# Multi-Objective Interpolation Training for Robustness to Label Noise

Diego Ortego, Eric Arazo, Paul Albert, Noel E. O'Connor and Kevin McGuinness
Insight Centre for Data Analytics, Dublin City University (DCU)
{diego.ortego, eric.arazo}@insight-centre.org

## Abstract

*Deep neural networks trained with standard cross-entropy loss memorize noisy labels, which degrades their performance. Most research to mitigate this memorization proposes new robust classification loss functions. Conversely, we propose a Multi-Objective Interpolation Training (MOIT) approach that jointly exploits contrastive learning and classification to mutually help each other and boost performance against label noise. We show that standard supervised contrastive learning degrades in the presence of label noise and propose an interpolation training strategy to mitigate this behavior. We further propose a novel label noise detection method that exploits the robust feature representations learned via contrastive learning to estimate per-sample soft-labels whose disagreements with the original labels accurately identify noisy samples. This detection allows treating noisy samples as unlabeled and training a classifier in a semi-supervised manner to prevent noise memorization and improve representation learning. We further propose MOIT+, a refinement of MOIT by fine-tuning on detected clean samples. Hyperparameter and ablation studies verify the key components of our method. Experiments on synthetic and real-world noise benchmarks demonstrate that MOIT/MOIT+ achieves state-of-the-art results. Code is available at https://git.io/JI40X.*

## 1. Introduction

Building a new dataset usually involves manually labeling every sample for the particular task at hand. This process is cumbersome and limits the creation of large datasets, which are usually necessary for training deep neural networks (DNNs) in order to achieve the required performance. Conversely, automatic data annotation based on web search and user tags [29, 22] leverages the use of larger data collections at the expense of introducing some incorrect labels. This label noise degrades DNN performance [3, 52] and this poses an interesting challenge that has recently gained a lot of interest in the research community [45, 41, 23, 50, 12, 1, 28, 55, 13, 31].

In image classification problems, label noise usually involves different noise distributions [22, 55]. In-distribution noise types consist of samples with incorrect labels, but whose image content belongs to the dataset classes. When in-distribution noise is synthetically introduced, it usually follows either an asymmetric or symmetric random distribution. The former involves label flips to classes with some semantic meaning, e.g., a cat is flipped to a tiger, while the latter does not. Furthermore, web label noise types are usually dominated by out-of-distribution samples where the image content does not belong to the dataset classes. Recent studies show that all label noise types impact DNN performance, although performance degrades less with web noise [22, 34].

Robustness to label noise is usually pursued by identifying noisy samples to: reduce their contribution in the loss [23, 11], correct their label [1, 28], or abstain their classification [42]. Other methods exploit interpolation training [53], regularizing label noise information in DNN weights [13], or small sets of correctly labeled data [18, 55]. However, most previous methods rely exclusively on classification losses and little effort has being directed towards incorporating similarity learning frameworks [32], i.e. directly learning image representations rather than a class mapping [45].

Similarity learning frameworks are very popular in computer vision for a variety of applications including face recognition [44], fine-grained retrieval [37], or visual search [35]. These methods learn representations for samples of the same class (positive samples) that lie closer in the feature space than those of samples from different classes (negative samples). Many traditional methods are based on sampling pairs or triplets to measure similarities [7, 19]. However, supervised and unsupervised contrastive learning approaches that consider a high number of negatives have recently received significant attention due to their success in unsupervised learning [5, 14, 27]. In the context of label noise, there are some attempts at training with simple similarity learning losses [45], but there are, to the best of our knowledge, no works exploring more recent contrastive learning losses [24].

This paper proposes Multi-Objective Interpolation Training (MOIT), a framework to robustly learn in the presence of label noise by jointly exploiting synergies between contrastive and semi-supervised learning. The former introduces a regularization of the contrastive loss in [24] to learn noise-robust representations that are key for accurately detecting noisy samples and, ultimately, for semi-supervised learning. The latter performs robust image classification and boosts performance. Our MOIT+ refinement further demonstrates that fine-tuning on the detected clean data can boost performance. MOIT/MOIT+ achieves state-of-the-art results across a variety of datasets (CIFAR-10/100 [26], mini-ImageNet [22], and mini-WebVision [29]) with both synthetic and real-world web label noise. Our main contributions are as follows:

1. A multi-objective interpolation training (MOIT) framework where supervised contrastive learning and semi-supervised learning help each other to robustly learn in the presence of both synthetic and web label noise under a single hyperparameter configuration.

2. An interpolated contrastive learning (ICL) loss that imposes linear relations both on the input and the contrastive loss to mitigate the performance degradation observed for the supervised contrastive learning loss in [24] when training with label noise.

3. A novel label noise detection strategy that exploits the noise-robust feature representations provided by ICL to enable semi-supervised learning. This detection strategy performs a $k$-nearest neighbor search to infer per-sample label distributions whose agreements with the original labels identify correctly labeled samples.

4. A fine-tuning strategy over detected clean data (MOIT+) that further boosts performance based on simple noise robust losses from the literature.

## 2. Related work

We briefly review recent image classification methods aiming at mitigating the effect of label noise on DNNs and recent contrastive learning methods.

**Noise rate estimation** Using a label noise transition matrix can mitigate label noise [36, 18, 48]. Patrini et al. [36] proposed to correct the softmax classification using a transition matrix. The estimation of this matrix is, however, challenging. The authors in [48] estimate the matrix by exploiting detected noisy samples that are similar to anchor points (i.e. highly reliable detected clean samples), while Hendrycks et al. [18] directly use a set of clean samples.

**Noisy sample rejection** Rejecting or reducing the contribution to the optimization objective of noisy samples can increase model robustness [23, 10, 45, 22]. Jiang et al. [23] propose a teacher-student framework where the teacher estimates per-sample weights to guide the student training. Defining per-sample weights is also exploited in [10] via an unsupervised estimation of data complexity. Nguyen et al. [33] iteratively refine a clean set to train on by measuring label agreements with ensembled network predictions. Cross-network disagreements and updates [11] lead to robust learning by training on selected clean data [51]. Also, [46] propose a loss for standard training together with cross-network consistency to select the clean samples to train on.

