Abstract—Learning from label proportions (LLP) is a widespread and important learning paradigm: only the bag-level proportional information of the grouped training instances is available for the classification task, instead of the instance-level labels in the fully supervised scenario. As a result, LLP is a typical weakly supervised learning protocol and commonly exists in practice. The data sensitivity and label scarcity restrict the use of the sensitive label information for real-world applications. In general, it is less laborious and more efficient to collect label proportions as the bag-level supervised information than the instance-level one. However, the hint for learning the discriminative feature representation is also limited as a less informative signal directly associated with the labels is provided, thus deteriorating the performance of the final instance-level classifier. In this article, delving into the label proportions, we bypass this weak supervision by leveraging generative adversarial networks (GANs) to derive an effective algorithm LLP-GAN. Endowed with an end-to-end structure, LLP-GAN performs approximation in the light of an adversarial learning mechanism without imposing restricted assumptions on distribution. Accordingly, the final instance-level classifier can be directly induced upon the discriminator with minor modification. Under mild assumptions, we give the explicit generative representation and prove the global optimality for LLP-GAN. In addition, compared with existing methods, our work empowers LLP solvers with desirable scalability inheriting from deep models. Extensive experiments on benchmark datasets and a real-world application demonstrate the vivid advantages of the proposed approach.

I. INTRODUCTION

Deep convolutional neural networks have demonstrated its advantages on a variety of challenging computer vision tasks, such as image recognition [16], [37], object detection [33], semantic segmentation [26], and image super-resolution [11]. However, this success largely depends on the large-scale annotated data, which is usually expensive and time-consuming to acquire. In practice, it is not feasible to build vast labeled training data for real-world applications, especially when the tasks are often of high complexity.

To address the bottleneck of the restrained data annotations in supervised learning, an alternative treatment is to exploit the weakly supervised learning paradigm, where substantially less directly supervised information is provided for the target learning tasks. A typical weakly supervised learning paradigm is multi-instance learning (MIL) [10]. A bunch of applications has been developed. For example, weakly supervised object detection (WSOD) has been widely studied and applied in real-world object detection [7], [13], [46]. On the other hand, in classification, learning from label proportions (LLP) is also built with a weakly supervised learning flavor but only providing the sample proportions of different categories. Training LLP aims to obtain a classifier for the new-come instances. Taking image classification as an example, in Fig. 1, we give an illustration of LLP, where data of three categories (dolphin, butterfly, and panda) are partitioned into four nonoverlapping groups. Then, the sizes of green, blue, and orange rectangles denote the label proportions of different categories in each group, respectively. In the learning phase, we seek a three-class classifier with only the sample feature information and class proportions in every group.

Up until now, solving LLP contributes to a number of real-world applications, such as demographic classification [1], video event detection [22], U.S. presidential election [39], domain adaptation [2], traffic flow prediction [24], and embryo implantation prediction [17]. Most of all, it can be employed in the field of medical and health informatics, such as disease prediction and auxiliary diagnosis, where people greatly pay attention to privacy protection. Accordingly, it is extremely difficult to obtain fully supervised data because this may result in information leakage for the patients.
Consequently, a relatively safe way to protect privacy is to pack the patients’ diagnosis data into many bags, and instead offer the bag-level supervision, and then provide them to the third party.

However, consider a generative model, this less informative supervision can hardly contribute to a satisfied classifier as a result of its insufficient implication of data distributions, e.g., the class-conditional probability $P(X | Y)$ and the marginal distribution $P(X)$. Even in a discriminative model, the indirect relationship between the instance $(X)$ and its response $(Y)$ is less determinative and leads to an underdetermined problem: Too many instance-level classifiers can satisfy proportional constraints exactly. In other words, in terms of instance classification, LLP is ill-posed. Despite that a number of achievements have been developed to resolve LLP accurately and efficiently, it is still of great challenge to design an effective instance-level learning scheme to significantly improve the performance of high-dimensional instances’ recognition, e.g., image classification, merely with the guide of bag-level proportions.

Generative models, on the other hand, are specialized in data representation and distribution inference. Apart from distribution estimation, more importantly, generative models are often applied to generate additional samples from an implicit or explicit approximate distribution [35]. Among them, generative adversarial networks (GANs) [14] have demonstrated its appealing advantages in both realistic data generation [23] and feature representation [32]. Unlike performing maximizing log-likelihood on the variational lower bound of unlabeled data, such as variational autoencoder (VAE) [19], [20], GANs seeks the equilibrium between two networks (discriminator and generator) by alternatively updates in an adversarial game. Specifically, with two players in an adversarial game, GAN utilizes the generator to synthesize realistic data from random noise and the discriminator to distinguish real samples from the generated ones. In doing so, the generator can eventually provide an implicit approximation that captures the essential statistics of data distribution.

In this article, we push the envelope further by focusing on applying GANs to LLP (see [31], [34], and [47] for more real-life applications). By referring group as bag, LLP also fits for learning with bags settings, which is primarily established in MIL [10]. To obtain an instance-level classifier, in our scheme, generated fake samples encourage the discriminator to not only detect the difference between the real and the fake instances but also distinguish true $K$ classes for real samples (through a $K + 1$-way classifier). Then, we can directly obtain the classifier upon the discriminator with minor modification. In practice, the performance improvement by incorporating the generative adversarial mechanism can be summarized as the following two reasons. On the one hand, it can better learn the data distributions $P(X)$ (the marginal) and $P(X, Y)$ (the joint) through GAN and, thus, further boost the LLP performance in the estimation of $P(Y | X)$ (the posterior distribution). On the other hand, the generative adversarial mechanism also plays the role of improving the feature representation ability for the discriminator, which is essential to build a competitive classifier.

This article is an extension of [25], and we provide more experimental results to further expound the mechanism of LLP-GAN and extensively demonstrate the capability of the proposed model to the application of privacy protection, which is proven by a real-world application. The main contributions of this article can be concluded as the following four aspects.

1) We propose a new algorithm called LLP-GAN, which first leverages the generative adversarial mechanism to boost the LLP.
2) Based on KL-divergence, we construct our LLP-GAN framework and study the global optimality for both generator and discriminator under mild assumptions.
3) In our algorithm, the generator learns data distributions through an adversarial scheme, without the common assumption that bags are independent and identical distributions.
4) Through in-depth empirical study with extensive experiments on benchmark datasets and a real-world case study, our approach achieves state-of-the-art performance, which is even comparable to a fully supervised situation when with small bag sizes.

