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

Self Supervision to Distillation for Long-Tailed Visual Recognition

Tianhao Li Limin Wang✉ Gangshan Wu
State Key Laboratory for Novel Software Technology, Nanjing University, China

ICCV 2021

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

DECOUPLING REPRESENTATION AND CLASSIFIER FOR LONG-TAILED RECOGNITION

Bingyi Kang[1,2], Saining Xie[1], Marcus Rohrbach[1], Zhicheng Yan[1], Albert Gordo[1], Jiashi Feng[2], Yannis Kalantidis[1]
[1]Facebook AI, [2]National University of Singapore
kang@u.nus.edu,{s9xie,mrf,zyan3,agordo,yannisk}@fb.com,elefjia@nus.edu.sg

Rebalance the classifier：

- Classifier Re-training (cRT)
  Re-train the classifier with class balanced sampling.

- τ -normalized classifier (τ -normalized)
  Adjusting the classifier weight norms.

$$\widetilde{w_i} = \frac{w_i}{||w_i||^{\tau}},$$

- **Learnable weight scaling (LWS)**
  Learning f_i on the training set.

$$\widetilde{w_i} = f_i * w_i, \text{where } f_i = \frac{1}{||w_i||^{\tau}}.$$

Introducton

Self **S**upervision to **D**istillation for Long-Tailed Visual Recognition (**SSD**):

Motivation:

The recent methods are incapable of **capturing tail class information** in the feature learning stage.

Method:

In this paper, we show that **soft label** can serve as a powerful solution to incorporate label correlation into a multi-stage training scheme for long-tailed recognition.

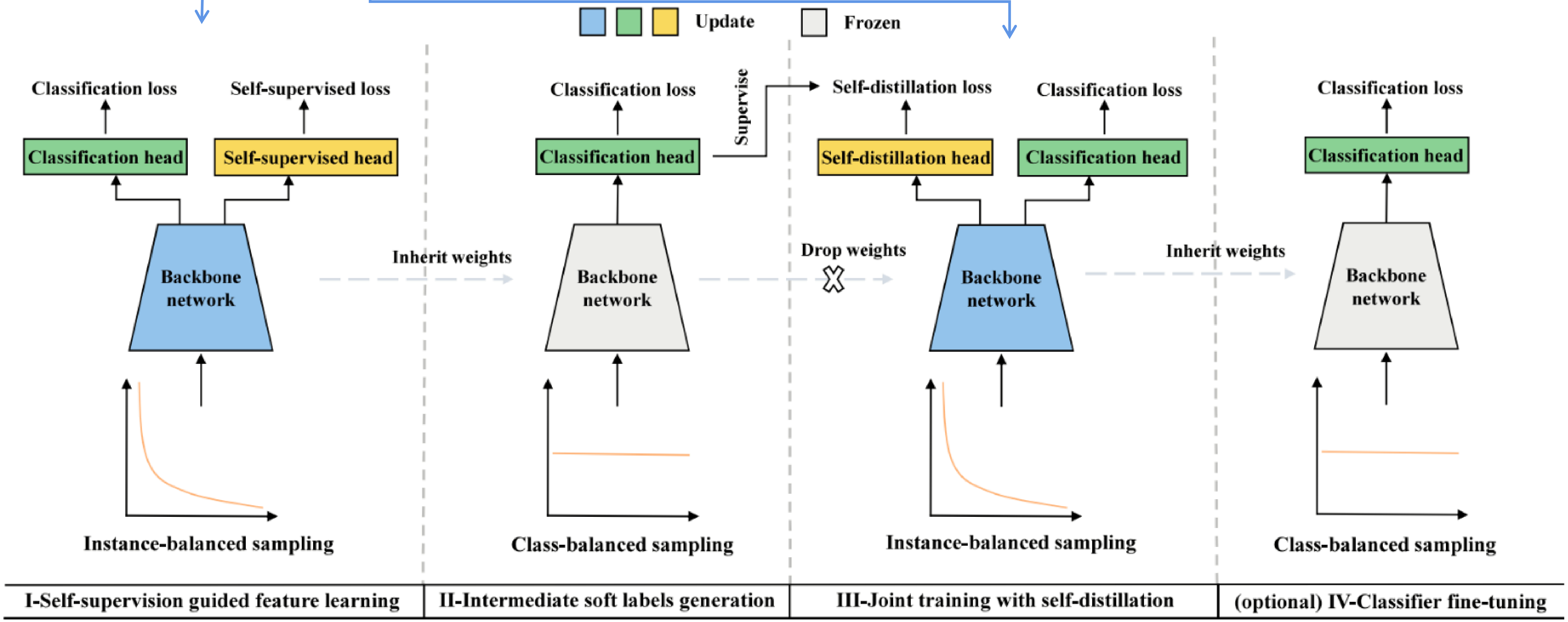
How to **generate** and **use** soft label?

