This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Forming Auxiliary High-confident Instance-level Loss to Promote Learning from Label Proportions

Tianhao Ma[1,2,*], Han Chen[1,2,*], Juncheng Hu[1,2], Yungang Zhu[1,2], Ximing Li[1,2,†]
[1]College of Computer Science and Technology, Jilin University, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
{matianhao2120, 2002chenhan}@gmail.com, {jchu, zhuyungang}@jlu.edu.cn, liximing86@gmail.com

Abstract

Learning from label proportions (LLP), i.e.*, a challenging weakly-supervised learning task, aims to train a classifier by using bags of instances and the proportions of classes within bags, rather than annotated labels for each instance. Beyond the traditional bag-level loss, the mainstream methodology of LLP is to incorporate an auxiliary instance-level loss with pseudo-labels formed by predictions. Unfortunately, we empirically observed that the pseudo-labels are are often inaccurate due to over-smoothing, especially for the scenarios with large bag sizes, hurting the classifier induction. To alleviate this problem, we suggest a novel LLP method, namely Learning from Label Proportions with Auxiliary High-confident Instance-level Loss (L^{2}P-AHIL). Specifically, we propose a dual entropy-based weight (DEW) method to adaptively measure the confidences of pseudo-labels. It simultaneously emphasizes accurate predictions at the bag level and avoids overly smoothed predictions. We then form high-confident instance-level loss with DEW, and jointly optimize it with the bag-level loss in a self-training manner. The experimental results on benchmark datasets show that L^{2}P-AHIL can surpass the existing baseline methods, and the performance gain can be more significant as the bag size increases. The implementation of our method is available at* https://github.com/TianhaoMa5/LLP-AHIL.

1. Introduction

During the past decades, supervised learning has been proven to gain significant success in diverse real-world applications, in which one can easily collect loads of training instances with precise supervision [21, 22]. Unfortunately, in many scenarios, it is challenging to collect a sufficient number of precisely labeled training instances due to various difficulties such as high annotation costs [34, 36] and privacy protection concerns [12]. Consequently, training instances with incomplete, inexact, and inaccurate supervision are often available only [50]. The demand for learning with such weak supervision facilitates the development of **W**eakly-**S**upervised **L**earning (**WSL**) [40], ranging many specific paradigms, *e.g.*, semi-supervised learning [18, 49], partial label learning [20, 27], and positive-unlabeled learning [10, 13], to name just a few.

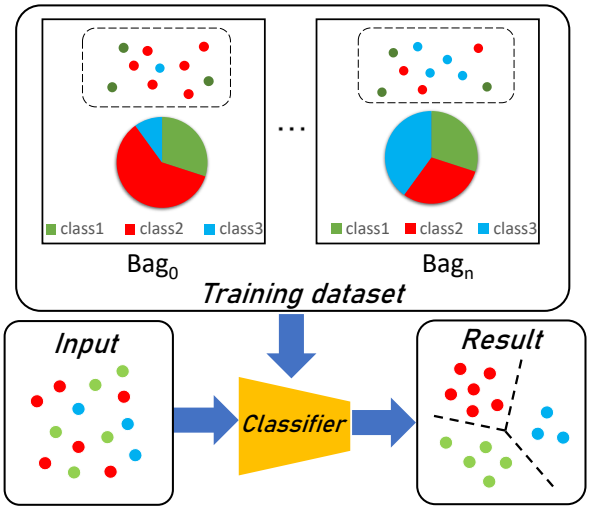


Figure 1. Before training, the data is grouped and only the proportion information is retained. Classifier training is performed using unlabeled data and proportions in bag units, and finally the instance classification task is completed.

Recently, an emerging paradigm of WSL, known as **L**earning from **L**abel **P**roportions (**LLP**) as shown in Fig. 1, has gained increasing attention within the machine learning community [2, 4, 11, 48], and it has been broadly applied across various computer vision domains such as re-

*Equal contribution.
†Corresponding author.

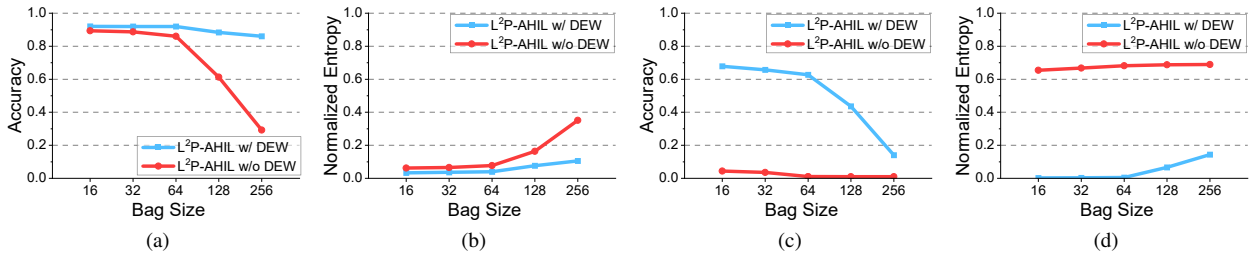


Figure 2. Results of preliminary experiments with different bag sizes on CIFAR-10 and CIFAR-100 (500 epochs). (a)(c) Averaged accuracy of pseudo-labels on CIFAR-10 and CIFAR-100. (b)(d) Averaged normalized entropy of pseudo-labels on CIFAR-10 and CIFAR-100. Higher normalized entropy values imply more smoothing results. Table 2 shows the adverse effects caused by these inaccurate pseudo-labels.

mote sensing image categorization [8, 19] and medical image processing [3]. Technically, LLP, as its name suggests, refers to the classification problem where a certain number of training instances are grouped into many bags with the corresponding proportion of classes, rather than annotated labels for each instance. Naturally, such sparse and inexact supervision from LLP causes significant difficulties in inducing effective classifiers for predicting labels of instances.

To handle the task of LLP, the basic method is to induce the classifier by optimizing the bag-level loss, measuring the difference between the ground-truth class proportion in a bag and the predicted class proportion formed by predictions of instances from the bag [46]. As suggested by [23, 24], optimizing the bag-level loss only can not guarantee inducing effective classifiers. Therefore, the mainstream methodology of LLP becomes to incorporate an auxiliary instance-level loss with pseudo-labels, and jointly optimize it with the bag-level loss in a self-training manner [9, 24, 25, 28]. However, in our preliminary experiments, we empirically observed that the pseudo-labels generated by this kind of methods are often inaccurate and even meaningless, especially for the scenarios with large bag sizes. As illustrated in Fig. 2, our early empirical results indicated that the pseudo-labels are inaccurate and tend to be overly smoothed, quantitatively measured by entropy values, as the bag size increases. Such low-quality pseudo-labels hurt the classifier induction, resulting in performance degradation.

