This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3180545

Extended T: Learning with Mixed Closed-set and Open-set Noisy Labels

Xiaobo Xia, Bo Han, Nannan Wang, Jiankang Deng, Jiatong Li, Yinian Mao, Tongliang Liu

Abstract—The *noise transition matrix* T, reflecting the probabilities that true labels flip into noisy ones, is of vital importance to model label noise and build statistically consistent classifiers. The traditional transition matrix is limited to model *closed-set* label noise, where noisy training data have true class labels *within* the noisy label set. It is unfitted to employ such a transition matrix to model *open-set* label noise, where some true class labels are *outside* the noisy label set. Therefore, when considering a more realistic situation, i.e., both closed-set and open-set label noises occur, prior works will give *unbelievable* solutions. Besides, the traditional transition matrix is mostly limited to model instance-independent label noise, which may not perform well in practice. In this paper, we focus on learning with the mixed closed-set and open-set noisy labels. We address the aforementioned issues by extending the traditional transition matrix to be able to model mixed label noise, and further to the cluster-dependent transition matrix to better combat the instance-dependent label noise in real-world applications. We term the proposed transition matrix as the cluster-dependent extended transition matrix. An unbiased estimator (i.e., extended T-estimator) has been designed to estimate the cluster-dependent extended transition matrix by only exploiting the noisy data. Comprehensive experiments validate that our method can better cope with realistic label noise, following its more robust performance than the prior state-of-the-art label-noise learning methods.

Index Terms—noise transition matrix, mixed noisy labels, instance-dependent label noise, deep clustering, robustness

1 Introduction

The success of deep networks largely relies on large-scale datasets with high-quality label annotations [1], [2], [3]. However, it is quite costly, time-consuming, or even infeasible to collect such data. Instead, in practice, many large-scale datasets are collected in cheap ways, e.g., from search engines or web crawlers. The obtained data in these ways inevitably contain noisy labels [4]. The presence of noisy labels adversely affects the model prediction and generalization performance [3], [5]. It is therefore of great importance to train deep networks robustly against noisy labels.

The types of noisy labels studied so far can be divided into two categories: *closed-set* and *open-set* noisy labels. The closed-set noise occurs when instances with incorrect labels have true class labels *within* the noisy label set [1]. Oppositely, the open-set noise occurs when instances with incorrect labels have true class labels *outside* the noisy label set [1]. Learning with closed-set noisy labels has been extensively studied, e.g., [6], [7], [8], [9]. In addition, there are some pioneer works focusing on learning with open-set noisy labels, e.g., [1], [10], [11]. All these methods are designed for handling the closed-set and open-set noisy labels *independently* and cannot handle the *mixed* closed-set and open-set noisy labels well. Nevertheless, it is more practical that the two types of noisy labels *exist simultaneously* in real-world applications. For example, many large-scale face recognition datasets are automatically collected via image search engines and crowdsourcing platforms. The face data in these datasets contain both two types of noisy labels [12].

One promising strategy for combating label noise is to *model label noise*. Compared with *model-free* methods which empirically work well but do not model label noise explicitly, *model-based* methods are more reliable, as optimal classifiers w.r.t. clean data are guaranteed [13], [14]. By utilizing the *noise transition matrix* which denotes the probabilities that clean labels flip into noisy ones, model-based methods have been verified to be able to deal with closed-set noise well, mainly with the kind of *class-dependent (instance-independent) closed-set label noise* [7], [15].

However, prior model-based methods have the following limitations, which make it hard for them to work well in practice. First, they cannot model open-set label noise and will provide *unbelievable* solutions when there exist mixed open-set and closed-set noise at the same time. Second, *instance-dependent label noise* is common in real-world applications as difficult instances are prone to have inaccurate labels [16]. It is *ill-posed* to learn the instance-dependent transition matrix by only exploiting the noisy training data as discussed in [14]. Therefore, when modeling instance-dependent label noise, the class-dependent transition matrix is always exploited to approximate the instance-dependent transition matrix. Unfortunately, the approximation error is large, especially when the label noise rate is high [14].

In this paper, we present a novel method for learning with the mixed (instance-dependent) closed-set and open-set label noise. The proposed method extends the traditional

- *X. Xia and T. Liu are with the Trustworthy Machine Learning Lab, School of Computer Science, Faculty of Engineering, University of Sydney, Australia (e-mail: xxia5420@uni.sydney.edu.au; tongliang.liu@sydney.edu.au).*
- *B. Han is with the Department of Computer Science, Hong Kong Baptist University, China (email: bhanml@comp.hkbu.edu.hk).*
- *N. Wang is with the State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, China (email: nnwang@xidian.edu.cn).*
- *J. Deng is with the Department of Computing, Imperial College London, UK (email: j.deng16@imperial.ac.uk).*
- *J. Li and Y. Mao are with Meituan, China (email: lijiatong.lee@outlook.com; maoyinian@meituan.com).*

Manuscript received April 19, 2005; revised August 26, 2015.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Imperial College London. Downloaded on June 17,2022 at 14:21:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3180545

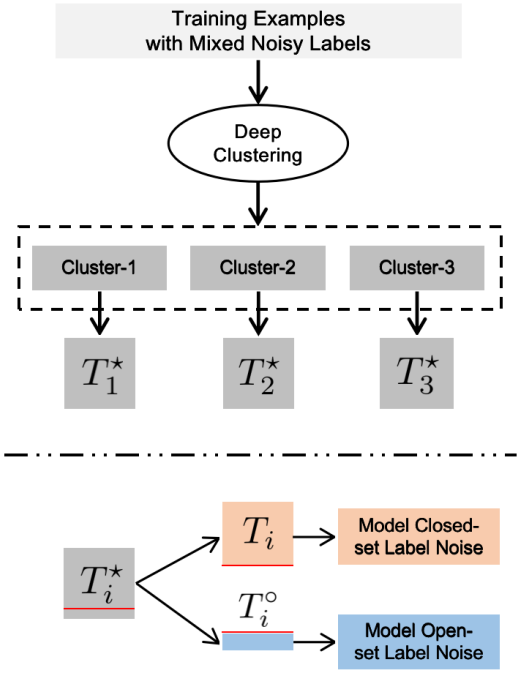


Fig. 1: The illustration of the *cluster-dependent extended transition matrices*. We first conduct clustering on *deep representations* of training examples and obtain different clusters (the top of this figure). Then the proposed method learns the transition matrix for different clusters and extends the traditional transition matrix to be able to model mixed label noise (the bottom of this figure). The transition matrices T_i and T_i° are concatenated vertically to form the extend transition matrix $T_i^\star$.

transition matrix to be able to model the mixed label noise and better approximate the instance-dependent label noise. Specifically, as all examples with open-set label noise have *out-of-distribution instances*, and we do not need to detect specific classes for them, we integrate all open-set classes as a *meta class*, which is paratactic with the other true classes in the closed set. Then we identify *anchor points* belonging to the meta class of the open-set and the true classes of the closed set. The extended transition matrix involving the meta class can be *unbiasedly* estimated by exploiting anchor points.

To further handle the instance-dependent label noise in reality, we exploit *cluster-dependent* transition matrices to better approximate the instance-dependent transition matrix. Specifically, we divide all training examples into several clusters (with the constraint that the cluster contains anchor points for the meta class of the open set and true classes of the closed set). The cluster-dependent transition matrix can then be unbiasedly estimated for each cluster. The training examples within the same cluster will share the same cluster-dependent transition matrix. The cluster-dependent transition matrices capture the geometric information of instances and thus can better approximate the instance-dependent transition matrix than the class-dependent transition matrix. The illustration of the proposed method is provided in Fig. 1.

1.1 Contributions

Before delving into details, we highlight the main contributions of this paper from three folds:

- We focus on learning with the mixed closed-set and open-set noisy labels and extend the traditional transition matrix to be able to model the mixed label noise, which solves the open problem in [17].
- We propose the cluster-dependent extended transition matrices to handle instance-dependent label noise in real-world applications, which produces a more reliable solution.
- We conduct comprehensive experiments on synthetic and real-world label-noise datasets to demonstrate that the proposed method achieves superior robustness over the baselines.

1.2 Organization

The rest of the paper is organized as follows. In Section 2 , we review related works on learning under label noise. In section 3, we introduce some notations and background knowledge. In Section 4, we introduce the proposed method in detail. Experimental results and analyses are provided in Section 5. Finally, we conclude the paper in Section 6. To improve readability, additional instructions and experimental results are provided in supplemental materials.

2 Related Work

In this section, we review prior works about learning with noisy labels (Section 2.1) and deep clustering (Section 2.2).

2.1 Learning with noisy labels

2.1.1 Learning with the noise transition matrix

The noise transition matrix plays an essential role in modeling label noise, which reflects the probabilities that true labels flip into other noisy ones. With the noise transition matrix, we can infer the clean class posterior probability with the noisy class posterior probability [6]. Thus, we can assign clean labels for given instances, even though only noisy training data are available. Lots of advanced methods borrow this idea and estimate the noise transition matrix to combat *closed-set* noisy labels [13]. Moreover, in order to reduce the estimation error of the noise transition matrix, a slack variable can be introduced to revise the initialized transition matrix [14], [18]. An intermediate class can be used to avoid directly estimating the noisy class posterior [7]. Besides, with a small trusted dataset, meta learning can be further employed [15].

