

Gray Learning from Non-IID Data with Out-of-distribution Samples

Zhilin Zhao, Longbing Cao, *Senior Member, IEEE*, and Chang-Dong Wang, *Senior Member, IEEE*

Abstract—The integrity of training data, even when annotated by experts, is far from guaranteed, especially for non-IID datasets comprising both in- and out-of-distribution samples. In an ideal scenario, the majority of samples would be in-distribution, while samples that deviate semantically would be identified as out-of-distribution and excluded during the annotation process. However, experts may erroneously classify these out-of-distribution samples as in-distribution, assigning them labels that are inherently unreliable. This mixture of unreliable labels and varied data types makes the task of learning robust neural networks notably challenging. We observe that both in- and out-of-distribution samples can almost invariably be ruled out from belonging to certain classes, aside from those corresponding to unreliable ground-truth labels. This opens the possibility of utilizing reliable complementary labels that indicate the classes to which a sample does not belong. Guided by this insight, we introduce a novel approach, termed *Gray Learning* (GL), which leverages both ground-truth and complementary labels. Crucially, GL adaptively adjusts the loss weights for these two label types based on prediction confidence levels. By grounding our approach in statistical learning theory, we derive bounds for the generalization error, demonstrating that GL achieves tight constraints even in non-IID settings. Extensive experimental evaluations reveal that our method significantly outperforms alternative approaches grounded in robust statistics.

Index Terms—Non-IID Data, Out-of-distribution Data, Gray Learning, Complementary Label, Generalization.

I. INTRODUCTION

DEEP neural networks trained on *in-distribution* data demonstrate powerful generalization capabilities when tested on samples from the same distribution [1], [2]. The training of such networks often necessitates large volumes of labeled data. To accumulate this data, original samples must be collected from various sources and subsequently annotated by experts [3], [4]. However, this raw data pool is not guaranteed to be clean. It may include *out-of-distribution* [5], [6] samples with semantic shifts originating from other distributions. These out-of-distribution samples do not belong to any classes within the in-distribution dataset and should, therefore, be discarded during the annotation process [7]. Despite this, due to limitations in expert knowledge or inadvertent errors,

these out-of-distribution samples can be misclassified as in-distribution and erroneously labeled [8]. As a result, the acquired training dataset becomes non-independent and identically distributed (non-IID), incorporating both in-distribution and out-of-distribution samples. Such non-IID data can distort the classification learning for in-distribution samples and inevitably impair the generalization capabilities of the trained network [9]–[11]. Therefore, deriving a robust network from such contaminated non-IID data is of paramount importance.

The primary challenge in training a robust network from non-IID data containing both in-distribution and out-of-distribution samples lies in leveraging the reliable information embedded within the unclean dataset. While the class labels for in-distribution samples are generally trustworthy, those for out-of-distribution samples are not. However, a network initialized randomly lacks the capacity to discern which samples are in-distribution and which are out-of-distribution. This presents a dilemma: Directly learning to classify by mapping inputs to their ground-truth labels would mislead the learning process because the out-of-distribution samples do not belong to the classes corresponding to their ground-truth labels. Nonetheless, we find that the complementary labels¹ [12], [13] for a given sample, be it in-distribution or out-of-distribution, are consistently reliable. Specifically, an in-distribution sample does not belong to any classes other than the one corresponding to its ground-truth label, and an out-of-distribution sample does not belong to any class. Therefore, the reliable information in this unclean non-IID dataset resides in the complementary labels. Based on this insight, *classification can be indirectly learned by training the network to reject mapping a sample to its corresponding complementary labels*.

Based on the above discussion, we propose a novel method, referred to as *gray learning* (GL), specifically designed for learning robust networks from non-IID data that includes both in-distribution and out-of-distribution samples. The core principle of GL is to learn from both ground-truth and complementary labels while adaptively adjusting the weights of the losses for these two types of labels based on prediction confidence. To elaborate, GL focuses on leveraging the ground-truth label for samples likely to be in-distribution, and complementary labels for those likely to be out-of-distribution. GL employs Maximum over Softmax Probabilities (MSP) [14] as a mechanism to calculate prediction confidence, which serves as a proxy for the likelihood that a given sample is in-distribution. Higher probabilities suggest that a sample is

¹Contrary to the definition of ground-truth labels that refer to the class a given input belongs to, complementary labels refer to the classes a given input does not belong to.

This work was supported in part by the Australian Research Council Discovery under Grant DP190101079 and in part by the Future Fellowship under Grant FT190100734.

Zhilin Zhao and Longbing Cao are with the Data Science Lab, School of Computing and DataX Research Centre, Macquarie University, Sydney, NSW 2109, Australia. E-mail: zhaozh17@hotmail.com, longbing.cao@gmail.com

Chang-Dong Wang is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China, Guangdong Province Key Laboratory of Computational Science, Guangzhou, China, and Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China. E-mail: changdongwang@hotmail.com

more likely to be in-distribution, whereas lower probabilities indicate the opposite. For each training sample, GL computes two distinct losses: one for its ground-truth label and another for its complementary labels. These loss weights are then adaptively adjusted based on the prediction confidence. During the early training stages, prediction confidences for all samples are generally low due to the network’s limited classification knowledge. As a result, the network primarily learns from the complementary labels for both in-distribution and out-of-distribution samples. However, as the training progresses, prediction confidence for in-distribution samples improves, widening the confidence gap between in- and out-of-distribution samples. This allows the network to increasingly rely on ground-truth labels for high-confidence samples that are more likely to be in-distribution, thereby further refining its ability to discriminate between in- and out-of-distribution samples based on their confidence levels.

The major contributions of this study are as follows:

- 1) A groundbreaking *gray learning* (GL) framework is proposed to effectively learn from non-IID data sets, which comprise a mix of in-distribution and out-of-distribution samples.
- 2) The generalization error bound is derived to elucidate the impact of out-of-distribution samples on model performance and to provide assurances of convergence.
- 3) Comprehensive experiments demonstrate that the proposed approach significantly surpasses existing methods grounded in robust statistics.

The remainder of the paper is organized in the following manner: Section II offers an overview of related work. Section III details the methodology of the proposed GL framework. Theoretical underpinnings and empirical results are presented in Sections IV and V, respectively. Finally, Section VI provides concluding remarks.

II. RELATED WORK

Non-IID data presents various non-IIDness and non-IID settings, becoming a critical challenge in both shallow and deep learning [15]–[20]. In this work, we explore a novel non-IID scenario characterized by the presence of both in-distribution and out-of-distribution samples. As there are no existing studies that specifically address this unique scenario, we draw insights from various related research fields that also grapple with non-IID data complexities. These include but are not limited to, out-of-distribution detection [7], outlier detection [21], domain generalization [22], learning with noisy labels [23].

A. Out-of-distribution Detection

Out-of-distribution detection aims to distinguish between in-distribution and out-of-distribution samples during the test phase [24] of networks, and various methods have been proposed for this purpose. Maximum over Softmax Probabilities (MSP) [14] uses a threshold-based detector that gauges sample confidence through maximum softmax outputs. Prior Networks (PN) [25] enhances this discrimination by incorporating real-world out-of-distribution samples during training to widen the

confidence gap between the two types of samples. Confidence-Calibrated Classifier (CCC) [26] employs a generative adversarial network to establish boundary points for in-distribution samples and treats these boundaries as indicative of out-of-distribution, thereby encouraging the network to assign lower confidence to such samples. IsoMax [6] refines MSP by resolving issues related to anisotropy and low entropy. Learning from Cross-class Vicinity Distribution (LCVD) [27] aims to diminish confidence in out-of-distribution samples by investigating the vicinity distributions associated with in-distribution samples. Notably, existing work in this area does not address the more complex scenario of training on non-IID data that includes both in- and out-of-distribution samples.

B. Outlier Detection

Outlier detection aims to identify and eliminate training samples that diverge significantly from the majority of the data, thereby enhancing the efficacy of downstream learning tasks [28]. For example, Iterative Learning (IL) [29], which builds upon the LOF algorithm [30], [31], iteratively identifies outlier samples from the training data and refines network training accordingly. In another study, feature coupling techniques [18] are employed to identify outliers in non-IID categorical data. Outlier Exposure (OE) [32] makes use of auxiliary anomalous data to enhance the performance of deep anomaly detectors by training them against a supplementary dataset comprising outliers. Moreover, Outlier Generation [33] features a two-level hierarchical latent space model built using autoencoders and variational autoencoders; this method aims to create synthetic but robust anomalies for training binary classifiers. However, it is important to note that these approaches do not specifically address the scenario where both training and test data exhibit non-IID characteristics and include both in- and out-of-distribution samples.

C. Domain Generalization

Domain generalization aims to train machine learning models capable of generalizing to previously unseen domains by utilizing data from a variety of known domains during the training process. Several notable methods have been proposed in this area. For instance, Domain-Adversarial Neural Network (DANN) [34] employs a gradient reversal layer to minimize domain discrepancies while maximizing task performance. Meta-Learning for Domain Generalization (MLDG) [35] framework leverages meta-learning techniques to enhance the network ability to generalize across multiple domains. Domain-Invariant Representation Learning (DIRL) [36] aims to identify domain-agnostic features by minimizing the mutual information between the learned features and the domain labels. Conditional Domain Adversarial Network (CDAN) [37] goes a step further by accounting for the conditional distribution of labels within each domain, thereby improving generalization performance. METABDRY [38] employs a combination of pointer networks, adversarial learning, and meta-learning to tackle the challenges posed by sparse boundary tags and a variable output vocabulary. However, it is worth noting that these domain generalization techniques operate under the

assumption that all training samples are accurately annotated, which contrasts with the non-IID scenarios we consider, where some training samples may contain incorrect labels.

