Semi-supervised Label Distribution Learning from Incomplete Annotations via Multi-view Fusion and Reliability-aware Filtering

Jinao Li[1,2], Jinyong Cheng[1,2,*], Kening Cui[1,2]
[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education,
Shandong Computer Science Center (National Supercomputer Center in Jinan),
Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[2]Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing,
Shandong Fundamental Research Center for Computer Science, Jinan, China
10431230216@stu.qlu.edu.cn, cjy@qlu.edu.cn, 10431230146@stu.qlu.edu.cn
*Corresponding Author: Jinyong Cheng Email: cjy@qlu.edu.cn

***Abstract*—Label Distribution Learning (LDL) models label ambiguity by assigning each instance a soft distribution over labels. In many real applications, however, obtaining complete label distributions is costly: annotators typically provide explicit description degrees for only a few dominant labels, leaving the remaining entries unobserved. This leads to incomplete label distributions and substantially weakens supervision. In this paper, we study semi-supervised label distribution learning under incomplete annotations and propose CoFuse-LDL, an EMA teacher–student framework that leverages unlabeled data via multi-view probability fusion and reliability-aware pseudo-distribution filtering. Specifically, the teacher and student predict distributions from weak and strong augmentations, and we fuse multi-branch predictions to obtain a robust pseudo-distribution. We then apply a confidence-based gate to filter unreliable pseudo targets and enforce distribution consistency on unlabeled samples with a KL divergence loss, while labeled samples are optimized using a masked cross-entropy on observed entries. Experiments on six LDL benchmarks with an observable ratio of 0.5 show that CoFuse-LDL consistently outperforms competitive baselines, demonstrating the benefit of fusing and filtering distribution-valued pseudo supervision when label distributions are only partially observed.**

***Index Terms*—Label Distribution Learning; Semi-supervised Learning; Incomplete Label Distributions; Teacher–Student Learning; Multi-view Fusion; Reliability-aware Filtering**

I. Introduction

Label Distribution Learning (LDL) [1] has become a widely used paradigm for handling label ambiguity by representing each instance with a continuous label distribution. In LDL, each label is associated with a description degree that reflects its relevance to the instance. Compared with conventional single-label or multi-label supervision, LDL offers a finer-grained representation that can capture mixed semantics and potential dependencies among labels. This formulation has shown effectiveness in a variety of computer vision and pattern recognition tasks, such as facial age estimation [2], visual sentiment distribution prediction [3], crowd opinion prediction on movies [4], emotion distribution modeling [5], and aesthetic beauty perception [6].

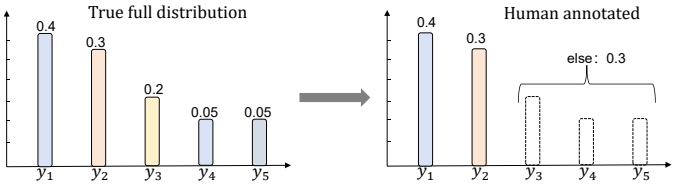

Fig. 1: Illustration of incomplete label distributions. The ground-truth distribution covers all labels, but in practice only description degrees of a few dominant labels are provided, and the remaining entries are unobserved or aggregated, yielding partial supervision.

Driven by its expressive power, LDL has witnessed rapid methodological progress, largely under the assumption that high-quality label distributions are available for training. For instance, Wen *et al.* [7] introduced ordinal label distribution learning, which preserves ordinal ranking relations by imposing cumulative consistency, and is particularly effective for ordinal regression. In addition, label enhancement can enrich supervision when annotations are indirect or weak; Jia *et al.* [8] exploited label correlations on local samples to recover latent distribution information from limited supervision. These studies suggest that, when label distributions are sufficiently informative and densely specified, LDL can effectively leverage soft supervision to improve generalization and robustness.

Despite these advantages, practical deployment of LDL is often hindered by the cost of acquiring complete and reliable distributions. In real annotation pipelines, annotators can usually identify a few dominant labels, yet it is difficult to assign precise description degrees to every label in a high-dimensional label space. As illustrated in Fig. 1, the resulting supervision is typically incomplete: only a sparse subset of labels is explicitly annotated, whereas the remaining entries are missing or coarsely aggregated. This partial observability weakens the supervision signal and introduces systematic bias toward the observed labels, leaving less salient labels severely under-supervised and

making optimization more fragile under imperfect supervision.

Meanwhile, modern applications provide abundant unlabeled data that are much easier to collect than full distributions. This motivates semi-supervised label distribution learning under incomplete supervision, where the goal is to bridge sparse, partially observed distributions and large-scale unlabeled samples. Compared to standard semi-supervised classification, SSL-LDL faces a distinctive difficulty: the target is a distribution rather than a discrete class, and pseudo supervision must allocate probability mass over both observed and unobserved labels. Without careful control, errors in pseudo-distributions can accumulate and propagate to unobserved entries, leading to biased distribution estimates and degraded performance.

