南京航空航天大學
Nanjing University of Aeronautics and Astronautics

# LTRL: Boosting Long-tail Recognition via Reflective Learning

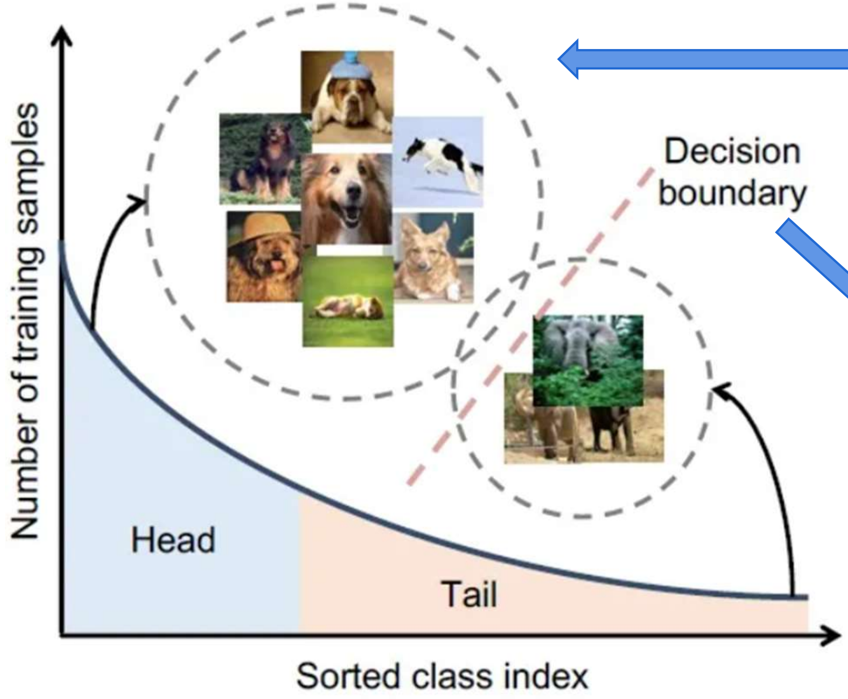
Qihao Zhao[1,2,⋆], Yalun Dai[3,⋆],
Shen Lin[4], Wei Hu[1], Fan Zhang[1,⋆⋆], and Jun Liu[2,5]

Beijing University of Chemical Technology, China
Singapore University of Technology and Design, Singapore
Nanyang Technological University, Singapore
Xidian University, China
Lancaster University, UK

ECCV 2024

# Introduction

## Long-tail Problems




feature space learned on these sampled is often larger than tail classes.

the decision boundary is usually biased towards dominant classes

The label distribution of a long-tailed dataset

# Introduction

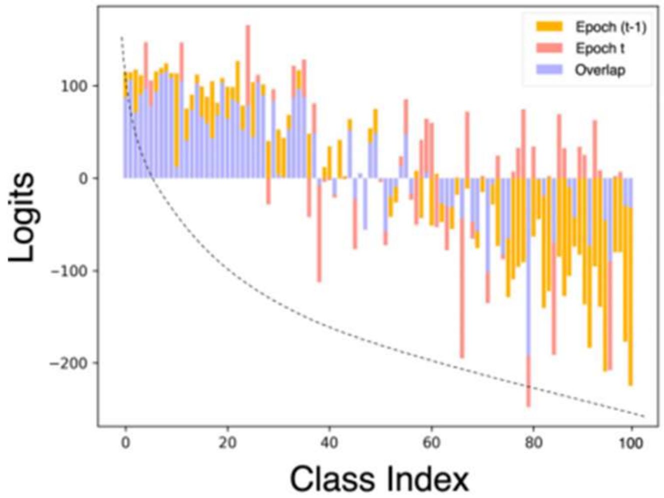
## Introducing Reflective Learning (RL)：

In the human classroom,top students habitually **review** studied knowledge post-class, **summarize** the connection between knowledge, and **correct** misconceptions after review summarize.

