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

Global and Local Mixture Consistency Cumulative Learning for Long-tailed Visual Recognitions

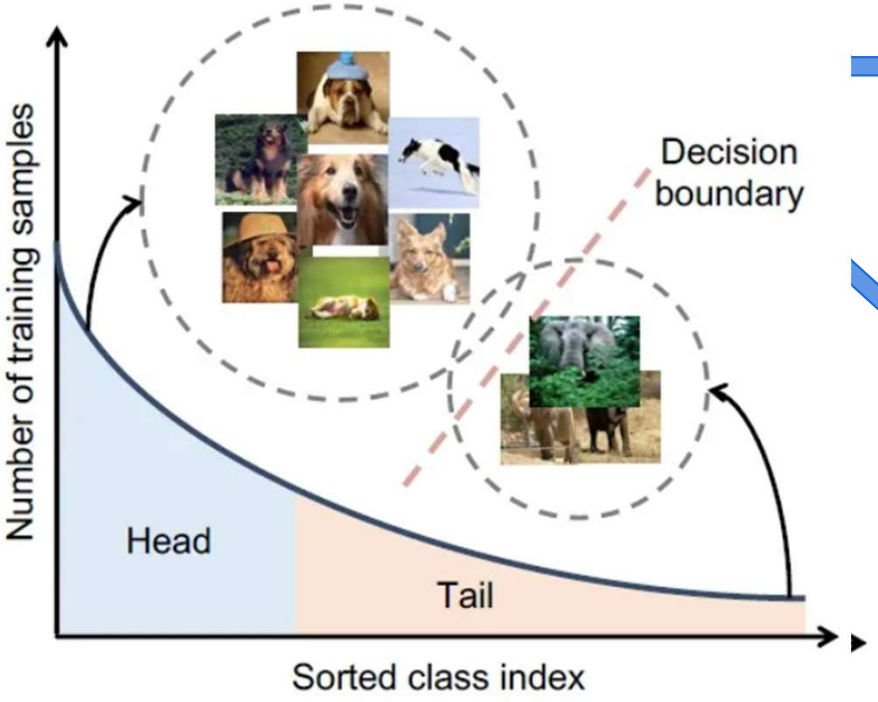
Fei Du[1,2,3], Peng Yang[1,3], Qi Jia[1,3], Fengtao Nan[1,2,3], Xiaoting Chen[1,3], Yun Yang[1,3*]
[1]National Pilot School of Software, Yunnan University, Kunming, China
[2]School of Information Science and Engineering, Yunnan University, Kunming, China
[3]Yunnan Key Laboratory of Software Engineering
{dufei,yangpeng,jiaqi,fengtaonan,chenxiaoting}@mail.ynu.edu.cn yangyun@ynu.edu.cn

CVPR 2023

Introduction

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

Long-tail Problems



The label distribution of a long-tailed dataset

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

the decision boundary is usually biased towards dominant classes

Introduction

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

Classical Rebalanced Strategies:

Re-sampling training data

oversample the tail class data or undersample the head classes

Designing cost-sensitive re-weighting loss functions

increases the loss weight of the tail classes to strengthen the tail class

Similarity:a two-stage training process

- Stage 1: trains the feature extractor on the original data distribution
- Stage 2: fixes the representation and trains a balanced classifier

Introduction

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

Motivation： Although multi-stage training significantly improves the performance of long-tail recognition, it also negatively increases the training tricks and overhead.