II. RELATED WORK

To the best of our knowledge, three end-to-end pipelines are recently proposed for LLP using deep neural networks as the backbone architectures. Specifically, DLLP [1] is the first end-to-end LLP algorithm, using KL-divergence of prior and posterior proportions as the objective. However, this method always stuck with poor performance with optimization only on the KL-divergence, especially with large bag sizes (e.g., $> 64$).

Similarly, LLP-VAT [41] introduces a consistency regularization technique in semisupervised learning to a multiclass

Fig. 1. Illustration of a three-class LLP.
LLP problem. Then, a new regularization based on virtual adversarial training [27] is employed to prevent overfitting. However, this method provides limited improvement to DLLP, as the feature representation is still based on the ill-posed bag-level KL-divergence loss.

Based on the individual labels of samples, Dulac-Arnold et al. [12] and Yu et al. [48] employ an instance-level cross-entropy loss as the objective. In detail, this method solves an accurate labeling problem with optimal transport (OT) [30], which can strictly conform to the proportional constraint. Then, alternating update and convex relaxation are exploited to mitigate the intractable combinatorial optimization. However, it eventually solves an unbalanced OT problem with entropic regularization, resulting in a relatively poor feature representation ability due to weak supervision and fail to intensively dig into the inherent class-related information, which is obviously essential to a competitive algorithm. In contrast, we approach this problem by leveraging a better feature extractor to obtain more discriminative representations from the data. The main insight of our work is to implement better representation learning through the adversarial mechanism, thus boosting the downstream discriminative task performance.

III. Preliminaries

This section provides the necessary preliminaries for the proposed approach, including the formal problem setting and related work with simple extensions.

A. Multiclass LLP

Before further discussion, we formally describe multiclass LLP. For simplicity, we assume that all the bags are disjoint, and let \( \mathcal{B}_i = \{x_1^i, x_2^i, \ldots, x_N^i\}, i = 1, 2, \ldots, n \), be the bags in training set. Then, bagging data are \( \mathcal{D} = \mathcal{B}_1 \cup \mathcal{B}_2 \cup \cdots \cup \mathcal{B}_n \), and \( \mathcal{B}_i \cap \mathcal{B}_j = \emptyset, \forall i \neq j \), where the total number of bags is \( n \).

Assuming that we have \( K \) classes, for a bag \( \mathcal{B}_i \), let \( p_i \) be a \( K \)-element vector, where the \( k \)th element \( p_{ik} \) is the proportion of instances belonging to the class \( k \) in that bag, with the constraint \( \sum_{k=1}^{K} p_{ik} = 1 \), i.e.,

\[
p_{ik} = \frac{\left| \{ j \in [1:N_i] : x_j^i \in \mathcal{B}_i, y_j^{i*} = k \} \right|}{|\mathcal{B}_i|}.
\]

Here, \( [1:N_i] = \{1, 2, \ldots, N_i\} \), and \( y_j^{i*} \) is the unaccessible ground-truth instance-level label of \( x_j^i \). In this way, we can denote the available training data as \( \mathcal{L} = \{(\mathcal{B}_i, p_i)\}_{i=1}^n \). The goal of LLP is to learn an instance-level classifier based on \( \mathcal{L} \).

B. Deep LLP With Bag-Level KL Divergence Loss

In terms of deep learning, DLLP is the first deep model for multiclass LLP [1]. Using DNN’s probabilistic classification outputs, it is straightforward to adapt cross-entropy loss in a bag-level manner by averaging the probability outputs in every bag as the proportion estimation. Inspired by Wang and Feng [43], DLLP reshapes standard cross-entropy loss by substituting instance-level labels with label proportions in order to meet the requirement of proportion consistency.

In detail, suppose that \( \tilde{p}_j^i = p_{\theta}(y|x_j^i) \) is the vector-valued DNNs output for \( x_j^i \), where \( \theta \) is the network parameter. Let \( \oplus \) be the elementwise summation operator. Then, the bag-level label proportion in the \( i \)th bag is obtained by incorporating the elementwise posterior probability

\[
\mathbf{\bar{p}}_i = \frac{1}{N_i} \oplus \mathbf{\hat{p}}_j^i = \frac{1}{N_i} \oplus p_{\theta}(y|x_j^i).
\]

Different from discriminant approaches, in order to smooth \( \max \) function [5], \( \mathbf{\bar{p}}_j^i \) is in a vector-type softmax manner to produce the classification probability distribution. Taking log as the elementwise logarithmic operator, the objective of DLLP can be intuitively formulated using the cross-entropy loss

\[
L_{\text{prop}} = -\sum_{j=1}^{n} \mathbf{p}_j^i \log(\mathbf{\bar{p}}_i). \tag{3}
\]

It penalizes the difference between prior and posterior probabilities in the bag level and commonly exists in GAN-based SSL [38]. However, as discussed, this bag-level objective may fail to perform well due to underdetermination.

C. Entropy Regularization for DLLP

Following the entropy regularization strategy [15], we can accordingly introduce an extra loss \( E_{\text{in}} \) to (3) with a tradeoff hyperparameter \( \lambda \) to constrain instance-level output distribution in a low entropy so as to provide a classification result with high confidence

\[
L = L_{\text{prop}} + \lambda E_{\text{in}} = -\sum_{i=1}^{n} \mathbf{p}_j^i \log(\mathbf{\bar{p}}_i) - \lambda \sum_{i=1}^{n} \sum_{j=1}^{N_i} (\mathbf{\hat{p}}_j^i)^\top \log(\mathbf{\bar{p}}_i). \tag{4}
\]

This straightforward extension of DLLP is similar to a KL divergence, taking care of bag- and instance-level consistencies simultaneously. It takes advantage of DNN’s output distribution to cater to the label proportions requirement, as well as minimizing output entropy as a regularization term to guarantee high true-fake belief. This is believed to be linked with an inherent maximum a posteriori (MAP) estimation [5] with certain prior distribution in network parameters. However, we will not look at the performance of this extension and consider not including it as a baseline because the experimental results empirically suggest: the original DLLP has already converged to the solution with fairly low instance-level entropy, which makes the proposed regularization term redundant. We will show the empirical results in Section V-B.