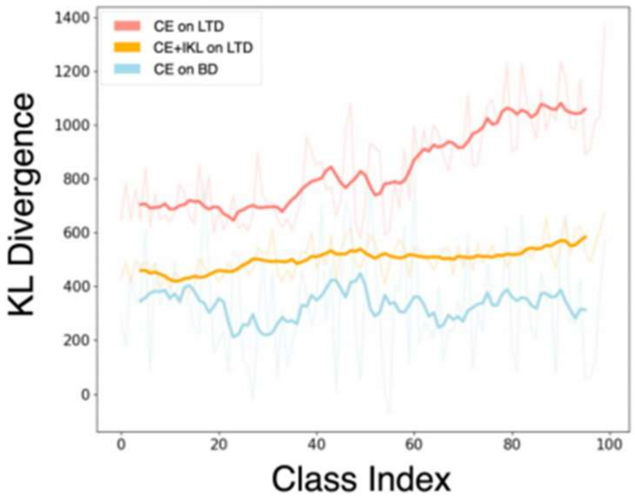
- **Review**： promote consistency between past and current predictions;
- **Summary**： summarize and utilize the relationships across classes;
- **Correction**： correct gradient conflicts in different learning methods

# Method-Knowledge Review

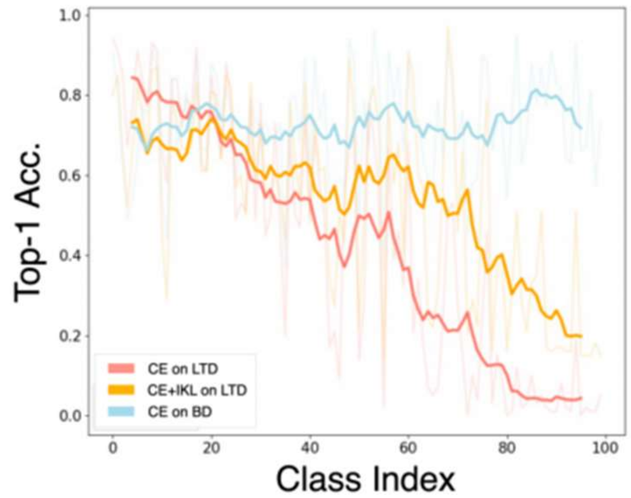
This analysis is conducted on CIFAR100-LT dataset with an Imbalanced Factor (IF) of 100:




(a) Logits between the same networks that across random adjacent epochs.




(b) Comparison of KL Divergence Across Classes for Different Methods and Data Distributions.




(c) Comparison of Top-1 accuracy Across Classes for Different Methods and Data Distributions.

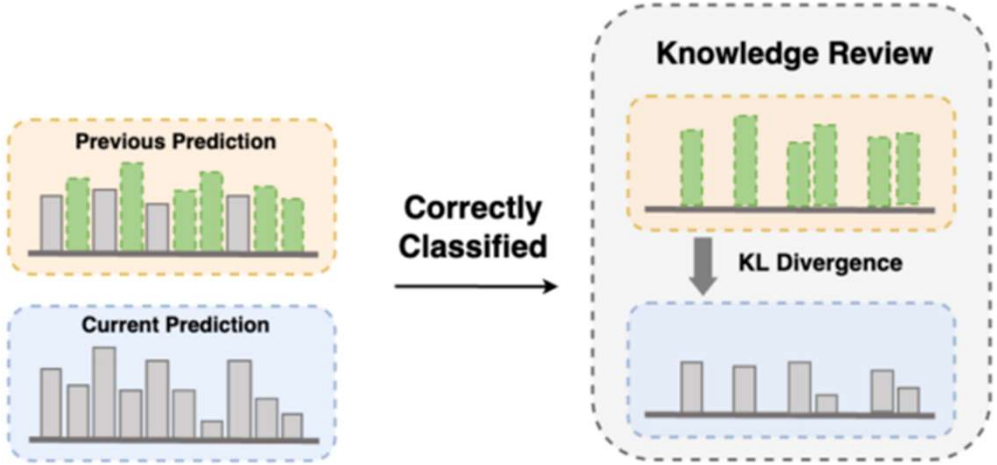
# Method-Knowledge Review

**Target:** Minimize the **KL divergence** of the previous and current epoch's prediction distribution.

$$\mathcal{L}_{KR} = \sum_{x_i \in \mathbb{D}} KL(p_{i,t-1}(x_i;\Theta_{t-1})||p_{i,t}(x_i;\Theta_t))$$

