DYNAMIC RELIABILITY CURRICULUM GUIDED REINFORCEMENT LEARNING FOR PARTIAL LABEL DISAMBIGUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Partial-label learning (PLL) addresses the problem where each training inatsnce is associated with a set of candidate labels, among which only one is the ground-truth label. The ambiguity of candidate labels brings great uncertainty to model optimization. Traditional disambiguation methods usually rely on classifier confidence or predefined label refinement rules. Such strategies may be affected by confirmation bias, leading to error accumulation and label noise propagation. To address these issues, this paper proposes a *dynamic reliability curriculum guided reinforcement learning framework for PLL*, called DC-RPLL. This method models label disambiguation as a Markov decision process, independently deciding for each inatsnce whether to *update*, *retain*, or *abstain*. Specifically, DC-RPLL first represents each candidate label set as a soft label distribution, jointly uses classifier predictions and local feature space structure to construct disambiguation labels. Meanwhile, the reliability of each inatsnce is estimated from neighborhood consistency, feature affinity, and soft label consistency to quantify the credibility of local structures. Based on this reliability, this paper designs dynamic curriculum thresholds so that highly reliable inatsnces enter policy updates early, while low-reliability inatsnces are gradually added later, thereby alleviating early error propagation. The Actor-Critic policy then guides policy learning according to the current confidence gain, neighborhood reliability, and label drift penalty, i.e., *update*, *retain*, and *abstain*. In addition, we develop a reliability-aware curriculum learning mechanism that gradually accepts inatsnces from high to low reliability during training, effectively suppressing early error propagation. Experiments on benchmark and real datasets verify the effectiveness of the proposed algorithm.

1 INTRODUCTION

Supervised deep learning has yielded unprecedented successes across various domains but remains heavily reliant on massive, precisely annotated datasets. In many real-world applications (Cour et al., 2011; Fan et al., 2026a), such as web image annotation, medical diagnosis, and crowdsourcing, acquiring precise ground-truth annotations is prohibitively expensive or structurally difficult due to subjective ambiguity. To alleviate this bottleneck, *partial-label learning* (PLL) has emerged as a compelling weakly supervised paradigm. In PLL, each training instance is associated with a candidate label set containing the unknown ground-truth label together with several false-positive labels. The fundamental challenge of PLL lies in resolving label ambiguity and distinguishing the true label from noisy candidate labels.

Intuitively, addressing the PLL problem hinges on effective disambiguation. To this end, label confidence estimation in PLL has been pursued along two principal directions, i.e., *non-deep disambiguation strategy* (NDS) (Berg et al., 2004; Luo & Orabona, 2010; Chai et al., 2019; Fuchs et al., 2024) and *deep disambiguation strategy* (DDS) (Yao et al., 2020; Fan et al., 2026c; 2021; Li et al., 2025). NDS approaches aim to gradually uncover the ground-truth label via iterative refinement of the disambiguated labels. DDS leverages regularization modules, data augmentation techniques and adaptive loss weights as stochastic optimization objectives for differentiating true labels from noisy candidates. Despite the progress made by existing PLL techniques, their label updating mechanisms are built on static rules and empirical heuristics. Directly trusting classifier predictions can result in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

a self-reinforcing error propagation process. When the classifier produces an incorrect prediction at an early training stage, treating this prediction as a corrected label may strengthen the corresponding error in subsequent iterations. This phenomenon is particularly problematic for ambiguous instances whose candidate sets contain multiple semantically similar classes.

To address these limitations, we propose a novel dynamic reliability curriculum guided reinforcement learning framework for PLL, termed DC-RPLL. The key idea is to progressively refine ambiguous candidate label distributions by integrating reliability-aware neighborhood propagation, adaptive curriculum learning, and reward-guided selective decision making, while explicitly controlling the risk of erroneous label correction. Specifically, we first construct a neighborhood propagation module based on K-nearest neighbors in the learned feature space, where semantically similar instances provide complementary soft-label information for disambiguating ambiguous candidate labels. To alleviate the adverse effect of unreliable neighbors, we further introduce a reliability estimation mechanism that jointly considers feature affinity and soft-label consistency, and adaptively regulates the contribution of neighborhood information to each instance. Subsequently, inspired by curriculum learning (Bengio et al., 2009), we develop an adaptive curriculum strategy that gradually decreases the reliability threshold, allowing the disambiguation process to evolve from highly reliable instances to increasingly ambiguous ones as the representation and classifier become more stable. To further determine whether a candidate-label distribution should be corrected, retained, or left unchanged, we formulate label disambiguation as a selective decision-making problem and employ an Actor-Critic reinforcement learning framework (Kaelbling et al., 1996) with three actions, namely *update*, *retain*, and *abstain*. The policy is guided by a self-supervised reward that jointly evaluates confidence improvement, entropy reduction, instance reliability, and label-distribution drift, thereby encouraging informative corrections while suppressing aggressive or unreliable label changes. Moreover, the correction strength is adaptively controlled according to instance reliability, such that highly reliable instances receive stronger refinement whereas uncertain instances are assigned conservative updates or abstention. Overall, DC-RPLL establishes a closed-loop disambiguation process that couples neighborhood consensus, reliability assessment, curriculum progression, and reinforced selective correction. Under local neighborhood consistency and bounded estimation-error assumptions, the proposed reliability-aware correction admits an error-contraction property, while the abstention mechanism prevents low-reliability instances from being forcibly corrected. The entire framework therefore provides a risk-aware and progressively refined solution to partial-label disambiguation, jointly improving label quality and classification reliability. The major contributions of this work are summarized as follows:

- We propose a novel dynamic reliability curriculum guided reinforcement learning framework for label disambiguation, effectively eliminating invalid label assignments and reducing erroneous disambiguation decisions.
- We design a self-supervised reward function that combines confidence improvement, entropy reduction, neighborhood reliability, and drift penalty, providing effective optimization guidance without ground-truth labels.
- We propose a curriculum learning mechanism that dynamically adjusts the admission threshold based on neighborhood propagation reliability, introducing instances from easy-to-hard to stabilize early training.
- We formulate label disambiguation as a selective decision process involving *update*, *retain*, and *abstain* actions and optimize the policy using a self-supervised confidence entropy reward.

