MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

SoREL: Soft-Label Refurbishment with Ensemble Learning for Noisy Long-Tailed Classification

Hsieh, Jun-Wei; Wu, Ying-Hsuan; Hsieh, Yi-Kuan; Li, Xin; Peng, Kuan-Chuan; Chang, Ming-Ching

TR2026-075 June 04, 2026

Abstract

Real-world datasets often suffer from both noisy labels and long-tailed distributions, where rare classes are more prone to annotation errors. Existing methods typically ad- dress these issues separately or rely on unreliable noise pre-screening, leading to biased learning and unstable optimization. We propose Soft-label Refurbishment with Ensemble Learning (SoREL), a two-stage framework that jointly handles label noise and class imbalance. In the first stage, SoREL performs robust soft-label refurbishment via contrastive learning for unbiased representation learning and a Balanced Noise-tolerant Cross-entropy (BANC) loss for stable pre-screening. In the second stage, refurbished soft labels guide multi-expert ensemble learning, where experts specialize in many-, medium-, and few-shot classes. Soft- label-based class statistics further refine loss weighting to better match the true data distribution. Experiments on simulated and real-world noisy long-tailed datasets demonstrate that SoREL achieves 91.80%/67.59% on CIFAR-10/100-LT and 77.74% / 81.40% on Food-101N and Animal-10N, significantly outperforming prior methods.

CVPR Findings 2026

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

SoREL: Soft-Label Refurbishment with Ensemble Learning for Noisy Long-Tailed Classification

Jun Wei Hsieh
National Yang Ming Chiao Tung University
jwhsieh@nycu.edu.tw

Ying-Hsuan Wu
National Yang Ming Chiao Tung University
vongola3088.ai10@nycu.edu.tw

Yi-Kuan Hsieh
National Yang Ming Chiao Tung University
yikuan0725@gmail.com

Xin Li
University at Albany, SUNY, NY
xin.li@mail.wvu.edu

Kuan-Chuan Peng
Mitsubishi Electric Research Laboratories
kpeng@merl.com

Ming-Ching Chang
University at Albany, SUNY, NY
mingching@gmail.com

Abstract

Real-world datasets often suffer from both noisy labels and long-tailed distributions, where rare classes are more prone to annotation errors. Existing methods typically address these issues separately or rely on unreliable noise pre-screening, leading to biased learning and unstable optimization. We propose Soft-label Refurbishment with Ensemble Learning (SoREL), a two-stage framework that jointly handles label noise and class imbalance. In the first stage, SoREL performs robust soft-label refurbishment via contrastive learning for unbiased representation learning and a Balanced Noise-tolerant Cross-entropy (BANC) loss for stable pre-screening. In the second stage, refurbished soft labels guide multi-expert ensemble learning, where experts specialize in many-, medium-, and few-shot classes. Soft-label-based class statistics further refine loss weighting to better match the true data distribution. Experiments on simulated and real-world noisy long-tailed datasets demonstrate that SoREL achieves 91.80%/67.59% on CIFAR-10/100-LT and 77.74% / 81.40% on Food-101N and Animal-10N, significantly outperforming prior methods.

Keywords: Long-Tail Learning; Noisy Label; Label Refurbishment; Defect Detection

1. Introduction

Deep learning has achieved remarkable success across vision and language tasks, largely enabled by large-scale datasets with clean, balanced annotations [30]. In practice, however, real-world data rarely meets these conditions. Annotation errors and heterogeneous sources introduce noisy labels [43], while natural category frequencies often follow a long-tailed distribution [49], in which a few classes dominate and many classes have only a handful of samples. These two challenges frequently co-occur, making robust model training particularly difficult.

Most prior work addresses noisy labels or long-tailed data individually, but their effectiveness drops when both problems are present. Methods for noisy labels typically assume balanced data, ignoring rare classes and becoming more vulnerable to noise [33]. Similar learning patterns between clean tail-class samples and mislabeled data can also lead to incorrect corrections, further exacerbating imbalance [2]. Conversely, long-tailed learning approaches often rely on re-sampling or re-weighting to balance class distributions [29]. When labels are noisy, however, these strategies may amplify errors and mislead the model, particularly for underrepresented classes.

Existing approaches struggle to maintain robustness when noisy labels and long-tailed distributions coexist. Many methods attempt to separate noisy from clean samples within tail classes for further processing [24, 40, 46], but pre-screening networks trained on long-tailed data are prone to misclassification. For example, noisy labels from head classes can resemble clean samples from tail classes, causing confusion. Other approaches [20] rely on feature distributions to identify noisy samples, yet this becomes unreliable for minority classes, where even a few mislabeled samples can dominate learning.

To overcome these limitations, we propose **Soft-label Refurbishment with Ensemble Learning (SoREL)**, a two-stage framework that jointly addresses label noise and class

imbalance. Instead of making hard decisions, SoREL performs *soft-label refurbishment* to integrate class distribution information, original annotations, and model predictions to refine all labels across the dataset. Motivated by the robustness of ensemble learning in long-tailed settings [8], these refined labels are leveraged to guide *multi-expert ensemble training*, providing a holistic and reliable framework for learning from data that is both noisy and imbalanced. The overall architecture is illustrated in Fig. 1.

The first stage of SoREL performs unsupervised contrastive learning to produce robust feature representations for all samples. This approach is naturally resistant to both class imbalance and label noise, producing unbiased embeddings [9, 23]. Based on these features, we introduce a **BAlanced Noise-tolerant Cross-entropy (BANC)** loss to train a balanced and noise-resilient classifier for data prescreening. The classifier outputs are then combined with original annotations, weighted by prediction confidence and class rarity, to generate refined soft labels. Unlike prior approaches that discard or overwrite labels based solely on predictions, our method adaptively blends predictions and annotations, producing soft labels that capture both reliability and class distribution information.

