UNICON: Combating Label Noise Through Uniform Selection and Contrastive Learning

Nazmul Karim[†] Mamshad Nayeem Rizve[‡] Nazanin Rahnavard[†] Ajmal Mian[§] Mubarak Shah[‡]
[†]Department of Electrical and Computer Engineering, UCF, USA
[‡]Center for Research in Computer Vision, UCF, USA
[§]Department of Computer Science and Software Engineering, UWA, Australia
{nazmul.karim18, nayeemrizve}@knights.ucf.edu, nazanin.rahnavard@ucf.edu
ajmal.mian@uwa.edu.au, shah@crcv.ucf.edu

Abstract

Supervised deep learning methods require a large repository of annotated data; hence, label noise is inevitable. Training with such noisy data negatively impacts the generalization performance of deep neural networks. To combat label noise, recent state-of-the-art methods employ some sort of sample selection mechanism to select a possibly clean subset of data. Next, an off-the-shelf semi-supervised learning method is used for training where rejected samples are treated as unlabeled data. Our comprehensive analysis shows that current selection methods disproportionately select samples from easy (fast learnable) classes while rejecting those from relatively harder ones. This creates class imbalance in the selected clean set and in turn, deteriorates performance under high label noise. In this work, we propose UNICON, a simple yet effective sample selection method which is robust to high label noise. To address the disproportionate selection of easy and hard samples, we introduce a Jensen-Shannon divergence based uniform selection mechanism which does not require any probabilistic modeling and hyperparameter tuning. We complement our selection method with contrastive learning to further combat the memorization of noisy labels. Extensive experimentation on multiple benchmark datasets demonstrates the effectiveness of UNICON; we obtain an 11.4% improvement over the current state-of-the-art on CIFAR100 dataset with a 90% noise rate. Our code is publicly available.[1]

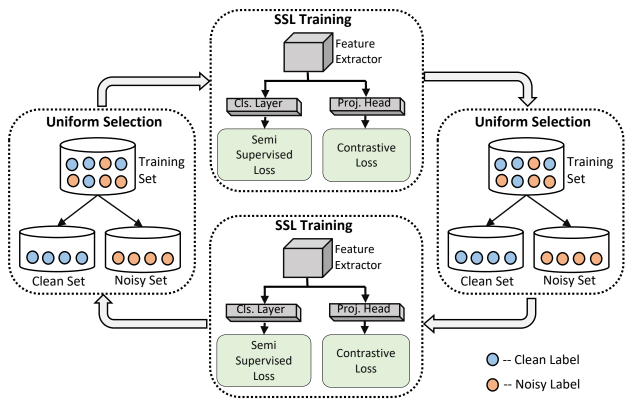


Figure 1. UNICON *training overview:* At each iteration, we employ a uniform selection technique to partition the training set into clean and noisy sets. Upon separation, we perform SSL-training with an additional contrastive loss function. The uniform selection and subsequent SSL-training is repeated until convergence.

Noise Rate (%)	90%	92%	95%	98%
DMix [25]	76.08	57.62	51.28	17.18
UNICON (Ours)	**90.81**	**87.61**	**80.82**	**50.63**

Table 1. Classification performance (%) of the proposed method on CIFAR10 under severe label noise.

1. Introduction

Deep neural networks (DNNs) have proven to be highly effective in solving various computer vision tasks [9, 18, 22, 36, 43, 48, 49, 53, 62]. Most state-of-the-art (SOTA) methods require supervised training with a large pool of annotated data [4, 8, 27, 28, 57]. Collecting and manually annotating such data is challenging and oftentimes very expensive. Most large-scale data collection techniques rely on open-source web data that can be automatically annotated using search engine queries and user tags [33, 52]. This annotation scheme inevitably introduces label noise [27, 57]. Training with such noisy labels is challenging since DNNs can effectively memorize arbitrary (noisy) labels over the course of training [2]. Combating label noise is one of the fundamental problems in deep learning [15, 24, 38, 46, 54, 54, 58, 60, 61, 65], and is the focus of this study.

Training with noisy label data has been the subject of

[1]https://github.com/nazmul-karim170/UNICON-Noisy-Label

many recent studies [12, 16, 31, 42, 45, 70]. Existing techniques can be categorized into two dominant groups: i) label correction, [11,40] and ii) sample separation [12,25,66]. The former approach requires the estimation of noise transition matrix, which is hard to estimate for high number of classes and in high noise scenarios. The latter approach tries to filter out the noisy samples from the clean ones based on the small-loss criterion [25], where the low-loss samples are assumed to have clean labels. Next, an off-the-shelf semi-supervised learning (SSL) technique [3,47] is used for training where the selected noisy samples are treated as unlabeled data. However, the selection process is usually biased towards easy classes as clean samples from the hard classes (e.g. cats and dogs can be considered as hard classes in CIFAR10 [21]) may produce high-loss values. This is more prominent at the early stage of training and can introduce class-disparity among the selected clean samples. Severe class-imbalance may lead to poor precision of sample selection, hence, sub-par classification performance.

In this work, we revamp the selection process from a more fundamental perspective. Our goal is to simplify the selection process by introducing an effective and scalable Jensen-Shannon divergence based sample separation mechanism. To address the disproportionate selection of easy and hard samples, we enforce a class-balance prior by selecting an equal number of clean samples from each class. Such a prior improves the overall quality of pseudo-labels, and hence, significantly boosts the performance of subsequent semi supervised learning-based training. In addition, we opt to employ unsupervised contrastive learning (CL) because of its inherent resistance (as labels are not required for training) to label noise memorization. We empirically show that unsupervised feature learning lowers memorization risk and improves the sample separation performance; especially under severe noise levels. We call this combined technique of UNIform selection and CONtrastive learning UNICON (shown in Fig. 1), which is found to be effective even in the presence of very high label noise (see Table 1). Our contributions are summarized as follows:

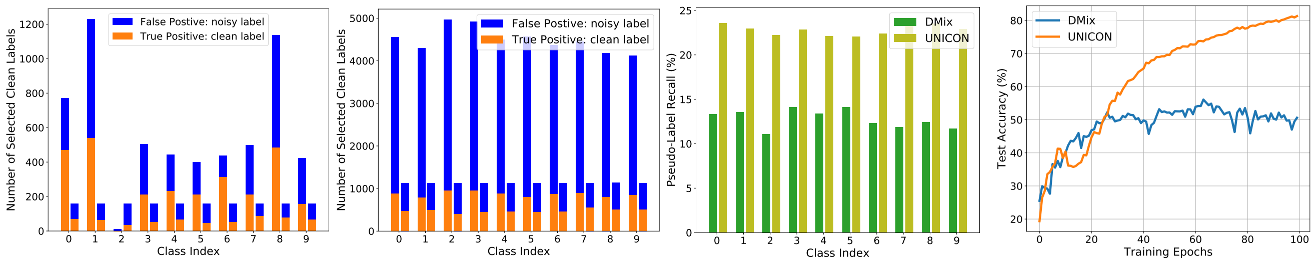
- We propose a simple yet effective uniform selection mechanism that ensures class-balancing among the selected clean samples. Through empirical analysis, we observe that class-uniformity helps in generating higher quality pseudo-labels for samples from all classes irrespective of their difficulty level.
- We further minimize the risk of label noise memorization by performing unsupervised feature learning using contrastive loss. This in turn boosts the sample separation performance.
- Our extensive experimentation demonstrates that UNICON achieves significant performance improvement over state-of-the-art methods, especially on datasets with severe label noise.

2. Related Work

Noisy label training has been studied extensively in recent works [26, 29, 35, 55, 72]. Wei et al. [56] proposed a regularization technique to learn from noisy labels. Another method called MentorNet [17] trains a student network by generating pseudo-labels using a pre-trained/mentor network. Based on their relationship in the feature space, Meta-cleaner [69] learns the confidence scores of noisy samples which are then used for obtaining cleaner representations. To deal with noisy labels, [32, 51, 64] gradually adjust the data labels based on the predicted labels given by the network. Some noisy label methods are based on loss correction [11, 14, 40] and noise-tolerant loss functions [5,71]. In [14], a noise transition matrix was estimated by correcting the loss obtained by a DNN trained on a noisy dataset. However, the performance of these methods deteriorates under high noise rates and large number of classes. Other approaches rely on the separation of clean samples from the noisy samples [10, 12, 25, 37, 51, 66]. A notable difference between these methods is the selection criteria of clean samples. A selection technique was proposed in [10] that utilizes prediction likelihoods to obtain separation.

Co-teaching [12] opts to train two networks simultaneously such that one network separates clean samples for the other network based on the small-loss criterion. The small-loss criterion suggests that samples with smaller loss tend to have clean labels. Therefore, one could separate samples on the training set based on their loss-values. DMix [25] proposed a hybrid framework to separate samples and uses a SSL technique [68] to concurrently train two networks. A modified training scheme for [25] was proposed in [35]. However, even for the same dataset, these methods employ different training settings and constraints under different noise rates and types. This limits their practical applications as prior knowledge of noise rate may not be available. Recently, a joint semi-supervised and contrastive learning-based technique was proposed in MOIT [39]. Jo-SRC [63] initially partitions the samples into clean and noisy sets before detecting in-distribution (ID) and OOD samples in the noisy set. However, it requires manual threshold adjustment for the separation during different epochs of the training. Furthermore, both [63] and [39] struggle to achieve good performance under high noise rates.

In contrast, our proposed method can handle severe label noise and requires minimal to no change in the hyperparameter settings under different label-noise scenarios (e.g. different noise rates, noise types etc.). We show how a minimalistic approach to the selection process can boost the classification performance significantly beating the sate-of-the-art methods in most cases. Furthermore, we achieve comparable performance to SOTA across different datasets which hints at the generalizability of our method.



(a) Selected clean set after warmup. (b) Selected clean set after 100 epochs. (c) Pseudo-label recall after 100 epochs. (d) Test accuracy (%).

Figure 2. A case of uniform and non-uniform selection for CIFAR10 under 90% noise rate. (a) Class distribution in $\mathbb{D}_{clean}$ after warmup (10 epochs of training). For each class index, the left and right bars indicate non-uniform (DMix [25]) and uniform selection (UNICON), respectively. (b) Class distribution after 100 epochs. UNICON selects clean samples with higher precision. (c) pseudo-label recall (%) after 100 epochs of training. Uniform selection criteria along with contrastive feature learning helps generating higher quality pseudo-labels with better recall. (d) This in turn boosts the test accuracy significantly.

3. Background

Let $\mathbb{D} = \{\mathcal{X}, \mathcal{Y}\} = \{(\mathbf{x}_0, \mathbf{y}_0), (\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$ denote the training set, where $\mathbf{x}_i$ is an image and $\mathbf{y}_i$ is the corresponding ground-truth label, and N is the total number of training samples. We instantiate the DNN model with a feature extractor (CNN backbone), $\mathbf{f}(.; \theta)$, with parameters θ; a classification layer, $\mathbf{h}(.; \phi)$, with parameters ϕ, and a projection head, $\mathbf{g}(., \psi)$, with parameters ψ for incorporating contrastive learning. For supervised training with ground-truth labels, we minimize cross-entropy (CE) loss, $\mathcal{L}_{CE}$, over the entire training set $\mathbb{D}$,

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i^T \log \hat{\mathbf{y}}_i, \tag{1}$$

where $\hat{\mathbf{y}}_i = \mathrm{softmax}(\mathbf{h}(\mathbf{f}(\mathbf{x}_i; \theta); \phi))$ is the softmax probability score of the network prediction corresponding to $\mathbf{x}_i$.

In this work, we consider the training set to be noisy i.e. some images are incorrectly labeled. It has been demonstrated that DNNs learn simpler patterns before memorizing the noisy labels [2]. Several studies [12, 25] utilize this observation and try to separate the clean samples from the noisy ones at the early stage of training. Such a separation scheme partitions the dataset into a clean subset, $\mathbb{D}_{clean}$, and a noisy subset, $\mathbb{D}_{noisy} = \mathbb{D} \setminus \mathbb{D}_{clean}$. After that, $\mathbb{D}_{clean}$ can be used for standard supervised training. To mitigate the impact of label noise, samples from $\mathbb{D}_{noisy}$ can be used for training without the corresponding noisy ground-truth labels. This training is generally performed in a semi-supervised manner where pseudo-labels are generated for the samples in $\mathbb{D}_{noisy}$.