2 Related Work

2.1 Partial Label Learning

Existing PLL methods can generally be categorized into *non-deep disambiguation strategy* (NDS) and *deep disambiguation strategy* (DDS). NDS methods resolve label ambiguity by exploiting structural relationships among instances and candidate labels (Chai et al., 2019). Based on the smoothness assumption that neighboring instances are likely to share the same class label, a variety of machine learning techniques have been developed for PLL, including KNN-based propagation (Luo & Orabona, 2010; Zhang et al., 2016), expectation maximization (Liu & Dietterich, 2012), low-rank

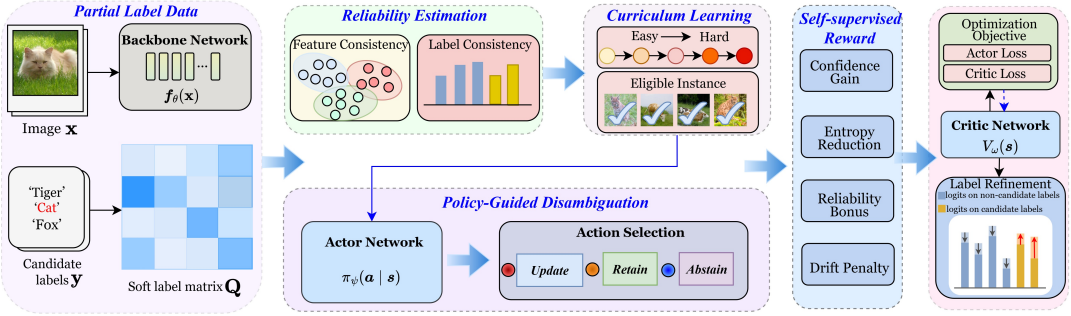


Figure 1: The framework of DC-RPLL

representation (Fan et al., 2022; Chen et al., 2013), graph-based optimization (Fan et al., 2021), and margin-based learning approaches (Chai et al., 2019; Fan et al., 2023). These methods iteratively refine latent labels by leveraging local consistency and global structural information. With the success of deep neural networks, DDS methods have become the dominant paradigm for PLL. Instead of explicitly performing iterative label disambiguation, DDS methods incorporate ambiguity-aware objectives into end-to-end optimization. Representative techniques include confidence regularization (Feng et al., 2025), label enhancement (Chai et al., 2019), prototype learning, consistency regularization, contrastive learning, and adaptive loss reweighting (Fan et al., 2026a; Wen et al., 2021). Recent studies further improve robustness by integrating evidential learning, uncertainty estimation, and semantic representation learning, enabling more effective exploitation of discriminative features under ambiguous supervision (Fan et al., 2026c). Most conventional PLL approaches assume that candidate labels are generated independently of instance characteristics. However, in many real-world applications, label ambiguity is inherently instance-dependent. To address this limitation, instance-dependent PLL offers a more realistic framework for ambiguous supervision. To tackle this challenge, methods such as contrastive learning, Bayesian models, and variational inference have been developed (Xu et al., 2021; 2024; Wu et al., 2024).

One prominent limitation in existing PLL techniques is their utilization of stationary heuristic schemes for label disambiguation, which makes them vulnerable to error propagation in scenarios with intense label uncertainty. Any initial label mistakes will be reinforced step by step, leading to persistent and significant performance loss.

2.1.1 Reinforcement Learning

RL is a sequential decision-making framework in which an agent learns an optimal policy through interactions with an environment and feedback in the form of rewards (Sutton et al., 1998). Unlike supervised learning, RL does not rely on explicit ground-truth annotations but instead optimizes long-term cumulative returns through trial-and-error exploration. Owing to its capability for adaptive exploration and delayed reward optimization, RL has achieved remarkable success in robotics, autonomous control, recommendation systems, and large-scale decision-making problems (Kober et al., 2013; Amin et al., 2021). Recent advances have significantly improved the efficiency and applicability of RL. *Preference-based reinforcement learning* (PbRL) learns reward functions directly from preference feedback rather than manually designed rewards, providing a more flexible paradigm for complex environments (Christiano et al., 2017). Building upon this idea, *reinforcement learning from human feedback* (RLHF) has become a key component in aligning *large language models* (LLMs) with human intentions, demonstrating the effectiveness of reward-guided policy optimization under weak supervision (Ouyang et al., 2022).

Despite recent advances, RL remains largely unexplored in PLL. Existing methods mainly rely on deterministic label updates based on prediction confidence or neighborhood consistency, which limit exploration of alternative label assignments and may suffer from confirmation bias. In contrast, RL naturally formulates label disambiguation as a sequential decision process, enabling adaptive candidate exploration and reward-guided refinement. Motivated by this, DC-RCPLL integrates reinforcement learning with candidate-aware policy exploration and reward-guided label refinement to reduce error propagation and improve disambiguation reliability.

3 DC-RPLL METHOD

3.1 PROBLEM FORMULATION

In PLL, we are given a training set $\mathcal{D} = \{(\mathbf{x}_i, S_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ is input instance and $S_i \subseteq \{1, 2, \ldots, C\}$ is the candidate label set. The true label $y_i \in S_i$ is unknown during training. The goal is to learn a classifier $\boldsymbol{f}_\theta : \mathcal{X} \to \Delta^C$ that maps inputs to class probability distributions with parameters θ, and simultaneously estimate a soft label matrix $\boldsymbol{Q} \in [0, 1]^{N \times C}$, where $\boldsymbol{q}_i \in \Delta^C$ is the disambiguated soft label for instance $\mathbf{x}_i$.

