t \in [0,1]$ denote the predicted probability that instance $\mathbf{x}_i$ is associated with label l_j at epoch t. The uncertainty of this prediction is assessed using binary entropy:

$$e_{ij}^t = -\left(\hat{y}_{ij}^t \log_2 \hat{y}_{ij}^t + (1-\hat{y}_{ij}^t)\log_2(1-\hat{y}_{ij}^t)\right), \quad (1)$$

which attains its maximum when $\hat{y}_{ij}^t = 0.5$. The metric also captures prediction variation over the most recent n_t epochs by averaging probability changes between adjacent epochs:

$$d_{ij}^t = \frac{1}{n_t - 1} \sum_{k=1}^{n_t-1} \left|\hat{y}_{ij}^{t-k+1} - \hat{y}_{ij}^{t-k}\right|, \quad (2)$$

where larger values reflect higher fluctuations.

The overall uncertainty of instance $\mathbf{x}_i$ with respect to label l_j at epoch t is obtained by aggregating Eq.(1) and Eq.(2):

$$u_{ij}^t = \lambda_1 e_{ij}^t + (1-\lambda_1) d_{ij}^t, \quad (3)$$

where $\lambda_1 \in [0,1]$ balances the contribution of current prediction uncertainty and temporal instability. Finally, the uncertainty scores u_{ij}^t for all training instances and labels collectively form the uncertainty matrix $\mathbf{U}^t \in \mathbb{R}^{n \times q}$.

Hardness Metric

The instantaneous training loss (e.g., binary cross-entropy) is a commonly used indicator for assessing instance difficulty [Zhou *et al.*, 2020; Zhou *et al.*, 2024a]. However, relying solely on model loss incurs the risk of overfitting to noisy instances. Prior studies found that predictions for noisy (outlier) samples tend to oscillate frequently during training, whereas genuinely hard examples are more likely to be misclassified consistently over long periods [Toneva *et al.*, 2018; Maini *et al.*, 2022]. Motivated by this observation, we disentangle hard samples from noisy ones by incorporating the prediction flip into the hardness metric.

Let $\tilde{y}_{ij}^t \in \{0,1\}$ be the binary relevance prediction of instance $\mathbf{x}_i$ for label l_j at epoch t. The prediction flip indicator between two consecutive epochs is defined as $f_{ij}^t = \left|\tilde{y}_{ij}^t - \tilde{y}_{ij}^{t-1}\right|$, capturing whether the predicted relevance of l_j for $\mathbf{x}_i$ changes across two successive epochs. To accumulate flip information and capture long-term prediction stability, we use an exponential moving average [Zhou *et al.*, 2020]:

$$\bar{f}_{ij}^t = \lambda_2 f_{ij}^t + (1-\lambda_2)\bar{f}_{ij}^{t-1}, \tag{4}$$

where $\lambda_2 \in (0,1]$ is the smoothing parameter. Larger value of $\bar{f}_{ij}^t$ indicates higher long-term prediction instability, suggesting a higher likelihood that the label y_{ij} is noisy rather than intrinsically difficult. Then, we define the hardness metric by integrating the current loss and prediction stability:

$$h_{ij}^t = \ell_{ij}^t\big(1-\bar{f}_{ij}^t\big), \tag{5}$$

where ℓ_{ij}^t denotes the loss of instance $\mathbf{x}_i$ with respect to label l_j at epoch t. This metric prioritizes sample-label pairs that consistently exhibit high loss with stable incorrect predictions, while suppressing those dominated by frequent prediction flips that are more likely caused by label noise. The hardness scores h_{ij}^t over all instances and labels constitute the hardness matrix $\mathbf{H}^t \in \mathbb{R}^{n\times q}$.

3.3 Dynamic Label Weight

To adaptively capture the evolving importance (e.g., uncertainty and hardness) of various labels during training, D2ACE introduces *dynamic label weights* that assess the priority of each label across epochs. Given a metric matrix $\mathbf{A} \in \{\mathbf{U}, \mathbf{H}\}$ representing either uncertainty or hardness at the current epoch[1], we first compute label-wise (column-wise) means and standard deviations:

$$\boldsymbol{\mu}_A = \frac{1}{n}\mathbf{1}_n^\top \mathbf{A}, \quad \boldsymbol{\sigma}_A = \sqrt{\frac{1}{n}\mathbf{1}_n^\top(\mathbf{A}\odot\mathbf{A}) - \boldsymbol{\mu}_A \odot \boldsymbol{\mu}_A}, \tag{6}$$

where $\mathbf{1}_n$ is a row vector of ones and $\odot$ denotes element-wise multiplication. Labels with higher mean values consistently exhibit stronger metric values across many instances, indicating overall importance for training, while labels with higher variance highlight the presence of instances with particularly large values that should be prioritized. We then combine these two factors using an element-wise exponential transformation to obtain label weights:

$$\mathbf{v}_A = \exp\Big(\tfrac{1}{2}(\boldsymbol{\mu}_A + \boldsymbol{\sigma}_A)\Big), \tag{7}$$

and scale the metric matrix column-wise to highlight important labels:

$$\mathbf{Q}_A = \mathbf{A}\,\mathrm{diag}(\mathbf{v}_A). \tag{8}$$

Finally, instance weights that reflect dynamic label-specific importance are obtained via row-wise aggregating the weighted matrix, i.e., $\boldsymbol{\delta}_A = \mathbf{Q}_A\mathbf{1}_q^\top \in \mathbb{R}^n$.

