# Learning from Noisy Label Proportions via Dynamic Noise-Aware Instance Selection

Peimin Ning, Jing Chai*, Gang Hu, Shunfang Wang
School of Information Science & Engineering, Yunnan University, Kunming, China

***Abstract*—Learning from Label Proportions (LLP) is a challenging weakly supervised learning problem that training instances are grouped as bags with only the instance proportions belonging to each class in each bag being accessible, leaving the individual instance labels unknown. Most existing LLP methods are designed under the assumption that the label proportions are accurately annotated, despite that this assumption is typically unsatisfied in real applications due to measurement or human errors. To address the gap between assumption and practice, we proposed a new LLP method by resorting to Dynamic Noise-Aware Instance Selection (LLP-DNAIS). Specifically, we design a module that fuses probabilistic consistency, feature consistency, predictive self-certainty, and proportion matching to calculate a confidence score for each instance, and then develop an adaptive thresholding strategy to dynamically evaluate the cleanliness of individual instances based on confidence scores, enabling differentiated training of clean and noisy instances. Finally, the bag-level and instance-level losses are integrated to jointly train the networks. Experimental results on noisy benchmark datasets demonstrate that LLP-DNAIS significantly outperforms existing baselines in classification accuracy.**

***Index Terms*—Weakly Supervised Learning, Learning from Label Proportions, Symmetric Noise**

## I. Introduction

The success of deep learning relies heavily on large-scale labeled data, however, acquiring accurate instance-level annotations is often impractical due to high costs and/or privacy constraints [1]–[3], which sparks a wave of research enthusiasm for Weakly Supervised Learning (WSL) [4]. Among the WSL community, Learning from Label Proportions (LLP) has emerged as a promising paradigm due to its unique mechanism [5]–[7]. In LLP, training instances are grouped into bags, and only the bag-level label proportions, i.e., the proportions of instances belonging to all classes in each training bag, rather than the individual instance-level labels, are accessible in the training phase, and the goal is to predict the labels of unseen testing instances as precise as possible. Thus, LLP leverages supervisory signals like "30% of images in a traffic image bag are cars" [8] to operate model design, suffering from the trouble of instance-level label ambiguities. LLP has been widely used in various fields, involving crop mapping and SAR image classification [1], [2] and social science (election behavior modeling [3]), addressing classification tasks with aggregate label information. Although some progress has been made, most existing LLP methods are built on the clean data assumption, i.e., the labels of instances can be accurately annotated (although unknown), and the bag-level label proportions are precise as well. However, in real scenarios, the instance-level errors are common due to reasons such as imprecision of sensors or mistake of annotators, and these errors can easily propagate to the bag level, thereby making existing methods suffering from significant performance degradation. To address this issue, we propose a robust LLP method through Dynamic Noise-Aware Instance Selection (LLP-DNAIS), which is resistant to labeling noise. LLP-DNAIS builds a multidimensional confidence module, consisting of probabilistic consistency, feature consistency, predictive self-certainty and proportion matching, to compute confidence scores for instances and by which evaluate the cleanliness of instances. An adaptive threshold is also developed to assign dynamic weights to instances, thereby distinguishing between clean and noisy instances. Finally, we jointly optimize the instance-level and bag-level losses to align model predictions with both instance reliability and bag proportional constraints. Experiments on noisy benchmarks show that LLP-DNAIS significantly outperforms existing LLP baselines.

The main contributions of this paper are listed as below.

- We point out that label noise is prevalent in real LLP data but existing LLP methods hardly considered this problem, leading to significant performance degradation.
- We propose a novel LLP method termed LLP-DNAIS that dynamically calculates instance confidence to distinguish between clean and noisy instances, lending itself to real LLP tasks with label noise.
- We conduct empirical experiments to verify the superiority of LLP-DNAIS in classification accuracy by comparing it with state-of-the-art LLP methods across different datasets and noisy levels.

## II. Related Work

The learnability of LLP was theoretically validated by the research in [9], which proved that instance-level classifiers can be effectively learned by using bag-level proportional information. Building on this, DLLP [10] proposed a bag-level loss that aligns the averaged predictions of the instances in a bag with the target label proportion, laying the foundation for subsequent methods. LLP-GAN [11] introduced a generator to create fake instances and leveraged adversarial learning to tackle LLP tasks. LLP-VAT [12] adopted consistency regularization to maintain prediction uniformity across different data augmentations. The researches in [13], [14] were inspired by self-training and introduced contrastive losses for auxiliary training. Despite leveraging self-supervised representation learning, these methods did not take the instance-level supervision into consideration, which might degrade