For learning with *open-set* noisy labels, true classes of some training data are *outside* the set of known classes. Recall the definition of the (traditional) transition matrix, the flip probabilities indicate the rates of the true classes flipped to the noisy ones. If we use the method of modeling closed-set label noise to model open-set label noise, we will mistakenly treat some unknown incorrect classes as true classes, which leads to poor classification performance. To the best of our knowledge, how to effectively model the mixed label noise in this situation is a new challenge, and there is no pioneer to solve the above problems.

2.1.2 Other methods of learning with noisy labels

We first briefly introduce other methods for dealing with closed-set and open-set noisy labels *separately* without modeling the noise explicitly, which include sample selection

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Imperial College London. Downloaded on June 17,2022 at 14:21:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3180545

[10], [19], [20], [21], reweighting examples [1], [22], [23], designing robust loss functions [8], [24], and (implicitly) adding regularization [25], [26], [27], etc.

We then introduce the pioneer works for dealing with mixed closed-set and open-set noisy labels without modeling the noise explicitly. EvidentialMix [17] focuses on synthetic mixed noisy labels and achieves promising performance by combining DivideMix [9] and the SL loss [28]. Note that our work is fundamentally different from EvidentialMix. The main reasons are as follows: (*i*) EvidentialMix combines several advanced approaches, but our work focuses on one, i.e., learning with the noise transition matrix; (*ii*) EvidentialMix works well, but does not model label noise. Our work models the mixed label noise explicitly and improves the reliability of the method.

2.1.3 Class-dependent noise vs instance-dependent noise

For closed-set noisy labels, class-dependent noise assumes that the label flip process only depends on the latent clean class of the instance. However, such an assumption is somewhat strong. Instead, instance-dependent noise is more practical, where the label flip process also depends on the instance. For example, in real-world datasets, an instance whose feature contains less information or is of poorer quality may be more prone to be labeled wrongly. Unfortunately, the case of instance-dependent noise has been less studied than class-dependent one [14]. We suggest that readers can refer [3] for more details of learning with noisy labels.

2.2 Deep clustering

As an unsupervised learning method, clustering has been widely used in various tasks. It aims to keep similar data points in the same cluster while dissimilar ones in different clusters. Clustering is proven to be able to find *representative data points* among all data points [29]. Benefitting from the power of deep learning, lots of approaches boost traditional clustering techniques, e.g., k-means and spectral clustering, by using deep models. They cluster on deep representations instead of original features as deep representations are lower dimensional and have a higher degree of discrimination [30].

3 Preliminaries

3.1 Problem definition

We consider a c-class classification problem. Let $\mathcal{X}$ and $\mathcal{Y}$ be the instance and label spaces, where $\mathcal{Y} = \{1, \dots, c\}$. We define the clean joint distribution of a pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ as $\mathcal{D}$. The training sample $S = \{(x_i, y_i)\}_{i=1}^{n}$ is drawn from $\mathcal{D}$, where n is the sample size.

For the *closed-set* noise problem with the noise rate $\alpha \in (0, 1)$, the examples in S are mislabeled with probability ξ. Incorrect labels is still within the label space $\mathcal{Y}$. For the *open-set* noise problem with the noise rate $\beta \in (0, 1)$, we need to define a new training set S' (with $S' \cap S = \varnothing$), where the label space for S' is represented by $\mathcal{Y}'$ (with $\mathcal{Y}' \cap \mathcal{Y} = \varnothing$)- this means that the instances in S' do not have labels in $\mathcal{Y}$. When learning with open-set noisy labels, a proportion β of instances in S are replaced with the instances in S'.

For the *mixed closed-set and open-set* noise problem, τ, $\rho \in (0, 1)$ is defined by mixing the two kinds of noise above. Specifically, a proportion τ of training examples in S are mislabeled. Among them, a proportion $\tau \times (1 - \rho)$ of training examples are corrupted by closed-set noise, and a proportion $\tau \times \rho$ of training examples are corrupted by open-set noise. We define a pair of random variables relating to noisy examples as $(X, \tilde{Y})$. Our aim is to train a robust classifier against mixed closed-set and open-set noisy labels, which can assign labels accurately to test data.

3.2 Inference with the noise transition matrix

We formally introduce the traditional transition matrix, i.e., $T \in [0, 1]^{c \times c}$, which is only capable of modeling the closed-set noise. The transition matrix generally depends on the instances and the true labels, i.e., $T_{ij}(x) = P(\tilde{Y} = j | Y = i, X = x)$ [16]. Note that the noisy class posterior $P(\tilde{Y}|X)$ can be estimated by using the noisy training data. With the transition matrix, we can bridge the noisy class posterior $P(\tilde{Y}|X)$ and the clean class posterior $P(Y|X)$ as follows:

$$P(\tilde{Y} = j | X = x) = \sum_{i=1}^{c} T_{ij}(x) P(Y = i | X = x). \quad (1)$$

Namely, when learning with noisy labels, if we have the access to the ground-truth transition matrix, we can infer $P(Y|X)$ with $P(\tilde{Y}|X)$. The transition matrix can be therefore used to build *statistically consistent* algorithm [6], [18]. Unfortunately, the instance-dependent transition matrix is unidentifiable without any assumption [14], [18]. The existing methods approximate the instance-dependent transition matrix by assuming that the noise transition matrix is *class-dependent* and *instance-independent*, i.e., $T(x) = P(\tilde{Y} = j | Y = i, X = x) = P(\tilde{Y} = j | Y = i)$. When there is no confusion, we will short-hand $T(x)$ as T for the class-dependent transition matrix. Moreover, the traditional transition matrix fails to handle the mixed closed-set and open-set noise since it *encodes no open-set class information*. We will discuss how to extend it to better handle the *instance-dependent mixed closed-set and open-set noise* in the next section.

4 Method

In this section, we first discuss how to model the mixed label noise by extending the traditional label noise transition matrix (Section 4.1). Then we present how to learn the cluster-dependent extended transition matrices for better handling the instance-dependent mixed label noise (Section 4.2). Finally, we show how to exploit the extended transition matrix to train a robust classifier (Section 4.3).

4.1 Class-dependent extended T

As stated in Section 3.2, the traditional noise transition matrix is a $c \times c$ matrix linking the noisy class information to the closed-set clean class information without considering the open-set clean class information. It is therefore limited to handling the open-set label noise problems. Taking the traditional class-dependent transition matrix as an example, we discuss how to extend it to handle the open-set label noise. Note that in the next subsection, we will discuss how to handle the instance-dependent mixed label noise.

Integration with a meta class. To model the open-set label noise, we introduce a *meta class* which is an integration

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Imperial College London. Downloaded on June 17,2022 at 14:21:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3180545

of all the possible open-set classes. The philosophy is that compared with the examples with closed-set label noise, we do not have to detect specific classes for the examples with the open-set label noise. Therefore, we integrate all open-set classes as a meta class. As shown in Fig. 1, we extend the traditional transition matrix to $(c+1) \times c$ dimensional, where the additional $1 \times c$ vector denoted by T° represents how the meta class (or the open-set class) flips into the closed-set classes, i.e., $P(\tilde{Y}=j|Y=m,X=x)$, where $j=1,...,c$, and m represents the meta class label. The extended transition matrix *encodes the open-set class information* and can be exploited to better reduce the side-effect of the open-set label noise.

Matrix estimation by anchor points. We then discuss how to estimate the extended transition matrix by exploiting *anchor points*. Anchor points are defined in the clean data domain [6]. Formally, an instance x is an anchor point for the class i if $P(Y=i|X=x)$ is equal to one or approaches one. Given an anchor point x, if $P(Y=i|X=x)=1$, we have that for $k \neq i$, $P(Y=k|X=x)=0$. Then, we have,

$$P(\tilde{Y}=j|X=x)=\sum_{k=1}^{c} T_{kj}P(Y=k|X=x)=T_{ij}. \quad (2)$$

The equation holds because the transition matrix is assumed to be class-dependent and instance-independent, i.e., $T_{ij}(x)=T_{ij}$. Therefore, the transition matrix T can be *unbiasedly* estimated via estimating the noisy class posteriors for the anchor point of each class (including the meta class). Note that the anchor point assumption is widely adopted in the literature of learning with noisy labels [6], [7], [13], [31]. We could follow them and assume the availability of anchor points. However, the assumption that anchor points are given may be strong for many real-world applications. We could relax the assumption by just assuming that the anchor points exist in the training data and then design algorithms to locate them. For the closed-set classes, corresponding anchor points can be detected effectively as did in [13], [18]. The main challenge we face is how to locate anchor points belonging to the meta class (or the open-set classes).

Anchor point detection. Prior work [32] has confirmed that deep representations even when trained with noisy labels still exhibit *clustering properties*, i.e., deep networks learn embeddings that tend to group clean examples of the same classes into the same clusters while pushing away the examples with corrupted labels outside these clusters. Note that the obtained deep representations are not sufficient for learning the class posteriors. Also, deep clustering has been verified to be effective for detecting *representative data points* among all data points, e.g., the *cluster centroid*. By the definition of anchor points, they are the *representatives* of each class, i.e., *they belong to specific classes surely*. Therefore, we exploit deep clustering and determine that the anchor points are the data points which are close to the centroid of the meta class cluster. Note that the proposed method is basically different from the previous work [33]. Specifically, the previous method exploits clusterable representations of features and uses up to third-order consensuses of noisy labels among neighbor representations to take the place of anchor points. In a contrast, our method exploits clustering to find the underlying anchor points.