3.4 Local Context-Aware Label Correlation Enhancement

To exploit label dependencies, we propose a *local context-aware label correlation enhancement* strategy that selectively amplifies informative metric scores in an instance-adaptive manner. Let $\mathbf{Y} = [\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top]^\top \in \{0,1\}^{n\times q}$ be the ground-truth label matrix of the training set. To prevent irrelevant labels from dominating correlation construction, we first apply relevance filtering by constructing a label-masked metric matrix $\mathbf{M}_A = \mathbf{Y}\odot\mathbf{A}$. This operation explicitly focuses on correlations among relevant labels (i.e., $y_{ij}=1$), thereby reducing the interference from abundant irrelevant (negative) labels and improving the ability to capture informative dependencies for positive label prediction [Yuan *et al.*, 2024]. Based on the masked metric, we estimate global label correlations using cosine similarity:

$$\mathbf{C}_A = \bar{\mathbf{M}}_A^\top \bar{\mathbf{M}}_A, \tag{9}$$

where $\bar{\mathbf{M}}_A = \mathbf{M}_A\mathrm{diag}(||[\mathbf{M}_A]_{:,1}||_2, \cdots, ||[\mathbf{M}_A]_{:,q}||_2)$ and $[\mathbf{M}_A]_{:,j}$ is the j-th column of $\mathbf{M}_A$. A larger value of $[C_A]_{ij}$ indicates stronger synchronous patterns of metric values between labels l_i and l_j at the dataset level.

To further incorporate *local label context* into global correlation modeling, we introduce a local label appearance matrix $\mathbf{Z} \in \{0,1\}^{n\times q}$:

$$Z_{ij} = \mathbb{I}\Big(\sum_{\mathbf{x}_m \in \mathcal{N}(\mathbf{x}_i)} y_{mj} \geq 1\Big), \tag{10}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and $\mathcal{N}(\mathbf{x}_i)$ identifies the local region of $\mathbf{x}_i$, which can be defined as $\mathbf{x}_i$ itself or its K-nearest neighbors. An entry $Z_{ij}=1$ indicates at least one instance in $\mathcal{N}(\mathbf{x}_i)$ associating with label l_j. We then integrate global dependency and local context to perform instance-adaptive correlation enhancement:

$$\mathbf{R}_A = \mathbf{M}_A \odot \big(\mathbf{Z}_A\mathbf{C}_A\big), \tag{11}$$

where only metric scores associated with relevant labels are amplified, and the degree of increase is determined by both global label correlations and their local presence around each instance. Our correlation enhancement focuses correlation estimation on relevant labels and explicitly conditions correlation propagation on local label context, addressing irrelevant label correlation dominance and neglect of instance-specific information in ML-Unc (see Appendix A.1 for a visual example). Finally, we obtain instance-level weights that capture local context-aware label correlations via row-wise summarization, i.e., $\boldsymbol{\phi}_A = \mathbf{R}_A\mathbf{1}_q^\top \in \mathbb{R}^n$.

Remarks: ML-Unc can be regarded as a degenerate variant of our label correlation enhancement, which does not consider label relevance filtering and defines the local region as the entire dataset (see Appendix A.2 for details).

[1]For simplicity, when referring to variables at the current epoch t, the superscript is omitted if there is no ambiguity.

3.5 Sampling Probability

At each epoch t, we first compute the *final instance weight* by integrating label dynamic importance and correlation-enhanced information:

$$\mathbf{w}_A = \boldsymbol{\delta}_A + \boldsymbol{\phi}_A, \tag{12}$$

where $\boldsymbol{\delta}_A$ captures label dynamics and $\boldsymbol{\phi}_A$ encodes label correlation-aware contributions.

To convert instance weights into sampling probabilities, we adopt the quantization index strategy [Song *et al.*, 2020a; Zhou *et al.*, 2024a]. Following the exponential decay strategy that captures the weight changes across epochs, the weight $[w_A]_i \in [0, 1]$ of $\mathbf{x}_i$ is first mapped to a quantization index:

$$Q([w_A]_i) = \left\lceil 1 - \frac{[w_A]_i}{\Delta} \right\rceil, \tag{13}$$

where $\Delta = \frac{1}{n}$ is the quantization step size. The sampling probability of $\mathbf{x}_i$ at the current epoch t is defined as:

$$P(\mathbf{x}_i|\mathbf{w}_A, t) = \frac{1/\exp(\log(s^{(t)})/n)^{Q([w_A]_i)}}{\sum_{j=1}^{n} 1/\exp(\log(s^{(t)})/n)^{Q([w_A]_j)}}, \tag{14}$$

which favors samples with larger weights. Besides, the selection pressure $s^{(t)}$ decays exponentially to prevent excessive exploitation in later training stages (Appendix D1).