We propose a novel dynamic reliability curriculum guided RL for ambiguous label disambiguation. PLL is formulated as a sequential decision-making problem, where an agent dynamically determines whether to *update*, *retain*, or *abstain* from correcting ambiguous candidate labels. Designing a reliable disambiguation policy is crucial for effectively resolving label ambiguity in PLL. *It is difficult to determine whether a candidate label should be corrected reliably in the presence of ambiguous supervision. This is because erroneous classifier predictions or unreliable neighborhood information may lead to incorrect label updates, which can be further propagated during subsequent training and result in confirmation bias.* To address this issue, we design a reliability-aware neighborhood propagation mechanism to generate soft label correction proposals and estimate their reliability from local feature and label consistency. Meanwhile, an adaptive curriculum mechanism is introduced to progressively regulate the correction of inatsnces from reliable to ambiguous ones. Based on these representations, an Actor-Critic policy is employed to dynamically select the *update*, *retain*, and *abstain* actions, while a self-supervised reward guides the policy toward reliable label refinement and suppresses excessive label drift. The overall framework of DC-RPLL is illustrated in Figure 1, and the detailed components of the proposed method are introduced in the following subsections.

3.2 RELIABILITY-AWARE NEIGHBORHOOD PROPAGATION MECHANISM

For each instance $\mathbf{x}_i$, the initial candidate distribution is $\boldsymbol{q}_{ic}^{(0)} = \frac{\mathbf{1}[c \in S_i]}{|S_i|}$. Therefore, $\sum_{c=1}^{C} \boldsymbol{q}_{ic}^{(0)} = 1$ and $\boldsymbol{q}_{ic}^{(0)} = 0$, $c \notin S_i$. The initial soft labels are evenly distributed on the candidate labels. The principle of doing so is that, in the absence of any prior information, all labels in the candidate set should be considered equally likely. Uniform initialization avoids artificially introducing preferences while ensuring that soft labels satisfy probability normalization. The deep classifier produces a predictive distribution $\boldsymbol{p}_i = \mathrm{softmax}(\boldsymbol{f}_\theta(\mathbf{x}_i))$ and extracts a feature representation $\mathbf{z}_i$. The classifier is trained using the current soft labels,

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \boldsymbol{q}_{ic} \log \boldsymbol{p}_{ic} \tag{1}$$

Thus, the current label distribution provides soft supervision while avoiding the use of unavailable ground-truth labels. As the soft labels gradually disambiguate, the supervised signals obtained by the classifier become increasingly accurate. To reduce instability, DC-RPLL does not immediately perform label refinement. Instead, the classifier first undergoes a warm-up stage, after which label refinement is periodically executed.

1) KNN Neighborhood Propagation: To mitigate the noisy labels influence, we through KNN attention to identify potential ground truth labels by utilizing complementary information between instances and candidate labels, making the model more resistant to ambiguous labels. The underlying assumption is that a label consistent with those of its k-nearest neighbors is more likely to have the same label. The neighborhood label proposal is calculated by $\tilde{\boldsymbol{q}}_i = \sum_{j \in \mathcal{N}_k(\mathbf{x}_i)} w_{ij} \boldsymbol{q}_j$, where $\mathcal{N}_k(\mathbf{x}_i)$ denotes the KNN set of instance $\mathbf{x}_i$ based on the Euclidean distances calculated in the embedding space $\mathbf{z}_i$, and the attention weights w_{ij} is computed using similarities of instance pairs in the embedding space,

$$w_{ij} = \frac{\exp(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{l \in \mathcal{N}_k(\mathbf{x}_i)} \exp(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_l)/\tau)} \tag{2}$$

where τ is a temperature hyperparameter, and $\mathrm{sim}(\cdot, \cdot)$ denotes the cosine similarity. This mechanism transfers label information from structurally similar instances while retaining the probabilistic nature of PLL labels.

2) Reliability Estimation: The reliability mechanism is designed to evaluate whether local neighborhood information provides consistent evidence. First, soft-label agreement between instance $\mathbf{x}_i$ and neighbor $\mathbf{x}_j$ is defined as $a_{ij} = \boldsymbol{q}_i^\top \boldsymbol{q}_j$. The weighted label agreement becomes $A_i = \sum_{j\in\mathcal{N}_k(\mathbf{x}_i)} w_{ij} a_{ij}$. Meanwhile, the weighted feature affinity is $F_i = \sum_{j\in\mathcal{N}_k(\mathbf{x}_i)} w_{ij} \mathbf{z}_i^\top \mathbf{z}_j$. The final reliability is calculated by,

$$r_i = \sigma\left(\frac{A_i + F_i}{\tau}\right) \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function. The resulting $r_i \in (0, 1)$ measures the consistency of the local neighborhood. A high value indicates that the instance is surrounded by feature-similar instances with compatible label distributions, while a low value indicates that neighborhood evidence is potentially unreliable.

3.3 Adaptive Curriculum Learning

We design a reliability-aware curriculum learning mechanism to implement easy-to-hard disambiguation, which is critical for stabilizing early training. Based on reliability r_i, this article designs a time-dependent reliability threshold $\eta_e = \eta_{\text{start}} + \frac{e}{E-1}(\eta_{\text{end}} - \eta_{\text{start}})$, where the current epoch is e, the total epoch is E, and the reliability threshold linearly decreases from $\eta_{\text{start}} = 0.80$ to $\eta_{\text{end}} = 0.35$. Hence the framework gradually changes from a conservative high-reliability regime to a more inclusive regime. A instance is admitted when $r_i \geq \eta_e$. Its curriculum level is

$$cur_i = \text{clip}\left(\frac{r_i - \eta_e}{1 - \eta_e}, 0, 1\right) \tag{4}$$

The key distinction from conventional curriculum learning is that the curriculum directly determines the legal action space,

$$\mathcal{A}_i^e = \begin{cases} abstain, & r_i < \eta_e, \\ update, retain, abstain, & r_i \geq \eta_e. \end{cases} \tag{5}$$

where $\mathcal{A}$ represents the action space. Therefore, low-reliability examples are explicitly prevented from undergoing premature label correction.

3.4 Deep Reinforcement Learning Disambiguation Strategy

The core innovation of DC-RPLL is modeling per-inatsnce label update as a *Markov decision process* (MDP) and solving it with a deep Actor-Critic network.

