Long-Tailed Out-of-Distribution Detection with Refined Separate Class Learning

Supplementary Material

A. More Discussion about Dynamic Class-wise Temperature Adjustment

A.1. Class Distributions

To better illustrate class distribution patterns, we select CIFAR10 (10 classes) and CIFAR100 (100 classes) [20] as representative examples. This choice is driven by ImageNet-1k's [6] extensive class count (1,000), which complicates direct visualization of distribution characteristics—whereas the more manageable class scales of CIFAR10 and CIFAR100 enable clear and intuitive demonstration of key patterns.

For the CIFAR10 training set with an imbalance ratio of $\rho = 100$, the number of samples per in-distribution (ID) class is represented as $\xi = (n_1, n_2, \ldots, n_{10})$ = (5000, 2997, 1796, 1077, 645, 387, 232, 139, 83, 50). Applying ℓ_2 normalization to ξ yields the normalized class distribution $\hat{\xi} = (\hat{n}_1, \hat{n}_2, \ldots, \hat{n}_{10})$ = (0.8005, 0.4798, 0.2875, 0.1724, 0.1033, 0.0620, 0.0371, 0.0223, 0.0133, 0.0080).

For the CIFAR100 training set with an imbalance ratio of $\rho = 100$, the number of samples per in-distribution (ID) class is represented as $\xi = (n_1, n_2, \ldots, n_{100})$ = (500, 477, 455, 434, 415, 396, 378, 361, 344, 328, 314, 299, 286, 273, 260, 248, 237, 226, 216, 206, 197, 188, 179, 171, 163, 156, 149, 142, 135, 129, 123, 118, 112, 107, 102, 98, 93, 89, 85, 81, 77, 74, 70, 67, 64, 61, 58, 56, 53, 51, 48, 46, 44, 42, 40, 38, 36, 35, 33, 32, 30, 29, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 15, 14, 13, 13, 12, 12, 11, 11, 10, 10, 9, 9, 8, 8, 7, 7, 7, 6, 6, 6, 6, 5, 5, 5, 5). After applying ℓ_2 normalization to ξ, the normalized class distribution $\hat{\xi} = (\hat{n}_1, \hat{n}_2, \ldots, \hat{n}_{100})$ = (0.2986, 0.2849, 0.2717, 0.2592, 0.2478, 0.2365, 0.2257, 0.2156, 0.2054, 0.1959, 0.1875, 0.1786, 0.1708, 0.1630, 0.1553, 0.1481, 0.1415, 0.1350, 0.1290, 0.1230, 0.1176, 0.1123, 0.1069, 0.1021, 0.0973, 0.0932, 0.0890, 0.0848, 0.0806, 0.0770, 0.0735, 0.0705, 0.0669, 0.0639, 0.0609, 0.0585, 0.0555, 0.0532, 0.0508, 0.0484, 0.0460, 0.0442, 0.0418, 0.0400, 0.0382, 0.0364, 0.0346, 0.0334, 0.0317, 0.0305, 0.0287, 0.0275, 0.0263, 0.0251, 0.0239, 0.0227, 0.0215, 0.0209, 0.0197, 0.0191, 0.0179, 0.0173, 0.0161, 0.0155, 0.0149, 0.0143, 0.0137, 0.0131, 0.0125, 0.0119, 0.0113, 0.0107, 0.0102, 0.0096, 0.0090, 0.0090, 0.0084, 0.0078, 0.0078, 0.0072, 0.0072, 0.0066, 0.0066, 0.0060, 0.0060, 0.0054, 0.0054, 0.0048, 0.0048, 0.0042, 0.0042, 0.0042, 0.0036, 0.0036, 0.0036, 0.0036, 0.0030, 0.0030, 0.0030, 0.0030).

Table 6. Impact of linear vs. square root functions in dynamic class-wise temperature adjustment on OOD detection performance and ID classification accuracy.