The second stage of SoREL tackles residual class imbalance through ensemble learning. The robust representations from the first stage are used to train three expert classifiers, each specializing in *many-shot*, *medium-shot*, and *few-shot* categories. Soft labels are further refined during this process, enhancing both robustness to long-tailed distributions and overall generalization. For the medium- and few-shot experts, loss weighting is guided by soft-label-based class distribution statistics. Unlike conventional hard-label-based weighting [13, 29], this strategy aligns optimization with the effective data distribution, improving expert calibration and overall model performance.

We evaluate SoREL on both the simulated and real-world noisy long-tailed datasets, including CIFAR-10/100-LT, Food-101N, and Animal-10N. SoREL consistently outperforms the state-of-the-art (SOTA) methods, reaching 91.80%/67.59% on CIFAR-10/100-LT and 77.74%/81.40% on Food-101N and Animal-10N, respectively, substantially beating prior long-tail (LT), noisy label (NL), and combined NL-LT approaches. The improvements are especially pronounced under high imbalance ratios and challenging noise conditions, with the gains up to 30% on CIFAR-100 with asymmetric noise. The ablation studies confirm the critical contributions of soft-label refurbishment, the BANC loss, and multi-expert ensemble learning, showing consistent performance improvements across many-, medium-, and few-shot classes, and demonstrating robustness, noise tolerance, and effective handling of imbalanced data distributions.

Our main contributions are summarized as follows:

- We propose the Soft-label Refurbishment with Ensemble Learning (SoREL) to address the challenging scenario where datasets exhibit both label noise and long-tailed distributions, which jointly compromise the performance.
- *Soft-Label Refurbishment:* We propose an advanced training strategy that combines contrastive learning-based prescreening with balanced and noise-tolerant predictions. By integrating original labels with prediction confidence and class distribution information, we generate robust soft labels that mitigate the effect of noisy annotations.
- *Multi-Expert Ensemble for Long-Tailed Learning:* Using the soft labels from the first stage, we train an ensemble of three experts specialized in many-shot, medium-shot, and few-shot categories. This strategy leverages soft-label-based class statistics to counter class imbalance, improving model robustness and generalization.
- *Beating the SOTA methods:* Our extensive experiments on the synthetic and real-world noisy long-tailed datasets show that SoREL consistently outperforms the SOTA methods, reaching 91.80%/67.59% on CIFAR-10/100-LT and 77.74%/81.40% on Food-101N and Animal-10N.

2. Related works

2.1. Label-noise learning

A common strategy for handling noisy data is to separate clean samples from noisy ones, reducing the impact of corrupted labels during training. The methods such as Co-Teaching [10] and DivideMix [21] achieve this via small-loss selection, while Jo-SRC [45] and UNICON [17] use Jensen-Shannon divergence to identify noisy samples.

Another approach focuses on correcting labels to prevent overfitting to incorrect annotations. SELFIE [32] identifies samples with consistently accurate predictions as refurbishable, and only these labels are corrected to minimize erroneous modifications. SEAL [4] computes the average softmax outputs of a deep neural network across training and retrains the network using these averaged soft labels.

Additional methods adjust the loss function [28, 31] or use regularization constraints [41] to mitigate the effect of noisy data. More recently, Robust Label Refurbishment [3] combines pseudo-labeling and confidence estimation to refurbish noisy labels, while Dynamic and Uniform Label Correction [44] employs a normalized Jensen-Shannon divergence to ensure both sample-wise adaptivity and class-wise uniformity in label correction.

2.2. Long-tail learning

Methods addressing long-tail data primarily focus on several strategies: (1) Re-balancing data distributions through resampling [25] or data augmentation [47, 51]; (2) Redesigning loss functions to improve robustness and generalization under long-tail distributions [1, 7, 29]; (3) Decoupling representation and classifier learning to reduce feature ex-

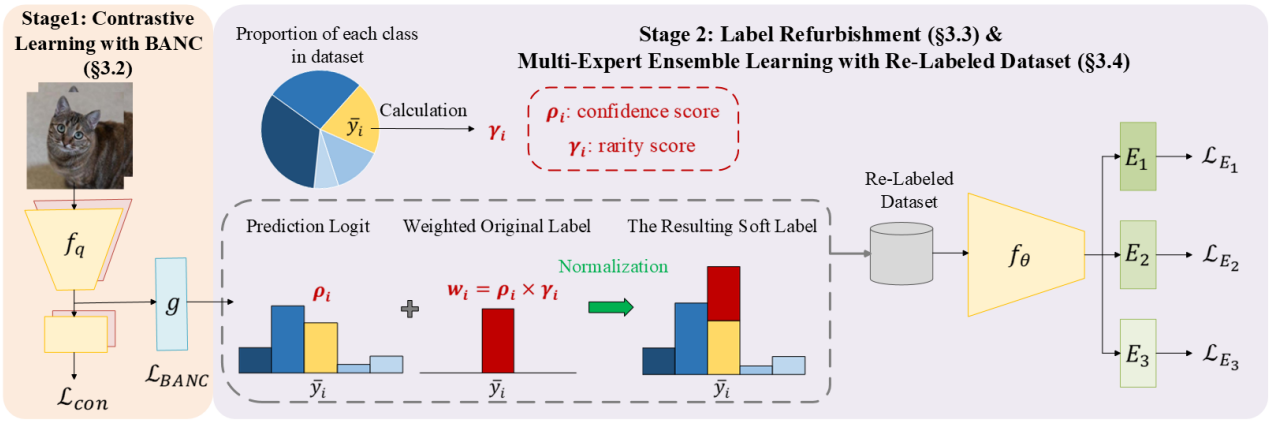

Figure 1. Our two-stage method, **Soft-label Refurbishment with Ensemble Learning (SoREL)**, tackles long-tail noisy-label learning. Stage 1 makes initial predictions via contrastive learning with a new BAlanced Noise-tolerant Cross-entropy (BANC) loss. Stage 2 refurbishes labels and ensembles three expert modules for long-tail classification.

tractor bias caused by class imbalance [16, 51]; (4) Leveraging transfer learning, transferring knowledge from head classes to tail classes [12, 37]; (5) Incorporating multi-expert models that integrate knowledge from different depths, enabling adaptive learning and mitigating negative effects on tail classes [15, 42].