$$KL(p_{i,t-1}||p_{i,t}) = \tau^2 \sum_{i=1}^{n} p_{i,t-1}(x_i;\Theta_{t-1}) log \frac{p_{i,t-1}(x_i;\Theta_{t-1})}{p_{i,t}(x_i;\Theta_t)}.$$

$$p_i(x_i;\Theta) = \frac{e^{(v_i^k/\tau)}}{\sum_c e^{(v_i^c/\tau)}}, \quad v_i = \{f(x_i;\Theta), W\}$$




**Optimize:** only transfer and distill the knowledge that is **correctly classified**

define a **correctly classified instances (CCI) set** containing all correctly classified instances as:

$$\mathbb{D}_{CCI} = \{x_i \in \mathbb{D} | argmax(p_i(x_i;\Theta)) == y_i\},$$

**Re-write:**

$$\mathcal{L}_{KR} = \frac{1}{||\mathbb{D}_{CCI}^{t-1}||} \sum_{x_i \in \mathbb{D}_{CCI}^{t-1}} KL(p_{i,t-1}(x_i;\Theta_{t-1})||p_{i,t}(x_i;\Theta_t))$$

# Method- Knowledge Summary

Humans are adept at summarizing connections and distinctions between knowledge.

However, under a long-tail distribution training setting, this supervision can mislead the model to **misclassify a tail class as a head class**. For example:

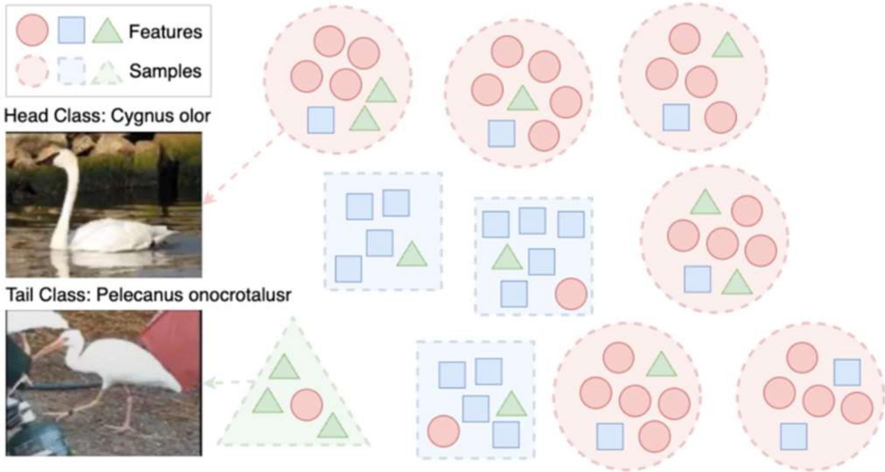



**Fig. 2:** Correlation of features among different samples in long-tailed data.

# Method- Knowledge Summary

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

**Solution: Knowledge Summary Module**

for C-th class, calculate the class center of $f_c$ by the median of all features across the C-th class:

$$f_c = Median_{x_i \in \mathbb{D}}(f(x_i; \Theta_{t-1}))$$

calculate the correlation feature label by **cosine similarity** and reconstruct the label $\hat{y}$:

$$M = \frac{f \cdot f^T}{||f|| \cdot ||f||}, \hat{y} = \alpha \cdot Y + (1-\alpha) \cdot M$$

