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

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

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

TR2026-074 June 04, 2026

Abstract

This supplementary material provides additional insights and experimental results that complement the main paper. In Section 2, pseudo codes of our Label Refurbishment with Label Raritity Algorithm are described in Algorithm 1. In Section 3, we present additional experimental results comparing our method with Representation CALibration (RCAL). Section 4 explores the noise resistance and bal- ancing capabilities of BANC and Symmetric Cross Entropy (SCE) losses based on the simulated CIFAR-100 dataset. To further clarify our design choices, Section 5 includes an ablation study on the influence of the scaling coefficient cc in Eq. (3) of the main paper. Additionally, 6 examines the impact of the hyperparameter a in Eq. (4) of the main paper. Finally, Section IV 7 details the Rarity Score gi from Eq. (6) of the main paper. These analyses provide a com- prehensive understanding of our proposed framework and its contributions to enhancing the effectiveness of long-tail noisy label learning.

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 Supplementary Material

Supplementary Material

1. Technical Appendices and Supplementary Material

This supplementary material provides additional insights and experimental results that complement the main paper. In §Section 2, pseudo codes of our Label Refurbishment with Label Raritity Algorithm are described in Algorithm 1. In §Section 3, we present additional experimental results comparing our method with Representation CALibration (RCAL). §Section 4 explores the noise resistance and balancing capabilities of BANC and Symmetric Cross Entropy (SCE) losses based on the simulated CIFAR-100 dataset. To further clarify our design choices, §Section 5 includes an ablation study on the influence of the scaling coefficient $\boldsymbol{c}$c in Eq. (3) of the main paper. Additionally, §6 examines the impact of the hyperparameter $\boldsymbol{\alpha}$ in Eq. (4) of the main paper. Finally, §Section IV 7 details the Rarity Score γ_i from Eq. (6) of the main paper. These analyses provide a comprehensive understanding of our proposed framework and its contributions to enhancing the effectiveness of long-tail noisy label learning.

2. Detailed Algorithm for Label Refurbishment with Label Rarity

Algorithm 1 shows the complete SoREL pseudo codes. The first stage of SoREL focuses on learning robust representations and producing reliable initial predictions. The contrastive feature learning generates embeddings resilient to label noise and class imbalance. In addition, the novel BAlanced Noise-tolerant Cross-entropy (BANC) loss is adopted to reduce residual label noise and encourage balanced learning. These components are then combined in Stage 2 to generate dependable initial predictions, forming a strong foundation for subsequent label refinement.

Algorithm 1 Label Refurbishment with Label Rarity

Input: Training data $\bar{\mathcal{D}} = \{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^{N}$, Backbone weight θ^b, Classifier weight θ^c, training model f, Stage 1: training epoch T_1, Stage 2: training epoch T_2;
Output: final model f^{final}
// 1) Pre-training with Contrastive Learning
for $t = 1, \ldots, T_1$ **do**
  $\mathcal{L}_{S1} = (1-\alpha)\mathcal{L}_{\text{con}}(\bar{\mathcal{D}}, f) + \alpha\mathcal{L}_{\text{BANC}}(\bar{\mathcal{D}}, f)$;
  Update weights:
  $\theta_t^b = \text{SGD}(L_E, \theta_{t-1}^b)$, $\theta_t^c = \text{SGD}(L_E, \theta_{t-1}^c)$;
end for
// 2) Robust Label Refurbishment
Loading the best θ_t^b, θ_t^c in f;
$l_i = f(\theta_t^b, \theta_t^c, \bar{\mathcal{D}})$;
// 2a) Logit Adjustment to get soft labels
if $l_i \neq \bar{y}_i$ **then**
  $\hat{\boldsymbol{y}}_i = \boldsymbol{s}_i(k)$; $\boldsymbol{s}_i(k) = p_i^k + w_i \bar{y}_i^k$;
else
  $\hat{\boldsymbol{y}}_i = \bar{y}_i$;
end if
// 2b) Multi-Experts Strategy
for $t = 1, \ldots, T_2$ **do**
  Load θ_t^b and θ_t^c into f;
  Define f_b: backbone; f_c^i: i-th classifier;
  $f_b = f(\theta_t^b)$, $f_c^1; f_c^2, f_c^3 = f(\theta_t^c)$;
  Compute ensemble loss:
  $L_E = L_{E1}(f_c^1(f_b, \bar{\mathcal{D}})) + L_{E2}(f_c^2(f_b, \bar{\mathcal{D}})) + L_{E3}(f_c^3(f_b, \bar{\mathcal{D}}))$;
  Update weights:
  $\theta_t^b = \text{SGD}(L_E, \theta_{t-1}^b)$;$\theta_t^c = \text{SGD}(L_E, \theta_{t-1}^c)$;
end for
return $f^{final}(\theta_t^b, \theta_t^c)$;