Recent advances include Local and Global Logit Adjustment [36], which trains expert models on the full dataset while enhancing inter-expert differentiation via logit adjustments. Additionally, [22] proposes a re-labeling method leveraging class prototypes and distribution matching for noisy long-tailed classification.

3. Our Proposed Method – SoREL

To address the joint challenges of label noise and long-tailed distributions, we propose **Soft-label Refurbishment with Ensemble Learning (SoREL)**, a unified framework that progressively enhances label quality and model robustness, effectively mitigating both noise and imbalance in real-world datasets. SoREL consists of two complementary stages.

The first stage of SoREL focuses on learning robust representations and producing reliable initial predictions. In § 3.1, contrastive feature learning generates embeddings resilient to label noise and class imbalance. § 3.2 introduces the novel BAlanced Noise-tolerant Cross-entropy (BANC) loss to reduce residual label noise and encourage balanced learning. These components are then combined in § 3.3 to generate dependable initial predictions, forming a strong foundation for subsequent label refinement.

Building on these initial predictions, the second stage of SoREL performs soft-label refurbishment and multi-expert ensemble learning to produce the final classifications. In § 3.4, predictions are adaptively refined using both confidence and class rarity, correcting noisy labels while preserving rare classes. In § 3.5, a three-expert ensemble is trained, with each expert specialized for many-, medium-, and few-shot categories. Soft-label-based class statistics are used to refine loss weighting, aligning training with the true data distribution and improving generalization across imbalanced and noisy datasets.

Notations: Scalars are denoted by lowercase letters, and vectors by lowercase boldface letters. We consider an imbalanced and noisily labeled training dataset $\bar{\mathcal{D}} = (\boldsymbol{x}_i, \bar{y}_i), i = 1, ..., N$, where $\boldsymbol{x}_i$ is the i-th instance and $\bar{y}_i \in [K]$ is its observed label, which may be incorrect. K is the number of classes. The corresponding true label, y_i, is unobservable. Let n_k be the number of training samples in class k. Without loss of generality, we assume the classes are sorted in descending order of size, *i.e.*, $n_1 \geq n_2 \geq \cdots \geq n_K$. The goal is to train a robust classifier using only the imbalanced, noisy dataset $\bar{\mathcal{D}}$ to accurately predict the labels of unseen instances.

3.1. Contrastive feature learning

To generate robust representations for data with noisy labels and long-tailed distributions, we employ self-supervised contrastive learning, which does not rely on original labels and thus avoids bias from incorrect annotations [9]. Prior work has shown that contrastive learning also mitigates the effects of class imbalance [23].

We adopt the MoCo [6] framework, which uses two networks with identical architectures: a query network and a key network, each consisting of a CNN encoder followed by a two-layer MLP. For each input $\boldsymbol{x}_i$, two random augmentations produce views $\boldsymbol{x}_i^q$ and $\boldsymbol{x}_i^k$, which are fed into the query and key networks to obtain embeddings $\boldsymbol{z}_i^q$ and $\boldsymbol{z}_i^k$. A large feature queue Q stores additional samples, helping to learn well-structured representations.

Unlike the original MoCo, we freeze the query encoder and update only the key encoder, which has been shown to

improve robustness [5]. Let $\mathcal{A}(i)$ denote the set of all representations in the queue and batch except $\boldsymbol{x}_i$. The contrastive loss is defined as:

$$\mathcal{L}_{con}(\boldsymbol{x}_i) = -\log \frac{\exp(\boldsymbol{z}_i^q \cdot \boldsymbol{z}_i^k/\tau)}{\sum_{\boldsymbol{z}_j \in \mathcal{A}(i)} \exp(\boldsymbol{z}_i^q \cdot \boldsymbol{z}_j/\tau)}, \quad (1)$$

where $\tau > 0$ is a temperature parameter. Minimizing this loss encourages the model to produce embeddings invariant to augmentations and robust to noise and imbalance.

3.2. The BANC loss

After learning robust features via self-supervised contrastive learning, we train a classifier to correct potential label errors in imbalanced datasets. Residual noisy labels may still persist, and existing contrastive learning methods often introduce additional losses to enhance noise resilience [48]. To address this, we propose the **BAlanced Noise-tolerant Cross-entropy (BANC)** loss, which improves robustness to label noise while promoting class balance. Unlike prior approaches that rely on data distribution statistics or separate feature guidance, BANC leverages non-target class scores to counteract biases caused by noisy labels and class imbalance.

BANC is inspired by Symmetric Cross Entropy (SCE) [38], which introduces a symmetric term to improve performance under noisy labels. Let $\boldsymbol{v}_i = f_q(\boldsymbol{x}_i)$ be the feature extracted from the query encoder f_q, $\bar{\boldsymbol{y}}_i = [\bar{y}_i^1, \ldots, \bar{y}_i^K]$ the one-hot label, and $g(\cdot)$ the classifier output logits $g(\boldsymbol{v}_i) = [p_i^1, \ldots, p_i^K]$. SCE is defined as:

$$\mathcal{L}_{SCE}(\boldsymbol{x}_i) = -\sum_{k=1}^{K} \bar{y}_i^k \log \bar{p}_i^k - \sum_{k=1}^{K} \log(\bar{y}_i^k)\bar{p}_i^k, \quad (2)$$

where $\bar{p}_i^k$ is the softmax-normalized probability. To avoid the undefined $\log(0)$ in SCE, we replace $-\log(\bar{y}_i^k)$ with a linear term $c(1-\bar{y}_i^k)$, where $c > 0$ is a scaling coefficient. The resulting BANC loss is:

$$\mathcal{L}_{BANC}(\boldsymbol{x}_i) = -\sum_{k=1}^{K} \bar{y}_i^k \log \bar{p}_i^k + \sum_{k=1}^{K} c\,(1-\bar{y}_i^k)\bar{p}_i^k. \quad (3)$$