IV. Adversarial Learning for LLP

Orthogonal to the bag-level loss discussed above, in this section, we focus on adversarial learning for LLP and propose LLP-GAN to devote GANs to harnessing LLP.

We illustrate the LLP-GAN framework in Fig. 2. First, the generator is employed to generate samples with input noise,
which is labeled as fake, and the discriminator yields class confidence maps for all the categories (including the fake one) by taking both fake and real samples as its inputs. This results in our adversarial loss in LLP-GAN. Second, we incorporate the proportions by adding the ordinary cross-entropy loss.

A. Objective Function of Discriminator

In LLP-GAN, to restate the key point, the discriminator is not only to identify whether a sample is from the real data or not but also to elaborately distinguish each real input’s label assignment as a $K$ classes classifier. However, we do not know any instance-level ground-truth label. Consequently, we conduct unsupervised adversarial learning, i.e., $L_{\text{unsup}}$ term.

Next, the main issue becomes how to exploit the proportional information to guide this unsupervised learning correctly. To this end, we replace the supervised information in semisupervised GANs, the cross-entropy term harnesses the label proportions consistency. In order to justify the nontriviality of this loss, we first look at its lower bound. More importantly, it is easier to perform the gradient method on the lower bound because it swaps the order of log and the summation operation. For brevity, the analysis will be done in a nonparametric setting, i.e., we assume that both data distributions have infinite capacity.

Definition 1: Suppose that $\mathcal{P}$ is a partition that divides the data space into $n$ disjoint sections. Let $p'_j(x), i = 1, 2, \ldots, n,$ be the marginal distributions with respect to elements in $\mathcal{P}$, respectively. Accordingly, $n$ bags in LLP training data spring from sampling upon $p'_j(x), i = 1, 2, \ldots, n.$ In the meantime, let $p(x, y)$ be the unknown holistic joint distribution.

We normalize the first $K$ classes in $p_D(\cdot|x)$ into the instance-level posterior probability $\bar{p}_D(\cdot|x)$ and compute $\bar{p}$ based on (2). Then, the ideal optimization problem for the discriminator of LLP-GAN is

$$
\max_D V(G, D) = L_{\text{unsup}} + L_{\text{sup}} = L_{\text{real}} + L_{\text{fake}} - \lambda C \mathbb{E}_L(p, \bar{p})
= \sum_{i=1}^{n} \mathbb{E}_{x \sim p'_i} [\log P_D(y \leq K|x)] + \mathbb{E}_{x \sim p} [\log P_D(K+1|x)]
= +\lambda \sum_{i=1}^{n} p_{i}^{T} \log(\bar{p}_{i})
$$

(5)

where $p_{g}(x)$ is the distribution of the synthesized data.

Remark 1: When $P_D(K + 1|x) \neq 1$, the normalized instance-level posterior probability $P_D(\cdot|x)$ is given by

$$
P_D(k|x) = \frac{P_D(k|x)}{1 - P_D(K+1|x)}, \quad k = 1, 2, \ldots, K.
$$

If $P_D(K + 1|x) = 1$, we have $\bar{P}_D(k|x) = 1/K, k = 1, 2, \ldots, K$.

Note that the weight $\lambda$ in (5) is to balance supervised and unsupervised terms, which is a slight modification of SSL with GANs [9, 36]. Intuitively, we reckon that proportional information is too weak to be comparable with supervised learning. Hence, a relatively large weight should be preferable in the experiments. However, large $\lambda$ may result in unstable GANs training. We study the effect of $\lambda$ in Section V-G. However, for simplicity, we fix $\lambda = 1$ in the following theoretical analysis on the discriminator.

Aside from identifying the first two terms in (5) as that in semisupervised GANs, the cross-entropy term harnesses the label proportions consistency. In order to justify the nontriviality of this loss, we first look at its lower bound. More importantly, it is easier to perform the gradient method on the lower bound because it swaps the order of log and the summation operation. For brevity, the analysis will be done in a nonparametric setting, i.e., we assume that both $D$ and $G$ have infinite capacity.

Remark 2 (Lower Bound Approximation): Let $p_i(k)$ be the class $k$ proportion in the $i$th bag. According to the idea of sampling methods and Jensen’s inequality, we have

$$
-C \mathbb{E}_L(p, \bar{p}) = \sum_{i=1}^{n} \sum_{k=1}^{K} p_i(k) \log \left( \frac{1}{N_i} \sum_{j=1}^{N_i} \bar{P}_D(k|x'_j) \right) \\
\leq \sum_{i=1}^{n} \sum_{k=1}^{K} p_i(k) \int p'_i(x) \bar{P}_D(k|x) \, dx \\
\geq \sum_{i=1}^{n} \sum_{k=1}^{K} p_i(k) \mathbb{E}_{x \sim p'_i} [\log \bar{P}_D(k|x)].
$$

(7)

The expectation in the last term can be approximated by Monte Carlo sampling. More importantly, similar
to EM mechanism [28] for mixture models, by approximating \(-C\mathcal{E}_{\mathcal{L}}(p, \mathcal{P})\) with its lower bound, we can perform gradient ascent independently on every sample. Hence, SGD can be applied.

As shown in (7), in order to facilitate gradient computation, we substitute cross-entropy in (5) by its lower bound and denote this approximate objective for discriminator by \(\tilde{V}(G, D)\).

**B. Optimal Discriminator and LLP Classifier**

Now, we give the optimal discriminator and the final classifier for LLP based on the analysis of \(\tilde{V}(G, D)\). First, we have the following result of the lower bound in (7).

**Lemma 1:** Maximization on lower bound in (7) induces an optimal discriminator \(D^*\) with posterior distribution \(\tilde{p}_D(y|x)\), which is consistent with prior distribution \(p_i(y)\) in each bag.

