Supplementary Material

1. Related Literature

1.1. Noisy Label Learning

Noisy Label Learning (NLL) has been studied extensively over the years. Previous NLL approaches mainly focus on, but are not limited to, the following aspects [32]: (1) Noise-robust loss design and adjustment [23, 26, 28, 30, 38, 48]; (2) Robust architectures [4, 17]; (3) Robust regularization [5, 10, 25, 45]. Recent SOTA NLL methods primarily mitigate the impact of label noise by distinguishing clean data from noisy data using well-designed criteria (sample selection), such as the *small-loss trick* [6, 19], Jensen-Shannon divergence [14, 42], and confidence-based methods [21]. However, when long-tailed data distributions are present, most of these metrics become ineffective, as clean data from tail classes exhibit similar training behaviors to noisy data.

1.2. Long-Tailed Learning

Long-Tailed Learning (LTL) methods primarily address the class imbalance problem through re-sampling [3, 13, 31] and re-weighting [29, 35, 37, 47] techniques. Additionally, recent work has emphasized the role of transfer learning in LTL methods [15, 36, 41, 43], leveraging many classes to aid the learning of few-shot classes in various ways. However, these methods are based on the assumption that all training data are clean, which does not hold true in real-world scenarios. Also, for transfer learning methods, the main consideration is to transfer knowledge from the head class to the tail class, but the fact that knowledge can be transferred between classes is usually ignored.

1.3. Long-Tailed Noisy Label Learning

Long-Tailed Noisy Label Learning (LTNLL) addresses the challenge of label noise in conjunction with long-tailed distributions. As discussed earlier, previous LTNLL methods can be broadly classified into two categories: discriminative methods [11, 12, 24, 34] and representational methods [1, 44, 46]. Among these, SFA [18] and RCAL [46], stand out as the current SOTA methods.

2. Technical Details

2.1. Noisy Label Learning

- **DivideMix** [19]. DivideMix simultaneously trains two networks, leveraging dataset co-division, label co-refinement, and co-guessing to achieve robustness against label noise.
- **DISC** [21]. DISC introduces a Dynamic Instance-specific Selection and Correction approach for noisy label learning (NLL). It effectively divides noisy data into subsets by setting instance-specific thresholds, thus mitigating label noise during model training.

2.2. Long-tailed Learning

- **cRT** [13]. This approach suggests that data imbalance does not hinder the learning of high-quality representations. Strong long-tailed recognition can be achieved by adjusting only the classifier, with the learning process decoupled into representation learning and classifier learning. cRT retrains the classifier using class-balanced sampling.
- **RIDE** [37]. RIDE reduces model variance through multiple experts, mitigates model bias with a distribution-aware diversity loss, and cuts computational cost with a dynamic expert routing module.
- **SADE** [47]. Building on RIDE, SADE introduces a novel test-time expert aggregation strategy that uses self-supervision to aggregate the learned experts, addressing unknown test class distributions.
- **DSCL** [41]. DSCL decouples two types of positive samples in self-contrastive learning (SCL) and optimizes their relationships toward distinct objectives to alleviate the impact of class imbalance. It proposes a patch-based self-distillation module that transfers knowledge from head to tail classes to address the underrepresentation of tail classes, utilizing patch-based features to identify shared visual patterns across instances.

2.3. Long-tailed Noisy Label Learning

- **HAR** [1]. HAR introduces a regularization technique that unifies the handling of noisy labels and class-imbalanced data. It assigns varying regularization strengths to data points, with higher uncertainty and lower density points receiving greater regularization.