*Corresponding Author: jingchai@aliyun.com

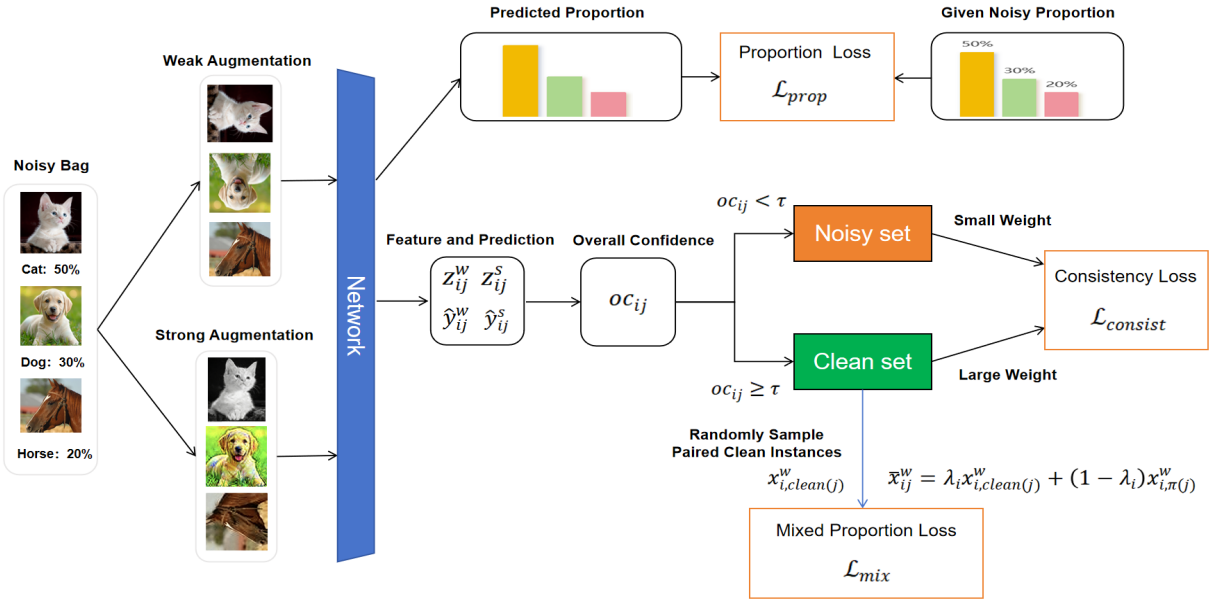



Fig. 1: Main structure of the proposed LLP-DNAIS method.

the ultimate learning performance. To enhance instance-level classification, several methods utilized model predictions to generate pseudo labels: SELF-LLP [15] aggregated the predictions from different epochs; ROT [16] and PLOT [17] resorted to the optimal transporting problem to generate pseudo labels; the work in [18] leveraged online pseudo-label generation with regret minimization; LLP-AHIL [20] used dual entropy-based weighting to measure the confidence of pseudo labels. Besides, MixBag [19] explored bag-level data augmentation to artificially increase the number of training bags and thus improved the final predictive accuracies. However, all these methods were built upon the clean data assumption, ignoring the existence of noises in many real-world scenarios.

## III. Method

### A. Formulation of LLP

Before formally discussing the proposed method, we first define the LLP problem by giving some useful notations. In brief, we consider a multi-classification issue with $C$ classes, where all training instances are organized into multiple disjoint bags. Thus, the training set $D$ can be expressed as $D = B_1 \cup B_2 \cup ... \cup B_N$, where $B_i$ denotes the $i$-th bag consisting of $M$ instances: $B_i = \{x_{i1}, x_{i2}, ..., x_{iM}\}$, where $x_{ij}$ represents the $j$-th instance in $B_i$. Given any bag $B_i$, its label proportion vector $p_i$ is also provided as $p_i = [p_{i1}, p_{i2}, \ldots, p_{iC}]^T$, where $p_{ik}$ denotes the proportion of instances belonging to the $k$-th class in $B_i$. Clearly, we have $\sum_{k=1}^{C} p_{ik} = 1$.

### B. Overview of LLP-DNAIS

The whole network consists of a feature extracting network $g_\theta(\cdot)$ parametered by $\theta$ and a classification network $f_\varphi(\cdot)$ parametered by $\varphi$. For any instance $x_{ij}$, both the weak and strong augmentations are adopted to produce a weakly augmented view $x_{ij}^w$ and a strongly augmented one $x_{ij}^s$. For $x_{ij}^w$, the extracted feature is $z_{ij}^w = g_\theta(x_{ij}^w)$, and the prediction is $\hat{y}_{ij}^w = f_\varphi(z_{ij}^w)$. Similarly, the extracted feature and prediction of $x_{ij}^s$ are $z_{ij}^s = g_\theta(x_{ij}^s)$ and $\hat{y}_{ij}^s = f_\varphi(z_{ij}^s)$, respectively. The design of LLP-DNAIS involves three key elements: instance confidence, which holistically evaluates the reliability of each instance from multiple perspectives; adaptive weighting, which distinguishes instances into clean and noisy ones via dynamic thresholds and assigns different weights to instances; and instance mixing, which generates mixed clean instances to fully utilize data for retraining. The overall structure of LLP-DNAIS is shown in Fig. 1.