Proof: Taking the aggregation with respect to one single bag, for example, the \(i\)th bag, we have

\[
\mathbb{E}_{x \sim p_{i}} \log p(x) = \mathbb{E}_{x \sim p_{i}} \log \frac{p(x, y)}{\hat{P}_D(x|y)} = \mathbb{E}_{x \sim p_{i}} \int p_i(y) \log \frac{p_i(y)p(x|y)}{\hat{P}_D(x|y)} \, dy \\
= \mathbb{E}_{x \sim p_{i}} \int p_i(y) \log \frac{\hat{P}_D(x|y)}{p_i(y)} \, dy + \mathbb{E}_{x \sim p_{i}} KL(p_i(y)||\hat{P}_D(y|x)) \\
= \sum_{k=1}^{K} p_i(k) \mathbb{E}_{x \sim p_{i}} \log \frac{\hat{P}_D(x|k)}{p_i(k)}.
\]

Here, we only consider \(x \sim p_{i}\), so \(p(x, y) = p_i(y)p(y|x)\) holds. Note that the last term in (8) is free of the discriminator, and the aggregation can be independently performed within every bag due to the disjoint assumption on bags. Then, maximizing the lower bound in (7) is equivalent to minimizing the expectation of KL-divergence between \(p_i(y)\) and \(\hat{P}_D(y|x)\).

Because of the infinite capacity assumption on discriminator and the nonnegativity of KL-divergence, we have

\[
D^* = \arg \min_D \mathbb{E}_{x \sim p_{i}} KL(p_i(y)||\hat{P}_D(y|x)) \\
\iff \hat{P}_{D^*}(y|x) \overset{a.e.}{=} p_i(y), \quad x \sim p_{i}(x) \quad (9)
\]

which concludes the proof.

Lemma 1 tells us that, if there is only one bag, then the final classifier \(\hat{P}_D(y|x)\) \(\overset{a.e.}{=} p(y)\). However, there are normally multiple bags in the LLP problem, and the final classifier will somewhat be a tradeoff among all the prior proportions \(p_i(y), i = 1, 2, \ldots, n\), unless the bags are independent of each other. Next, we will show how the adversarial learning on the discriminator helps to determine the formulation of this tradeoff in a weighted aggregation.

**Theorem 1:** For fixed \(G\), the optimal discriminator \(D^*\) for \(\tilde{V}(G, D)\) satisfies

\[
P_D^*(y = k|x) = \frac{\sum_{i=1}^{n} p_i(k)p_{i}(x)}{\sum_{i=1}^{n} p_{i}(x) + p_s(x)} \quad (10)
\]

for all \(k = 1, \ldots, K\).

**Proof:** According to (5) and (7), given any generator \(G\), we have

\[
\tilde{V}(G, D) = \sum_{i=1}^{n} \mathbb{E}_{x \sim p_{i}} [\log (1 - P_D(K + 1|x))] + \mathbb{E}_{x \sim p_s} [\log P_D(K + 1|x)] \\
+ \sum_{i=1}^{n} \sum_{k=1}^{K} p_i(k) \mathbb{E}_{x \sim p_{i}} [\log P_D(k|x)]
\]

\[
= \int \left[ \sum_{i=1}^{n} p_{i}(x) \left[ \log \sum_{k=1}^{K} P_D(k|x) \\
+ \sum_{k=1}^{K} p_i(k) \log \frac{P_D(k|x)}{1 - P_D(K + 1|x)} \right] + p_s(x) \log \left[ 1 - \sum_{k=1}^{K} P_D(k|x) \right] \right] \, dx. \quad (11)
\]

By taking the derivative of the integrand, we find the solution in [0, 1] for maximization as (10).

**Remark 3 (Beyond the Incontinuity of \(p_s\)):** According to Arjovsky and Bottou [3], the problematic scenario is that the generator is a mapping from a low-dimensional space to a high-dimensional one. This will result in the density of \(p_g(x)\) infeasible. However, based on the definition of \(\hat{P}_D(y|x)\) in (6), we have

\[
\hat{P}_{D^*}(y|x) = \frac{\sum_{i=1}^{n} p_i(y)p_{i}(x)}{\sum_{i=1}^{n} p_{i}(x)} = \sum_{i=1}^{n} w_i(x)p_i(y) \quad (12)
\]

Hence, the final classifier is independent to \(p_s(x)\). Furthermore, (12) explicitly expresses the normalized weights of the aggregation

\[
w_i(x) = \frac{p_{i}(x)}{\sum_{i=1}^{n} p_{i}(x)} \quad (13)
\]

**Remark 4 (Relationship to One-Side Label Smoothing):** Notice that the optimal discriminator \(D^*\) is also related to the one-sided label smoothing mentioned in [36], which is inspired by [40] and shown to reduce the vulnerability of neural networks to adversarial examples [44].

In particular, in our model, we only smooth labels of real data (multiclass) in the discriminator by setting the targets as the prior proportions \(p_i(y)\) in corresponding bags.

**C. Objective Function of Generator**

Normally, for the generator, we should solve the following optimization problem with respect to \(p_s\):

\[
\min_G \tilde{V}(G, D^*) = \min_G \mathbb{E}_{x \sim p} \log P_D^*(K + 1|x). \quad (14)
\]

Let us denote

\[
C(G) := \max_D \tilde{V}(G, D) = \tilde{V}(G, D^*). \quad (15)
\]

Since \(\tilde{V}(G, D)\) is convex in \(p_s\) and the supremum of a set of convex functions is also convex, we have the following sufficient and necessary condition of global optimality.
**Theorem 2:** The global minimum of $C(G)$ is achieved if and only if $p_\epsilon = 1/n \sum_{i=1}^n p_d^i$.  