State Representation. For each instance $\mathbf{x}_i$, the RL state consists of $\boldsymbol{s}_i = [\boldsymbol{q}_i, \boldsymbol{p}_i, \tilde{\mathbf{q}}_i, r_i, conf(\mathbf{q}_i), \bar{H}_i, d_i, cur_i]$, where $conf(\boldsymbol{q}_i) = \max_c \boldsymbol{q}_{ic}$ is the current label confidence, $\bar{H}_i = -\frac{1}{\log C}\sum_{c=1}^{C} \boldsymbol{q}_{ic} \log \boldsymbol{q}_{ic}$ is the normalized entropy, $d_i = \frac{|S_i|}{C}$ is candidate-set density, and cur_i is curriculum level. This design offers a sufficient statistic that characterizes discriminability, uncertainty, and learning dynamics, thereby playing a critical role in enabling reliable action-selection policies.

Actor–Critic Policy. The state is encoded as $\mathbf{h}_i = \phi_\psi(\boldsymbol{s}_i)$, where ϕ_ψ is a multilayer nonlinear encoder. The Actor predicts $\pi_\psi(\boldsymbol{a}_i|\boldsymbol{s}_i) = \text{softmax}(W_A h_i + b_A)$, while the Critic estimates $V_\omega(\boldsymbol{s}_i) = W_V h_i + b_V$, where W_A, W_v, b_A and b_V are the parameters to be learned. The action space is,

$$\mathcal{A}_i = \begin{cases} update, & \boldsymbol{a}_i = 0, \\ retain, & \boldsymbol{a}_i = 1, \\ abstain, & \boldsymbol{a}_i = 2, \end{cases} \tag{6}$$

where action $\boldsymbol{a}_i = 0$ denotes blending the current soft label with the propagated proposal to advance disambiguation, action $\boldsymbol{a}_i = 1$ denotes keeping the current soft label unchanged, which is suitable when the proposal yields no clear improvement. Action $\boldsymbol{a}_i = 2$ denotes postponing disambiguation for this inatsnce, which is suitable for low-reliability instances that are not yet ready for updating.

The inatsnce-level Actor-Critic policy enables each inatsnce to make decisions independently based on its own state, rather than relying on global rules. The three-action design provides a flexible

exploration strategy that helps avoid error accumulation, high-reliability inatsnces can be updated, medium-reliability inatsnces can be kept, and low-reliability inatsnces can abstain. Meanwhile, the curriculum mask ensures that the policy processes only trustworthy inatsnces in the early stage, avoiding the negative effects of erroneous disambiguation.

Self-supervised Reward Function Design. A critical challenge for RL-based PLL is that ground-truth labels are unavailable for reward calculation. We design a fully self-supervised multi-dimensional reward function aligned with disambiguation objectives. For each inatsnce, a candidate update label is first constructed as,

$$\alpha_i = \mathrm{clip}(\lambda_r r_i, \alpha_{\min}, \alpha_{\max}) \tag{7}$$

$$\hat{\boldsymbol{q}}_i = \mathrm{Normalize}_{\mathcal{Y}_i}\left((1-\alpha_i)\boldsymbol{p}_i + \alpha_i \tilde{\boldsymbol{q}}_i\right). \tag{8}$$

where λ_r, $\alpha_{\min}$ and $\alpha_{\max}$ are balance parameters. This formulation indicates that the higher the reliability, the more the KNN-propagated label is trusted. Meanwhile, the lower the reliability, the more the prediction is trusted. Then, the confidence gain and entropy decrease are defined as

$$\mathrm{gain}_i = \mathrm{conf}(\hat{\boldsymbol{q}}_i) - \mathrm{conf}(\boldsymbol{q}_i) + \lambda_H\left(H(\boldsymbol{q}_i) - H(\hat{\boldsymbol{q}}_i)\right), \tag{9}$$

where λ_H is the entropy reward weight. The rewards for the three actions are respectively

$$R_i^U = \mathrm{gain}_i + \lambda_R(r_i - 0.5) - \lambda_D\|\tilde{q}_i - q_i\|_1 \tag{10}$$

$$R_i^R = -\lambda_+ \max(\mathrm{gain}_i, 0) + \lambda_- \max(-\mathrm{gain}_i, 0) \tag{11}$$

$$R_i^A = \lambda_A(1 - r_i) - \lambda_P r_i \tag{12}$$

where λ_R is the reliability reward weight, λ_D is the drift penalty weight, λ_+ and λ_- are the retain reward weights, λ_A and λ_P are the abstention reward and penalty, respectively.

The update reward encourages confidence improvement, entropy reduction, high reliability, and low drift. The retain reward assigns a positive reward when the update gain is negative, thereby avoiding blind updates. The abstain reward gives a positive reward under low reliability and a penalty under high reliability, preventing the policy from degenerating into always abstaining. The ground-truth-label-free reward design enables the policy to learn disambiguation decisions without leaking supervision information. The reward terms correspond to disambiguation quality, stability, and reliability, respectively, guiding the policy to strike a balance between exploration and conservatism.

Reliability-Adaptive Label Correction. When Update is selected, the proposal does not directly replace the current distribution. Instead, the update coefficient is,

$$\beta_i = \mathrm{clip}[\beta_{\min} + (\beta_{\max} - \beta_{\min})r_i, 0, 1] \tag{13}$$

where $\beta_{\max}$ and $\beta_{\min}$ are balance parameters. The new label distribution is

$$q_i^{e+1} = (1 - \beta_i)q_i^e + \beta_i \tilde{q}_i \tag{14}$$

which yields $r_i \uparrow \Rightarrow \beta_i \uparrow$. Thus, highly reliable inatsnces receive stronger corrections, whereas uncertain inatsnces retain more of their previous label information. If Retain or Abstain is selected, $q_i^{e+1} = q_i^e$. This mechanism prevents uncertain instances from undergoing unnecessarily large distributional changes.

