Decoupled Optimisation for Long-Tailed Visual Recognition

Cong Cong*[1], Shiyu Xuan[2], Sidong Liu[3], Shiliang Zhang[2], Maurice Pagnucco[1], Yang Song[1]

[1]School of Computer Science and Engineering, University of New South Wales, Sydney, Australia
[2]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University, China
[3] Australian Institute of Health Innovation, Macquarie University, Sydney, Australia
z3414050@ad.unsw.edu.au, {morri,yang.song1}@unsw.edu.au, sidong.liu@mq.edu.au, shiyu_xuan@stu.pku.edu.cn, slzhang.jdl@pku.edu.cn

Abstract

When training on a long-tailed dataset, conventional learning algorithms tend to exhibit a bias towards classes with a larger sample size. Our investigation has revealed that this biased learning tendency originates from the model parameters, which are trained to disproportionately contribute to the classes characterised by their sample size (e.g., many, medium, and few classes). To balance the overall parameter contribution across all classes, we investigate the importance of each model parameter to the learning of different class groups, and propose a multistage parameter Decouple and Optimisation (DO) framework that decouples parameters into different groups with each group learning a specific portion of classes. To optimise the parameter learning, we apply different training objectives with a collaborative optimisation step to learn complementary information about each class group. Extensive experiments on long-tailed datasets, including CIFAR100, Places-LT, ImageNet-LT, and iNaturaList 2018, show that our framework achieves competitive performance compared to the state-of-the-art.

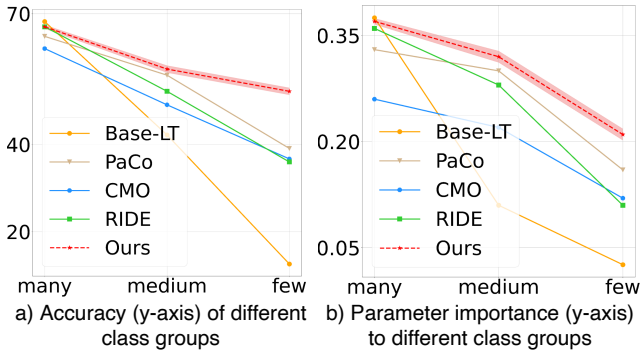


Figure 1: a) Classification accuracies of different class groups comparing the baseline ResNet model, PaCo (Cui et al. 2021), CMO (Park et al. 2022), RIDE (Wang et al. 2020), and our proposed method. b) Averaged parameter importance for different class groups comparing the various approaches. We notice a positive correlation between accuracy and parameter importance. More visualisation results are provided in Supplementary Material.

Introduction

Real-world datasets normally exhibit a long-tailed class distribution, where certain classes possess a large number of samples, while rarer classes are characterised by a limited sample size (Zhang et al. 2021c). Such class imbalances pose a significant challenge when training Deep Convolutional Neural Networks as the head classes with dominant amounts of instances tend to overwhelm the model learning on tail classes by influencing the learning of most of the gradients, consequently leading to subpar performance on minority classes. This is a critical issue, especially in domains like autonomous driving (Yurtsever et al. 2020) and computer-aided diagnostics (Marrakchi, Makansi, and Brox 2021; Cong et al. 2022a,b) where models trained on long-tailed datasets are required to demonstrate high performance across all classes.

Early attempts to alleviate this issue include upsampling rare classes (Zhang et al. 2021b), knowledge transfer (Wang et al. 2021a; Li et al. 2021) and loss re-weighting (Wang et al. 2021c; Ren et al. 2020). Recent works discover that improving feature quality, especially via self-supervised learning (Liu et al. 2021a; Li et al. 2022b), can effectively enhance model performance on imbalanced datasets. Some studies find that feature learning and classifier learning favour different learning strategies, therefore applying a two-stage decoupled learning scheme further improves performance (Kang et al. 2019; Zhou et al. 2020). Moreover, the model ensemble based on multi-expert learning represents the state-of-the-art, with each expert model focusing on distinct partitions of the data distribution (Zhang et al. 2022).