**Noisy label correction** Correcting noisy labels to replace or balance their influence is widely used in previous works [38, 12, 1, 30]. Bootstrapping loss [38] correction approaches exploit a perceptual term that introduces reliance on a new label given by either the model prediction with fixed [38] or dynamic [1] importance, or class prototypes [12]. More recently, Liu et al. [30] introduced a perceptual term that maximizes the inner product between the model output and the targets without need for per-sample weights.

**Noisy label rejection** Rejecting the original labels by relabeling all samples with the network predictions [41] or learned label distributions [50] mitigates the effect of label noise. Recently, several approaches perform semi-supervised learning [9, 25, 34] by treating detected noisy samples as unlabeled, thus rejecting their labels while exploiting the image content. Their main differences are in the noise detection mechanism: Ding et al. [9] exploit high certainty agreements between the network predictions and labels, Kim et al. [25] use high softmax probabilities after performing negative learning, and Ortego et al. [34] look at the agreements between the original and relabeled labels using [41].

**Other label noise methods** Zhang et al. [53] proposed an interpolation training strategy, *mixup*, that greatly prevents label noise memorization and has been adopted by many other methods [1, 28, 22, 34, 30]. Harutyunyan et al. [13] quantify the amount of memorized information via the Shannon mutual information between neural network weights and the vector of all training labels, and encourage this to be small. Thulasidasan et al. [42] add an abstention class to be predicted by noisy samples due to an abstention penalty introduced in the loss. Robust loss functions are studied in several works by jointly exploiting the benefits of mean absolute error and cross-entropy losses [54], a generalized version of mutual information insensitive to noise [49], or [31] combinations of robust loss functions that mutually boost each other. Furthermore, several strategies to prevent memorization can be exploited together and DivideMix [28] is a good example as it uses interpolation training, cross-network agreements, semi-supervised learning, and label correction.

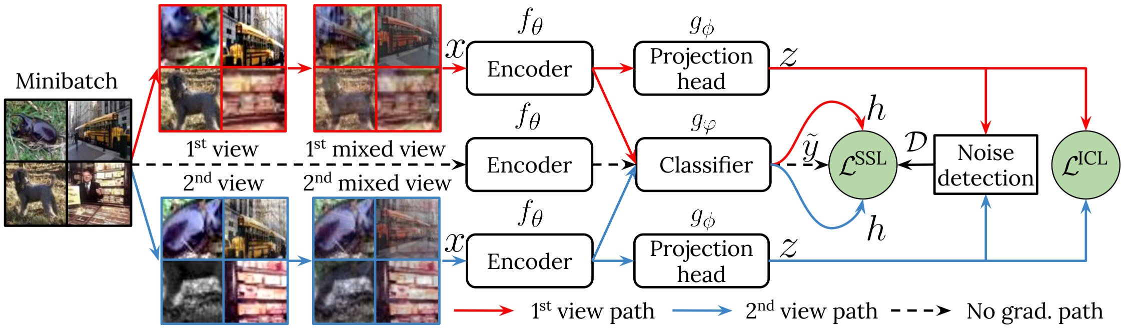

Figure 1. Multi-Objective Interpolation Training (MOIT) for improved robustness to label noise. We interpolate samples and impose the same interpolation in the supervised contrastive learning loss $\mathcal{L}^{ICL}$ and the semi-supervised classification loss $\mathcal{L}^{SSL}$ that we jointly use during training. Label noise detection is performed at every epoch to enable semi-supervised learning and its result is used after training to fine-tune the encoder and classifier to further boost performance.

**Contrastive representation learning** Recent works in self-supervised learning [27, 5, 24] have demonstrated the potential of contrastive based similarity learning frameworks for representation learning. These methods maximize (minimize) similarities of positive (negative) pairs. Adequate data augmentation [43], large amounts of negative samples via large batch size [5] or memory banks [14, 47], and careful network architecture designs [6] are usually important for better performance. Regarding the label noise scenario for image classification, no works explore the impact of incorrect labels on contrastive learning and only Wang et al. [45] incorporate a simple similarity learning objective.

## 3. Method

We target learning robust feature representations in the presence of label noise. In particular, we adopt the contrastive learning approach from [24] and randomly sample $N$ images to apply two random data augmentation operations to each, thus generating two data views. The resulting training minibatch $\{(x_i, y_i)\}_{i=1}^{2N}$ of image-label pairs $x_i$ and $y_i$ consists of $2N$ images. Every image is mapped to a low-dimensional representation $z_i$ by learning an encoder network $f_\theta$ and a projection network $g_\phi$ with parameters $\theta$ and $\phi$. In particular, an intermediate embedding $v_i = f_\theta(x_i)$ is generated and subsequently transformed into the representation $w_i = g_\phi(v_i)$. Finally, $z_i = w_i / \|w_i\|_2$ is the $L_2$-normalized low-dimensional representation used to learn based on the per-sample loss:

$$\mathcal{L}_i(z_i, y_i) = \frac{1}{2N_{y_i} - 1} \sum_{j=1}^{2N} \mathbb{1}_{i \neq j} \, \mathbb{1}_{y_i = y_j} P_{i,j}, \qquad (1)$$

$$P_{i,j} = -\log \frac{\exp(z_i \cdot z_j / \tau)}{\sum_{r=1}^{2N} \mathbb{1}_{r \neq i} \exp(z_i \cdot z_r / \tau)}, \qquad (2)$$

where $P_{i,j}$ denotes the $j$-th component of the temperature $\tau$ scaled softmax distribution of inner products $z_i \cdot z_j$ of representations from the positive pair of samples $x_i$ and $x_j$, which can be interpreted as a probability. $P_{i,j}$ is aggregated in Eq. 1 across all $N_{y_i}$ samples $x_j$ in the minibatch sharing label with $x_i$ ($y_i = y_j$) except for the self-contrast case ($i = j$), as defined by the indicator function $\mathbb{1}_B \in \{0, 1\}$ that returns 1 when condition $B$ is fulfilled and 0 otherwise. Minimizing $\mathcal{L}_i$ implies adjusting $f_\theta$ and $g_\phi$ to pull together the feature representations $z_i$ and $z_j$ when they share the same label ($y_i = y_j$), while pushing them apart when they do not. Also, the gradient analysis in [24] reveals that Eq. 1 focuses on hard positives/negatives rather than easy ones. Note that having two data views implies that $\mathcal{L}_i$ contains an unsupervised contribution equivalent to the *NT-Xent* loss [5].