A common approach to exploit unlabeled data is consistency regularization with a teacher–student architecture, typically implemented with an EMA teacher and confidence-thresholded pseudo supervision under weak and strong augmentations [9]–[11]. However, directly transferring generic semi-supervised pipelines to label distribution learning is often ineffective in the incomplete-distribution regime, mainly for three reasons. First, the pseudo target in LDL is a high-dimensional probability vector. Small mass shifts across labels can accumulate into nontrivial distribution drift, which increases confirmation bias and destabilizes self-training, especially when supervision is sparse or partial [10]. Second, incomplete label distributions provide supervision only on a subset of dominant labels, which can bias the predictor toward observed entries and yield systematically skewed pseudo-distributions on unlabeled data. Such bias can be amplified over iterations and forms a self-reinforcing error loop [12]. Third, distribution predictions can vary substantially across stochastic augmentations and across teacher and student branches, producing noisy pseudo targets and slowing down LDL optimization.

To address these challenges, we propose CoFuse-LDL, a reliability-guided framework that combines multi-view probability fusion with pseudo-distribution filtering for semi-supervised LDL under incomplete label distributions, as shown in Fig. 2. CoFuse-LDL maintains an EMA teacher and a student network and generates multiple distribution predictions from weakly and strongly augmented views. Instead of adopting a single branch as pseudo supervision, a fusion module aggregates multi-branch outputs into a refined pseudo-distribution with reduced variance. A reliability-aware filtering strategy then selects confident pseudo targets to regularize unlabeled samples via KL-based distribution matching, while labeled samples are optimized with a masked cross-entropy objective on observed description degrees.

Our main contributions are summarized as follows:

- We study semi-supervised label distribution learning under incomplete label distributions, where only a subset of description degrees is observable for each labeled sample, and unlabeled data must be exploited without introducing severe distribution drift.
- We propose CoFuse-LDL, which constructs robust distribution-valued pseudo targets via multi-view fusion and suppresses unreliable pseudo supervision via reliability-aware filtering, improving training stability under incomplete supervision.
- Experiments on multiple LDL benchmarks with an observable-ratio protocol demonstrate consistent improvements over competitive baselines, validating the effectiveness of fusing and filtering pseudo-distributions for learning from incomplete supervision.

II. Related Work

A. Label Distribution Learning

Label Distribution Learning (LDL) [1] is a learning paradigm that maps each instance to a distribution of description degrees over a set of labels, where each value indicates how well a label describes the instance. This paradigm was first introduced by Geng [1] to handle scenarios involving label ambiguity. Traditional LDL methods fall into three categories: problem transformation (PT), algorithm adaptation (AA), and specialized algorithms [1]. Representative examples include PT-SVM [2], PT-Bayes [2], AA-KNN [1], AA-BP [1], *etc*.

Recent works on LDL have explored various strategies to address label ambiguity. LDL-END [13] improves the kNN-based framework by jointly learning data-dependent weights and incorporating neighborhood label correlations. Adaptive weighted ranking-oriented LDL [14] introduces a label ranking mechanism to improve robustness under noisy distributions. Beyond algorithm design, LDL has also been applied to many real-world applications, such as age estimation [15], emotion analysis [16], multimodal fusion [17], head pose estimation [18], *etc.*

B. Incomplete Label Distribution Learning

In many real annotation pipelines, complete label distributions are difficult to obtain. Annotators often provide description degrees for only a few dominant labels, while the remaining entries are unobserved or merged into a coarse group. This setting is commonly referred to as incomplete label distribution learning (ILDL) [19].

Existing ILDL methods mainly aim to recover the missing description degrees from partially observed supervision. A typical strategy is to exploit sample relatedness, such that instances close in the feature space tend to share similar distributions, enabling missing entries to be inferred from neighbors [1]. More recent approaches formulate ILDL as a completion problem and reconstruct missing degrees by modeling structural regularities of the distribution matrix, such as low rank correlations and feature manifold constraints [19]. Related challenging settings, such as imbalanced label distribution learning, further highlight that biased supervision can degrade distribution prediction on tail labels [20].

Despite these efforts, most ILDL studies focus on recovery within the labeled set and do not directly leverage abundant unlabeled data. In practice, incompleteness can induce biased supervision, where tail labels receive few explicit signals, and reconstruction errors may propagate to the predictor. Different from prior ILDL methods, our work studies a semi supervised setting and exploits unlabeled samples through