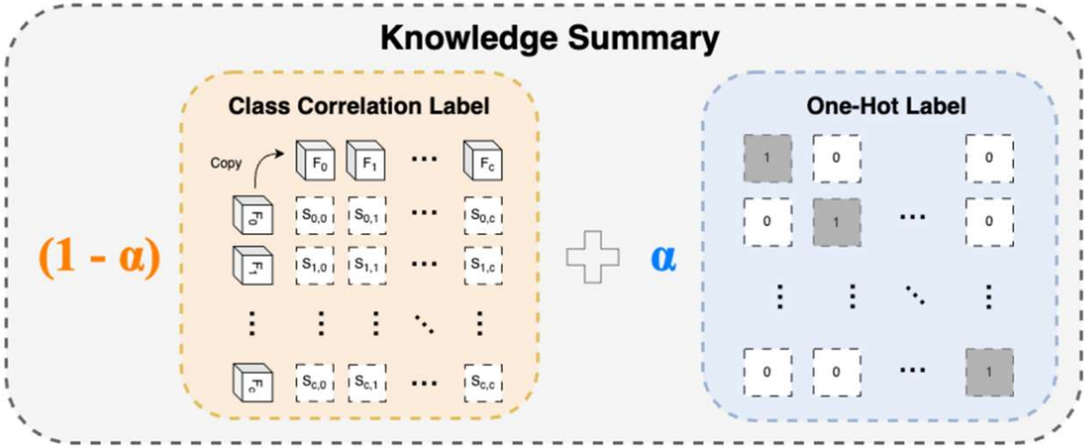
$\alpha$ is a hyperparameter

M $\in (0,1)$ is the feature similarity matrix

$Y$ is the label y after extending to the label matrix

**KS loss:**

$$\mathcal{L}_{KS} = \frac{1}{||\mathbb{D}||} \sum_{x_i \in \mathbb{D}} CrossEntropy(p(x_i; \Theta_t), \hat{y})$$

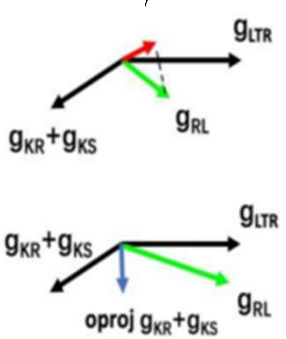


# Method- Knowledge Correction

During the training process, our proposed KR and KS modules can easily combine with the existing LTR methods. Therefore, the overall loss ($\mathcal{L}_{RL}$) for implementation consists of two parts, the existing $\mathcal{L}_{LTR}$ loss for long-tailed recognition and our $\mathcal{L}_{KR}$, $\mathcal{L}_{KS}$ for KR and KS modules, respectively.

It is expressed as:

$$\mathcal{L}_{RL} = \mathcal{L}_{LTR} + (\mathcal{L}_{KR} + \mathcal{L}_{KS}) \tag{9}$$



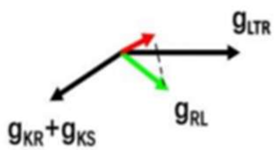

direction conflicts between the two gradients occur when $\cos\alpha_{ij} < 0$

# Method- Knowledge Correction

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

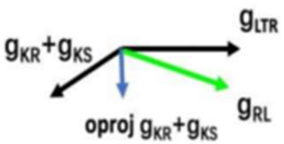
## Solution: Knowledge Correction Module

To address this issue, we introduce knowledge correction (KC) to mitigate conflicts by projecting gradients when negative transfer occurs. Negative transfer between two gradients $g_i$ and $g_j$ is identified when $\cos\alpha(g_i, g_j) < 0$. Following this identification, each gradient is projected onto the orthonormal plane of the other gradients to eliminate harmful conflicts. Therefore, we have the formula for projecting the gradient $\mathcal{L}_{LTR}$ onto the orthonormal plane of gradient $\mathcal{L}_{KR} + \mathcal{L}_{KS}$ as:




$$\hat{g}_{KR+KS} := g_{KR+KS} - \frac{cos(g_{KR+KS}, g_{LTR})}{\|g_{LTR}\|^2} \cdot g_{LTR} \tag{10}$$

final gradient update formula:




$$g_{RL} = \begin{cases} \hat{g}_{KR+KS} + g_{LTR}, \text{if } \cos(g_{KR+KS}, g_{LTR}) < 0 \\ g_{KR+KS} + g_{LTR}, \text{otherwise} \end{cases}$$