In the presence of label noise, Eq. 1 incorrectly selects positive/negative samples, which degrades the feature representation $z$ (see Tab. 2). To overcome this limitation and perform robust image classification under label noise conditions, we propose a Multi-Objective Interpolation Training (MOIT) framework that consists of: *i)* a regularization technique to prevent memorization when training with the supervised contrastive learning loss (Sec. 3.1), *ii)* a semi-supervised classification strategy based on a novel label noise detection strategy that exploits the noise-robust representation $z$ to measure agreements with the original labels $y$ and tag noisy samples as unlabeled (Sec. 3.2), and *iii)* a classifier refinement on clean data to boost classification performance (Sec. 3.3). Fig. 1 shows an overview of MOIT.

### 3.1. Interpolated Contrastive Learning

Interpolation training strategies have demonstrated excellent performance in classification frameworks [53, 40, 39], and have further shown promising results to prevent label noise memorization [53, 1, 28, 22]. Inspired by this success, we propose Interpolated Contrastive Learning (ICL), a novel adaptation of mixup data augmentation [53] for supervised contrastive learning. ICL performs convex combinations of pairs of samples as

$$x_i = \lambda x_a + (1 - \lambda)\, x_b, \tag{3}$$

where $\lambda \in [0, 1] \sim \text{Beta}\,(\alpha, \alpha)$ and $x_i$ denotes the training sample that combines two minibatch samples $x_a$ and $x_b$, and imposes a linear relation in the contrastive loss:

$$\mathcal{L}_i^{MIX} = \lambda \mathcal{L}_i\,(z_i, y_a) + (1 - \lambda)\, \mathcal{L}_i\,(z_i, y_b)\,. \tag{4}$$

The first and second terms in Eq. 4 consider, respectively, positive samples from the class $y_a$ ($y_b$) given by the first (second) sample $x_a$ ($x_b$). The selection of positive/negative samples involves considering a unique class for every mixed example. However, in most cases the input samples contain two classes as a result of the interpolation, where $\lambda$ determines the dominant one. We assign this dominant class to every sample for positive/negative sampling. Intuitively, ICL makes it harder to pull together clean and noisy samples with the same label, as noisy samples are interpolated with either another clean sample that provides a clean pattern beneficial for training or another noisy sample that makes it harder to memorize the noisy pattern.

**Memory bank** The number of positives and negatives selected for contrastive learning depends on the minibatch size and the number of dataset classes. Therefore, unless a large minibatch is used during training, few positive and negative samples are selected, which negatively affects the training process [24]. To address limitations in computing resources, we introduce the memory bank proposed in [47] to perform robust similarity learning despite using relatively small minibatches compared to those in [24]. In particular, we define a memory to store the last $M$ feature representations from previous minibatches and define a loss term $\mathcal{L}_i^{MEM}$ similar to $\mathcal{L}_i^{MIX}$ in Eq. 4. While $\mathcal{L}_i^{MIX}$ is estimated contrasting the $2N$ minibatch samples across them, $\mathcal{L}_i^{MEM}$ contrasts the $2N$ samples with the $M$ memory samples, thus extending the number of positive and negative samples. The final ICL loss then aggregates the average batch and memory losses:

$$\mathcal{L}^{ICL} = \mathcal{L}^{MIX} + \mathcal{L}^{MEM}. \tag{5}$$

Sec. 4.3 shows the benefits of using ICL loss instead of the original loss in [24] and Sec. 4.4 demonstrates the effect of the memory bank on the overall method.

## 3.2. Semi-Supervised Classification

The goal is to predict a class $c \in \{1, \ldots, C\}$ by learning a second mapping $h(v) = g_\varphi\,(v)$ to the class space, where $C$ is the number of classes. Naïvely training a classifier in the presence of label noise leads to noise memorization [3, 52], which degrades the performance. Semi-supervised learning, where noisy labels are discarded, can mitigate this memorization [9, 28, 34]. We, therefore, propose to jointly adopt semi-supervised learning with ICL. The former boosts the performance achievable by the latter, while the latter enables accurate label noise detection necessary for good performance in the former.

**Label noise detection** We propose to measure agreements between the feature representation $z_i$ (robust to label noise) and the original label $y_i$ to identify mislabeled samples. To quantify this agreement, we start by estimating a class probability distribution from the representation $z_i$ by doing a $k$-nearest neighbor (k-NN) search:

$$p\,(c \mid x_i) = \frac{1}{K} \sum_{\substack{k=1 \\ x_k \in \mathcal{N}_i}}^{K} \mathbb{1}_{y_k \neq c}, \tag{6}$$

where $\mathcal{N}_i$ denotes the neighbourhood of $K$ closest images to $x_i$ according to the feature representation $z$. Eq. 6 then counts the number of samples per class in the local neighborhood $\mathcal{N}_i$ and normalizes the counts to estimate a probability distribution. This distribution can be interpreted as a soft-label that can be compared with the original label to identify potential disagreements, i.e. noisy samples. However, the labels $y$ might be noisy, thus biasing the estimation of $p$. We, therefore, estimate a corrected distribution $\hat{p}$ using:

$$\hat{p}\,(c \mid x_i) = \frac{1}{K} \sum_{\substack{k=1 \\ x_k \in \mathcal{N}_i}}^{K} \mathbb{1}_{\hat{y}_k \neq c}, \tag{7}$$

where we introduce corrected labels $\hat{y}$ that are estimated taking the dominant label in $\mathcal{N}_i$, i.e. $\hat{y} = \arg\max_c\, p\,(c \mid x)$. Finally, the disagreement between the corrected distribution $\hat{p}\,(c \mid x_i)$ and the label noise distribution given by the original label $y_i$ is measured by the cross-entropy

$$d_i = -y_i^T \log\,(\hat{p})\,, \tag{8}$$

where $T$ denotes the transpose operation. The higher $d_i$, the higher the disagreement between distributions and the more likely $x_i$ is a noisy sample. We select clean samples for each class $c$ based on $d_i$ using:

$$\mathcal{D}_c = \{(x_i, y_i) : d_i \leq \gamma_c\}\,, \tag{9}$$

where $\gamma_c$ is a per-class threshold on $d_i$, which is dynamically defined to ensure a balanced clean set across classes. To perform this balancing, we use the median of per-class agreements between the corrected label $\hat{y}_i$ and the original label $y_i$ across all classes. Sec. 4.4 illustrates the importance of this balancing strategy as well as the corrected distribution $\hat{p}$ over $p$ for achieving better performance. Note that a k-NN noise detection that resembles Eq. 6 has been recently proposed in [4]. However, we differ in that we propose a corrected version in Eq. 7 that surpasses the straightforward k-NN of Eq. 6 (see Tab. 3), we use k-NN during training, and always avoid using a trusted clean set.