D. Learning with Noisy Labels

Learning with noisy labels aims to develop robust models that can be trained effectively even when some of the training samples have incorrect in-distribution labels [39], [40]. This stands in contrast to our focus on non-IID scenarios, where the training data includes out-of-distribution samples that are incorrectly annotated as in-distribution. Various methods exist for tackling the issue of noisy labels. Data cleaning techniques [41] identify and correct samples with label noise. MentorNet [42] employs a pre-trained auxiliary network to guide the selection of clean samples during training. Decoupling [43] and Co-teaching [44] both utilize dual networks; Decoupling updates each network based on samples that yield differing predictions, whereas Co-teaching trains each network on a subset of samples with low loss, as chosen by the other network. Another avenue of research explores noise-tolerant loss functions. For instance, Mean-Absolute Error (MAE) [45] is a symmetric loss that has been proven robust against various types of label noise. Symmetric Cross-Entropy Learning (SL) [46] enhances traditional cross-entropy loss by adding a noise-robust reverse cross-entropy term, thus making it more symmetric and robust. Bootstrapping [47] avoids the need for explicit noise modeling by convexly combining the original training labels with the network current predictions. Self-Reweighting from Class Centroids (SRCC) [48] dynamically adjusts the contribution of each sample based on its proximity to class centroids learned online.

E. Other Related Areas

The non-IID scenarios considered in this study also intersect with several cutting-edge research areas. For instance, open-set recognition [49] accounts for out-of-distribution samples by categorizing them into an additional class during training. However, unlike our approach, open-set recognition explicitly identifies the distribution of each training sample. Scene graphs [50] provide a semantic understanding of what is normal or expected in a given type of scene. If an object or relationship appears that does not fit into the established scene graph, it could potentially be flagged as an out-of-distribution item. Unsupervised machine translation [51] improves translation performance by incorporating multi-modal information from visual content and ensure the trained model can generalize to a different form of data that was not seen during training. Zero-shot learning [52], [53] transfer knowledge from seen to unseen activities is similar to identifying out-of-distribution examples based on learned in-distribution data. Specifically, zero-shot temporal activity detection [52] aims to detect activities that have never been seen during training, and ZeroNAS [53] conducts a differentiable generative adversarial networks architecture search in a specially designed search space for zero-shot learning.

The GL method incorporates elements of Curriculum Learning [54] and Negative Learning [23]. Curriculum Learning,

as established by previous works [54], begins with the network learning simpler aspects of a task and progressively incorporates more challenging examples. In a similar vein, Self-Paced Learning (SPL) [55], [56] integrates curriculum design directly into the learning process. Unlike traditional supervised learning methods that update networks based on labeled samples, SPL allows the network itself to dynamically dictate the pace of the curriculum, ranging from simpler to more complex samples. In contrast to these traditional methods, Negative Learning employs complementary labels to reduce the likelihood of erroneous label information affecting the network training. This approach provides a counterpoint to the more standard techniques of supervised learning. On a theoretical level, we employ classical generalization bounds, specifically those based on the Rademacher complexity of a hypothesis class [57]. The Rademacher complexity serves as a measure of uniform convergence rates. Supporting this theoretical underpinning, Golowich et al. [58] have provided bounds on the Rademacher complexity for neural networks, assuming norm constraints on the parameter matrix of each layer. Additionally, we draw inspiration from domain adaptation theories, particularly in the use of divergence measures [59] to analyze the discrepancy between in-distribution and out-of-distribution samples.

III. GRAY LEARNING

In the specific scenario we examine, the training dataset is characterized by a non-IID amalgamation of both in-distribution and out-of-distribution samples. In contrast, the test dataset is exclusively composed of in-distribution samples. It is crucial to note that the out-of-distribution samples within the training data are semantically divergent from the in-distribution samples, indicating that they do not align with any recognized in-distribution class. While in-distribution samples carry accurate and reliable annotations, the out-of-distribution samples are misleadingly labeled, as though they belong to in-distribution categories. This distinction underscores the reliability of the labels for in-distribution samples, in stark contrast to the dubious nature of the labels attached to out-of-distribution samples.

As depicted in Fig. 1, GL method employs a phased approach to segregate and address in-distribution and out-of-distribution samples distinctively during the training regimen. Specifically, for each training instance, the GL algorithm performs the following steps:

- 1) Calculates a prediction confidence score, which serves as an estimate of the likelihood that the sample belongs to either the in- or out-of-distribution category.
- 2) Assesses loss functions based on both the ground-truth labels and their complementary counterparts.
- 3) Dynamically adjusts the weights assigned to the two types of losses in accordance with the prediction confidence score.

In this manner, GL is inclined to rely more heavily on ground-truth labels for samples that exhibit high confidence scores, and on complementary labels for those with low confidence scores. Generally speaking, samples with high confidence are

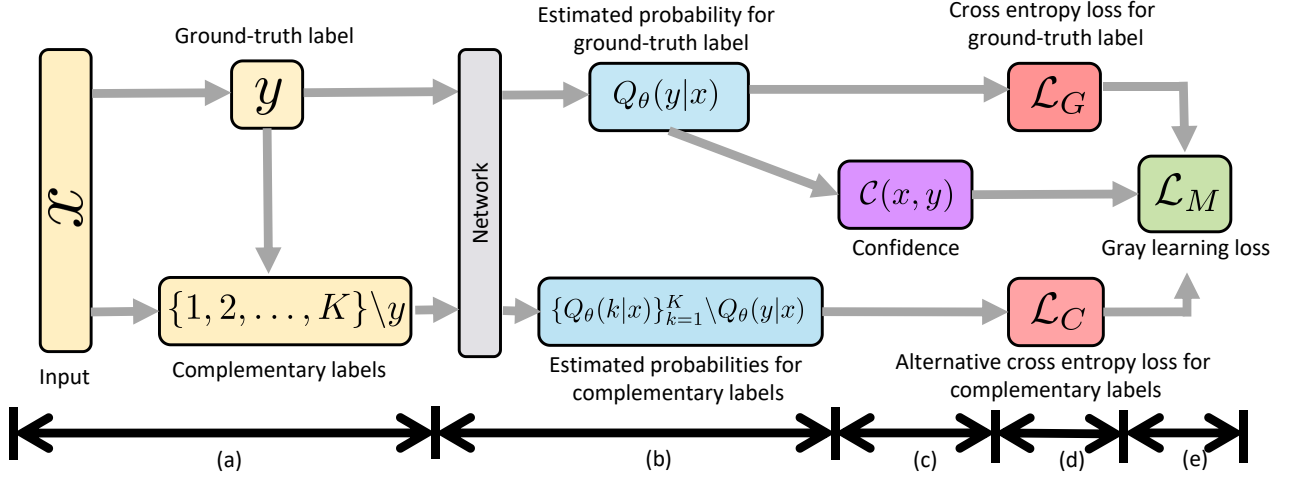


Fig. 1. The framework of gray learning. (a): Input x corresponds to the ground-truth label y and the complementary labels $\{1, 2, \dots, K\} \setminus y$. The complementary labels of a training sample are inferred from its ground-truth label, which indicate the classes the training sample does not belong to. (b): The network outputs the probabilities $Q_\theta(y|x)$ and $\{Q_\theta(k|x)\}_{k=1}^{K-1} \setminus Q_\theta(y|x)$ for the corresponding labels. (c): The confidence $\mathcal{C}(x, y)$ is based on the probability of the ground-truth label. (d): The loss functions \mathcal{L}_G and \mathcal{L}_C are for the ground-truth label and the complementary labels, respectively. We obtain the loss function \mathcal{L}_M of gray learning by using $\mathcal{C}(x, y)$ to adaptively adjust the weights for \mathcal{L}_G and \mathcal{L}_C where a sample with higher confidence provides a higher weight for \mathcal{L}_G , or vice versa.

likely to be in-distribution given that such samples form the majority of the training set. Conversely, samples manifesting low confidence scores are considered ambiguous in the initial stages of training but are increasingly likely to be categorized as out-of-distribution as the model evolves.

A. Setup

Let \mathcal{X} represent the input space and \mathcal{Y} the label space. We define \mathcal{Y} as a finite set containing K classes, formally $\mathcal{Y} = \{1, 2, \dots, K\}$. Our training dataset, denoted as \mathcal{D} , consists of N samples: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$. These samples are drawn from an unknown distribution \mathcal{P} , which is defined over the Cartesian product $\mathcal{X} \times \mathcal{Y}$.

For a given loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$, and any function $h : \mathcal{X} \rightarrow \mathcal{Y}$ that belongs to the hypothesis class \mathcal{H} [60], we introduce the expected risk:

$$\epsilon_{\mathcal{P}}(\mathcal{L}, h) = \mathbb{E}_{(x, y) \sim \mathcal{P}} \mathcal{L}(h(x), y), \quad (1)$$

and the empirical risk:

$$\hat{\epsilon}_{\mathcal{P}}(\mathcal{L}, h) = \frac{1}{N} \sum_{(x, y) \sim \mathcal{D}} \mathcal{L}(h(x), y). \quad (2)$$

We specify the distribution of in-distribution samples as \mathcal{P}_I and that of out-of-distribution samples as \mathcal{P}_O . Then, the overall distribution \mathcal{P}_M of the training dataset can be modeled as a weighted mixture of these two distributions:

$$\mathcal{P}_M = (1 - \alpha)\mathcal{P}_I + \alpha\mathcal{P}_O \quad (3)$$

where $\alpha \in [0, 1]$ serves as a component parameter that controls the proportion of out-of-distribution samples in \mathcal{P}_M . We denote an in-distribution sample as $(x^I, y^I) \sim \mathcal{P}_I$ and an out-of-distribution sample as $(x^O, y^O) \sim \mathcal{P}_O$. The training

dataset \mathcal{D}_M can be considered as a union of N_I in-distribution and N_O out-of-distribution samples:

$$\mathcal{D}_M = \mathcal{D}_I \cup \mathcal{D}_O \quad (4)$$

where $\mathcal{D}_I = \{(x_i^I, y_i^I)\}_{i=1}^{N_I}$ and $\mathcal{D}_O = \{(x_i^O, y_i^O)\}_{i=1}^{N_O}$. It is important to note that the total number of samples N is the sum of N_I and N_O , i.e., $N_I + N_O = N$.