Algorithm 1: Meta Class Detection Algorithm

```
Input: Clustering results on deep representations,
  corresponding noisy labels of deep representations;
for i = 1, ..., c do
    Detect the cluster if the maximum number of the
      class label in it is i;
    Assign the class label i to this cluster;
    Remove the cluster that have a corresponding
      class label from the detection queue;
end
4: Output: The cluster including training examples
  with the meta class.
```

In this paper, we utilize the deep k-means cluster technique [34] to detect anchor points. Given the training sample $(x_1,\ldots,x_n)$, the number of clusters is set to $k=c+1$. The reason is that we have integrated all the possible open-set classes as a meta class. There are total $c+1$ classes in the training set. With an initialized deep network Ψ, we can obtain deep representations of the training sample $(\Psi(x_1),\ldots,\Psi(x_n))$. Here, we use a noisy validation set to obtain Ψ that has the highest accuracy on the noisy validation set. Note that the clean labels are dominating in noisy classes and that noisy labels are random, the accuracy on the noisy validation set and the accuracy on the clean test data set are positively correlated. The noisy validation set can therefore be used. Then, deep representations of instances can be obtained with the selected Ψ. In this way, we can obtain robust deep representations for detecting meta classes and anchor points. We use clustering on such deep representations since they are lower dimensional and have a higher degree of discrimination [30]. For k-means, we formulate the loss function as:

$$\ell_k=\sum_{i=1}^{n}\sum_{k=1}^{c+1} M_{ik}\|\Psi(x_i)-\mu_k\|_2^2, \quad (3)$$

where M is the cluster matrix with $M_{ik}=1$ if $\Psi(x_i)$ belongs to the k-th cluster, otherwise $M_{ik}=0$. The symbol μ_k represents the k-th cluster centroid. After clustering, we use an iterative strategy to assign class labels for obtained clusters. Specifically, as for a cluster that includes noisy training examples, correct labels in it are still *diagonally dominant* [32]. Thus, we can accurately assign class labels to each cluster based on the maximum number of the class label in this cluster. When we finish assigning closed-set class labels to c clusters, we regard the remaining one as the cluster of training examples with the meta class. We provide visualization of clustering results in Section 5.2.4 to verify the effectiveness of this way and support our claims. The algorithm flow of determining the meta class is provided in Algorithm 1.

In the following, to detect the anchor points belonging to the closed-set classes, we follow [13], [14], [18]. Then, we use these anchor points to estimate the transition matrix T for modeling the closed-set label noise with Eq. (2). For detecting the anchor points belonging to the meta class, we determine that the anchor points are the data points that are *close to the centroid of the meta class cluster*, which has been explained above from the perspective of representatives. As for the

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Imperial College London. Downloaded on June 17,2022 at 14:21:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3180545

anchor point detection, we only rely on the data points that are close to the centroid of the meta class cluster. Thus, we do not need that clustering can perform perfectly for further assigning labels, which is too strict for complex data. Then with Eq. (2), we use the noisy class posterior probabilities of the anchor point to estimate the transition matrix for modeling the open-set label noise. We denote the transition matrix for the open-set label noise as $T^{\circ} \in [0,1]^{1\times c}$.

When the estimation of the transition matrix for both types of label noise is finished, we combine T and T° to obtain the extended transition matrix $T^{\star} \in [0,1]^{(c+1)\times c}$ to model the mixed label noise.

4.2 Cluster-dependent extended T

We have presented how to model the mixed label noise by using the class-dependent extended transition matrix and how to estimate the extended transition matrix. However, in the real-world, label noise is more likely to be instance-dependent [14]. To handle this problem, we propose to use cluster-dependent extended transition matrices to better model the instance-dependent label noise, which is based on the intuition that *the instances which have similar features are more prone to have a similar label flip process* [14], [16]. We thus can employ the same extended transition matrix to model the mixed label noise for the instances which have similar features. We term such extended transition matrices as cluster-dependent extended transition matrices.

We now show how to learn the cluster-dependent transition matrices as follows. Consider the training examples $(x_1,\ldots,x_n)$, we cluster on their deep representations, i.e., $(\Psi(x_1),\ldots,\Psi(x_n))$ again to obtain different clusters. The total number of the clusters is set to a small number, i.e., z. Note that in Section 4.1, we have discussed how to detect anchor points of closed-set classes and meta classes. Since we need to estimate the cluster-dependent transition matrix for each cluster, after deep clustering and obtaining z clusters, we need to ensure that there are anchor points for each classes in each cluster for accurate estimation. Therefore, when we set the value of z for clustering, we need to ensure the existence of anchor points in each cluster with such a value of z. The overall procedure to learn the cluster-dependent extended transition matrices is summarized in Algorithm 2. After that, the training examples within the same cluster will share the same cluster-dependent transition matrix.

Algorithm 2: Cluster-dependent Transition Matrices Learning Algorithm

Input: Noisy training sample S_t, noisy validation sample S_v, the number of cluster-dependent transition matrices z;
Train a deep model by using the noisy data $\mathcal{S}_t$ and $\mathcal{S}_v$;
Get the deep representations of the examples by employing the trained deep network;
Detect the meta class as shown in Algorithm 1;
Detect anchor points used for estimation with clustering;
Cluster on the deep representations of the examples to obtain z clusters;
for $i = 1,\ldots,z$ **do**
  Estimate T_i for the closed-set label noise;
  Estimate T_i° for the open-set label noise;
  Obtain the cluster-dependent transition matrix $T_i^{\star}$ as discussed in Section 4.1;
end
Output: $T_1^{\star},\ldots,T_z^{\star}$.

4.3 Learning with importance reweighting

In the previous subsection, we have presented the methods about how to learn the cluster-dependent extended transition matrices to model the mixed instance-dependent label noise. With the learned extended transition matrix $T_i^{\star}(i=1,\ldots,z)$, we employ the *importance reweighting* technique [6] to train a robust classifier against mixed noisy labels. For the c-class classification problem under the mixed label noise, by exploiting the cluster-dependent extended transition matrices $T^{\star}$ (we hide the index for simplifying), the empirical risk can be formulated as:

$$\tilde{R}_{\tilde{\ell},n} = \frac{1}{n}\sum_{i=1}^{n} \frac{g_{\tilde{y}_i}(x_i)}{(T^{\star\top}g)_{\tilde{y}_i}(x_i)}\ell(f(x_i),\tilde{y}_i), \tag{4}$$

where $\ell : \mathbb{R}^c \times \mathcal{Y} \to \mathbb{R}_+$ is a surrogate loss function for c-class classification, e.g., the *cross-entropy loss*. Here, $g(x)$ is the output of the softmax layer. We use $\arg\max_{j\in\{1,\ldots,c\}} g_j(x)$ to assign labels for the test data. Note that during training, the $T^{\star}$ is determined according to the cluster to which the example x_i belongs. As we detect the anchor points from the noisy training data, as did in [13], [14], [18], data points that are similar to anchor points will be detected if there are no anchor points available. Also, deep networks may have poor confidence calibration. Then, the cluster-dependent extended transition matrices will be poorly estimated. To handle the issues, we follow [18] to revise the cluster-dependent extended transition matrices, which helps lead to a better classifier. We term the systemic proposed method for training a robust classifier against mixed label noise as *Extended T*. In more detail, Extended T-i means that the number of the cluster-dependent transition matrices is set to i. In the next section, we show that the proposed method can cope with mixed closed-set and open-set noisy labels well.

5 Experiments

In this section, we first introduce the methods for comparison in the experiments (Section 5.1). We then introduce the details of the experiments on synthetic datasets (Section 5.2). The experiments on real-world datasets are finally presented (Section 5.3 and Section 5.4).

5.1 Comparison methods

We compare the proposed method with the multiple advanced methods: (1) CE, which trains the deep models with the standard cross entropy loss on noisy datasets. (2) GCE [11], which handles label noise by exploiting the negative Box-Cox transformation. (3) PCE [8], which boosts the standard cross entropy loss with a partial trick. (4) PGCE [8], which boosts GCE with a partial trick. (5) APL [35], which combines two robust loss functions which boost each other.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Imperial College London. Downloaded on June 17,2022 at 14:21:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3180545

(6) DMI [36], which handles label noise from the perspective of the information theory. (7) NLNL [37], which proposes a novel learning method called Negative Learning (NL) to reduce the side effect of label noise. (8) Co-teaching [20], which trains two networks simultaneously and exchanges the selected examples for network updating. (9) Co-teaching+ [10], which trains two networks simultaneously and finds confident examples among the prediction disagreement data. (10) JoCor [38], which reduces the diversity of networks to improve the robustness. (11) S2E [4], which utilizes AutoML to handle label noise. (12) Forward [13], which estimates the class-dependent transition matrix to correct the training loss. (13) T-Revision [18], which introduces a slack variable to revise the estimated transition matrix and leads to a better classifier. Note that we do not compare with some state-of-the-art methods like SELF [39], DivideMix [9], and EvidentialMix [17]. It is because their proposed methods are aggregations of multiple advanced approaches, e.g., sample selection, semi-supervised learning, and co-training, while this work only focuses on one. Therefore, the comparison is not fair. We implement all methods with default parameters by PyTorch and conduct all the experiments on NVIDIA Tesla V100 GPUs.