**Semi-supervised learning** We learn the classifier by performing semi-supervised learning where samples in $\mathcal{D}$ are considered as labeled and the remaining samples as unlabeled. To leverage these unlabeled samples, pseudo-labeling [2] based on interpolated samples is applied by defining the objective

$$\mathcal{L}_i^{SSL} = -\lambda \tilde{y}_a^T \log(h_i) - (1-\lambda)\, \tilde{y}_b^T \log(h_i), \tag{10}$$

where the pseudo-label $\tilde{y}_a$ ($\tilde{y}_b$) for $x_a$ ($x_b$) is estimated as

$$\tilde{y}_a = \begin{cases} y_a, & x_a \in \mathcal{D}_c \\ \bar{h}_a, & x_a \notin \mathcal{D}_c \end{cases}, \tag{11}$$

where $\bar{h}_a$ is the softmax prediction for image $x_a$ without data augmentation. The final Multi-Objective Interpolation Training (MOIT) optimizes the loss:

$$\mathcal{L}^{MOIT} = \mathcal{L}^{ICL} + \mathcal{L}^{SSL}. \tag{12}$$

In summary, the proposed MOIT framework enables robust training in the presence of label noise by learning robust representations via contrastive learning that help in achieving successful noise detection that discards noisy labels and enables semi-supervised learning for classification. Note that the method needs to learn useful features before performing accurate noise detection; thus we start training with $\tilde{y} = y, \forall x$ in $\mathcal{L}^{SSL}$, i.e. a normal supervised training. We start doing semi-supervised learning once reasonable features to search for reliable nearest neighbors in Eq. 7 are learned and the clean sample detection is made reliable. We assume that good features are available soon after reducing the learning rate, given that there is little risk of overfitting noisy labels at earlier epochs when using a high learning rate, as often reported in the literature [41, 50, 1].

### 3.3. Classification refinement

Supervised pre-training on relatively clean datasets such as ImageNet [8] has proved to mitigate label noise memorization [17, 22]. We, therefore, refine our MOIT predictions by fine-tuning $f_\theta$ and re-training $g_\varphi$ on our detected clean set $\mathcal{D}$ using a constant low learning rate. We name this fine-tuning stage MOIT+. We train using mixup [53] and later introduce hard bootstrapping loss correction [38] to deal with possible low amounts of label noise present in $\mathcal{D}$, thus defining the following training objective:

$$\mathcal{L}_i^{MOIT+} = -\lambda \left[ (\delta y_a + (1-\delta)\, \tilde{y}_a)^T \log(h_i) \right] - \\ (1-\lambda) \left[ (\delta y_b + (1-\delta)\, \tilde{y}_b)^T \log(h_i) \right], \tag{13}$$

where $\lambda$ is the mixing coefficient from [53] as we are interpolating images as explained in Eq. 3, and $\delta$ is a weight to balance the contribution of the original labels ($y_a$ and $y_b$) or the network predictions ($\tilde{y}_a$ and $\tilde{y}_b$). This training objective is similar to that used in [1], but different in that we do not train from scratch using all data, or need to infer per sample $\delta$ weights. Instead, we set $\delta = 0.8$ as done in [38] to give more importance to the original labels, which is reasonable given that the training uses the detected clean data $\mathcal{D}$. Note that $\tilde{y}_a = \arg\max_c \bar{h}_a$ ($\tilde{y}_b = \arg\max_c \bar{h}_b$) is the network prediction for $x_a$ ($x_b$) without data augmentation. As commented before, MOIT+ starts with a mixup training without bootstrapping (i.e. $\delta = 1.0$) during the initial epochs to allow adequate re-training of $g_\varphi$ before trusting its predictions.

## 4. Experiments

We first run experiments on the standard benchmarks for synthetic noise in CIFAR-100 [26] aiming at analyzing the different components of our method. We further perform comparative evaluations against related work using synthetic label noise in CIFAR-10/100, controlled web noise in mini-ImageNet [22], and the uncontrolled web noise from the WebVision dataset [29].

### 4.1. Datasets

The CIFAR-10/100 datasets [26] contain 50K (10K) small resolution images for training (test). For hyperparameter and ablation studies, we keep 5K training samples for validation using their correct labels. However, to facilitate comparison with related work, we train with the full 50K samples and use the 10K test set for evaluation (reporting accuracy in the last epoch). For noise addition, we follow the criteria in [50]: symmetric noise is introduced by randomly flipping the labels of a percentage of the training set to incorrect labels; asymmetric noise uses label flips to incorrect classes "truck → automobile, bird → airplane, deer → horse, cat → dog" in CIFAR-10, whereas in CIFAR-100 label flips are done circularly within the super-classes.

Jiang et al. [22] propose to use mini-ImageNet and Stanford Cars to introduce both web and symmetric in-distribution noise in a controlled manner with different noise ratios. We adopt the mini-ImageNet web noise dataset for evaluation in a real scenario with several ratios, which consists of 100 classes with 50K (5K) samples for training (validation). For further evaluation against web noise, we adopt the mini-WebVision dataset [28] that uses the top-50 classes from the Google image subset of WebVision [29].

