Supplementary Material for DaSC

Jae Soon Baik[1], In Young Yoon[1], Kun Hoon Kim[1], and Jun Won Choi[2*]

[1] Hanyang University, Korea
{jsbaik, inyoungyoon, khkim}@spa.hanyang.ac.kr,
[2] Seoul National University, Korea
junwchoi@snu.ac.kr

A Experimental Results with Higher Imbalance Ratio

In this section, we provide the performance of the proposed method compared to other baselines on long-tailed CIFAR and mini-ImageNet datasets with an imbalance ratio of 0.01 in Tabs. 1 and 2. The performances of other methods were taken from [8]. For methods not reported in [8], we reproduced their performances. For all experimental setups, the proposed DaSC consistently outperforms other baseline methods. Specifically, on CIFAR-10 with symmetric noise of 0.4, DaSC achieves 2.37% and 8.01% better performance over SFA and TABASCO, respectively. Similarly, for CIFAR-10 with asymmetric noise of 0.2, DaSC outperforms SFA and TABASCO by 4.93% and 10.58%, respectively. For CIFAR-10N with human annotations, DaSC achieves significant performance improvements of 6.97% and 6.07% over SFA and TABASCO, respectively. These results demonstrate the effectiveness and robustness of DaSC against extremely noisy labels and long-tailed distributions.

B Ablation Study on Hyperparameters

In Figs. 1 and 2, we present experimental results for the confidence threshold τ_c, temperature parameter for temperature scaling τ_t, and memory bank size $|M|$ on the long-tailed CIFAR-10 dataset with asymmetric noise of 0.2 and symmetric noise of 0.4, respectively. Additionally, Figs. 3 and 4 provide experimental results with other hyperparameters, including the temperature parameter τ_s for SBCL, the temperature parameter τ_m for MIDL, the coefficient λ_{SBCL} for SBCL, and the coefficient λ_{MIDL} for MIDL. These results show that DaSC maintains robust performance across various hyperparameter changes.

*Corresponding author

Table 1: Performance of the proposed method compared to baseline methods on the long-tailed version of CIFAR datasets with (a) symmetric noise and (b) asymmetric noise. The best results are shown in bold.

Dataset		CIFAR-10		CIFAR-100	
Imbalance Ratio		0.01			
Noise Ratio		0.4	0.6	0.4	0.6
Baseline	CE	47.81	28.04	21.99	15.51
LT	LA [9]	42.63	36.37	21.54	13.14
	LDAM [2]	45.52	35.29	18.81	12.65
	IB [10]	49.07	32.54	20.34	12.10
	BBN [13]	45.22	31.63	17.31	12.83
NL	DivideMix [7]	32.42	34.73	36.20	26.29
	UNICON [5]	61.23	54.69	32.09	24.82
	TCL [4]	56.13	45.88	33.29	24.39
NL-LT	MW-Net [11]	46.62	39.33	19.65	13.72
	RoLT [12]	60.11	44.23	23.51	16.61
	HAR [1]	51.54	38.28	20.21	14.89
	ULC [3]	45.22	50.56	33.41	25.69
	SFA [6]	67.98	54.70	37.69	30.02
	TABASCO [8]	62.34	55.76	36.91	26.25
	DaSC	**70.35**	**58.49**	**41.12**	**33.65**