We conduct extensive empirical analysis to investigate the effectiveness of partitioning the dataset into $\mathbb{D}_{clean}$, and $\mathbb{D}_{noisy}$ subsets. We find that the typical construction of $\mathbb{D}_{clean}$ creates disparity or *imbalance* among classes [12, 25, 63]. Fig. 2a (left bars) depicts such a case where the $\mathbb{D}_{clean}$ for noisy CIFAR10 (90% noise rate) contain class imbalance when we employ a recently proposed method, DMix [25]. To be specific, we observe that 1228 samples are selected from class-1, whereas only 10 samples from class-2 are selected. However, the imbalance among true positives (TPs) are of particular importance as the quality of pseudo-labels for $\mathbb{D}_{noisy}$ relies heavily on them. Methods such as [25] attempt to address this issue by selecting more clean samples which in turn increases the false positive or noisy labels count (Fig. 2b (left bars)) while drastically decreasing the precision. As the selected clean set $\mathbb{D}_{clean}$ contains many false positives, supervised training on such a set leads to memorization. Consequently, the recall of the subsequent pseudo-labels drops drastically; as shown by the pseudo-label recall in Fig. 2c (left bars). In this way, the selection mechanism negatively impacts the SSL-Training and reduces the average classification accuracy (Fig. 2d).

We propose to address these problems by a simple and effective technique of uniform selection (Fig. 2a (right bars)). Furthermore, we employ contrastive feature learning to learn better unsupervised features irrespective of the quality of ground-truth or pseudo-labels. Details of our proposed method are presented in the following section.

4. Proposed Method

We propose UNICON with a unique sample-selection approach as well as simple but effective modification to the SSL-Training. UNICON improves precision (Fig. 2b (right bars)) as well as pseudo-label recall (Fig. 2c (right bars)) over training. Fig. 2d shows that our hybrid framework of uniform selection and SSL training improves the classification performance significantly. Next, we present our uniform sample selection strategy in Sec. 4.1, and our proposed SSL training method with contrastive learning in Sec 4.2.

4.1. Uniform Sample Selection

During the partitioning of $\mathbb{D}$, we opt to enforce class-balancing in $\mathbb{D}_{clean}$ by selecting/filtering R portion of samples from *each class*, where we define R as the filter rate. Fig. 3 shows our proposed selection mecha-

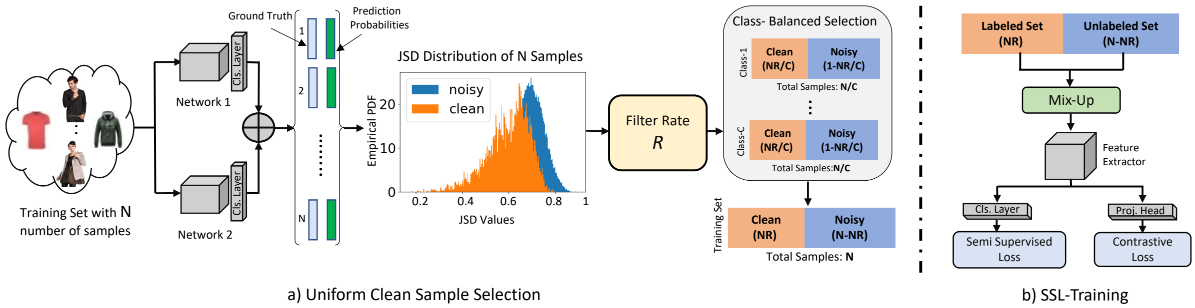


Figure 3. *Proposed Selection Mechanism and SSL-Training:* a) For selection, we ensemble the predictions of both networks to calculate JSD. After estimating the filter rate R from JSD distribution, we take equal number (NR/C) of samples from each class. b) We consider separated clean and noisy sets as labeled and unlabeled data only to employ Mix-up [67] based SSL-training with contrastive loss. On top of classification (Cls.) layer, we add a projection (Proj.) head to facilitate contrastive learning. We train both networks sequentially.

nism in which we feed $\mathbb{D}$ to two networks with parameters $(\theta^{(1)}, \phi^{(1)}, \psi^{(1)})$ and $(\theta^{(2)}, \phi^{(2)}, \psi^{(2)})$. For $\mathbf{x}_i$, the average prediction probabilities from both networks can be denoted as $\mathbf{p}_i = [\mathbf{p}_i^1, \mathbf{p}_i^2, \ldots, \mathbf{p}_i^C]$, and the corresponding ground-truth label as $\mathbf{y}_i = [\mathbf{y}_i^1, \mathbf{y}_i^2, \ldots, \mathbf{y}_i^C]$; here, C is the total number of classes. To construct the clean, $\mathbb{D}_{clean}$, and noisy, $\mathbb{D}_{noisy}$, subsets, we compute the disagreement/divergence between the ground-truth labels, $\mathbf{y}_i$, and the predicted probabilities, $\mathbf{p}_i$. To this end, we use Jensen-Shannon divergence (JSD), d_i, as a measure of disagreement. The JSD is defined as,

$$\begin{aligned} d_i &= \mathrm{JSD}(\mathbf{y}_i, \mathbf{p}_i) \\ &= \frac{1}{2}\mathrm{KLD}(\mathbf{y}_i || \frac{\mathbf{y}_i + \mathbf{p}_i}{2}) + \frac{1}{2}\mathrm{KLD}(\mathbf{p}_i || \frac{\mathbf{y}_i + \mathbf{p}_i}{2}), \end{aligned} \tag{2}$$

where $\mathrm{KLD}(.)$ is the Kullback-Leibler divergence function.

Previous works use different divergence measures to construct the clean and noisy subsets. Authors in [12, 25] apply CE loss-based divergence measure for selection. [25] uses a similar divergence measure and fits a Gaussian mixture model (GMM) on the normalized CE values for partitioning. In contrast, we opt to employ JSD-based selection since it does not require normalization and probabilistic modelling. Besides, unlike CE loss, JSD is symmetric by design and the value ranges from 0 to 1.