Building upon the concept of discrepancy between source and target domains [59], we introduce a metric $d_{\mathcal{H}}^{\mathcal{L}}(\mathcal{P}_I, \mathcal{P}_O)$ to quantify the *discrepancy* between in-distribution and out-of-distribution sample distributions. This is formalized as:

$$d_{\mathcal{H}}^{\mathcal{L}}(\mathcal{P}_I, \mathcal{P}_O) = \sup_{h \in \mathcal{H}} (|\epsilon_{\mathcal{P}_I}(\mathcal{L}, h) - \epsilon_{\mathcal{P}_O}(\mathcal{L}, h)|). \quad (5)$$

In essence, a low discrepancy value suggests that a hypothesis $h \in \mathcal{H}$ performs comparably under both \mathcal{P}_I and \mathcal{P}_O , and vice versa. For the mixture distribution \mathcal{P}_M , we estimate the conditional distribution $\mathcal{P}_M(y|x)$ by a parameterized distribution $Q_\theta(y|x)$ with model parameter θ . The softmax value of label y is formulated as:

$$Q_\theta(y|x) = \frac{\exp h_\theta^y(x)}{\sum_{k \in [K]} \exp h_\theta^k(x)}. \quad (6)$$

In the context of deep learning, the hypothesis h_θ corresponds to a parameterized neural network. $h_\theta(x) = \{h_\theta^k(x)\}_{k=1}^K$ gives the network output for an input x .

B. Objective Function

The algorithmic procedure for the GL method is further delineated in Algorithm 1.

1) *Confidence Calculation*: Unlike Maximum Softmax Probability (MSP), a conventional method for out-of-distribution detection that calculates confidence scores for unlabeled samples during the testing phase, GL differs by focusing on labeled samples within the training phase. In essence, while softmax distributions of MSP serve as a basis for discerning in-distribution versus out-of-distribution samples, GL adapts this approach to suit its unique training context.

Specifically, we modify the confidence calculation methodology of MSP to take into account ground-truth labels. The confidence of a training input x with its associated label y is then defined as follows:

$$\mathcal{C}(x, y) = Q_\theta(y|x). \quad (7)$$

Given that in-distribution samples constitute the majority of the training dataset, a sample (x, y) with a high confidence score is more likely to be an in-distribution sample. However, it is important to note that a low-confidence score does not definitively categorize a sample as either in-distribution or out-of-distribution. This is because, during the early stages of training, all samples, regardless of their true distribution, tend to exhibit low confidence scores.

2) *Cross Entropy Loss for Ground-truth Labels*: For a sample (x, y) that exhibits high confidence and is thus likely to be an in-distribution instance, we employ the standard cross-entropy loss \mathcal{L}_G . This loss function maps the input x to its corresponding ground-truth label y , mathematically represented as,

$$\mathcal{L}_G(x, y) = -\log Q_\theta(y|x). \quad (8)$$

For samples that are truly in-distribution, this cross-entropy loss \mathcal{L}_G is effective because the ground-truth labels are reliable indicators of the actual classes. However, when it comes to out-of-distribution samples, the application of \mathcal{L}_G can be detrimental. This is because these samples are erroneously tagged with labels from in-distribution classes, to which they do not actually belong. Consequently, \mathcal{L}_G is ill-suited for a training dataset comprising a non-IID mixture of in- and out-of-distribution samples, especially if the objective is to train a robust network.

3) *Alternative Cross Entropy Loss for Complementary Labels*: For low-confidence samples (x, y) , the determination of whether they belong to in- or out-of-distribution classes is ambiguous, especially in the early stages of training when most samples exhibit low confidence. To address this, we introduce the concept of complementary labels. These labels are useful for both in- and out-of-distribution samples, as neither set belongs to the classes represented by these complementary labels.

For a given training input x with its associated label y , we define the corresponding set of complementary labels as:

$$\mathcal{Z}(x, y) = \{1, 2, \dots, K\} \setminus y. \quad (9)$$

Subsequently, we employ an alternative cross-entropy loss \mathcal{L}_C , designed to negate the corresponding complementary labels. This loss function is mathematically formulated as

$$\mathcal{L}_C(x, \mathcal{Z}(x, y)) = - \sum_{y' \in \mathcal{Z}(x, y)} \log(1 - Q_\theta(y'|x)). \quad (10)$$

For an in-distribution sample, \mathcal{L}_C helps identify the ground-truth label by actively excluding the complementary labels. While out-of-distribution samples can also benefit from this approach by improving their confidence on the misleading ground-truth label and excluding the complementary labels. Accordingly, \mathcal{L}_C is less risky than \mathcal{L}_G in terms of providing incorrect label information, as it does not directly map inputs to ground-truth labels.

4) *Adaptively Weighting Loss*: GL obviates the need for selecting in-distribution samples based on confidence during training. Unlike traditional methods, GL capitalizes on the use of complementary labels, allowing it to learn effectively from a non-IID dataset comprising both in- and out-of-distribution samples.

For high-confidence samples, GL applies a higher weight to the cross-entropy loss \mathcal{L}_G , which targets ground-truth labels, thereby widening the confidence gap between in- and out-of-distribution samples. Conversely, for low-confidence samples, GL emphasizes the alternative cross-entropy loss \mathcal{L}_C , which targets complementary labels. This allows the model to glean accurate label information even when the ground-truth labels may be untrustworthy.

To balance these objectives, GL employs a weighted loss function $\mathcal{L}_M(x, y)$, defined as follows:

$$\mathcal{L}_M(x, y) = \mathcal{C}(x, y)\mathcal{L}_G(x, y) + (1 - \mathcal{C}(x, y))\mathcal{L}_C(x, \mathcal{Z}(x, y)). \quad (11)$$

Here, $\mathcal{C}(x, y)$ serves as the weighting factor, derived from the prediction confidence according to Eq. (7). The expected risk under this loss function $\epsilon_{\mathcal{P}_M}(\mathcal{L}_M, h)$ is then given by

$$\epsilon_{\mathcal{P}_M}(\mathcal{L}_M, h) = \int \mathcal{L}_M(x, y) d\mathcal{P}_M. \quad (12)$$

To empirically estimate this risk, we obtain N samples from the mixture distribution \mathcal{P}_M which includes N_I in-distribution samples and N_O out-of-distribution samples. The empirical risk of the expected risk $\mathcal{L}_M(x, y)$ is defined as:

$$\hat{\epsilon}_{\mathcal{P}_M}(\mathcal{L}_M, h) = \frac{1}{N} \sum_{(x, y) \sim \mathcal{D}_M} \mathcal{L}_M(x, y). \quad (13)$$

As the training progresses, GL dynamically adjusts the weighting between \mathcal{L}_G and \mathcal{L}_C based on the evolving confidence levels of the samples. Specifically, when the confidence level of a sample (x, y) is low, GL focuses on learning from the complementary labels, as it is uncertain whether the sample is in- or out-of-distribution. Once the confidence level rises, indicating likely in-distribution status, GL shifts its focus towards the ground-truth label for more direct and effective learning. This iterative strategy enables GL to increasingly segregate in- and out-of-distribution samples as training progresses, culminating in a network that is robust and versatile.

Algorithm 1 Gray Learning

```

1: Input: training dataset  $\mathcal{D}_M$ 
2: repeat
3:   for  $(x, y)$  in  $\mathcal{D}_M$  do
4:     Obtain its complementary label set:  $\mathcal{Z}(x, y) = \{1, 2, \dots, K\} \setminus y$ 
5:     Calculate the confidence:  $\mathcal{C}(x, y) = Q_\theta(y|x)$ 
6:     Estimate the loss for the ground-truth label:  $\mathcal{L}_G(x, y) = -\log Q_\theta(y|x)$ 
7:     Estimate the loss for complementary labels:  $\mathcal{L}_C(x, \mathcal{Z}(x, y)) = -\sum_{y' \in \mathcal{Z}(x, y)} \log(1 - Q_\theta(y'|x))$ 
8:     Obtain the adaptively weighting loss
           
$$\mathcal{L}_M(x, y) = \mathcal{C}(x, y)\mathcal{L}_G(x, y) + (1 - \mathcal{C}(x, y))\mathcal{L}_C(x, \mathcal{Z}(x, y))$$

9:   end for
10:  Estimate the empirical risk  $\hat{\epsilon}_{\mathcal{P}_M}(\mathcal{L}_M, h_\theta)$ 
11:  Obtain gradients  $\nabla_\theta \hat{\epsilon}_{\mathcal{P}_M}(\mathcal{L}_M, h_\theta)$  to update parameters  $\theta$ 
12: until convergence
13: Output: parameterized network  $h_\theta$ 

```

IV. THEORETICAL GUARANTEES

In this section, we present the theoretical results that guarantee the efficacy of the GL method. Unlike standard methods that solely optimize the conventional cross-entropy loss, GL is designed to work with non-IID datasets that include a mixture of both in-distribution and out-of-distribution samples. Specifically, in the training phase, in-distribution samples come with reliable annotations, while out-of-distribution samples are erroneously treated as in-distribution and are labeled as such.

We juxtapose GL against a baseline approach herein referred to as the standard method. The standard method employs the traditional cross-entropy loss \mathcal{L}_G for optimization, while GL uses the weighted loss \mathcal{L}_M . We define the hypothesis that minimizes the loss function for the standard method on the mixed distribution \mathcal{P}_M of in- and out-of-distribution samples as \hat{h}_M . Analogously, the hypothesis that minimizes the loss function for GL under the same distribution is defined as \tilde{h}_M . They are:

$$\begin{aligned} \hat{h}_M &= \arg \min_{h \in \mathcal{H}} \hat{\epsilon}_{\mathcal{P}_M}(\mathcal{L}_G, h), \\ \tilde{h}_M &= \arg \min_{h \in \mathcal{H}} \hat{\epsilon}_{\mathcal{P}_M}(\mathcal{L}_M, h). \end{aligned} \quad (14)$$

It is worth noting that the test data comprise exclusively in-distribution samples. Therefore, both \hat{h}_M from the standard method and \tilde{h}_M from GL aspire to approximate the optimal hypothesis h_I^* tailored for the in-distribution sample. The optimal hypothesis h_I^* is formally defined as:

$$h_I^* = \arg \min_{h \in \mathcal{H}} \epsilon_{\mathcal{P}_I}(\mathcal{L}_G, h). \quad (15)$$

With this theoretical framework, we set the stage for further discussions and proofs, which will quantify the relative advantages of using the GL method in specific scenarios involving non-IID data.

A. Generalization Error of Standard Method.

In the context of a neural network trained via the standard method over a mixed distribution \mathcal{P}_M comprising both in-distribution and out-of-distribution samples, the following theorem establishes an upper bound for the expected risk when

evaluated on the target in-distribution \mathcal{P}_I . For a more detailed mathematical derivation, readers are referred to Appendix 1.