- **RoLT/RoLT+** [39]. RoLT distinguishes mislabeled examples from rare examples by designing a class-dependent noise detector based on the distance to class prototypes. RoLT+ further enhances robustness by employing semi-supervised methods.
- **PCL** [34]. Prototypical Classifier (PCL) does not require additional parameters for the embedding network. Unlike conventional classifiers, which tend to be biased toward head classes, PCL provides balanced predictions across all classes, even in class-imbalanced training datasets. By leveraging this feature, noisy labels can be detected by thresholding the confidence scores produced by PCL, with the threshold dynamically adjusted during training. A sample reweighting strategy is used to mitigate the impact of noisy labels.
- **RCAL/RCAL+** [46]. RCAL employs a representation calibration framework that adjusts the means and covariances of tail classes by weighted averages of their nearest head classes. RCAL+ further improves robustness through the use of semi-supervised methods.
- **TABASCO** [24]. TABASCO addresses label-noise learning in intrinsically long-tailed data. It introduces a two-stage bi-dimensional sample selection process to better separate clean and noisy samples, particularly in tail classes. TABASCO features two complementary separation metrics, overcoming the limitations of using a single metric in sample separation.
- **SFA** [18]. SFA uses a distance-based sample selection algorithm to identify clean instances, guiding the training process. As class prototypes derived from inaccurate supervision may be unreliable, SFA initially selects high-confidence instances to compute the class prototypes, updating them using a running average.
- **OT** [22]. This method proposes a pseudo-labeling approach that uses class prototypes to match distributions. By employing optimal transport (OT), it aligns the training samples with class prototypes to mitigate the effects of noise and imbalance. A filtering criterion is applied to extract a clean and balanced subset of the dataset, which helps in training a more robust model.

3. Datasets and Implementation Details

Simulated Noisy and Imbalanced Datasets. We validate IBC on CIFAR-10 [16] and CIFAR-100 [16] with varying noise rates and imbalance ratios. CIFAR-10 contains 10 classes, with 50,000 training images and 10,000 test images, each of size 32×32. CIFAR-100 has 100 classes, with 50,000 training images and 10,000 test images, also of size 32×32.

To simulate realistic conditions, we first create imbalanced versions of CIFAR-10 and CIFAR-100, followed by label noise injection. We apply long-tailed class imbalance by reducing the number of examples in each class using an exponential function:

$$n_k = n_0 \cdot k^v,$$

where n_k is the number of instances in the k-th class, n_0 is the original number of instances, and $v \in (0, 1)$. The imbalance ratio ρ is defined as the ratio between the sample size of the most frequent (head) class and the most scarce (tail) class. We simulate label noise by following the method in [18], where the probability that the true label i is corrupted to the noisy label j is given by:

$$T_{ij}(x) = \begin{cases} 1 - \eta, & \text{if } i = j, \\ \frac{n_j}{n - n_i}\eta, & \text{otherwise}, \end{cases}$$

where η is the noise rate, and n_i and n_j are the number of instances in classes i and j, respectively. We explore noise rates $\eta \in \{0.2, 0.5\}$ and imbalance ratios $\rho \in \{50, 100, 200\}$ in our experiments.

Real-world Noisy and Imbalanced Datasets. We also evaluate IBC on real-world datasets, including WebVision [20] and Clothing1M [40]. WebVision consists of 2.4 million images crawled from the web, with 1,000 concepts shared with ImageNet ILSVRC12. Following the "mini" setting in [2], we use the first 50 classes of the Google resized image subset and name it mini-WebVision. We test the trained network on the WebVision-50 validation set and the ILSVRC12 validation set. Clothing1M contains 1 million training images, and 50k, 14k, and 10k images with clean labels for training, validation, and testing, with 14 classes. Following the setting in [46], we exclude the 50k and 14k clean data in our experiment.

Implementation Details. All implementation codes are in PyTorch [27]. For both CIFAR-10 and CIFAR-100, we use a Pre-Act ResNet-18 [8] architecture and apply standard weak augmentations for all images. We utilize the official MoCo implementation [9] in PyTorch. For WebVision-50, we use the Inception-ResNet-v2 [33] backbone. For Clothing1M, we use ResNet-50 [7] as the backbone network. The following hyperparameters were tuned across experiments: - For MoCo training, we use SGD with momentum of 0.9, weight decay of 5×10^{-4}, queue size of 4096, learning rate of 0.03, batch size of 64, and 4096 training epochs. - For CIFAR-10 and CIFAR-100, we warm up for 30 epochs, use SGD with momentum of 0.9, weight decay of 5×10^{-4}, and set the batch size to 64 for 200 training epochs. The initial learning rate is set to 0.02, reduced by a factor of 10 after 150 epochs. For CIFAR-10, we choose k from $\{1, 2, 3\}$ and δ from $\{0.1, 0.2, 0.3\}$; for CIFAR-100, k is chosen from $\{10, 20, 30, 40, 50\}$ and δ from $\{0.1, 0.2, 0.3\}$. - For mini-WebVision, we use SGD with momentum of 0.9,