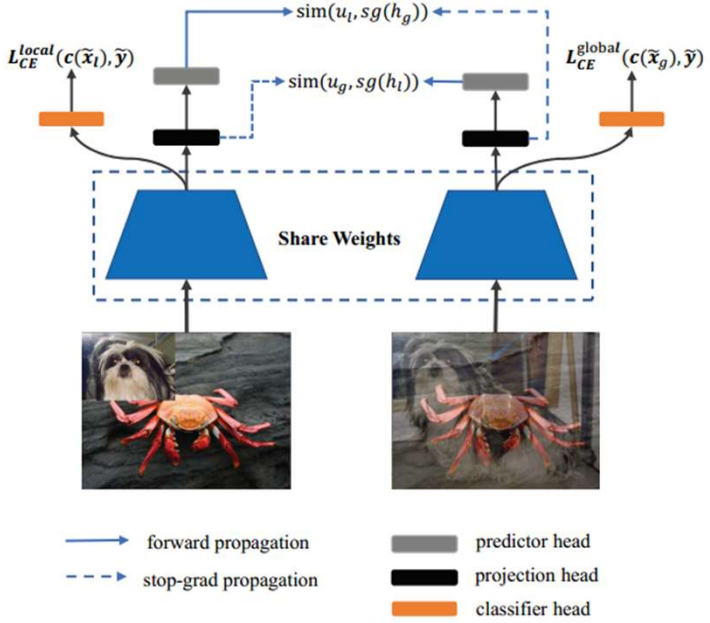


Figure 1. An overview of our GLMC: two types of mixed-label augmented images are processed by an encoder network and a projection head to obtain the representation h_g and h_l. Then a prediction head transforms the two representations to output u_g and u_l. We minimize their negative cosine similarity as an auxiliary loss in the supervised loss. $sg(\cdot)$ denotes stop gradient operation.

Solution:
an efficient one-stage training strategy called Global and Local Mixture Consistency cumulative learning framework (GLMC)

Core ideas:

- A global and local mixture consistency loss
- A cumulative head-tail soft label reweighted loss

Method

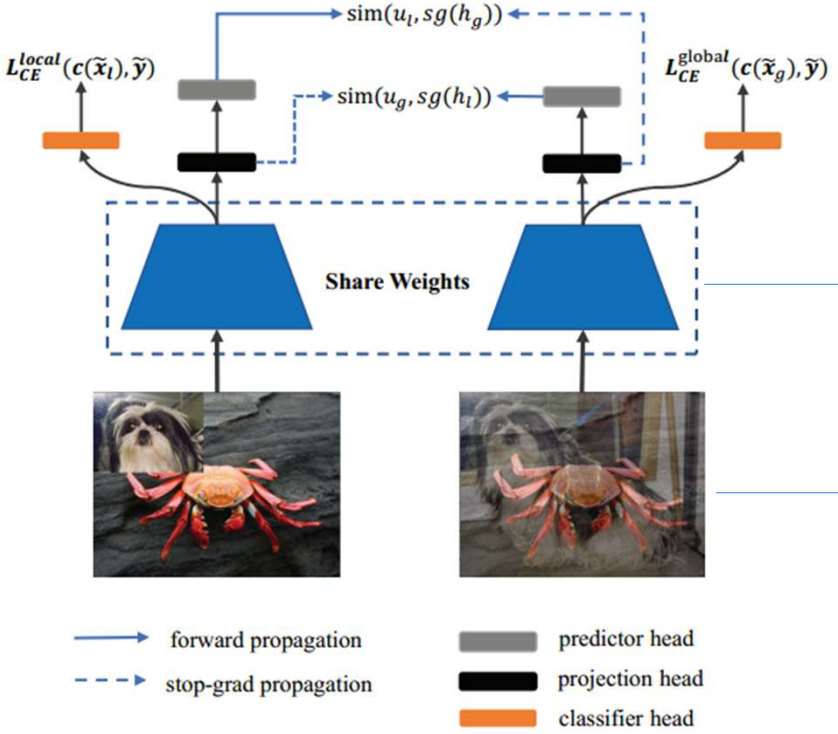


Figure 1. An overview of our GLMC: two types of mixed-label augmented images are processed by an encoder network and a projection head to obtain the representation h_g and h_l. Then a prediction head transforms the two representations to output u_g and u_l. We minimize their negative cosine similarity as an auxiliary loss in the supervised loss. $sg(\cdot)$ denotes stop gradient operation.

train the encoder using a standard supervised task and a self-supervised task in a multi-task learning way.

data augmentation method:MixUp and CutMix

Method

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

Self-Supervised Learning Branch

Target:to maximize the cosine similarity of global and local mixtures in representation space to obtain contrastive consistency

minimize their negative cosine similarity:

$$sim(u_g, h_l) = -\frac{u_g}{\|u_g\|} \cdot \frac{h_l}{\|h_l\|}$$

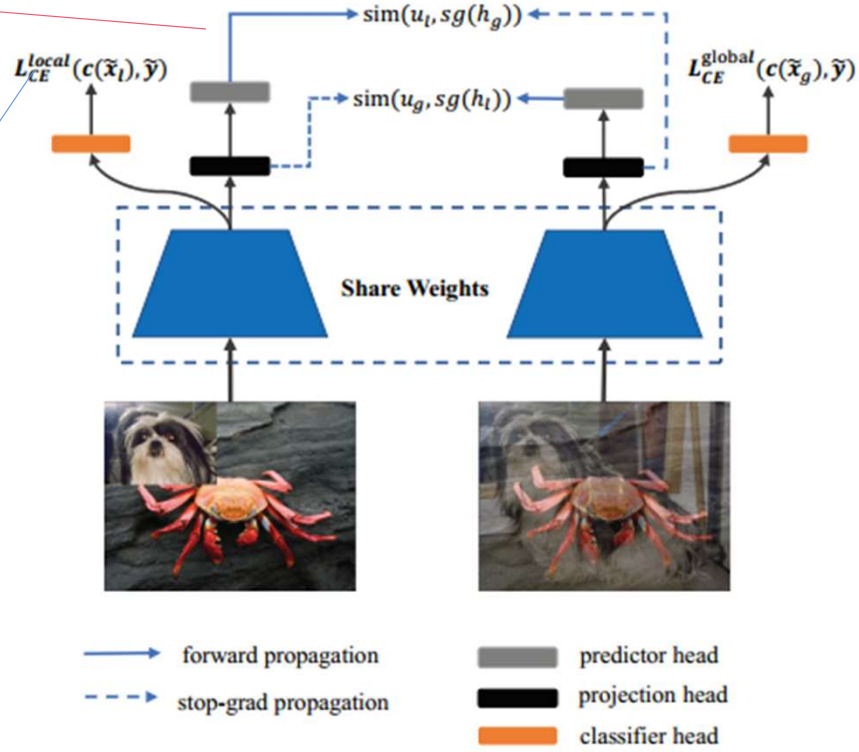
The loss function is defined as:

$$\pounds_{sim} = sim(u_g, sg(h_l)) + sim(u_l, sg(h_g))$$

Supervised Learning Branch.

calculate the mixed-label cross-entropy loss:

$$\pounds_c = -\frac{1}{2N}\sum_{i=1}^{N}(\tilde{p}_g^i(logf(\tilde{x}_g^i)) + \tilde{p}_l^i(logf(\tilde{x}_l^i)))$$



Method

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

Class-Balanced Learning

Class-Balanced learning. The design principle of class reweighting is to introduce a weighting factor inversely proportional to the label frequency and then strengthen the learning of the minority class. Following [44], the weighting factor w_i is define as:

$$w_i = \frac{C \cdot (1/r_i)^k}{\sum_{i=1}^{C}(1/r_i)^k} \tag{7}$$

where r_i is the i-th class frequencies of the training dataset, and k is a hyper-parameter to scale the gap between the head and tail classes. Note that $k = 0$ corresponds to no reweighting and $k = 1$ corresponds to class-balanced method [9]. We change the scalar weights to the one-hot vectors form and mix the weight vectors of the two images:

$$\tilde{w} = \lambda w_i + (1 - \lambda) w_j. \tag{8}$$

given a train dataset: $D = \{(x_i, y_i, w_i)\}_{i=1}^{N}$,

rebalanced loss: $\pounds_{cb} = -\frac{1}{2N}\sum_{i=1}^{N}\tilde{w}^i(\tilde{p}_g^i(logf(\tilde{x}_g^i)) + \tilde{p}_l^i(logf(\tilde{x}_l^i)))$