$\mathcal{D}_{\text{in}}^{\text{train}}$	Modulation function	AUROC↑	AUPR↑	FPR95↓	ACC↑
CIFAR10-LT	Linear Function	92.15	91.71	32.23	76.77
	Square Root Function	**93.46**	**93.28**	**27.73**	**79.14**
CIFAR100-LT	Linear Function	75.05	68.66	64.53	41.72
	Square Root Function	**76.11**	**70.76**	**62.86**	**42.81**
ImageNet-LT	Linear Function	78.62	77.95	74.28	45.80
	Square Root Function	**78.90**	**78.10**	**73.04**	**46.19**

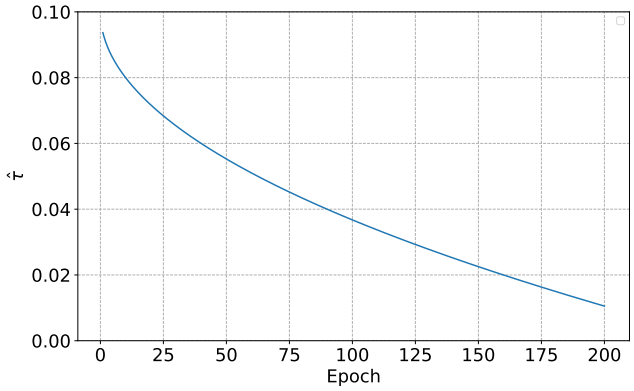
A.2. The design of Dynamic Class-wise Temperature Adjustment

Incorporating Training Epochs and Class Sample Sizes. In long-tailed scenarios, head classes (with abundant samples) and tail classes (with fewer samples) exhibit differing learning dynamics. To address this, our adjustment mechanism considers both the current training epoch (e) and the total training epochs (E), ensuring that temperature modulation evolves as training progresses. Additionally, the normalized class sample size, denoted as $\hat{n}_{c(\boldsymbol{x})}$, allows the temperature to adapt based on the relative abundance of each class.

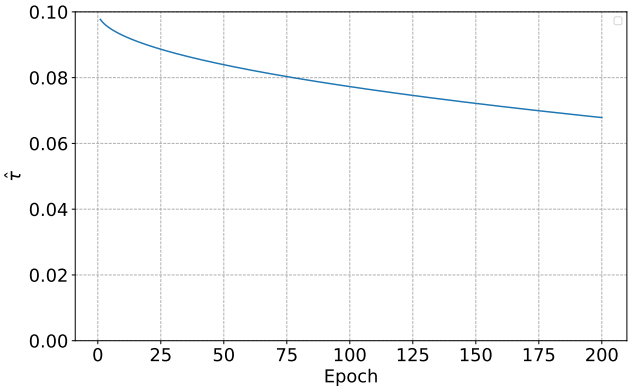
Rationale for the Square Root Function. The inclusion of the square root function serves to moderate the influence of class sample size on temperature adjustment. In long-tailed distributions, the disparity between head and tail class sample sizes can be substantial. A linear adjustment might overemphasize this difference, leading to overly aggressive temperature scaling for tail classes. By applying the square root, we achieve a more balanced modulation, ensuring that temperature adjustments are neither too drastic for tail classes nor too minimal for head classes. As illustrated in Tab. 6, compared with linear functions ($\hat{\tau}_{c(\boldsymbol{x})} = \tau \cdot \left[1 - \frac{e}{E} \cdot \hat{n}_{c(\boldsymbol{x})}\right]$), using the square root function (in Eq. (4))) is more beneficial for OOD detection and ID classification. In practice, this approach aligns with practices in statistical modeling where square root transformations are employed to stabilize variance and normalize distributions [1, 21, 44].

A.3. The variation of the temperature for head and tail class

We examine the variation in temperature on the CIFAR10-LT benchmark, where the percentage (k) of tail classes is set to 0.6. The head classes encompass n_1, n_2, n_3, n_4, while the tail classes consist of $n_5, n_6, n_7, n_8, n_9, n_{10}$. To comprehensively understand the temperature adjustment strategy during training, we plot the variations of the tempera-



(a) Temperature variation of head class n_1



(b) Temperature variation of tail class n_5

Figure 5. The variations of the temperature value $\hat{\tau}$ with respect to head and tail class in CIFAR10-LT. (a) illustrates the variation of the temperature value $\hat{\tau}$ of the head class n_1. (b) shows the variation of temperature value $\hat{\tau}$ of the tail class n_5.

ture value $\hat{\tau}$, defined in Eq. (4), for the head class n_1 and tail class n_5. The default values of $\tau = 0.1$ and $E = 200$ are utilized for this analysis. As depicted in Fig. 5a, concerning the head class n_1, the temperature value $\hat{\tau}$ gradually diminishes from 0.1 to 0.01 during training. The reduction in $\hat{\tau}$ can enhance the pushing effect between the OOD samples and the head classes, thereby amplifying the disparity between the head class and OOD samples.