Actor–Critic Optimization. The temporal-difference target is $y_i = R_i + \gamma V(\boldsymbol{s}_i')$, where γ is the discount factor, and $\boldsymbol{s}_i'$ is the next state after label update. The advantage estimate is $A_i^{dv} = y_i - V(\boldsymbol{s}_i)$. Advantage measures how much better the taken action is compared to the average expected value. The Actor objective is,

$$\mathcal{L}_A = -\frac{1}{N}\sum_{i=1}^{N} \log \pi_\psi(\boldsymbol{a}_i|\boldsymbol{s}_i) A_i^{dv} - \lambda_E \mathcal{H}(\pi_\psi) \tag{15}$$

where $\mathcal{H}(\pi) = -\sum_a \pi(\boldsymbol{a}|\boldsymbol{s}) \log \pi(\boldsymbol{a}|\boldsymbol{s})$. The Critic is optimized using,

$$\mathcal{L}_C = \frac{1}{N}\sum_{i=1}^{N} \mathrm{SmoothL1}(V(\boldsymbol{s}_i), y_i) \tag{16}$$

Table 1: Accuracy comparisons on real-world partially labeled datasets.

Dataset	Methods							
	DC-RPLL	**CC**	**PRODEN**	**LW**	**CAVL**	**VALEN**	**ED-PLL**	**MR-PLL**
Lost	**76.89 ± 0.26%**	69.29 ± 0.31%	65.17 ± 0.72%	67.11 ± 0.78%	63.11 ± 0.72%	71.56 ± 0.80%	76.78 ± 0.57%	75.85 ± 0.42%
MSRCv2	**60.57 ± 0.33%**	48.57 ± 0.29%	52.00 ± 0.46%	47.15 ± 0.75%	52.84 ± 0.16%	45.89 ± 0.55%	59.85 ± 0.33%	58.85 ± 0.28%
Birdsong	**76.15 ± 0.27%**	56.31 ± 0.81%	63.38 ± 0.52%	66.50 ± 0.73%	70.50 ± 0.92%	72.30 ± 0.49%	74.94 ± 0.17%	75.35 ± 0.34%
Soccer Player	**60.08 ± 0.22%**	49.12 ± 0.34%	51.45 ± 0.19%	49.15 ± 0.38%	54.27 ± 0.37%	53.91 ± 0.12%	58.23 ± 0.46%	56.84 ± 0.28%
Yahoo!News	**68.55 ± 0.26%**	52.47 ± 0.34%	59.06 ± 0.38%	47.03 ± 0.37%	63.86 ± 0.58%	67.73 ± 0.32%	62.34 ± 0.45%	64.98 ± 0.37%

The policy therefore learns from the relative quality of different label actions rather than from explicit ground-truth labels. Actor-Critic learns long-term decisions, enabling the policy to focus not only on the current soft label quality but also on the value of the next state. Entropy regularization encourages exploration and prevents the policy from collapsing prematurely. The Critic provides advantage estimates, reducing the variance of the policy gradient. The overall learning process of the proposed DC-RPLL method is summarized in **Appendix B.1**, and a theoretical analysis of the objective function is provided in **Appendix A**.

4 Experiments

4.1 Experimental Details

Datasets To comprehensively evaluate DC-RPLL, we conduct experiments on four benchmark datasets, including **MNIST** (LeCun et al., 1998), **F-MNIST** (Xiao et al., 2017), **K-MNIST** (Clanuwat et al., 2018), and **CIFAR-10** (Krizhevsky et al., 2009). Following the standard PLL protocol (Yao et al., 2021), candidate label sets are generated by independently adding incorrect labels with probability q (Fan et al., 2024). We consider $q \in \{0.1, 0.3, 0.5\}$ to represent low, moderate, and high ambiguity levels, respectively. We further evaluate DC-RPLL on five real-world PLL datasets, including **Lost** (Cour et al., 2011), **BirdSong** (Briggs et al., 2012), **MSRCv2** (Liu & Dietterich, 2012), **Soccer Player** (Zeng et al., 2013), and **Yahoo!News** (Guillaumin et al., 2010). Detailed dataset statistics are provided in **Appendix B.2**.

Comparison Methods. To better verify the superiority of the DC-RPLL algorithm, we compare the proposed method with the state-of-the-art PLL methods including CC (Feng et al., 2020), PRODEN (Yao et al., 2021), LW (Lv et al., 2020), CAVL (Zhang et al., 2021), VALEN (Xu et al., 2024), ED-PLL (Fan et al., 2026c) and MR-PLL (Fan et al., 2026b). Each comparison method sets parameters according to suggested settings in the original papers, and the details are shown in **Appendix B.3**.

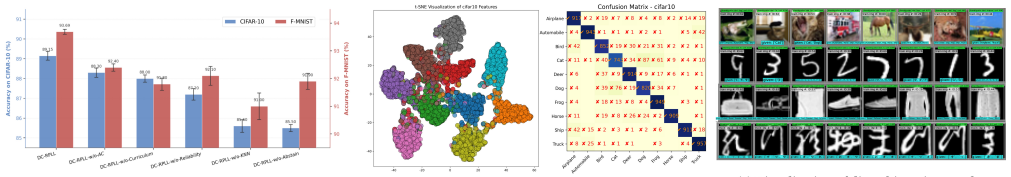
Implementation Details. The implementation of DC-RPLL is based on PyTorch (Paszke et al., 2019), and all experiments are conducted on an NVIDIA GeForce RTX 4090 GPU. The backbone network uses an MLP and ResNet-18 to apply to real-world partially labeled datasets and benchmark datasets, which outputs both class logits and feature representations for classification and neighborhood-based label refinement. The Actor-Critic module consists of two-layer multilayer perceptrons with $ReLU$ activations and a hidden dimension of 256. All models are trained for 300 epochs with a batch size of 256. SGD with a momentum of 0.9 is adopted to optimize the classifier, while the actor and critic use a learning rate of 10^{-3}. For reliability-aware neighborhood propagation, we set the number of nearest neighbors to $K = 8$ and the temperature parameter to $\tau = 0.5$. The reliability-aware fusion coefficient, entropy gain coefficient, reliability reward coefficient, and distribution-drift penalty are set to $\lambda_r = 0.25$, $\lambda_H = 0.7$, $\lambda_R = 0.25$, and $\lambda_D = 0.1$, respectively. The discount factor is set to $\gamma = 0.9$, with an entropy regularization coefficient of $\lambda_E = 0.01$. $\lambda_+ = 0.25$, $\lambda_- = 0.5$, $\lambda_A = 0.25$ and $\lambda_P = 0.2$. The curriculum threshold is linearly annealed from $\eta_{\text{start}} = 0.80$ to $\eta_{\text{end}} = 0.35$, allowing the correction process to gradually expand from highly reliable instances to more ambiguous ones. The correction strength is adaptively determined according to the reliability score, with $\beta_{\min} = 0.10$ and $\beta_{\max} = 0.75$. A 20-epoch warm-up period is used before label refinement, and the neighborhood graph and reliability estimates are updated every 5 epochs.