After measuring the divergence, $\mathbf{d} = \{d_i : i \in (1, \ldots, N)\}$, for all the samples, we compute a cutoff divergence value, d_{cutoff}, which can be expressed as,

$$d_{cutoff} = \begin{cases} d_{avg} - (d_{avg} - d_{min})/\tau, & \text{if } d_{avg} \geq d_{\mu} \\ d_{avg}, & \text{otherwise} \end{cases} \tag{3}$$

where d_{avg} is the average over all values in $\mathbf{d}$, d_{min} is the lowest divergence score, τ is the filter coefficient, and d_{μ} is an adjustment threshold. Finally, we determine R as the percentage of samples that have JSDs lower than d_{cutoff}.

There are two major benefits of this particular design of d_{cutoff}. First, we determine the value of d_{cutoff} based on the network prediction scores (as JSD depends on prediction probabilities) which eliminates the requirement of manual per-dataset tuning. The second benefit stems from the same source, i.e., d_{cutoff} is determined from prediction scores. This ensures that if the network prediction scores are consistently low (high d_{avg}), d_{cutoff} will encourage a conservative selection of $\mathbb{D}_{clean}$; which helps in avoiding noisy sample selection at the early stage of training.

Algorithm 1: Uniform Clean Sample Selection

Input: training set $\mathbb{D} = (\mathcal{X}, \mathcal{Y})$, number of samples N, number of classes C
for $i = 1$ **to** N **do**
  $\mathbf{p}_i = (\hat{\mathbf{y}}_i^{(1)} + \hat{\mathbf{y}}_i^{(2)})/2$
  $d_i = JSD(\mathbf{p}_i, \mathbf{y}_i)$ (see Eq. (2))
Determine the cutoff distance, d_{cutoff} using Eq. (3)
$\mathbf{d}_R \leftarrow \{d_i < d_{cutoff} : i \in (1, \ldots, N)\}$
Determine filter rate, $R = |\mathbf{d}_R|/N$
$\mathbb{D}_{clean} = \{\}$
`// Uniform Selection`
for $j = 1$ **to** C **do**
  $\mathbf{d}^{(j)}_{filtered} \leftarrow$ Lowest R portion of $\mathbf{d}^{(j)}$
  $\mathbb{D}^{(j)}_{clean} \leftarrow \{(\mathbf{x}_t^{(j)}, \mathbf{y}_t^{(j)}) : \forall \ d_t^{(j)} \in \mathbf{d}^{(j)}_{filtered}\}$
  $\mathbb{D}_{clean} \leftarrow \mathbb{D}_{clean} \cup \mathbb{D}^{(j)}_{clean}$
$\mathbb{D}_{noisy} \leftarrow \mathbb{D} \setminus \mathbb{D}_{clean}$
Output: $\mathbb{D}_{noisy}, \mathbb{D}_{clean}$

In the next step, we create class-specific partitions, $\{\mathbf{d}^{(1)}, \mathbf{d}^{(2)}, \ldots, \mathbf{d}^{(C)}\}$, where $\mathbf{d}^{(j)}$ indicates the JSDs for class j. Motivated by the small-loss criterion [25], we define UNICON selection criterion as follows:

UNICON Selection Criterion: *For each class j, if the difference $\mathbf{d}_i^{(j)}$ falls within the lowest R portion of all values in $\mathbf{d}^{(j)}$, we consider $\mathbf{x}_i^{(j)}$ to have a clean label.*
Here, $i \in \{1, 2, \ldots, N_j\}$, N_j is the total number of samples in class j, and $\mathbf{x}_i^{(j)}$ is the i-th image belonging to the j-th class with JSD of $\mathbf{d}_i^{(j)}$.

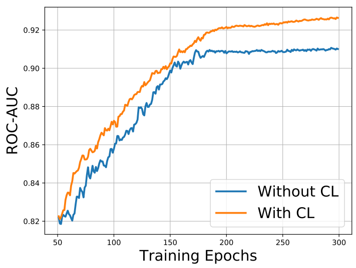


Figure 4. ROC-AUC score of clean sample selection with and without contrastive learning. As CL helps preventing the memorization, the clean samples are being detected with better precision. Here we considered CIFAR10 with 90% noise rate.

Finally, following the UNICON selection criterion, we aggregate all the selected clean and noisy samples from each class to form $\mathbb{D}_{clean}$ and $\mathbb{D}_{noisy}$ with cardinalities of NR and $N(1-R)$, respectively. In cases where the total number of available samples (both clean and noisy) for any class falls below NR/C, we take all the available samples in that class for $\mathbb{D}_{clean}$. Algorithm 1 summarizes our selection method. Note that a previous technique named Jo-SRC [63] has employed JSD for clean sample detection. However, our sample selection process differs significantly from Jo-SRC. For instance, the selection threshold in [63] needs to be manually fine-tuned during different epochs of the training while UNICON automatically adjusts the filter rate, R, based on the network prediction scores; making our proposed selection method hyperparameter independent.

4.2. SSL-Training

Fig. 3 shows the details of our SSL-Training with semi-supervised and contrastive loss. Following FixMatch [47], we perform semi-supervised learning with the samples from $\mathbb{D}_{noisy}$. To this end, we generate two copies of each sample with a weak and a strong augmentation. Pseudo-labels are generated from the weakly augmented copy for computing a semi-supervised loss, $\mathcal{L}_{semi}$, on the strongly augmented copy. We also apply MixUp [67] augmentation between the samples from $\mathbb{D}_{clean}$ and $\mathbb{D}_{noisy}$; for the $\mathbb{D}_{noisy}$ samples, we use the pseudo-labels obtained from weakly augmented copy. However, feature or representation learning in such a SSL manner still bears the risk of noise memorization. During training, DNNs memorize certain portion of noisy samples irrespective of the sample selection technique. The presence of such noisy samples in the clean subset, will lead to noisy SSL training. To address this issue, we incorporate *contrastive learning* (CL) [6, 19] into our SSL training pipeline to facilitate feature learning without relying on labels/pseudo-labels. Such an unsupervised feature learning scheme further mitigates the risk of noisy label memorization since it does not rely on imperfect separation of clean and noisy samples as well as incorrect pseudo-labels generated during SSL training. Thus, incorporation of CL improves the performance of our proposed selection technique, as shown by the area under the curve (AUC) of Receiver Operating Characteristics (ROC) in Fig. 4. In our work, we employ contrastive loss only for the samples in the unlabeled set, $\mathbb{D}_{noisy}$. To this end, we employ the projection head $\mathbf{g}(.;\psi)$ to obtain feature projections $\mathbf{z}_i = \mathbf{g}(\mathbf{f}(\mathbf{x}_{i,1};\theta);\psi)$, and $\mathbf{z}_j = \mathbf{g}(\mathbf{f}(\mathbf{x}_{i,2};\theta);\psi)$ of the differently augmented copies $(\mathbf{x}_{i,1}, \mathbf{x}_{i,2})$ of input $\mathbf{x}_i$. The contrastive loss function [6, 19] can be expressed as