### 4.2. Training details

We use a PreAct ResNet-18 (PRN-18) [16] as encoder network in CIFAR following [1], while for mini-ImageNet we use the ResNet-18 (RN-18) from [20] used in mini-ImageNet for few-shot learning. For mini-WebVision we use a standard

Table 1. Training details. We always train from scratch. LR: Learning rate. B: Bootstrapping.

| | CIFAR | mini-ImageNet | mini-WebVision |
|---|---|---|---|
| Resolution | $32 \times 32$ | $84 \times 84$ | $224 \times 224$ |
| Batch size | 128 | 64 | 64 |
| Mem. size | 20K | 100K | 50K |
| Network | PRN-18 | RN-18 | RN-18 |
| Epochs | 250 | 130 | 130 |
| Optimizer | SGD, momentum 0.9, weight decay $10^{-4}$ | | |
| Initial LR | 0.1 | 0.1 | 0.1 |
| LR decay | 125, 200 | 80, 105 | 80, 105 |
| Decay factor | $\times 0.1$ | $\times 0.1$ | $\times 0.1$ |
| SSL epoch | 130 | 85 | 85 |
| Decay factor | $\times 0.1$ | $\times 0.1$ | $\times 0.1$ |
| Epochs (MOIT+) | 70 | 50 | 50 |
| LR (MOIT+) | 0.001 (not reduced) | | |
| B epoch (MOIT+) | 30 | 20 | 20 |

RN-18 [15]. We do not evaluate using other frameworks in mini-ImageNet or WebVision [22, 28] due to limitations of our computing resources. We, conversely, re-run the official implementation of top-performing and recent methods [53, 28, 30] in our framework. As projection head and classifier, we always use a linear layer that maps, respectively, to a feature vector $z$ of 128 dimensions and to the class space.

Table 1 presents the training details for MOIT and MOIT+. We interpolated input samples as proposed in [53] with $\alpha = 1$ (i.e. $\lambda$ is sampled from a uniform distribution), and apply standard strong data augmentations to achieve successful contrastive learning[1] in MOIT: random resized crops, horizontal flips, color jitter and gray scale transformations. For MOIT+ and all other methods, mixup as well as standard augmentations are used (CIFAR: random horizontal flips and random 4 pixel translations, mini-ImageNet and mini-WebVision: random resized crops and random horizontal flips). We double the epochs in MOIT+ for 80% noise in CIFAR-10/100 as there are few selected clean samples, which make epochs extremely short. We always use $\tau = 0.1$ temperature scaling for contrastive learning and increase the memory size in mini-ImageNet and mini-WebVision to deal with reduced batch size. Note that MOIT+ finetunes the model in the last epoch when training MOIT.

In practice, the noise ratio and distribution are not usually known a-priori; we therefore use a common configuration for training our method (mixup $\alpha$, $k$-NN parameter $K$, loss function, $\mathcal{D}$ balancing criterion, $\delta$ for MOIT+), and only modify typical hyperparameters (batch size, memory and epochs). We use the official implementations of DivideMix (DMix) [28] and ELR [30]. However, DMix adopts specific configurations for different datasets and even for different noise ratios and types in the same dataset. To perform as fair as possible a comparison without degrading DMix results, we select a single parametrization of DMix in every dataset based on the most repeated configuration in [28]. This affects the CIFAR configuration (CIFAR-10: $\lambda_u = 0$, CIFAR-100: $\lambda_u = 150$) as mini-WebVision has a unique configuration that we also adopt for mini-ImageNet. We run DMix and ELR for the same number of epochs as our method respecting suggested learning rates and equip ELR with *mixup* for a fair comparison with DMix and our method that both use interpolation training. Note that ELR+ in [30] uses mixup, but we do not use it for comparison as it involves using a second network and a weight averaging.

Table 2. Weighted k-NN evaluation in CIFAR-100.

| | | Symmetric | | Asymmetric | |
|---|---|---|---|---|---|
| | 0% | 40% | 80% | 10% | 40% |
| SCL | 72.66 | 58.32 | 41.00 | 71.11 | 68.00 |
| ICL | 75.30 | 66.38 | 53.60 | 74.34 | 72.04 |
| MOIT | **75.76** | **67.42** | **55.58** | **74.86** | **72.60** |

### 4.3. Supervised contrastive learning and label noise

We start by analyzing supervised contrastive learning behavior in the presence of label noise and how introducing interpolation training impacts the learned representations. We evaluate the quality of representations using a *weighted k-NN* ($k = 200$) evaluation typical in unsupervised learning [21]. Tab. 2 reports this evaluation using the embedding $z$ extracted after the projection head (model from the last training epoch) and the true labels in the training set. This experiments show that Supervised Contrastive Learning (SCL) [24] performance degrades when there is label noise (the noise-free accuracy of 72.66 decreases). The proposed regularization using Interpolated Contrastive Learning (ICL) mitigates label noise drops and outperforms SCL in the noise-free case, validating the utility of imposing a interpolated behavior in the contrastive loss. Note that ICL and MOIT (joint ICL and semi-supervised classification) perform worse in the asymmetric case than in the symmetric case. This occurs due to the former having label flips that keep some semantic meaning (e.g. cat→dog), while the latter does not (e.g. cat→truck). Semantic noise is more informative during ICL, which leads to better performance and less room for semi-supervised learning improvement in MOIT compared to ICL. We train SCL and ICL using a memory bank for 350 epochs with initial learning rate of 0.1, divided by 10 at epochs 200 and 300. Note that contrastive learning frameworks tend to be very sensitive to hyperparameters [5, 14, 24] (learning rate, temperature, data augmentation, etc.), a behavior that we also observed when training them alone in the presence of label noise. We experimentally found that averaging the contrastive losses of the minibatch

[1] https://github.com/HobbitLong/SupContrast

Table 3. Classification accuracy for different noise detection strategies and $K$ values for 40% asymmetric noise in CIFAR-100.

|  | $K$ | 5 | 10 | 25 | 50 | 100 | 150 | 200 | 250 | 300 | 350 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| k-NN ($p$) | Acc. | 59.42 | 61.74 | 64.84 | 66.10 | 67.18 | 67.42 | 67.46 | 67.68 | 67.14 | 66.94 |
| k-NN ($\hat{p}$) | Acc. | 62.28 | 65.30 | 68.58 | 70.56 | 71.16 | 71.22 | 71.24 | **71.42** | 70.98 | 70.80 |