D2ACE considers two types of complementary instance weights, namely uncertainty weights $\mathbf{w}_U$ and hardness weights $\mathbf{w}_H$, and obtains two corresponding types of sampling distributions based on Eq.(14). To adaptively balance their contributions during training, we introduce an epoch-wise Bernoulli random variable $\beta^t \sim \text{Bernoulli}(p_\beta^t)$ to determine whether uncertainty-based ($\beta^t = 1$) or hardness-based ($\beta^t = 0$) sampling is used for each instance. The joint sampling probability of instance $\mathbf{x}_i$ is defined as:

$$\begin{aligned} P(\mathbf{x}_i, \beta_t) &= P(\beta_t)\, P(\mathbf{x}_i \mid \beta_t) \\ &= \begin{cases} p_\beta^t\, P(\mathbf{x}_i \mid w_U, t), & \beta_t = 1, \\ (1 - p_\beta^t)\, P(\mathbf{x}_i \mid w_H, t), & \beta_t = 0. \end{cases} \end{aligned} \tag{15}$$

Marginalizing over β^t yields the *Bernoulli mixture of two metrics-aware sampling* distributions:

$$P(\mathbf{x}_i \mid t) = p_\beta^t\, P(\mathbf{x}_i \mid \mathbf{w}_U, t) + (1 - p_\beta^t)\, P(\mathbf{x}_i \mid \mathbf{w}_H, t) \tag{16}$$

The mixing coefficient p_β^t linearly decays across epochs as:

$$p_\beta^t = p_\beta^{t_{\text{start}}} + \frac{t - t_{\text{start}}}{t_{\text{end}} - t_{\text{start}}} \left(p_\beta^{t_{\text{end}}} - p_\beta^{t_{\text{start}}} \right), \tag{17}$$

where $p_\beta^{t_{\text{start}}}$ and $p_\beta^{t_{\text{end}}}$ denote the mixing probabilities at the initial and final epochs, respectively. We set $p_\beta^{t_{\text{start}}} > p_\beta^{t_{\text{end}}}$, which progressively shifts the sampling focus from uncertain to hard instances. Specifically, early epochs emphasize uncertain instances to explore ambiguous regions and prevent premature convergence, while later epochs prioritize persistently misclassified hard instances to refine challenging decision boundaries and correct remaining systematic errors after initial learning.

Algorithm 1: Training by D2ACE Batch Selection

Input: Training set: $\mathcal{D}$, # epochs: T, # warm-up epochs T_w, batch size: b, Bernoulli mixing coefficient in each epoch $\{p_\beta^t\}_{t=T_w+1}^{T}$, initialized model parameters Θ
Output: Trained model Θ

```
for t = 1 to T do
    for k = 1 to n/b do
        if t < T_w then
            B ← Sample a batch randomly;
        else
            U ← Update uncertainty by Eq.(1)-(3) ;
            H ← Update hardness by Eq.(4)-(5) ;
            δ_H, δ_U ← Update label dynamic weights by
              Eq.(6)-(8) ;
            φ_H, φ_U ← Update label correlation enhanced
              weights by Eq.(9)-(11) ;
            w_U, w_H ← δ_U + φ_U, δ_H + φ_H ;
            B ← Sample a batch based on Eq.(15);
        Forward with B and compute loss ;
        Backward and update Θ by Adam optimizer;
```

3.6 Algorithm

Algorithm 1 presents the training procedure of a deep multi-label model with D2ACE batch selection. After a warm-up stage with randomly sampling, mini-batches are selected via a Bernoulli mixture of uncertainty- and hardness-based distributions, controlled by the epoch-wise mixing coefficient p_β^t. The **computational complexity** of our method is $O(n^2 d + T(nqn_t + \sigma(\mathbf{Y})^2 n + \sigma(\mathbf{Z})q)$, where $\sigma(\cdot)$ denotes the number of nonzero elements. It benefits from label sparsity and is more efficient than ML-Unc that scales quadratically with q (see Appendix B for details). The Algorithm 1 is **guaranteed to converge** in the same sense as generic Adam [Zou *et al.*, 2019] (see Appendix C for analysis and proof).

4 Experiments

4.1 Experiment Setup

Table 2 shows 15 multi-label datasets spanning various domains and input formats used in the experiments. The tabular multi-label datasets are from the MULAN repository [Tsoumakas *et al.*, 2011], where each instance is represented as a feature vector. VOC2007 [Everingham *et al.*, 2010] and MS-COCO [Lin *et al.*, 2014] are two multi-label image datasets, which consist of raw images.