Our analysis shows that each model parameter holds varying degrees of importance in learning different class groups. In long-tailed classification, model parameters tend to be generally more important for the classes with many samples, *i.e.*, the *many* class group. The measurement of the importance of a parameter here is similar to that used during model pruning, *i.e.*, being estimated by the changes in the final loss induced by removing it from the model. Fig. 1 (b) shows the averaged parameter importance across vari-

*Corresponding author

ous *class groups* through different long-tailed learning methods. A positive correlation is found between the accuracy of each class group and the average importance of parameters for that class group. When the naive ResNet model (**Base**) is trained on the long-tailed ImageNet dataset, it exhibits a highly biased parameter importance towards the *many* group, resulting in notably higher accuracy for this group compared to the others. Conversely, the long-tailed methods demonstrate enhanced performance on the *medium* and *few* groups as well as increased parameter importance on these specific class groups.

The aforementioned observations illustrate the advantages of rebalancing parameter importance across class groups to improve imbalanced classification. This motivates us to propose a novel multi-stage optimisation framework aimed at achieving equilibrium in parameter importance across all classes. Specifically, this framework decouples the model parameters into different subsets, each optimised to cater to a specific class group in the training set. In each stage, we first apply a Collaborative Parameter Optimisation (CPO) procedure that is designed to improve parameter importance to a particular group of classes. Following this, we employ a Taylor-guided Parameter Decoupling (TPO) method to select parameters that hold greater importance with regard to the learning of the current preferred class group. Parameters of less importance are then re-initialised and used for optimisation in the subsequent stage. At the end of each stage, a sub-model is constructed, comprising parameters that exhibit high importance to the current stage of learning. In the inference phase, the outputs of these sub-models are aggregated using an Instance-level Test-time learning mechanism to obtain the final prediction. As depicted in Fig. 1, our approach demonstrates a well-balanced parameter importance across all class groups, resulting in a substantial enhancement in performance. Specifically, our contributions are summarised as follows:

- We propose a multi-stage parameter decoupling and optimisation (DO) framework that well balances the importance of parameters across all classes.
- Our framework employs a Collaborative Parameter Optimisation (CPO) procedure that adopts a group-enhanced sampling strategy and a compensation loss to enforce model parameters to learn complementary information about different classes.
- We employ a Taylor-guided Parameter Decoupling (TPD) method that adopts Taylor expansions to approximate the parameter importance and use it to select important parameters for different groups of classes.
- A novel instance-level test time learning algorithm is proposed to obtain more precise predictions when assembling from models with different expertise.
- Extensive experiments on the CIFAR100 (Krizhevsky, Hinton et al. 2009), ImageNet-LT (Liu et al. 2019), Places-LT (Liu et al. 2019), and iNaturalist18 (Van Horn et al. 2018) show that our method achieves superior performance over recent methods with performance improvement in *many*, *medium* and *few* groups.

Related Work

Long-tailed learning aims to train models on datasets that follow a long-tailed class distribution. Existing algorithms can be roughly categorised into *single model imbalance learning* and *multi-expert imbalance learning*.

Single Model Imbalance Learning. These works can be further divided into three subcategories: *re-balancing, knowledge transfer*, and *multi-stage learning*. *Re-balancing*, which enhances the impact of minority classes in the model training procedure, is normally achieved via class resampling (Zhang et al. 2021b) and loss re-weighting (Wang et al. 2021b,c; Ren et al. 2020). These approaches assign higher weights to minority class samples at either the category or instance level. Other studies attempt to transfer knowledge from the majority classes to knowledge-starved minority classes via distribution calibration (Wang et al. 2021a; Liu, Li, and Sun 2022) or augmentation (Chu et al. 2020; Li et al. 2021). *Multi-stage learning* is an effective training scheme for long-tailed classification, as feature learning and classifier learning favour different training strategies (Kang et al. 2019). Self-supervised learning (SSL) has been employed to improve feature quality in previous studies (Li et al. 2022b; Cui et al. 2021), demonstrating that SSL produces more robust features to class imbalance and substantially enhances model performance in long-tailed classification. Other methods have been proposed to improve calibration between the two learning stages. For instance, (Li, Wang, and Wu 2021) implements an extra self-distillation stage to better incorporate label correlation in multi-stage learning. Moreover, Zhang et al. (Zhang et al. 2021a) improve the current two-stage methods using a lightweight distribution alignment module for calibrating the classification scores.