Table 4. Effect on classification accuracy of the balancing strategy for the clean set $\mathcal{D}$ in CIFAR-100. A: Asymmetric. S: Symmetric.

|  | Unbalanced | Min | Max | Median |
|---|---|---|---|---|
| A-40% | 69.58 | 52.88 | 62.58 | **71.42** |
| S-40% | 66.28 | 63.26 | 66.12 | **66.58** |

Table 5. Ablation study for MOIT and MOIT+ in CIFAR-100. A: Asymmetric, S: Symmetric, SSL: semi-supervised learning, M: memory, B: Balanced clean set, r-t C: Re-training classifier, s-DA: strong data augmentation.

|  | S-40% | A-40% |
|---|---|---|
| (MOIT) w/o SSL | 62.82 | 53.73 |
| (MOIT) w/o M | 66.10 | 68.88 |
| (MOIT) w/o B | 66.28 | 69.58 |
| MOIT | **66.58** | **71.42** |
| (MOIT+) w/o r-t C | 69.54 | 73.32 |
| (MOIT+) w/ s-DA | 67.98 | 71.90 |
| MOIT+ | **70.68** | **73.58** |

$\mathcal{L}^{MIX}$ and the memory $\mathcal{L}^{MEM}$ helped convergence in SCL and ICL and used it in this experiment. Adding a classification objective, as done in the proposed MOIT method, stabilizes this behavior and achieves better representations than SCL and ICL alone (see Tab. 2).

### 4.4. Label noise detection analysis

We exploit the feature representation $z$ by searching the closest $K$ neighbors to estimate a corrected soft-label $\hat{p}$ in Eq. 7 and measure agreements with the original labels $y$. Tab. 3 shows that using this corrected soft-label $\hat{p}$ (bottom) rather than the soft-label $p$ from Eq. 6 (top) results in better performance due to improved label noise detection: precision and recall for $\hat{p}$ are 90.83 and 87.84 compared to 80.20 and 84.43 for $p$. The method is also not very sensitive to the value of $K$ once it is set to a high enough value. We adopt $K = 250$ for the remaining experiments. We further study the effect of balancing the clean set $\mathcal{D}$ (see Tab. 4). In particular, we experiment by balancing with the minimum (Min), maximum (Max), or median (used by our method) number of agreements between corrected $\hat{y}$ and original $y$ labels across classes. The median consistently outperforms the others as it poses a better trade-off than the *Min* (*Max*), which restricts (extends) the samples to select in classes with many (few) agreements. Here the unbalanced criterion considers as clean all samples that satisfy $\hat{y} = y$.

### 4.5. Joint training ablation study

Tab. 5 illustrates the effect of removing key components of our method on classification accuracy. Removing semi-supervised learning (SSL) involves training the classifier using *mixup*, which results in substantial degradation due to label noise memorization. Removing the memory (M) decreases performance due to the limited batch size used (128), which provides few positives/negatives for supervised contrastive learning with 100 classes. Not balancing (B) the clean set $\mathcal{D}$ to perform SSL also decreases performance. The criterion used to select clean samples without balancing was to select every sample $x$ satisfying the agreement $\hat{y} = y$ as studied in Sec. 4.4. Regarding the classifier refinement done by MOIT+, re-training the classifier (r-t C) and avoiding the use of strong data augmentation impact performance. The former might prevent some slight memorization behavior in the classifier occurring during MOIT, while the latter avoids the strong data augmentation that harms classification accuracy but is required for successful contrastive learning.

### 4.6. Synthetic label noise evaluation

Tables 6 and 7 evaluate the performance of MOIT and MOIT+ in, respectively, CIFAR-10 and CIFAR-100 for different levels of symmetric and asymmetric noise and report average accuracy for each dataset to ease comparison. We compare against some relevant and recent methods from the literature [53, 1, 50, 49, 28, 30] and demonstrate that MOIT and MOIT+ achieve state-of-the-art results. We achieve especially robust results for asymmetric noise, which is more realistic than symmetric as label flips are done considering semantic similarities between classes. We run DMix (evaluation done without ensembling both networks) and ELR, while using the remaining results from [34], which used the same network architecture and label noise criterion. DMix [28] and, especially, ELR outperform our method for some noise levels, but experience important drops at high noise levels, which penalize the average performance. We stress that our label noise criterion (also adopted in [23, 45, 34]) considers 40% noise as 0.4 probability of flipping the label to an *incorrect* class, and not to any class as reported in the DMix and ELR papers [28, 30], which results in 40% being more challenging in our setup.

Table 6. Performance in CIFAR-10 with symmetric and asymmetric noise. (*) Denotes that we have run the algorithm.

| | Symmetric | | | | Asymmetric | | | Avg. |
|---|---|---|---|---|---|---|---|---|
| | 0% | 20% | 40% | 80% | 10% | 30% | 40% | |
| CE | 93.85 | 78.93 | 55.06 | 33.09 | 88.81 | 81.69 | 76.04 | 72.50 |
| Mix [53] | **95.96** | 84.76 | 66.07 | 20.38 | 93.30 | 83.26 | 77.74 | 74.50 |
| DB [1] | 79.18 | 93.82 | 92.26 | 15.53 | 89.58 | 92.20 | 91.20 | 79.11 |
| DMI [49] | 93.88 | 88.33 | 83.24 | 43.67 | 91.11 | 91.16 | 83.99 | 82.20 |
| PCIL [50] | 93.89 | 92.72 | 91.32 | 55.99 | 93.14 | 92.85 | 91.57 | 87.35 |
| DRPL [34] | 94.08 | 94.00 | 92.27 | 61.07 | **95.50** | 92.98 | 92.84 | 88.96 |
| DMix* [28] | 94.27 | **95.12** | **94.11** | 35.36 | 93.77 | 92.47 | 90.04 | 85.02 |
| ELR* [30] | 95.49 | 94.49 | 92.56 | 38.23 | 95.25 | **94.66** | 92.88 | 86.22 |
| MOIT | 95.17 | 92.88 | 90.55 | 70.53 | 93.50 | 93.19 | 92.27 | 89.73 |
| MOIT+ | 95.65 | 94.08 | 91.95 | **75.83** | 94.23 | 94.31 | **93.27** | **91.33** |