5.2 Experiments on synthetic noisy datasets

5.2.1 Datasets and implementation details

Datasets. We evaluate the robustness of the proposed method on synthetic *CIFAR-10* [40], which is popularly used in learning with noisy labels. Original *CIFAR-10* consists of 50,000 training images and 10,000 test images with 10 classes. The size of images is 32×32×3. Note that the original *CIFAR-10* contains clean training labels. We thus corrupt the training data manually to generate noisy labels. Specifically, for class-dependent closed-set noise, we consider the symmetric noise in the main paper [13]. For instance-dependent closed-set noise, we borrow noise generation in [14]. For open-set noise, we follow [1] and borrow the images from *SVHN* [41], *CIFAR-100* [40], and *ImageNet32* (32×32×3 *ImageNet* images) [42] to act as outside images. Note that only images whose labels exclude 10 classes in *CIFAR-10* are considered. We then use the outside images to replace some training images in *CIFAR-10*. This corresponds to the setting of Type I in [1]. For Type II open-set noise, we consider the case where images damaged by Gaussian random noise.

For the mixed noise that consists of class-dependent closed-set noise, the overall noise rate τ ranges in $\{0.2, 0.4, 0.6, 0.8\}$. The proportion ρ of open-set noise ranges in $\{0.25, 0.5, 0.75\}$. For the mixed noise that consists of instance-dependent closed-set noise, the overall noise rate τ ranges in $\{0.2, 0.4, 0.6\}$. The range of the proportion ρ is not changed. The purpose of doing so is to ensure that clean labels is diagonally dominated in noisy classes [14], [16]. As we focus on learning with mixed noisy labels in this paper, we do not set ρ to be zero. We leave out 10% of noisy training data as a validation set. Note that even validation set is noisy, it still can be used for model selection effectively [39]. To avoid randomness, results are reported over five trials.

Implementation. We employ a PreAct ResNet-18 network [43]. For learning the transition matrix, we follow the optimization method in [13], [18]. We next use the SGD optimizer with momentum 0.9, batch size 128, and weight decay 5×10^{-4} to initialize the network. The initial learning rate is set to 10^{-2}, and divided by 10 after the 40th epochs and 80th epochs. 100 epochs are set totally. Following [18], we then exploit the Adam optimizer with a learning rate 5×10^{-7} to revise the transition matrix. Typical data augmentations including random crop and horizontal flip are applied.

We use up-to-three cluster-dependent transition matrices, i.e., Extended T-3. We consider it is possible that the proposed method show higher performance if we set a larger number of clusters, e.g., Extended T-5 or T-6. However, too many cluster-dependent transition matrices may make optimization difficult. It is hard to achieve an optimal choice of the number of used cluster-dependent transition matrices theoretically. In more detail, in the first perspective of data, it is difficult to achieve an optimal choice of the number of the clusters. In the second perspective of learning models, it is difficult to determine the optimal choice, since we need an accurate quantification for the model capacity and approximation power. The quantification is still mysterious for the whole community, even though learning with clean labels. In this paper, we empirically determine the number of cluster-dependent transition matrices to justify our claims.

5.2.2 Analyses of classification accuracy

We report comprehensive experimental results with *class-dependent* closed-set noise in Table 1 and 2 and results with *instance-dependent* closed-set noise in Table 3 and 4. More experimental results are provided in Appendix A.

For *CIFAR-10+SVHN* with class-dependent closed-set noise, we can clearly see that the proposed method consistently outperforms the prior state-of-the-art approaches for learning with mixed label noise. Specifically, in the cases of high label noise rates, e.g., $\tau = 0.6$ and $\tau = 0.8$, our method Extended T-3 achieves the best classification performance. This means that the proposed cluster-dependent transition matrices can reduce the estimation error brought by randomness and uncertainty in the estimation process, which helps lead to better classifiers. For *CIFAR-10+Gaussian* with class-dependent closed-set noise, we can see that when the noise rate is low, e.g., $\tau = 0.2$, DMI outperforms the proposed method slightly. When the noise rate increases, the proposed method consistently outperforms all baselines. To sum up, the synthetic experiments reveal that our method is powerful in handling mixed label noise, particularly in the case of high noise rates.

For *CIFAR-10+SVHN* with *instance-dependent* closed-set noise, in the case where $\tau = 0.2$ and $\rho = 0.25$, S2E achieves the best performance. In the other cases, the proposed method surpasses the baselines. The experimental results on *CIFAR10+Gaussian* is similar. Note that the proposed Extended T-3 almost outperforms Extended T and Extended T-2 all the time. Such results show that the cluster-dependent transition matrices can better approximate the instance-dependent transition matrices than the class-dependent transition matrix.

5.2.3 Discussion on classification accuracy

We further discuss the negative impact of closed/open-set label noise based on the above experimental results. As

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Imperial College London. Downloaded on June 17,2022 at 14:21:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3180545

Method τ	0.2			0.4			0.6			0.8		
ρ	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
CE	90.01 ±0.12	89.82 ±0.35	90.18 ±0.16	87.82 ±0.35	88.20 ±0.16	88.27 ±0.17	83.18 ±0.65	84.91 ±0.11	86.44 ±0.67	75.07 ±0.85	78.44 ±0.92	81.72 ±0.34
GCE	90.71 ±0.20	90.56 ±0.22	90.88 ±0.20	90.04 ±0.21	90.06 ±0.25	89.54 ±0.14	85.82 ±0.22	86.21 ±0.26	86.26 ±0.16	82.57 ±0.76	83.80 ±0.28	84.12 ±0.29
PCE	90.55 ±0.10	90.04 ±0.12	90.36 ±0.37	88.73 ±0.70	88.75 ±0.54	87.29 ±0.61	84.05 ±0.88	85.02 ±0.30	85.26 ±0.93	83.15 ±0.36	82.33 ±1.27	82.29 ±1.48
PGCE	90.62 ±0.10	90.04 ±0.28	90.46 ±0.42	89.22 ±0.36	90.17 ±0.83	90.16 ±0.50	85.77 ±0.42	86.01 ±0.67	86.04 ±1.09	80.36 ±1.44	82.63 ±0.92	83.12 ±0.97
APL	90.23 ±0.17	90.53 ±0.14	90.01 ±0.38	90.21 ±0.25	88.25 ±0.29	88.21 ±0.43	84.27 ±0.39	84.51 ±0.37	84.52 ±0.29	82.95 ±0.18	82.02 ±0.50	82.11 ±0.36
DMI	90.71 ±0.10	90.04 ±0.33	90.58 ±0.17	89.87 ±0.72	89.94 ±0.27	89.02 ±0.47	85.96 ±0.85	85.21 ±0.64	85.76 ±0.95	81.95 ±1.07	82.01 ±0.78	80.27 ±0.60
NLNL	89.85 ±0.19	89.72 ±0.32	89.98 ±0.27	87.39 ±0.28	87.80 ±0.50	88.78 ±0.29	82.90 ±0.55	84.58 ±0.41	84.37 ±0.72	76.39 ±0.59	79.65 ±0.54	78.39 ±1.08
Co-teaching	90.50 ±0.07	90.62 ±0.15	90.87 ±0.13	86.64 ±0.25	87.48 ±0.40	87.31 ±0.28	75.10 ±0.48	76.25 ±1.08	78.78 ±2.48	46.36 ±2.22	49.20 ±2.89	53.44 ±2.34
Co-teaching+	88.46 ±0.54	89.51 ±0.38	88.32 ±0.40	86.39 ±0.51	86.71 ±0.21	84.92 ±0.56	63.18 ±4.87	65.29 ±9.84	56.41 ±8.83	10.07 ±1.07	17.06 ±8.07	15.38 ±2.93
JoCor	88.25 ±0.06	89.15 ±0.21	89.15 ±0.45	84.16 ±1.07	82.12 ±1.03	82.02 ±0.74	67.29 ±1.23	69.02 ±1.72	71.70 ±1.73	43.93 ±0.32	42.82 ±1.31	40.12 ±3.44
S2E	89.42 ±1.35	89.68 ±1.13	89.87 ±1.80	88.24 ±2.48	88.99 ±1.94	88.78 ±1.57	81.16 ±2.49	85.44 ±1.62	85.48 ±2.32	57.45 ±4.17	74.16 ±4.52	78.39 ±3.46
Forward	89.37 ±0.14	89.27 ±0.86	89.64 ±0.80	87.54 ±0.24	87.76 ±0.54	87.01 ±0.39	80.19 ±2.72	83.21 ±0.89	83.92 ±1.98	78.05 ±2.02	80.32 ±1.84	78.66 ±1.72
T-Revision	90.23 ±0.14	89.97 ±0.23	90.02 ±0.14	88.68 ±0.24	88.79 ±0.29	89.02 ±0.47	85.07 ±1.03	85.37 ±1.09	85.42 ±0.83	81.04 ±2.04	81.36 ±0.97	82.98 ±1.17
Extended T	90.86 ±0.13	**90.89 ±0.28**	90.78 ±0.16	**90.94 ±0.22**	**90.72 ±0.37**	90.68 ±0.38	87.34 ±0.38	86.92 ±0.93	87.18 ±0.75	83.68 ±0.47	83.94 ±1.02	84.83 ±1.42
Extended T-2	**90.92 ±0.08**	90.58 ±0.54	**91.03 ±0.22**	90.73 ±0.28	90.54 ±0.30	90.42 ±0.51	87.32 ±0.77	87.03 ±1.07	87.08 ±0.99	84.02 ±0.76	84.02 ±1.08	84.77 ±1.53
Extended T-3	90.89 ±0.17	90.77 ±0.18	91.02 ±0.17	90.67 ±0.42	90.67 ±0.41	**90.72 ±0.69**	**87.48 ±0.63**	**87.19 ±0.92**	**87.29 ±0.73**	**84.42 ±0.93**	**84.31 ±0.94**	**84.88 ±1.07**

TABLE 1: Mean and standard deviations of test accuracies (%) on *CIFAR-10+SVHN* with *class-dependent* closed-set noise. The experimental results with the best mean are bolded.