Framework

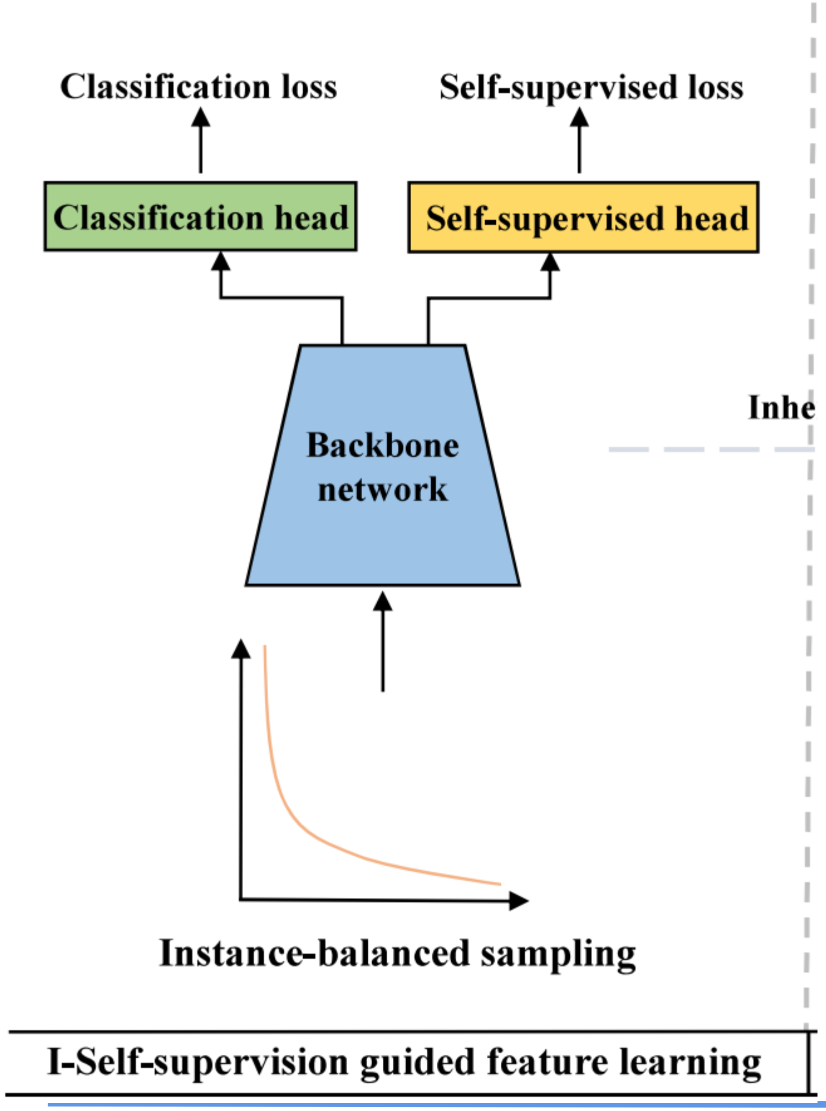
南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Self Supervision to Distillation (SSD)

LWS



I-Self-supervision guided feature learning



Stage 1:

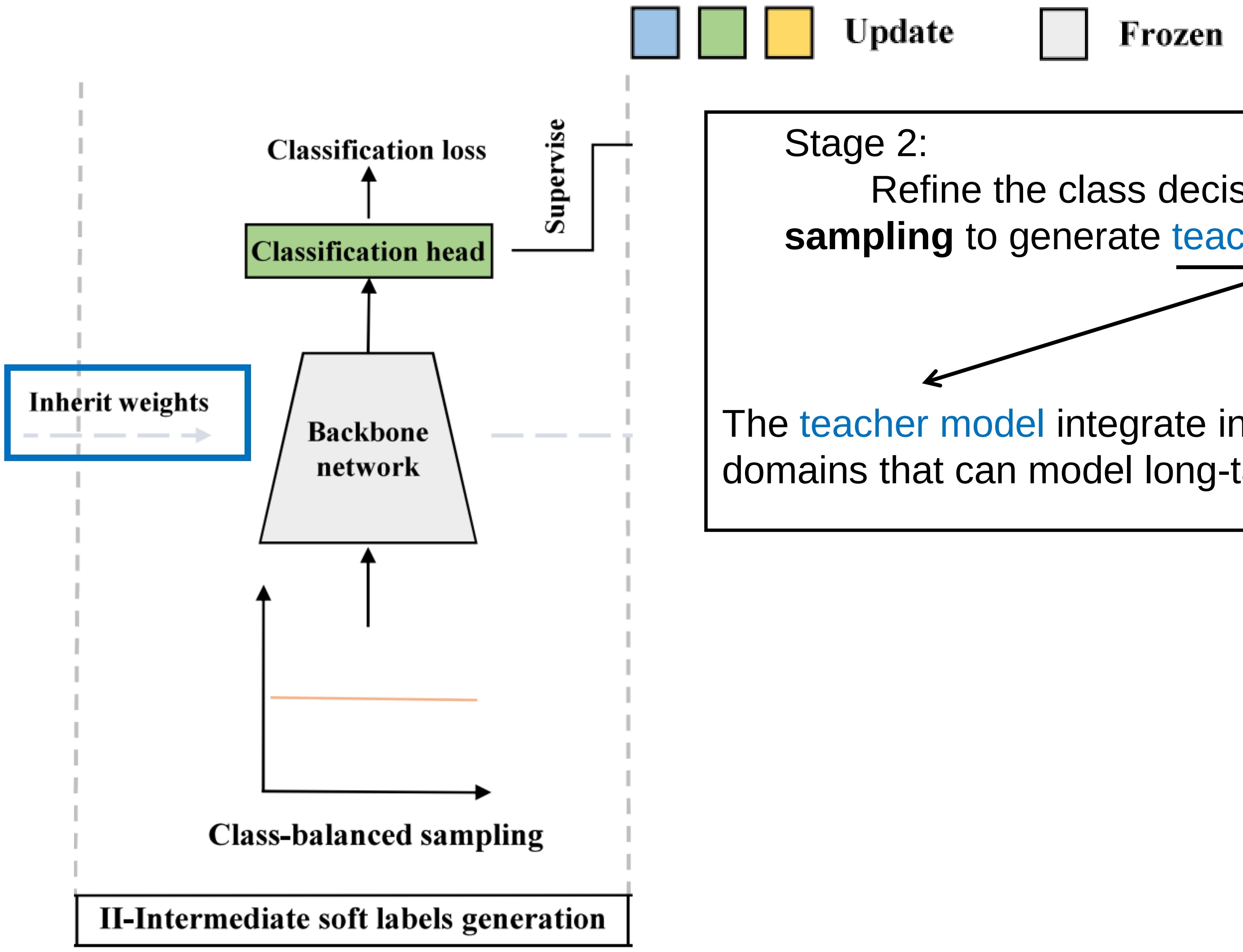
Train an initial feature network under **label supervision** and **self-supervision** jointly using **instance-balanced sampling**.

$$\mathcal{L} = \alpha_1 \mathcal{L}_{sup}(\mathbf{x}; \theta, \omega_{sup}) + \underline{\alpha_2 \mathcal{L}_{self}(\mathbf{x}, \mathbf{y}; \theta, \omega_{self}),} \quad (1)$$

Instance discrimination: (MOCO)

$$\mathcal{L}_{self} = -\log\left(\frac{\exp(\mathbf{v}_i \mathbf{v'}_i / \tau)}{\exp(\mathbf{v}_i \mathbf{v'}_i / \tau) + \sum_K \exp(\mathbf{v}_i \mathbf{v'}_k / \tau)}\right), \quad (2)$$

II-Intermediate soft labels generation.