We compare D2ACE with seven batch selection methods, including Random, Active [Chang *et al.*, 2017], Recent [Song *et al.*, 2020a], DIHCL [Zhou *et al.*, 2020], Balance [Hand *et al.*, 2018], Hard-Imb [Zhou *et al.*, 2024a], and ML-Unc [Zhou *et al.*, 2025]. The Random uses the default mini-batch sampling implemented in PyTorch. For Active, Recent, and DIHCL that are tailored for single-label data, we adapt them to the multi-label scenario by computing the instance weight as the aggregation of label-wise importance scores. For tabular datasets, we evaluate all batch selection methods on CLIF [Hang and Zhang, 2021], DELA [Hang and Zhang, 2024], and PACA [Hang *et al.*, 2022] base models.

Dataset		n	d	q	*Card*	*Dens*	Domain
tabular	cal500	502	68	174	26.04	0.15	music
	birds	645	260	19	1.01	0.05	audio
	enron	1702	1001	53	3.38	0.06	text
	scene	2407	294	6	1.07	0.18	vision
	yeast	2417	103	14	4.24	0.30	biology
	Corel5k	5000	499	374	3.52	0.01	vision
	rcv1subset1	6000	944	101	2.88	0.03	text
	rcv1subset2	6000	944	101	2.63	0.03	text
	rcv1subset3	6000	944	101	2.61	0.03	text
	bibtex	7395	1836	159	2.40	0.02	text
	yahoo-Arts	7484	2314	25	1.67	0.07	text
	yahoo-Business	11214	2192	28	1.47	0.06	text
	mediamill	43907	120	101	4.38	0.04	vision
image	VOC2007	9963	-	20	1.45	0.07	vision
	MS-COCO	123287	-	80	2.91	0.03	vision

Table 2: Multi-label datasets.

For image datasets, we use HST [Chen *et al.*, 2024] and ML-GCN [Chen *et al.*, 2019] with a ResNet-101 backbone pre-trained on ImageNet [Deng *et al.*, 2009] and standard preprocessing following [Chen *et al.*, 2024]. The hyperparameters of batch selection baselines and base deep multi-label models follow their respective original implementations. The specific configurations for all methods are listed in Appendix D.

For tabular datasets, we evaluate performance using Macro-F1, Macro-AUC, and Ranking-Loss [Zhang and Zhou, 2013] that cover both classification and label ranking aspects. Results are reported as the average over stratified five-fold cross-validation [Sechidis *et al.*, 2011]. For image datasets, we use the Mean Average Precision (mAP) as the evaluation metric and report average results on the pre-defined split test set over five runs with different random seeds. In each fold (run), we record the test result at the epoch that achieves the best validation performance. The source code is available at https://github.com/iunanyou/D2ACE.

4.2 Results

Tabular Data Results. As shown in Table 3, D2ACE is the best method across three base models and three evaluation metrics, and significantly outperforms the baselines in 61 out of 63 cases according to the Wilcoxon signed-rank test. The success of our method stems from jointly modeling stage-wise metric utility and dynamic label importance, as well as enhancing instance selection by exploiting label correlations under local label context. Based on detailed results presented in Table 4 and Appendix E, our method demonstrates more prominent advantages on large-scale datasets. Multi-label batch selection methods that explicitly account for label characteristics generally outperform single-label approaches, yet remain inferior to our method. ML-Unc propagates label correlations dominated by irrelevant labels and relies solely on uncertainty, which may neglect extremely hard instances in later training. Hard-Imb emphasizes hardness, making it prone to overfitting noisy samples in early training, while both Hard-Imb and Balance treat label imbalance as a static priority throughout training. In Macro-F1, Balance, DIHCL, and Recent achieve the second-best performance with different base models. This is because Balance explicitly enforces label-balanced mini-batches, and DIHCL and Recent assess instance importance based on binary label predictions, which are more closely aligned with the F1 metric than prediction probabilities used in ML-Unc and Active or loss-based criteria adopted by Hard-Imb.

Metric	Base	Random	Active	Recent	DIHCL	Balance	Hard-Imb	ML-Unc	**D2ACE**
Macro -AUC	CLIF	7.08*	5.54*	5.77*	3.08*	4.15*	5.23*	3.54†	**1.31**
	DELA	6.46*	4.85*	5.38−	5.31*	4.92*	3.92*	2.92*	**2.00**
	PACA	5.31†	4.69*	4.69*	5.31*	4.38†	5.23*	4.00	**2.15**
Macro -F1	CLIF	6.62*	5.08*	4.69*	3.00*	5.31*	5.31*	4.00†	**1.69**
	DELA	6.46*	4.77*	3.23	5.92*	5.38*	4.00*	4.23*	**1.92**
	PACA	5.54*	5.54*	5.69*	5.38*	3.77*	4.00*	4.31†	**1.69**
Ranking -Loss	CLIF	6.85*	5.77*	4.85*	3.54*	4.31*	4.92*	3.00*	**1.69**
	DELA	5.77*	4.85*	5.46*	4.92*	4.77*	3.62*	4.46*	**1.46**
	PACA	5.69†	5.31*	5.54*	4.92*	4.77†	3.38†	3.85†	**2.08**

Table 3: Average ranks and Wilcoxon signed-rank test results on multi-label tabular datasets. The **best**, second, and third ranks are highlighted. † (∗) denotes that our method significantly outperforms baseline at $p < 0.05$ ($p < 0.01$) levels.