4.2 Experimental Results

Comparisons with State-of-the-arts methods. On the 40 real-world experiments (5 datasets $\times$ 8 methods) in Table 1, DC-RPLL again achieves the best mean accuracy on **all** datasets, with an average improvement of 0.86% over the strongest competitor. The largest margin appears on Soccer

Table 2: Comparative disambiguation accuracy (mean ± std) across datasets.

Dataset	q	Method							
		DC-RPLL	**CC**	**PRODEN**	**LW**	**CAVL**	**VALEN**	**ED-PLL**	**MR-PLL**
CIFAR-10	0.1	**90.52 ± 0.13%**	85.35 ± 0.19%	84.12 ± 0.36%	81.43 ± 0.13%	83.37 ± 0.07%	83.21 ± 0.10%	89.57 ± 0.11%	90.49 ± 0.14%
	0.3	**89.15 ± 0.22%**	82.50 ± 0.24%	83.36 ± 0.46%	80.95 ± 0.17%	77.53 ± 0.34%	82.87 ± 0.26%	87.95 ± 0.32%	88.82 ± 0.27%
	0.5	**87.42 ± 0.15%**	79.33 ± 0.41%	77.52 ± 0.18%	78.72 ± 0.17%	73.25 ± 0.17%	81.88 ± 0.20%	85.72 ± 0.19%	84.72 ± 0.23%
MNIST	0.1	**99.62 ± 0.08%**	98.57 ± 0.38%	98.91 ± 0.03%	98.76 ± 0.03%	98.95 ± 0.04%	98.75 ± 0.02%	99.27 ± 0.12%	99.62 ± 0.32%
	0.3	**99.56 ± 0.12%**	98.28 ± 0.23%	98.71 ± 0.02%	98.64 ± 0.02%	98.90 ± 0.15%	98.66 ± 0.03%	99.16 ± 0.09%	99.39 ± 0.12%
	0.5	**99.52 ± 0.17%**	97.86 ± 0.19%	98.67 ± 0.03%	98.41 ± 0.03%	98.71 ± 0.04%	98.32 ± 0.02%	98.75 ± 0.15%	99.25 ± 0.13%
F-MNIST	0.1	**94.57 ± 0.28%**	90.83 ± 0.17%	88.27 ± 0.03%	88.18 ± 0.05%	90.32 ± 0.08%	90.72 ± 0.06%	92.28 ± 0.17%	94.41 ± 0.32%
	0.3	**93.69 ± 0.09%**	89.92 ± 0.41%	87.83 ± 0.04%	88.02 ± 0.04%	89.77 ± 0.04%	90.16 ± 0.06%	90.88 ± 0.22%	93.45 ± 0.41%
	0.5	**93.03 ± 0.14%**	88.76 ± 0.16%	87.01 ± 0.03%	87.66 ± 0.06%	88.92 ± 0.11%	89.51 ± 0.04%	89.46 ± 0.49%	92.81 ± 0.27%
K-MNIST	0.1	**98.49 ± 0.24%**	92.52 ± 0.26%	91.11 ± 0.03%	92.78 ± 0.04%	93.73 ± 0.16%	92.67 ± 0.06%	96.82 ± 0.31%	97.58 ± 0.34%
	0.3	**98.45 ± 0.12%**	91.17 ± 0.04%	90.67 ± 0.04%	91.28 ± 0.03%	93.57 ± 0.11%	91.75 ± 0.22%	94.89 ± 0.18%	96.87 ± 0.23%
	0.5	**98.31 ± 0.09%**	89.44 ± 0.48%	89.00 ± 0.05%	86.55 ± 0.06%	91.57 ± 0.22%	90.40 ± 0.03%	91.52 ± 0.24%	96.36 ± 0.41%



(a) Ablation study of the DC-RPLL (b) T-SNE Visualization and confusion matrix of the DC-RPLL (c) Visualization of disambiguation results

Figure 2: Validity verification of DC-RPLL.

Player, and on Yahoo!News DC-RPLL exceeds PRODEN, LW, CC, and CAVL by 9.49%, 21.52%, 16.08%, and 4.69%, respectively. Compared with more complex methods such as VALEN and MR-PLL, DC-RPLL still performs better on Birdsong ($+3.85\%$ and $+0.80\%$) and MSRCv2 ($+14.68\%$ and $+1.72\%$), indicating that the gains arise from principled reliability integration rather than model capacity alone. Across all 96 benchmark experiments (4 datasets $\times$ 3 partial rates $\times$ 8 methods) in Table 2, DC-RPLL achieves the best disambiguation accuracy in **100%** of cases. On CIFAR-10 at $q = 0.5$, DC-RPLL reaches 87.42%, outperforming the strongest baseline by 1.70%. As q increases from 0.1 to 0.5, DC-RPLL gracefully degraded, with only 3.10% on CIFAR-10 and 0.18% on K-MNIST, while CC and other baselines decreased by 6.02% and 3.39%, respectively. This robustness to label ambiguity stems from the dynamic reliability curriculum and Actor-Critic decision mechanism, which prevent low-reliability instances from contaminating soft labels when candidate sets are large.