3. Comparison with RCAL

In addressing the simultaneous challenges of noisy labels and long-tailed distributions, Representation CALibration (RCAL) [1] stands out among current state-of-the-art methods as the only approach that does not adopt the strategy of segregating noisy and clean samples. Therefore, we discuss RCAL as a distinct case. Similar to our approach, RCAL employs contrastive learning to enhance feature representation in its initial stage. However, it diverges by incorporating representation calibration as a subsequent optimization step, while our method trains the classifier using the proposed BAlanced Noise-tolerant Cross (BANC) entropy loss. Additionally, our method integrates label refurbishment and multi-expert ensemble learning to further address noisy label and long-tailed distribution challenges.

Following the methodology in the RCAL paper, we simulated various noise rates and imbalance ratios on the CIFAR-10 and CIFAR-100 datasets, conducting experiments to compare our results with those reported by RCAL. As shown in Table 1, despite employing a similar feature learning strategy, our method consistently outperforms RCAL across

all scenarios. Notably, at an imbalance ratio of 100, the performance gap is significant, with improvements of approximately 15% on CIFAR-10 and 30% on CIFAR-100. These results underscore the effectiveness of our two-stage approach in addressing the dual challenges of noisy labels and long-tailed distributions.

4. Comparison between BAlanced Noise-tolerant Cross entropy (BANC) and Symmetric Cross Entropy (SCE)

As discussed in Sec. 3.1 of the main paper, our proposed BAlanced Noise-tolerant Cross (BANC) entropy is inspired by Symmetric Cross Entropy (SCE). We introduce a linear function, $c, (1 - \bar{y}_i^k)$, in the symmetric term and use the scaling coefficient c to adjust the loss penalty for mispredictions. This enhancement improves tolerance to noisy labels and enables the model to learn each class more balanced. This section compares the noise resistance and balancing capabilities of BANC and SCE losses through experiments on the simulated CIFAR-100 dataset. Table 2 summarizes their performance under various conditions, including imbalance ratios of 10 and 100, with symmetric noise rates of 0.4 and 0.6. We also evaluate their performance across three class sub-groups: many-shot, medium-shot, and few-shot classes. When the imbalance ratio is 10, the BANC loss consistently outperforms SCE across all noise rates. At an imbalance ratio of 100, the BANC loss shows significantly better performance than SCE, particularly in medium-shot and few-shot classes, resulting in higher overall accuracy. These results demonstrate that the BANC loss not only provides superior noise resistance but also effectively balances learning across different class distributions.

5. Influence of the Scaling Coefficient c in Eq. (3) of Main Paper

The scaling coefficient c in the BANC loss regulates the penalty for mispredictions, influencing BANC's effectiveness in handling noisy labels under class imbalance. To identify the optimal value of c, we conducted three experiments on the Food101 and Animal10 datasets, each with an imbalance ratio of 100. As shown in Fig. 1 and Fig. 2, the best performance is achieved when $c = 6$.

6. Influence of α in Eq. (4) of Main Paper

The hyperparameter α controls the contribution of contrastive loss $\mathcal{L}_{con}$ and the BANC loss $\mathcal{L}_{BANC}$. We conducted an experiment here to investigate the most suitable value of α on the simulated CIFAR100 dataset with symmetric noise rate 0.4 and imbalance ratio 100. Fig. 3 shows that the best performance is achieved when $\alpha = 0.2$.

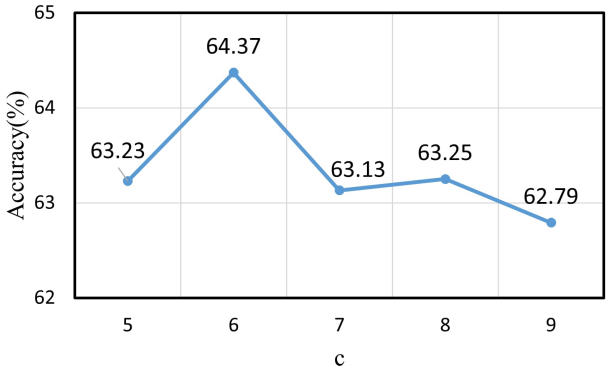

Figure 1. The impact of the scaling coefficient c on classification accuracy during our research on Food101. The optimal result is observed when $c = 6$. Please note that this result represents predictions from the first stage and does not incorporate label refurbishment and multi-expert ensemble learning.