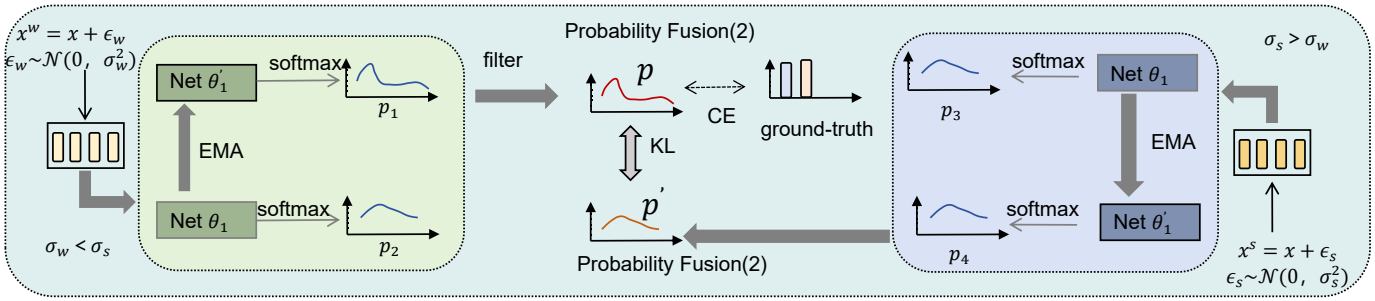

Fig. 2: Overview of CoFuse-LDL. A student and an EMA teacher take weakly and strongly augmented views to produce multiple distribution predictions. Probability fusion aggregates these predictions into a robust pseudo-distribution. Reliability-aware filtering retains high-confidence pseudo targets for KL-based consistency on unlabeled data, while labeled samples are trained with a masked cross-entropy objective on observed description degrees.

distribution valued pseudo supervision. We generate pseudo distributions with an EMA teacher student scheme [9], adopt weak and strong augmentations following the consistency and confidence paradigm [10], [11], reduce variance by fusing multi view predictions, and suppress confirmation bias by filtering unreliable pseudo targets.

III. Method

We study semi-supervised label distribution learning with incomplete label distributions. CoFuse-LDL adopts an exponential moving average teacher–student scheme and exploits unlabeled data via fused pseudo-distributions with reliability-aware filtering.

A. Problem Setup and Overall Objective

Let $\mathcal{D}_l = \{(x_i, \tilde{\boldsymbol{y}}_i, \boldsymbol{m}_i)\}_{i=1}^{N_l}$ be the labeled set and $\mathcal{D}_u = \{x_j\}_{j=1}^{N_u}$ be the unlabeled set, where samples in $\mathcal{D}_u$ contain no label distributions. Each input $x \in \mathcal{X}$ is associated with a label distribution over K labels. A valid distribution lies in the probability simplex

$$\Delta^{K-1} = \Big\{ \boldsymbol{y} \in \mathbb{R}^K \;\Big|\; y_k \geq 0,\ \sum_{k=1}^{K} y_k = 1 \Big\}. \tag{1}$$

For labeled sample i, the supervision is incomplete and represented by $(\tilde{\boldsymbol{y}}_i, \boldsymbol{m}_i)$, where $\tilde{\boldsymbol{y}}_i \in [0,1]^K$ stores the observed description degrees and $\boldsymbol{m}_i \in \{0,1\}^K$ is a binary mask. Specifically, $m_{ik} = 1$ if $\tilde{y}_{ik}$ is observed and $m_{ik} = 0$ otherwise. Unobserved entries are treated as missing rather than imputed. The observable ratio is denoted by $\alpha \in (0,1]$ and is fixed to $\alpha = 0.5$ in our experiments, meaning that each labeled distribution keeps the top $\lceil \alpha K \rceil$ description degrees ranked by magnitude and masks the rest. Since $\tilde{\boldsymbol{y}}_i$ stores partial degrees, it is not required that $\sum_k \tilde{y}_{ik} = 1$.

We learn a student predictor $f_\theta : \mathcal{X} \to \mathbb{R}^K$ with parameters θ, producing logits $\boldsymbol{z}_\theta(x) = f_\theta(x)$. Throughout this section, all predictive distributions are computed by a temperature-scaled softmax:

$$\boldsymbol{p}_\theta(x) = \mathrm{softmax}(\boldsymbol{z}_\theta(x)/T), p_{\theta,k}(x) = \frac{\exp(z_{\theta,k}(x)/T)}{\sum_{r=1}^{K} \exp(z_{\theta,r}(x)/T)}, \tag{2}$$

where $T > 0$ is the temperature parameter and $T = 1$ recovers the standard softmax.

Training minimizes a labeled loss and an unlabeled regularization loss:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda \mathcal{L}_{\text{u}}, \tag{3}$$

where $\lambda \geq 0$ controls the strength of unlabeled regularization. Let $\mathcal{B}_l \subset \mathcal{D}_l$ and $\mathcal{B}_u \subset \mathcal{D}_u$ be mini-batches with sizes $|\mathcal{B}_l|$ and $|\mathcal{B}_u|$.