# Experiments

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

| Method | CIFAR-100-LT | | |
|---|---|---|---|
| IF | 10 | 50 | 100 |
| Softmax | 59.1 | 45.6 | 41.4 |
| BBN | 59.8 | 49.3 | 44.7 |
| BSCE | 61.0 | 50.9 | 46.1 |
| RIDE | 61.8 | 51.7 | 48.0 |
| SADE | 63.6 | 53.9 | 49.4 |
| Softmax+RL | 59.6 | 46.2 | 41.9 |
| BSCE+RL | 64.5 | 52.2 | 47.9 |
| RIDE+RL | 62.4 | 53.1 | 48.8 |
| SADE+RL | 64.5 | 55.4 | 50.7 |
| BSCE† | 63.0 | - | 50.3 |
| PaCo† | 64.2 | 56.0 | 52.0 |
| SADE† | 65.3 | 57.3 | 53.2 |
| MDCS† | - | - | 56.1 |
| BSCE+RL† | 64.6 | - | 51.2 |
| PaCo+RL† | 65.1 | 57.1 | 52.8 |
| SADE+RL† | **66.8** | **59.1** | 54.7 |
| MDCS+RL† | - | - | **57.3** |

**Table 1:** Comparisons on CIFAR100-LT datasets with the IF of 10, 50, and 100. †denotes models trained with RandAugment [9] for 400 epochs.

| Method | Many | Medium | Few | All |
|---|---|---|---|---|
| Softmax | 68.1 | 41.5 | 14.0 | 48.0 |
| Decouple-LWS | 61.8 | 47.6 | 30.9 | 50.8 |
| BSCE | 64.1 | 48.2 | 33.4 | 52.3 |
| LADE | 64.4 | 47.7 | 34.3 | 52.3 |
| PaCo | 63.2 | 51.6 | 39.2 | 54.4 |
| RIDE | 68.0 | 52.9 | 35.1 | 56.3 |
| SADE | 66.5 | 57.0 | 43.5 | 58.8 |
| Softmax+RL | 68.6 | 42.0 | 14.7 | 48.6 |
| BSCE+RL | 65.6 | 49.7 | 37.9 | 54.8 |
| PaCo+RL | 64.0 | 52.5 | 42.1 | 56.4 |
| RIDE+RL | 68.9 | 54.1 | 38.6 | 59.0 |
| SADE+RL | 66.3 | 58.3 | 47.8 | 60.2 |
| PaCo† | 67.5 | 56.9 | 36.7 | 58.2 |
| SADE† | 67.3 | 60.4 | 46.4 | 61.2 |
| MDCS† | 72.6 | 58.1 | 44.3 | 61.8 |
| PaCo+RL † | 67.4 | 57.3 | 37.8 | 58.8 |
| SADE+RL † | 67.9 | **61.2** | **47.8** | 62.0 |
| MDCS+RL† | **72.7** | 59.5 | 46.0 | **62.7** |

**Table 2:** Comparisons on ImageNet-LT. † denotes models trained with RandAugment [9] for 400 epochs.

Baselines.
re-balancing: cRT,LWS
multi-branch models: BBN,BSCE,LDAM
ensemble learning:NCL,RIDE,SADE

# Experiments

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

| Method | Many | Medium | Few | All |
|---|---|---|---|---|
| Softmax | 46.2 | 27.5 | 12.7 | 31.4 |
| BLS | 42.6 | 39.8 | 32.7 | 39.4 |
| LADE | 42.6 | 39.4 | 32.3 | 39.2 |
| RIDE | 43.1 | 41.0 | 33.0 | 40.3 |
| SADE | 40.4 | 43.2 | 36.8 | 40.9 |
| Softmax+RL | **46.1** | 28.0 | 15.6 | 32.8 |
| BLS+RL | 43.0 | 40.3 | 34.8 | 41.1 |
| LADE+RL | 42.8 | 39.7 | 35.5 | 41.8 |
| RIDE+RL | 43.1 | 41.9 | 36.9 | 42.1 |
| SADE+RL | 41.0 | 44.3 | **38.7** | 42.2 |
| PaCo† | 36.1 | 47.2 | 33.9 | 41.2 |
| PaCo+RL † | 36.4 | **47.7** | 36.6 | **42.8** |