### C. Proportion Loss

In LLP, only the bag-level label proportions are provided as supervisory signals, thus the proportion loss $\mathcal{L}_{prop}$, which computes the discrepancy between the given and predicted label proportions, is typically adopted as the bag-level loss:

$$\mathcal{L}_{prop} = \frac{1}{N}\sum_{i=1}^{N}\left[\mathrm{D}_{KL}(p_i^n \| \bar{\hat{y}}_i^w) + \lambda \| p_i^n - \bar{\hat{y}}_i^w \|_2\right] \quad (1)$$

where $\bar{\hat{y}}_i^w = \frac{1}{M}\sum_{j=1}^{M} \hat{y}_{ij}^w$ is obtained by aggregating and averaging the predictions of all instances within the bag $B_i$.

Here, the proportion loss is computed using Kullback-Leibler (KL) divergence with an auxiliary $l_2$ regularization term. The reason of adopting $l_2$ regularization is that the KL divergence compels the network output to approach $p_i$, which might mislead network training when $p_i$ is noisy, and the $l_2$ loss relaxes this compulsion due to its symmetry.

### D. Instance Confidence

According to the studies in noisy label learning [21], the networks tend to first fit reliable clean instances, and then the noisy ones. Thereby, clean instances are typically assigned with high confidences to indicate their high reliability, whereas the confidences of noisy instances are relatively low. Inspired by this, we adopt multiple strategies (probabilistic consistency,

feature consistency, predictive self-certainty and proportion matching) to conduct holistic confidence assessment.

*1) Probabilistic and Feature Consistency:* Given the weakly and strongly augmented views of $x_{ij}$, we compute the cosine similarity between their predictions and that between their extracted features, thus yield the probabilistic consistency $con_{ij}^{p}$ and feature consistency $con_{ij}^{f}$ respectively as follows:

$$con_{ij}^{p} = \frac{\cos\left(\hat{y}_{ij}^{w}, \hat{y}_{ij}^{s}\right) + 1}{2} \tag{2}$$

$$con_{ij}^{f} = \frac{\cos\left(z_{ij}^{w}, z_{ij}^{s}\right) + 1}{2} \tag{3}$$

The overall consistency $con_{ij}$ is the weighted average of the above two consistencies:

$$con_{ij} = w_p con_{ij}^{p} + (1-w_p) con_{ij}^{f} \tag{4}$$

where $w_p$ is a weighting parameter controlling the ratio of probabilistic consistency.

*2) Predictive Self-Certainty:* The entropy of the prediction itself reflects the uncertainty of the prediction, thus we may first normalize the entropy of the prediction, and then use one minus the normalized entropy to indicate the self-certainty of the prediction, which can be expressed as:

$$cer_{ij} = 1 - \frac{entropy_{ij}}{log\left(C\right)} = 1 + \frac{\sum_{k=1}^{C} \hat{y}_{ij,k}^{w} log\left(\hat{y}_{ij,k}^{w}\right)}{log\left(C\right)} \tag{5}$$

where the denominator $log(C)$ is the maximum self-entropy of a $C$-dimensional probability vector, $\hat{y}_{ij,k}^{w}$ denotes the $k$-th element of the vector $\hat{y}_{ij}^{w}$. Clearly, $cer_{ij}$ lies in the range [0,1].

*3) Proportion Matching:* In addition, We also compute the bag-level proportion matching degree for the $i$-th bag to evaluate the alignment between predicted and true class proportions, ensuring that higher values (closer to 1) indicate smaller differences (better alignment).

$$\text{mat}_i = 1 - \|p_i - \hat{y}_i^{w}\|_2 \tag{6}$$

where $\|\cdot\|_2$ denotes the $l_2$ norm.

*4) Overall Instance Confidence:* The overall confidence $oc_{ij}$ for the instance $x_{ij}$ is calculated by integrating the above probabilistic and feature consistencies, predictive self-certainty and proportion matching:

$$oc_{ij} = \alpha con_{ij} + \beta cer_{ij} + (1-\alpha-\beta) mat_i \tag{7}$$

where $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$ are the weighting parameters. Note that since $con_{ij}$, $cer_{ij}$, and $mat_i$ all lie in [0,1], the overall confidence $oc_{ij}$ is also between 0 and 1.