Method τ	0.2			0.4			0.6			0.8		
ρ	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
CE	89.05 ±0.40	89.02 ±0.31	88.93 ±0.76	87.65 ±0.31	86.26 ±0.98	86.37 ±0.52	82.62 ±0.83	82.15 ±0.47	81.09 ±1.74	74.22 ±1.42	76.72 ±1.90	77.83 ±1.29
GCE	90.01 ±0.22	90.30 ±0.47	90.04 ±0.27	88.68 ±0.72	88.65 ±0.72	87.88 ±0.94	85.24 ±0.92	86.24 ±0.95	86.02 ±0.73	80.75 ±0.85	81.29 ±1.72	81.74 ±1.96
PCE	90.21 ±0.12	89.37 ±0.25	90.33 ±0.86	88.12 ±0.77	87.88 ±0.92	88.12 ±0.35	86.32 ±1.01	85.21 ±0.84	86.02 ±0.77	78.92 ±0.66	77.92 ±3.63	78.22 ±1.04
PGCE	90.05 ±0.27	90.92 ±0.37	90.22 ±0.63	89.38 ±0.26	89.05 ±0.29	88.68 ±0.74	86.05 ±1.82	87.02 ±0.77	85.36 ±0.92	80.12 ±1.25	81.09 ±2.71	80.20 ±2.67
APL	88.62 ±0.77	89.05 ±0.66	89.12 ±0.37	89.02 ±0.46	88.15 ±0.49	87.92 ±0.47	84.22 ±0.95	85.12 ±0.76	85.19 ±1.06	79.33 ±1.43	80.15 ±0.98	81.12 ±1.07
DMI	**90.45 ±0.13**	90.82 ±0.16	**91.05 ±0.28**	89.37 ±0.63	89.32 ±0.72	88.36 ±1.96	85.47 ±0.92	85.15 ±0.67	85.88 ±0.70	76.45 ±0.92	76.37 ±2.03	76.73 ±2.47
NLNL	89.32 ±0.27	89.34 ±0.58	88.79 ±0.47	87.21 ±0.26	87.05 ±0.34	86.92 ±0.80	83.11 ±0.38	83.15 ±0.67	82.18 ±0.94	75.24 ±0.73	78.63 ±0.62	77.45 ±1.57
Co-teaching	90.42 ±0.08	90.10 ±0.27	89.97 ±0.25	86.52 ±0.42	87.02 ±0.59	87.22 ±0.47	78.65 ±0.93	79.85 ±1.02	77.78 ±1.25	50.92 ±2.77	51.33 ±2.06	52.70 ±1.89
Co-teaching+	89.45 ±0.28	89.27 ±0.19	88.65 ±0.41	86.05 ±0.52	85.77 ±0.82	86.07 ±0.39	75.62 ±2.73	76.88 ±2.95	75.62 ±0.95	30.98 ±9.83	30.06 ±9.77	28.65 ±8.84
JoCor	88.15 ±0.27	88.45 ±0.53	88.29 ±0.27	81.22 ±0.73	82.16 ±0.83	83.05 ±1.04	66.27 ±0.95	68.22 ±1.93	67.45 ±1.80	42.86 ±2.36	43.71 ±3.05	44.05 ±1.52
S2E	90.21 ±0.17	89.37 ±0.98	89.92 ±0.40	88.27 ±1.73	88.43 ±1.63	87.93 ±1.99	80.59 ±4.06	81.22 ±4.24	81.77 ±3.81	70.54 ±5.64	72.88 ±5.13	76.46 ±3.95
Forward	88.25 ±0.32	89.02 ±0.45	89.10 ±0.96	87.24 ±0.77	86.73 ±0.92	86.30 ±1.00	84.15 ±0.46	84.16 ±0.68	84.25 ±0.82	77.58 ±2.63	78.94 ±2.72	78.92 ±3.02
T-Revision	89.26 ±0.88	89.65 ±0.39	89.67 ±0.47	89.05 ±0.34	88.56 ±0.45	88.06 ±0.92	85.67 ±1.03	85.26 ±0.90	85.21 ±0.79	80.25 ±1.91	80.77 ±2.60	80.73 ±2.41
Extended T	90.42 ±0.18	90.80 ±0.31	**90.30 ±0.43**	89.78 ±0.77	90.03 ±0.52	**90.05 ±0.62**	87.02 ±0.92	**88.24 ±1.02**	86.33 ±0.64	83.83 ±0.85	**84.52 ±1.90**	83.14 ±0.66
Extended T-2	90.33 ±0.63	**91.05 ±0.09**	91.02 ±0.18	89.23 ±0.74	89.73 ±0.72	90.02 ±0.72	**88.31 ±1.02**	88.01 ±1.17	86.52 ±1.01	**84.22 ±1.16**	83.92 ±0.97	83.92 ±1.09
Extended T-3	90.32 ±0.71	91.03 ±0.18	90.53 ±0.56	**90.01 ±0.17**	**90.03 ±0.71**	89.88 ±0.47	87.55 ±1.37	87.02 ±0.83	**86.67 ±0.89**	83.89 ±2.03	84.36 ±1.82	**84.05 ±1.43**

TABLE 2: Mean and standard deviations of test accuracies (%) on *CIFAR-10+Gaussian* with *class-dependent* closed-set noise. The experimental results with the best mean are bolded.

we can see, when we increase the proportion of the open-set noisy labels (with a fixed τ), there is no significant degradation for the classification performance of baselines and the proposed method. Additionally, in some cases, the classification performance is improved with an increasing ρ. We conduct some discussions for this.

For the training examples with closed- or open-set incorrect labels, they all have *out-of-distribution inputs*. Specifically,

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Imperial College London. Downloaded on June 17,2022 at 14:21:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3180545

Method	τ	0.2			0.4			0.6		
	ρ	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
CE		89.72 ±0.30	89.02 ±0.45	88.70 ±0.25	83.12 ±1.25	85.03 ±0.62	84.32 ±0.68	65.80 ±4.74	81.04 ±2.77	82.46 ±1.92
GCE		89.14 ±0.21	89.06 ±0.35	88.27 ±0.71	84.71 ±0.92	84.88 ±0.90	85.02 ±0.73	72.62 ±3.75	81.02 ±0.94	83.02 ±0.67
PCE		88.36 ±0.40	88.24 ±0.42	88.49 ±0.47	84.44 ±0.77	85.87 ±0.99	85.74 ±0.84	72.67 ±3.26	82.06 ±0.74	82.74 ±1.08
PGCE		89.02 ±0.16	88.02 ±0.72	87.69 ±1.46	84.06 ±1.30	85.22 ±1.06	85.62 ±0.88	70.64 ±4.79	81.05 ±1.92	82.02 ±1.45
APL		88.62 ±0.70	87.92 ±0.62	88.50 ±0.46	82.05 ±0.42	82.68 ±1.33	82.94 ±1.77	62.06 ±9.87	78.22 ±5.42	78.03 ±4.46
DMI		90.04 ±0.12	89.76 ±0.46	89.04 ±0.16	83.97 ±0.79	84.89 ±0.77	85.05 ±0.90	71.83 ±4.33	81.54 ±1.40	83.88 ±1.17
NLNL		87.33 ±0.62	88.16 ±0.64	88.79 ±0.75	79.22 ±2.52	81.22 ±0.79	82.02 ±0.83	57.14 ±8.72	76.88 ±3.76	77.27 ±4.02
Co-teaching		89.34 ±0.45	89.67 ±0.20	89.20 ±0.73	78.25 ±3.02	80.15 ±2.93	81.22 ±0.97	69.66 ±4.27	75.22 ±2.04	77.83 ±0.96
Co-teaching+		88.25 ±0.45	88.10 ±0.94	88.20 ±0.37	83.36 ±1.07	82.97 ±1.43	84.65 ±0.92	43.83 ±7.81	70.30 ±2.91	74.65 ±4.87
JoCor		88.01 ±0.25	87.04 ±0.82	88.07 ±0.42	80.22 ±0.52	80.50 ±1.02	81.62 ±0.77	65.02 ±2.04	73.63 ±1.95	75.28 ±0.95
S2E		**90.45 ±0.91**	89.78 ±1.43	90.04 ±1.73	85.62 ±1.71	86.04 ±1.90	86.02 ±1.41	70.29 ±6.20	81.95 ±3.09	82.09 ±2.71
Forward		88.01 ±0.46	88.12 ±0.29	88.10 ±0.62	83.12 ±1.17	84.15 ±1.26	85.14 ±0.76	68.90 ±1.52	79.33 ±3.66	81.37 ±2.42
T-Revision		89.47 ±0.76	89.52 ±0.83	89.66 ±0.44	86.15 ±0.92	86.17 ±1.24	86.92 ±1.07	73.06 ±2.96	83.98 ±0.95	84.77 ±2.10
Extended T		89.88 ±0.36	89.67 ±0.19	89.70 ±0.52	86.58 ±0.94	86.62 ±1.09	87.15 ±1.05	74.96 ±2.02	84.09 ±1.72	85.92 ±1.72
Extended T-2		89.93 ±0.19	90.01 ±0.43	90.14 ±0.36	86.72 ±0.78	87.01 ±0.82	87.31 ±0.90	75.33 ±2.01	84.55 ±0.78	85.94 ±1.29
Extended T-3		90.20 ±0.52	**90.10 ±0.47**	**90.25 ±0.38**	**87.26 ±1.28**	**87.39 ±1.21**	**87.58 ±1.37**	**76.01 ±1.33**	**84.77 ±1.09**	**86.00 ±1.90**