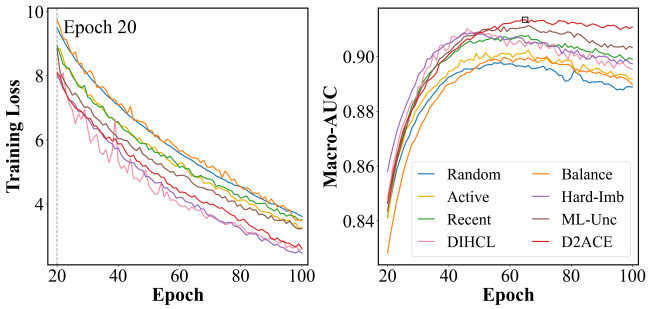


Figure 2: The convergence curves and Macro-AUC on validation set of CLIF using various batch selection methods on the bibtex dataset.

Image Data Results. Table 5 reports the mAP results of various batch selection methods under HST and ML-GCN base models on two multi-label image datasets. D2ACE consistently achieves the best performance across both base models, with more pronounced improvements on the larger MS-COCO dataset. DIHCL ranks second overall, primarily due to its temporally smoothed hardness estimation, which is more robust than relying solely on the loss at the current epoch (e.g., Hard-Imb) for complex image classification tasks. ML-Unc follows next, as it incorporates recent uncertainty variations together with label correlations.

4.3 Analysis

Convergency. As shown in Figure 2, hardness-based methods, including DIHCL and Hard-Imb, converge fastest, as they prioritize high-loss instances during training. However, their validation performance declines sharply after 50 epochs, indicating a tendency to overfit outliers. In contrast, D2ACE converges faster than uncertainty-based approaches (ML-Unc, Recent, and Active), especially in later training stages, due to its gradual shift in emphasis toward difficult instances. Moreover, it achieves the best validation performance near epoch 60 (marked by the black square), surpassing all baselines, and remains stable thereafter. Uncertainty-based methods exhibit mild performance degradation after reaching their peak validation performance. Balance converges at a speed similar to Random but with larger fluctuations, mainly due to random filtering of negative samples.