To solve this problem, in this paper, we propose a novel LLP method, namely **L**earning from **L**abel **P**roportions with **A**uxiliary **H**igh-confident **I**nstance-level **L**oss (**L^{2}P-AHIL**). Specifically, we suggest a dual entropy-based weight (DEW) method to adaptively evaluate the confidence of pseudo-labels from both bag and instance levels. At the bag level, it emphasizes accurate class proportion predictions measured the difference between entropy values of ground-truth proportions and predicted proportions from the class-pivoted perspective. At the instance level, it avoids overly smoothed predictions of instances, directly measured by the entropy values. We then form a high-confident instance-level loss with DEW, and jointly optimize it with the bag-level loss in a self-training manner. The experimental results on benchmark datasets demonstrate that L^{2}P-AHIL can outperform the existing baseline methods, and the performance gain can be more significant as the bag size increases.

In a nutshell, the major contributions of this paper are outlined as follows:

- We quantitatively evaluated the problem of inaccurate and meaningless pseudo-labels of the existing LLP art, as the big size increases.
- To solve this problem, we propose a novel LLP method named L^{2}P-AHIL, which adaptively measures the confidences of pseudo-labels.
- We conduct extensive experiments to empirically validate the effectiveness of L^{2}P-AHIL even with larger bag sizes.

2. Related Work

Learning from label proportions. [46] theoretically demonstrated that an instance-level classifier can be trained using bag-level proportion information. Building on this theory, DLLP [1] proposed a bag-level loss that aligns the average of all instances' predictions within a bag with the target distribution. Many methods have been developed based on this loss. LLP-GAN [23] introduced a generator to produce fake instances, increasing the class count to $K+1$, enabling the model to classify these fake instances as the $(K+1)^{\text{th}}$ class. LLP-VAT [41] uses a consistency regularization loss to maintain uniform predictions for an instance across various data augmentations. [29, 45] introduced self-training with a contrastive learning loss to aid training. These methods incorporate self-supervised representation learning, but without instance-level supervised training, it is difficult to train a well-performing instance-level classifier.

To improve instance-level classification performance,

some approaches employ the model's predictions as pseudo-labels for self-training. SELF-LLP [25] employs the model's predictions from different epochs as pseudo-labels for training. ROT and PLOT [9, 24] utilize pseudo-labels generated through optimal transport for training. In the study by [28], they employed an online pseudo-labeling method with regret minimization. However, due to the significant smoothing of model predictions, especially with larger bag sizes, it naturally leads to the generation of numerous incorrect pseudo-labels. In our approach, we utilize pseudo-labels for self-training and adaptively adjust the weight of each instance's pseudo-label based on both bag-level and instance-level entropy values. This prevents unreliable pseudo-labels from compromising the model's performance.

From another perspective, the works of [5, 15, 31, 33, 37, 38, 48] are grounded in statistical learning theory. However, these approaches have significant limitations. For instance, the method proposed in [37, 38] heavily relies on the assumption that an example can belong to multiple bags. On the other hand, [5] introduced an unbiased estimator, but their method suffered from severe overfitting due to negative empirical loss.

Adaptive weighting techniques in weakly-supervised learning. In the domain of semi-supervised learning, many methods utilize an adaptive weighting approach for pseudo-labels. [17] adaptively adjust the weight of each class by estimating the learning difficulty of each class. Influence [35] introduces the concept of adjusting weights using an algorithm based on the influence function. SoftMatch [6] employs a truncated Gaussian function to weight samples according to their confidence levels. [14] adaptively adjust the weight of each instance based on the entropy of the pseudo-labels for each instance. However, many adaptive weighting strategies are not specifically tailored for LLP. Directly applying these methods to LLP may not yield optimal performance, as demonstrated in Sec. 4. In our method, we measure the confidences of pseudo-labels based on the DEW method, adaptively adjusting the weight of each instance accordingly.

3. Proposed L^{2}P-AHIL Method

In this section, we introduce the proposed LLP method named L^{2}P-AHIL.

Task formulation of LLP. We now introduce essential notations to formulate the LLP task. Typically, we are given an LLP training dataset of N proportion-labeled bags $\mathcal{D} = \{(\mathbf{B}_i, \mathbf{p}_i)\}_{i=1}^{N}$, where each $\mathbf{B}_i = \{\mathbf{x}_{i,j}\}_{j=1}^{M}$ denotes a bag of M instances represented by $\mathbf{x}$. Its corresponding proportion label $\mathbf{p}_i \in \triangle^{C-1}$ denotes the proportion of C pre-defined classes in the bag. The value of $m_i^c = p_{i,c}M$ represents the number of instances of class c in $\mathbf{B}_i$. Generally, all bags are disjoint, *i.e.*, $\mathbf{B}_i \cap \mathbf{B}_k = \varnothing, \forall i \neq k$. The goal of LLP is to induce a classifier over $\mathcal{D}$ with weakly supervised signals, *i.e.*, the bag-level label proportion only.

3.1. Overview

Overall speaking, our L^{2}P-AHIL method is denoted as $f_{\mathbf{W}}(\cdot)$, parameterized by $\mathbf{W}$, which ingests an instance $\mathbf{x}$ and outputs its prediction $\hat{\mathbf{y}} = f_{\mathbf{W}}(\mathbf{x})$. We follow the methodology of incorporating an auxiliary instance-level loss with pseudo-labels. For the instance-level loss, we adopt the commonly used "Weak and Strong Augmentation" strategy in the semi-supervised learning domain [39] to enhance the model's representational learning capabilities. For each instance $\mathbf{x}$, we transform it into a weakly augmented version $\mathbf{x}^{\mathrm{w}}$ and a strongly augmented version $\mathbf{x}^{\mathrm{s}}$, where $\mathbf{x}^{\mathrm{w}}$ is specifically used to form the corresponding pseudo-label and $\mathbf{x}^{\mathrm{s}}$ is used as a training sample. The predictions for them are represented by $\hat{\mathbf{y}}^{\mathrm{w}} = f_{\mathbf{W}}(\mathbf{x}^{\mathrm{w}})$ and $\hat{\mathbf{y}}^{\mathrm{s}} = f_{\mathbf{W}}(\mathbf{x}^{\mathrm{s}})$, respectively. With the predictions of all instances, we can formulate the objective of L^{2}P-AHIL with respect to $\mathbf{W}$ as follows:

$$\mathcal{L}(\mathbf{W}) = \mathcal{L}_{\mathrm{b}}(\mathbf{p}, \hat{\mathbf{y}}^{\mathrm{w}}) + \lambda \mathcal{L}_{\mathrm{i}}(\tilde{\mathbf{y}}, \hat{\mathbf{y}}^{\mathrm{s}}), \tag{1}$$

where $\mathcal{L}_{\mathrm{b}}(\mathbf{p}, \hat{\mathbf{y}}^{\mathrm{w}})$ is the bag-level proportion loss; $\mathcal{L}_{\mathrm{i}}(\tilde{\mathbf{y}}, \hat{\mathbf{y}}^{\mathrm{s}})$ is the instance-level self-training loss and $\tilde{\mathbf{y}}$ denotes the pseudo-label; λ is coefficient parameter used to balance the two losses. The overall pipeline of the proposed model is illustrated in Fig. 3a.