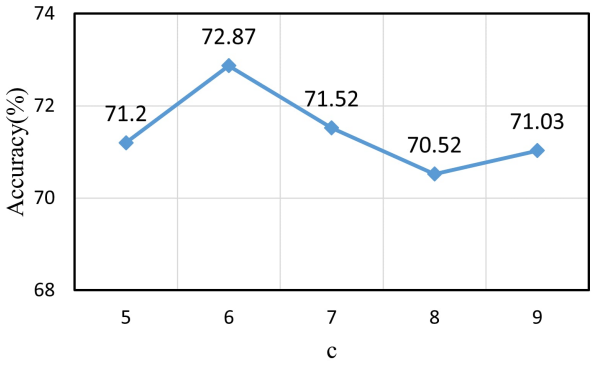

Figure 2. The impact of the scaling coefficient c on classification accuracy during our research on Animal10. The optimal result is observed when $c = 6$. Please note that this result represents predictions from the first stage and does not incorporate label refurbishment and multi-expert ensemble learning.

7. Illustration of γ_i in Eq. (6) of Main Paper

The rarity score γ_i is estimated as a function inversely proportional to the proportion of class $\bar{y}_i^k$ in the dataset. Let $n_{\bar{y}_i^k}$ denote the number of samples belonging to class $\bar{y}_i^k$, and let the proportion of class $\bar{y}_i^k$ be $h_i = \frac{n_{\bar{y}_i^k}}{N}$. We use a normal distribution function with zero mean to model γ_i: when h_i is close to zero, γ_i approaches the maximum value 1, and when h_i is greater than 0.5, γ_i quickly decays to zero. From Fig. 4, the variance σ of Eq. (6) in the main paper is estimated as 0.2.

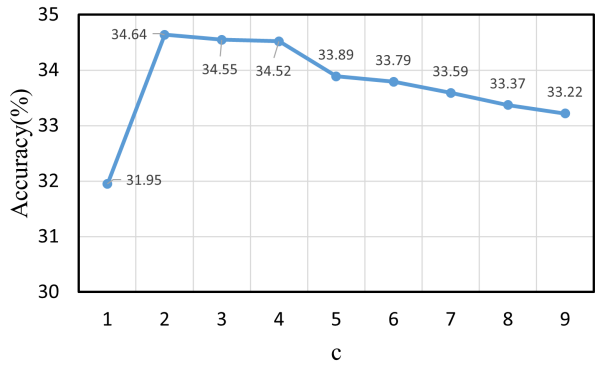

Figure 3. The effect of hyperparameter α on classification accuracy is investigated, and the optimal result is observed when $\alpha = 0.2$. Note that this outcome solely reflects predictions and does not encompass label refurbishment and multi-expert ensemble learning.

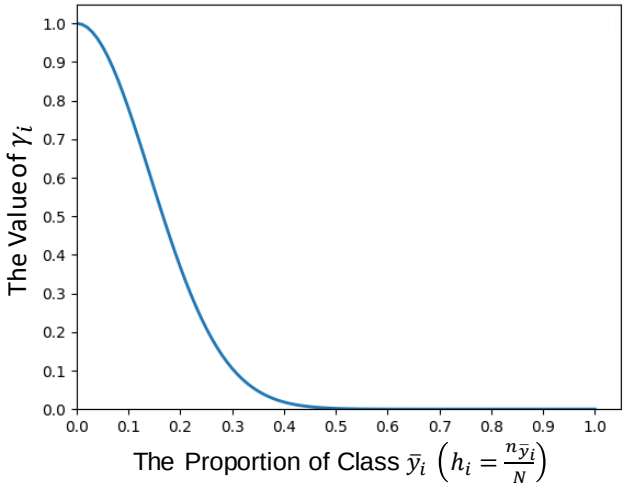

Figure 4. Plot of the rarity score γ_i, which is inversely proportional to the number of samples in class k.

References

[1] 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. 1

Table 1. Test accuracy (%) on simulated CIFAR-10 and CIFAR-100 with varying noise rates and imbalance ratios. Bold shows the highest score.