The BANC loss explicitly introduces label error tolerance and class-balancing effects. When $c > 0$ and $\bar{y}_i^k = 0$, the term $c(1-\bar{y}_i^k)$ imposes a small penalty for mispredictions, effectively mitigating the adverse effects of label noise. Similarly, for class-imbalanced scenarios, this penalty encourages the model to prioritize minority classes, fostering balanced learning across categories. Additional experiments in the supplement validate these effects. For initial prediction, the BANC loss is combined with the contrastive loss $\mathcal{L}_{con}$ [9, 23] to train a classifier that is both noise-resilient and balanced across classes.

3.3. Combined loss for initial prediction

The goal of the first stage of the SoREL framework is to produce robust embeddings and reliable initial predictions. To this end, we leverage contrastive learning to obtain noise-resistant feature representations and use the BANC loss to train a classifier resilient to label noise and class imbalance. These objectives are combined into a single loss for the first stage, denoted as $\mathcal{L}S1$:

$$\mathcal{L}_{S1} = (1-\alpha)\mathcal{L}_{con} + \alpha\mathcal{L}_{BANC}, \quad (4)$$

where α is a hyperparameter controlling the relative contribution of the two terms, with a default value of 0.2. The effect of α on model performance is analyzed in the supplement. The resulting initial predictions form a strong foundation for subsequent soft-label refinement.

3.4. Soft-label refurbishment

We propose a label refurbishment strategy to handle noisy labels and long-tailed data distributions, which is later incorporated into a multi-expert ensemble. This process refines initial predictions by generating new soft labels whenever the predicted class differs from the original label. The new soft labels are formed by combining the predicted logits with a weighted version of the original label, where the weight reflects both prediction confidence and class rarity. This ensures reliable relabeling while preserving rare-class information, producing soft labels that balance noise correction with minority-class preservation. An example of this process is illustrated in Fig. 2.

Formally, let $l_i = \arg\max_k p_i^k$ denote the predicted class for sample $\boldsymbol{x}_i$, and let $\bar{\boldsymbol{y}}_i$ be its original one-hot label. The refurbished soft label $\hat{\boldsymbol{y}}_i$ is defined as:

$$\hat{\boldsymbol{y}}_i = \begin{cases} \bar{\boldsymbol{s}}_i, & \text{if } l_i \neq \bar{y}_i, \\ \bar{\boldsymbol{y}}_i, & \text{otherwise,} \end{cases} \quad \text{and } \bar{\boldsymbol{s}}_i(k) = \frac{\boldsymbol{s}_i(k)}{\sum_{k=1}^{K} \boldsymbol{s}_i(k)}, \quad (5)$$

where $\boldsymbol{s}_i(k) = p_i^k + w_i\bar{y}_i^k$ and w_i is the weight assigned to the original label of $\boldsymbol{x}_i$, as explained next.

Weighting original labels: We assign a weight w_i to the original label based on two principles: (1) Lower confidence in the predicted class implies a higher chance of labeling error, so the original label receives a smaller weight. (2) If the original label belongs to a rare class, it is more valuable despite possible prediction errors, so the weight is increased. Formally, the weight is computed as the product of a confidence score and a rarity score:

$$w_i = \rho_i \times \gamma_i, \quad (6)$$

where $\rho_i = p_i^{\bar{y}_i}$ is the predicted confidence for the original label, and γ_i reflects class rarity. The rarity score is inversely

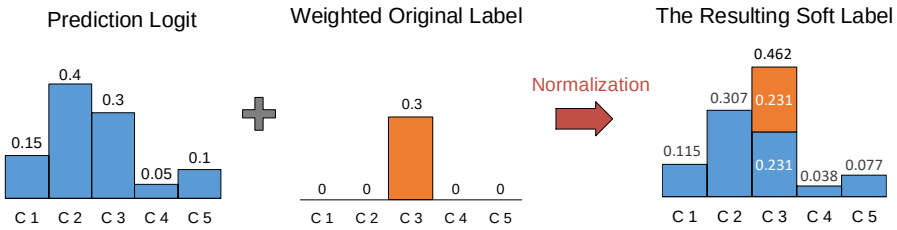

Figure 2. Label refurbishment involves the use of soft labels determined by the confidence of the original labels. In this example, the prediction score of C3 is updated after normalization.

related to the class frequency: for class $\bar{y}_i$, let $n_{\bar{y}_i}$ be the number of samples and $h_i = n_{\bar{y}_i}/N$, the proportion in the dataset. When h_i is close to 0, γ_i approaches the maximum value of 1, and when h_i is greater than 0.5, γ_i quickly decays to 0. We model γ_i using a zero-mean Gaussian-like function:

$$\gamma_i = \exp\left(-\frac{h_i^2}{\sigma^2}\right). \tag{7}$$

where the variance σ (default 0.2) controls the decay rate. This design ensures rare classes retain higher weights while common classes are down-weighted, enabling soft-label refurbishment to respect long-tailed distributions. Further details are in the supplement.

Unlike existing methods that rely solely on prediction confidence, SoREL combines a sample's predicted confidence and inherent class rarity to generate new soft labels. This strategy adapts to classes with varying sample sizes and ensures that each sample's label remains close to its true class, improving model robustness and accuracy across diverse distributions while enhancing resistance to long-tail effects and label noise.