Bag-level proportion loss. Following [1], L^{2}P-AHIL continues to utilize this loss for bag-level proportion learning. It is defined as the cross-entropy between prior and posterior proportions which can be formulated as follows:

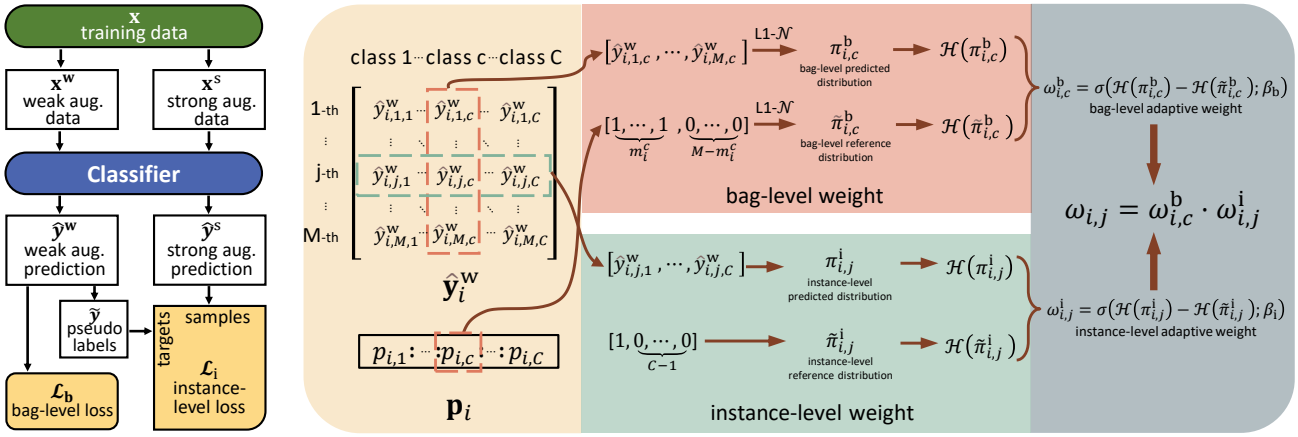
$$\mathcal{L}_{\mathrm{b}}(\mathbf{p}, \hat{\mathbf{y}}^{\mathrm{w}}) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{H}(\mathbf{p}_i, \bar{\mathbf{p}}_i), \tag{2}$$

where $\bar{\mathbf{p}}_i = [\bar{p}_{i,c}]_{c=1}^{C}$ denotes the predicted proportion in $\mathbf{B}_i$, and $\mathcal{H}(\cdot, \cdot)$ denotes the cross-entropy loss. The value of $\bar{p}_{i,c}$ is computed by averaging the predicted probabilities of class c across all instances in $\mathbf{B}_i$. It can be represented by the formula:

$$\bar{p}_{i,c} = \frac{1}{M} \sum_{j=1}^{M} \hat{y}_{i,j,c}^{\mathrm{w}}, \tag{3}$$

where $\hat{y}_{i,j,c}^{\mathrm{w}}$ represents the predicted probability of class c for $\mathbf{x}_{i,j}^{\mathrm{w}}$.

Instance-level self-training loss. We define the pseudo-labels $\tilde{\mathbf{y}}$ as the hardened version of $\hat{\mathbf{y}}^{\mathrm{w}}$. This means that for



(a) Model pipeline

(b) The process of DEW calculation

Figure 3. (a) The plot of the model's pipeline with the "Weak and Strong Augmentation" strategy (aug. denotes augmentation). When calculating the instance-level loss $\mathcal{L}_{\mathrm{i}}$, $\hat{\mathbf{y}}^{\mathrm{s}}$ is used as input samples and $\tilde{\mathbf{y}}$ as targets. Adaptive weight components are omitted for visual clarity. (b) The plot of the detail process of DEW. The process involves: 1) acquiring predicted and reference distributions at both bag and instance levels (red and green pathways, respectively), 2) calculating entropy for these distributions, and 3) determining adaptive weights through a mapping function that integrates information from both levels. L1-$\mathcal{N}$ indicates L1-Normalization operations.

each $\mathbf{x}_{i,j}$, its pseudo label $\tilde{\mathbf{y}}_{i,j} = [\tilde{y}_{i,j,c}]_{c=1}^{C}$ has one class with a probability of 1, while the probabilities for all other classes are set to 0. Formally, this can be expressed as:

$$\tilde{y}_{i,j,c} = \begin{cases} 1, & \text{if } c = \arg\max(\hat{\mathbf{y}}^{\mathrm{w}}_{i,j}), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Using the strongly augmented predictions $\hat{\mathbf{y}}^{\mathrm{s}}$ as training samples and the pseudo labels $\tilde{\mathbf{y}}$ as training targets, we employ cross entropy as the loss function to calculate the instance-level loss:

$$\mathcal{L}_{\mathrm{i}}(\tilde{\mathbf{y}}, \hat{\mathbf{y}}^{\mathrm{s}}) = \frac{1}{N \cdot M} \sum_{i=1}^{N} \sum_{j=1}^{M} \omega_{i,j} \mathcal{H}(\tilde{\mathbf{y}}_{i,j}, \hat{\mathbf{y}}^{\mathrm{s}}_{i,j}), \tag{5}$$

where $\omega_{i,j}$ denotes the adaptive weight for $\mathbf{x}_{i,j}$ according to its confidence. How to measure this confidence is the key to L^{2}P-AHIL, and we will describe the details in the following parts.