Similarly, as illustrated in Fig. 5b, the temperature value $\hat{\tau}$ for tail class n_5 progressively decreases. Smaller $\hat{\tau}$ reinforce the pulling effect within each tail class, making tail classes and OOD samples more discernible. Notable is the less pronounced decrease in the magnitude of $\hat{\tau}$ for the tail class n_5 compared to the head class n_1. Given the scarcity of tail class samples in the training data, an excessively low $\hat{\tau}$ may not effectively capture the nuanced differences between tail classes, which could have a negative impact on the accuracy of the classification.

B. The Informative Outlier Mining Algorithm

Algorithm 1 outlines the informative outlier mining process. In each training epoch, for a batch of ID samples of size $\mathcal{B}$, we randomly sample $3 \times \mathcal{B}$ outliers from the auxiliary OOD training set. Each outlier is then assigned a score using the outlier scoring function defined in Eq. (5), which reflects its tendency to resemble head, tail, or neither class types. We then sort these outliers in descending order by score to group them into three categories:

- *Tail-class-like OOD*: This category refers to outliers that exhibit characteristics resembling tail-class samples. If an outlier belongs to this category, it indicates a higher probability of being incorrectly predicted as tail class.
- *Head-class-like OOD*: This category pertains to outliers that possess attributes similar to head-class samples. If an outlier falls into this category, it suggests an ease of being predicted as head class.
- *Neutral OOD*: Outliers in this category lack distinctive features aligning with either head or tail classes, resulting in no clear prediction bias. Their ambiguous characteristics make them particularly challenging for the model to distinguish.

By employing informative outlier mining, we aim to identify outliers that provide valuable information for long-tailed OOD detection. This approach enables the selection of outliers that possess specific characteristics related to either tail or head classes, facilitating the improvement of OOD detection performance under long-tailed settings.

C. More Experiment Settings

C.1. Datasets

In-distribution training and test sets ($\mathcal{D}_{\text{in}}^{\text{train}}$, $\mathcal{D}_{\text{in}}^{\text{test}}$) We use three popular long-tailed image classification datasets as $\mathcal{D}_{\text{in}}^{\text{train}}$, including CIFAR10-LT, CIFAR100-LT [2], and ImageNet-LT [28]. The original version of CIFAR10 and CIFAR100 contains 50,000 training images and 10,000 validation images of size 32×32 with 10 and 100 classes, respectively. CIFAR10-LT and CIFAR100-LT are the imbalanced version of them, which reduce the number of training examples per class and keep the validation set unchanged. The imbalance ratio ρ denotes the ratio between sample sizes of the most frequent class and least frequent class. ImageNet-LT is a large-scale dataset in long-tail recognition, which truncates the balanced version ImageNet [6]. ImageNet-LT has 1,000 classes, which contain 115,846 training images with the number of per-class training data ranging from 5 to 1,280, and 20,000 validation images with a balanced class size.

OOD training set ($\mathcal{D}_{\text{out}}^{\text{train}}$) TinyImages80M [43] contains 80 million images with a size of 32×32. We use a sub-

Table 7. Ablation on the percentage (k) of tail classes. Experiments are conducted on CIFAR100-LT using ResNet-18. The first row ($k = 100\%$) means that $\mathcal{L}_{\text{head}}$ is not used and $\mathcal{L}_{\text{tail}}$ is applied to all ID training samples. The last row ($k = 0\%$) denotes the exclusion of $\mathcal{L}_{\text{tail}}$, with $\mathcal{L}_{\text{head}}$ being applied to all ID training samples. All values are percentages averaged over six OOD test datasets. Bold numbers are the best results.

k	AUROC ↑	AUPR ↑	FPR95 ↓	ACC ↑
100%	74.87	60.09	63.63	40.25
70%	75.33	69.62	64.16	**43.23**
60%	**76.11**	**70.76**	**62.86**	42.81
50%	75.64	70.48	63.82	42.10
40%	75.25	70.39	64.75	41.08
0%	75.06	69.72	64.57	43.02

Table 8. Ablation on different model structures trained on CIFAR100-LT. All values are percentages averaged over six OOD test datasets. Bold numbers are the best results.