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\kappa)}{\sum_{b=1}^{2B} \mathbb{1}_{b\neq i} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_b)/\kappa)}, \tag{4}$$

$$\mathcal{L}_{\mathcal{C}} = \frac{1}{2B} \sum_{b=1}^{2B} [\ell_{2b-1,2b} + \ell_{2b,2b-1}], \tag{5}$$

where $\mathbb{1}_{b\neq i}$ is an indicator function that gives a 1 iff $b \neq i$, κ is a temperature constant, B is the number of samples in mini-batch, and $\text{sim}(\mathbf{z}_i, \mathbf{z}_j)$ can be expressed as the cosine similarity between $\mathbf{z}_i$ and $\mathbf{z}_j$. The total loss function we minimize is

$$\mathcal{L}_{tot} = \mathcal{L}_{semi} + \lambda_{\mathcal{C}} \mathcal{L}_{\mathcal{C}}, \tag{6}$$

where $\lambda_{\mathcal{C}}$ is contrastive loss coefficient. Additional details of the contrastive learning as well as the rest of our SSL-Training scheme is provided in the *supplementary material*.

5. Experimental Settings

5.1. Datasets

CIFAR10/100: The CIFAR-10/100 datasets [21] contain 50K training and 10K test images. In general, it is difficult to control or determine the noise characteristics; e.g. noise rate, in natural datasets. Therefore, synthetic noise models are commonly used for the evaluation of noise-robust algorithms. In our work, we employ two types of noise models: symmetric and asymmetric. For symmetric noise model, an r portion of samples from one particular class are uniformly distributed to all other classes. On the other hand, the design of asymmetric label noise follows the structure of real mistakes that take place in CIFAR10 [26]: "Truck$\rightarrow$ Automobile, Bird $\rightarrow$ Airplane, Deer $\rightarrow$ Horse, Cat $\rightarrow$ Dog". For CIFAR100, we use label flips for each class to the next one within the super-classes.

Tiny-ImageNet [23]: This dataset is a smaller version of the original ImageNet in terms of the number of classes and the image resolution. There are in total 200 classes containing 500 images per class. The image size is 64×64.

Clothing1M: Clothing1M is a large-scale real-world dataset with noisy labels [57]. It contains 1M images from 14 different cloth-related classes. Since the labels are produced by the seller provided surrounding texts of the images, a large portion of confusing classes (e.g., Knitwear and Sweater) are mislabeled.

Method	CIFAR-10				CIFAR-100			
	20%	50%	80%	90%	20%	50%	80%	90%
CE	86.8	79.4	62.9	42.7	62.0	46.7	19.9	10.1
LDMI [59]	88.3	81.2	43.7	36.9	58.8	51.8	27.9	13.7
M-Up [68]	95.6	87.1	71.6	52.2	67.8	57.3	30.8	14.6
PCIL [64]	92.4	89.1	77.5	58.9	69.4	57.5	31.1	15.3
JPL [20]	93.5	90.2	35.7	23.4	70.9	67.7	17.8	12.8
MOIT [39]	94.1	91.1	75.8	70.1	75.9	70.1	51.4	24.5
DMix [25]	**96.1**	94.6	92.9	76.0	77.3	74.6	60.2	31.5
ELR [30]	95.8	94.8	93.3	78.7	77.6	73.6	60.8	33.4
UNICON	96.0	**95.6**	**93.9**	**90.8**	**78.9**	**77.6**	**63.9**	**44.8**

Table 2. Test accuracies (%) obtained by different techniques under symmetric noise. Our class balance with contrastive loss strategy improves performance at almost every noise level. Results for previous techniques were copied from their respective papers.

Webvision [27]: This dataset contains 2.4 million images (obtained from Flickr and Google) that are categorized into the same 1,000 classes as in the ImageNet ILSVRC12. Following the previous studies [25, 35], we use the first 50 classes of the Google image subset as the training data.

5.2. Training Details

We use the PreAct ResNet18 [13] architecture for CIFAR10, CIFAR100, and Tiny-ImageNet. For Clothing1M and WebVision datasets, we take a ResNet50 network [13] pre-trained on ImageNet and a InceptionResNetV2 network [50] which is trained from scratch. We modify these architectures with a projection head, that produces a embedding vector of size 128, to facilitate contrastive learning.

For CIFAR-10 and CIFAR-100, optimization is performed using stochastic gradient descent (SGD) optimizer with the following settings: an initial learning rate (LR) of 0.02, a weight decay of $5e^{-4}$, a value of 0.9 for the momentum, and a batch size of 64. For CIFAR-10 and CIFAR-100, we train each network for around 300 epochs while linearly decaying the learning rate (lr-decay) by 0.1 per 120 epochs. Following [25], a warmup period of 10 and 30 epochs was employed before starting the selection and SSL-Training. For Tiny-ImageNet, we use an initial LR of 0.01, a weight decay of $1e^{-3}$ with a batch size of 32. We train the network for 350 epochs and the lr-decay rate is 0.1/100 epochs. The warmup period is 15 epochs. For Clothing1M, we choose an initial LR of 0.002 and a weight decay of $1e^{-3}$. We employ the same settings as Tiny-ImageNet for WebVision. The total number of training epochs is 100 and the lr-decay rate is 0.1/40 epochs.

For data augmentations, we follow Auto-augment policy described in [7]. For CIFAR-10 and CIFAR100, we use CIFAR10-Policy and we apply ImageNet-Policy to Tiny-ImageNet. As these policies are transferable from one dataset to another, ImageNet-Policy is employed for both Clothing1M and Webvision dataset.