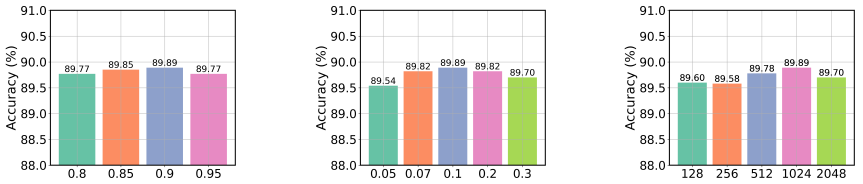
(a) Symmetric noise

Dataset		CIFAR-10		CIFAR-100	
Imbalance Ratio		0.01			
Noise Ratio		0.2	0.4	0.2	0.4
Baseline	CE	56.56	44.64	25.35	17.89
LT	LA [9]	58.78	43.37	32.16	22.67
	LDAM [2]	61.25	40.85	29.22	18.65
	IB [10]	56.28	42.96	31.15	23.40
	BBN [13]	54.51	51.15	25.19	21.68
NL	DivideMix [7]	41.12	42.79	38.46	29.69
	UNICON [5]	53.53	34.05	34.14	30.72
	TCL [4]	60.58	49.66	40.18	30.54
NL-LT	MW-Net [11]	62.19	45.21	27.56	20.40
	RoLT [12]	54.81	50.26	32.96	-
	HAR [1]	62.42	51.97	27.90	20.03
	ULC [3]	41.14	22.73	34.07	25.04
	SFA [6]	68.63	52.16	41.89	33.33
	TABASCO [8]	62.98	54.04	40.35	33.15
	DaSC	**73.56**	**58.45**	**43.52**	**35.12**

(b) Asymmetric noise

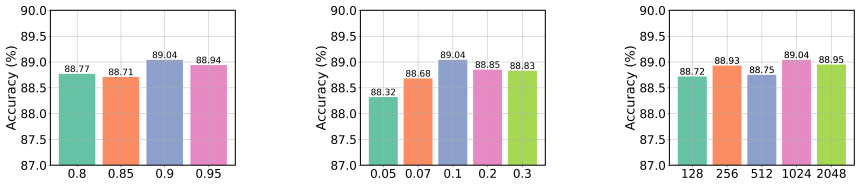
Table 2: Performance of the proposed method compared to baseline methods on the long-tailed version of CIFAR with human annotations and Red mini-ImageNet dataset. The best results are shown in bold.

Dataset		Red		10N	100N
Imbalance Ratio		≈ 0.01		0.01	
Noise Ratio		0.2	0.4	≈ 0.4	
Baseline	CE	30.88	31.46	49.31	25.28
LT	LA [9]	10.32	9.56	50.09	26.39
	LDAM [2]	14.30	15.64	50.36	30.17
	IB [10]	16.72	16.34	56.41	31.55
	BBN [13]	30.92	30.30	52.98	25.06
NL	DivideMix [7]	33.00	34.72	30.67	31.34
	UNICON [5]	31.86	31.12	59.47	37.06
	TCL [4]	37.24	35.70	61.70	39.56
NL-LT	MW-Net [11]	30.74	31.12	54.95	31.80
	RoLT [12]	15.78	16.90	61.23	33.48
	HAR [1]	32.60	31.30	56.84	32.34
	ULC [3]	34.24	34.84	43.89	35.71
	SFA [6]	36.70	35.52	63.64	40.83
	TABASCO [8]	37.20	37.12	64.54	39.30
	DaSC	**40.26**	**39.72**	**70.61**	**44.59**



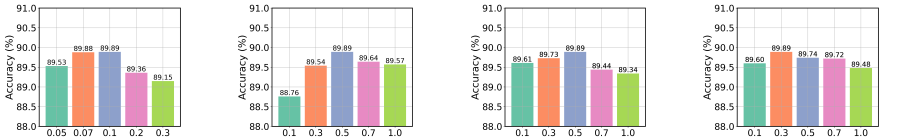
(a) Confidence threshold τ_c **(b)** Temperature parameter τ_T **(c)** Memory bank size $|M|$

Fig. 1: Performance comparison with various hyperparameter configurations. The performance was evaluated on long-tailed CIFAR-10 with asymmetric noise of 0.2.



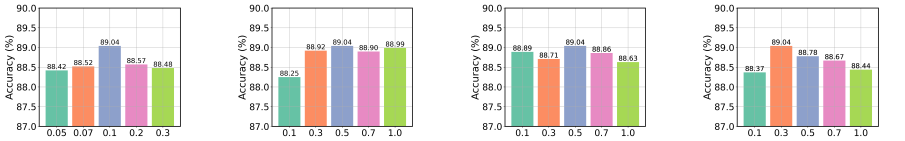
(a) Confidence threshold τ_c **(b)** Temperature parameter τ_T **(c)** Memory bank size $|M|$

Fig. 2: Performance comparison with various hyperparameter configurations. The performance was evaluated on long-tailed CIFAR-10 with symmetric noise of 0.4.