Multi-expert Imbalance Learning. Existing single model approaches reduce model bias for the minority classes but increase the model variance across all classes, leading to decreased accuracy for majority classes (Wang et al. 2020). Therefore, multi-expert imbalance learning frameworks are proposed, e.g., RIDE (Wang et al. 2020), allowing the multiple expert models to capture complementary knowledge. Following this line of research, NCL (Li et al. 2022a) is proposed to enhance knowledge transfer between experts via an online distillation module, SADE (Zhang et al. 2022) explicitly focuses each expert on different data distributions, is employed to fuse experts' outputs using a self-supervised test-time aggregation mechanism. Moreover, SHIKE (Jin et al. 2023) incorporates features from different layers to exploit information encoded at different depths of a network, and BalPoE (Aimar et al. 2023) encourages an unbiased and well-calibrated ensemble via logit adjustment and Mixup.

Continual Learning. Our work draws inspiration from Continual learning, where model parameters are continuously adapted to accommodate non-stationary data distributions. Current approaches can be classified into three categories. *Regularisation*-based methods, which use extra regularization terms, are proposed to strike a balance between the previous and current tasks (Kirkpatrick et al. 2017). While they have shown effectiveness, they may encounter diffi-

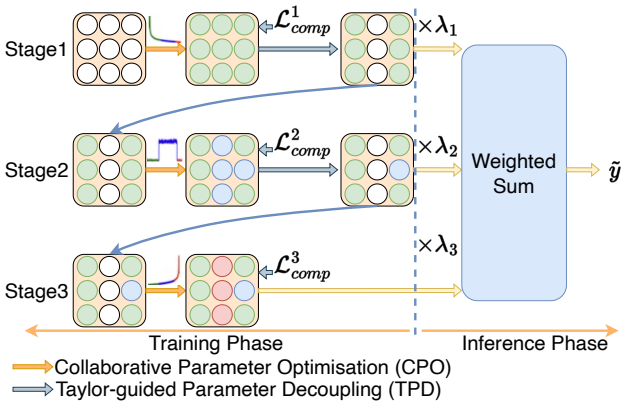


Figure 2: The overall workflow of DO. During training, a multi-stage training schema is used. In each stage, we first apply a CPO step, which strategically targets a set of learnable parameters and trains them to carry complementary information related to a specific group of classes. Then a TPD step is applied to reserve important parameters for the current learning stage and set the remaining parameters as learnable parameters for the subsequent stage. During the inference phase, instance-level test-time learning is used to obtain aggregation weights for fusing the outputs from each sub-model to get the final prediction.

culties when dealing with challenging settings or complex datasets (Mai et al. 2022; Wu et al. 2019). *Rehearsal*-based techniques, on the other hand, use a compact memory buffer or employ an additional generative model to store or generate representative data from previous tasks (Shin et al. 2017). Numerous recent studies have enhanced this concept by integrating knowledge distillation (Chaudhry et al. 2021) or SSL (Pham, Liu, and Hoi 2021). Nevertheless, the applicability of these approaches is generally constrained by the substantial memory requirements. The *Architecture*-based methods focus on constructing task-specific parameters. The model architecture can be fixed (Mallya, Davis, and Lazebnik 2018; Jin and Kim 2022) or dynamic (Hung et al. 2019; Ostapenko et al. 2019) in size when allocating parameters to each task.

Differences from previous works The single model methods show improved performance in minority classes, but might decrease the accuracy for majority classes. While multi-expert models can alleviate this issue, they lack efficient interaction between sub-models and they typically fix an expert's capacity at model initialisation. In contrast, our work innovates upon conventional multiple expert training by dynamically allocating model parameters into sub-groups and explores how to efficiently improve minority classes' performance without compromising the accuracy of majority classes and additionally enhance the parameter interaction. We have incorporated the concept of architecture-based continual learning in our approach. However, our focus is on designing more reliable criteria to quantify parameter importance and balance their importance across all classes in order to enhance long-tailed classification.

Methodology

Since there is a positive correlation between the accuracy of each class group and the corresponding parameter importance, we propose a novel optimisation framework named Decoupled Optimisation (DO) for long-tailed visual recognition to balance parameter importance across class groups explicitly. Given a classification model $\mathcal{F}$ parameterised with θ, we conduct training on a long-tailed dataset $\mathcal{D}$ in T stages and decouple θ into T different groups $\{\theta_1, \theta_2, \cdots, \theta_T\}$ where each focuses on a specific group of classes $\{Y_1, Y_2, \cdots, Y_T\}$.