Theorem 1. Assume (1) \mathcal{H} is the class of real-valued networks of depth d over the domain \mathcal{X} , and $x \in \mathcal{X}$ is upper bounded by B , i.e., for any x , $\|x\| \leq B$; (2) the Frobenius norm of the weight matrices W_1, \dots, W_d are at most M_1, \dots, M_d ; (3) the loss function \mathcal{L} is L -Lipschitz continuous w.r.t. $h \in \mathcal{H}$ and $|\mathcal{L}(h, y)| \leq c$ for all y and $h \in \mathcal{H}$; (4) the activation function is 1-Lipschitz, positive-homogeneous, and applied element-wise (such as the ReLU). For any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\begin{aligned} &\epsilon_{\mathcal{P}_I}(\mathcal{L}_G, \hat{h}_M) - \epsilon_{\mathcal{P}_I}(\mathcal{L}_G, h_I^*) \\ &\leq 2\alpha d_{\mathcal{H}}(\mathcal{P}_I, \mathcal{P}_O) \\ &\quad + \frac{4BL(\alpha\sqrt{N_I} + (1-\alpha)\sqrt{N_O})}{\sqrt{N_I N_O}} (\sqrt{2d \ln 2} + 1) \prod_{i=1}^d M_i \\ &\quad + \frac{8c(\alpha\sqrt{N_I} + (1-\alpha)\sqrt{N_O})}{\sqrt{N_I N_O}} \sqrt{2 \ln(16/\delta)}. \end{aligned}$$

In analyzing the performance trade-offs, we identify three contributing terms that influence the gap between the minimizer \hat{h}_M of the empirical risk $\hat{\epsilon}_{\mathcal{P}_M}(\mathcal{L}_G, h)$ on the non-IID samples and the optimal h_I^* of the expected risk $\epsilon_{\mathcal{P}_I}(\mathcal{L}_G, h)$ on the in-distribution samples. The first term addresses the intrinsic distributional differences between in-distribution and out-of-distribution samples. The coefficient α quantitatively signifies this impact. Intuitively, a higher prevalence of out-of-distribution samples in the training data will degrade the in-distribution classification performance, aligning with our empirical expectations. The second term is crucial and pertains to the characteristics of both the neural network architecture and the cross-entropy loss function being used. It encapsulates how well the model can generalize from the training data to unseen in-distribution samples. The impact of this term can vary based on hyperparameter settings, network depth, and other architectural nuances. The third term is associated with the stochastic nature of the training data sampled from \mathcal{P}_M . A

larger dataset can mitigate the sampling bias, thereby reducing the gap between \hat{h}_M and h_I^* .

B. Equivalent form of Gray Learning

In order to derive the generalization error for our proposed GL method, we reframe the objective function, as expressed in Eq. (12), into an equivalent formulation. The comprehensive derivation of this equivalent form is available in Appendix 2.

Theorem 2. *For training samples drawn from the mixture distribution \mathcal{P}_M , the expected risk of GL can be rewritten as:*

$$- \int \log \mathcal{L}_G(x, y) dP_M,$$

s.t. $r(\theta, x, y) \leq \lambda, \forall (x, y) \sim \mathcal{P}_M$,

where

$$r(\theta, x, y) = (1 - Q_\theta(y|x)) \log Q_\theta(y|x) (1 - Q_\theta(y|x)) \\ - (1 - Q_\theta(y|x)) \sum_{k=1}^K \log(1 - Q_\theta(y|x)),$$

and $\lambda > 0$.

We observe that the objective function employed by the GL method can be construed as a constrained form of the conventional cross-entropy loss, where each training sample (x, y) is subjected to a regularizer $r(\theta, x, y)$ capped by an upper bound λ . This parameter λ dynamically modulates the influence of the regularizer during the training process, allowing for more nuanced learning behavior. Interestingly, as λ tends toward infinity, the GL method reverts to the baseline approach, effectively neutralizing its specialized handling of in-distribution and out-of-distribution samples.

C. Generalization Error of Gray Learning

For a network trained using the GL method on the mixed distribution \mathcal{P}_M encompassing both in-distribution and out-of-distribution samples, the ensuing theorem delineates a theoretical upper bound on the expected risk for the target in-distribution \mathcal{P}_I . Additionally, the theorem specifies conditions under which the GL method outperforms the standard approach. The full derivation of this theorem can be found in Appendix 3.

Theorem 3. *Following the conditions of Theorem 1, let $\log \sum_{k=1}^K \exp(h_\theta^k(x)) \leq z$ for any x . For any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$\epsilon_{\mathcal{P}_I}(\mathcal{L}_G, \tilde{h}_M) - \epsilon_{\mathcal{P}_I}(\mathcal{L}_G, h_I^*) \\ \leq 2\alpha d_{\mathcal{H}}(\mathcal{P}_I, \mathcal{P}_O) \\ + \frac{4BLK(\alpha\sqrt{N_I} + (1-\alpha)\sqrt{N_O})}{\sqrt{N_I N_O}} (c + \log(2\lambda - 2)) \\ + \frac{8c(\alpha\sqrt{N_I} + (1-\alpha)\sqrt{N_O})}{\sqrt{N_I N_O}} \sqrt{2 \ln(16/\delta)}.$$

This bound is tighter than that of the baseline method if

$$\lambda \leq 1 + \frac{1}{2} \exp\left(\frac{B(\sqrt{2d \ln 2} + 1) \prod_{i=1}^d M_i}{L\sqrt{K}} - z\right).$$

It is noteworthy that the performance disparity between the optimal hypothesis \tilde{h}_M , which minimizes the empirical risk under GL for non-IID data, and h_I^* , the minimizer of the expected risk for in-distribution samples, is conceptually akin to the gap elucidated for the standard method in Theorem 1. The first term of this gap is attributed to the unavoidable distribution discrepancy between in-distribution and out-of-distribution samples, as captured by Eq. (5). The divergence in the second term arises from the contrasting empirical risks between the standard and GL methods. Intriguingly, if the regularizer $r(\theta, x, y)$ adheres to the criteria outlined in Theorem 3, GL exhibits robust learning capabilities even when faced with non-IID data. The parameter λ serves as an implicit, adaptively-adjusted variable throughout the training regimen. Owing to their ability to effectively model complex data structures, deep neural networks can iteratively minimize λ to fulfill the stated condition in real-world scenarios.

V. EXPERIMENT RESULTS

In the absence of directly comparable methods specifically designed for handling non-IID data comprised of both in- and out-of-distribution samples during training, we validate the efficacy of our proposed GL method². For our evaluations, we compare GL against a baseline approach, which we refer to as the standard method, as well as alternative techniques drawn from the field of robust statistics. The standard method exclusively employs traditional cross-entropy loss for optimization and lacks a dedicated mechanism for addressing out-of-distribution samples.

We conduct an exhaustive set of experiments to assess the robustness and versatility of GL. These experiments include an analysis of the effects of varying proportions, sources, and labels of out-of-distribution samples when mixed with in-distribution training samples. Additionally, we extend our evaluation to multiple network architectures, assess the calibration capabilities, and carry out an ablation study to identify key factors contributing to the performance.

A. Settings

In our experimental evaluation, we assess performance of GL across both imagery and tabular data, the details of which are summarized in TABLE III. For the non-IID data, the proportion of out-of-distribution samples is controlled by the component parameter α . Unless otherwise stated, we set $\alpha = 0.1$ as a default value for our experiments. To generate complementary labels for each training sample, we take all labels other than the ground-truth label as its complementary set. Our evaluation metrics include *Classification Accuracy* for gauging the discriminative capabilities and *Expected Calibration Error* (ECE) [61] for measuring the predictive confidence calibration. ECE quantifies the difference between the network predicted confidence and its actual classification accuracy. A well-calibrated model will exhibit high confidence for

²The source code is publicly available at: <https://github.com/Lawliet-zzl/GL>.

TABLE I
CLASSIFICATION ACCURACY ON IMAGERY DATA. ALL VALUES ARE IN PERCENTAGE, AND BOLDFACE VALUES SHOW THE RELATIVELY BETTER CLASSIFICATION PERFORMANCE.

In-distribution	Out-of-distribution label	Standard	MAE	BT	SPL	IL	SL	NL	SRCC	LCVD	GL
CIFAR10	Specific	93.8	94.1	95.1	89.9	94.8	94.7	94.8	94.7	94.2	95.2
	Random	93.7	94.3	95.1	90.9	95.1	94.9	94.5	94.7	94.4	95.3
SVHN	Specific	95.0	96.3	96.0	92.0	96.1	96.4	96.1	96.4	95.2	96.7
	Random	95.2	96.7	96.4	93.5	96.5	96.6	96.4	96.2	95.3	96.9
CIFAR100	Specific	66.3	69.9	67.6	69.2	66.8	70.8	70.3	70.3	71.1	77.6
	Random	67.1	69.3	67.5	69.2	67.1	70.7	70.5	71.2	72.8	77.2

TABLE II
CLASSIFICATION ACCURACY ON TABULAR DATA. ALL VALUES ARE IN PERCENTAGE, AND BOLDFACE VALUES SHOW THE RELATIVELY BETTER CLASSIFICATION PERFORMANCE.

In-distribution	Standard	MAE	BT	SPL	IL	SL	NL	SRCC	LCVD	GL
Abalone	79.1	81.6	80.8	49.4	80.3	80.9	79.6	80.2	80.6	82.1
Arrhythmia	77.9	78.5	78.3	81.0	78.6	79.8	79.2	81.5	81.2	83.4
Gene	39.2	38.1	38.8	38.3	37.9	35.7	40.6	41.2	42.3	44.8
Iris	59.0	60.0	64.0	50.0	66.0	64.0	60.0	75.9	77.2	95.0
Skyserver	68.6	68.4	68.8	68.7	68.7	68.6	68.8	72.1	74.1	78.9
Speech	53.0	52.8	53.8	44.6	53.8	52.3	53.8	53.2	52.3	56.0
Stellar	56.8	56.8	56.6	56.8	56.7	56.9	56.5	66.3	65.4	76.1
WineQT	59.0	57.0	56.0	48.0	56.1	55.8	55.5	60.2	60.7	62.5
Average	61.6	61.7	62.1	54.6	62.3	61.8	61.8	64.2	66.3	72.4

TABLE III
STATISTICS OF DATASETS.