For labeled data, we fit only the observed entries and normalize the observed degrees on their support to obtain a proper target distribution. Define the observed mass

$$c_i = \sum_{k=1}^{K} m_{ik} \tilde{y}_{ik}, \tag{4}$$

and the normalized observed target

$$\bar{y}_{ik} = \begin{cases} \tilde{y}_{ik}/c_i, & \text{if } m_{ik} = 1 \text{ and } c_i > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Denote by $\boldsymbol{p}_{s,i}^w$ the student prediction on the weakly augmented view of x_i defined in (11). The supervised loss is a masked cross-entropy on the observed support:

$$\mathcal{L}_{\text{sup}}(\mathcal{B}_l) = \frac{1}{|\mathcal{B}_l|} \sum_{(x_i, \tilde{\boldsymbol{y}}_i, \boldsymbol{m}_i) \in \mathcal{B}_l} \left(-\sum_{k=1}^{K} m_{ik}\, \bar{y}_{ik}\, \log p_{s,ik}^w \right). \tag{6}$$

B. Teacher–Student Learning with Fusion and Filtering

We generate two stochastic views of each input x using a weak augmentation operator $\mathcal{A}_w(\cdot)$ and a strong augmentation operator $\mathcal{A}_s(\cdot)$:

$$x^w = \mathcal{A}_w(x), \qquad x^s = \mathcal{A}_s(x), \tag{7}$$

where $\mathcal{A}_s$ applies stronger perturbations than $\mathcal{A}_w$.

Besides the student f_θ, we maintain a teacher network $f_{\theta'}$ with parameters θ'. The teacher is updated by exponential moving average

$$\theta' \leftarrow \mu\, \theta' + (1 - \mu)\, \theta, \tag{8}$$

where $\mu \in [0, 1)$ is the EMA coefficient.

Given the two views, we compute temperature-scaled distributions from student and teacher:

$$\boldsymbol{p}_s^w(x) = \mathrm{softmax}(f_\theta(x^w)/T), \boldsymbol{p}_s^s(x) = \mathrm{softmax}(f_\theta(x^s)/T), \tag{9}$$

$$\boldsymbol{p}_t^w(x) = \mathrm{softmax}(f_{\theta'}(x^w)/T), \boldsymbol{p}_t^s(x) = \mathrm{softmax}(f_{\theta'}(x^s)/T), \tag{10}$$

where the temperature parameter T is shared across all branches. For labeled sample x_i, we write $\boldsymbol{p}_{s,i}^w = \boldsymbol{p}_s^w(x_i)$ and define

$$\boldsymbol{p}_{s,i}^w = \mathrm{softmax}(f_\theta(x_i^w)/T). \tag{11}$$

To stabilize distribution-valued pseudo supervision under incomplete supervision, we fuse predictions from the teacher and the student weak view to form a pseudo target. Define the prediction set

$$\mathcal{P}(x) = \{\boldsymbol{p}_t^w(x), \boldsymbol{p}_t^s(x), \boldsymbol{p}_s^w(x)\}. \tag{12}$$

We compute a fused pseudo-distribution $\hat{\boldsymbol{q}}(x) \in \Delta^{K-1}$ via a convex combination:

$$\hat{\boldsymbol{q}}(x) = \sum_{\boldsymbol{p}\in\mathcal{P}(x)} w_{\boldsymbol{p}}(x)\, \boldsymbol{p}, w_{\boldsymbol{p}}(x) \geq 0, \sum_{\boldsymbol{p}\in\mathcal{P}(x)} w_{\boldsymbol{p}}(x) = 1, \tag{13}$$

where $w_{\boldsymbol{p}}(x)$ is the fusion weight for branch $\boldsymbol{p}$. In the non-adaptive setting, we use uniform weights $w_{\boldsymbol{p}}(x) = 1/|\mathcal{P}(x)|$.

We filter unreliable pseudo targets using a confidence-based gate. The confidence score is defined as

$$s(x) = \max_{k\in\{1,\ldots,K\}} \hat{q}_k(x), \tag{14}$$

and the gate is

$$g(x) = \mathbf{1}[s(x) \geq \tau], \tag{15}$$

where $\tau \in (0,1)$ is the confidence threshold and $\mathbf{1}[\cdot]$ is the indicator function.

For unlabeled samples $x_j \in \mathcal{B}_u$, we enforce the student prediction on the strong view to match the fused pseudo target by KL divergence:

$$\mathcal{L}_{\mathrm{u}}(\mathcal{B}_u) = \frac{1}{\max(1,G)} \sum_{x_j\in\mathcal{B}_u} g(x_j)\, \mathrm{KL}(\mathrm{sgrad}(\hat{\boldsymbol{q}}(x_j)) \parallel \boldsymbol{p}_s^s(x_j)), \tag{16}$$

where $\mathrm{sgrad}(\cdot)$ blocks gradients through the pseudo target so that $\mathcal{L}_{\mathrm{u}}$ updates only the student parameters θ. Here $G = \sum_{x_j\in\mathcal{B}_u} g(x_j)$ denotes the number of selected unlabeled samples in the mini-batch. The KL divergence between $\boldsymbol{a}, \boldsymbol{b} \in \Delta^{K-1}$ is

$$\mathrm{KL}(\boldsymbol{a}\|\boldsymbol{b}) = \sum_{k=1}^{K} a_k \log \frac{a_k}{b_k}. \tag{17}$$

The overall mini-batch objective follows (3):

$$\mathcal{L}(\mathcal{B}_l, \mathcal{B}_u) = \mathcal{L}_{\mathrm{sup}}(\mathcal{B}_l) + \lambda\, \mathcal{L}_{\mathrm{u}}(\mathcal{B}_u). \tag{18}$$

At training iteration $t \in \{1, \ldots, T_{\max}\}$, we update θ by gradient-based optimization on $\mathcal{L}(\mathcal{B}_l, \mathcal{B}_u)$, and then update θ' by (8). The pseudo code of the proposed framework is provided in **Algorithm 1**.