Analysis of ablation study. To evaluate the contribution of each component in DC-RPLL, we conduct ablation studies on CIFAR-10 ($q = 0.3$) and F-MNIST. As shown in Figure 2(a), ablation experiments were performed on seven specific parts, i.e., DC-RPLL-w/o-AC replaces the Actor-Critic policy with a fixed rule that updates when $r_i \geq \eta_e$ and abstains otherwise, DC-RPLL-w/o-Curriculum removes the dynamic threshold by setting $\eta_e = 0$ so that all instances are admitted, DC-RPLL-w/o-Reliability fixes the update strength to $\beta_i = 0.5$ instead of adapting it to r_i, DC-RPLL-w/o-KNN drops the KNN propagation and uses the classifier prediction directly as the proposal and DC-RPLL-w/o-Abstain removes the abstain action and forces the policy to choose between $update$ and $retain$. Overall, the ablation results verify that all components work synergistically to improve the robustness and effectiveness of DC-RPLL. Removing any component leads to performance degradation, demonstrating the effectiveness of the Actor-Critic policy, dynamic curriculum, reliability-adaptive updating, KNN propagation, and abstain action. These components jointly enable reliable instance-wise decision-making, progressive label correction, effective neighborhood information utilization, and robust handling of ambiguous instances, contributing to the overall performance of DC-RPLL.

Visualization of disambiguation results. To further illustrate the effectiveness of DC-RPLL, we visualize the learned feature representations on **CIFAR-10** using T-SNE (Maaten & Hinton, 2008). As shown in Figure 2(b) and (c), DC-RPLL learns more discriminative features with clearer class boundaries and better clustering structures. By leveraging reliability-aware neighborhood propagation, instances from the same class become more compact, while features from different classes are better separated. This enhanced representation facilitates similarity estimation among instances, thereby improving ambiguous label detection and disambiguation. Moreover, the reliability score r_i provides a calibrated signal for assessing the credibility of candidate labels and dynamically refining the disambiguation process through the dynamic curriculum and the Actor-Critic policy. The visualization of confusion matrix and disambiguation results further confirms the effectiveness of DC-RPLL in achieving accurate label disambiguation.

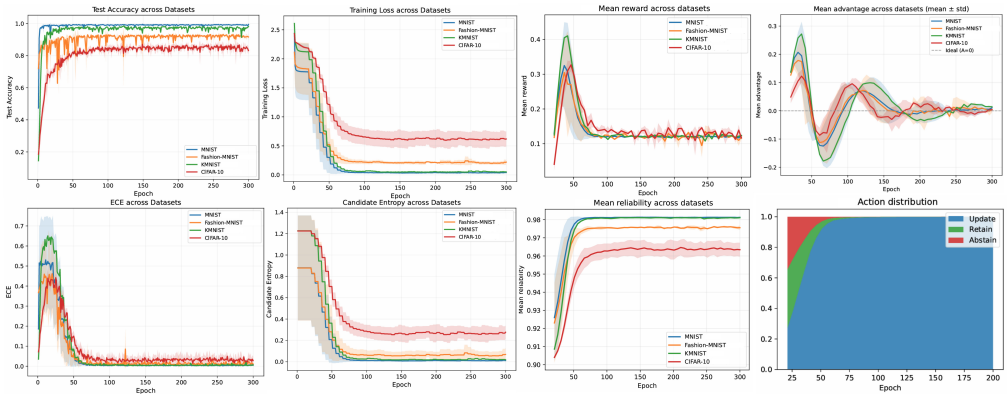


Figure 3: Effectiveness of DC-RPLL.

Analysis of the Effectiveness of DC-RPLL. As shown in Figure 3, the proposed method exhibits stable training dynamics on MNIST, F-MNIST, KMNIST, and CIFAR-10. Test accuracy rises quickly early and then stabilizes. Candidate entropy first remains high, then drops rapidly and stays low, indicating progressive label disambiguation aligned with accuracy gains. *Expected Calibration Error* (ECE) also decreases, reflecting improved calibration and more stable reliability estimation. In policy behavior, early training favors *abstain*/*retain* to avoid aggressive updates, while *update* gradually increases later, reducing confirmation bias. In RL optimization, Mean Reward rises quickly and then stabilizes, while Mean Advantage shifts from oscillation to near zero, indicating stable Actor–Critic optimization. Overall, the method forms a consistent loop, performance improvement, entropy reduction, better calibration, higher reliability, and a shift from *abstain*/*retain* to Update, consistent with the Reliability-aware Curriculum design. Furthermore, the computational complexity of the DC-RPLL can reach $\mathcal{O}\big(\frac{E-W}{T}N^2d\big)$. The detailed calculation process is shown in **Appendix B.5**. Figure 3 illustrates the performance of DC-RPL under various parameter configurations, DC-RPL achieves optimal performance at $k = 8$, initial curriculum threshold $\eta_{\text{start}} = 0.8$, final curriculum threshol $\eta_{\text{end}} = 0.35$, reliability reward weight $\lambda_R = 0.25$ and entropy-reduction weight $\lambda_H = 0.7$. These results further demonstrate the effectiveness of the DC-RPLL algorithm among its various components.

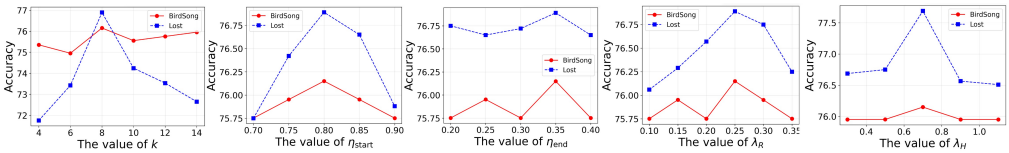


Figure 4: Parameter sensitivity analysis of DC-RPLL.

5 Conclusion

In this paper, we proposed a novel *dynamic reliability curriculum guided reinforcement learning framework for PLL* (DC-RPLL) framework for reliability-driven selective label disambiguation in PLL. The highlight of DC-RPLL is that it integrates reliability-aware label propagation, an easy-to-hard curriculum, and reinforcement learning to enable progressive and selective correction of ambiguous labels. Specifically, DC-RPLL first combines deep classifier predictions with KNN-based local label propagation to generate structure-aware candidate-label proposals, and then employs feature affinity and soft-label consistency to estimate instance reliability and dynamically balance classifier predictions with neighborhood information. Furthermore, the proposed curriculum strategy prioritizes high-reliability instances for disambiguation while restricting low-reliability instances to conservative Abstain actions. An Actor-Critic policy is subsequently employed to select among $update$, $retain$, and $abstain$ actions, guided by a confidence–entropy-based self-supervised reward. Finally, reliability-aware update strengths enable gradual rather than aggressive label refinement, thereby reducing error accumulation and confirmation bias. Extensive experiments demonstrate the effectiveness of DC-RPLL across benchmark and real-world datasets, highlighting the potential of combining reliability-aware curriculum learning and reinforcement learning for robust partial label disambiguation.