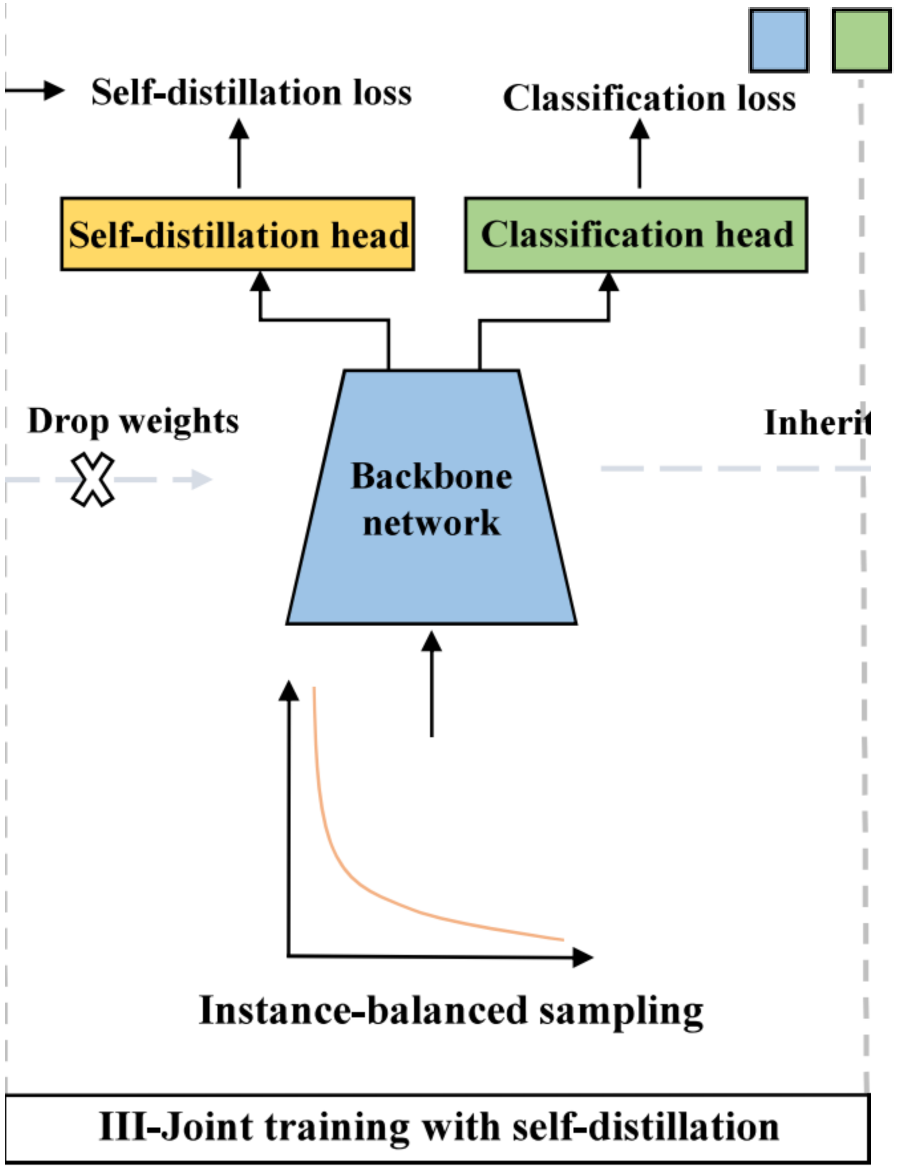
Stage 2:

Refine the class decision boundaries with **class balanced sampling** to generate teacher model by **fixing** the feature backbone.

The teacher model integrate information from both **label** and **data** domains that can model long-tailed distribution effectively.

III-Joint training with self-distillation.

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



Stage 3:

Train a self-distillation network with two classification heads under the supervision of both **soft** labels from previous stages and **hard** labels from the original training set.