Algorithm 1 CoFuse-LDL: teacher–student learning with fusion and filtering

Require: Labeled set $\mathcal{D}_l = \{(x_i, \tilde{\boldsymbol{y}}_i, \boldsymbol{m}_i)\}_{i=1}^{N_l}$, unlabeled set $\mathcal{D}_u = \{x_j\}_{j=1}^{N_u}$; student θ, teacher θ'; EMA μ; unlabeled weight λ; threshold τ; temperature T; augmentations $\mathcal{A}_w, \mathcal{A}_s$; fusion weights $w_{\boldsymbol{p}}(\cdot)$; optimizer Opt; iterations $T_{\max}$

Ensure: Trained student parameters θ

Initialize θ and set $\theta' \leftarrow \theta$
for $t = 1$ to $T_{\max}$ **do**
  Sample mini-batches $\mathcal{B}_l \subset \mathcal{D}_l$ and $\mathcal{B}_u \subset \mathcal{D}_u$
  $\mathcal{L}_{\mathrm{sup}} \leftarrow 0,\ \mathcal{L}_{\mathrm{u}} \leftarrow 0,\ G \leftarrow 0$
  for each $(x_i, \tilde{\boldsymbol{y}}_i, \boldsymbol{m}_i) \in \mathcal{B}_l$ **do**
    $x_i^w \leftarrow \mathcal{A}_w(x_i), \quad \boldsymbol{p}_{s,i}^w \leftarrow \mathrm{softmax}(f_\theta(x_i^w)/T)$
    $c_i \leftarrow \sum_k m_{ik}\tilde{y}_{ik}$
    $\mathcal{L}_{\mathrm{sup}} \mathrel{+}= -\sum_{k:\, m_{ik}=1} \frac{\tilde{y}_{ik}}{\max(c_i,\epsilon)} \log p_{s,ik}^w$
  end for
  $\mathcal{L}_{\mathrm{sup}} \leftarrow \mathcal{L}_{\mathrm{sup}}/|\mathcal{B}_l|$
  for each $x_j \in \mathcal{B}_u$ **do**
    $x_j^w \leftarrow \mathcal{A}_w(x_j), \quad x_j^s \leftarrow \mathcal{A}_s(x_j)$
    $\boldsymbol{p}_s^w \leftarrow \mathrm{softmax}(f_\theta(x_j^w)/T),$
    $\boldsymbol{p}_s^s \leftarrow \mathrm{softmax}(f_\theta(x_j^s)/T)$
    $\boldsymbol{p}_t^w \leftarrow \mathrm{softmax}(f_{\theta'}(x_j^w)/T),$
    $\boldsymbol{p}_t^s \leftarrow \mathrm{softmax}(f_{\theta'}(x_j^s)/T)$
    $\hat{\boldsymbol{q}} \leftarrow \sum_{\boldsymbol{p}\in\{\boldsymbol{p}_t^w, \boldsymbol{p}_t^s, \boldsymbol{p}_s^w\}} w_{\boldsymbol{p}}(x_j)\, \boldsymbol{p}$
    $g \leftarrow \mathbf{1}[\max_k \hat{q}_k \geq \tau], \quad G \leftarrow G + g$
    $\mathcal{L}_{\mathrm{u}} \mathrel{+}= g \cdot \mathrm{KL}(\mathrm{stopgrad}(\hat{\boldsymbol{q}}) \parallel \boldsymbol{p}_s^s)$
  end for
  $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{sup}} + \lambda\, \mathcal{L}_{\mathrm{u}}/\max(1, G)$
  $\theta \leftarrow \mathrm{Opt}(\theta, \nabla_\theta \mathcal{L})$
  $\theta' \leftarrow \mu\, \theta' + (1-\mu)\, \theta$
end for
return θ

IV. Experimental Setup

In this section, we evaluate the proposed CoFuse-LDL under incomplete label distributions, where only a subset of description degrees is observable for each training sample. We compare CoFuse-LDL with standard LDL baselines, specialized incomplete-LDL methods, and representative semi-supervised paradigms adapted to distribution prediction.

A. Datasets and Evaluation

We conduct experiments on six widely used label distribution learning benchmarks: *SCUT-FBP* [24], *Flickr-LDL* [3], *Movie* [25], *Emotion6* [5], *Natural Scene* [1], and *RAF-ML* [26]. To simulate incomplete label distributions, we assume annotators provide description degrees only for dominant labels: for each training sample with a ground-truth distribution over K labels, we rank its entries and retain the top $\lceil \alpha K \rceil$ labels as observed, construct a binary mask to mark observed entries, and treat the remaining entries as missing; unless otherwise stated, we set $\alpha = 0.5$. CoFuse-LDL requires both labeled and unlabeled training samples, so for each dataset we split

TABLE I: Chebyshev distance (↓) on ILDL benchmarks with incomplete label distributions (Observable Ratio = 0.5). Only the top 50% labels by description degree are observed; the rest are masked.