Type	Dataset	# instances	# features	# labels
Imagery	CIFAR10	60000	1024	10
	SVHN	99289	1024	10
	CIFAR100	60000	1024	100
Tabular	Abalone	4177	8	3
	Arrhythmia	87553	187	5
	Gene	801	20531	5
	Iris	150	150	3
	Skyserver	100000	17	3
	Speech	3960	12	6
	Stellar	100000	16	3
	WineQT	1143	11	6

correctly classified samples and low confidence for misclassifications. In evaluating ECE, we divide the confidence range into 20 bins to capture detailed calibration behavior across different levels of predictive confidence.

1) *Imagery Data*: In our experiments, we utilize a diverse range of datasets to create non-IID training data, incorporating both in-distribution and out-of-distribution samples. Specifically, for in-distribution data, we employ CIFAR10 [62], SVHN [63], and CIFAR100 [62]. We choose Mini-Imagenet [64] as our source for out-of-distribution data, using 100 classes unless otherwise specified.

For CIFAR10 and SVHN, each featuring 10 classes, we segment Mini-Imagenet into 10 balanced subsets according to class labels. For CIFAR100, which contains 100 classes, we

use all samples from Mini-Imagenet as the out-of-distribution data. We introduce two types of labeling for these out-of-distribution samples: *specific labels* and *random labels*. In the specific-labeling scheme, all out-of-distribution samples from the same Mini-Imagenet class receive the same in-distribution label. In contrast, the random-labeling scheme assigns a randomly selected in-distribution label to each out-of-distribution sample.

We preprocess the training data using standard data augmentation techniques, including resizing, random cropping, and random horizontal flipping. Our evaluation spans six different neural network architectures: ResNet18 [65], VGG19 [66], ShuffleNetV2 [67], MobileNetV2 [68], SqueezeNet [69], and DenseNet121 [70]. Parameters are updated using Stochastic Gradient Descent [71], with an initial learning rate of 0.1, reduced by a factor of 10 at epochs 100 and 150. All models are trained for 200 epochs with mini-batches of 128 samples.

2) *Tabular Data*: We extend our evaluation to include small tabular datasets, encompassing two from the UCI repository³ (Abalone and Iris) as well as six from Kaggle⁴ (Arrhythmia, Gene, Skyserver, Speech, Stellar, and WineQT). In each case, we designate samples from the smallest class as out-of-distribution and the remaining samples as in-distribution. For these tabular datasets, we employ shallow, fully-connected neural networks with a two-layer architecture featuring 128 ReLU units in each hidden layer. We use the Adam optimizer [72] with default hyperparameters, training the models over 10 epochs in mini-batches of 16 samples.

³<https://archive.ics.uci.edu/>

⁴<https://www.kaggle.com/>

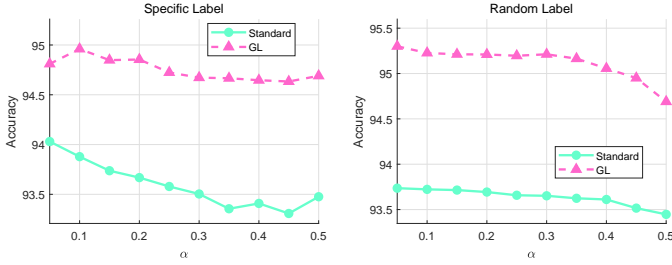


Fig. 2. The effect of the proportion of out-of-distribution samples.

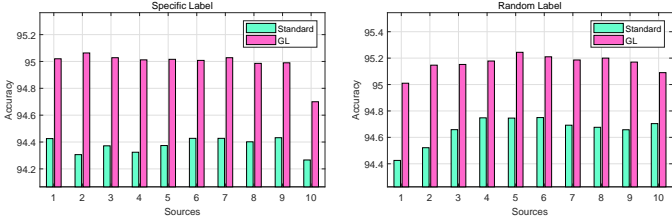


Fig. 3. The effect of the out-of-distribution inputs. An index in $\{1, \dots, 10\}$ represents a source of out-of-distribution inputs.

B. Comparison Results

This study pioneers the investigation into non-IID data that encompasses both in-distribution and out-of-distribution samples. Given the absence of direct competitors, we benchmark GL against various alternative methods derived from related research domains. These methods are also suitable for handling the non-IID data under study. Specifically, the alternative methods include Mean-Absolute Error (MAE) [45], Bootstrapping [47], Iterative Learning (IL) [29], and Symmetric cross entropy Learning (SL) [46], Self-Reweighting from Class Centroids (SRCC) [48], Learning from Cross-class Vicinity Distribution (LCVD) [27]. The other related methods include Self-Paced Learning (SPL) [56] and Negative Learning (NL) [23]. For the image datasets, we deploy the ResNet18 architecture and designate samples from Mini-Imagenet as out-of-distribution. In the case of CIFAR10 and SVHN, we confine ourselves to the first source of Mini-Imagenet to ensure an equivalent number of classes. For CIFAR100, all sources from Mini-Imagenet are incorporated.

The comparison results for imagery data are outlined in TABLE I. Our findings reveal that Self-Paced Learning (SPL) consistently underperforms in terms of classification accuracy across all examined combinations of in-distribution datasets and out-of-distribution labels. This outcome suggests that SPL is not effective at discerning between in-distribution and out-of-distribution samples during the training process. In contrast, BT and IL show a marked improvement, specifically 4.41% gains on CIFAR10 and SVHN. These results indicate that strategies that filter out out-of-distribution samples and avoid directly fitting to non-IID data can enhance the robustness of learned models. NL only marginally outperforms the straightforward MAE method by 0.46%, suggesting that merely utilizing complementary labels is not sufficient for capturing the nuances required for accurate classification. Remarkably, our proposed GL method outshines all competitors, achieving a 7.2% improvement over the worst-performing SPL and a 3.3%

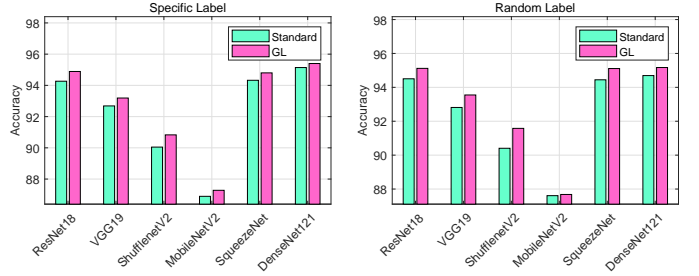


Fig. 4. The effect of the network architectures.

improvement over the next best method, SL. This superiority can be attributed to capability of GL to ameliorate the limitations of both SPL and NL by adaptively reweighting training samples based on their confidence levels and reconsidering the ground-truth labels for high-confidence samples.

The comparative analysis for small tabular data is presented in TABLE II. Notably, GL significantly outperforms the competing methods, showing an average classification accuracy improvement of 16.7% across all datasets. Previous studies have highlighted that neural networks are prone to overfitting when dealing with small, unreliable tabular data [20], [73], leading to poor generalization on test in-distribution samples. GL addresses this issue by adaptively utilizing ground-truth labels for in-distribution samples and complementary labels for out-of-distribution samples. The use of complementary labels mitigates the risk of introducing erroneous label information, enhancing the robustness. Consequently, GL demonstrates superior performance even when applied to small tabular datasets.

C. Effect of Out-of-distribution Samples

We investigate the impact of both the proportion parameter α and the source variations of out-of-distribution samples within the training set. For this analysis, we utilize the CIFAR10 dataset and employ a ResNet18 architecture for training the networks. To assess the efficacy of labeling strategies for out-of-distribution samples, we separately examine the cases of specific and random labels. Our comparative study pits the performance of GL against that of the standard method, which employs a cross-entropy loss function for handling non-IID samples.

1) *Effect of the proportion α* : To explore the effects of a high value of the proportion parameter α , we designate samples from SVHN as out-of-distribution samples for the CIFAR10 dataset and vary α within the range of $[0.05, \dots, 0.5]$. The resulting experimental findings are presented in Fig. 2. For the standard approach, we observe a decline in classification performance as the proportion of out-of-distribution samples increases, regardless of whether specific or random labels are employed. This sensitivity to the presence of out-of-distribution samples confirms the theoretical insights provided in Theorem 1. Specifically, a high value of α exacerbates the divergence between the optimal solution achieved on clean, in-distribution samples and the empirical minimizer obtained using the standard method on non-IID samples. In contrast, the

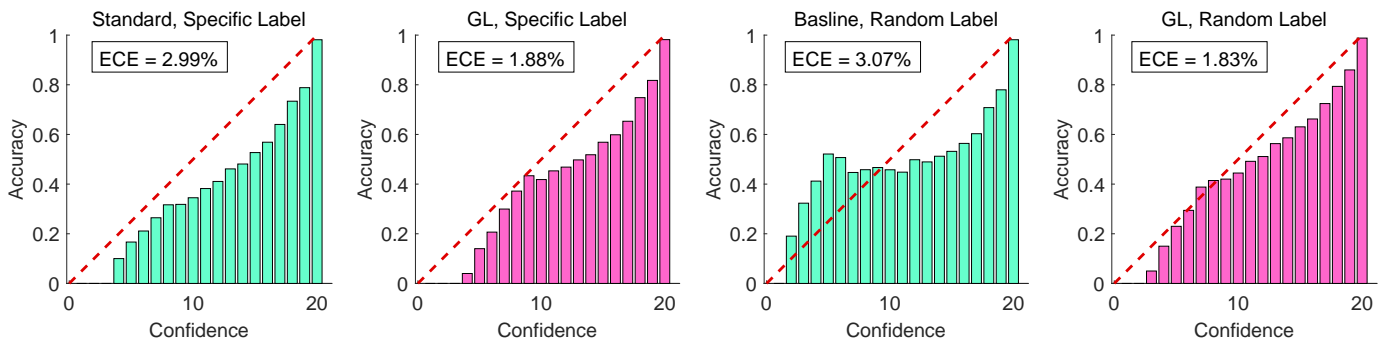


Fig. 5. The calibration results on CIFAR10. The confidence is equally divided into 20 intervals, and each bar represents the expected accuracy of samples whose confidence values are in the same interval. The red dotted diagonal indicates the perfect calibration.

performance of GL remains stable across varying proportions of out-of-distribution samples, indicating its robustness to such variations. During training, the upper bound of λ , as described in Theorem 2, is progressively reduced to fulfill the criteria specified in Theorem 3. By adaptively adjusting λ , GL effectively harmonizes the two terms on the right-hand side of the bound outlined in Theorem 3. Hence, GL is adept at robustly learning network models from non-IID data sets that include out-of-distribution samples.