Model	Method	AUROC↑	AUPR↑	FPR95↓	ACC↑
ResNet-34	OE [14]	73.11	67.55	68.15	39.00
	PASCL [47]	73.66	67.99	67.15	39.54
	EAT [50]	73.81	68.94	67.22	**43.26**
	COCL [31]	74.40	**70.57**	68.17	42.06
	TSCL [10]	74.57	68.15	64.59	40.18
	RSCL(Ours)	**75.05**	69.61	**64.10**	42.54
ResNet-50	OE [14]	72.70	67.62	69.77	37.86
	PASCL [47]	73.07	68.10	69.63	37.74
	EAT [50]	73.04	68.56	68.98	41.21
	COCL [31]	74.55	70.76	69.73	40.70
	TSCL [10]	73.73	68.14	66.27	38.05
	RSCL(Ours)	**75.59**	**70.96**	**64.11**	**42.02**

set of random 300K images from TinyImages80M as $\mathcal{D}_{\text{out}}^{\text{train}}$ for CIFAR10-LT and CIFAR100-LT. For ImageNet-LT, we use ImageNet-Extra as $\mathcal{D}_{\text{out}}^{\text{train}}$ following [47]. ImageNet-Extra contains 517,711 images belonging to 500 classes from ImageNet-22k [6], but having not overlapping with the 1,000 in-distribution classes in ImageNet-LT.

OOD test set ($\mathcal{D}_{\text{out}}^{\text{test}}$) We use SC-OOD benchmark [51] as true OOD data for CIFAR10-LT and CIFAR100-LT following [10, 31, 47, 50]. The SC-OOD benchmark contains six datasets: Texture [4], SVHN [34], CIFAR [20], Tiny ImageNet [22], LSUN [53], and Places365 [57]. For ImageNet-LT, we use ImageNet-1k-OOD constructed by [47] as $\mathcal{D}_{\text{out}}^{\text{test}}$. ImageNet-1k-OOD has 50,000 OOD test images from 1,000 classes randomly selected from ImageNet-22k [6] (with 50 images in each class), which is of the same size as the in-distribution test set. The 1,000 classes in ImageNet-1k-OOD are not ovelapped with either the 1,000 in-distribution classes in ImageNet-LT or the 500 OOD training classes in ImageNet-Extra.

C.2. Implementation Details

In this section, we provide more implementation details about our approach RSCL. For experiments on CIFAR10-LT and CIFAR100-LT, we train the ResNet18 model for 200 epochs using Adam optimizer with initial learning rate 1×10^{-3} and batch size 256. The learning rate is decayed to 0 using a cosine annealing learning rate scheduler. For experiments on ImageNet-LT, we follow the settings in [31, 47, 50] and use ResNet50 as backbone. We train the model for 100 epochs using SGD optimizer with initial learning rate 0.1 and batch size 60. We decay the learning rate by a factor of 10 at epoch 60 and 80. On all datasets, we set $\tau = 0.1$ following [10, 47]. Empirical analysis leads us to configure the loss weight hyperparameters as follows: the percentage of tail classes $k = 0.6$, loss weight parameter $\alpha = 0.05$, $\beta = 0.05$, $\gamma = 0.1$.

For other hyper-parameters in the compared methods, we use the values suggested in the original papers. Significantly, the implementation of the original PASCL [47] and EAT [50] approaches incorporated fine-tuning techniques. However, as our RSCL approach does not involve fine-tuning technique, to ensure a fair comparison, we re-run the PASCL and EAT without using the fine-tuning techniques. Furthermore, the original COCL approach [31] integrated a calibration technique at inference time, and the FPR95 evaluation metric used by COCL is different from PASCL, EAT, and TSCL. Therefore, we re-run the COCL approach utilizing our evaluation metrics and refrain from using the calibration technique at inference time. In particular, even when compared with the original COCL and EAT method, our RSCL still achieves superior OOD detection performance. All experiments in this paper are conducted on a single NVIDIA A100 GPU.

D. More Ablation Study

Abaltion on the percentage (k) of tail classes. The percentage (k) of tail classes emerges as a crucial hyper-parameter governing the delineating between head and tail classes. Ablation analyses concerning k are consolidated in Tab. 7. Optimal results materialize when k is calibrated to approximately 60%, a setting that aligns with the default value employed in our experimental setup. Notably, the results observed at $k = 60\%$ markedly outperform those at $k = 100\%$ (without $\mathcal{L}_{\text{head}}$) and $k = 0\%$ (without $\mathcal{L}_{\text{tail}}$), emphasizing the importance of our proposed approach in enhancing model performance. Furthermore, our approach exhibits notable robustness across a broad spectrum of k values (e,g, $k \in [50\%, 70\%]$).