Method

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

Cumulative Class-Balanced Learning

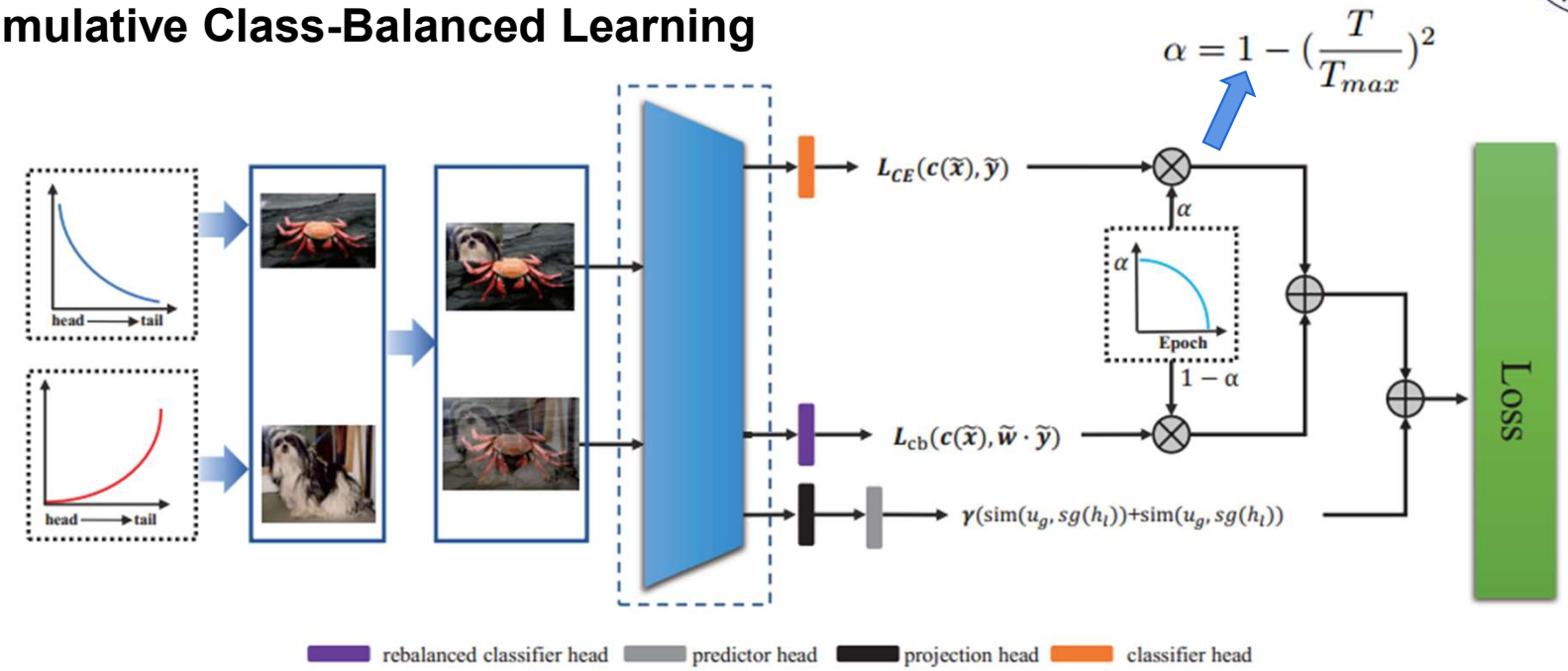


Figure 2. An illustration of the cumulative class-balanced learning pipeline. We apply uniform and reversed samplers to obtain head and tail data, and then they are synthesized into head-tail mixture samples by MixUp and CutMix. The cumulative learning strategy adaptively weights the rebalanced classifier and the conventional classifier by epochs.

The total loss is defined as: $\pounds_{total} = \alpha \pounds_c + (1 - \alpha) \pounds_{cb} + \gamma \pounds_{sim}$

Method

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

Extra Works: Finetuning Classifier Weights

The author use MaxNorm to finetune the classifier in the second stage (MaxNorm restricts weight norms within a ball of radius δ):

$$\Theta^* = \underset{\Theta}{argmin} F(\Theta; D), \ s.t. ||\theta_k||_2^2 \leq \delta^2 \qquad (12)$$

this can be solved by applying projected gradient descent (PGD). For each epoch (or iteration), PGD first computes an updated θ_k and then projects it onto the norm ball:

$$\theta_k \leftarrow min(1, \delta / ||\theta_k||_2) * \theta_k \qquad (13)$$

Experiments

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

Table 1. Top-1 accuracy (%) of ResNet-32 on CIFAR-10-LT and CIFAR-100-LT with different imbalance factors [100, 50, 10]. GLMC consistently outperformed the previous best method only in the one-stage.