2) *Effect of sources*: We utilize samples from Mini-Imagenet as the out-of-distribution data for the CIFAR10 dataset and partition Mini-Imagenet into ten distinct subsets based on class order, with each subset comprising samples from ten classes. Each of these subsets serves as a unique source of out-of-distribution samples mixed into the training data, and we evaluate the impact of each source individually. The findings are presented in Fig. 3. Our results indicate that the choice of out-of-distribution source can significantly influence classification outcomes. This variability can be attributed to the distributional discrepancy $d_{\mathcal{H}}(\mathcal{P}_I, \mathcal{P}_O)$ between in-distribution and out-of-distribution samples, as delineated in Theorem 1 and Theorem 3. Across all the different sources, GL consistently outperforms the standard method, highlighting its robustness to a wide array of out-of-distribution samples. This superior performance is achieved by adaptive adjustment of the upper bound of $r(\theta, x, y)$ for each sample, effectively narrowing the gap as outlined in Theorem 3.

D. Effect of Network Architectures

We employ various deep network architectures, including ResNet18, VGG19, ShufflenetV2, MobileNetV2, Senet18, and DenseNet121, to train models using clean data, consisting of in-distribution samples from CIFAR10 and out-of-distribution samples from the first subset of Mini-Imagenet. With a fixed value of $\alpha = 0.1$, we compare the classification performance of GL with the standard method. The findings are summarized in Fig. 4. Our results demonstrate that the choice of network architecture influences classification performance, with residual networks (ResNet18 and DenseNet121) showing superior outcomes. Importantly, GL consistently outperforms the standard method across all architectures, achieving improvements of 0.55% with specific labels and 0.64% with random labels. This attests to the general applicability of GL. The

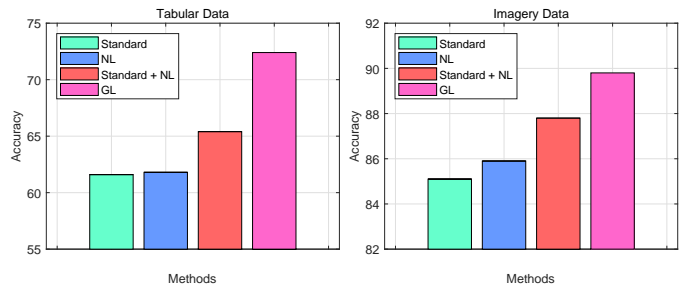


Fig. 6. Results of the ablation study. Each bar for tabular data presents the average classification accuracy across the eight tabular datasets. Each bar for imagery data presents the average classification accuracy across the three imagery datasets with specific and random out-of-distribution labels.

method extends the conventional cross-entropy loss function to accommodate the influence of out-of-distribution samples, making it versatile across different network architectures. Furthermore, as indicated by Theorem 3, the effectiveness of GL is architecture-agnostic, provided the networks are sufficiently powerful to meet the low upper bound λ condition for each training sample.

E. Calibration

We evaluate the ECE of both the standard method and GL on CIFAR10, utilizing the ResNet18 architecture, the first subset of Mini-Imagenet as the out-of-distribution samples, and a component parameter of $\alpha = 0.1$. The findings are depicted in Fig. 5. The standard method registers ECE values of 2.99% and 3.07% for out-of-distribution samples with specific and random labels, respectively. In contrast, GL demonstrates significantly improved calibration with ECE values of 1.88% and 1.83%. These results highlight the poor calibration of the standard method and the near-perfect calibration achieved by GL, thereby affirming the reliability of predictions. The standard method is disproportionately impacted by out-of-distribution samples during training, often leading it to produce high-confidence but incorrect predictions. In stark contrast, the robustness of GL to such samples is achieved by judiciously leveraging complementary labels, thereby mitigating the risk of propagating incorrect label information.

F. Ablation Study

In order to scrutinize the importance of adaptively adjusting the weights for ground-truth and complementary label losses based on prediction confidence, we conduct an ablation study. For context, GL can be conceptualized as dynamically weighting the cross-entropy loss of the standard method and the alternative cross-entropy loss from NL based on prediction confidence. To this end, we compare GL with three other configurations: the standard method, NL, and a composite approach that uniformly blends the losses of standard and NL methods (referred to as Standard + NL). The outcomes of the ablation study across both tabular and imagery datasets are illustrated in Fig. 6.

On tabular data, GL outperforms the standard, NL, and Standard + NL methods by margins of 17.5%, 17.1% and 10.7%, respectively. Similarly, on imagery data, the respective performance gains for GL are 5.52%, 4.54% and 2.28%. These consistent findings across both data types offer valuable insights. When compared to the standard and NL methods, it becomes evident that relying solely on either ground-truth or complementary labels is insufficient for robustly training networks on non-IID data that include both in- and out-of-distribution samples. These methods lack the finesse to differentiate between the two types of samples, implicitly trusting all labels equally. In comparison to the Standard + NL approach, the results suggest that adaptive weighting strategy of GL, which takes prediction confidence into account, is instrumental in its superior performance. This dynamic approach enables GL to focus predominantly on ground-truth labels for in-distribution samples and complementary labels for out-of-distribution samples, thereby widening the confidence gap between these two categories.

VI. CONCLUSION

This study represents the first effort to train neural networks robustly on non-IID data comprising both in-distribution and out-of-distribution samples. Notably, these two types of samples are intermixed, and the out-of-distribution samples are incorrectly annotated with in-distribution labels. We introduce a novel Gray Learning (GL) approach that adaptively learns from both ground-truth labels and complementary labels, dynamically adjusting the loss weights for each based on prediction confidence. To substantiate the efficacy of GL, we derive generalization bounds rooted in the Rademacher complexity of the hypothesis class, demonstrating that GL yields tighter bounds compared to conventional methods that rely solely on cross-entropy loss. Empirical evaluations confirm that our approach is robust against varying in-distribution and out-of-distribution sample proportions, different sources, and alternative labeling schemes for out-of-distribution samples, outperforming competing methods based on robust statistics. Theoretical results indicate that the discrepancy between in-distribution and out-of-distribution samples is crucial for the generalization bound. Consequently, quantifying and minimizing this discrepancy present intriguing and promising avenues for future research.

APPENDIX A PROOF OF THEOREM 1

To proceed, we introduce the following generalization bounds:

Lemma 1 ([71]). *Assume that for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $h \in \mathcal{H}$, we have that $|\mathcal{L}(h(x), y)| \leq c$. Let the optimal $h^* = \arg \min_{h \in \mathcal{H}} \epsilon_{\mathcal{P}}(\mathcal{L}, h)$ and $\hat{h} = \arg \min_{h \in \mathcal{H}} \hat{\epsilon}_{\mathcal{P}}(\mathcal{L}, h)$. With the probability of at least $1 - \delta$, we have*

$$|\hat{\epsilon}_{\mathcal{P}}(\mathcal{L}, h) - \epsilon_{\mathcal{P}}(\mathcal{L}, h)| \leq 2\mathcal{R}(\mathcal{L} \circ \mathcal{H} \circ \mathcal{D}) + 4c\sqrt{\frac{2\ln(4/\delta)}{N}}$$

where $\mathcal{R}(l \circ \mathcal{H} \circ \mathcal{D})$ is the Rademacher complexity of \mathcal{H} with respect to \mathcal{L} and \mathcal{D} .

For any $h \in \mathcal{H}$ and any \mathcal{L}

$$\begin{aligned} & |\epsilon_{\mathcal{P}_M}(\mathcal{L}, h) - \epsilon_{\mathcal{P}_I}(\mathcal{L}, h)| \\ &= |\alpha\epsilon_{\mathcal{P}_O}(\mathcal{L}, h) + (1 - \alpha)\epsilon_{\mathcal{P}_I}(\mathcal{L}, h) - \epsilon_{\mathcal{P}_I}(\mathcal{L}, h)| \\ &= \alpha|\epsilon_{\mathcal{P}_O}(\mathcal{L}, h) - \epsilon_{\mathcal{P}_I}(\mathcal{L}, h)| \\ &\leq \alpha d_{\mathcal{H}}(\mathcal{P}_I, \mathcal{P}_O). \end{aligned} \tag{16}$$

According to Lemma 1, with the probability of at least $1 - \delta$, for any $h \in \mathcal{H}$,

$$\begin{aligned} & |\epsilon_{\mathcal{P}_M}(\mathcal{L}, h) - \hat{\epsilon}_{\mathcal{P}_M}(\mathcal{L}, h)| \\ &\leq (1 - \alpha)|\epsilon_{\mathcal{P}_I}(\mathcal{L}, h) - \hat{\epsilon}_{\mathcal{P}_I}(\mathcal{L}, h)| + \alpha|\epsilon_{\mathcal{P}_O}(\mathcal{L}, h) - \hat{\epsilon}_{\mathcal{P}_O}(\mathcal{L}, h)| \\ &\leq (1 - \alpha) \underbrace{\left(2\mathcal{R}(\mathcal{L} \circ \mathcal{H} \circ \mathcal{D}_I) + 4c\sqrt{\frac{2\ln(8/\delta)}{N_I}} \right)}_{\triangleq \mathfrak{B}_1(8/\delta)} \\ &\quad + \alpha \underbrace{\left(2\mathcal{R}(\mathcal{L} \circ \mathcal{H} \circ \mathcal{D}_O) + 4c\sqrt{\frac{2\ln(8/\delta)}{N_O}} \right)}_{\triangleq \mathfrak{B}_2(8/\delta)}. \end{aligned} \tag{17}$$