Method	*Movie*	*SCUT-FBP*	*Emotion6*	*Flickr_LDL*	*RAF-ML*	*Natural Scene*
SA-BFGS [1]	0.3984 ± 0.010•	0.7852 ± 0.041•	0.8841 ± 0.021•	0.9214 ± 0.018•	0.8125 ± 0.019•	0.7241 ± 0.024•
EDL-LRL [8]	0.3812 ± 0.012•	0.3841 ± 0.025•	0.4421 ± 0.009•	0.6125 ± 0.008•	0.5012 ± 0.015•	0.4682 ± 0.026•
LDL-LDM [21]	0.5124 ± 0.031•	0.4352 ± 0.048•	0.4982 ± 0.018•	0.6024 ± 0.010•	0.5621 ± 0.031•	0.5014 ± 0.025•
RDA [20]	0.2432 ± 0.008•	0.3214 ± 0.018•	0.3854 ± 0.009•	0.5621 ± 0.008•	0.4125 ± 0.008•	0.4021 ± 0.023•
SMC-LDL [19]	0.2294 ± 0.009•	0.3015 ± 0.014•	0.3782 ± 0.007•	0.5482 ± 0.009•	0.3984 ± 0.010•	0.3912 ± 0.021•
NoReg-ILDL [22]	0.2185 ± 0.010•	0.2941 ± 0.021•	0.3712 ± 0.008•	0.5421 ± 0.009•	0.3842 ± 0.012•	0.3854 ± 0.024•
FixMatch-LDL* [10]	0.2241 ± 0.011•	0.2854 ± 0.020•	0.3684 ± 0.010•	0.5312 ± 0.012•	0.3752 ± 0.013•	0.3712 ± 0.020•
FlexMatch-LDL* [11]	0.2115 ± 0.009•	0.2712 ± 0.018•	0.3612 ± 0.009•	0.5241 ± 0.010•	0.3684 ± 0.012•	0.3621 ± 0.019•
MeanTeacher-LDL* [9]	0.2054 ± 0.011•	0.2641 ± 0.024•	0.3542 ± 0.010•	0.5214 ± 0.011•	0.3621 ± 0.014•	0.3541 ± 0.021•
ReMixMatch-LDL* [23]	0.1985 ± 0.009•	0.2521 ± 0.019•	0.3421 ± 0.009•	0.5142 ± 0.009•	0.3512 ± 0.012•	0.3421 ± 0.019•
CoFuse-LDL (Ours)	**0.1825 ± 0.024**	**0.2214 ± 0.031**	**0.3214 ± 0.019**	**0.5012 ± 0.010**	**0.3354 ± 0.028**	**0.3184 ± 0.021**

TABLE II: KL divergence (↓) on ILDL benchmarks with incomplete label distributions (Observable Ratio = 0.5). Only the top 50% labels by description degree are observed; the rest are masked.

Method	*Movie*	*SCUT-FBP*	*Emotion6*	*Flickr_LDL*	*RAF-ML*	*Natural Scene*
SA-BFGS [1]	0.9542 ± 0.065•	15.210 ± 4.821•	24.125 ± 1.254•	31.254 ± 1.842•	21.425 ± 1.521•	5.842 ± 0.452•
EDL-LRL [8]	0.8841 ± 0.052•	0.9541 ± 0.124•	1.6251 ± 0.142•	11.245 ± 5.124•	1.4521 ± 0.124•	3.125 ± 1.842•
LDL-LDM [21]	1.9542 ± 0.214•	1.2541 ± 0.184•	1.9421 ± 0.121•	2.9542 ± 0.182•	2.1241 ± 0.210•	1.984 ± 0.195•
RDA [20]	0.3124 ± 0.018•	0.5214 ± 0.045•	0.8842 ± 0.031•	1.9542 ± 0.142•	0.8521 ± 0.028•	1.342 ± 0.075•
SMC-LDL [19]	0.2984 ± 0.016•	0.4982 ± 0.039•	0.8421 ± 0.029•	1.8841 ± 0.131•	0.8124 ± 0.026•	1.295 ± 0.068•
NoReg-ILDL [22]	0.2842 ± 0.015•	0.4852 ± 0.038•	0.8241 ± 0.028•	1.8241 ± 0.124•	0.7842 ± 0.024•	1.254 ± 0.062•
FixMatch-LDL* [10]	0.2751 ± 0.013•	0.4721 ± 0.036•	0.8124 ± 0.026•	1.7952 ± 0.118•	0.7621 ± 0.031•	1.231 ± 0.058•
FlexMatch-LDL* [11]	0.2654 ± 0.012•	0.4682 ± 0.035•	0.7952 ± 0.025•	1.7854 ± 0.112•	0.7412 ± 0.030•	1.214 ± 0.055•
MeanTeacher-LDL* [9]	0.2541 ± 0.014•	0.4521 ± 0.034•	0.7842 ± 0.024•	1.7241 ± 0.105•	0.7142 ± 0.031•	1.184 ± 0.051•
ReMixMatch-LDL* [23]	0.2312 ± 0.012•	0.4215 ± 0.032•	0.7421 ± 0.022•	1.6541 ± 0.098•	0.6841 ± 0.028•	1.124 ± 0.045•
CoFuse-LDL (Ours)	**0.2054 ± 0.082**	**0.3842 ± 0.051**	**0.6842 ± 0.082**	**1.5241 ± 0.142**	**0.6241 ± 0.185**	**0.924 ± 0.045**