TABLE 3: Mean and standard deviations of test accuracies (%) on *CIFAR-10+SVHN* with *instance-dependent* closed-set noise. The experimental results with the best mean are bolded.

Method	τ	0.2			0.4			0.6		
	ρ	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
CE		87.33 ±0.21	88.90 ±0.73	89.25 ±0.38	84.33 ±0.70	85.02 ±0.77	84.06 ±1.05	67.92 ±4.09	80.05 ±3.05	81.25 ±2.32
GCE		89.20 ±0.37	89.06 ±0.11	89.27 ±0.53	84.17 ±0.37	85.06 ±0.83	85.04 ±0.95	73.40 ±3.08	80.24 ±1.27	81.33 ±1.64
PCE		88.92 ±0.30	88.92 ±0.31	89.06 ±0.78	84.64 ±0.96	85.32 ±0.71	84.92 ±0.46	73.25 ±2.37	81.06 ±3.08	80.25 ±2.93
PGCE		89.62 ±0.33	89.65 ±0.72	89.31 ±1.07	83.92 ±1.67	84.92 ±0.47	85.01 ±0.90	70.06 ±6.72	80.35 ±2.62	81.37 ±1.69
APL		88.09 ±0.87	88.21 ±0.39	89.05 ±0.92	83.26 ±0.72	84.64 ±1.58	84.72 ±1.21	65.82 ±5.09	78.66 ±1.81	79.65 ±1.93
DMI		89.03 ±1.02	89.33 ±0.40	89.63 ±0.59	84.79 ±0.61	84.63 ±0.98	84.61 ±0.92	71.82 ±3.77	82.45 ±2.94	82.77 ±2.61
NLNL		88.02 ±0.36	88.44 ±0.29	88.45 ±0.24	80.11 ±2.05	81.66 ±0.91	82.35 ±1.23	58.83 ±5.04	75.32 ±2.83	77.36 ±1.81
Co-teaching		89.93 ±0.20	89.76 ±0.74	90.01 ±0.46	82.35 ±1.10	80.12 ±1.17	81.15 ±0.94	67.35 ±4.31	73.62 ±0.95	78.68 ±1.62
Co-teaching+		88.26 ±0.35	89.65 ±0.42	89.33 ±0.59	82.79 ±2.04	79.82 ±3.17	82.33 ±2.51	50.77 ±6.63	72.77 ±3.59	75.29 ±5.46
JoCor		88.09 ±0.32	88.52 ±0.93	88.03 ±0.45	80.65 ±0.96	80.23 ±1.05	81.68 ±1.22	63.83 ±2.91	71.05 ±3.15	76.44 ±3.17
S2E		89.70 ±1.03	89.75 ±1.07	89.93 ±1.25	85.24 ±1.47	83.77 ±2.04	84.25 ±1.97	66.82 ±3.90	78.68 ±2.73	81.37 ±2.15
Forward		88.54 ±0.60	88.27 ±0.48	88.36 ±0.59	82.07 ±2.06	83.75 ±1.26	83.32 ±1.73	68.17 ±3.96	78.27 ±3.27	81.05 ±1.98
T-Revision		89.62 ±0.84	89.77 ±0.92	90.05 ±0.79	85.22 ±0.99	85.06 ±1.07	85.34 ±1.87	73.42 ±2.94	82.30 ±1.95	83.08 ±1.49
Extended T		89.77 ±0.26	90.05 ±0.64	90.08 ±0.62	85.39 ±0.65	86.08 ±1.24	85.38 ±1.27	73.92 ±2.89	83.17 ±1.42	83.45 ±1.63
Extended T-2		89.92 ±0.64	**90.19 ±0.73**	90.14 ±0.62	86.01 ±1.26	86.19 ±1.43	85.62 ±1.54	74.05 ±1.83	83.62 ±2.01	83.92 ±1.99
Extended T-3		**89.97 ±0.90**	90.10 ±0.58	**90.16 ±0.36**	**86.27 ±0.92**	**86.36 ±1.89**	**86.01 ±1.53**	**74.55 ±1.04**	**84.01 ±1.75**	**84.19 ±1.84**

TABLE 4: Mean and standard deviations of test accuracies (%) on *CIFAR-10+Gaussian* with *instance-dependent* closed-set noise. The experimental results with the best mean are bolded.

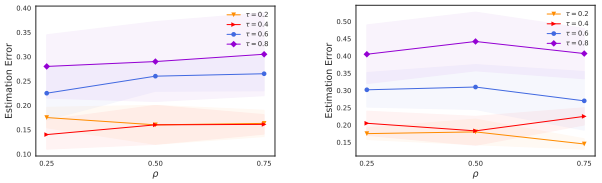
for the training examples with closed-set (*resp.* open-set) incorrect labels, they have out-of-distribution labels (*resp.* instances). Empirically, open-set noisy labels would be more harmful to degenerate the classification performance if deep models are severely overfitted [1]. This seems to a bit contradict the results in this paper. However, this contradiction comes from different experimental settings compared with [1]. To be specific, in this paper, we follow a standard machine learning paradigm [44] and use a noisy validation set for early stopping, which exploits the memorization effect of deep models [20], [45]. Although the examples with open-set incorrect labels may bring more serious degradation

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Imperial College London. Downloaded on June 17,2022 at 14:21:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3180545

to the performance if the networks severely overfit them, in the early training stage, e.g., before early stopping, the networks mainly fit the examples with clean labels [4], [45]. The examples with open-set incorrect labels are more likely identified as *outliers* and hard to be fitted. Therefore, they do not do much harm. Such explanations and analyses are similar to those in [1], which are mainly conducted with experiments.

5.2.4 Estimation and visualization results



(a) (b)

Fig. 2: Illustrations of the transition matrix estimation errors. Figure (a) illustrates the estimation error for modeling open-set label noise. Figure (b) illustrates the estimation error for modeling mixed label noise. The error bar for standard deviation in each figure has been shaded.

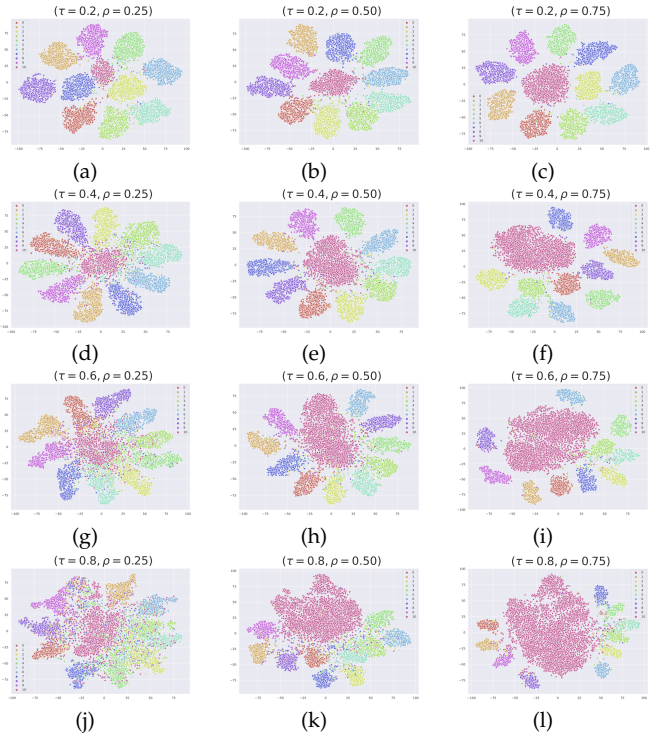


Fig. 3: Illustrations of the visualization results on the *CIFAR-10+SVHN* dataset with class-dependent closed-set noise.