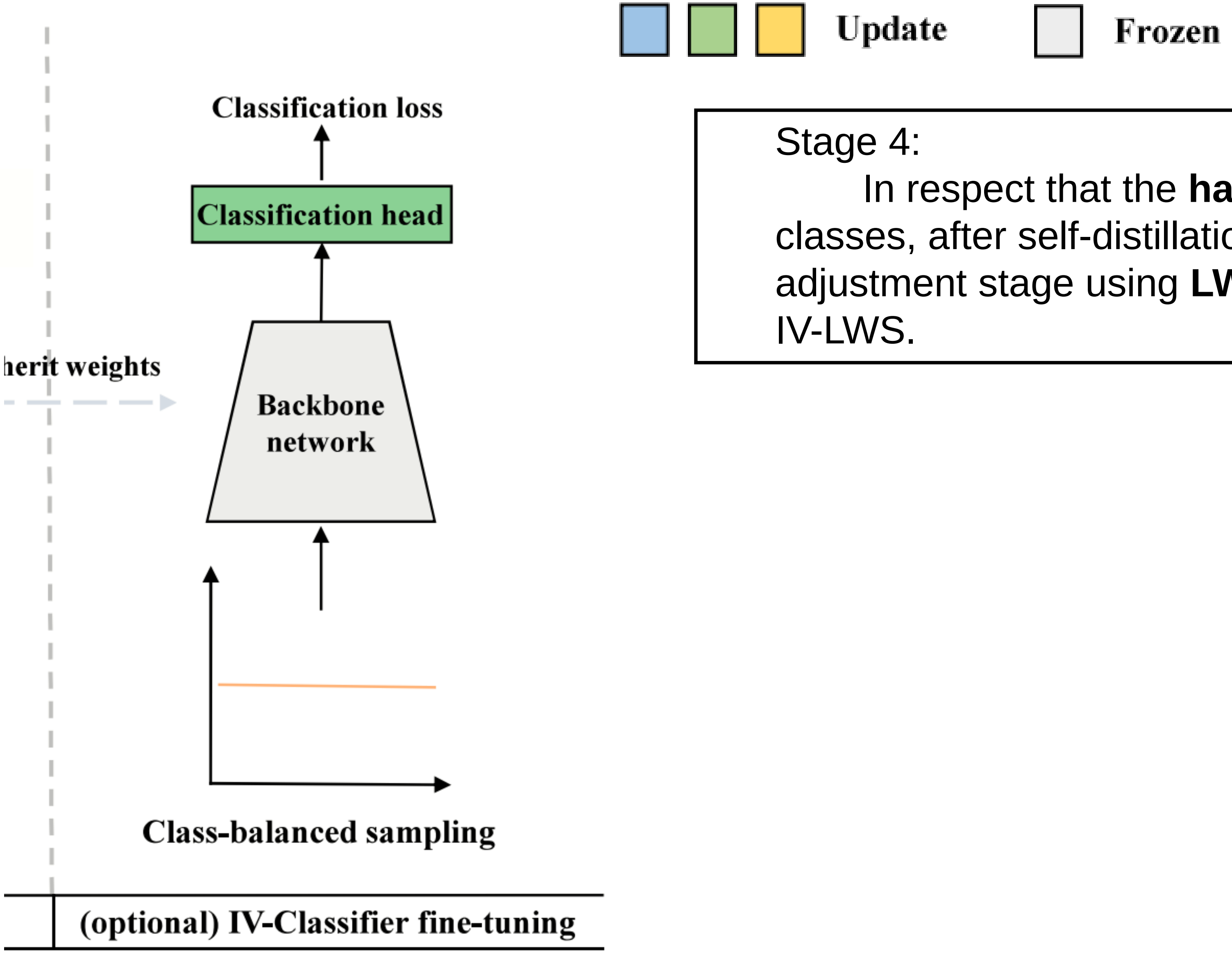
$$\mathcal{L} = \lambda_1 \mathcal{L}_{ce}(\mathbf{y}, \mathbf{z}^{hard}) + \lambda_2 \underline{\mathcal{L}_{kd}(\tilde{\mathbf{y}}, \mathbf{z}^{soft})}, \quad (5)$$

$$\mathcal{L}_{kd}(\underline{\tilde{\mathbf{y}}}, \mathbf{z}^{soft}) = -T^2 \sum_{i=1}^{C} \underline{\tilde{y}_i} \log\left(\frac{\exp(z_i^{soft}/T)}{\sum_{k=1}^{C} \exp(z_k^{soft}/T)}\right). \quad (4)$$

teacher

$$\tilde{y}_i = \frac{\exp(\tilde{z}_i/T)}{\sum_{k=1}^{C} \exp(\tilde{z}_k/T)}, \quad (3)$$