**Table 3:** Comparisons on Places-LT, starting from an ImageNet pre-trained ResNet-152. †denotes models trained with RandAugment [9] for 400 epochs.

| Method | Many | Medium | Few | All |
|---|---|---|---|---|
| Softmax | 74.7 | 66.3 | 60.0 | 64.7 |
| BLS | 70.9 | 70.7 | 70.4 | 70.6 |
| LADE† | 64.4 | 47.7 | 34.3 | 52.3 |
| MiSLAS | 71.7 | 71.5 | 69.7 | 70.7 |
| RIDE | 71.5 | 70.0 | 71.6 | 71.8 |
| SADE | 74.5 | 72.5 | 73.0 | 72.9 |
| Softmax+RL | 75.4 | 67.1 | 61.1 | 65.5 |
| BLS+RL | 68.8 | 72.5 | 75.9 | 73.1 |
| LADE+RL | 64.8 | 48.9 | 36.6 | 73.6 |
| RIDE+RL | 71.4 | 70.9 | 74.8 | 73.6 |
| SADE+RL | 74.7 | 73.1 | **77.8** | 74.2 |
| PaCo† | 69.5 | 73.4 | 73.0 | 73.0 |
| SADE† | 75.5 | 73.7 | 75.1 | 74.5 |
| NCL† | 72.7 | 75.6 | 74.5 | 74.9 |
| PaCo+RL† | 69.6 | 73.4 | 75.9 | 73.6 |
| SADE+RL† | **75.7** | 74.1 | **77.8** | 75.3 |
| NCL+RL† | 72.5 | **76.7** | **77.8** | **76.5** |

**Table 4:** Comparisons on iNaturalist 2018. † denotes models trained with RandAugment [9] for 400 epochs.

Many (with more than 100 images)
Medium (with 20 to 100 images)
Few (with less than 20 images)

# Experiments

| Method | Resnet-50 | ResNeXt-50 | Swin-T | Swin-S |
|---|---|---|---|---|
| Softmax | 41.6 | 44.4 | 42.6 | 42.9 |
| OLTR | - | 46.3 | - | - |
| $\tau$-norm | 46.7 | 49.4 | - | - |
| cRT | 47.7 | 49.9 | - | - |
| LWS | 47.3 | 49.6 | - | - |
| LDAM | - | - | 50.6 | 49.5 |
| RIDE | 54.9 | 56.4 | 56.3 | 54.2 |
| Softmax+RL | 45.8 | 47.3 | 43.7 | 43.6 |
| $\tau$-norm+RL | 47.3 | 50.5 | - | - |
| cRT+RL | 48.5 | 51.2 | - | - |
| LWS+RL | 48.5 | 50.5 | - | - |
| LDAM+RL | - | - | 52.1 | 50.3 |
| RIDE+RL | **56.8** | **58.7** | **59.1** | **55.6** |

**Table 5:** Comparisons on ImageNet-LT with different backbones.

| Method | Many | Med | Few | All |
|---|---|---|---|---|
| Softmax | 66.1 | 37.3 | 10.6 | 41.4 |
| OLTR | 61.8 | 41.4 | 17.6 | - |
| $\tau$-norm | 65.7 | 43.6 | 17.3 | 43.2 |
| cRT | 64.0 | 44.8 | 18.1 | 43.3 |
| LDAM | 61.5 | 41.7 | 20.2 | 42.0 |
| RIDE | 69.3 | 49.3 | 26.0 | 48.0 |
| SADE | 60.3 | 50.2 | 33.7 | 49.4 |
| Softmax+RL | 66.8 | 37.9 | 11.2 | 41.9 |
| LDAM+RL | 62.4 | 42.4 | 28.3 | 49.2 |
| RIDE+RL | **69.9** | 50.4 | 28.1 | 49.2 |
| SADE+RL | 60.4 | **50.8** | **35.5** | **50.7** |