We report the estimation error of the transition matrix T° for open-set noise, and of the transition matrix $T^{\star}$ for mixed noise. Note that only our method extends the traditional transition matrix to model the open-set noise and further model the mixed noise. Therefore, we do not report the estimation results of other model-based methods. The experiments are conducted on *CIFAR10+SVHN*. Here, we consider controlled class-dependent closed-set noise. It is because we can show the estimation results more clearly. The estimation errors are calculated with ℓ_1 norm. The results are presented in Fig. 2.

Note that we use clustering on deep representations as discussed. Here we visualize the results for better justification. The experiments are also conducted on *CIFAR-10+SVHN* with class-dependent closed-set noise. We exploit 2D t-SNE [46]. The visualizations are provided in Fig. 3. We simply set the meta class to class 10. We can see, under our settings, deep representations exhibit *clustering properties* [32]. In fact, for the proposed method, we do not need perfect clustering performance, as we only aim to detect a small number of anchor points belonging to the meta class, which are close to the centroid of the meta class cluster. It is possible that there are few data points belonging to the closed-set classes in the meta class cluster. We can choose multiple data points in the meta class cluster to estimate the transition matrix, which can alleviate the mentioned issue [31]. All experimental results verify the effectiveness of the proposed method and justify our claims well.

5.3 Experiments on *Clothing1M*

5.3.1 Implementation details

We verify the effectiveness of our method on the real-world noisy dataset *Clothing1M* [2]. *Clothing1M* has 1M images with real-world noisy labels, and 50k, 14k, 10k images with clean labels for training, validating, testing, but with 14 classes. Note that we do not use the 50k and 14k clean data in all the experiments, since it is more practical that there is no available clean data. For preprocessing, we resize the image to 256×256, crop the middle 224×224 as input, and perform normalization. We leave 10% noisy training data as a validation set for model selection.

Following prior works [7], [13], [18], we exploit a ResNet-50 pre-trained on ImageNet. As did in [14], we only exploit 1M noisy data to initialize the network and estimate the transition matrices. For initialization, we use SGD with momentum 0.9, weight decay 10^{-3}, batch size 32, and run with learning rates 10^{-3} and 10^{-4} for 5 epochs each. For revising the transition matrices, Adam is used and the learning rate is changed to 5×10^{-7}.

5.3.2 Experimental results

The experimental results on *Clothing1M* are shown in Table 5. As can be seen, the proposed method significantly outperforms the baselines. Moreover, the cluster-dependent transition matrices make the networks more robust against real-world noise. Extended T-3 can achieve a +1.24% improvement over Extended T.

5.4 Experiments on real-world face datasets

5.4.1 Datasets and implementation details

Training data. The training datasets include VggFace-2 [47] and MS1MV0 [48]. VggFace-2 is a noisy dataset collected from Google image search. MS1MV0 is a raw dataset with a large amount of noisy labels [49], [50].

Test data. We use four popular benchmarks as test datasets, including CFP [51], AgeDB [52], CALFW [53], and CPLFW [54]. For test data used for experiments on real-world face

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Imperial College London. Downloaded on June 17,2022 at 14:21:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3180545

Method	CE	GCE	PCE	PGCE	APL	DMI	NLNL	Co-teaching
Accuracy	68.88	69.45	69.48	69.93	54.46	70.12	43.92	67.94
Method	Co-teaching+	JoCor	S2E	Forward	T-Revision	Extended T	Extended T-2	Extended T-3
Accuracy	66.52	69.06	68.03	69.91	70.97	71.35	71.82	**72.59**

TABLE 5: Test accuracies (%) of different methods training on *Clothing1M*. The best results are bolded.

datasets, CFP consists of face images of celebrities in frontal and profile views. AgeDB contains images annotated with accurate to the year, noise-free labels. CALFW considers a more general cross-age situation and provides a face image set with large intra-class variations. CPLFW is similar to CALFW, but considers a cross-pose case. The important statistics of the datasets are summarized in Table 6.

	Datasets	#Identities	#Images
Train	VggFace-2 [47]	9.1K	3.3M
	MS1MV0 [48]	100K	10M
Test	CFP [51]	500	7,000
	AgeDB [52]	568	16,488
	CALFW [53]	5,749	12,174
	CPLFW [54]	5,749	11,652

TABLE 6: Face datasets for training and testing.

Data processing. We follow ArcFace [55] to generate the normalised face crops by exploiting five facial points (two eyes, nose tip, and two mouth corners) predicted by RetinaFace [56]. The size of the face crops is 112×112.

Network structure and activation function. To be fair, in this paper, we employ the same architecture and the activation function for testing different baselines. We use MobileFaceNet [57] and Arc-Softmax [55], which are popular in the face recognition task. The dimension of the face embedding feature is 512. For the angular margin m and feature scale s, we set 0.5 and 32, respectively.

Training and testing. At the training stage, we train the deep models with SGD with momentum 0.9, with a total batch size 512 on 4 GPUs parallelly and weight decay 5×10^{-4}. For learning an initial classifier, the learning rate is initially 0.1 and divided by 10 at the 5th, 10th, and 15th epochs. We set 20 epochs in total. For learning the classifier and slack variable, Adam is used and the learning rate is changed to 5×10^{-7}. At the test stage, we use MobileFaceNet to extract the 512-D feature embeddings of test face images. We follow the unrestricted with labelled outside data protocol [58] to report the verification performance on test face datasets.

5.4.2 Experimental results

We use two training datasets, i.e., VggFace-2 and MS1MV0, to separately train the deep networks. In Table 7 and 8, we show the results of the proposed method and baselines on CFP [51], AgeDB [52], CALFW [53], and CPLFW [54], respectively. We can observe that most of the results obtained by training on VggFace-2 are higher than the results on MS1MV0. It is because that MS1MV0 contains more noisy labels, and therefore is more challenging [12], [50]. To our method, we can effectively model the mixed noise, and thus can achieve higher performance than the baselines. Specifically, when training on VggFace-2, Extended T consistently outperforms the baselines. The improvement of classification performance brought by changing the number of clusters is not too obvious. It is because that the noise rate of VggFace-2 is low [55]. When training on MS1MV0, we can see that the proposed method surpasses the baselines again. Specifically, compared with Forward and T-Revision, the proposed method leads them clearly. Compared with the methods which empirically work well, the proposed method still outperforms them. It is worth noting that the cluster-dependent transition matrices bring a significant performance improvement. The improvement shows that the cluster-dependent transition matrices can model the complex label noise more accurately in this realistic scenario.

Method	CFP	AgeDB	CALFW	CPLFW	Ave.
CE	95.30	92.69	89.94	85.97	90.98
GCE	94.26	91.06	89.98	85.28	90.15
PCE	94.05	91.33	90.19	84.62	90.05
PGCE	93.92	91.25	90.39	84.77	90.08
APL	94.16	90.29	88.32	85.09	89.47
DMI	94.39	92.83	90.06	85.86	90.79
NLNL	86.74	87.63	84.79	80.06	84.81
Co-teaching	95.47	92.53	89.58	85.32	90.73
Co-teaching+	95.26	92.01	85.12	85.19	89.40
JoCor	92.32	90.97	84.58	82.11	87.50
S2E	92.09	91.64	89.70	84.63	89.52
Forward	95.07	92.40	89.10	85.79	90.59
T-Revision	95.38	92.79	89.86	85.94	90.99
Extended T	95.57	93.06	90.33	86.24	91.30
Extended T-2	95.59	**93.15**	90.42	86.36	91.38
Extended T-3	**95.73**	93.14	**90.67**	**86.52**	**91.51**

TABLE 7: Test accuracies (%) of different methods training on VggFace-2. The best results are bolded.

Method	CFP	AgeDB	CALFW	CPLFW	Ave.
CE	88.79	92.35	91.36	82.92	88.86
GCE	89.02	91.87	91.32	82.77	88.75
PCE	89.33	90.62	91.72	83.06	88.68
PGCE	90.36	91.25	91.94	83.75	89.33
APL	87.06	91.55	91.29	82.02	87.98
DMI	89.02	92.18	90.94	83.01	88.79
NLNL	83.06	85.47	84.21	74.02	81.69
Co-teaching	91.25	93.05	91.24	84.02	89.89
Co-teaching+	91.36	91.87	90.93	84.53	89.67
JoCor	86.16	89.30	88.09	79.84	85.85
S2E	91.52	91.61	91.03	81.47	88.91
Forward	90.02	92.32	91.28	82.95	89.14
T-Revision	90.29	92.49	91.60	83.82	89.55
Extended T	90.33	92.59	91.65	83.91	89.62
Extended T-2	91.65	93.05	92.30	84.67	90.42
Extended T-3	**92.08**	**93.71**	**92.61**	**85.31**	**90.93**

TABLE 8: Test accuracies (%) of different methods training on MS1MV0. The best results are bolded.