Dataset	Macro-AUC ↑								Ranking-Loss ↓							
	Random	Active	Recent	DIHCL	Balance	Hard-Imb	ML-Unc	**D2ACE**	Random	Active	Recent	DIHCL	Balance	Hard-Imb	ML-Unc	**D2ACE**
cal500	0.5849(4)	0.5826(6)	0.5818(7)	0.5869(3)	0.5827(5)	0.5807(8)	**0.5884(1)**	0.5879(2)	0.2318(6)	0.2313(4)	0.2317(5)	**0.2300(1)**	0.2310(2)	0.2318(6)	0.2311(3)	0.2318(6)
birds	0.8026(6)	0.8037(5)	0.8006(8)	0.8025(7)	0.8046(3)	0.8045(4)	0.8050(2)	**0.8090(1)**	0.1777(7)	0.1710(2)	0.1822(8)	0.1773(6)	0.1728(4)	0.1739(5)	**0.1705(1)**	0.1720(3)
enron	0.7704(7)	0.7751(4)	0.7748(5)	0.7729(6)	0.7757(2)	0.7695(8)	0.7753(3)	**0.7766(1)**	0.1205(5)	0.1213(7)	0.1209(6)	**0.1146(1)**	0.1242(8)	0.1166(2)	0.1186(4)	0.1170(3)
scene	0.9481(8)	0.9496(6)	0.9489(7)	**0.9519(1)**	0.9504(3)	0.9503(5)	0.9504(3)	0.9510(2)	0.0671(7)	0.0669(6)	0.0665(5)	0.0650(2)	0.0661(3)	0.0671(7)	0.0664(4)	**0.0629(1)**
yeast	0.7226(7)	0.7236(3)	0.7230(5)	0.7217(8)	0.7234(4)	0.7229(6)	0.7242(2)	**0.7305(1)**	0.1654(5)	0.1648(4)	0.1670(7)	0.1682(8)	0.1646(3)	0.1658(6)	0.1645(2)	**0.1616(1)**
Corel5k	0.7582(8)	0.7603(7)	0.7614(5)	0.7629(2)	0.7619(4)	0.7606(6)	0.7621(3)	**0.7647(1)**	0.1590(8)	0.1584(5)	0.1564(2)	0.1589(6)	0.1576(4)	0.1589(6)	0.1568(3)	**0.1544(1)**
rcvsubset1	0.9243(8)	0.9257(4)	0.9256(5)	0.9269(2)	0.9250(7)	0.9259(3)	0.9254(6)	**0.9271(1)**	0.0668(8)	0.0664(6)	0.0655(3)	0.0641(2)	0.0667(7)	0.0662(5)	0.0655(3)	**0.0633(1)**
rcvsubset2	0.9211(8)	0.9221(6)	0.9222(5)	0.9232(2)	0.9223(3)	0.9217(7)	0.9223(3)	**0.9238(1)**	0.0719(8)	0.0709(7)	0.0694(2)	0.0696(4)	0.0696(4)	0.0698(6)	0.0694(2)	**0.0675(1)**
rcvsubset3	0.9168(8)	0.9176(6)	0.9188(4)	**0.9220(1)**	0.9185(5)	0.9172(7)	0.9205(3)	0.9211(2)	0.0738(6)	0.0743(8)	0.0739(7)	0.0705(2)	0.0728(3)	0.0736(5)	0.0733(4)	**0.0697(1)**
bibtex	0.9031(8)	0.9036(4)	0.9034(6)	0.9045(2)	0.9035(5)	0.9037(3)	0.9034(6)	**0.9061(1)**	0.0879(8)	0.0862(7)	0.0861(6)	0.0857(3)	0.0858(5)	0.0857(3)	0.0855(2)	**0.0839(1)**
yahoo-Arts1	0.7441(7)	0.7438(8)	0.7479(5)	0.7495(2)	0.7465(6)	0.7482(4)	0.7486(3)	**0.7512(1)**	0.1578(8)	0.1572(7)	0.1569(5)	0.1565(4)	0.1564(3)	0.1569(5)	0.1563(2)	**0.1528(1)**
yahoo-Business1	0.8063(7)	0.8070(5)	0.8047(8)	**0.8115(1)**	0.8076(3)	0.8064(6)	0.8072(4)	**0.8115(1)**	0.1339(6)	0.1334(4)	0.1336(5)	0.1332(2)	0.1346(8)	0.1332(2)	0.1343(7)	**0.1267(1)**
mediamill	0.8025(6)	0.8012(8)	0.8027(5)	0.8043(3)	0.8028(4)	**0.8129(1)**	0.8018(7)	0.8108(2)	0.0441(7)	0.0442(8)	0.0438(2)	0.0439(5)	0.0438(2)	0.0440(6)	0.0438(2)	**0.0429(1)**
Avg(Rank)	7.08	5.54	5.77	3.08	4.15	5.23	3.54	**1.31**	6.85	5.77	4.85	3.54	4.31	4.92	3.00	**1.69**

Table 4: The Macro-AUC and Ranking-Loss results of CLIF using various batch selection methods on multi-label tabular datasets.

Dataset	HST								ML-GCN							
	Random	Active	Recent	DIHCL	Balance	Hard-Imb	ML-Unc	**D2ACE**	Random	Active	Recent	DIHCL	Balance	Hard-Imb	ML-Unc	**D2ACE**
VOC2007	0.9300(5)	0.9294(8)	0.9308(4)	0.9310(3)	0.9298(7)	0.9300(5)	0.9314(2)	**0.9320(1)**	0.9293(6)	0.9285(8)	0.9295(5)	0.9308(2)	0.9289(7)	0.9298(4)	0.9305(3)	**0.9314(1)**
MS-COCO	0.8012(8)	0.8054(7)	0.8127(5)	0.8186(2)	0.8065(6)	0.8182(3)	0.8179(4)	**0.8252(1)**	0.7962(8)	0.8025(6)	0.8074(4)	0.8103(2)	0.8021(7)	0.8055(5)	0.8085(3)	**0.8142(1)**
Avg(Rank)	6.50	7.50	4.50	2.50	6.50	4.00	3.00	**1.00**	7.00	7.00	4.50	2.00	7.00	4.50	3.00	**1.00**

Table 5: The mAP results of HST and ML-GCN using various batch selection methods on multi-label image datasets.

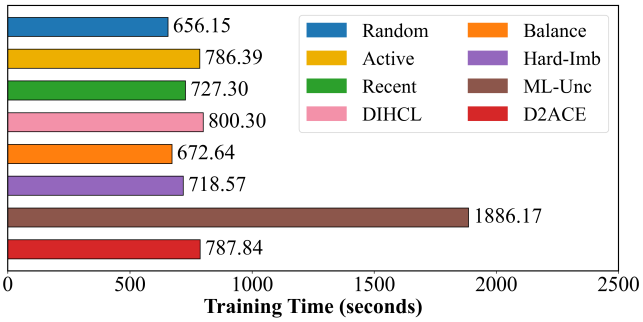


Figure 3: Training time of CLIF using various batch selection methods on the bibtex dataset (100 epochs).