**Table 6:** Comparisons on CIFAR-100-LT(IF=100) with different sample sizes.

# Experiments



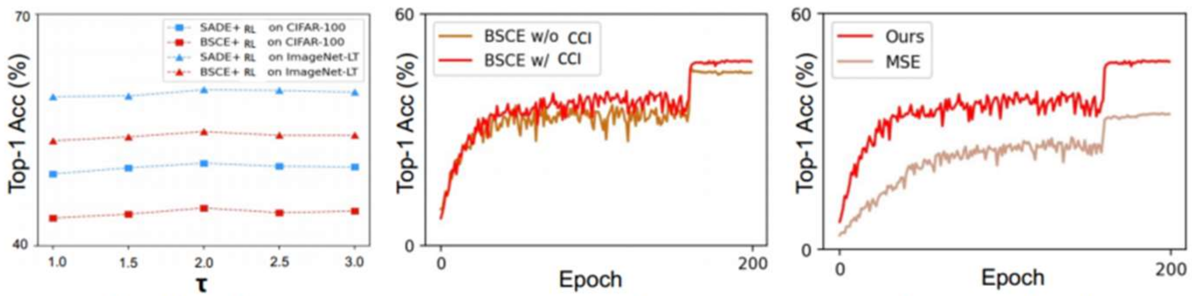
| Method | CIFAR100-LT | ImageNet-LT | iNaturalist 2018 |
|---|---|---|---|
| Decouple | 43.8 | 47.9 | 67.7 |
| Mixup | 45.1 | 51.5 | 70.0 |
| MiSLAS | 47.0 | 52.7 | 71.6 |
| WD + WD & Max | 53.6 | 53.9 | 70.2 |
| Decouple + RL | 50.9 | 54.5 | 72.8 |
| MiSLAS & RL | 53.1 | 56.0 | **74.2** |
| WD & RL + WD & Max & RL | **56.8** | **56.7** | 73.5 |

**Table 7:** Results of comparing and combining our method with other regularization-based methods.

# Component Analysis and Ablation Study

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

**The effective of temperature $\tau$.** The temperature parameter $\tau$ is introduced to soften the previous predictions, allowing the current model to learn from a smoother, more generalized distribution. By adjusting the temperature parameter during training, we can control the trade-off between accuracy and generalization to optimize the current prediction. Higher temperature values lead to better generalization but lower accuracy, while lower temperature values lead to better accuracy but less generalization. In Figure. 5 (a), we show several settings of $\tau$ on the CIFAR-100LT (IF=100) and ImageNet-LT, we observe that when the $\tau$ set to 2, the models achieve the best performance.




(a) The effect of τ. (c) The effect of CCI. (d) Direct matching logits.

**Fig. 5:** Figure (a): The effect of temperature $\tau$ for different methods and datasets. Figure (b): The effect of our CCI. Figure (c): The effect of directing matching logits.

# Component Analysis and Ablation Study



**The effectiveness of our components KR, KS and KC.** Our proposed method is fundamentally composed of two primary components: Knowledge Review (KR) and Knowledge Summary (KS). As shown in Tab 8, the KR component is designed to enforce consistency across all categories. As a result, it notably enhances the accuracy of the tail classes, but this comes at the expense of a slight reduction in the accuracy of the head classes. In contrast, KS facilitates learning across all categories by leveraging the inherent feature correlations, compensating for the minor drawbacks introduced by KS, and ensuring an overall improved performance.

| Method | | | ImageNet-LT | | iNaturalist 2018 | |
|---|---|---|---|---|---|---|
| KR | KS | KC | RIDE | SADE | RIDE | SADE |
| - | - | - | 56.3 | 58.8 | 71.8 | 72.9 |
| ✓ | - | - | 58.0 | 59.7 | 72.4 | 73.3 |
| - | ✓ | - | 58.4 | 59.3 | 72.7 | 73.6 |
| ✓ | ✓ | - | 58.6 | 60.0 | 72.9 | 73.8 |
| ✓ | ✓ | ✓ | **59.0** | **60.2** | **73.6** | **74.2** |

**Table 8: Ablation study on the components of our methods.** Comparisons with different component combinations.

# Thanks