data into training, validation, and test subsets with a ratio of 80%/10%/10%, and then remove label distributions from part of the training samples to form unlabeled data while keeping the rest as labeled data; labeled samples further undergo the observable-ratio masking, whereas unlabeled samples are used without any label distribution.

We report four distribution distances (Chebyshev, Clark, Canberra, and Kullback–Leibler; lower is better) and two similarities (Cosine and Intersection; higher is better), using 10-fold cross-validation with mean ± standard deviation; statistical significance is assessed by a two-tailed t-test at the 0.05 level, where • indicates CoFuse-LDL significantly outperforms the compared method. We compare against standard label distribution learning baselines (SA-BFGS [1], EDL-LRL [8], LDLSF [27], LDL-LCLR [27], Adam-LDL-SCL [8], and LDL-LDM [21]), incomplete-distribution methods (NoReg-ILDL [22], SMC-LDL [19], and RDA [20]), and semi-supervised baselines adapted to distribution learning (FixMatch-LDL* [10], FlexMatch-LDL* [11], MeanTeacher-LDL* [9], and ReMixMatch-LDL* [23]).

B. Implementation Details and Results

All experiments are implemented in PyTorch and conducted on a single NVIDIA GeForce RTX 4060 GPU. We train models with a batch size of 64 and an initial learning rate of 0.001 for up to 300 epochs. Unless otherwise stated, we fix the observable ratio to $\alpha = 0.5$, meaning that for each labeled training sample only the top 50% labels ranked by description degree are treated as observed and the remaining entries are masked as missing.

TABLE III: Ablation study of CoFuse-LDL on the *Movie* dataset with observable ratio 0.5. MF and PF denote multi-view fusion and pseudo-distribution filtering. Bold values indicate the best performance.

Index	MF	PF	Cheb↓	Clark↓	Can↓	KL↓	Cos↑
1			0.2241	0.7962	1.5492	0.2751	0.8162
2	✓		0.2085	0.7241	1.3524	0.2312	0.8421
3	✓	✓	**0.1825**	**0.6842**	**1.1321**	**0.2054**	**0.8725**

We report results under the same observable-ratio protocol for all compared methods.

Tables I and II report the Chebyshev distance and KL divergence on six benchmarks. CoFuse-LDL consistently achieves the best performance across all datasets. To avoid over-claiming a single fixed baseline as the strongest competitor, we compare against the best non-ours method for each dataset. For Chebyshev distance, CoFuse-LDL improves over the best competing method by 0.0160 on *Movie* (0.1825 vs. 0.1985), 0.0307 on *SCUT-FBP* (0.2214 vs. 0.2521), 0.0207 on *Emotion6* (0.3214 vs. 0.3421), and 0.0237 on *Natural Scene* (0.3184 vs. 0.3421). Similar gains are observed in KL divergence, where CoFuse-LDL reduces KL compared with the best competing method by 0.0258 on *Movie* (0.2054 vs. 0.2312), 0.0373 on *SCUT-FBP* (0.3842 vs. 0.4215), 0.0579 on *Emotion6* (0.6842 vs. 0.7421), and 0.2000 on *Natural Scene* (0.9240 vs. 1.1240). These results indicate that CoFuse-LDL improves both worst-case label mismatch, reflected by Chebyshev distance, and

overall distribution alignment, reflected by KL divergence, supporting the effectiveness of probability fusion and reliability-aware filtering when label distributions are only partially observed.

C. Ablation Study

We study the effect of key components in our framework on the *Movie* dataset under incomplete label distributions with observable ratio 0.5. As reported in Table III, starting from a basic teacher–student consistency baseline, introducing multi-view fusion improves distribution prediction quality by reducing the Chebyshev distance from 0.2241 to 0.2085 and the KL divergence from 0.2751 to 0.2312, while increasing the Cosine similarity from 0.8162 to 0.8421. Further enabling pseudo-distribution filtering yields the best performance, bringing the Chebyshev distance down to 0.1825 and the KL divergence down to 0.2054, and improving the Cosine similarity to 0.8725. Overall, fusion stabilizes distribution-level pseudo targets by aggregating complementary predictions, and filtering improves robustness by suppressing unreliable pseudo supervision. The two components are complementary under incomplete supervision.