Table 7. Performance in CIFAR-100 with symmetric and asymmetric noise. (*) Denotes that we have run the algorithm.

| | Symmetric | | | | Asymmetric | | | Avg. |
|---|---|---|---|---|---|---|---|---|
| | 0% | 20% | 40% | 80% | 10% | 30% | 40% | |
| CE | 74.34 | 58.75 | 42.92 | 8.29 | 68.10 | 53.28 | 44.46 | 50.02 |
| Mix [53] | 77.90 | 66.40 | 52.20 | 13.21 | 72.40 | 57.63 | 48.07 | 55.40 |
| DB [1] | 64.79 | 69.11 | 62.78 | 45.67 | 67.09 | 58.59 | 47.44 | 59.35 |
| DMI [49] | 74.44 | 58.82 | 53.22 | 20.30 | 68.15 | 54.15 | 46.20 | 53.61 |
| PCIL [50] | 77.75 | 74.93 | 68.49 | 25.41 | 76.05 | 59.29 | 48.26 | 61.45 |
| DRPL [34] | 71.84 | 71.16 | 72.37 | **52.95** | 72.03 | 69.30 | 65.69 | 67.91 |
| DMix* [28] | 67.41 | 71.39 | 70.83 | 49.52 | 69.53 | 68.28 | 50.99 | 63.99 |
| ELR* [30] | **78.01** | **75.90** | **72.89** | 36.83 | 77.08 | 74.61 | 71.25 | 69.51 |
| MOIT | 75.83 | 72.78 | 67.36 | 45.63 | 75.49 | 73.34 | 71.55 | 68.85 |
| MOIT+ | 77.07 | 75.89 | 70.88 | 51.36 | **77.43** | **75.13** | **74.04** | **71.69** |

Table 8. Performance evaluation on controlled web noise in mini-ImageNet. We run all methods.

| | | 0% | 20% | 40% | 80% |
|---|---|---|---|---|---|
| Mix [53] | Best | 61.18 | 57.76 | 52.88 | 38.32 |
| | Last | 58.96 | 54.60 | 50.40 | 37.32 |
| DMix [28] | Best | 57.80 | 55.86 | 55.44 | 41.12 |
| | Last | 55.84 | 50.30 | 50.94 | 35.42 |
| ELR [30] | Best | 63.12 | 61.48 | 57.32 | 41.68 |
| | Last | 57.38 | 58.10 | 50.62 | 41.68 |
| MOIT | Best | 67.18 | 64.82 | 61.76 | 46.40 |
| | Last | 64.72 | 63.14 | 60.78 | 45.88 |
| MOIT+ | Best | **68.28** | **64.98** | **62.36** | **47.80** |
| | Last | 67.82 | 63.10 | 61.16 | 46.78 |

### 4.7. Web label noise evaluation

Table 9. Performance evaluation in mini-WebVision. We run all methods.

| | Mix [53] | DMix [28] | ELR [30] | MOIT | MOIT+ |
|---|---|---|---|---|---|
| Best | 74.96 | 76.08 | 73.00 | 78.36 | **78.76** |
| Last | 73.76 | 74.64 | 71.88 | 77.76 | **78.72** |

Tables 8 and 9 illustrate the superior performance of MOIT/MOIT+ when training in the presence of web label noise in mini-ImageNet [22] and mini-WebVision [28]. The results demonstrate that MOIT/MOIT+ are robust to web noise and that they do not need careful re-parametrization depending on the noise level or distribution to achieve state-of-the-art performance. The results in Tab. 8 further confirm that the improvements are consistent across noise levels. It is interesting to observe that, although MOIT+ consistently outperforms MOIT, the improvements compared to CIFAR experiments tend to be smaller. We think that a plausible explanation is the dominance of out-of-distribution samples in web-noise, which makes label correction via semi-supervised learning less beneficial. Note that we run M, DMix, EReg, and MOIT for the same number of epochs (130) in both mini-ImageNet and mini-WebVision.

## 5. Conclusion

This paper proposes Multi-Objective Interpolation Training (MOIT), an approach for image classification with deep neural networks that robustly learns in the presence of both synthetic and web label noise. The key idea of MOIT is to combine supervised contrastive learning and classification in such a way that they are both robust to label noise. Interpolated Contrastive Learning regularization enables learning label noise robust representations that are used to estimate a soft-label distribution whose agreement with the original label allows identification of correctly labeled samples. MOIT then treats the remaining samples as unlabeled and trains a label noise robust image classifier in a semi-supervised manner. We further propose MOIT+, a refinement of our model by fine-tuning the model while re-training the image classifier. We conduct experiments in CIFAR-10/100 with synthetic label noise and in mini-ImageNet and mini-WebVision with web noise to demonstrate that MOIT and MOIT+ achieve state-of-the-art results when training deep neural networks with different noise distributions and levels. Future work will explore instance-dependent label noise as well as how to simplify the contrastive learning framework by using class prototypes.

### Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant number SFI/15/SIRG/3283 and SFI/12/RC/2289 P2.

## References

[1] E. Arazo, D. Ortego, P. Albert, N. O'Connor, and K. McGuinness. Unsupervised Label Noise Modeling and Loss Cor-

rection. In *International Conference on Machine Learning (ICML)*, 2019.

[2] E. Arazo, D. Ortego, P. Albert, N.E. O'Connor, and K. McGuinness. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. In *International Joint Conference on Neural Networks (IJCNN)*, 2020.

[3] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M.S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien. A Closer Look at Memorization in Deep Networks. In *International Conference on Machine Learning (ICML)*, 2017.

[4] D. Bahri, H. Jiang, and M. Gupta. Deep k-NN for Noisy Labels. In *International Conference on Machine Learning (ICML)*, 2020.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning (ICML)*, 2020.

[6] X. Chen, H. Fan, R. Girshick, and K. He. Improved Baselines with Momentum Contrastive Learning. *arXiv:2003.04297*, 2020.