6 Conclusion

In this paper, we investigate into learning with mixed closed-set and open-set noisy labels, which is more practical but lacks systematic study in current works. We extend the traditional transition matrices to be able to model mixed label noise and exploit the cluster-dependent transition matrix to better handle the instance-dependent label noise. Empirical evaluations on synthetic and real-world datasets show the effectiveness of the proposed method for modeling label noise and leading to better classifiers. We believe that our work will urge the research community to explore the robustness of algorithms in this realistic noisy label scenario.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Imperial College London. Downloaded on June 17,2022 at 14:21:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3180545

REFERENCES

[1] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *CVPR*, pages 8688–8696, 2018.

[2] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, pages 2691–2699, 2015.

[3] Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W Tsang, James T Kwok, and Masashi Sugiyama. A survey of label-noise representation learning: Past, present and future. *arXiv preprint arXiv:2011.04406*, 2020.

[4] Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James T Kwok. Searching to exploit memorization effect in learning with noisy labels. In *ICML*, 2020.

[5] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

[6] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2016.

[7] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. In *NeurIPS*, 2020.

[8] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? In *ICLR*, 2020.

[9] Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2020.

[10] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W Tsang, and Masashi Sugiyama. How does disagreement benefit co-teaching? In *ICML*, 2019.

[11] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, pages 8778–8788, 2018.

[12] Xiaobo Wang, Shuo Wang, Jun Wang, Hailin Shi, and Tao Mei. Co-mining: Deep face recognition with noisy labels. In *ICCV*, pages 9358–9367, 2019.

[13] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, pages 1944–1952, 2017.

[14] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In *NeurIPS*, 2020.

[15] Jun Shu, Qian Zhao, Zengben Xu, and Deyu Meng. Meta transition adaptation for robust deep learning with noisy labels. *arXiv preprint arXiv:2006.05697*, 2020.

[16] Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance-and label-dependent label noise. In *ICML*, 2020.

[17] Ragav Sachdeva, Filipe R Cordeiro, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Evidentialmix: Learning with combined open-set and closed-set noisy labels. In *WACV*, 2020.

[18] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *NeurIPS*, pages 6838–6849, 2019.

[19] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, pages 2309–2318, 2018.

[20] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, pages 8527–8537, 2018.

[21] Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor W Tsang, and Masashi Sugiyama. Sigua: Forgetting may make learning with noisy labels more robust. In *ICML*, 2020.

[22] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, pages 4331–4340, 2018.

[23] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, pages 1919–1930, 2019.

[24] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *ICML*, 2020.

[25] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *CVPR*, pages 839–847, 2017.

[26] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *Advances in Neural Information Processing Systems*, pages 5596–5605, 2017.

[27] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *ICLR*, 2021.

[28] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *NeurIPS*, pages 3179–3189, 2018.

[29] Holger Teichgraeber and Adam R Brandt. Clustering methods to find representative periods for the optimization of energy systems: An initial framework and comparison. *Applied Energy*, 239:1283–1293, 2019.

[30] Xu Yang, Cheng Deng, Feng Zheng, Junchi Yan, and Wei Liu. Deep spectral clustering using dual autoencoder network. In *CVPR*, pages 4066–4075, 2019.

[31] Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. In *ECCV*, pages 68–83, 2018.

[32] Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, and Jinwoo Shin. Robust inference via generative classifiers for handling noisy labels. In *ICML*, pages 3763–3772, 2019.

[33] Zhaowei Zhu, Yiwen Song, and Yang Liu. Clusterability as an alternative to anchor points when learning with noisy labels. In *ICML*, 2021.

[34] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[35] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *ICML*, 2020.

[36] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *NeurIPS*, pages 6222–6233, 2019.

[37] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *ICCV*, pages 101–110, 2019.

[38] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, 2020.

[39] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling. In *ICLR*, 2020.

[40] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, 2009.

[41] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y.Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[42] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A down-sampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645, 2016.

[44] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2018.

[45] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. *arXiv preprint arXiv:1706.05394*, 2017.

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

[47] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, pages 67–74, 2018.

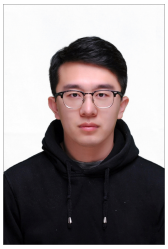
[48] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, pages 87–102, 2016.

[49] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *ECCV*, pages 765–780, 2018.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Imperial College London. Downloaded on June 17,2022 at 14:21:20 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3180545

[50] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *CVPR*, 2020.
[51] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *WACV*, pages 1–9, 2016.
[52] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *CVPRW*, pages 51–59, 2017.
[53] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.
[54] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5, 2018.
[55] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019.
[56] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, pages 5203–5212, 2020.
[57] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *CCBR*, pages 428–438, 2018.
[58] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008.



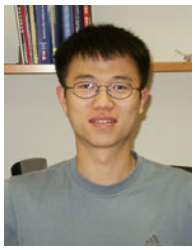
Xiaobo Xia received the B.E. degree in telecommunications engineering from Xidian University, in 2020. He is currently pursuing the Ph.D. degree in computer science from University of Sydney. He has published two papers on the NeurIPS conference (one spotlight), one paper on the ICLR conference, and one paper on the ICML conference as the first/co-first author. He also serves as the reviewer for top-tier conferences such as ICML, NeurIPS, ICLR, CVPR, and IJCAI. His research interest lies in machine learning, with a particular emphasis on weakly-supervised learning and kernel methods.

Bo Han is an Assistant Professor of Computer Science at Hong Kong Baptist University, and a Visiting Scientist at RIKEN Center for Advanced Intelligence Project (RIKEN AIP). He was a Postdoc Fellow at RIKEN AIP (2019-2020). He received his Ph.D. degree in Computer Science from University of Technology Sydney (2015-2019). During 2018-2019, he was a Research Intern with the AI Residency Program at RIKEN AIP. His research interests lie in machine learning and deep learning, especially weakly-supervised learning and adversarial learning. He has served as area chairs of NeurIPS'20 and ICLR'21, senior program committee of IJCAI'21, and program committees of ICML, AISTATS, UAI, AAAI, IJCAI and ACML. He received the RIKEN BAIHO Award (2019), RGC Early Career Scheme (2020) and NSFC Young Scientists Fund (2020).

Nannan Wang received the B.E. degree in information and computation science from Xi'an University of Posts and Telecommunications in 2009. He received his Ph.D. degree in information and telecommunications engineering in 2015. Now, he works with the state key laboratory of integrated services networks at Xidian University. From September 2011 to September 2013, he has been a visiting Ph.D. student with the University of Technology, Sydney, NSW, Australia. His current research interests include computer vision, pattern recognition, and machine learning. He has published more than 100 papers in refereed journals and proceedings including IEEE TPAMI/IJCV/ICML/NeurIPS/ICLR/AAAI, etc.

Jiankang Deng is a Ph.D. candidate in the Intelligent Behaviour Understanding Group (IBUG) at Imperial College London (ICL), supervised by Stefanos Zafeiriou and funded by the Imperial President's PhD Scholarships. He is in the project of EPSRC FACER2VM (Face Matching for Automatic Identity Retrieval, Recognition, Verification and Management). His Ph.D. research topic is face analysis (face detection, face alignment, face recognition and face generation). During his PhD studies, he has organised the Menpo 2D Challenge (CVPR 2017), the Menpo 3D Challenge (ICCV 2017) and Lightweight Face Recognition Challenge (ICCV 2019). He also won many academic challenges, such as ILSVRC Object Detection and Tracking 2017, Activity-Net Untrimmed Video Classification 2017, iQIYI Celebrity Video Identification Challenge 2018, Disguised Face Recognition Challenge 2019. He is a reviewer in prestigious computer vision journals and conferences including T-PAMI, IJCV, CVPR, ICCV and ECCV. He is the main contributor of the widely used open-source platform Insightface.

Jiatong Li received his dual-Ph.D. degree from the School of Information and Electronics, Beijing Institute of Technology, as well as the Faculty of Engineering and Information Technology, University of Technology Sydney. He was a Post-Doctoral Research Fellow with China Academy of Electronics & Information technology. He is now a Research Scientist at Meituan. His research interests include image processing, computer vision, robotics vision, especially with deep / machine learning and signal processing algorithms.



Yinian Mao received his B.S.E. from Tsinghua University in 2001, and Ph.D. from University of Maryland, College Park in 2006, both in electrical engineering. He is Senior Director of Technology at Meituan, leading Meituan's autonomous drone delivery business unit. Before joining Meituan, he was founder and CEO of Airlango Technology, and was senior staff engineer at Qualcomm. His research includes robotics, transport layer optimization, adaptive streaming, and chipset and system security. Dr. Mao is a member of IEEE and holds more than 30 patents worldwide.

Tongliang Liu is a Lecturer (Assistant Professor) with the School of Computer Science at the University of Sydney. He is also a Visiting Scientist at RIKEN AIP. His current research interests include weakly supervised learning and adversarial learning. He has authored and co-authored more than 60 research articles including the NeurIPS, ICML, CVPR, ECCV, AAAI, IJCAI, KDD, ICME, IEEE T-PAMI, T-NNLS, and T-IP, with best paper awards, e.g., the 2019 ICME Best Paper Award. He is a recipient of Discovery Early Career Researcher Award (DECRA) from Australian Research Council (ARC) and was shortlisted for the J. G. Russell Award by Australian Academy of Science (AAS) in 2019.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Imperial College London. Downloaded on June 17,2022 at 14:21:20 UTC from IEEE Xplore. Restrictions apply.