Soft labels are generated only for samples whose predicted class differs from the original label, rather than discarding or directly overwriting them. By combining prediction confidence and class rarity, SoREL corrects mismatches while avoiding worsening long-tail imbalances. Although label noise can distort the rarity scores in Eq. (7), integrating both factors produces soft labels that better reflect the true class distribution. Unlike conventional methods [13, 29, 40] that rely solely on observed distributions, this strategy aligns samples more closely with their true classes. Samples with severe deviations between noisy and clean rarity distributions are considered unreliable and excluded.

3.5. Three-expert ensemble learning

After generating the refined soft labels, we move to the next stage of training, focusing on ensemble learning to better handle long-tail distributions. While the first stage improves label quality using contrastive learning and the BANC loss, the dataset remains imbalanced. To address this, we adopt a three-expert model that leverages a shared backbone and specialized classifiers to mitigate long-tail effects. Specifically, the model consists of: (1) a shared backbone f_θ and (2) three expert classifiers E_1, E_2, and E_3, as illustrated in Fig. 1. The backbone is initialized from the query encoder f_q trained in the first stage, ensuring robust and unbiased feature representations.

The backbone f_θ is frozen during expert training to preserve the learned representations. Each expert is specialized for different subsets of classes: E_1 targets many-shot classes, E_2 balances medium-shot classes for overall accuracy, and E_3 focuses on few-shot classes to compensate for underrepresented categories. Each expert is trained with a *shot-adaptive loss* specifically designed for its target class group, as detailed in the following paragraphs.

Many-shot expert E_1: E_1 is trained using standard softmax cross-entropy loss. Since many-shot classes dominate the dataset, gradients naturally bias predictions toward these classes. Let $g_{E_1}(\boldsymbol{v}_i)$ denote the logits predicted by E_1 for feature $\boldsymbol{v}_i = f_q(\boldsymbol{x}_i)$. Let $g^k_{E_1}(\boldsymbol{v}_i)$ be the k-th entry of $g_{E_1}(\boldsymbol{v}_i)$. Similarly, $\hat{y}_i^k$ is the k-th entry of the soft label $\hat{\boldsymbol{y}}_i$. Let $\Phi(\cdot)$ denote the softmax normalization. The loss is:

$$\mathcal{L}_{E_1}(\boldsymbol{x}_i) = -\sum_{k=1}^{K} \hat{y}_i^k \log \Phi\left(g^k_{E_1}(\boldsymbol{v}_i)\right). \tag{8}$$

Medium-shot expert E_2: E_2 is trained using a loss similar to the balanced softmax [29], where the loss is weighted according to class frequency to promote balanced performance across all classes. Let n_k denote the sample size for class k. Incorporating $\log n_k$ adjusts the gradients for medium-shot classes, improving overall accuracy as shown in Table 6. The loss is:

$$\mathcal{L}_{E_2}(\boldsymbol{x}_i) = -\sum_{k=1}^{K} \hat{y}_i^k \log \Phi\left(g^k_{E_2}(\boldsymbol{v}_i) + \log n_k\right). \tag{9}$$

Few-shot expert E_3: E_3 specializes in rare classes. Inspired by the diversity loss [50], we amplify the class weighting with $\log n_k^2$ to prioritize few-shot classes, improving recall for underrepresented classes as in Table 6. Its loss is:

$$\mathcal{L}_{E_3}(\boldsymbol{x}_i) = -\sum_{k=1}^{K} \hat{y}_i^k \log \Phi\left(g^k_{E_3}(\boldsymbol{v}_i) + \log n_k^2\right). \tag{10}$$

Training hard samples with soft labels: We use the refined soft labels $\hat{\boldsymbol{y}}_i$ as ground truth for hard samples, including

minority-class samples and those with noisy labels. This approach improves learning efficiency, prevents early saturation, and enhances model robustness and generalization.

Sample size calculation: For experts E_2 and E_3, we weigh classes based on effective sample sizes estimated from soft labels. Let $\hat{\boldsymbol{y}}_i = [y_i^1, y_i^2, \ldots, y_i^k, \ldots, y_i^K]$ be the soft label for sample i. The sample size for class k is computed as:

$$n_k = \sum_{i=1}^{N} y_i^k. \tag{11}$$

This method leverages the probabilistic information in soft labels, providing a more accurate estimate of class membership distributions.

Multi-expert inference: During inference, each expert produces a score for the input, and the final class is determined by averaging the three expert outputs. The complete SoREL pseudo code is presented in Algorithm 1 in the supplement.

4. Experiments

4.1. Experimental setup

Datasets: To rigorously evaluate the effectiveness of SoREL, we conducted experiments under both the simulated and real-world conditions characterized by long-tail distributions and label noise. In such settings, class imbalance arises when certain categories have far fewer samples than others, posing significant challenges for model learning. To create realistic long-tail scenarios, we followed established experimental protocols from prior long-tailed learning studies [1, 7], adjusting the degree of imbalance to control the ratio between the most and least represented classes.

For real-world evaluation, we used benchmark datasets that naturally exhibit noisy labels and long-tail characteristics, including **Animal-10N** [32] and **Food-101N** [19]. We focused on Animal-10N and Food-101N to highlight the robustness of SoREL in challenging real-world settings. To analyze performance under controlled imbalance, we applied synthetic long-tail adjustments to these datasets following prior protocols.

We further evaluated SoREL on the simulated noisy and long-tailed datasets derived from CIFAR-10 [18] and CIFAR-100 [18]. After constructing a long-tail distribution, we introduced controlled label noise under the two commonly used settings: *symmetric* and *asymmetric* noise [28, 35]. In the symmetric setting, a portion of training labels was randomly replaced with any other class label, simulating uniform noise. In contrast, asymmetric noise was generated through label flipping between semantically similar categories, for example replacing truck with automobile, bird with airplane, deer with horse, and cat with dog in CIFAR-10, while similar flips occurred within super-classes in CIFAR-100. The noise ratio indicates the proportion of mislabeled samples relative to the total dataset size.