weight decay of 1×10^{-3}, batch size of 32, and 100 training epochs. The initial learning rate is 0.01, reduced by a factor of 10 after 50 epochs, with k chosen from $\{5, 10, 15\}$ and δ from $\{0.1, 0.2, 0.3\}$. - For Clothing1M, we use Adam with a fixed learning rate of 0.001, batch size of 256, and 200 training epochs. We select k from $\{1, 3, 5\}$ and δ from $\{0.1, 0.2, 0.3\}$.

4. Algorithm

Algorithm 1 Pseudo-code of IBC

Input: training dataset $\tilde{\mathcal{D}} = \{(x_i, y_i)\}_{i=1}^{n}$, encoder network f, classifiers E_h, E_m, E_t, holistic model parameter Θ, pre-training epochs T_p, total training epochs T_{max}, redistribution strength δ and k-nearest neighbor selection parameter k.
for $t = 1, \ldots, T_p$ **do**
    Pre-train the encoder network f with MoCo [9].
end for
for $t = 0, \ldots, T_{max}$ **do**
    for $k = 1, \ldots, K$ **do**
        Compute class prototypes $\mathbf{C}_k^{(t)}$ by Eq. (3).
        Compute Euclidean distance $dist(\mathbf{C}_k^{(t)}, x_i)$ by Eq. (4).
        Obtain $\mathcal{D}_{\text{clean}}^k$ and $\mathcal{D}_{\text{noisy}}^k$ by Eq. (6).
    end for
    $\mathcal{D}_{\text{clean}} = \bigcup_{k=1}^{K} \mathcal{D}_{\text{clean}}^k$, $\mathcal{D}_{\text{noisy}} = \bigcup_{k=1}^{K} \mathcal{D}_{\text{noisy}}^k$.
    Compute cosine similarity and k-NN prototypes for each instance by Eq. (7) and Eq. (8).
    Construct three shot-specific soft labels: $\mathbf{y}_i^{(h)}$, $\mathbf{y}_i^{(m)}$, $\mathbf{y}_i^{(t)}$ for each instance by Eq. (9).
    Forward clean instances losses $\mathcal{L}_{\text{clean}} = \mathcal{L}_h + \mathcal{L}_m + \mathcal{L}_t$ by Eq. (10) $\sim$ Eq. (12).
    Forward noisy instances losses $\mathcal{L}_{\text{noisy}} = \text{MixMatch}(\mathcal{D}_{\text{clean}}, \mathcal{D}_{\text{noisy}}, f)$
    Backward total loss $\mathcal{L} = \mathcal{L}_{\text{clean}} + \mathcal{L}_{\text{noisy}}$.
    Update parameters: $\Theta_t = \text{SGD}(\mathcal{L}, \Theta_{t-1})$.
end for
return parameters of Θ.