IV-Classifier fine-tuning

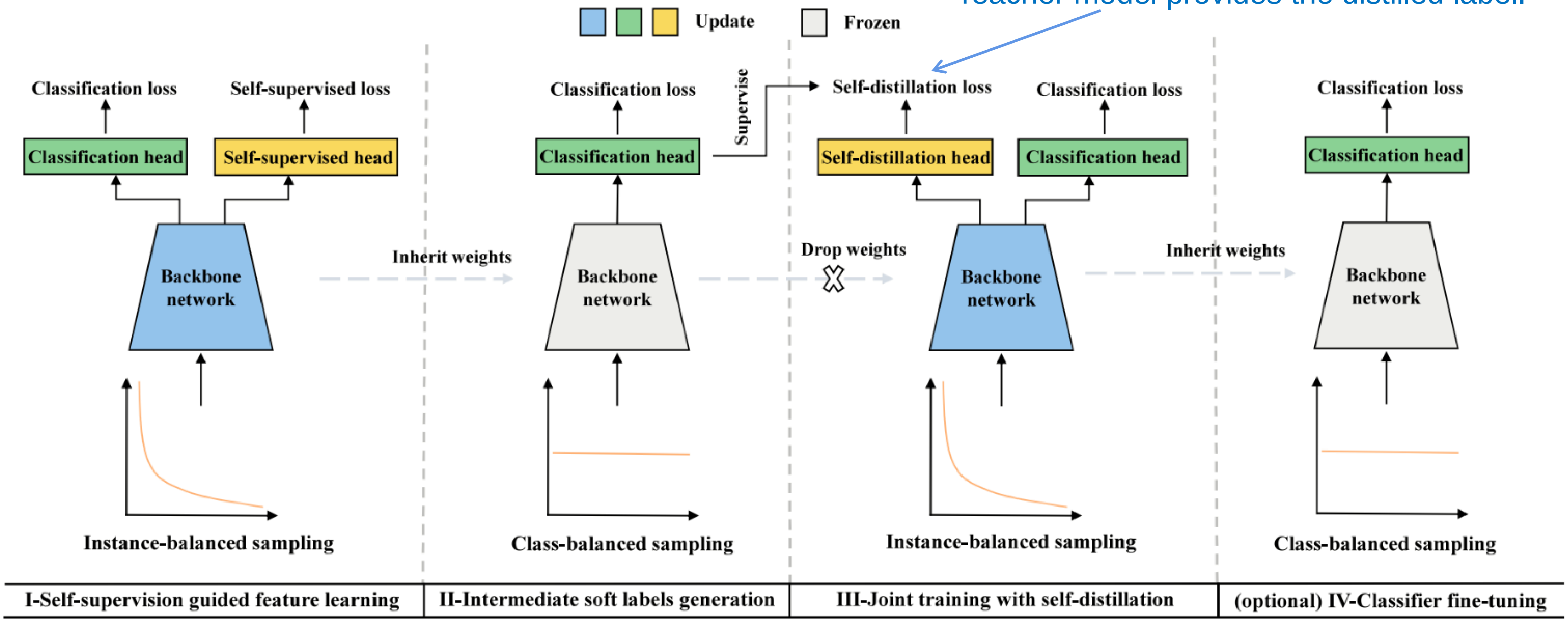


Stage 4:

In respect that the **hard** classifier is still biased to head classes, after self-distillation, we propose run the **soft** classifier adjustment stage using **LWS** for further improvement, termed as IV-LWS.

Framework

Teacher model provides the distilled label.



Fine-tune under the class-balanced setting to generate **teacher** model.

Ablation study on ImageNet-LT

- The effectiveness of the **stage I** (Self-supervision guided feature learning).

Methods	1.5×	I	II	III-hard (test)	III-soft (test)	IV-LWS	Many	Medium	Few	Overall
CE							66.9	38.0	8.1	45.1
	✓						67.9	39.5	9.5	46.3
LWS							61.1	48.0	31.5	50.7
	✓						63.4	48.6	32.3	52.1
Our SSD	✓	✓					69.8	42.8	11.0	48.9
	✓	✓	✓				64.9	51.1	34.0	54.1
	✓		✓			✓	66.0	50.8	34.2	54.4
	✓	✓	✓	✓			**71.1**	46.1	15.6	51.6
	✓	✓	✓		✓		67.1	52.8	33.3	55.7
	✓	✓	✓			✓	66.8	**53.1**	**35.4**	**56.0**

+1.9% +3.3% +1.6% +2.6%

Ablation study on ImageNet-LT

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

- The effectiveness of the **stage II** (Fine-tune under the class-balanced setting to generate teacher model).

Methods	1.5×	I	II	III-hard (test)	III-soft (test)	IV-LWS	Many	Medium	Few	Overall
CE							66.9	38.0	8.1	45.1
	✓						67.9	39.5	9.5	46.3
LWS							61.1	48.0	31.5	50.7
	✓						63.4	48.6	32.3	52.1
Our SSD	✓	✓					69.8	42.8	11.0	48.9
	✓	✓	✓				64.9	51.1	34.0	54.1
	✓		✓			✓	66.0	50.8	34.2	54.4
	✓	✓	✓	✓			**71.1**	46.1	15.6	51.6
	✓	✓	✓		✓		67.1	52.8	33.3	55.7
	✓	✓	✓			✓	66.8	**53.1**	**35.4**	**56.0**

+5.2%

Ablation study on ImageNet-LT

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

- The effectiveness of the **stage III** (Joint training with self-distillation).