Method	CIFAR-10			CIFAR-100		
	10%	30%	40%	10%	30%	40%
CE	88.8	81.7	76.1	68.1	53.3	44.5
LDMI [59]	91.1	91.2	84.0	68.1	54.1	46.2
M-Up [68]	93.3	83.3	77.7	72.4	57.6	48.1
JPL [20]	94.2	92.5	90.7	72.0	68.1	59.5
PCIL [64]	93.1	92.9	91.6	76.0	59.3	48.3
DMix* [25]	93.8	92.5	91.7	71.6	69.5	55.1
ELR* [30]	**95.4**	94.7	93.0	77.3	74.6	73.2
MOIT [39]	94.2	94.1	93.2	77.4	75.1	74.0
UNICON	95.3	**94.8**	**94.1**	**78.2**	**75.6**	**74.8**

Table 3. Experimental results on CIFAR10 and CIFAR100 with asymmetric noise. UNICON sees consistent improvement for CIFAR100 dataset under different asymmetric noise settings. $(*)$ indicates that we run the algorithm.

6. Experimental Results

We present the performance of UNICON under different label noise scenarios. We start with the synthetic noisy label datasets (e.g. CIFAR10, CIFAR100 and TinyImageNet) and move on to the real world noisy datasets (e.g. WebVision, Clothing1M). For experiments, we consider symmetric noise rates of 20%, 50%, 80%, and 90% and asymmetric noise rates of 10%, 30%, and 40%.

CIFAR10 and CIFAR100 datasets: Table 2 shows the average test accuracies for these datasets. In case of CIFAR10, from moderate to severe label noise, UNICON performs consistently better than the baseline methods. For 90% noise rate, we achieve a significantly better performance improvement over the state-of-the art. For high noise rate, techniques like [25] usually fail due to high number of false positives. However, for low noise rate (20%), [25] performs slightly better than ours. Low noise rate indicates more clean samples are available for supervised learning. One possible explanation could be that the scarcity of unlabeled data (i.e. $|\mathbb{D}_{noisy}| < |\mathbb{D}_{clean}|$) makes contrastive feature learning less effective. We have also conducted experiments under the asymmetric noise scenario. In case of asymmetric noise, each class is not equally affected by label noise. This makes the selection of clean samples a bit more challenging. However, UNICON achieves similar performance gain as symmetric noise which is shown in Table 3. Note that there is an exception at 10% noise rate as [30] obtains 0.1% better accuracy than UNICON.

Table 2 and 3 contain the average test accuracies for CIFAR100 dataset. UNICON shows similar effectiveness against label noise in CIFAR100 obtaining an accuracy improvement of 11.4% for 90% noise rate. This improvement is consistent under different noise settings. While ELR [30], DMix [25] and MOIT [39] show some level of resistance to noisy labels for low noise rate, the performances are not consistent for high noise rate. Furthermore, the asymmetric noise performance of our method are also superior than other baseline methods in Table 3.

Noise (%)	0		20		50	
Alg.	Best	Avg.	Best	Avg.	Best	Avg.
Standard CE	57.4	56.7	35.8	35.6	19.8	19.6
Decoupling [34]	-	-	37.0	36.3	22.8	22.6
F-correction [41]	-	-	44.5	44.4	33.1	32.8
MentorNet [17]	-	-	45.7	45.5	35.8	35.5
Co-teaching+ [66]	52.4	52.1	48.2	47.7	41.8	41.2
M-correction [1]	57.7	57.2	57.2	56.6	51.6	51.3
NCT [44]	62.4	61.5	58.0	57.2	47.8	47.4
UNICON	**63.1**	**62.7**	**59.2**	**58.4**	**52.7**	**52.4**

Table 4. Test accuracies (%) on Tiny-ImageNet dataset under symmetric noise settings. We report the results for other methods directly from [44] with the highest (Best) and the average (Avg.) test accuracy (%) over the last 10 epochs.

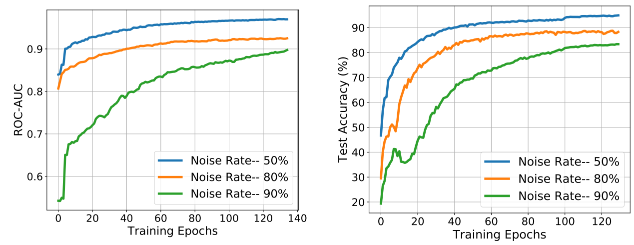
Method	Backbone	Test Accuracy
Standard CE	ResNet-50	69.21
Joint-Optim [51]	ResNet-50	72.00
MetaCleaner [69]	ResNet-50	72.50
MLNT [26]	ResNet-50	73.47
PCIL [64]	ResNet-50	73.49
JPL [20]	ResNet-50	74.15
DMix [25]	ResNet-50	74.76
ELR [30]	ResNet-50	74.81
UNICON	ResNet-50	**74.98**

Table 5. Experimental results on Clothing1M dataset. Results for previous techniques were copied from their respective papers.

Dataset	WebVision		ILSVRC12	
Method	*Top-1*	*Top-5*	*Top-1*	*Top-5*
D2L [32]	62.68	84.00	57.80	81.36
MentrorNet [17]	63.00	81.40	57.80	79.92
Co-Teaching [12]	63.58	85.20	61.48	84.70
Iterative-CV [55]	65.24	85.34	61.60	84.98
DivideMix [25]	77.32	91.64	75.20	90.84
ELR [30]	77.78	91.68	70.29	89.76
MOIT [39]	**78.76**	-	-	-
UNICON	77.60	**93.44**	**75.29**	**93.72**

Table 6. Experimental results on Webvision and ILSVRC12. All methods are trained on the Webvision while evaluated on both Webvsion and ILSVRC12 validation set. Results for baseline methods are taken from [30] and [39]. MOIT [39] does not evaluate their method on ILSVRC12 and did not provide top-5 accuracies.

TinyImageNet Dataset: Table 4 presents the performance comparison of UNICON and other state of the art methods. Even with no label noise, Tiny-ImageNet remains a challenging benchmark dataset to deal with. It becomes more challenging under the presence of label noise. One of the baseline methods M-correction [1] uses



(a) ROC-AUC score vs epochs. (b) Test Accuracy vs Epochs

Figure 5. ROC-AUC score and test accuracy (%) on CIFAR10 with different noise rates. As the model becomes more precise in selection, the test-time performance improves accordingly.