### *E. Adaptive weight strategy*

Based on the overall confidence, we separate training instances into a clean set $N_{clean}$ and a noisy set $N_{noisy}$. via dynamic thresholds, and then assign these instances with adaptive weights. The threshold $\tau$ lies in $[0.35, 0.7]$ and dynamically grows w.r.t. the increase of epochs ($\tau_0 = 0.35, \tau_1 = 0.45, \tau_2 = 0.6, \tau_{\text{max}} = 0.7; T_1 = 5, T_2 = 30, T_3 = 100$):

$$\tau(T) = \begin{cases} \tau_0 + \dfrac{\tau_1 - \tau_0}{T_1 - 0} \cdot T, & 0 \leq T < T_1, \\ \tau_1 + \dfrac{\tau_2 - \tau_1}{T_2 - T_1} \cdot (T - T_1), & T_1 \leq T < T_2, \\ \tau_2 + \dfrac{\tau_{\text{max}} - \tau_2}{T_3 - T_2} \cdot (T - T_2), & T_2 \leq T < T_3, \\ \tau_{\text{max}}, & T \geq T_3. \end{cases} \tag{8}$$

The set separation and the calculation of adaptive weight $w_{ij}$ of $x_{ij}$ are given as:

$$\begin{cases} x_{ij} \in N_{clean}, w_{ij} = 1, & if\ oc_{ij} \geq \tau \\ x_{ij} \in N_{noisy}, w_{ij} = \frac{oc_{ij}}{10\tau}, & if\ oc_{ij} < \tau \end{cases} \tag{9}$$

According to Eq.(9), clean instances are weighted more heavily than noisy ones, where the factor of 10 in the denominator for noisy instance weights is an empirical value determined by preliminary experiments to scale their weights to a reasonable low range [0,1], thus mitigating the interference of noisy instances on model training.

### *F. Mixed Proportion Loss*

We further mix high-confident clean instances to enhance data's utilization rate, and compute the corresponding mixed label proportion loss. For the $j$-th clean instance $x_{i,clean(j)}$ in $B_i$ (here $clean(j)$ denotes the original index of this clean instance in $B_i$), we first generate the mixing weight $\lambda_i$ from a Beta distribution with shape parameters $\alpha = 2$ and $\beta = 2$ (i.e., Beta(2,2)), then randomly select another clean instance $x_{i,\pi(j)}$ from $B_i$ (here $\pi(j)$ denotes the original index of this randomly selected clean instance in $B_i$). Mixed instances are then generated via pixel-/feature-wise weighted fusion according to the mixing weights, which is formulated as:

$$\bar{x}_{ij}^{w} = \lambda_i x_{i,clean(j)}^{w} + (1-\lambda_i) x_{i,\pi(j)}^{w} \tag{10}$$

By feeding all clean and mixed clean instances in $B_i$ into the network, averaging their predictions, we obtain:

$$\hat{y}_{i,clean}^{w} = \frac{1}{|N_{clean}^{i}|} \sum_{j=1}^{|N_{clean}^{i}|} f_{\varphi}\left(g_{\theta}\left(x_{i,clean(j)}^{w}\right)\right) \tag{11}$$

$$\hat{y}_{i,mix}^{w} = \frac{1}{|N_{clean}^{i}|} \sum_{j=1}^{|N_{clean}^{i}|} f_{\varphi}\left(g_{\theta}\left(\bar{x}_{ij}^{w}\right)\right) \tag{12}$$

where $N_{clean}^{i}$ denotes the clean instance set selected from $B_i$, $\hat{y}_{i,clean}^{w}$ and $\hat{y}_{i,mix}^{w}$ denote the predicted clean label proportion and mixed clean label proportion w.r.t. $B_i$, respectively. By replacing $p_i$ and $\hat{y}_i^{w}$ in Eq.(1) with $\hat{y}_{i,clean}^{w}$ and $\hat{y}_{i,mix}^{w}$, respectively, we obtain the following mixed proportion loss:

$$\mathcal{L}_{mix} = \mathcal{L}_{prop}\left(\hat{y}_{i,clean}^{w}, \hat{y}_{i,mix}^{w}\right) \tag{13}$$

TABLE I

CLASSIFICATION ACCURACIES OF DIFFERENT METHODS WITH DIFFERENT SYMMETRIC NOISE LEVELS.