**Proof:** Denote $p_d := \sum_{i=1}^n p_d^i$. Then, according to Theorem 1, we can reformulate $C(G)$ as

$$C(G) = \sum_{i=1}^n \mathbb{E}_{x \sim p_d^i} \log \frac{p_d(x)}{p_\epsilon(x) + p_d(x)} + \mathbb{E}_{x \sim p_\epsilon} \log \frac{p_\epsilon(x)}{p_d(x) + p_\epsilon(x)}$$

$$= + \sum_{i=1}^n \sum_{k=1}^n p_i(k) \mathbb{E}_{x \sim p_d^i} \log \tilde{P}_D(k|x)$$

$$= 2 \cdot \text{JSD}(p_d^i | p_d) - 2 \log 2$$

$$= - \sum_{i=1}^n \mathbb{E}_{x \sim p_d^i} [\text{CE}(p_i(y), \tilde{P}_D(y|x))]$$

where JSD$(\|\|)$ and CE$(\cdot, \cdot)$ are the Jensen–Shannon divergence and cross-entropy between two distributions, respectively. However, note that $p_d$ is the summation of $n$ independent distributions, so $(1/n)p_d$ is a well-defined probabilistic density. Then, we have

$$C(G^*) = \min_{G} C(G) = n \log(n) - (n + 1) \log(n + 1)$$

$$- \sum_{i=1}^n \mathbb{E}_{x \sim p_d^i} [\text{CE}(p_i(y), \tilde{P}_D(y|x))] \iff p_{\epsilon} \leq \frac{1}{n} p_d.$$

(17)

This concludes the proof.

**Remark 5:** When there is only one bag, the first two terms in (17) will degenerate as $n \log(n) - (n + 1) \log(n + 1) = -2 \log 2$, which adheres to results in original GANs. On the other hand, the third term manifests the uncertainty on instance labels, which is concealed in the form of label proportions.

**Remark 6:** According to the analysis above, ideally, we can obtain the Nash equilibrium between the discriminator and the generator, i.e., the solution pair $(G^*, D^*)$ satisfies

$$\tilde{V}(G^*, D^*) \geq \tilde{V}(G^*, D) \quad \forall D$$

$$\tilde{V}(G^*, D^*) \leq \tilde{V}(G, D^*) \quad \forall G.$$  

(18)

However, as shown in [9], a well-trained generator would lead to the inefficiency of supervised information. In other words, the discriminator would possess the same generalization ability as merely training it on $L_{prop}$. Hence, we apply feature matching (FM) to the generator and obtain its alternative objective by matching the expected value of the features (statistics) on an intermediate layer of the discriminator [36]

$$L(G) = \| \mathbb{E}_{x \sim p_d} f(x) - \mathbb{E}_{x \sim p_\epsilon} f(x) \|_2^2.$$  

(19)

In fact, FM is similar to the perceptual loss for style transfer in a concurrent work [18], and the goal of this improvement is to impede the “perfect” generator resulting in unstable training and discriminator with low generalization.

### D. LLP-GAN Algorithm

So far, we have clarified the objective functions of both discriminator and generator in LLP-GAN. When accomplishing the training stage, the discriminator can be put into effect as the final classifier. The strict proof for algorithm convergence is similar to that in [14]. Because $\max_D \tilde{V}(G, D)$ is convex in $G$, and the subdifferential of $\max_D \tilde{V}(G, D)$ contains that of $\tilde{V}(G, D^*)$ in every step, the line search method (stochastic) gradient descent eventually converges [6].

We present the LLP-GAN algorithm that coincides with the algorithm of the original GAN [14].

**Algorithm 1 LLP-GAN Training Algorithm**

**Input:** The training set $L = \{(B_i, y_i)\}_{i=1}^L$; $L$: number of total iterations; $\lambda$: weight parameter.

**Output:** The parameters of the final discriminator $D$.  

**Set $m$ to the total number of training data points.**

**for** $i = 1: L$ **do**

- Draw $m$ samples $\{z^{(1)}, z^{(2)}, \ldots, z^{(m)}\}$ from a simple-to-sample noise prior $p(z)$ (e.g., $N(0, I)$).  
- Compute $\{G(z^{(1)}, G(z^{(2)}), \ldots, G(z^{(m)})\}$ as sampling from $p_\epsilon(z)$.
- Fix the generator $G$ and perform gradient ascent on parameters of $D$ in $\tilde{V}(G, D)$ for one step.
- Fix the discriminator $D$ and perform gradient descent on parameters of $G$ in $L(G)$ for one step.

**end**

Return parameters of discriminator $D$ in the last step.

### V. EXPERIMENTS

In this section, we conduct extensive numerical experiments mainly on image datasets to demonstrate the capability of the proposed LLP-GAN in solving the LLP problem. First, synthetic data “two moons” is employed to intuitively illustrate the advantage of our approach. Then, we evaluate the performance of our method on four benchmark datasets, Kuzushi-MNIST [8], Fashion-MNIST [45], CIFAR-10, and CIFAR-100 [21], and conduct in-depth comparisons with several recently proposed LLP algorithms. Besides, we evaluate our method and other SOTA LLP models on real-world datasets and empirically study the effect of hyperparameter.

#### A. Experimental Setup

1) **Label Proportion Generation:** As there are no off-the-shelf learning from label proportions datasets, the first challenge is how to generate the proportion-based supervised information using the ordinary fully supervised datasets. Following the approach in [48], in this article, we obtain the proportional information by randomly dividing the training data into different bags and compute the corresponding label proportions of each bag with the known instance-level labels. To keep up the same settings in the previous work, the bag size is fixed as 16, 32, 64, and 128. We conceal the accessible instance-level labels and instead provide the bag-level label proportions for LLP. Note that we still need the instance-level labels in test data to justify the effectuality of the learned classifier.

2) **Training Settings:** We employ Pytorch as the deep learning platform for all the experiments. For K-MNIST and F-MNIST, a five-hidden-layer fully connected MLP network
is employed, while a 13-layer max-pooling-based network is used for CIFAR-10 and CIFAR-100. For the optimizer, we select ADAM with the corresponding parameter settings as $\beta_1 = 0.5$ and $\beta_2 = 0.999$. For all the datasets, the initial learning rate is consistently set as $1e-4$ and divided by 2 every 100 epochs when bag sizes are 16 and 32. For bag sizes of 64 and 128, we constantly fix the learning rate as $1e-4$ to prevent underfitting. On the other hand, in order to avoid overfitting, we follow the commonly practical data augmentation to pad the image with four pixels on each side, take a random $32 \times 32$ crop, and horizontally flip randomly with the probability of 0.5 for CIFAR-10 and CIFAR-100. In this article, the final performance is obtained by running the model five times with 300 epochs.

B. D LLP With Entropy Regularization

Although DLLP with entropy regularization is a side contribution of this work, as claimed before, we consider not including it as a baseline. The reason is the experimental result on MNIST in Fig. 3 suggests the solution of the original DLLP has consistently converged to the one with fairly low instance-level entropy, which makes the regularization redundant.