V. Conclusion

This paper presents our framework for learning label distributions from incomplete annotations. The core idea is to construct reliable distribution-level supervision when only the top-ranked description degrees are observable. By combining multi-view fusion to form stable pseudo targets and pseudo-distribution filtering to suppress unreliable supervision, our method improves both robustness and generalization under incomplete label distributions. Experiments on six LDL benchmarks with observable ratio 0.5 show consistent gains over standard LDL baselines, specialized incomplete-LDL methods, and strong semi-supervised baselines across multiple distance and similarity metrics. In future work, we will extend the framework to more challenging missing patterns and noise conditions, explore stronger uncertainty-aware reliability modeling, and evaluate the approach on broader datasets and real-world incomplete annotation scenarios.

References

[1] Xin Geng, "Label distribution learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1734–1748, 2016.

[2] Xin Geng, Chao Yin, and Zhi-Hua Zhou, "Facial age estimation by learning from label distributions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2401–2412, 2013.

[3] Jufeng Yang, Ming Sun, and Xiaoxiao Sun, "Learning visual sentiment distributions via augmented conditional probability neural network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

[4] Xin Geng and Peng Hou, "Crowd opinion prediction on movies: From review summaries to sentiment distributions," *IEEE Transactions on Multimedia*, vol. 17, no. 6, pp. 765–777, 2015.

[5] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C. Gallagher, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[6] Shu Liu, Enquan Huang, Ziyu Zhou, Yan Xu, Xiaoyan Kui, Tao Lei, and Hongying Meng, "Lightweight facial attractiveness prediction using dual label distribution," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 17, no. 1, pp. 195–207, 2025.

[7] Changsong Wen, Xin Zhang, Xingxu Yao, and Jufeng Yang, "Ordinal label distribution learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[8] Xiuyi Jia, Zechao Li, Xiang Zheng, Weiwei Li, and Sheng-Jun Huang, "Label distribution learning with label correlations on local samples," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1619–1631, 2021.

[9] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *arXiv preprint arXiv:1703.01780*, 2017.

[10] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A. Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Cheng-Wen Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems*, 2020.

[11] Bowen Zhang, Yidong Wang, Wenxin Hou, Hing-Bin Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," in *Advances in Neural Information Processing Systems*, 2021.

[12] Pan Du, Suyun Zhao, Puhui Tan, Zisen Sheng, Zeyu Gan, Hong Chen, and Cuiping Li, "Towards a theoretical understanding of semi-supervised learning under class distribution mismatch," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 6, pp. 4853–4868, 2025.

[13] Nanqiao You, Xiaoqiao Zhao, Zhihai Gao, and Xuyan Song, "Explainable label distribution learning by exploiting neighborhood," *International Journal of Machine Learning and Cybernetics*, vol. 16, no. 1, pp. 173–188, 2025.

[14] Xiuyi Jia, Tian Qin, Yunan Lu, and Weiwei Li, "Adaptive weighted ranking-oriented label distribution learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 5, pp. 6629–6643, 2023.

[15] Yuntao Shou, Xiangyong Cao, Huan Liu, and Deyu Meng, "Masked contrastive graph representation learning for age estimation," *Pattern Recognition*, vol. 157, pp. 110884, 2025.

[16] Yujin Kang and Yoon-Sik Cho, "Beyond single emotion: Multi-label approach to conversational emotion recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[17] Shuai Liu, Peng Gao, Yating Li, Weina Fu, and Weiping Ding, "Multi-modal fusion network with complementarity and importance for emotion recognition," *Information Sciences*, vol. 619, pp. 679–694, 2023.

[18] Jiaman Li, Karen Liu, and Jiajun Wu, "Ego-body pose estimation via ego-head pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[19] Li Zhang and Fan Wang, "Structured matrix completion for incomplete label distribution learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[20] Xingyu Zhao, Yuexuan An, Ning Xu, Jing Wang, and Xin Geng, "Imbalanced label distribution learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[21] Jing Wang and Xin Geng, "Learn the highest label and rest label description degrees," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2021.

[22] Xiang Li and Songcan Chen, "No regularization is needed: Efficient and effective incomplete label distribution learning," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024, pp. 4470–4478.

[23] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel, "Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring," in *International Conference on Learning Representations (ICLR)*, 2020.

[24] Duorui Xie, Lingyu Liang, Lianwen Jin, Jie Xu, and Mengru Li, "Scut-fbp: A benchmark dataset for facial beauty perception," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, 2015.

[25] Xin Geng and Peng Hou, "Pre-release prediction of crowd opinion on movies by label distribution learning," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2015.

[26] Shan Li and Weihong Deng, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *International Journal of Computer Vision*, vol. 127, no. 6, pp. 884–906, 2019.

[27] Tingting Ren, Xiuyi Jia, Weiwei Li, Lei Chen, and Zechao Li, "Label distribution learning with label-specific features," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2019.