| Dataset | Method | Noise Level | | | | |
|---|---|---|---|---|---|---|
| | | 20% | 30% | 40% | 50% | 60% |
| CIFAR-10 | DLLP | 74.70 | 62.20 | 53.90 | 44.49 | 32.80 |
| | LLP-VAT | 69.51 | 57.23 | 54.34 | 47.25 | 40.24 |
| | ROT | 72.17 | 62.92 | 58.62 | 49.90 | 31.31 |
| | $L^2$P-AHIL | 82.56 | 72.69 | 63.93 | 43.93 | 35.43 |
| | LLP-DNAIS | **86.19** | **80.85** | **75.88** | **70.08** | **59.19** |
| CIFAR-100 | DLLP | 53.70 | 44.48 | 37.14 | 24.32 | 15.30 |
| | LLP-VAT | 51.44 | 35.07 | 33.03 | 19.78 | 13.44 |
| | ROT | 47.56 | 39.18 | 28.38 | 16.11 | 9.59 |
| | $L^2$P-AHIL | 58.55 | 43.57 | 37.78 | 21.24 | 13.74 |
| | LLP-DNAIS | **60.47** | **49.63** | **42.49** | **27.91** | **18.61** |
| KMNIST | DLLP | 90.62 | 87.22 | 81.62 | 74.62 | 53.22 |
| | LLP-VAT | 90.22 | 85.40 | 79.29 | 68.81 | 63.20 |
| | ROT | 92.11 | 88.44 | 82.39 | 76.43 | 63.02 |
| | $L^2$P-AHIL | **97.66** | 93.81 | 84.24 | 70.28 | 55.48 |
| | LLP-DNAIS | 96.77 | **95.58** | **95.24** | **92.77** | **89.92** |
| Fashion-MNIST | DLLP | 85.04 | 83.16 | 80.29 | 76.99 | 73.22 |
| | LLP-VAT | 86.29 | 84.11 | 82.12 | 81.13 | 75.18 |
| | ROT | 89.09 | 86.57 | 82.12 | 77.72 | 73.23 |
| | $L^2$P-AHIL | 90.41 | 86.62 | 81.56 | 75.07 | 60.79 |
| | LLP-DNAIS | **91.66** | **90.95** | **88.44** | **88.03** | **87.16** |

TABLE II

ABLATION STUDY RESULTS.

| Noise Level | Method | Dataset | |
|---|---|---|---|
| | | CIFAR-10 | CIFAR-100 |
| 20% | LLP-DNAIS | **86.19** | **60.47** |
| | Variant 1 | 73.74 | 53.05 |
| | Variant 2 | 85.74 | 58.77 |
| | Variant 3 | 84.49 | 54.78 |
| | Variant 4 | 84.21 | 58.62 |
| 30% | LLP-DNAIS | **80.85** | **49.63** |
| | Variant 1 | 67.86 | 46.95 |
| | Variant 2 | 79.34 | 48.50 |
| | Variant 3 | 79.56 | 48.05 |
| | Variant 4 | 79.57 | 48.96 |
| 40% | LLP-DNAIS | **75.88** | **42.49** |
| | Variant 1 | 56.84 | 37.73 |
| | Variant 2 | 72.36 | 39.23 |
| | Variant 3 | 73.98 | 40.08 |
| | Variant 4 | 74.96 | 42.46 |
| 50% | LLP-DNAIS | **70.08** | **27.91** |
| | Variant 1 | 53.13 | 22.41 |
| | Variant 2 | 69.35 | 27.10 |
| | Variant 3 | 68.31 | 23.31 |
| | Variant 4 | 68.81 | 27.90 |
| 60% | LLP-DNAIS | **59.19** | **18.61** |
| | Variant 1 | 38.65 | 11.56 |
| | Variant 2 | 21.09 | 16.61 |
| | Variant 3 | 57.38 | 15.77 |
| | Variant 4 | 28.40 | 12.84 |

### G. Instance-Level Consistency Loss

We compute the weighted instance-level cross-entropy between the weakly and strongly augmented views of the same instance to measure the consistency of prediction, with the adaptive weight $w_{ij}$ in Eq.(8) being adopted:

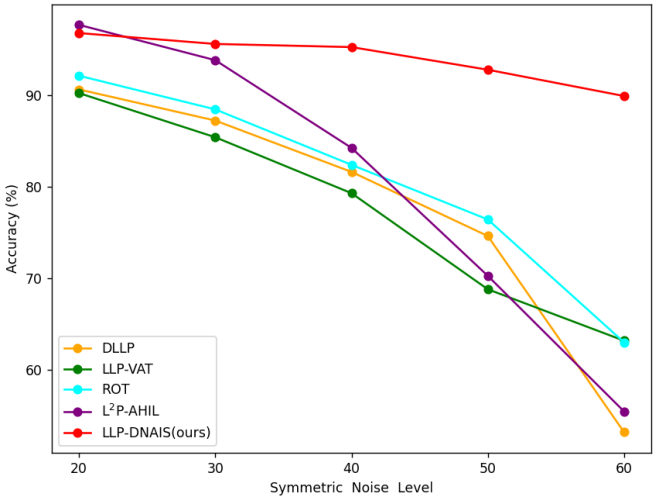



Fig. 2: Results on KMNIST for different methods.