C. Synthetic Data

In this part, we first demonstrate the advantage of applying generative adversarial mechanisms to LLP through a toy example. In detail, we employ the well-known “two-moon” dataset to compare the performance of DLLP [1], LLP-VAT [41], and LLP-GAN. All the approaches that use the same pipeline neural network with three hidden layers are employed with ReLU as the activation function. In order to build LLP training data, we generate 2000 data points in total and equally divide them into 40 bags, with each one containing 50 points.

The results are shown in Fig. 4, where instances in different classes are in red and cyan, respectively. For DLLP in Fig. 4(a), blobs of data are misclassified for both categories, while the proportion of misclassification data points reduces when incorporating the generative adversarial mechanism in Fig. 4(c). When it comes to stronger baseline LLP-VAT in Fig. 4(b), a little improvement is obtained compared with DLLP. However, it is obviously inferior to LLP-GAN.

This synthetic experiment indicates the advantage of introducing the generative adversarial mechanism to LLP.

As shown in the figure, the generative adversarial training helps to better learn the data distributions $P(X)$ (the marginal) and $P(X, Y)$ (the joint), which provides important manifold information of the data, i.e., two moons are disconnect and the points in the same moon are intrinsically more related with a closer distance compared with the points in different moons. This discovery by the adversarial training, thus, further boosts the LLP performance in the estimation of $P(Y|X)$ (the posterior distribution). On the other hand, the generative adversarial mechanism also improves the feature representation for the discriminator, which is essential to build a competitive classifier. As is shown, the samples are not linearly demonstrated but in the fashion of two moons. Compared to other methods, our LLP-GAN can better learn the manifold information.

D. Results on KMNIST and FMNIST

In this part, we choose datasets KMNIST and FMNIST to demonstrate the advantage of our model. Specifically, KMNIST is designed for classifying classical Japanese literature, containing 60000 training data and 10000 test data. FMNIST proposes a task to classify the fashion of products, consisting of 60000 training examples and 10000 test examples.

We employ a fully connected network for classification (discriminator in our model), and the generator is also constructed with a fully connected layer. In practice, the performance of GAN is strongly related to the structure of the generator, and a sophisticated design often results in a better performance. However, this is beyond the scope of this article. For these two datasets, we just construct the generator by four hidden layers with ReLU as the activation function, where we also employ batch normalization to stabilize the training. Besides, there is no data augmentation for these two datasets with just pulling each spatially 2-D instance into a 1-D vector and followed by fully connected layers.

We first conduct a comprehensive comparison for KMNIST and FMNIST in Table I, which provides the accuracy for DLLP, LLP-VAT, and LLP-GAN with different bag sizes.
including 16, 32, 64, and 128. Considering the randomness in the results, we conduct the experiment five times and report the average accuracy along with the corresponding standard deviation. Specifically, we can observe that the advantage of LLP-GAN is significant compared with DLLP and LLP-VAT for KMNIST. Particularly, this advantage becomes even more apparent when the bag size is relatively bigger. In general, a bigger bag size means more challenging for the corresponding task. That is to say, our model is more effective under extremely hard circumstances. This indicates that the implicit learning of data distributions with GAN in an unsupervised manner can improve the feature representation for the downstream discriminative tasks, including LLP studied in this article.

In practice, one good reason for introducing a generative adversarial mechanism to the classification model is that it can learn better feature representation for the downstream classification. To show this, we apply t-SNE [42] to visualize the embeddings of DLLP and LLP-GAN on the test data of KMNIST with the same model settings, as well as the bag size 64. More specifically, we extract the features in the last layer for both DLLP and LLP-GAN after 200 epoch training. From Fig. 6, the learned representations of LLP-GAN are more significantly concentrated to clusters, especially for the center part by comparing Fig. 6(a) with Fig. 6(b). This visualization is consistent with our analysis.

**E. Results on CIFAR-10 and CIFAR-100**

In this part, we perform DLLP, LLP-VAT, and LLP-GAN on the convolution-based network on the datasets CIFAR-10 and CIFAR-100. Considering the randomness in the results, we conduct the experiment five times and report the average accuracy along with the corresponding standard deviation. Specifically, we can observe that the advantage of LLP-GAN is significant compared with DLLP and LLP-VAT for KMNIST. Particularly, this advantage becomes even more apparent when the bag size is relatively bigger. In general, a bigger bag size means more challenging for the corresponding task. That is to say, our model is more effective under extremely hard circumstances. This indicates that the implicit learning of data distributions with GAN in an unsupervised manner can improve the feature representation for the downstream discriminative tasks, including LLP studied in this article.

### TABLE I

**TEST ACCURACY RATES AND STANDARD DEVIATIONS (%) ON KMNIST AND FMNIST WITH DIFFERENT BAG SIZES. THE RESULTS ARE OBTAINED WITH FIVE RUNS**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Bag Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>K-MNIST</td>
<td>DLLP [1]</td>
<td>92.35 (0.22)</td>
</tr>
<tr>
<td></td>
<td>LLP-VAT [41]</td>
<td>92.45 (0.17)</td>
</tr>
<tr>
<td></td>
<td>LLP-GAN</td>
<td>92.75 (0.19)</td>
</tr>
<tr>
<td>F-MNIST</td>
<td>DLLP [1]</td>
<td>88.71 (0.21)</td>
</tr>
<tr>
<td></td>
<td>LLP-VAT [41]</td>
<td>88.37 (0.28)</td>
</tr>
<tr>
<td></td>
<td>LLP-GAN</td>
<td><strong>88.81</strong> (0.16)</td>
</tr>
</tbody>
</table>

Fig. 5. Convergence curves of different algorithms on KMNIST and FMNIST with bag sizes of 64 and 128. (Best viewed in color.) (a) KMNIST-64. (b) KMNIST-128. (c) FMNIST-64. (d) FMNIST-128.