Applying the bound Eq. (16) and Eq. (17), we have the following, with the probability at least $1 - \delta$,

$$\begin{aligned} & \epsilon_{\mathcal{P}_I}(\mathcal{L}, \hat{h}_M) \\ &\leq \epsilon_{\mathcal{P}_M}(\mathcal{L}, \hat{h}_M) + \alpha d_{\mathcal{H}}(\mathcal{D}_I, \mathcal{D}_O) \\ &\leq \hat{\epsilon}_{\mathcal{P}_M}(\hat{h}_M) + \alpha d_{\mathcal{H}}(\mathcal{D}_I, \mathcal{D}_O) + \mathfrak{B}_1(8/\delta) + \mathfrak{B}_2(8/\delta) \\ &\leq \hat{\epsilon}_{\mathcal{P}_M}(h_I^*) + \alpha d_{\mathcal{H}}(\mathcal{D}_I, \mathcal{D}_O) + \mathfrak{B}_1(8/\delta) + \mathfrak{B}_2(8/\delta) \end{aligned} \tag{18}$$

Now applying the bound Eq. (17) and Eq. (16), we have the following with the probability at least $1 - \delta$,

$$\begin{aligned} & \epsilon_{\mathcal{P}_I}(\mathcal{L}, \hat{h}_M) \\ &\leq \epsilon_{\mathcal{P}_M}(h_I^*) + \alpha d_{\mathcal{H}}(\mathcal{D}_I, \mathcal{D}_O) + 2\mathfrak{B}_1(16/\delta) + 2\mathfrak{B}_2(16/\delta) \\ &\leq \epsilon_{\mathcal{P}_I}(h_I^*) + 2\alpha d_{\mathcal{H}}(\mathcal{D}_I, \mathcal{D}_O) + 2\mathfrak{B}_1(16/\delta) + 2\mathfrak{B}_2(16/\delta). \end{aligned} \tag{19}$$

Because the loss function \mathcal{L} is L -Lipschitz continuous w.r.t. $h \in \mathcal{H}$. According to the Talagrand's contraction lemma [74], we have

$$\begin{aligned} \mathcal{R}(\mathcal{L} \circ \mathcal{H} \circ \mathcal{D}_I) &\leq L \mathbb{E}_{\sigma \in \{\pm 1\}^{N_I}} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{N_I} \sigma_i h(x_i^I) \right] \\ \mathcal{R}(\mathcal{L} \circ \mathcal{H} \circ \mathcal{D}_O) &\leq L \mathbb{E}_{\sigma \in \{\pm 1\}^{N_O}} \left[\sup_{h \in \mathcal{H}} \sum_{j=1}^{N_O} \sigma_j h(x_j^O) \right]. \end{aligned} \quad (20)$$

To bound $\mathcal{R}(\mathcal{L} \circ \mathcal{H} \circ \mathcal{D}_I)$ and $\mathcal{R}(\mathcal{L} \circ \mathcal{H} \circ \mathcal{D}_O)$ further, we require the following lemma,

Lemma 2 ([58]). *Let \mathcal{H} be the class of real-valued networks of depth d over the domain \mathcal{X} and $x \in \mathcal{X}$ is upper bounded by B , i.e., for any x , $\|x\| \leq B$. Assume the Frobenius norm of the weight matrices W_1, \dots, W_d are at most M_1, \dots, M_d . Let the activation function be 1-Lipschitz, positive-homogeneous, and applied element-wise (such as the ReLU). Then,*

$$\mathbb{E}_{\sigma \in \{\pm 1\}^N} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^N \sigma_i h(x_i) \right] \leq \sqrt{N} B (\sqrt{2d \ln 2} + 1) \prod_{i=1}^d M_i.$$

We complete the proof by substituting Eq. (20) into Eq. (19) and applying Lemma 2.

APPENDIX B PROOF OF THEOREM 2

Rewriting $\mathcal{L}_M(x, y)$, we have

$$\begin{aligned} \mathcal{L}_M(x, y) &= -Q_\theta(y|x) \log Q_\theta(y|x) \\ &\quad - (1 - Q_\theta(y|x)) \sum_{k=1}^K \log(1 - Q_\theta(k|x)) \\ &\quad + (1 - Q_\theta(y|x)) \log(1 - Q_\theta(y|x)). \end{aligned} \quad (21)$$

For the sum of the first and the third terms in Eq. (21), we have

$$\begin{aligned} &(1 - Q_\theta(y|x)) \log(1 - Q_\theta(y|x)) - Q_\theta(y|x) \log Q_\theta(y|x) \\ &= \log \frac{(1 - Q_\theta(y|x))^{1 - Q_\theta(y|x)}}{Q_\theta(y|x)^{Q_\theta(y|x)}} \\ &= \log \frac{(Q_\theta(y|x)(1 - Q_\theta(y|x)))^{1 - Q_\theta(y|x)}}{Q_\theta(y|x)} \\ &= -\log Q_\theta(y|x) + (1 - Q_\theta(y|x)) (Q_\theta(y|x)(1 - Q_\theta(y|x))). \end{aligned} \quad (22)$$

We complete the proof by substituting Eq. (22) into Eq. (21).

APPENDIX C PROOF OF THEOREM 3

We consider the constraint $r(\theta, x, y) \leq \lambda, \forall (x, y) \sim \mathcal{P}_M$ to bound the Rademacher complexities $\mathcal{R}(\mathcal{L} \circ \mathcal{H} \circ \mathcal{D}_I)$ and $\mathcal{R}(\mathcal{L} \circ \mathcal{H} \circ \mathcal{D}_O)$. Due to the affection of the constraint on exploring the hypothesis class \mathcal{H} , we need to obtain the upper bound of $h_\theta^y(x), \forall (x, y) \sim \mathcal{P}_M$ to bound the Rademacher complexities.

To process, we calculate the upper bound of $\sum_{k=1}^K \log(1 - Q_\theta(k|x))$ in $r(\theta, x, y)$ with $\sum_{k=1}^K Q_\theta(k|x) = 1$. We solve the constrained optimization problem by forming a Lagrangian and introducing Lagrange multiplier μ . Accordingly, we define

$$\mathcal{G} = \sum_{k=1}^K \log(1 - Q_\theta(k|x)) + \mu \left(\sum_{k=1}^K Q_\theta(k|x) - 1 \right). \quad (23)$$

The partial derivatives of \mathcal{G} with respect to $Q_\theta(k|x)$ and λ are

$$\frac{\partial \mathcal{G}}{\partial Q_\theta(k|x)} = \frac{1}{Q_\theta(k|x) - 1} + \mu = 0, \forall k \in [K] \quad (24)$$

and

$$\frac{\partial \mathcal{G}}{\partial \lambda} = \sum_{k=1}^K Q_\theta(k|x) - 1 = 0. \quad (25)$$

According to Eq. (24) and Eq. (25), we can obtain the maximum value of \mathcal{G} when $Q_\theta(k|x) = 1/K, \forall k \in [K]$, i.e.,

$$\sum_{k=1}^K \log(1 - Q_\theta(k|x)) \leq K \log \left(1 - \frac{1}{K} \right). \quad (26)$$

According to the basic inequality $x^y \geq \frac{x}{x+y}, \forall x > 0, y \in (0, 1)$ and $Q_\theta(y|x) \in (0, 1)$, we have

$$\begin{aligned} &(1 - Q_\theta(y|x)) \log Q_\theta(y|x) (1 - Q_\theta(y|x)) \\ &= \log(Q_\theta(y|x) (1 - Q_\theta(y|x)))^{(1 - Q_\theta(y|x))} \\ &\geq \frac{Q_\theta(y|x) (1 - Q_\theta(y|x))}{Q_\theta(y|x) (1 - Q_\theta(y|x)) + 1 - Q_\theta(y|x)} \\ &= \frac{Q_\theta(y|x)}{1 + Q_\theta(y|x)} \geq \frac{Q_\theta(y|x)}{2}. \end{aligned} \quad (27)$$

We obtain the upper bound of $h_\theta^y(x)$ by combining Eq. (26), Eq. (27) and the assumption $\log \sum_{k=1}^K \exp(h_\theta^k(x)) \leq z$,

$$h_\theta^y(x) \leq \log(2\lambda - 2) + z, \forall (x, y) \sim \mathcal{P}_M. \quad (28)$$

According to Jensen's inequality, Khintchine-Kahane inequality [75] and Eq. (28), we have

$$\begin{aligned} \mathbb{E}_{\sigma \in \{\pm 1\}^N} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^N \sigma_i h(x_i) \right] &\leq \sup_{h \in \mathcal{H}} \sqrt{\sum_{i=1}^N \|h(x_i)\|^2} \\ &\leq LK \sqrt{N} (\log(2\lambda - 2) + z). \end{aligned} \quad (29)$$

We complete the proof by substituting Eq. (22) and Eq. (29) into Eq. (21).

REFERENCES

- [1] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *5th International Conference on Learning Representations*, 2017, pp. 1–15.
- [2] Z. Allen-Zhu, Y. Li, and Y. Liang, "Learning and generalization in over-parameterized neural networks, going beyond two layers," in *Advances in Neural Information Processing Systems* 32, 2019, pp. 6155–6166.
- [3] J. Zhang and X. Wu, "Multi-label inference for crowdsourcing," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2738–2747.
- [4] P. Awasthi, A. Blum, N. Haghtalab, and Y. Mansour, "Efficient PAC learning from the crowd," in *Proceedings of the 30th Conference on Learning Theory*, vol. 65, 2017, pp. 127–150.