3.2. Dual Entropy-based Weight

Based on the assumption that lower entropy values imply high confidence in LLP, we propose a DEW weight for computing $\omega_{i,j}$. Specifically, we consider the confidence from both bag and instance levels. Accordingly, we design $\omega_{i,j}$ as a product of a bag-level adaptive weight ω^{b} and an instance-level adaptive weight ω^{i} as follows:

$$\omega_{i,j} = \omega^{\mathrm{b}}_{i,c} \cdot \omega^{\mathrm{i}}_{i,j}, \tag{6}$$

where $\omega_{i,j}$ represents the adaptive weight for $\mathbf{x}_{i,j}$ and here, $c = \arg\max \hat{\mathbf{y}}^{\mathrm{w}}_{i,j}$ denotes the class with the highest probability. We calculate $\omega^{\mathrm{b}}_{i,c}$ and $\omega^{\mathrm{i}}_{i,j}$ based on the entropy differences between the predicted distribution and the reference distribution at the bag and instance level, respectively. The detailed process of DEW is shown in Fig. 3b.

3.2.1 Bag-level Adaptive Weight

At the bag level, the predicted distribution for class c is constructed by aggregating the predictions from all instances within bag $\mathbf{B}_i$. Specifically, we first collect the unnormalized predictions $\left[\hat{y}^{\mathrm{w}}_{i,j,c}\right]_{j=1}^{M}$ into a vector denoted as $\pi'^{\mathrm{b}}_{i,c}$. To obtain a valid probability distribution (i.e., summing to 1) at the bag level, we apply L1 normalization to $\pi'^{\mathrm{b}}_{i,c}$, resulting in the bag-level predicted distribution $\pi^{\mathrm{b}}_{i,c} = \pi'^{\mathrm{b}}_{i,c} / \|\pi'^{\mathrm{b}}_{i,c}\|_1$.

Based on practical knowledge, we define bag-level reference distribution as $\tilde{\pi}^{\mathrm{b}}_{i,c}$, which is composed as follows: there are m^c_i instances predicted as class c with the highest probability, i.e., $1/m^c_i$, while the remaining $(M - m^c_i)$ instances have the lowest probability, i.e., 0. It can be expressed as follows:

$$\begin{aligned} \tilde{\pi}^{\mathrm{b}}_{i,c} &= \text{L1-Normalization}\Big(\big[\underbrace{1, \cdots, 1}_{m^c_i}, \underbrace{0, \cdots, 0}_{M - m^c_i}\big]\Big) \\ &= \Big[\underbrace{\frac{1}{m^c_i}, \cdots, \frac{1}{m^c_i}}_{m^c_i}, \underbrace{0, \cdots, 0}_{M - m^c_i}\Big]. \end{aligned} \tag{7}$$

The entropy of the bag-level reference distribution can then be calculated as $\mathcal{H}(\tilde{\pi}^{\mathrm{b}}_{i,c}) = \log m^c_i$. As the predicted dis-

tribution approaches the reference distribution, the adaptive weight tends toward 1; conversely, it tends toward 0 when the distributions diverge. Therefore, we introduce $\sigma(x;\beta) = \exp\left(-\frac{x^2}{\beta}\right)$ as a mapping function, which allows us to derive the bag-level adaptive weight:

$$\begin{aligned}\omega^{\mathrm{b}}_{i,c} &= \sigma\left(\mathcal{H}(\pi^{\mathrm{b}}_{i,c}) - \mathcal{H}(\tilde{\pi}^{\mathrm{b}}_{i,c}); \beta_{\mathrm{b}}\right) \\ &= \exp\left\{-\frac{\left[\mathcal{H}(\pi^{\mathrm{b}}_{i,c}) - \mathcal{H}(\tilde{\pi}^{\mathrm{b}}_{i,c})\right]^2}{\beta_{\mathrm{b}}}\right\} \\ &= \exp\left\{-\frac{\left[\mathcal{H}(\pi^{\mathrm{b}}_{i,c}) - \log m^c_i\right]^2}{\beta_{\mathrm{b}}}\right\},\end{aligned} \tag{8}$$

where $\beta_{\mathrm{b}} > 0$ is a hyper parameter that controls the degree of smoothness of the mapping function.

3.2.2 Instance-level Adaptive Weight

At the instance level, the predicted distribution is derived from the prediction results of individual instances. Specifically, for each instance $\mathbf{x}_{i,j}$, we define instance-level predicted distribution $\pi^{\mathrm{i}}_{i,j} = \left[\hat{y}^{\mathrm{w}}_{i,j,c}\right]_{c=1}^{C}$ to measure the confidence of the instance.

For the instance-level reference distribution, we consider a label to have high confidence when its predicted probabilities are concentrated on a single class. Therefore, we define the instance-level reference distribution as $\tilde{\pi}^{\mathrm{i}}_{i,j}$, where only one class has a probability of 1, and the probabilities of all other classes are set to 0, *i.e.*,

$$\tilde{\pi}^{\mathrm{i}}_{i,j} = \big[1, \underbrace{0, \ldots, 0}_{C-1}\big]. \tag{9}$$

It is straightforward to calculate that the entropy of instance-level reference distribution is 0, *i.e.*, $\mathcal{H}(\tilde{\pi}^{\mathrm{i}}_{i,j}) = 0$. With the same mapping function as the bag level, the instance-level adaptive weight is obtained with a hyper parameter $\beta_{\mathrm{i}} > 0$:

$$\begin{aligned}\omega^{\mathrm{i}}_{i,j} &= \sigma\left(\mathcal{H}(\pi^{\mathrm{i}}_{i,j}) - \mathcal{H}(\tilde{\pi}^{\mathrm{i}}_{i,j}); \beta_{\mathrm{i}}\right) \\ &= \exp\left\{-\frac{\left[\mathcal{H}(\pi^{\mathrm{i}}_{i,j}) - \mathcal{H}(\tilde{\pi}^{\mathrm{i}}_{i,j})\right]^2}{\beta_{\mathrm{i}}}\right\} \\ &= \exp\left[-\frac{\mathcal{H}(\pi^{\mathrm{i}}_{i,j})^2}{\beta_{\mathrm{i}}}\right].\end{aligned} \tag{10}$$

4. Experiments

4.1. Experimental Setup

Datasets. We assessed L^{2}P-AHIL on five benchmark datasets for LLP: Fashion-MNIST [44], SVHN [30], CIFAR-10, CIFAR-100 [16] and Mini-ImagNet [43]. Fashion-MNIST is a dataset consisting of 60,000 training and 10,000 test examples. Each example is a 28x28 grayscale image, representing 10 categories of fashion items. The SVHN dataset is composed of 32x32 RGB images of digits, featuring 73,257 training examples and 26,032 testing examples, along with an additional 531,131 training samples that we do not utilize in our experiments. CIFAR-10 and CIFAR-100 datasets each contain 50,000 training and 10,000 testing examples, with every example being a 32x32 color natural image. They are categorized into 10 and 100 classes, respectively. MiniImageNet is a subset of the ImageNet dataset, consisting of 100 selected classes. Each class contains 80 images for training and 20 images for testing, with all images resized to 64x64.