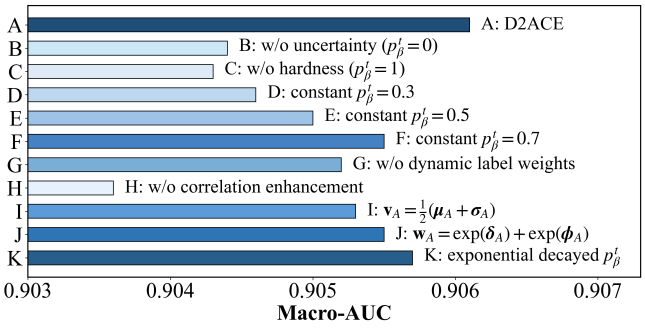


Figure 4: Ablation study of D2ACE with CLIF on the bibtex dataset.

Efficiency. As shown in Figure 3, Random is the most time-efficient method, followed by Balance which only requires the computation of label statistics. D2ACE leverages sparse matrix operations to model label correlations, attaining comparable training time to label dependency-agnostic methods (Active, DIHCL) while exhibiting higher efficiency than the correlation-based ML-Unc.

Ablation Study. Figure 4 presents the ablation results of D2ACE under different variants, each removing or modifying a key component. Relying solely on either hardness (B) or uncertainty (C), as well as fixing the metric priority (D–F), leads to consistent performance degradation, demonstrating both the complementarity of the two metrics and the importance of stage-wise adjustment of metric emphasis. Removing dynamic label weighting (G) also reduces performance, highlighting the necessity of capturing evolving label significance during training. Most notably, excluding label correlation enhancement (H) causes the largest performance decline, which indicates the critical role of leveraging label correlations under local context. A linear form for positive label weights $\mathbf{v}_A$ (I) is less effective than the exponential form in amplifying informative signals. Conversely, exponential transformation of instance weights $\mathbf{w}_A$ (J) yields overly uniform sampling probabilities. Exponential decay for p_β^t (K) is less suitable, as it slows the shift from exploration to exploitation.

Further empirical analyses concerning parameter sensitivity, along with case studies on stage-wise sampling and label dynamic weighting, are detailed in Appendix F.

5 Conclusion

Batch selection in multi-label classification is challenging due to evolving metric utility, dynamic label importance, as well as sparse and instance-specific label correlations. To address these, we propose D2ACE, a batch selection framework that does not affect the convergence of Adam optimizers. It handles metric dynamics by combining uncertainty and noise-resistant hardness through stage-wise Bernoulli mixture sampling, and captures label dynamics with epoch-wise weighting that adjusts label priorities based on current metrics. It also employs local context-aware correlation enhancement to focus on relevant labels and exploit instance-specific dependencies. Extensive experiments on tabular and image benchmarks demonstrate that D2ACE consistently outperforms state-of-the-art batch selection methods across deep MLC models, while being more computationally efficient than the label correlation-based baseline.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62302074, 62402077), Natural Science Foundation of Chongqing (CSTB2025NSCQ-GPX1269, CSTB2023NSCQ-MSX0613), the Science and Technology Research Program of Chongqing Municipal Education Commission (KJQN202300631).

References

[Bai *et al.*, 2024] Peihao Bai, Guanghui Li, Jiawei Luo, and Cheng Liang. Deep learning model for protein multi-label subcellular localization and function prediction based on multi-task collaborative training. *Briefings in Bioinformatics*, 25(6):bbae568, 2024.

[Chai *et al.*, 2024] Yuyang Chai, Zhuang Li, Jiahui Liu, Lei Chen, Fei Li, Donghong Ji, and Chong Teng. Compositional generalization for multi-label text classification: A data-augmentation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17727–17735, 2024.

[Chang *et al.*, 2017] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*, 2017.

[Chen *et al.*, 2019] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019.

[Chen *et al.*, 2024] Tianshui Chen, Tao Pu, Lingbo Liu, Yukai Shi, Zhijing Yang, and Liang Lin. Heterogeneous semantic transfer for multi-label recognition with partial labels. *International Journal of Computer Vision*, 132(12):6091–6106, 2024.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[Everingham *et al.*, 2010] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[Guo *et al.*, 2023] Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. Texts as images in prompt tuning for multi-label image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2808–2817, 2023.

[Hand *et al.*, 2018] Emily Hand, Carlos Castillo, and Rama Chellappa. Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6878–6885, 2018.

[Hang and Zhang, 2021] Jun-Yi Hang and Min-Ling Zhang. Collaborative learning of label semantics and deep label-specific features for multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9860–9871, 2021.

[Hang and Zhang, 2024] Jun-Yi Hang and Min-Ling Zhang. Dual perspective of label-specific feature learning for multi-label classification. *ACM Transactions on Knowledge Discovery from Data*, 19(1):1–30, 2024.

[Hang *et al.*, 2022] Jun-Yi Hang, Min-Ling Zhang, Yanghe Feng, and Xiaocheng Song. End-to-end probabilistic label-specific feature learning for multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6847–6855, 2022.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Lanchantin *et al.*, 2021] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16478–16488, 2021.