$$\mathcal{L}_{ce} = \frac{-w_{ij}}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{C} \hat{y}_{ij,k}^{w} log\left(\hat{y}_{ij,k}^{s}\right) \tag{14}$$

We further enforce the consistency by adding a $l_1$ loss, thus the instance-level consistency loss can be expressed as:

$$\mathcal{L}_{consist} = \mathcal{L}_{ce} + \frac{\lambda_c}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \parallel \hat{y}_{ij}^{w} - \hat{y}_{ij}^{s} \parallel_1 \tag{15}$$

### H. Total Loss

The total loss is obtained by summarizing the above losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{prop}} \mathcal{L}_{\text{prop}} + \mathcal{L}_{\text{mix}} + \mathcal{L}_{\text{consist}} \tag{16}$$

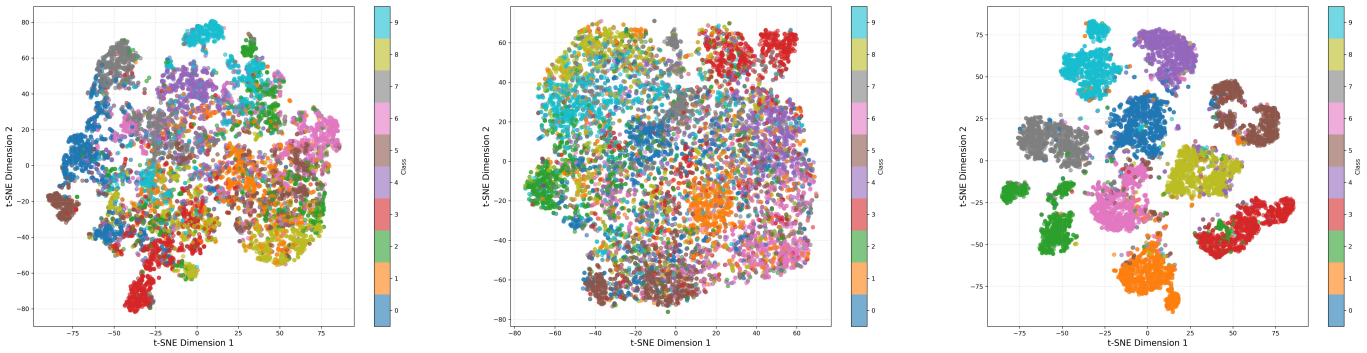
## IV. EXPERIMENTS

### A. Datasets and baselines

**Datasets**. Four benchmark datasets are employed: Fashion-MNIST [22], KMNIST [23], CIFAR-10 and CIFAR-100 [24].

**Baselines**. A) DLLP [10]: using aggregated model outputs to estimate label proportions; B) LLP-VAT [12]: involving consistency regularization to ensure different augmentations of the same instance yield similar predictions; C) ROT [16]: generating pseudo-labels via optimal transport to design instance-level loss besides bag-level loss; D) $L^2$P-AHIL [20]: forming an auxiliary instance-level classification loss by identifying and leveraging high-confident instances.

We compare LLP-DNAIS with the four state-of-the-art LLP methods above, all of which rely on the clean data assumption (precisely annotated bag-level proportions with no noise) and lack noise-aware modules for label proportion noise. For a fair comparison, we unify key training hyperparameters (e.g., initial learning rate, epochs, batch/bag size, optimizer settings) across all baselines and our method, while keeping their original optimal method-specific hyperparameters (e.g., loss weights, regularization coefficients) unchanged.




(a) DLLP (b) L$^2$P-AHIL (c) LLP-DNAIS(ours)

Fig. 3: T-SNE visualization of image representations on KMNIST with 60% symmetric noise.

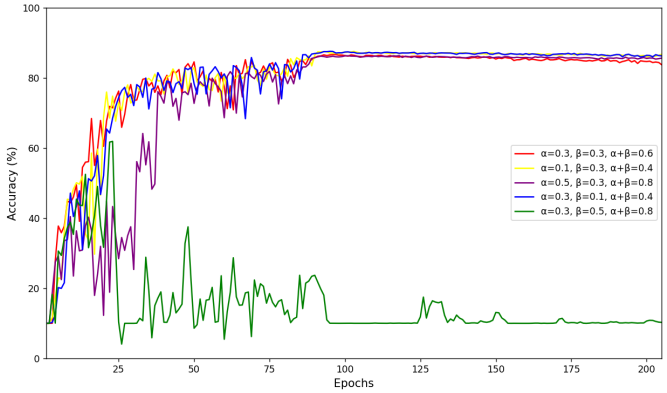



Fig. 4: Parameter sensitivity analysis of LLP-DNAIS on Fashion-MNIST with 60% symmetric noise.