[7] S. Chopra and R. and Y. LeCun Hadsell. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[8] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[9] Y. Ding, L. Wang, D. Fan, and B. Gong. A Semi-Supervised Two-Stage Approach to Learning from Noisy Labels. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[10] S. Guo, W. Huang, H. Zhang, C. Zhuang, D. Dong, M.R. Scott, and D. Huang. CurriculumNet: Weakly Supervised Learning from Large-Scale Web Images. In *European Conference on Computer Vision (ECCV)*, 2018.

[11] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[12] J. Han, P. Luo, and X. Wang. Deep Self-Learning From Noisy Labels. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[13] H. Harutyunyan, K. Reing, G. V. Steeg, and A. Galstyan. Improving Generalization by Controlling Label-Noise Information in Neural Network Weights. In *International Conference on Machine Learning (ICML)*, 2020.

[14] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Identity Mappings in Deep Residual Networks. In *European Conference on Computer Vision (ECCV)*, 2016.

[17] D. Hendrycks, K. Lee, and M. Mazeika. Using Pre-Training Can Improve Model Robustness and Uncertainty. In *International Conference on Machine Learning (ICML)*, 2019.

[18] D. Hendrycks, M. Mazeika, D. Wilson, and K. Gimpel. Using Trusted Data to Train Deep Networks on Labels Corrupted by Severe Noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[19] E. Hoffer and N. Ailon. Deep metric learning using Triplet network. *arXiv:1412.6622*, 2018.

[20] Y. Hu, V. Gripon, and S. Pateux. Leveraging the Feature Distribution in Transfer-based Few-Shot Learning. *arXiv:2006.03806*, 2020.

[21] J. Huang, Q. Dong, S. Gong, X. Zhu. Unsupervised Deep Learning by Neighbourhood Discovery. In *Int. Conf. on Mach. Learn. (ICML)*, 2019.

[22] L. Jiang, D. Huang, M. Liu, and W. Yang. Beyond Synthetic Noise: Deep Learning on Controlled Noisy Labels. In *International Conference on Machine Learning (ICML)*, 2020.

[23] L. Jiang, Z. Zhou, T. Leung, L.J. Li, and L. Fei-Fei. MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels. In *International Conference on Machine Learning (ICML)*, 2018.

[24] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised Contrastive Learning. *arXiv:2004.11362*, 2020.

[25] Y. Kim, J. Yim, J. Yun, and J. Kim. NLNL: Negative Learning for Noisy Labels. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[26] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[27] P. H. Le-Khac, G. Healy, and A. F. Smeaton. Contrastive Representation Learning: A Framework and Review. *IEEE Access*, 8:193907–193934, 2020.

[28] J. Li, R. Socher, and S.C.H. Hoi. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[29] W. Li, L. Wang, W. Li, E. Agustsson, and L. Van Gool. WebVision Database: Visual Learning and Understanding from Web Data. *arXiv: 1708.02862*, 2017.

[30] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda. Early-Learning Regularization Prevents Memorization of Noisy Labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[31] X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, and J. Bailey. Normalized Loss Functions for Deep Learning with Noisy Labels. In *International Conference on Machine Learning (ICML)*, 2020.

[32] K. Musgrave, S. Belongie, and S.-N. Lim. A Metric Learning Reality Check. In *European Conference on Computer Vision (ECCV)*, 2020.

[33] D. T. Nguyen, C. K. Mummadi, T. P. N. Ngo, T. H. P. Nguyen, L. Beggel, and T. Brox. SELF: Learning to Filter Noisy Labels with Self-Ensembling. In *International Conference on Learning Representations (ICLR)*, 2020.

[34] D. Ortego, E. Arazo, P. Albert, N. O'Connor, and K. McGuinness. Towards Robust Learning with Different Label Noise

Distributions. In *International Conference on Pattern Recognition (ICPR)*, 2020.

[35] D. Picard A. Histace E. Klein P. Jacob. Metric Learning With HORDE: High-Order Regularizer for Deep Embeddings. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[36] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[37] Q. Qian, L. Shang, B. Sun, J. Hu, H. Li, and R. Jin. SoftTriple Loss: Deep Metric Learning Without Triplet Sampling. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[38] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *International Conference on Learning Representations (ICLR)*, 2015.

[39] D. Han S. J. Oh S. Chun J. Choe Y. Yoo S. Yun. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[40] R. Takahashi, T. Matsubara, and K. Uehara. RICAP: Random Image Cropping and Patching Data Augmentation for Deep CNNs. In *Asian Conference on Machine Learning (ACML)*, 2018.

[41] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa. Joint Optimization Framework for Learning with Noisy Labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[42] S. Thulasidasan, T. Bhattacharya, J. Bilmes, G. Chennupati, and J. Mohd-Yusof. Combating Label Noise in Deep Learning Using Abstention. In *International Conference on Machine Learning (ICML)*, 2019.

[43] Y. Tian, C. Sun, B. Poole, P. Krishnan, C. Schmid, and P. Isola. What Makes for Good Views for Contrastive Learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[44] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[45] Y. Wang, W. Liu, X. Ma, J. Bailey, H. Zha, L. Song, and S.-T. Xia. Iterative Learning With Open-Set Noisy Labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[46] H. Wei, L. Feng, X. Chen, and B. An. Combating noisy labels by agreement: A joint training method with co-regularization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[47] H. Zhang W. Huang M. R. Scott X. Wang. Cross-Batch Memory for Embedding Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[48] 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 (NeurIPS)*, 2019.

[49] Y. Xu, P. Cao, Y. Kong, and Y. Wang. L_DMI: An Information-theoretic Noise-robust Loss Function. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[50] K. Yi and J. Wu. Probabilistic End-To-End Noise Correction for Learning With Noisy Labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[51] X. Yu, B. Han, J. Yao, G. Niu, I. W. Tsang, and M. Sugiyama. How does Disagreement Help Generalization against Label Corruption? In *International Conference on Machine Learning (ICML)*, 2019.

[52] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires re-thinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

[53] H. Zhang, M. Cisse, Y.N. Dauphin, and D. Lopez-Paz. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

[54] Z. Zhang and M. Sabuncu. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[55] Z. Zhang, H. Zhang, S. O. Arik, H. Lee, and T. Pfister. Distilling Effective Supervision From Severe Label Noise. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.