Methods	1.5×	I	II	III-hard (test)	III-soft (test)	IV-LWS	Many	Medium	Few	Overall
CE							66.9	38.0	8.1	45.1
	✓						67.9	39.5	9.5	46.3
LWS							61.1	48.0	31.5	50.7
	✓						63.4	48.6	32.3	52.1
Our SSD	✓	✓					69.8	42.8	11.0	48.9
	✓	✓	✓				64.9	51.1	34.0	54.1
	✓		✓			✓	66.0	50.8	34.2	54.4
	✓	✓	✓	✓			**71.1**	46.1	15.6	51.6
	✓	✓	✓		✓		67.1	52.8	33.3	55.7
	✓	✓	✓			✓	66.8	**53.1**	**35.4**	**56.0**

+1.6

Ablation study on ImageNet-LT

- The effectiveness of the **stage IV** (Classifier fine-tuning).

Methods	1.5×	I	II	III-hard (test)	III-soft (test)	IV-LWS	Many	Medium	Few	Overall
CE							66.9	38.0	8.1	45.1
	✓						67.9	39.5	9.5	46.3
LWS							61.1	48.0	31.5	50.7
	✓						63.4	48.6	32.3	52.1
Our SSD	✓	✓					69.8	42.8	11.0	48.9
	✓	✓	✓				64.9	51.1	34.0	54.1
	✓		✓			✓	66.0	50.8	34.2	54.4
	✓	✓	✓	✓			**71.1**	46.1	15.6	51.6
	✓	✓	✓		✓		67.1	52.8	33.3	55.7
	✓	✓	✓		√	✓	66.8	**53.1**	**35.4**	**56.0**

Ablation study on distillation

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

- Study on different self-distillation strategies

  (1) **Coupled** self-distillation which is the conventional way of knowledge distillation and trains **a single classifier** using **both hard and soft** labels;
  (2) **Single** self-distillation, which only use **soft** labels to train the classifier;
  (3) **Our** train **two classifiers** using **hard** and **soft** labels separately.

Methods	Many	Medium	Few	Overall
Plain	67.9	39.5	9.5	46.3
Teacher model	64.9	51.1	**34.0**	54.1
Coupled	68.6	49.1	23.8	53.2
Single	67.4	52.0	31.3	55.1
Our III-hard	**71.1**	46.1	15.6	51.6
Our III-soft	67.1	**52.8**	33.3	**55.7**

+0.6%

Table 5. Top-1 accuracy of different self-distillation strategies on the test set of ImageNet-LT.

Hard labels might be able to provide comple mentary knowledge for feature learning.

Experiments

- Unlike conventional knowledge distillation that uses temperature to **smooth the label distribution of a single image**, we consider taking it to **flatten the data distribution of the entire dataset** by suppressing the frequency of head classes.