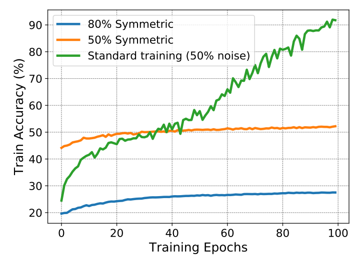


Figure 6. Training accuracy at different epochs. Low accuracy indicates that the networks do not memorize the noisy labels even after long training. In contrast to UNICON, standard CE loss-based training leads to a high training accuracy (should be ∼50%), i.e., complete memorization of noisy labels.

a loss-correction technique to tackle noisy labels while NCT [44] leverages from collaborative learning of two networks. However, both methods underperform compared to our method. Table 4 shows that UNICON gains around 1% performance improvement over SOTA for all noise rates.

Clothing1M Dataset: Table 5 presents performance comparison on this real world noisy labeled dataset. We achieve 0.17% performance improvement over ELR [30]. The performance improvement for clothing1M sometimes depends on the length of warmup, as longer period of standard CE-based training can lead to memorization. In our training, we use a warm-up period of 2,000 steps.

WebVision Dataset: We present our experimental results on this dataset in Table 6. While validating, MOIT [39] sees SOTA *Top-1* accuracy while our method achieves the best *Top-5* accuracy. We obtain around 1.5% improvement over SOTA (MOIT [39] did not provide *Top-5* accuracy.) Furthermore, UNICON secures SOTA *Top-1* and *Top-5* accuracies on ILSVRC12 validation set. While the gain in *Top-1* accuracy is not significant, we achieve a performance improvement of 1.88% over DMix [25] in *Top-5* accuracy.

6.1. Ablation Studies

In this section, we conduct an ablation study of UNICON under different training settings.

Sample Selection Performance: In general, the precision of clean sample selection directly impacts the overall

Dataset	CIFAR10						CIFAR100					
Noise Rate	50%		80%		90%		50%		80%		90%	
Method	Best	Last	Best	Last	Best	Last	Best	Last	Best	Last	Best	Last
UNICON w/o balancing	94.28	94.06	91.41	91.16	85.49	85.28	75.26	75.01	60.51	60.16	39.87	39.02
UNICON w/o CL	94.92	94.24	91.67	91.21	87.28	86.34	75.75	75.09	60.54	60.17	41.83	41.11
UNICON w/o ensemble	95.20	94.91	92.38	92.11	88.84	88.18	76.28	76.10	62.98	62.11	42.36	41.56
UNICON	**95.61**	**95.24**	**93.97**	**93.97**	**90.81**	**89.95**	**77.63**	**76.91**	**63.98**	**63.13**	**44.82**	**44.51**

Table 7. Ablation study with different training settings. Both contrastive loss and class-imbalance affects the performance significantly; especially for high noise rates. Ensembling the outputs of both network during separation seems to improve the performance as well. Test results at last epoch are also shown here.

performance of any selection-based noisy label technique. Likewise, the success of UNICON depends on how well it can separate the clean samples. Fig. 5a shows the the ROC-AUC score of our selection mechanism under different noise settings. It can be observed that UNICON sees a steady rise in the precision irrespective of the noise level. In case of high noise rate, it is usual for the network to get confused between clean and noisy samples. However, our separation approach proves to be effective even under such scenario. With improved precision, the network learns better discriminative features from labeled data and generalizes well to the unlabeled data. Through the generation of quality pseudo-labels, UNICON improves the classification accuracy significantly (Fig. 5b).

Effect of Contrastive Learning: CL is one of the key components of our framework. Table 7 indicates the impact of CL in overall performance of our method. As CL is resistant to label noise memorization, it boosts the performance significantly even in high label noise scenarios. For CIFAR10 and CIFAR100, with 90% noise rate, UNICON without CL sees 3.53% and 2.99% drop in test accuracies respectively. We explain more on contrastive learning and its impact in the *supplementary material.*

Effect of Ensemble and balancing: During selection, we take the average of both network's predictions instead of depending on just one network [12]. This seems to improve the performance significantly in case of high noise rate (see Table 7). However, taking the feedback from both networks bears the risk of confirmation bias over the course of training [25]. We prevent that by training one network at a time. During the same training epoch, we perform the separation again before training the other network. Table 7 also contains the performance of our method without balancing. The significant decrease in classification accuracies underlines the importance of class-balance prior. The effectiveness of UNICON in combating memorization can be observed in Fig. 6.

7. Limitations of UNICON

In this work, to combat label noise we employ a class-balance prior. The prior helps in combating artificial imbalance caused by current state-of-the-art selection methods. This prior can be restrictive in some extreme scenarios where the dataset itself exhibits extreme imbalance. However, in such cases, it is possible to update our prior accordingly based on the class distribution of the dataset. Since knowing the dataset distribution in advance is equally restrictive we do not explore this direction in this study. Additionally, even though we provide a general solution for combating label noise, our solution is particularly effective under high label noise. Therefore, it is possible to outperform our proposed method on datasets which do not contain a significant amount of label noise. However, we emphasize that such success can be attributed to superior training strategy and complicated design whereas our simple solution is more general and provides reasonable results even for such low noise rate scenarios.

8. Conclusion

In this work, we proposed UNICON, a simple yet effective solution for combating label noise. Our proposed uniform selection technique effectively addresses often overlooked but critical shortcoming of selection based state-of-the-art methods. Furthermore, our constrastive feature learning approach provides a fundamental solution to combat memorization of noisy label. Equipped with these two components, our method selects clean samples more precisely over the course of training by reducing the class-disparity among the true positives and CL-based unsupervised feature learning. Network trained on high precision clean samples generates higher quality pseudo-labels for the noisy label data and the overall process improves the high noise level performance significantly. UNICON achieves $\sim$10% performance improvement over state-of-the-art on 90% noisy CIFAR10 and CIFAR100. Through extensive empirical analysis, we show the effectiveness of our method under different noise scenarios.