References

[1] Kaidi Cao, Yining Chen, Junwei Lu, Nikos Arechiga, Adrien Gaidon, and Tengyu Ma. Heteroskedastic and imbalanced deep learning with adaptive regularization. In *International Conference on Learning Representations*, 2021. 1

[2] Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *International Conference on Machine Learning*, pages 1062–1070. PMLR, 2019. 2

[3] Chengjian Feng, Yujie Zhong, and Weilin Huang. Exploring classification equilibrium in long-tailed object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3397–3406, 2021. 1

[4] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *International Conference on Learning Representations*, 2017. 1

[5] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Machine Learning*, 2015. 1

[6] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, pages 8535–8545, 2018. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016. 2

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 2, 3

[10] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019. 1

[11] 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. 1

[12] Shenwang Jiang, Jianan Li, Ying Wang, Bo Huang, Zhang Zhang, and Tingfa Xu. Delving into sample loss curve to embrace noisy and imbalanced data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7024–7032, 2022. 1

[13] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. 1

[14] Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9666–9676, 2022. 1

[15] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13896–13905, 2020. 1

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

[17] Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, and Jinwoo Shin. Robust inference via generative classifiers for handling noisy labels. In *International Conference on Machine Learning*, pages 3763–3772. PMLR, 2019. 1

[18] 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 *IJCAI*, pages 3902–3910, 2023. 1, 2

[19] Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020. 1

[20] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data, 2017. 2

[21] Yifan Li, Hu Han, Shiguang Shan, and Xilin Chen. Disc: Learning from noisy labels via dynamic instance-specific selection and correction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24070–24079, 2023. 1

[22] Zhuo Li, He Zhao, Zhen Li, Tongliang Liu, Dandan Guo, and Xiang Wan. Extracting clean and balanced subset for noisy long-tailed classification, 2024. 2

[23] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2016. 1

[24] Yang Lu, Yiliang Zhang, Bo Han, Yiu-Ming Cheung, and Hanzi Wang. Label-noise learning with intrinsically long-tailed data. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1369–1378, 2023. 1, 2

[25] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pages 6448–6458. PMLR, 2020. 1

[26] Yueming Lyu and Ivor W. Tsang. Curriculum loss: Robust learning and generalization against label corruption. In *International Conference on Learning Representations*, 2020. 1

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, 2019. 2

[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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2233–2241, 2017. 1

[29] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, pages 4175–4186, 2020. 1

[30] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pages 4334–4343. PMLR, 2018. 1

[31] Jiang-Xin Shi, Tong Wei, Yuke Xiang, and Yu-Feng Li. How re-sampling helps for long-tail learning? In *Advances in Neural Information Processing Systems*, pages 75669–75687, 2023. 1

[32] 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*, 34(11):8135–8153, 2023. 1

[33] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017. 2

[34] Yu-Feng Li Tong Wei, Jiang-Xin Shi and Min-Ling Zhang. Prototypical classifier for robust class-imbalanced learning. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2022. 1, 2

[35] Duc-Quang Vu, Trang T. T. Phung, Jia-Ching Wang, and Son T. Mai. Lcsl: Long-tailed classification via self-labeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11):12048–12058, 2024. 1

[36] Chaozheng Wang, Shuzheng Gao, Pengyun Wang, Cuiyun Gao, Wenjie Pei, Lujia Pan, and Zenglin Xu. Label-aware distribution calibration for long-tailed classification. *IEEE Transactions on Neural Networks and Learning Systems*, 35 (5):6963–6975, 2022. 1

[37] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021. 1

[38] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 322–330, 2019. 1

[39] Tong Wei, Jiang-Xin Shi, Wei-Wei Tu, and Yu-Feng Li. Robust long-tailed learning under label noise, 2021. 2

[40] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2691–2699, 2015. 2

[41] Shiyu Xuan and Shiliang Zhang. Decoupled contrastive learning for long-tailed recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6396–6403, 2024. 1

[42] 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5188–5197, 2021. 1

[43] Lingjie Yi, Jiachen Yao, Weimin Lyu, Haibin Ling, Raphael Douady, and Chao Chen. Representation learning for long tail recognition via feature space re-construction. In *International Conference on Learning Representations*, 2025. 1

[44] 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

[45] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 1

[46] Manyi Zhang, Xuyang Zhao, Jun Yao, Chun Yuan, and Weiran Huang. When noisy labels meet long tail dilemmas: A representation calibration method. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15844–15854, 2023. 1, 2

[47] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. In *Advances in Neural Information Processing Systems*, pages 34077–34090, 2022. 1

[48] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, 2018. 1