	Method	CIFAR-10-LT			CIFAR-100-LT		
		IF=100	50	10	100	50	10
	CE	70.4	74.8	86.4	38.3	43.9	55.7
rebalance classifier	BBN [45]	79.82	82.18	88.32	42.56	47.02	59.12
	CB-Focal [9]	74.6	79.3	87.1	39.6	45.2	58
	LogitAjust [29]	80.92	-	-	42.01	47.03	57.74
	weight balancing [1]	-	-	-	53.35	57.71	68.67
augmentation	Mixup [42]	73.06	77.82	87.1	39.54	54.99	58.02
	RISDA [6]	79.89	79.89	79.89	50.16	53.84	62.38
	CMO [32]	-	-	-	47.2	51.7	58.4
self-supervised pretraining	KCL [18]	77.6	81.7	88	42.8	46.3	57.6
	TSC [25]	79.7	82.9	88.7	42.8	46.3	57.6
	BCL [47]	84.32	87.24	91.12	51.93	56.59	64.87
	PaCo [8]	-	-	-	52	56	64.2
	SSD [26]	-	-	-	46	50.5	62.3
ensemble classifier	RIDE (3 experts) + CMO [32]	-	-	-	50	53	60.2
	RIDE (3 experts) [37]	-	-	-	48.6	51.4	59.8
one-stage training	**ours**	**92.34**	**94.37**	**94.92**	**55.88**	**61.08**	**70.74**
finetune classifier	**ours + MaxNorm [1]**	**94.18**	**95.13**	**95.7**	**57.11**	**62.32**	**72.33**

Experiments

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

Table 2. Top-1 accuracy (%) on ImageNet-LT dataset. Comparison to the state-of-the-art methods with different backbone. † denotes results reproduced by [47] with 180 epochs.

Method	Backbone	ImageNet-LT			
		Many	Med	Few	All
CE	ResNet-50	64	33.8	5.8	41.6
CB-Focal [9]	ResNet-50	39.6	32.7	16.8	33.2
LDAM [3]	ResNet-50	60.4	46.9	30.7	49.8
KCL [18]	ResNet-50	61.8	49.4	30.9	51.5
TSC [25]	ResNet-50	63.5	49.7	30.4	52.4
RISDA [6]	ResNet-50	-	-	-	49.3
BCL (90 epochs) [47]	ResNeXt-50	67.2	53.9	36.5	56.7
BCL (180 epochs) [47]	ResNeXt-50	67.9	54.2	36.6	57.1
PaCo† (180 epochs) [8]	ResNeXt-50	64.4	55.7	33.7	56.0
Balanced Softmax† (180 epochs) [34]	ResNeXt-50	65.8	53.2	34.1	55.4
SSD [26]	ResNeXt-50	66.8	53.1	35.4	56
RIDE (3 experts) + CMO [32]	ResNet-50	66.4	53.9	35.6	56.2
RIDE (3 experts) [37]	Swin-S	66.9	52.8	37.4	56
weight balancing + MaxNorm [1]	ResNeXt-50	62.5	50.4	41.5	53.9
ours	ResNeXt-50	**70.1**	52.4	30.4	56.3
ours + MaxNorm [1]		60.8	**55.9**	**45.5**	56.7
ours + BS [34]		64.76	55.67	42.19	**57.21**

Experiments

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

Table 3. Top-1 accuracy (%) on full ImageNet dataset with ResNet-50 backbone.

Method	Augmentation	Top-1 acc
vanilla	Simple Augment	76.4
vanilla	MixUp [42]	77.9
vanilla	CutMix [41]	78.6
Supcon [21]	RandAugment	78.4
PaCo [8]	Simple Augment	78.7
PaCo [8]	RandAugment	79.3
ours	MixUp + CutMix	**80.2**

Table 4. Top-1 accuracy (%) on full CIFAR-10 and CIFAR-100 dataset with ResNet-50 backbone.

Method	CIFAR-10	CIFAR-100
vanilla	94.85	75.28
MixUp [42]	95.95	77.99
CutMix [41]	95.41	78.03
SupCon [21]	96	76.5
PaCo [8]	-	79.1
ours	**97.23**	**83.05**

GLMC utilizes a global and local mixture consistency loss as an auxiliary loss in supervised loss to improve the robustness of the model, which can be added to the model as a plug-and-play component. To verify the effectiveness of GLMC under a balanced setting, we conduct experiments on full ImageNet and full CIFAR. They are indicative to compare GLMC with the related state-of-the-art methods (MixUp [42], CutMix [41], PaCo [8], and SupCon [8]). Note that under full ImageNet and CIFAR, we remove the cumulative reweighting and resampling strategies customized for long-tail tasks.