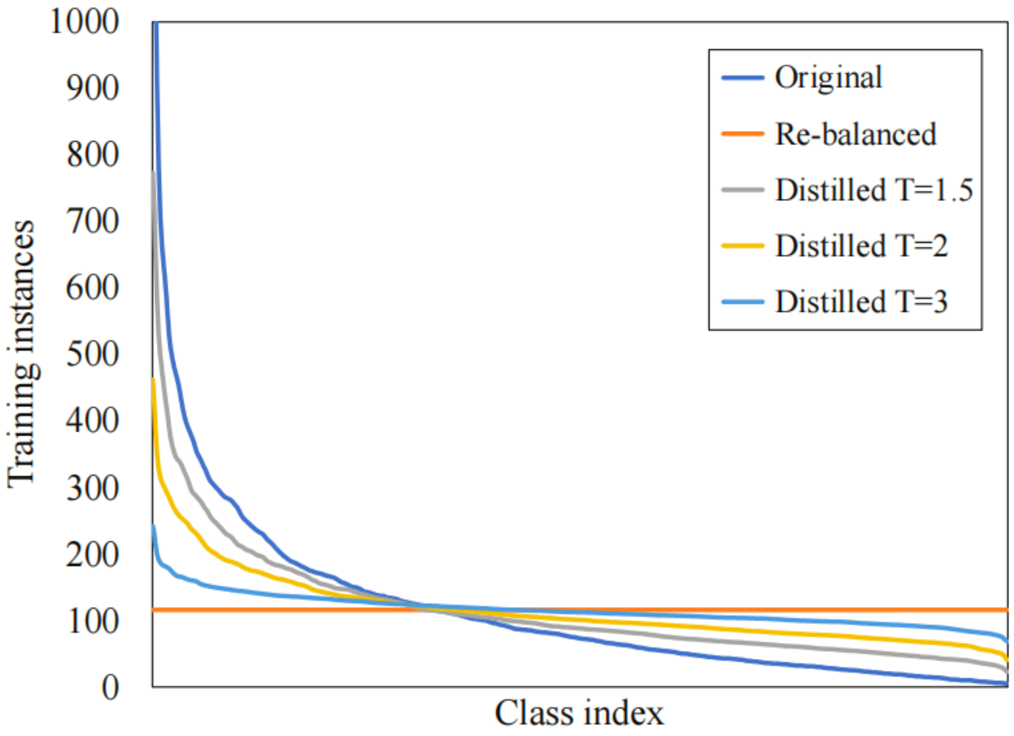


Figure 5. Visualization for different training strategy on ImageNet-LT dataset. *Original*, *Re-balanced* and *Distilled* denote distribution for original long-tailed data, after class-balanced sampling and distilled label.

Experiments

Methods	Many	Medium	Few	Overall
Cross Entropy	65.9	37.5	7.7	44.4
OLTR [28]	-	-	-	46.3
NCM [20]	56.6	45.3	28.1	47.3
cRT [20]	61.8	46.2	27.4	49.6
LWS [20]	60.2	47.2	30.3	49.9
De-confound [35]	62.7	48.8	31.6	51.8
cRT*	62.6	46.9	27.9	50.3
LWS*	61.1	48.0	31.5	50.7
SSD (ours)	64.2 (+3.1)	50.8 (+2.8)	34.5 (+3.0)	53.8 (+3.1)
cRT*‡	64.2	47.7	27.8	51.3
LWS*‡	63.4	48.6	32.3	52.1
SSD (ours)‡	**66.8** (+3.4)	**53.1** (+4.5)	**35.4** (+3.1)	**56.0** (+3.9)

Table 1. Top-1 accuracy on ImageNet-LT dataset. Comparison to the state-of-the-art methods with ResNeXt-50 as backbone. We report absolute improvements against LWS with the same hyper-parameters. * indicates our reproduced results with the released code. Results marked with ‡are trained with 1.5× scheduler.

Experiments

Methods	Imbalance factor		
	100	50	10
Cross Entropy (CE)*	39.1	44.0	55.8
Focal [27]	38.4	44.3	55.8
LDAM-DRW [2]	42.0	46.6	58.7
LWS* [20]	42.3	46.0	58.1
CE-DRW [48]	41.5	45.3	58.2
CE-DRS [48]	41.6	45.5	58.1
BBN [48]	42.6	47.0	59.1
M2m [23]	43.5	-	57.6
LFME [41]	43.8	-	-
Domain Adaption [19]	44.1	49.1	58.0
De-confound [35]	44.1	50.3	59.6
SSD (ours)	**46.0**	**50.5**	**62.3**

Table 2. Top-1 accuracy on CIFAR100-LT dataset with the imbalance factor of 100, 50 and 10. We compare with state-of-the-art methods with ResNet-32 as backbone network. * indicates our reproduced results with the released code.

Experiments

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

Methods	Top-1 Acc.	
	1×	2×
CB-Focal [2]	61.1	-
LDAM [2]	64.6	-
LDAM+DRW [2]	68.0	-
LDAM+DRW† [2]	64.6	66.1
τ-norm‡ [20]	65.6	69.3
cRT‡ [20]	65.2	68.5
LWS‡ [20]	65.9	69.5
CE-DRW [48]	63.7	-
CE-DRS [48]	63.6	-
BBN [48]	66.3	69.6
FSA [6]	65.9	-
LWS‡* [20]	66.6	69.5
SSD (ours)‡	**69.3**	**71.5**

Table 3. Top-1 accuracy on iNaturalist 2018 dataset with 1× and 2× schedulers and comparison to state-of-the-art methods with ResNet-50 as backbone. * indicates our reproduced results. Results marked by † are cited from [48]. 2× means using 200 epochs training scheduler for methods marked by ‡ and 180 epochs for other methods.

THANKS