### B. Implementation details

**Bag Generation**. First, we inject instance-level symmetric label noise into training instances at five levels (20%, 30%, 40%, 50%, 60%): each instance's ground-truth label is flipped to a random non-true class with equal probability at the preset noise rate, with a fixed random seed ensuring reproducibility. The full dataset is then shuffled and partitioned into fixed-size non-overlapping bags (each with $M$ instances), where noisy bag-level label proportions are empirically calculated from the noisy instance labels and thus deviate from clean proportions based on unaltered labels. Notably, these bags are randomly constructed before training and fixed across epochs for stable supervision and experimental reproducibility.

**Training**. The WRN-28-2 [25] network is used as the backbone. The models are trained using Stochastic Gradient Descent (SGD) [26], with a momentum of 0.9 and a weight decay of 5e-4. The initial learning rate is set to $\eta_0$ = 0.02, following a cosine decay schedule [27] . We fix the batch size as 32, the bag size $M = 16$, and the proportion loss weight $\lambda_{prop} = 8.0$. When calculating the instance confidence, we set $\alpha = 0.3, \beta = 0.3, (1 - \alpha - \beta) = 0.4$.

### C. Results and Analysis

We present experimental results in Table I, where the reported results are the averages of three independent runs. It is observed that LLP-DNAIS significantly outperforms the baselines across most benchmarks and noise levels. Compared with L$^2$P-AHIL which adaptively evaluates pseudo-label confidence via dual entropy weights, our method is only slightly inferior to it on KMNIST under 20% noise level. In all other cases, our method consistently outperforms all baselines.

LLP-DNAIS exhibits much slighter performance decline as noise level increases. E.g., on KMNIST (Fig. 2) its accuracy drops by merely 6.85% when noise rises from 20% to 60%, while other methods suffer the declines exceeding 25%.

LLP-DNAIS learns more distinguishable representations. Fig. 3 visualizes t-SNE [28] embeddings on KMNIST with 60% symmetric noise, with two baselines included for comparison. While baselines fail to learn discriminative representations in noisy settings, LLP-DNAIS yields well-separated clusters, demonstrating its superiority in high-quality representation learning under noisy label-proportion scenarios.

### D. Ablation study

We conduct ablation studies on CIFAR-10 and CIFAR-100 with various noise levels, by comparing LLP-DNAIS with four variants: (1) fixing the overall instance confidence at 0.5; (2) treating all instances as clean, with a fixed threshold and no clean/noisy distinction; (3) removing the mixed proportion loss; and (4) removing the $l_2$ loss from the proportion loss. As shown in Table II, LLP-DNAIS consistently outperforms all variants, which justifies the effectiveness of these modules in promoting the generalized learning performance.

### E. Parameter Sensitivity

We conduct sensitivity analysis of the weighting parameters $\alpha$ and $\beta$ (the 3rd parameter is $1 - \alpha - \beta$) when calculating instance confidence, evaluated on Fashion-MNIST with 60% symmetric noise, where $\alpha$ and $\beta$ are tested over 0.1, 0.3, 0.5. As shown in Fig. 4, classification accuracy keeps relatively stable when $\alpha$ and $\beta$ vary with smaller $\alpha+\beta$ values, whereas the stability degrades significantly for larger $\alpha + \beta$ values. It even leads to performance collapse when $\alpha = 0.3$ and $\beta = 0.5$, attributed to relatively high $\beta$ and excessively low $1 - \alpha - \beta$. The candidate reason is that in early training stage, the certainty of instance-level prediction is low, and assigning heavy weights (high $\beta$) on predictive self-certainty scores will reduce the weights of proportion-matching scores (if $\alpha$ is set to medium values), thereby significantly misleading the subsequent model training.

## V. Conclusion

This paper designs a noise-aware LLP method under symmetric noisy scenarios, addressing the core challenges of

proportion deviation propagated from unknown yet corrupted instance-level labels. The design consists of three cores: multi-dimensional instance-confidence evaluation, which includes synthesizing probabilistic and feature consistency, predictive self-certainty and proportion matching; dynamic threshold-based adaptive weight assignment; and a hybrid loss function that integrates the bag-level proportion loss and mixed proportion loss, and the instance-level consistency loss. Experimental results on multiple benchmarks validate the effectiveness of our method, demonstrating state-of-the-art performance across various symmetric noise settings. In future work, we would like to extend the method to asymmetric noise scenarios.

## ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. 62166046).

## REFERENCES

[1] L. E. Cué La Rosa, D. A. B. Oliveira, and P. Ghamisi, "Learning crop-type mapping from regional label proportions in large-scale SAR and optical imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[2] Y. Ding, Y. Li, and W. Yu, "Learning from label proportions for SAR image classification," *EURASIP Journal on Advances in Signal Processing*, vol. 41, pp. 1–12, 2017.