(a) SBCL Temp. τ_s **(b)** SBCL Coef. λ_{SBCL} **(c)** MIDL Temp. τ_m **(d)** MIDL Coef. λ_{MIDL}

Fig. 3: Performance comparison with various hyperparameter configurations. The performance was evaluated on long-tailed CIFAR-10 with asymmetric noise of 0.2.



(a) SBCL Temp. τ_s **(b)** SBCL Coef. λ_{SBCL} **(c)** MIDL Temp. τ_m **(d)** MIDL Coef. λ_{MIDL}

Fig. 4: Performance comparison with various hyperparameter configurations. The performance was evaluated on long-tailed CIFAR-10 with symmetric noise of 0.4.

C Performance of Different Representations and Model Prediction for Noisy Sample Selection.

Table 3 presents the performance of DaSC using different representations and model predictions for detecting correctly labeled samples. For representations, we use either $f(x(i))$ from the backbone network or $z'(i)$ from the MLP projector. For model predictions in DaCC, we employ $\hat{p}^c(i)$ from the conventional classifier or $\hat{p}^b(i)$ from the balanced classifier.

The results show that the DaSC using $z'(i)$ and $\hat{p}^c(i)$ outperforms all other setups. Using the low-dimensional representation from the MLP projector yields better performance than using the representation from the backbone network. Additionally, predictions from the conventional classifier are more effective than those from the balanced classifier. This is likely due to the early training instability of the balanced classifier, which is trained using an estimate of the data distribution at each epoch, impacting the accurate identification of correctly labeled samples.

Table 3: Performance of DaSC with different representation and model prediction strategies for class centroid estimation.

Representation	Model Prediction	CIFAR-10		CIFAR-100	
		Sym.	Asym.	Sym.	Asym.
		0.4	0.2	0.4	0.2
$f(x(i))$	$\hat{p}^c(i)$	88.90	89.10	61.16	62.85
$z'(i)$	$\hat{p}^b(i)$	88.84	89.57	61.26	62.26
$z'(i)$	$\hat{p}^c(i)$	**89.04**	**89.89**	**61.85**	**63.22**

D Ablation Study on Subset $\mathcal{D}^I$

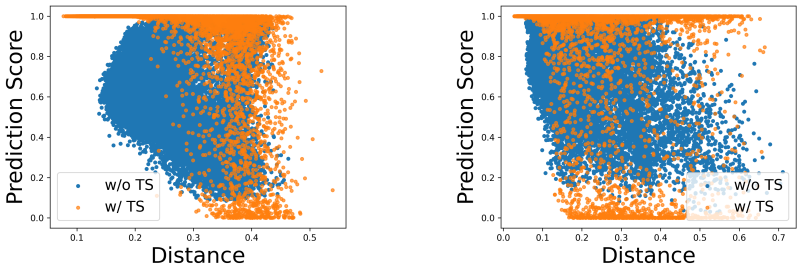
DaSC leverages samples from a specific subset $\mathcal{D}^I$ rather than directly from the training dataset $\mathcal{D}$. This strategy enhances performance by leveraging a variety of classes to estimate class centroid while filtering out unreliable samples due to noisy labels and long-tailed distributions. As shown in Tab. 4, using samples from the subset $\mathcal{D}^I$ achieves better performance than using them directly from the training dataset $\mathcal{D}$.

Table 4: Performance comparison of using $\mathcal{D}$ versus $\mathcal{D}^I$ used in DaCC.

Sample Set	CIFAR-10		CIFAR-100	
	Sym.	Asym.	Sym.	Asym.
	0.4	0.2	0.4	0.2
$\mathcal{D}$	87.89	89.31	61.26	62.56
$\mathcal{D}^I$	**89.04**	**89.89**	**61.85**	**63.22**

E Effect of Temperature Scaling

The proposed DaSC employs temperature scaling to mitigate the inherent bias in model predictions. To explore its impact, Fig. 5 presents the prediction score of each sample relative to the distance from the closest class centroid. These results indicate that temperature scaling assigns higher weights to reliable samples closer to the centroid (*i.e.*, higher prediction scores), highlighting their importance, while giving lower weights to unreliable samples farther from the centroids.