Bag generation. For each dataset, with a specified bag size M, we randomly select M samples from the training set $\mathcal{D}$ to form a bag $\mathbf{B}_i$. The samples in distinct bags do not overlap. We compute the class proportion information within each bag to guide the training, but real labels are not used during training. Following prior studies, we choose M from the set $\{16, 32, 64, 128, 256\}$. As every dataset contains an equal number of samples across classes, the bags created through this method exhibit relatively balanced proportions, leading to more challenging LLP tasks, as evidenced by [46].

4.2. Implementation Details

For SVHN, Fashion-MNIST, and CIFAR-10, we used the WRN-28-2 [47] architecture as the encoder, and for CIFAR-100, the WRN-28-8. For MiniImageNet, we employed the ResNet-18 architecture. The classifier consists of a single linear layer. Each training step utilizes a batch size equal to the bag size multiplied by the number of bags, amounting to 1024. Models are trained using Stochastic Gradient Descent (SGD) [32], with a momentum of 0.9 and a weight decay of 5e-4 for WRN-28-2, 1e-3 for WRN-28-8, and 1e-4 for ResNet-18. The initial learning rate is set to $\eta_0 = 0.03$ for all datasets, except for MiniImageNet, where it is set to $\eta_0 = 0.05$, following a cosine learning rate decay schedule [26] as $\eta = \eta_0 \cos\left(\frac{7\pi k}{16K}\right)$, where k is the current training step and K is the total number of steps, set to 2^{20}. We train our models for 1024 epochs. Weak augmentations include random horizontal flips and random cropping. For SVHN, in accordance with [39], horizontal flips are replaced with random translations up to 12.5% in both directions. Strong augmentations utilize RandAugment [7]. The setting for fully supervised learning is the same as mentioned above.

Baselines. A) LLPFC [48]: This method offers a theoretically grounded approach to LLP by reducing it to learning with label noise, employing the forward correction (FC) loss. B) DLLP [1]: Traditional LLP methods primarily rely

Table 1. Classification accuracy (mean±std) on CIFAR-10, CIFAR-100, SVHN, Fashion-MNIST and MiniImageNet for different bag sizes. The highest accuracy is highlighted in bold. The symbol "*" indicates results for FLMm are reproduced directly from its original publication [45], while the dash symbol "-" signifies either the inapplicability of LLPFC to CIFAR-100 or the absence of reported results for SVHN, Fashion-MNIST, and MiniImageNet in FLMm's experiments.

Dataset	Model	Bag Size					Fully Supervised
		16	32	64	128	256	
CIFAR-10	LLPFC	84.10±0.19	71.70±0.78	52.71±0.36	20.78±0.70	18.79±0.21	96.05±0.33
	DLLP	91.59±1.52	88.61±0.90	79.76±1.45	64.95±0.01	44.87±1.13	
	LLP-VAT	91.80±0.08	89.11±0.22	78.75±0.46	63.89±0.19	46.93±0.71	
	ROT	94.86±0.68	94.34±0.65	93.97±0.96	92.23±0.81	63.10±0.84	
	SoftMatch	**95.24±0.12**	**95.25±0.14**	94.23±0.18	**93.87±0.22**	48.20±0.26	
	FLMm*	92.34	92.00	91.74	91.54	91.29	
	L^{2}P-AHIL	94.96±0.13	95.00±0.11	**94.58±0.21**	93.64±0.20	**92.88±0.53**	
CIFAR-100	LLPFC	-	-	-	-	-	79.89±0.14
	DLLP	71.28±1.56	69.92±2.86	53.58±1.60	25.86±2.15	8.82±0.94	
	LLP-VAT	73.85±0.22	71.62±0.07	65.31±0.33	37.36±0.63	2.79±0.67	
	ROT	72.74±0.08	69.31±0.22	17.48±0.86	11.02±0.79	2.86±1.11	
	SoftMatch	**80.14±0.12**	2.40±0.15	2.04±0.10	2.12±0.13	1.98±0.20	
	FLMm*	66.16	65.59	64.07	61.25	**57.10**	
	L^{2}P-AHIL	78.65±0.28	**77.30±0.50**	**76.52±0.23**	**72.21±0.37**	23.56±2.13	
SVHN	LLPFC	93.04±0.21	23.26±0.63	21.28±0.23	20.54±0.37	19.58±0.09	97.77±0.03
	DLLP	96.90±0.50	96.93±0.23	96.64±0.32	95.51±0.04	94.34±0.12	
	LLP-VAT	96.88±0.03	96.68±0.01	96.38±0.10	95.29±0.17	92.18±0.29	
	ROT	95.54±0.10	94.78±0.13	96.75±0.11	26.00±0.43	12.15±0.57	
	SoftMatch	22.39±0.11	19.68±0.13	19.60±0.12	19.64±0.14	19.57±0.16	
	FLMm*	-	-	-	-	-	
	L^{2}P-AHIL	**97.91±0.02**	**97.88±0.01**	**97.74±0.06**	**97.67±0.17**	**96.98±0.31**	
Fashion-MNIST	LLPFC	88.40±0.23	85.85±0.03	73.63±0.48	28.36±0.76	20.03±0.47	96.39±0.02
	DLLP	94.20±0.02	93.70±0.39	93.18±0.22	91.70±0.21	89.62±0.46	
	LLP-VAT	94.69±0.20	94.17±0.16	93.25±0.18	92.30±0.13	89.51±0.51	
	ROT	94.25±0.17	93.68±0.22	92.53±0.46	91.84±0.19	90.14±0.31	
	SoftMatch	95.85±0.22	**95.86±0.25**	95.18±0.21	**94.73±0.20**	93.29±0.19	
	FLMm*	-	-	-	-	-	
	L^{2}P-AHIL	**96.93±0.23**	95.78±0.15	**95.27±0.13**	94.19±0.14	**93.51±0.49**	
MiniImageNet	LLPFC	-	-	-	-	-	73.95±0.22
	DLLP	64.53±0.41	55.37±0.38	27.57±0.20	9.06±0.14	3.40±0.14	
	LLP-VAT	64.17±0.34	54.36±0.29	30.96±0.24	9.69±0.17	4.90±0.09	
	ROT	67.02±0.34	27.49±0.38	6.01±0.30	3.50±0.10	1.75±0.13	
	SoftMatch	2.02±0.23	1.86±0.24	1.95±0.20	1.72±0.33	1.87±0.25	
	FLMm*	-	-	-	-	-	
	L^{2}P-AHIL	**70.26±0.26**	**59.81±0.21**	**37.51±0.16**	**16.91±0.15**	**7.46±0.08**	

on the model's output predictions to estimate the proportions of bags. C) LLP-VAT [41]: This method is based on consistency regularization, where the same sample under different data augmentations should yield similar prediction results. D) ROT [9]: This method uses optimal transport to generate pseudo-labels for instance-level training combined with bag-level proportion loss. To ensure fairness, we employed the same data augmentation methods used in our approach for instance-level training. E) SoftMatch [6]: This method is one of the most important adaptive weighting techniques in weakly-supervised learning. We adopt the truncated Gaussian function, proposed in [6], as the instance-level adaptive weight. F) FLMm [45]: This method entails pre-training with contrastive learning using a larger encoder, ResNet-50, followed by fine-tuning utilizing the FLMm approach. $*$) Fully supervised: Our fully-supervised approach uses instance-level ground-truth labels for training.