- [5] Z. Zhao, L. Cao, and K.-Y. Lin, "Revealing the distributional vulnerability of discriminators by implicit generators," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8888–8901, 2023.
- [6] D. Macêdo, T. I. Ren, C. Zanchettin, A. L. I. Oliveira, and T. B. Ludermir, "Entropic out-of-distribution detection: Seamless detection of unknown examples," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 6, pp. 2350–2364, 2022.
- [7] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. A. DePristo, J. V. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," in *Advances in Neural Information Processing Systems* 32, 2019, pp. 14 680–14 691.
- [8] Y. Yan, R. Rosales, G. Fung, S. Ramanathan, and J. G. Dy, "Learning from multiple annotators with varying expertise," *Mach. Learn.*, vol. 9, no. 3, pp. 291–327, 2014.
- [9] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. C. Courville, Y. Bengio, and S. Lacoste-Julien, "A closer look at memorization in deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 233–242.
- [10] N. Konstantinov and C. Lampert, "Robust learning from untrusted sources," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 3488–3498.
- [11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* 25, 2012, pp. 1106–1114.
- [12] T. Ishida, G. Niu, W. Hu, and M. Sugiyama, "Learning from complementary labels," in *Advances in Neural Information Processing Systems* 30, 2017, pp. 5639–5649.
- [13] Y. Zhang, F. Liu, Z. Fang, B. Yuan, G. Zhang, and J. Lu, "Learning from a complementary-label source domain: Theory and algorithms," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 12, pp. 7667–7681, 2022.
- [14] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *5th International Conference on Learning Representations*, 2017, pp. 1–12.
- [15] L. Cao, "Non-iidness learning in behavioral and social data," *Comput. J.*, vol. 57, no. 9, pp. 1358–1370, 2014.
- [16] A. Pentina and C. H. Lampert, "Lifelong learning with non-i.i.d. tasks," in *Advances in Neural Information Processing Systems*, vol. 28, 2015, pp. 1–9.
- [17] S. Jian, L. Cao, K. Lu, and H. Gao, "Unsupervised coupled metric similarity for non-iid categorical data," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 9, pp. 1810–1823, 2018.
- [18] G. Pang, L. Cao, and L. Chen, "Homophily outlier detection in non-iid categorical data," *Data Min. Knowl. Discov.*, vol. 35, no. 4, pp. 1163–1224, 2021.
- [19] M. Luo, F. Chen, D. Hu, Y. Zhang, J. Liang, and J. Feng, "No fear of heterogeneity: Classifier calibration for federated learning with non-iid data," *Advances in Neural Information Processing Systems* 34, pp. 5972–5984, 2021.
- [20] C. Zhu, L. Cao, and J. Yin, "Unsupervised heterogeneous coupling learning for categorical representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 533–549, 2022.
- [21] Y. Liu, Z. Li, C. Zhou, Y. Jiang, J. Sun, M. Wang, and X. He, "Generative adversarial active learning for unsupervised outlier detection," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1517–1528, 2020.
- [22] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396–4415, 2023.
- [23] V. S. Sheng, J. Zhang, B. Gu, and X. Wu, "Majority voting and pairing with multiple noisy labeling," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 7, pp. 1355–1368, 2019.
- [24] S. Liang, Y. Li, and R. Srikanth, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *6th International Conference on Learning Representations*, 2018, pp. 1–27.
- [25] A. Malinin and M. J. F. Gales, "Predictive uncertainty estimation via prior networks," in *Advances in Neural Information Processing Systems* 31, 2018, pp. 7047–7058.
- [26] K. Lee, H. Lee, K. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," in *6th International Conference on Learning Representations*, 2018, pp. 1–16.
- [27] Z. Zhao, L. Cao, and K.-Y. Lin, "Out-of-distribution detection by cross-class vicinity distribution of in-distribution data," *IEEE Trans. Neural Networks Learn. Syst.*, pp. 1–12, 2023.
- [28] S. Bhatia, B. Hooi, L. Akoglu, S. Chatterjee, X. Jiang, and M. Gupta, "ODD: Outlier detection and description," in *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2021, pp. 4108–4109.
- [29] Y. Wang, W. Liu, X. Ma, J. Bailey, H. Zha, L. Song, and S. Xia, "Iterative learning with open-set noisy labels," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8688–8696.
- [30] M. M. Breunig, H. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 2000, pp. 93–104.
- [31] M. Sugiyama and K. M. Borgwardt, "Rapid distance-based outlier detection via sampling," in *Advances in Neural Information Processing Systems* 26, 2013, pp. 467–475.
- [32] D. Hendrycks, M. Mazeika, and T. G. Dietterich, "Deep anomaly detection with outlier exposure," in *7th International Conference on Learning Representations*, 2019, pp. 1–18.
- [33] A. R. Rivera, A. Khan, I. E. I. Bekkouch, and T. S. Sheikh, "Anomaly detection based on zero-shot outlier synthesis and hierarchical feature distillation," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 1, pp. 281–291, 2022.
- [34] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, pp. 2096–2030, 2016.
- [35] D. Li, Y. Yang, Y. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press, 2018, pp. 3490–3497.
- [36] M. Levi, I. Attias, and A. Kontorovich, "Domain invariant adversarial learning," *Trans. Mach. Learn. Res.*, vol. 2022, pp. 1–25, 2022.
- [37] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Advances in Neural Information Processing Systems* 31, 2018, pp. 1647–1657.
- [38] J. Li, S. Shang, and L. Chen, "Domain generalization for named entity boundary detection via metalearning," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 9, pp. 3819–3830, 2021.
- [39] N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Advances in Neural Information Processing Systems* 26, 2013, pp. 1196–1204.
- [40] X. Xia, T. Liu, N. Wang, B. Han, C. Gong, G. Niu, and M. Sugiyama, "Are anchor points really indispensable in label-noise learning?" in *Advances in Neural Information Processing Systems* 32, 2019, pp. 6835–6846.
- [41] H. Köhler and S. Link, "Qualitative cleaning of uncertain data," in *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, 2016, pp. 2269–2274.
- [42] L. Jiang, Z. Zhou, T. Leung, L. Li, and L. Fei-Fei, "MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 2309–2318.
- [43] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. W. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in Neural Information Processing Systems* 31, 2018, pp. 8536–8546.
- [44] E. Malach and S. Shalev-Shwartz, "Decoupling "when to update" from "how to update"," in *Advances in Neural Information Processing Systems* 30, 2017, pp. 960–970.
- [45] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 1919–1925.
- [46] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *International Conference on Computer Vision*, 2019, pp. 322–330.
- [47] S. E. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," in *3rd International Conference on Learning Representations*, 2015, pp. 1–11.
- [48] F. Ma, Y. Wu, X. Yu, and Y. Yang, "Learning with noisy labels via self-reweighting from class centroids," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 11, pp. 6275–6285, 2022.
- [49] H. Zhang, A. Li, J. Guo, and Y. Guo, "Hybrid models for open set recognition," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 102–117.
- [50] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, "A comprehensive survey of scene graphs: Generation and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1–26, 2023.
- [51] M. Li, P. Huang, X. Chang, J. Hu, Y. Yang, and A. Hauptmann, "Video pivoting unsupervised multi-modal machine translation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3918–3932, 2023.

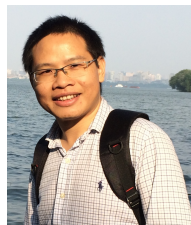
- [52] L. Zhang, X. Chang, J. Liu, M. Luo, Z. Li, L. Yao, and A. Hauptmann, "TN-ZSTAD: transferable network for zero-shot temporal activity detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3848–3861, 2023.
- [53] C. Yan, X. Chang, Z. Li, W. Guan, Z. Ge, L. Zhu, and Q. Zheng, "Zeronas: Differentiable generative adversarial networks search for zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9733–9740, 2022.
- [54] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 41–48.
- [55] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Advances in Neural Information Processing Systems 23*, 2010, pp. 1189–1197.
- [56] L. Jiang, D. Meng, S. Yu, Z. Lan, S. Shan, and A. G. Hauptmann, "Self-paced learning with diversity," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 2078–2086.
- [57] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, 2002.
- [58] N. Golowich, A. Rakhlin, and O. Shamir, "Size-independent sample complexity," in *Conference On Learning Theory*, vol. 75, 2018, pp. 297–299.
- [59] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, no. 1-2, pp. 151–175, 2010.
- [60] S. M. Kakade, K. Sridharan, and A. Tewari, "On the complexity of linear prediction: Risk bounds, margin bounds, and regularization," in *Advances in Neural Information Processing Systems 21*, 2008, pp. 793–800.
- [61] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 1321–1330.
- [62] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.
- [63] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011, pp. 1–9.
- [64] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [66] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.
- [67] N. Ma, X. Zhang, H. Zheng, and J. Sun, "ShuffleNet V2: practical guidelines for efficient CNN architecture design," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 122–138.
- [68] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.
- [69] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size," in *CoRR*, 2016, pp. 1–13.
- [70] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269.
- [71] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning From Theory to Algorithms*. Cambridge University Press, 2014.
- [72] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations*, 2015, pp. 1–15.
- [73] M. M. Bejani and M. Ghatee, "A systematic review on overfitting control in shallow and deep neural networks," *Artif. Intell. Rev.*, vol. 54, no. 8, pp. 6391–6438, 2021.
- [74] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. MIT Press, 2018.
- [75] J. Hoffmann-Jorgensen, J. Kuelbs, and M. B. Marcus, *On the Rademacher Series, Probability in Banach spaces*, 9. Springer Science & Business Media, 2012, vol. 35.



Zhilin Zhao received his Ph.D. degree from the University of Technology Sydney in 2022. Before that, he earned his B.S. and M.S. degrees from the School of Data and Computer Science at Sun Yat-Sen University, China, in 2016 and 2018, respectively. He is currently a Post-Doctoral Fellow at Macquarie University, Australia. His research interests encompass generalization analysis, distribution discrepancy estimation, and out-of-distribution detection.



Longbing Cao received a PhD degree in pattern recognition and intelligent systems at Chinese Academy of Sciences in 2002 and another PhD in computing sciences at University of Technology Sydney in 2005. He is the Distinguished Chair Professor in AI at Macquarie University and an Australian Research Council Future Fellow (professorial level). His research interests include AI and intelligent systems, data science and analytics, machine learning, behavior informatics, and enterprise innovation.



Chang-Dong Wang received the Ph.D. degree in computer science in 2013 from Sun Yat-sen University, Guangzhou, China. He joined Sun Yat-sen University in 2013, where he is currently an associate professor with School of Computer Science and Engineering. His current research interests include machine learning and data mining. He has published over 80 scientific papers in international journals and conferences such as IEEE TPAMI, IEEE TKDE, IEEE TCYB, IEEE TNNLS, KDD, AAAI, IJCAI, CVPR, ICDM, CIKM and SDM. His ICDM 2010 paper won the Honorable Mention for Best Research Paper Awards. He won 2012 Microsoft Research Fellowship Nomination Award. He was awarded 2015 Chinese Association for Artificial Intelligence (CAAI) Outstanding Dissertation. He is an Associate Editor in Journal of Artificial Intelligence Research (JAIR).