Robustness under different model structures. To evaluate the effectiveness of our approach across different model structures, we conduct ablation experiments on ResNet-34 and ResNet-50, with the results presented in Tab. 8. Noteworthy is the consistent outperformance of our approach over OE [14], PASCL [47], EAT [50], COCL [31], and TSCL [10] when applied to ResNet-50, as evidenced by superior OOD detection performance and ID

Table 9. Average FPR95 on CIFAR100-LT using ResNet18 depending on hyperparmeter α, β and γ. All values are percentages averaged over six OOD test datasets. Bold numbers are the best results.

Average FPR95						
α	β					
	0.01		0.05		0.1	
	γ					
	0.05	0.1	0.05	0.1	0.05	0.1
0.01	66.05	65.48	65.45	65.38	65.61	64.69
0.05	64.45	64.94	64.21	**62.86**	64.65	63.48
0.1	63.69	63.61	63.91	63.69	63.43	63.47

Table 10. Accuracy on CIFAR100-LT using ResNet18 depending on hyperparmeter α, β and γ. All values are percentages averaged over six OOD test datasets. Bold numbers are the best results.

ACC						
α	β					
	0.01		0.05		0.1	
	γ					
	0.05	0.1	0.05	0.1	0.05	0.1
0.01	42.48	**43.60**	42.34	43.26	41.72	42.82
0.05	41.63	42.83	41.72	42.81	41.34	42.71
0.1	41.41	42.33	41.01	42.53	40.78	42.30

classification accuracy. For instance, on ResNet-50, our approach achieves an improvement in average AUROC by **1.04**%, a reduction in average FPR95 by **5.62**%, an enhancement in ID classification accuracy by **1.32**% compared to COCL. Similarly, promising trends are observed on ResNet-34. These findings underscore the robust performance of our approach, independent of the specific model structures used.

Ablation on loss weight α, β and γ Tab. 9 and Tab. 10 show the impact of different loss weights on both OOD detection performance and ID classification accuracy. The hyperparameter α is set to strike a balance between learning from ID classes and outlier classes. A higher value assigned to α indicates a model preference towards outliers. Elevating α leads to enhanced OOD detection performance but may compromise ID classification accuracy. Hence, selecting an appropriate α becomes crucial. Furthermore, β and γ are designated to balance the differentiation between OOD samples and head/tail samples, with a slightly higher value for γ compared to β often proving advantageous for both OOD detection and ID classification. Finally, a configuration where α = 0.05, accompanied by corresponding values of β = 0.05 and γ = 0.1, is generally recommended and serves as the default setting in our experiment settings.

Robustness under different imbalance ratios (ρ). We have showcased the excellent performance of RSCL on ρ = 100 in Section 5.2. Furthermore, RSCL exhibits robust performance across various imbalance ratios such as ρ = 50 and ρ = 20. The results in Tab. 11 underscore that RSCL consistently outperforms OE and PASCL by a considerable margin in terms of OOD detection performance and ID classification accuracy. Moreover, RSCL demonstrates superior performance compared to recent methods such as TSCL, EAT, and COCL.

Table 11. Robustness under different imbalance ratios (ρ). All values are percentages averaged over six OOD test datasets. Bold numbers are the best results.

Imbalance Ratio	Method	**AUROC**↑	**AUPR**↑	**FPR95**↓	**ACC**↑
$\rho = 50$	OE	74.07	68.38	66.68	43.47
	PASCL	74.35	68.51	66.49	44.00
	EAT	75.45	70.02	63.62	46.17
	COCL	76.07	71.42	65.49	46.61
	TSCL	75.59	69.34	62.95	44.22
	RSCL (Ours)	**76.50**	**70.79**	**61.56**	**46.64**
$\rho = 20$	OE	76.51	70.57	63.06	50.98
	PASCL	77.34	70.80	61.81	50.13
	EAT	77.43	72.47	62.14	53.76
	COCL	78.52	**73.27**	59.94	53.67
	TSCL	77.64	71.46	60.11	50.69
	RSCL (Ours)	**78.71**	73.19	**58.12**	**53.80**