All comparison methods except FLMm, including the fully-supervised approach, are reproduced by us. On the same dataset, we use the same experimental settings. In our method, hyper parameter settings are $\lambda = 0.5$, $\beta_{\mathrm{b}} = 1$, $\beta_{\mathrm{i}} = 1$, and specially for CIFAR-100 and Mini-ImageNet, $\beta_{\mathrm{b}} = 5$, $\beta_{\mathrm{i}} = 5$.

4.3. Results and Analysis

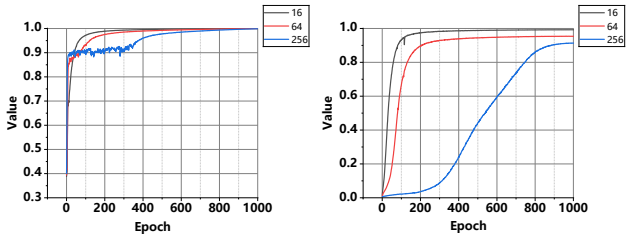
L^{2}P-AHIL achieves SOTA results. We present results on five benchmark datasets in Table 1. We observe that L^{2}P-AHIL significantly outperforms the majority of baseline methods across all benchmark datasets and for every bag

Table 2. Classification accuracy (mean±std) of ablation experiments with respect to weight. The highest accuracy is highlighted in bold.

Dataset	Weight		Bag Size				
	ω^{b}	ω^{i}	16	32	64	128	256
CIFAR-10	✗	✗	94.60±0.23	94.01±0.31	93.61±0.30	86.58±0.54	51.90±0.35
	✓	✗	94.63±0.33	94.51±0.29	94.21±0.34	92.85±0.35	78.03±0.25
	✗	✓	**95.00±0.08**	94.46±0.24	94.31±0.18	92.23±0.27	65.54±0.24
	✓	✓	94.96±0.13	**95.00±0.11**	**94.58±0.21**	**93.64±0.20**	**92.88±0.53**
CIFAR-100	✗	✗	77.54±0.37	8.91±2.49	3.05±0.94	2.23±0.87	1.96±0.95
	✓	✗	76.68±0.46	75.40±0.52	72.97±0.23	65.09±0.13	21.59±0.09
	✗	✓	77.97±0.86	76.86±0.32	70.99±0.74	69.48±0.19	2.89±1.09
	✓	✓	**78.65±0.28**	**77.30±0.50**	**76.52±0.23**	**72.21±0.37**	**23.56±2.13**

size. Compared with FLMm, which utilizes a stronger pre-trained encoder, our method's performance is only lower on CIFAR-100 with a bag size of 256. In other bag sizes, our method significantly surpasses FLMm, particularly on CIFAR-100 with bag sizes of 16 and 32, where we exceed their performance by 12.49% and 11.71%, respectively. For methods other than FLMm, our results exhibit a more gradual decline with increasing bag sizes. For instance, as the bag size increases from 16 to 256 on CIFAR-10, L^{2}P-AHIL's performance drops by only 2.08%, whereas other methods show a decline of over 40%. Compared to ROT, which also uses pseudo-label self-training, we consistently achieve significantly better results in all cases. SoftMatch performs well in certain cases, particularly when the bag size is small, such as 16 on CIFAR-10/100. However, as the bag size increases, its performance drops significantly. For example, when the bag size is 128 on CIFAR-100, the accuracy falls to only 2.12%. Although SoftMatch uses a similar technique to our method, the adaptive weighting it employs fails to address the issue of over-smoothing. More importantly, L^{2}P-AHIL achieves results comparable to those of fully supervised learning models. For example, at a bag size of 16, L^{2}P-AHIL's performance on simpler datasets like Fashion-MNIST and SVHN matches that of fully supervised learning. On more complex datasets such as CIFAR-10 and CIFAR-100, there's only a slight performance degradation, lower than 2%. For the more challenging Mini-ImageNet dataset, L^{2}P-AHIL achieves the highest accuracy across all bag sizes.

Variations in adaptive weighting during training. As shown in Fig. 4, the adaptive weights in L^{2}P-AHIL increase progressively throughout the training process, potentially reaching high values to fully leverage the information from pseudo-labels. Simultaneously, L^{2}P-AHIL effectively measures the confidences of pseudo-labels, efficiently minimizing the interference of meaningless pseudo-labels in the training process and achieving state-of-the-art results.



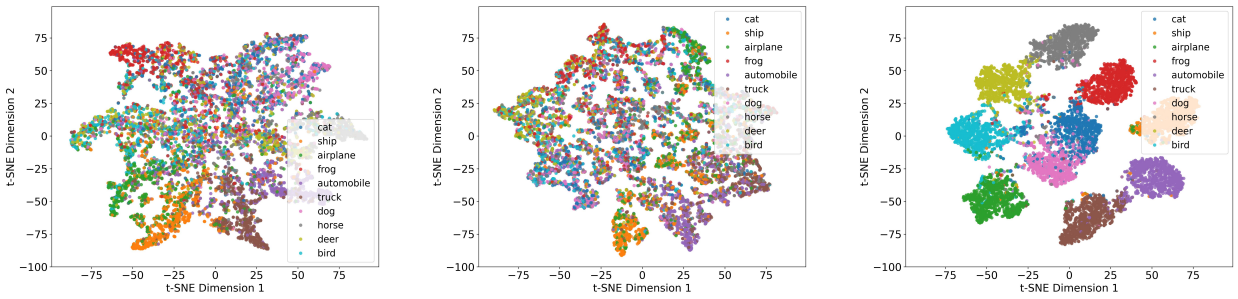
(a) CIFAR-10 (b) CIFAR-100

Figure 4. Evolution of adaptive weights across training epochs for bag sizes 16, 64 and 256 on CIFAR-10 and CIFAR-100.