Dataset	Imbalance Ratio	10					100				
CIFAR-10	Noise Rate	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
	ERM	80.41	75.61	71.94	70.13	63.25	64.41	62.17	52.94	48.11	38.71
	ERM-DRW	81.72	77.61	71.94	70.13	63.25	66.74	62.17	52.94	48.11	38.71
	LDAM	84.59	82.37	77.48	71.41	60.30	71.46	66.26	58.34	46.64	36.66
	LDAM-DRW	85.94	83.73	80.20	74.87	67.93	76.58	72.28	66.68	57.51	43.23
	CRT	80.22	76.15	74.17	70.05	64.15	61.54	59.52	54.05	50.12	36.73
	NCM	82.33	74.73	74.76	68.43	64.82	68.09	66.25	60.91	55.47	42.61
	MiSLAS	87.58	85.21	83.39	76.16	72.46	75.62	71.48	67.90	62.04	54.54
	Co-teaching	80.30	78.54	68.71	57.10	46.77	55.58	50.29	38.01	30.75	22.85
	CDR	81.68	78.09	73.86	68.12	62.24	60.47	55.34	46.32	42.51	32.44
	Sel-CL+	86.47	85.11	84.41	80.35	77.27	72.31	71.02	65.70	61.37	56.21
	HAR-DRW	84.09	82.43	80.41	77.43	67.39	70.81	67.88	48.59	54.23	42.80
	RoLT	85.68	85.43	83.50	80.92	78.96	73.02	71.20	66.53	57.86	48.98
	RoLT-DRW	86.24	85.49	84.11	81.99	80.05	76.22	74.92	71.08	63.61	55.06
	RCAL	88.09	86.46	84.58	83.43	80.80	78.60	75.81	72.76	69.78	65.05
	Our	**94.51**	**93.06**	**92.45**	**91.28**	**89.43**	**91.20**	**88.86**	**87.12**	**86.00**	**85.88**
Dataset	Imbalance Ratio	10					100				
CIFAR-100	Noise Rate	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
	ERM	48.54	43.27	37.43	32.94	26.24	31.81	26.21	21.79	17.91	14.23
	ERM-DRW	50.38	45.24	39.02	34.78	28.50	34.49	28.67	23.84	19.47	14.76
	LDAM	51.77	48.14	43.27	36.66	29.62	34.77	29.70	25.04	19.72	14.19
	LDAM-DRW	54.01	50.44	45.11	39.35	32.24	37.24	32.27	27.55	21.22	15.21
	CRT	49.13	42.56	37.80	32.18	25.55	32.25	26.31	21.48	20.62	16.01
	NCM	50.76	45.15	41.31	35.41	29.34	34.89	29.45	24.74	21.84	16.77
	MiSLAS	57.72	53.67	50.04	46.05	40.63	41.02	37.40	32.84	26.95	21.84
	Co-teaching	45.61	41.33	36.14	32.08	25.33	30.55	25.67	22.01	16.20	13.45
	CDR	47.02	40.64	35.37	30.93	24.91	27.20	25.46	21.98	17.33	13.64
	Sel-CL+	55.68	53.52	50.92	47.57	44.86	37.45	36.79	35.09	31.96	28.59
	HAR-DRW	51.04	46.24	41.23	37.35	31.30	33.21	26.29	22.57	18.98	14.78
	RoLT	54.11	51.00	47.42	44.63	38.64	35.21	30.97	27.60	24.73	20.14
	RoLT-DRW	55.37	52.41	49.31	46.34	40.88	37.60	32.68	30.22	26.58	21.05
	RCAL	57.50	54.85	51.66	48.91	44.36	41.68	39.85	36.57	33.36	30.26
	Our	**77.90**	**75.32**	**74.66**	**72.54**	**70.53**	**74.05**	**71.47**	**71.37**	**69.03**	**66.31**

Table 2. Testing accuracy (%) on simulated CIFAR-100. Bold shows the highest score.

Imbalance Ratio	10							
Noise Rate (**Sym.**)	0.4				0.6			
Loss	Many	Medium	Few	All	Many	Medium	Few	All
SCE	79.88	62.33	**100.00**	73.64	78.32	51.54	**100.00**	69.75
BANC	**80.68**	**62.52**	**100.00**	**74.30**	**78.45**	**51.65**	**100.00**	**69.98**
Imbalance Ratio	100							
Noise Rate (**Sym.**)	0.4				0.6			
Loss	Many	Medium	Few	All	Many	Medium	Few	All
SCE	**91.82**	76.26	36.92	71.17	**92.87**	71.93	22.97	65.87
BANC	90.77	**76.38**	**41.28**	**72.03**	92.31	**72.74**	**23.84**	**66.64**