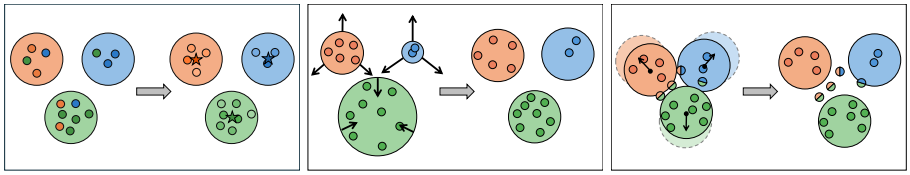
(a) Sym. 0.4 **(b)** Asym. 0.2

Fig. 5: Prediction scores of each sample versus the distance to the closest class centroid. 'TS' denotes the temperature scaling.

F Intuitive explanations of main components

Our method achieves the following objectives for improving sample selection: 1) accurately obtaining the class probability for all class samples, 2) enhancing the representation skewed by class imbalance, and 3) improving the discrimination among low-confidence samples. We devise three methods, DaCC, SBCL, and MIDL, to achieve these goals. Fig. 6 visually explains how each main component works.

Fig. 6a illustrates that DaCC uses samples from all classes based on predictions rather than their noisy label to estimate class centroids, allowing the effective use of samples from all classes. DaCC weights samples according to predicted scores for more effective use of reliable samples. Fig. 6b demonstrates that SBCL balances skewed representation using reliable label information from high-confidence samples. Fig. 6c illustrates that MIDL enhances class discrimination by using diverse negative keys with mixup samples in a self-supervised manner.



(a) DaCC **(b)** SBCL **(c)** MIDL

Fig. 6: Intuitive illustration of main components in DaSC: (a) DaCC, (b) SBCL, and (c) MIDL.

References

1. Cao, K., Chen, Y., Lu, J., Aréchiga, N., Gaidon, A., Ma, T.: Heteroskedastic and imbalanced deep learning with adaptive regularization. In: ICLR (2021)
2. Cao, K., Wei, C., Gaidon, A., Aréchiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: NeurIPS. pp. 1565–1576 (2019)
3. Huang, Y., Bai, B., Zhao, S., Bai, K., Wang, F.: Uncertainty-aware learning against label noise on imbalanced datasets. In: AAAI. pp. 6960–6969 (2022)
4. Huang, Z., Zhang, J., Shan, H.: Twin contrastive learning with noisy labels. In: CVPR. pp. 11661–11670 (2023)
5. Karim, N., Rizve, M.N., Rahnavard, N., Mian, A., Shah, M.: Unicon: Combating label noise through uniform selection and contrastive learning. In: CVPR. pp. 9676–9686 (2022)
6. Li, H.T., Wei, T., Yang, H., Hu, K., Peng, C., Sun, L.B., Cai, X.L., Zhang, M.L.: Stochastic feature averaging for learning with long-tailed noisy labels. In: IJCAI (2023)
7. Li, J., Socher, R., Hoi, S.C.H.: Dividemix: Learning with noisy labels as semi-supervised learning. In: ICLR (2020)
8. Lu, Y., Zhang, Y., Han, B., Cheung, Y.m., Wang, H.: Label-noise learning with intrinsically long-tailed data. In: ICCV. pp. 1369–1378 (2023)
9. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. In: ICLR (2021)
10. Park, S., Lim, J., Jeon, Y., Choi, J.Y.: Influence-balanced loss for imbalanced visual classification. In: ICCV. pp. 735–744 (2021)
11. Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., Meng, D.: Meta-weight-net: Learning an explicit mapping for sample weighting. In: NeurIPS. pp. 1917–1928 (2019)
12. Wei, T., Shi, J.X., Tu, W.W., Li, Y.F.: Robust long-tailed learning under label noise. arXiv preprint arXiv:2108.11569 (2021)
13. Zhou, B., Cui, Q., Wei, X.S., Chen, Z.M.: Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In: CVPR. pp. 9719–9728 (2020)