Baselines: We compare SoREL against three categories of representative approaches: (1) **Long-tail learning (LT)** methods: LA [26], LDAM [1], BBN [52], LWS [16], and IB [27]. (2) **Label-noise learning methods (NL)** methods: DivideMix [21], Co-learning [34], JoCoR [39], and UNICON [17]. (3) **Joint noisy-label and long-tail (NL-LT)** methods: HAR [2], RoLT [40], ULC [14], MW-Net [31], H2E [46], and TABASCO [24]. Additional comparisons with RCAL [48] are provided in the supplement.

Implementation details: For fair comparison, we adopt the same backbone architectures as used in prior studies. Following [24], we use ResNet18 [11] as the backbone for CIFAR and Animal-10N, and following [46], we use ResNet50 for Food-101N. The initial learning rate is set to 0.02 for all datasets except Animal-10N, where it is 0.001. We employ a cosine annealing schedule for learning rate decay and train all models from scratch using SGD with momentum 0.9 and weight decay 5×10^{-4}. The batch size is 128 in Stage 1 and 512 in Stage 2, with each stage trained for 200 epochs.

All experiments are conducted using the same parameter configuration: $c = 6$, $\alpha = 0.2$, and $\tau = 0.2$. The comparative results are reported based on the values provided by TABASCO [24] and H2E [46]. The effect of the scaling coefficient c is analyzed in the supplement.

4.2. Evaluation results

Results on simulated CIFAR-10/100: Table 1 compares classification accuracy under different symmetric noise rates on CIFAR-10 and CIFAR-100. Our SoREL consistently outperforms all baselines under an imbalance ratio of 10 across all noise levels. On the more challenging CIFAR-100 dataset, SoREL surpasses TABASCO by over 10%.

Table 2 reports results under asymmetric noise rates of 0.2 and 0.4, also with imbalance ratio 10. SoREL achieves the best performance in all cases, maintaining a 20% to 40% margin over competing methods. As noise severity increases, SoREL shows notably smaller accuracy drops, showing strong robustness to noise in long-tailed settings. These results suggest that the methods relying solely on noisy sample identification may degrade as imbalance grows, while SoREL remains stable through joint label refinement and rarity modeling.

Results on Food-101N and Animal-10N: Table 3 presents the evaluation results on the real-world noisy and imbalanced datasets, Food-101N and Animal-10N. The evaluation sets for these datasets are clean and do not exhibit class imbalance. In all scenarios, SoREL consistently outperforms the SOTA approaches. Notably, on the larger-scale Food-101N dataset with higher noise levels, SoREL achieves the most robust performance, maintaining strong accuracy even as the imbalance ratio increases. These results demonstrate the

Table 1. The testing accuracy (%) on simulated CIFAR-10 and CIFAR-100 with **symmetric** noise. The highest scores are marked in **bold**.

Imbalance Ratio		10				100			
Dataset		CIFAR-10		CIFAR-100		CIFAR-10		CIFAR-100	
Noise Rate (**Symmetric**)		0.4	0.6	0.4	0.6	0.4	0.6	0.4	0.6
Baseline	CE	71.67	61.16	34.53	23.63	47.81	28.04	21.99	15.51
LT	LA	70.56	54.92	29.07	23.21	42.63	36.37	21.54	13.14
	LDAM	70.53	61.97	31.30	23.13	45.52	35.29	18.81	12.65
	IB	73.24	62.62	32.40	25.84	49.07	32.54	20.34	12.10
NL	DivideMix	82.67	80.17	54.71	44.98	32.42	34.73	36.20	26.29
	UNICON	84.25	82.29	52.34	45.87	61.23	54.69	32.09	24.82
NL-LT	HAR	77.44	63.75	38.17	26.09	51.54	38.28	20.21	14.89
	RoLT	81.62	76.58	42.95	32.59	60.11	44.23	23.51	16.61
	ULC	84.46	83.25	54.91	44.66	45.22	50.56	33.41	25.69
	MW-Net	70.90	59.85	32.03	21.71	46.62	39.33	19.65	13.72
	TABASCO	85.53	84.83	56.52	45.98	62.34	55.76	36.91	26.25
	OT [22]	86.38	83.86	56.71	48.07	-	-	-	-
Ours	SoREL	**89.45**	**86.34**	**64.55**	**57.63**	**78.32**	**69.85**	**45.03**	**36.34**

Table 2. The testing accuracy (%) on simulated CIFAR-10 and CIFAR-100 with **asymmetric** noise. The highest scores are marked in **bold**.

Imbalance Ratio		10				100			
Dataset		CIFAR-10		CIFAR-100		CIFAR-10		CIFAR-100	
Noise Rate (**Asymmetric**)		0.2	0.4	0.2	0.4	0.2	0.4	0.2	0.4
Baseline	CE	79.90	62.88	44.45	32.05	56.56	44.64	25.35	17.89
LT	LA	71.49	59.88	39.34	28.49	58.78	59.88	39.34	28.49
	LDAM	74.58	62.29	40.06	33.26	61.25	40.85	29.22	18.65
	IB	73.49	58.36	45.02	35.25	56.28	42.96	31.15	23.40
NL	DivideMix	80.92	69.35	58.09	41.99	41.12	42.79	38.46	29.69
	UNICON	72.81	69.04	55.99	44.70	53.53	34.05	34.14	30.72
NL-LT	HAR	82.85	69.19	48.50	33.20	62.42	51.97	27.90	20.03
	RoLT	73.30	58.29	48.19	39.32	54.81	50.26	32.96	-
	ULC	74.07	73.19	54.45	43.20	41.14	22.73	34.07	25.04
	MW-Net	79.34	65.49	42.52	30.42	62.19	45.21	27.56	20.04
	TABASCO	82.10	80.57	59.39	50.51	62.98	54.04	40.35	33.15
	OT [22]	85.49	-	60.45	52.08	-	-	-	-
Ours	SoREL	**91.80**	**88.17**	**67.59**	**52.78**	**85.60**	**78.17**	**49.20**	**35.69**

effectiveness of SoREL in handling real-world noisy and long-tailed distributions.