L^{2}P-AHIL learns more distinguishable representations. We visualize the image representations generated by the feature encoder using t-SNE [42] in Fig. 5. Different colors correspond to different ground-truth class labels. We use the CIFAR-10 dataset with a bag size 256 and compare the t-SNE embeddings of three classic methods. For the traditional method DLLP, training the model solely with the bag-level proportion loss does not lead to effective representation learning. Although ROT introduces an instance-level loss through pseudo labels, it still falls short of achieving sufficient representation learning. In contrast, L^{2}P-AHIL generates well-separated clusters and more distinct representations, demonstrating its superior ability to learn high-quality representations.

4.4. Ablation Study

In this section, we perform an ablation study to assess the influence of two central elements of L^{2}P-AHIL: bag-level adaptive weight ω^{b} and instance-level adaptive weight ω^{i}. Experiments were conducted on the CIFAR-10 and CIFAR-100 datasets, with results presented in Table 2. The first row shows that removing both adaptive weights. With smaller bag sizes and fewer classes, the pseudo-labels generated by the model were relatively accurate, and the classification performance did not decline significantly, as seen with bag



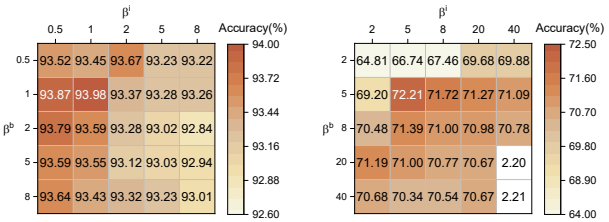
(a) DLLP features (b) ROT features (c) L^{2}P-AHIL features(ours)

Figure 5. t-SNE visualization of learned feature spaces on CIFAR-10 with bag size 256. Distinct colors correspond to different classes, where (a) DLLP and (b) ROT baselines exhibit overlapping clusters, while (c) our method L^{2}P-AHIL demonstrates clearer inter-class separation and tighter intra-class clustering, indicating more discriminative feature learning.

sizes from 16 to 64 on CIFAR-10. However, for larger bag sizes, such as 256, accuracy plummeted to merely 51.90%. On the more complex CIFAR-100, with a greater number of classes, the model's performance deteriorated substantially. On CIFAR-100, we observe that the model's performance drops sharply between adjacent bag sizes. The large number of classes and the reduction in supervisory information make the model more prone to generating inaccurate pseudo-labels, a phenomenon also observed in other weakly supervised domains. Without adaptive weights, the model is heavily impacted by incorrect pseudo-labels, leading to a significant performance drop. The second and third rows indicate that the model's accuracy decreases when the instance-level and bag-level adaptive weights are removed, respectively. This degradation intensifies with increasing bag sizes, highlighting that reliance on a single weight (bag-level or instance-level) inadequately captures pseudo-label confidence.

4.5. Parameter Sensitivity

In this section, we investigate how β_{b} and β_{i}, two key hyper parameters in L^{2}P-AHIL that control the mapping functions at bag-level and instance-level respectively, influence model performance. We conduct experiments on CIFAR-10 and CIFAR-100 with a bag size 128 (Fig. 6). For the 100-class CIFAR-100 task, the best result is achieved with $\beta_{\mathrm{b}} = 5$ and $\beta_{\mathrm{i}} = 5$. For the 10-class task, represented by CIFAR-10, the optimal result is obtained with $\beta_{\mathrm{b}} = 1$ and $\beta_{\mathrm{i}} = 1$. We selected our hyper parameters β_{b} and β_{i} based on these tests. When β_{b} and β_{i} are small, the weights are highly sensitive to changes in entropy, and some accurate pseudo-labels may be assigned low weights, preventing the model from fully utilizing them. For example, when β_{b} = 2 and β_{i} = 2, the accuracy on CIFAR-100 is only 64.81%. Conversely, when β_{b} and β_{i} are too large, many smooth but inaccurate pseudo-labels receive high weights, which hampers classifier training. For instance, with β_{b} = 40 and β_{i} = 40, the model fails to converge on CIFAR-100.



(a) CIFAR-10 (b) CIFAR-100

Figure 6. The plots of joint sensitivity tests for hyper parameters β_{b} and β_{i} on CIFAR-10 and CIFAR-100 with a bag size 128.

5. Conclusion

In this paper, we evaluated the problem of imprecise pseudo-labels in existing LLP methods. Addressing these challenges, we introduced L^{2}P-AHIL, a novel approach that measures pseudo-label confidence using both bag-level and instance-level entropy values, and then adaptively adjusts the pseudo-label weight for each instance. Our extensive experiments showed that L^{2}P-AHIL achieves state-of-the-art performance and remains effective even with larger bag sizes. We also conducted comprehensive ablation studies, confirming the effectiveness of each weight component in our method.

6. Limitations

The major limitation of L^{2}P-AHIL is that the hyper parameters of the adaptive weight is relatively sensitive to the datasets with various class numbers. Besides, our test dataset is considered classical and easily accessible, yet it has not been evaluated on larger, more complex datasets.

References

[1] Ehsan Mohammady Ardehaly and Aron Culotta. Co-training for demographic classification using deep learning from label proportions. In *IEEE International Conference on Data Mining Workshops*, pages 1017–1024, 2017. 2, 3, 5

[2] Takanori Asanomi, Shinnosuke Matsuo, Daiki Suehiro, and Ryoma Bise. Mixbag: Bag-level data augmentation for learning from label proportions. In *IEEE/CVF International Conference on Computer Vision*, pages 16570–16579, 2023. 1

[3] Gerda Bortsova, Florian Dubost, Silas Ørting, Ioannis Katramados, Laurens Hogeweg, Laura Thomsen, Mathilde Wille, and Marleen de Bruijne. Deep learning from label proportions for emphysema quantification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 768–776. Springer, 2018. 2

[4] Anand Brahmbhatt, Rishi Saket, and Aravindan Raghuveer. Pac learning linear thresholds from label proportions. In *Advances in Neural Information Processing Systems*, 2023. 1

[5] Róbert Busa-Fekete, Heejin Choi, Travis Dick, Claudio Gentile, and Andrés Muñoz Medina. Easy learning from label proportions. In *Advances in Neural Information Processing Systems*, 2023. 3

[6] Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality tradeoff in semi-supervised learning. In *International Conference on Learning Representations*, 2023. 3, 6

[7] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3008–3017, 2020. 5

[8] Yongke Ding, Yuanxiang Li, and Wenxian Yu. Learning from label proportions for sar image classification. *Eurasip Journal on Advances in Signal Processing*, 2017:1–12, 2017. 2