[Li *et al.*, 2023] Jiaxuan Li, Xiaoyan Zhu, and Jiayin Wang. Adaboost.c2: Boosting classifiers chains for multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8580–8587, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755, 2014.

[Lin *et al.*, 2023] Nankai Lin, Guanqiu Qin, Gang Wang, Dong Zhou, and Aimin Yang. An effective deployment of contrastive learning in multi-label text classification. In *Findings of the Association for Computational Linguistics*, pages 8730–8744, 2023.

[Liu *et al.*, 2021] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W Tsang. The emerging trends of multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 7955–7974, 2021.

[Liu *et al.*, 2025] Bin Liu, Ao Zhou, Bingkun Wei, Jin Wang, and Grigorios Tsoumakas. Oversampling multi-label data based on natural neighbor and label correlation. *Expert Systems with Applications*, 259:125257, 2025.

[Loshchilov and Hutter, 2015] Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*, 2015.

[Maini *et al.*, 2022] Pratyush Maini, Saurabh Garg, Zachary Lipton, and J Zico Kolter. Characterizing datapoints via second-split forgetting. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*, pages 30044–30057, 2022.

[Ning *et al.*, 2025] Zhihan Ning, Zhixing Jiang, and David Zhang. Exploiting meta-learned confidences for imbalanced multilabel learning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 10242–10256, 2025.

[Sechidis *et al.*, 2011] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 145–158, 2011.

[Shou *et al.*, 2023] Xiao Shou, Tian Gao, Dharmashankar Subramanian, Debarun Bhattacharjya, and Kristin P Bennett. Concurrent multi-label prediction in event streams. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9820–9828, 2023.

[Song *et al.*, 2020a] Hwanjun Song, Minseok Kim, Sundong Kim, and Jae-Gil Lee. Carpe diem, seize the samples uncertain "at the moment" for adaptive batch selection. In *Proceedings of the ACM International Conference on Information & Knowledge Management*, pages 1385–1394, 2020.

[Song *et al.*, 2020b] Hwanjun Song, Sundong Kim, Minseok Kim, and Jae-Gil Lee. Ada-boundary: Accelerating dnn training via adaptive boundary batch selection. *Machine Learning*, 109(9):1837–1853, 2020.

[Toneva *et al.*, 2018] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

[Tsoumakas *et al.*, 2010] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1079–1089, 2010.

[Tsoumakas *et al.*, 2011] Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. Mulan: A java library for multi-label learning. *The Journal of Machine Learning Research*, 12:2411–2414, 2011.

[Wu *et al.*, 2024] Xiangyu Wu, Qing-Yuan Jiang, Yang Yang, Yi-Feng Wu, Qing-Guo Chen, and Jianfeng Lu. Tai++: Text as image for multi-label image classification by co-learning transferable prompt. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 5226–5234, 2024.

[Yang *et al.*, 2020] Liang Yang, Xi-Zhu Wu, Yuan Jiang, and Zhi-Hua Zhou. Multi-label learning with deep forest. In *Proceedings of the European Conference on Artificial Intelligence*, pages 1634–1641, 2020.

[Yuan *et al.*, 2024] Zhixiang Yuan, Kaixin Zhang, and Tao Huang. Positive label is all you need for multi-label classification. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2024.

[Zhang and Zhou, 2007] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

[Zhang and Zhou, 2013] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2013.

[Zhang *et al.*, 2018] Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. Binary relevance for multi-label learning: An overview. *Frontiers of Computer Science*, 12(2):191–202, 2018.

[Zhao *et al.*, 2021] Wenting Zhao, Shufeng Kong, Junwen Bai, Daniel Fink, and Carla Gomes. Hot-vae: Learning high-order label correlation for multi-label classification via attention-based variational autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15016–15024, 2021.

[Zhao *et al.*, 2025] Xingyu Zhao, Yuexuan An, Ning Xu, Lei Qi, and Xin Geng. Interactive fusion label enhancement for multi-label learning. *ACM Transactions on Knowledge Discovery from Data*, 19(7):1–23, 2025.

[Zhou *et al.*, 2020] Tianyi Zhou, Shengjie Wang, and Jeffrey Bilmes. Curriculum learning by dynamic instance hardness. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*, pages 8602–8613, 2020.

[Zhou *et al.*, 2024a] Ao Zhou, Bin Liu, Zhaoyang Peng, Jin Wang, and Grigorios Tsoumakas. Multi-label adaptive batch selection by highlighting hard and imbalanced samples. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 265–281, 2024.

[Zhou *et al.*, 2024b] Ao Zhou, Bin Liu, Jin Wang, Kaiwei Sun, and Kelin Liu. Aemlo: Autoencoder-guided multi-label oversampling. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 107–124. Springer, 2024.

[Zhou *et al.*, 2025] Ao Zhou, Bin Liu, Jin Wang, and Grigorios Tsoumakas. Batch selection for multi-label classification guided by uncertainty and dynamic label correlations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 22902–22909, 2025.

[Zou *et al.*, 2019] Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11127–11135, 2019.