4.3. Ablation study and discussions

On simulated CIFAR-100: We conducted a comprehensive ablation study on the simulated CIFAR-100 dataset with an asymmetric noise rate of 0.4, evaluating each component of SoREL under the imbalance ratios of 10 and 100 (see Table 4). The study yields the following insights:

1. **Contrastive learning baseline:** Even under noisy labels and long-tailed distributions, self-supervised contrastive learning demonstrates strong robustness.
2. **BANC loss:** Incorporating BANC on top of contrastive learning improves performance by roughly 3%, enhancing the reliability of initial predictions.
3. **Label refurbishment and multi-expert ensemble:** Applying these components further boosts performance by about 4%. Even when cross-entropy is used in the first stage, label refurbishment and ensemble learning still provide a 1% gain, highlighting their efficacy and stability.
4. **Impact of label refurbishment:** Label refurbishment alone is critical. Without it, ensemble learning performance drops substantially—by 7% and 2% under imbalance ratios of 10 and 100, respectively. This underscores the importance of correcting initial labels, which otherwise can distort class sample statistics and negatively affect the weighted loss during expert training.

The following ablation studies further evaluate the individual contributions of label refurbishment and multi-expert ensemble learning.

Effectiveness of label refurbishment: We performed additional experiments on the simulated CIFAR-100 with an imbalance ratio of 100 to assess the impact of label refurbishment. Comparisons were made among first-stage predictions, second-stage predictions with label refurbishment, and second-stage predictions without it. Since omitting label refurbishment can negatively affect the loss weighted by class sample size, the second-stage classifier was trained using only cross-entropy for a fair assessment. Table 5 shows that incorporating label refurbishment improves second-stage performance by approximately 2%, confirming the effective-

Table 3. The testing accuracy (%) on Food-101N and Animal-10N with varying imbalance ratios. The highest scores are marked in **bold**.

Dataset		Food-101N			Animal-10N		
Imbalance Ratio		20	50	100	20	50	100
Baseline	CE	57.21	49.94	44.71	66.10	59.94	53.02
LT	LA	62.81	55.42	52.30	69.08	67.78	61.89
	LDAM	61.35	59.29	48.61	75.40	72.82	68.21
	BBN	63.44	57.89	53.16	72.14	70.26	60.08
	LWS	61.29	54.42	51.10	71.16	69.35	62.40
NL	DivideMix	69.46	57.15	42.80	72.43	65.77	47.60
	Co-learning	53.76	45.92	35.10	61.70	52.76	43.23
	JoCoR	49.07	32.98	33.49	51.29	44.02	37.19
NL-LT	HAR	59.95	52.45	46.12	71.92	68.43	62.19
	CL+LA	50.16	42.18	39.13	54.14	46.23	41.92
	Co-teaching-WBL	58.04	52.12	53.97	72.43	71.06	66.60
	H2E	70.35	63.69	58.66	77.04	74.94	66.58
Ours	SoREL	**77.74**	**73.14**	**70.60**	**81.40**	**78.88**	**75.48**

Table 4. The ablation study of testing accuracy (%) on the simulated CIFAR-100. CL: self-supervised contrastive learning; Re-Label: label refurbishment; Multi-Exp: multi-expert ensemble Learning. Symbol ✗ in CL denotes training with standard supervised learning. Symbol ✗ in BANC indicates that cross-entropy loss is used.

Component				Imbalance Ratio	
CL	BANC	Re-Label	Multi-Expert	10	100
✗	✗	✗	✗	32.05	17.89
✓	✗	✗	✗	43.86	34.43
✓	✗	✓	✓	44.94	35.40
✓	✓	✗	✗	47.37	34.64
✓	✓	✓	✓	**51.70**	**35.69**

Table 5. The ablation study of testing accuracy (%) on the simulated CIFAR-100 with imbalance ratio 100. Noise rates are asymmetric at 0.4 and symmetric at 0.6. Re-Label: label refurbishment.

Stage	**Asymmetric:** 0.4	**Symmetric:** 0.6
Stage 1	34.63	33.79
Stage 2 w/o Re-Label	33.41	23.68
Stage 2	**35.69**	**36.34**

Table 6. Ablation study of test accuracy(%) of many / medium / few classes on the simulated CIFAR-100 with an symmetric noise 0.6 and imbalance ratio 100.

Model	Many	Medium	Few	All
Expert E_1	**85.00**	**64.80**	15.47	34.13
Expert E_2	72.50	61.45	20.33	35.76
Expert E_3	36.00	49.26	**23.82**	32.96
Ensemble	75.55	62.17	20.74	**36.34**

ness of this strategy.

Effectiveness of multi-expert ensemble learning: We further evaluated our three-expert ensemble on the simulated CIFAR-100 with an imbalance ratio of 100, focusing on three class subgroups: (1) many-shot classes with more than 100 training images, (2) medium-shot classes with 20–100 images, and (3) few-shot classes with fewer than 20 images. Table 6 shows that Expert E_1 excels in many-shot classes, Expert E_3 specializes in few-shot classes, and Expert E_2 provides balanced performance across all subgroups. The ensemble of the three experts enhances overall performance, with notable improvements observed in each class subgroup.

5. Conclusions

We propose Soft-label Refurbishment with Ensemble Learning (SoREL), a two-stage framework to address the challenges of long-tailed classification with noisy labels. SoREL integrates a soft label refurbishment strategy with multi-expert ensemble learning. Soft-label refurbishment mitigates the impact of noisy annotations using a noise-tolerant mechanism, while the three expert classifiers, specializing in many-shot, medium-shot, and few-shot classes, effectively handle class imbalance. Each expert is trained with adaptive loss functions tailored to the frequency of its target classes. Compared to prior methods based on bi-dimensional sample selection and representation calibration, SoREL demonstrates superior robustness, accuracy, and generalization across diverse datasets. Our extensive experiments confirm the effectiveness and stability of SoREL.