Fig. 6. 2-D embeddings for the test data of KMNIST with t-SNE, with each color denoting one class. (Best viewed in color.) (a) KMNIST and DLLP. (b) KMNIST and LLP-GAN.

training process. In particular, for the dataset KMNIST, DLLP, and LLP-VAT can converge quickly to reach a relatively good performance within few epochs, and then, their error rates keep stable. On the other hand, our method can ultimately obtain a better result, in spite of a little slower convergence. An explanation for this phenomenon is that it needs more iterations to reach an equilibrium between the generator and discriminator. Meanwhile, it again indicates that a better feature representation obtained by the generative adversarial mechanism can boost the downstream LLP performance. Note that LLP-VAT is a strong baseline, which also incorporates the adversarial mechanism as a regularization to LLP by minimizing the effect of adversarial samples. However, its improvement is relatively limited compared with DLLP, especially for small bag sizes, such as 64.

In practice, one good reason for introducing a generative adversarial mechanism to the classification model is that it can learn better feature representation for the downstream classification. To show this, we apply t-SNE [42] to visualize the embeddings of DLLP and LLP-GAN on the test data of KMNIST with the same model settings, as well as the bag size 64. More specifically, we extract the features in the last layer for both DLLP and LLP-GAN after 200 epoch training. From Fig. 6, the learned representations of LLP-GAN are more significantly concentrated to clusters, especially for the center part by comparing Fig. 6(a) with Fig. 6(b). This visualization is consistent with our analysis.

E. Results on CIFAR-10 and CIFAR-100

In this part, we perform DLLP, LLP-VAT, and LLP-GAN on the convolution-based network on the datasets CIFAR-10 and CIFAR-100.
and CIFAR-100. Specifically, CIFAR-10 and CIFAR-100 are computer-vision datasets generally used for object recognition with 60000 RGB images belonging to ten and 100 categories, respectively. A 13-layer convolution-based network is employed for these two datasets, and the generator is also designed based on deconvolution structure [29].

The final results are displayed in Table II with the best results in bold. Compared with the relatively less challenging tasks KMNIST and FMNIST, the advantage of our model is more impressive on CIFAR-10 and CIFAR-100, with significant improvement in test accuracy, especially under big bag size. In particular, there is a significant performance gap between DLLP and our method when the bag size is big. However, for smaller bag sizes, the improvement is very limited; even the performance is worse when using the adversarial mechanism. An explanation for this situation is that the generative adversarial mechanism hardly affects weakly supervised learning when the supervision information is slightly impaired. In the LLP scenario, when the bag size is small enough, the information on the label is almost equivalent to the fully supervised scenario. Nonetheless, it is worth challenging LLP models under big bag sizes because extremely limited supervision is in accordance with real-world applications.

Table II

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Bag Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>DLLP [1]</td>
<td>87.69 (0.54)</td>
</tr>
<tr>
<td></td>
<td>LLP-VAT [41]</td>
<td><strong>88.36 (0.29)</strong></td>
</tr>
<tr>
<td></td>
<td>LLP-GAN</td>
<td>86.97 (0.22)</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>DLLP [1]</td>
<td>61.95 (0.54)</td>
</tr>
<tr>
<td></td>
<td>LLP-VAT [41]</td>
<td><strong>65.21 (0.39)</strong></td>
</tr>
<tr>
<td></td>
<td>LLP-GAN</td>
<td>61.66 (0.49)</td>
</tr>
</tbody>
</table>

Note that the noise input is fixed for the corresponding generated images in Fig. 7(a) and (b) with only different epochs. The generated images demonstrate that our approach can stably learn a comparable generator. In other words, the generated samples suggest that the adversarial mechanism works well in our setting.

F. Real-World Application

To further investigate the performance of LLP-GAN in the real-world application, we evaluate our method on a real-life text dataset. Specifically, this dataset is related to a text classification task, which comes from a mobile phone notes app platform. The dataset is collected from many anonymous users, with each user providing several note data records. In total, we have 4000 records, and each record is related to one of 20 categories, including address, beauty, book, and education. However, when collecting the data, most of the users are not willing to offer the specific category information and are instead satisfied with providing packeted category proportions. On the other hand, we cannot annotate the data manually because they are already encoded into a numerical value. In other words, we are facing tabular data instead of the original text data. As a result of the above two reasons, we only have a weakly supervised dataset with label proportions, and the corresponding text classification task can be typically solved using LLP models.

The final test accuracy with respect to private fully supervised test data is displayed in Table III with the best results.
TABLE III
TEST ACCURACY RATES ON THE REAL-WORLD APPLICATION WITH DIFFERENT BAG SIZES

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Bag Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>D LLP [1]</td>
<td>90.12 (0.65)</td>
</tr>
<tr>
<td>Real</td>
<td>LLP-VAT [41]</td>
<td>90.21 (0.31)</td>
</tr>
<tr>
<td></td>
<td>LLP-GAN</td>
<td>90.47 (0.87)</td>
</tr>
<tr>
<td></td>
<td>Supervised model</td>
<td>92.88 (0.35)</td>
</tr>
</tbody>
</table>

![Accuracy rate convergence curves of LLP-GAN with different hyperparameters](image)

Fig. 8. Accuracy rate convergence curves of LLP-GAN with different hyperparameters, where the x-axis represents the epochs. (Best viewed in color.) (a) FMNIST. (b) KMNIST. (c) CIFAR-10. (d) CIFAR-100.

marked in bold. Similar to the previous results on benchmark datasets, LLP-GAN can consistently obtain a better performance compared with other methods, especially with larger bag sizes. In practice, the performance is close to that of the fully supervised situation under small bag sizes. In this way, we not only protect individual privacy but also provide relatively efficient information to attain a promising classification. Consequently, an important issue will be to control the bag size appropriately so that we can learn useful knowledge from the weak supervision in the data, as well as keeping the privacy intact.

G. Effect of Hyperparameter

In our model, only one hyperparameter $\lambda$ is involved, which is to trade off between the unsupervised learning of GAN and proportional supervised cross-entropy part, in the objective of the discriminator. Accordingly, we investigate the effect of different choices of the hyperparameter on the final performance. Specifically, we evaluate the performance under three orders of magnitude settings, including 0.1, 1, and 10 on FMNIST, KMNIST, CIFAR-10, and CIFAR-100 with the bag size 64. In general, harder task tends to be more sensitive to hyperparameter selection.

We provide the error rate convergence curves in Fig. 8, where different colors denote different $\lambda$ settings. At the first glance, we find that all four datasets are very sensitive to the hyperparameters with a significant performance gap with different $\lambda$’s. Furthermore, bigger $\lambda$ is prone to result in a better performance for all the four datasets.