Acknowledgement: Professor Ajmal Mian is the recipient of an Australian Research Council Future Fellowship Award (project number FT210100268) funded by the Australian Govt. This work is partly supported by the National Science Foundation under Grant No. CCF-1718195.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*, pages 312–321. PMLR, 2019. 7

[2] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pages 233–242. PMLR, 2017. 1, 3

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. 2

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 1

[5] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples, 2018. 2

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 5

[7] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data, 2019. 6

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[9] Fernando Diaz. Integration of news content into web results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 182–191, 2009. 1

[10] Yifan Ding, Liqiang Wang, Deliang Fan, and Boqing Gong. A semi-supervised two-stage approach to learning from noisy labels. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1215–1224. IEEE, 2018. 2

[11] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. 2016. 2

[12] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *arXiv preprint arXiv:1804.06872*, 2018. 2, 3, 4, 7, 8

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[14] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *arXiv preprint arXiv:1802.05300*, 2018. 2

[15] Wei Hu, Yangyu Huang, Fan Zhang, and Ruirui Li. Noise-tolerant paradigm for training face recognition cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11887–11896, 2019. 1

[16] Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. In *International Conference on Machine Learning*, pages 4804–4815. PMLR, 2020. 2

[17] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels, 2018. 2, 7

[18] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5830–5840, 2021. 1

[19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 5

[20] Youngdong Kim, Juseung Yun, Hyounguk Shon, and Junmo Kim. Joint negative and positive learning for noisy labels, 2021. 6, 7

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 5

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1

[23] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 5

[24] Hengduo Li, Zuxuan Wu, Chen Zhu, Caiming Xiong, Richard Socher, and Larry S Davis. Learning from noisy anchors for one-stage object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10588–10597, 2020. 1

[25] Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning, 2020. 1, 2, 3, 4, 6, 7, 8

[26] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan Kankanhalli. Learning to learn from noisy labeled data, 2019. 2, 5, 7

[27] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017. 1, 6

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[29] Chang Liu, Han Yu, Boyang Li, Zhiqi Shen, Zhanning Gao, Peiran Ren, Xuansong Xie, Lizhen Cui, and Chunyan Miao. Noise-resistant deep metric learning with ranking-based instance selection, 2021. 2

[30] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels, 2020. 6, 7

[31] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International Conference on Machine Learning*, pages 6226–6236. PMLR, 2020. 2

[32] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah Erfani, Shutao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *International Conference on Machine Learning*, pages 3355–3364. PMLR, 2018. 2, 7

[33] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. A new baseline for image annotation. In *European conference on computer vision*, pages 316–329. Springer, 2008. 1

[34] Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". *arXiv preprint arXiv:1706.02613*, 2017. 7

[35] Kento Nishi, Yi Ding, Alex Rich, and Tobias Höllerer. Augmentation strategies for learning with noisy labels, 2021. 2, 6

[36] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 1

[37] Curtis G Northcutt, Tailin Wu, and Isaac L Chuang. Learning with confident examples: Rank pruning for robust classification with noisy labels. *arXiv preprint arXiv:1705.01936*, 2017. 2

[38] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6913–6922, 2021. 1

[39] Diego Ortego, Eric Arazo, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Multi-objective interpolation training for robustness to label noise, 2021. 2, 6, 7

[40] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017. 2

[41] Giorgio Patrini, Alessandro Rozza, Aditya Menon, Richard Nock, and Lizhen Qu. Making neural networks robust to label noise: a loss correction approach. *arXiv preprint arXiv:1609.03683*, 2(8), 2016. 7

[42] Yuntao Qu, Shasha Mo, and Jianwei Niu. Dat: Training deep networks robust to label-noise by matching the feature distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6821–6829, 2021. 2

[43] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016. 1

[44] Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. Noisy concurrent training for efficient learning under label noise, 2020. 7

[45] Karishma Sharma, Pinar Donmez, Enming Luo, Yan Liu, and I Zeki Yalniz. Noiserank: Unsupervised label noise reduction with dependence models. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 737–753. Springer, 2020. 2

[46] Yunhang Shen, Rongrong Ji, Zhiwei Chen, Xiaopeng Hong, Feng Zheng, Jianzhuang Liu, Mingliang Xu, and Qi Tian. Noise-aware fully webly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11326–11335, 2020. 1

[47] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 2, 5

[48] Dehua Song, Yunhe Wang, Hanting Chen, Chang Xu, Chunjing Xu, and DaCheng Tao. Addersr: Towards energy efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15648–15657, 2021. 1

[49] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16519–16529, 2021. 1

[50] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 6

[51] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels, 2018. 2, 7

[52] Chih-Fong Tsai and Chihli Hung. Automatically annotating images with keywords: A review of image annotation systems. *Recent Patents on Computer Science*, 1(1):55–68, 2008. 1

[53] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 943–952, 2021. 1

[54] Xiaobo Wang, Shuo Wang, Jun Wang, Hailin Shi, and Tao Mei. Co-mining: Deep face recognition with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9358–9367, 2019. 1

[55] Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels, 2018. 2, 7

[56] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization, 2020. 2

[57] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015. 1, 5

[58] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pages 10687–10698, 2020. 1

[59] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. l_dmi: An information-theoretic noise-robust loss function, 2019. 6

[60] Cheng Xue, Qi Dou, Xueying Shi, Hao Chen, and Pheng-Ann Heng. Robust learning at noisy labeled medical images: Applied to skin lesion classification. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1280–1283. IEEE, 2019. 1

[61] Fengxiang Yang, Zhun Zhong, Zhiming Luo, Yuanzheng Cai, Yaojin Lin, Shaozi Li, and Nicu Sebe. Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4855–4864, 2021. 1

[62] Yazhou Yao, Tao Chen, Guo-Sen Xie, Chuanyi Zhang, Fumin Shen, Qi Wu, Zhenmin Tang, and Jian Zhang. Non-salient region object mining for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2623–2632, 2021. 1

[63] Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-src: A contrastive approach for combating noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5192–5201, 2021. 2, 3, 5

[64] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels, 2019. 2, 6, 7

[65] Qing Yu, Atsushi Hashimoto, and Yoshitaka Ushiku. Divergence optimization for noisy universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2515–2524, 2021. 1

[66] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W. Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption?, 2019. 2, 7

[67] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 4, 5

[68] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2018. 2, 6

[69] Weihe Zhang, Yali Wang, and Yu Qiao. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7373–7382, 2019. 2, 7

[70] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. *arXiv preprint arXiv:2103.07756*, 2021. 2

[71] Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018. 2

[72] Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. Meta label correction for noisy label learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021. 2