Limitations: Despite strong performance, SoREL has some limits: The two-stage training can be time consuming on very large datasets; heuristic rarity estimation and class grouping may limit its generalization to complex or dynamic distributions; it assumes known class frequencies, often unavailable in evolving datasets; and it still needs broader testing on diverse real-world domains to fully assess robustness.

Future work will focus on jointly optimizing the two stages of SoREL in an end-to-end manner to improve training efficiency and reduce computational cost. We also plan to explore automated strategies for determining class grouping and adaptive weighting to better handle dynamic or unknown distributions. Extending SoREL to additional real-world noisy datasets and other modalities, such as video or multi-label data, will further validate its robustness and practical applicability. In addition, integrating more complex uncertainty estimation or self-supervised learning techniques may enhance resilience to extreme noise and imbalance.

References

[1] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 2, 6

[2] Kaidi Cao, Yining Chen, Junwei Lu, Nikos Arechiga, Adrien Gaidon, and Tengyu Ma. Heteroskedastic and imbalanced deep learning with adaptive regularization. *arXiv preprint arXiv:2006.15766*, 2020. 1, 6

[3] Mingcai Chen, Hao Cheng, Yuntao Du, Ming Xu, Wenyu Jiang, and Chongjun Wang. Two wrongs don't make a right: Combating confirmation bias in learning with label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14765–14773, 2023. 2

[4] Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11442–11450, 2021. 2

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 4

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[7] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 2, 6

[8] Yingxiao Du and Jianxin Wu. No one left behind: Improving the worst categories in long-tailed learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15804–15813, 2023. 2

[9] Aritra Ghosh and Andrew Lan. Contrastive learning improves model robustness under label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2703–2708, 2021. 2, 3, 4

[10] 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. *Advances in neural information processing systems*, 31, 2018. 2

[11] 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

[12] Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei. Distilling virtual examples for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 235–244, 2021. 3

[13] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6626–6636, 2021. 2, 5

[14] Yingsong Huang, Bing Bai, Shengwei Zhao, Kun Bai, and Fei Wang. Uncertainty-aware learning against label noise on imbalanced datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6960–6969, 2022. 6

[15] Yan Jin, Mengke Li, Yang Lu, Yiu-ming Cheung, and Hanzi Wang. Long-tailed visual recognition via self-heterogeneous integration with knowledge excavation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23695–23704, 2023. 3

[16] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 3, 6

[17] Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9676–9686, 2022. 2, 6

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[19] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5447–5456, 2018. 6

[20] Hao-Tian Li, Tong Wei, Hao Yang, Kun Hu, Chong Peng, Li-Bo Sun, Xun-Liang Cai, and Min-Ling Zhang. Stochastic feature averaging for learning with long-tailed noisy labels. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 3902–3910, 2023. 1

[21] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR 2020*, 2020. 2, 6

[22] Zhuo Li, He Zhao, Anningzhe Gao, Dandan Guo, Tsung-Hui Chang, and Xiang Wan. Prototype-oriented clean subset extraction for noisy long-tailed classification. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2025. 3, 7

[23] Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*, 2021. 2, 3, 4

[24] Yang Lu, Yiliang Zhang, Bo Han, Yiu-ming Cheung, and Hanzi Wang. Label-noise learning with intrinsically long-tailed data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1369–1378, 2023. 1, 6

[25] 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

[26] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020. 6

[27] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 735–744, 2021. 6

[28] 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, 6

[29] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020. 1, 2, 5

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[31] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32, 2019. 2, 6

[32] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*, pages 5907–5915. PMLR, 2019. 2, 6

[33] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 1

[34] Cheng Tan, Jun Xia, Lirong Wu, and Stan Z Li. Co-learning: Learning from noisy labels with self-supervision. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1405–1413, 2021. 6

[35] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5552–5560, 2018. 6

[36] Yingfan Tao, Jingna Sun, Hao Yang, Li Chen, Xu Wang, Wenming Yang, Daniel Du, and Min Zheng. Local and global logit adjustments for long-tailed learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11783–11792, 2023. 3

[37] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020. 3

[38] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 322–330, 2019. 4

[39] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13726–13735, 2020. 6

[40] Tong Wei, Jiang-Xin Shi, Wei-Wei Tu, and Yu-Feng Li. Robust long-tailed learning under label noise. *arXiv preprint arXiv:2108.11569*, 2021. 1, 5, 6

[41] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*, 2020. 2

[42] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 247–263. Springer, 2020. 3

[43] 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

[44] Yuanzhuo Xu, Xiaoguang Niu, Jie Yang, Ruiyi Su, Jian Zhang, Shubo Liu, and Steve Drew. Revisiting interpolation for noisy label correction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20):21833–21841, 2025. 2

[45] 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

[46] Xuanyu Yi, Kaihua Tang, Xian-Sheng Hua, Joo-Hwee Lim, and Hanwang Zhang. Identifying hard noise in long-tailed sample distribution. In *European Conference on Computer Vision*, pages 739–756. Springer, 2022. 1, 6

[47] Yuhang Zang, Chen Huang, and Chen Change Loy. Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3457–3466, 2021. 2

[48] Manyi Zhang, Xuyang Zhao, Jun Yao, Chun Yuan, and Weiran Huang. When noisy labels meet long tail dilemmas: A representation calibration method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15890–15900, 2023. 4, 6

[49] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[50] Qihao Zhao, Chen Jiang, Wei Hu, Fan Zhang, and Jun Liu. Mdcs: More diverse experts with consistency self-distillation for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11597–11608, 2023. 5

[51] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16489–16498, 2021. 2, 3

[52] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020. 6