To further analyze this issue, in Fig. 8(a) and (b), we provide convergence curves of FMNIST and KMNIST under different settings of $\lambda$ within 300 training epochs. When it comes to smaller $\lambda$, i.e., $\lambda = 0.1$, fluctuation of the curve is relatively severer and with an unsatisfying accuracy at the last epoch. On the other hand, when setting $\lambda$ as 1, for the FMNIST dataset, a comparable performance is achieved with $\lambda = 10$, while, for KMNIST, its performance (the cyan curve) is evidently between the green ($\lambda = 0.1$) and the red ($\lambda = 10$) curves.

When it comes to CIFAR-10 and CIFAR-100, in Fig. 8(c) and (d), we can observe a similar situation as Fig. 8(b). The performance improves along with $\lambda$ increasing. Besides, it also indicates that the convergence speed may be sensitive to the choice of $\lambda$, and it is hard to reach a convergence within 300 epochs when setting $\lambda$ as 1 for CIFAR-10 and CIFAR-100. To sum up, in most cases, the relative large hyperparameter, i.e., $\lambda = 10$, is a good choice, leading to an acceptable and stable performance within a limited training time.

This experiment indicates that the proportion loss is actually dominant to the final accuracy classification, and a small $\lambda$ may not guarantee the convergence because it hardly meets the requirement of proportional information, thus degrading the final performance. However, with too large $\lambda$, the generative adversarial mechanism may have little influence on the learning course, which will result in futile distribution estimation and inferior feature representation learning. Therefore, a proper $\lambda$ is extremely necessary to the final performance. Based on these considerations, we choose $\lambda = 10$ throughout this article.

H. Discussion on Experimental Results

Note that the new results are slightly different from the original paper from [25] for CIFAR-10 and CIFAR-100. This is mainly caused by the following two reasons. On the one hand, the final results in the current work are obtained within 300 or 400 epochs with data augmentation, while the original results in [25] are conducted with no data augmentation but more training epochs. In practice, we find that the convergence speed of LLP-GAN is relatively slow for CIFAR-10 and CIFAR-100, and a better performance can be obtained with much more training epochs. On the other hand, the bag distribution is also with some differences to [25] because we...
adopt a more reasonable data generation course, as introduced in [41].

Furthermore, we can observe that the advantage of LLP-GAN is more obvious when the tasks become more challenging. For example, the performance gap between our method with DLLP and LLP-VAT is apparently larger when the bag size is set as 128. Meanwhile, the advantage is more remarkable for CIFAR-10 and CIFAR-100, compared with the other two datasets, as the two CIFAR tasks are much harder. In particular, for the dataset CIFAR-100, DLLP becomes unacceptable when the bag size increases, while our approach can properly tackle these situations.

VI. CONCLUSION

This article proposed a new algorithm LLP-GAN for the LLP problem in virtue of the adversarial learning using a GAN structure. Consequently, our method is superior to existing methods in the following three aspects. First, it demonstrates nice theoretical properties that are innately in accordance with GANs. Second, LLP-GAN can produce a probabilistic classifier, which benefits from the generative model and meets the proportion consistency naturally. Third, on account of equipping CNNs, our algorithm is suitable for large-scale problems, especially for computer vision datasets. In addition, the experiments on four benchmark datasets and one real-world application verify all these advantages of our approach.

Nevertheless, limitations in our method can be summarized in four aspects. First, learning complexity in the sense of PAC has not been involved in this study. That is to say, we cannot evaluate the performance under limited data or without test data. Second, there is no guarantee of algorithm robustness to data perturbations, notably when the proportions are imprecisely provided. Third, varying GAN models (such as WGAN [4]) are not fully considered, and their performance is still unknown. In addition, in many real-world applications, the bags are built based on certain features, such as education levels and job titles, rather than randomly established. Hence, a practical issue will be to ensure good performance under more challenging nonrandom bag assignments. To overcome these drawbacks, shed light on the promising improvement of our current work.

ACKNOWLEDGMENT

Bo Wang would like to acknowledge that the main part of this research was conducted during his visit to Texas A&M University and thank Dr. Xia Hu for his hosting and insightful discussions.

REFERENCES


Jiahin Liu received the Ph.D. degree from the Department of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China, in 2019. He is currently an Assistant Professor with the School of Information and Electronics, Beijing Institute of Technology, Beijing. His research interests include weakly supervised machine learning and computer vision.

Bo Wang received the master’s degree from the Beijing Institute of Technology, Beijing, China, in 2010, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, in 2014. He was also a Visiting Scholar with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX, USA, in 2019. He is currently an Associate Professor with the School of Information and Technology and Management, University of International Business and Economics, Beijing. His principal research interests include statistical machine learning, optimization-based data mining, computer vision, and deep learning methods in finance.

Hanyuan Hang received the Ph.D. degree in statistical machine learning from the University of Stuttgart, Stuttgart, Germany, in 2015. He is currently an Assistant Professor with the Department of Applied Mathematics, University of Twente, Enschede, The Netherlands. His research papers have been published in the top machine learning conference, International Conference on Machine Learning (ICML), and several top statistics and machine learning journals, such as Journal of Machine Learning Research, Annals of Statistics, and Neural Computation. His main research interests lie in developing machine learning algorithms and their theory.

Huadong Wang received the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2017. He currently holds a post-doctoral position at the Department of Computer Science and Technology, Tsinghua University, Beijing. His research interests include machine learning and natural language processing.

Zhiquan Qi received the master’s and Ph.D. degrees from the College of Science, China Agricultural University, Beijing, China, in 2006 and 2011, respectively. He is currently an Associate Research Fellow with the Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing. His current research interests include deep learning and applications in weak label learning.

Yong Shi (Senior Member, IEEE) received the B.S. degree in mathematics from Shandong Normal University, Jinan, China, in 1994, the master’s degree in applied mathematics from the Beijing Institute of Technology, Beijing, China, in 1997, and the Ph.D. degree in management science and engineering from China Agricultural University, Beijing, in 2005. He is currently a Professor with the Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing. He has published four books on Support Vector Machines (SVMs).