REFERENCES

Susan Amin, Maziar Gomrokchi, Harsh Satija, Herke Van Hoof, and Doina Precup. A survey of exploration methods in reinforcement learning. *arXiv preprint arXiv:2109.00157*, 2021.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Tamara Berg, Alexander Berg, Jaety Edwards, and David Forsyth. Who's in the picture. *Advances in neural information processing systems*, 17, 2004.

Forrest Briggs, Xiaoli Z Fern, and Raviv Raich. Rank-loss support instance machines for miml instance annotation. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 534–542, 2012.

Jing Chai, Ivor W Tsang, and Weijie Chen. Large margin partial label machine. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2594–2608, 2019.

Yi-Chen Chen, Vishal M Patel, Jaishanker K Pillai, Rama Chellappa, and P Jonathon Phillips. Dictionary learning from ambiguously labeled data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 353–360, Portland Oregon, USA, 2013.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *The Journal of Machine Learning Research*, 12:1501–1536, 2011.

Jinfu Fan, Yang Yu, Zhongjie Wang, and Jinyi Gu. Partial label learning based on disambiguation correction net with graph representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):4953–4967, 2021.

Jinfu Fan, Yang Yu, and Zhongjie Wang. Addressing label ambiguity imbalance in candidate labels: Measures and disambiguation algorithm. *Information Sciences*, 612:1–19, 2022.

Jinfu Fan, Zhencun Jiang, Yuanqing Xian, and Zhongjie Wang. A multi-class partial hinge loss for partial label learning. *Applied Intelligence*, 53(23):28333–28348, 2023.

Jinfu Fan, Linqing Huang, Chaoyu Gong, Yang You, Min Gan, and Zhongjie Wang. Kmt-pll: K-means cross-attention transformer for partial label learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):2789–2800, 2024.

Jinfu Fan, Jiangnan Li, Linqing Huang, Min Gan, and CL Philip Chen. Partial label zero-shot learning with semantic mining and instance–label alignment. *Engineering Applications of Artificial Intelligence*, 177:114854, 2026a.

Jinfu Fan, Jiangnan LI, Xiaohui Zhong, Kangrui Ren, and Linqing Huang. Mutual information guided reinforcement learning for ambiguous label disambiguation. In *IJCAI 2026*, 2026b.

Jinfu Fan, Jiangnan Li, Xiaohui Zhong, Kangrui Ren, Zhencun Jiang, Min Gan, Tianhao Gu, and Linqing Huang. Evidential deep partial label learning to quantify disambiguation uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24770–24779, 2026c.

Jian Feng, Linqing Huang, JiangNan Li, Kangrui Ren, JinFu Fan, QingKai Bu, Min Gan, and CL Philip Chen. Partial label learning via mutual information representation learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, and Masashi Sugiyama. Provably consistent partial-label learning. *Advances in neural information processing systems*, 33: 10948–10960, 2020.

Tobias Fuchs, Florian Kalinke, and Klemens Böhm. Partial-label learning with a reject option. *arXiv preprint arXiv:2402.00592*, 2024.

Xiuwen Gong, Nitin Bisht, and Guandong Xu. Does label smoothing help deep partial label learning? In *Forty-first International Conference on Machine Learning*, 2024.

Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multiple instance metric learning from automatically labeled bags of faces. In *European conference on computer vision*, pp. 634–647. Springer, 2010.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Zhiyong Li, Linqing Huang, Tianhao Gu, Qingkai Bu, Fuyu Qi, and Jinfu Fan. Partial multi-label learning via k-means graph transformer. *Knowledge-Based Systems*, 325:114017, 2025.

Liping Liu and Thomas Dietterich. A conditional multinomial mixture model for superset label learning. *Advances in Neural Information Processing Systems*, 25:548–556, 2012.

Jie Luo and Francesco Orabona. Learning from candidate labeling sets. *Advances in Neural Information Processing Systems*, 23:1504–1512, 2010.

Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In *international conference on machine learning*, pp. 6500–6510. PMLR, 2020.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Hongwei Wen, Jingyi Cui, Hanyuan Hang, Jiabin Liu, Yisen Wang, and Zhouchen Lin. Leveraged weighted loss for partial label learning. In *International conference on machine learning*, pp. 11091–11100. PMLR, 2021.

Dong-Dong Wu, Deng-Bao Wang, and Min-Ling Zhang. Distilling reliable knowledge for instance-dependent partial label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15888–15896, 2024.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Ning Xu, Congyu Qiao, Xin Geng, and Min-Ling Zhang. Instance-dependent partial label learning. *Advances in Neural Information Processing Systems*, 34:27119–27130, 2021.

Ning Xu, Congyu Qiao, Yuchen Zhao, Xin Geng, and Min-Ling Zhang. Variational label enhancement for instance-dependent partial label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Yao Yao, Jiehui Deng, Xiuhua Chen, Chen Gong, Jianxin Wu, and Jian Yang. Deep discriminative cnn with temporal ensembling for ambiguously-labeled image classification. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pp. 12669–12676, 2020.

Yao Yao, Chen Gong, Jiehui Deng, and Jian Yang. Network cooperation with progressive disambiguation for partial label learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part II*, pp. 471–488. Springer, 2021.

Zinan Zeng, Shijie Xiao, Kui Jia, Tsung-Han Chan, Shenghua Gao, Dong Xu, and Yi Ma. Learning by associating ambiguously labeled images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 708–715, 2013.

Fei Zhang, Lei Feng, Bo Han, Tongliang Liu, Gang Niu, Tao Qin, and Masashi Sugiyama. Exploiting class activation value for partial-label learning. In *International conference on learning representations*, 2021.

Min-Ling Zhang, Bin-Bin Zhou, and Xu-Ying Liu. Partial label learning via feature-aware disambiguation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1335–1344, San Francisco, USA, 2016.