Experiments

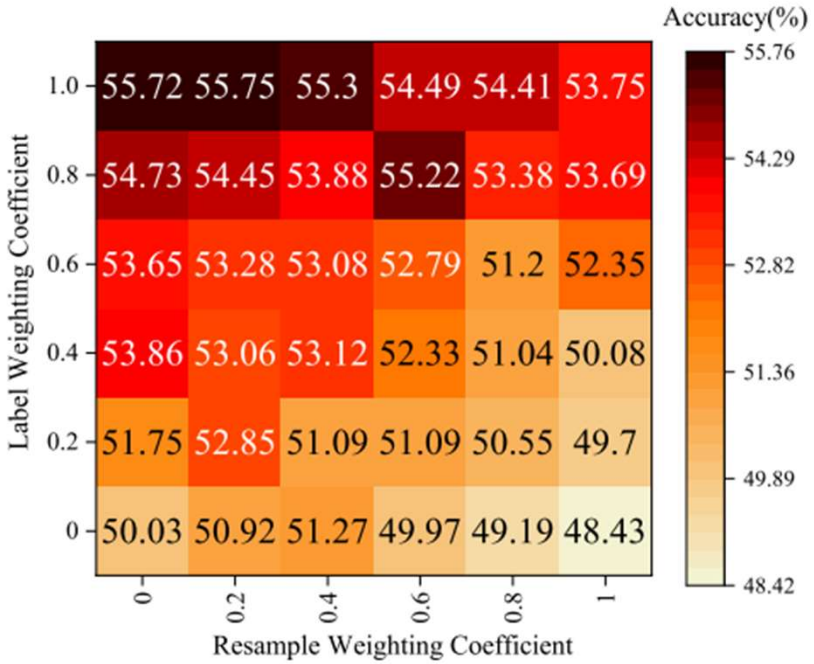


Figure 3. Confusion matrices of different label reweighting and resample coefficient k on CIFAR-100-LT with an imbalance ratio of 100.

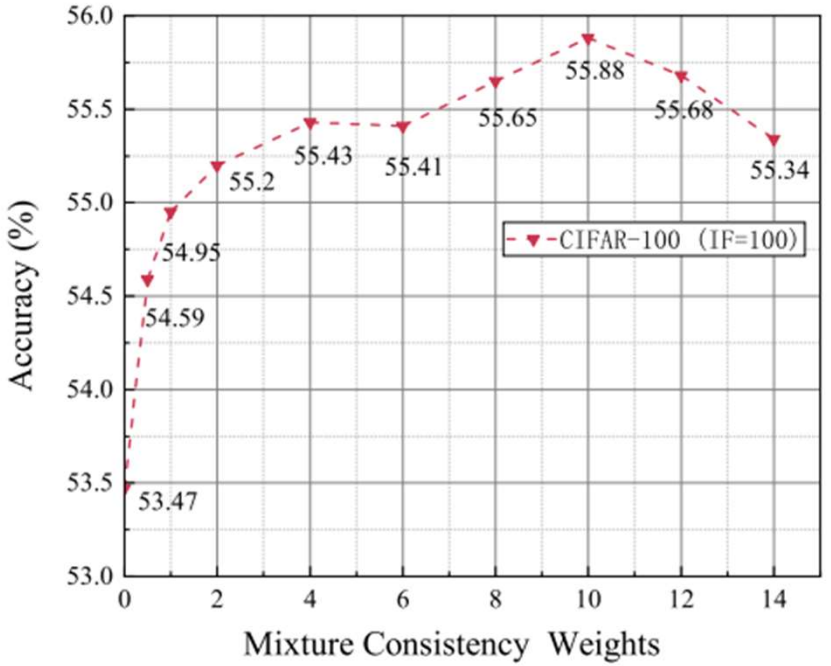


Figure 4. Different global and local mixture consistency weights on CIFAR-100-LT (IF = 100) .

Experiments

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

The effect of each component
- Global and Local Mixture Consistency Learning
- Cumulative Class-Balanced reweighting.

Table 5. Ablations of the different key components of GLMC architecture. We report the accuracies (%) on CIFAR100-LT (IF=100) with ResNet-32 backbone. Note that all model use one-stage training.

Global and Local Mixture Consistency	Cumulative Class-Balanced	Accuracies(%)
×	×	38.3
×	✓	44.63
✓	×	50.11
✓	✓	55.88

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

Thanks