[9] Gabriel Dulac-Arnold, Neil Zeghidour, Marco Cuturi, Lucas Beyer, and Jean-Philippe Vert. Deep multi-class learning from label proportions. In *International Conference on Machine Learning*, 2019. 2, 3, 6

[10] Tianyu Guo, Chang Xu, Jiajun Huang, Yunhe Wang, Boxin Shi, Chao Xu, and Dacheng Tao. On positive-unlabeled classification in gan. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8382–8390, 2020. 1

[11] Shreyas Havaldar, Navodita Sharma, Shubhi Sareen, Karthikeyan Shanmugam, and Aravindan Raghuveer. Learning from label proportions: Bootstrapping supervised learners via belief propagation. In *NeurIPS Workshop on Regulatable Machine Learning*, 2023. 1

[12] Jerónimo Hernández-González, Inaki Inza, Lorena Crisol-Ortíz, María A. Guembe, María J. Iñarra, and Jose A. Lozano. Fitting the data from embryo implantation prediction: Learning from label proportions. *Statistical Methods in Medical Research*, 27(4):1056–1066, 2018. 1

[13] Yu-Guan Hsieh, Gang Niu, and Masashi Sugiyama. Classification from positive, unlabeled and biased negative data. In *International Conference on Machine Learning*, pages 2820–2829, 2019. 1

[14] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5070–5079, 2019. 3

[15] Ryoma Kobayashi, Yusuke Mukuta, and Tatsuya Harada. Learning from label proportions with instance-wise consistency. *arXiv*, 2022. 3

[16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[17] Zhengfeng Lai, Chao Wang, Henrry Gunawan, Sen-Ching S. Cheung, and Chen-Nee Chuah. Smoothed adaptive weighting for imbalanced semi-supervised learning: Improve reliability against unknown distribution data. In *International Conference on Machine Learning*, pages 11828–11843, 2022. 3

[18] Changchun Li, Ximing Li, and Jihong Ouyang. Semi-supervised text classification with balanced deep representation distributions. In *Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 5044–5053, 2021. 1

[19] Fan Li and Graham Taylor. Alter-cnn: An approach to learning from label proportions with application to ice-water classification. In *Neural Information Processing Systems Workshops (NIPSW) on Learning and privacy with incomplete data and weak supervision*, pages 100–106, 2015. 2

[20] Ximing Li, Yuanzhi Jiang, Changchun Li, Yiyuan Wang, and Jihong Ouyang. Learning with partial labels from semi-supervised perspective. In *AAAI Conference on Artificial Intelligence*, pages 8666–8674, 2023. 1

[21] Yukang Lin, Ronghui Li, Kedi Lyu, Yachao Zhang, and Xiu Li. Rich: Robust implicit clothed humans reconstruction from multi-scale spatial cues. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 193–206, 2023. 1

[22] Yukang Lin, Haonan Han, Chaoqun Gong, Zunnan Xu, Yachao Zhang, and Xiu Li. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6715–6724, 2024. 1

[23] Jiabin Liu, Bo Wang, Zhiquan Qi, Yingjie Tian, and Yong Shi. Learning from label proportions with generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 7167–7177, 2019. 2

[24] Jiabin Liu, Bo Wang, Xin Shen, Zhiquan Qi, and Yingjie Tian. Two-stage training for learning from label proportions. In *International Joint Conference on Artificial Intelligence*, pages 2737–2743, 2021. 2, 3

[25] Jiabin Liu, Zhiquan Qi, Bo Wang, Yingjie Tian, and Yong Shi. Self-llp: Self-supervised learning from label proportions with self-ensemble. *Pattern Recognition*, 129:108767, 2022. 2, 3

[26] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 5

[27] 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*, pages 6500–6510, 2020. 1

[28] Shinnosuke Matsuo, Ryoma Bise, Seiichi Uchida, and Daiki Suehiro. Learning from label proportion with online pseudo-label decision by regret minimization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2023. 2, 3

[29] Jay Nandy, Rishi Saket, Prateek Jain, Jatin Chauhan, Balaraman Ravindran, and Aravindan Raghuveer. Domain-agnostic contrastive representations for learning from label proportions. In *ACM International Conference on Information & Knowledge Management*, pages 1542–1551, 2022. 2

[30] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. *IEEE Transactions on Geoscience and Remote Sensing*, 2011. 5

[31] Giorgio Patrini, Richard Nock, Tibério S. Caetano, and Paul Rivera. (almost) no label no cry. In *Advances in Neural Information Processing Systems*, pages 190–198, 2014. 3

[32] Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. 5

[33] Novi Quadrianto, Alexander J. Smola, Tibério S. Caetano, and Quoc V. Le. Estimating labels from label proportions. In *International Conference on Machine Learning*, pages 776–783, 2008. 3

[34] Raúl Ramos-Pollán and Fabio A. González. Lightweight learning from label proportions on satellite imagery. *arXiv*, 2023. 1

[35] Zhongzheng Ren, Raymond A. Yeh, and Alexander G. Schwing. Not all unlabeled data are equal: Learning to weight data in semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2020. 3

[36] Laura Elena Cue La Rosa, Dario Augusto Borges Oliveira, and Pedram Ghamisi. Learning crop type mapping from regional label proportions in large-scale sar and optical imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 1

[37] Rishi Saket. Learnability of linear thresholds from label proportions. In *Advances in Neural Information Processing Systems*, pages 6555–6566, 2021. 3

[38] Rishi Saket. Algorithms and hardness for learning linear thresholds from label proportions. In *Advances in Neural Information Processing Systems*, 2022. 3

[39] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, 2020. 3, 5

[40] 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. 1

[41] Kuen-Han Tsai and Hsuan-Tien Lin. Learning from label proportions with consistency regularization. In *Asian Conference on Machine Learning*, pages 513–528, 2020. 2, 6

[42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 7

[43] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 5

[44] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *arXiv*, 2017. 5

[45] Haoran Yang, Wanjing Zhang, and Wai Lam. A two-stage training framework with feature-label matching mechanism for learning from label proportions. In *Asian Conference on Machine Learning*, pages 1461–1476, 2021. 2, 6

[46] Felix X. Yu, Krzysztof Choromanski, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. On learning from label proportions. *arXiv*, 2014. 2, 5

[47] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv*, 2016. 5

[48] Jianxin Zhang, Yutong Wang, and Clayton Scott. Learning from label proportions by learning with label noise. In *Advances in Neural Information Processing Systems*, 2022. 1, 3, 5

[49] Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14451–14461, 2022. 1

[50] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018. 1