[3] T. Sun, D. Sheldon, and B. O'Connor, "A probabilistic approach for learning with label proportions applied to the US presidential election," in *Proceedings of the IEEE International Conference on Data Mining*, pp. 445–454, 2017.

[4] M. Sugiyama, H. Bao, T. Ishida, N. Lu, T. Sakai, and G. Niu, *Machine Learning from Weak Supervision: An Empirical Risk Minimization Approach*, MIT Press, Cambridge, Massachusetts, USA, 2022.

[5] A. Brahmbhatt, R. Saket, and A. Raghuveer, "PAC learning linear thresholds from label proportions," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[6] S. Havaldar, N. Sharma, S. Sareen, K. Shanmugam, and A. Raghuveer, "Learning from label proportions: Bootstrapping supervised learners via belief propagation," in *NeurIPS Workshop on Regulatable Machine Learning*, 2023.

[7] J. Zhang, Y. Wang, and C. Scott, "Learning from label proportions by learning with label noise," *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[8] T. Liebig, M. Stolpe, and K. Morik, "Distributed traffic flow prediction with label proportions: from in-network towards high performance computation with MPI," in *Proceedings of the International Conference on Mining Urban Data*, pp. 36–43, 2015.

[9] F. X. Yu, K. Choromanski, S. Kumar, T. Jebara, and S.-F. Chang, "On learning from label proportions," arXiv preprint arXiv:1402.5701, 2014.

[10] E. M. Ardehaly and A. Culotta, "Co-Training for Demographic Classification Using Deep Learning from Label Proportions," in *Proceedings of the IEEE International Conference on Data Mining Workshops*, pp. 1017–1024, 2017.

[11] J. Liu, B. Wang, H. Hang, H. Wang, Z. Qi, Y. Tian, and Y. Shi, "LLP-GAN: A GAN-based algorithm for learning from label proportions," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8377–8388, 2023.

[12] K. H. Tsai and H. T. Lin, "Learning from label proportions with consistency regularization," in *Proceedings of the Asian Conference on Machine Learning*, vol.129, pp. 513–528, 2020.

[13] J. Nandy, R. Saket, P. Jain, J. Chauhan, B. Ravindran, and A. Raghuveer, "Domain agnostic contrastive representations for learning from label proportions," in *Proceedings of the ACM International Conference on Information and Knowledge Management*, pp. 1542–1551, 2022.

[14] H. Yang, W. Zhang, and W. Lam, "A two-stage training framework with feature-label matching mechanism for learning from label proportions," in *Proceedings of the Asian Conference on Machine Learning*, vol.157, pp. 1461–1476, 2021.

[15] J. Liu, Z. Qi, B. Wang, Y. Tian, and Y. Shi, "SELF-LLP: Self-supervised learning from label proportions with self-ensemble," *Pattern Recognition*, vol. 129, p. 108767, 2022.

[16] G. Dulac-Arnold, N. Zeghidour, M. Cuturi, L. Beyer, and J.-P. Vert, "Deep multi-class learning from label proportions," in *Proceedings of the International Conference on Machine Learning*, vol. 97, pp. 1758–1767, 2019.

[17] J. Liu, B. Wang, X. Shen, Z. Qi, and Y. Tian, "Two-stage training for learning from label proportions," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 2737–2743, 2021.

[18] T. Asanomi, S. Matsuo, D. Suehiro, and R. Bise, "MixBag: Bag-level data augmentation for learning from label proportions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16570–16579, 2023.

[19] S. Matsuo, R. Bise, S. Uchida, and D. Suehiro, "Learning from label proportions with online pseudo label decision by regret minimization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023.

[20] T. Ma, H. Chen, J. Hu, Y. Zhu, and X. Li, "Forming auxiliary high-confident instance-level loss to promote learning from label proportions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20592–20601, 2025.

[21] S. Ren, J. Dai, Y. Zhou, C. Dan, and J. Z. Kolter, "Learning from noisy labels via conditional distributionally robust optimization," in *Proceedings of the International Conference on Machine Learning*, vol. 202, pp. 28657–28676, 2023.

[22] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms," arXiv preprint arXiv:1708.07747, 2017.

[23] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, "Deep learning for classical Japanese literature," arXiv preprint arXiv:1812.01718, 2018.

[24] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Technical Report TR-2009-0, University of Toronto, Toronto, Ontario, Canada, 2009.

[25] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference*, pp. 87.1–87.12, 2016.

[26] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[27] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with warm restarts," in *Proceedings of the International Conference on Learning Representations*, 2017.

[28] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.