Fig. 2 shows the workflow of our framework. Following previous long-tail studies (Liu et al. 2019), which define three class groups (*many*, *medium* group, and *few* groups) on $\mathcal{D}$, we set T=3 and start learning from the *many* group in the first stage and then gradually move to the groups with fewer samples. In each stage t, a Collaborative Parameter Optimisation (CPO) process is firstly conducted to encourage θ_t to carry important information in representing the specific class group t. The importance of a parameter $w_i \in \theta_t$ to class y is approximate using the first-order Taylor expansion around w_i, *i.e.*, $\mathcal{E}^y_{w_i} = (w_i g^y_{w_i})^2$, where $g^y_{w_i}$ represents the first-order derivatives with regard to the class y. Once θ_t is optimised in that stage, we apply a Taylor-guided Parameter Decoupling (TPD) method based on parameter importance to decouple θ_t into the important parts $\bar{\theta}_t$ and unimportant parts $\hat{\theta}_t$. Then, $\bar{\theta}_t$ are fixed, and $\hat{\theta}_t$ are re-initialised for further optimisation for other groups of classes, except for the last stage, *i.e.*, $\bar{\theta}_T$=θ_T. To conduct the CPO process in the next stage t, $\bar{\theta}_{t-1}$ from previous stages is also activated. This offers a twofold benefit. Firstly, parameters optimised across distinct stages synergistically interact, amplifying the overall performance. Secondly, the reservoir of knowledge from prior stages contributes to the learning in the present stage. This setting is especially important for learning the classes with fewer samples since they are less represented by the learned parameters.

To optimally balance parameter importance during inference, two operations are adopted: 1) at the end of each training stage, $\{\bar{\theta}_{1\cdots t-1} \cup \bar{\theta}_t\}$ are stored as a sub-model and 2) an instance-level test-time learning algorithm is applied to obtain the aggregation weight λ_t for these T sub-models based on their prediction stability. The final result is the weighted sum of these T sub-models.

Collaborative Parameter Optimisation

The CPO process is proposed to explicitly improve parameter importance about a particular class group. This is done with a group-preferred sampling strategy and optimised with a compensation loss. Let's define N_i as the number of images in i-th class, K as the total number of classes, and $L = [N_i / \sum_{j \in K} N_j : i \in 1 \cdots K]$ is a list containing label frequencies.

Group-preferred sampling strategy. The model changes its learning preferences by altering the sampling ratio per class. In our approach, we use different sampling strategies, $p_t(x, y)$, for each stage, where (x, y) denotes a data sample

x and its corresponding class label y. Specifically, we follow the original long-tailed distribution for data sampling for *many*-preferred learning in Stage 1, *i.e.*, $p_1(x,y)$=$L[y]$ and an inverse long-tailed distribution, *i.e.*, $p_3(x,y)$=$1/N_y$ for sampling during *few*-preferred learning in Stage 3. For the learning of *medium* group classes during Stage 2, we introduce the medium-enhanced ratio ρ_m and define p_2 using the following equation:

$$p_2(x,y) = \begin{cases} \rho_m & y \in Y_2 \\ 1-\rho_m & else \end{cases} \tag{1}$$

Here, ρ_m controls the degree of how strongly we want to enhance the *medium* group learning. Further discussions and insights regarding the impact of ρ_m are provided in the ablation studies.

Compensation loss. The commonly used cross-entropy loss treats each class equally, which may lead to sub-optimal performance when the objective is to emphasise learning from specific categories. To address this limitation, we propose incorporating a compensation term α_t to dynamically enhance the importance of certain model parameters concerning the currently preferred class group.

$$\mathcal{L}^t_{comp} = \frac{1}{n}\sum_{x\in D} -y \log \sigma(\mathcal{F}_{\theta_t}(x) - \log \alpha_t) \tag{2}$$

where σ is the softmax function and α_t alternates between stages according to:

$$\alpha_t = \begin{cases} 1.0 & t=1 \\ \rho_m & t=2 \\ \tau(y) & t=3 \end{cases} \tag{3}$$

Here $\tau(y)$=$\mathcal{R}(L)[y]$ where $\mathcal{R}(\cdot)$ denotes the reverse order operation. The compensation term α_t serves as a margin, exerting a stronger regularisation that encourages the model to prioritise learning on the currently preferred class group.

Taylor-guided Parameter Decoupling

Not all parameters have equal importance to learning, and removing those low-importance parameters may not significantly impact the model's performance (Denil et al. 2013). The importance of a parameter can be estimated by the changes in the loss induced by removing it from the model. We design a Taylor-guided Parameter Decoupling (TPD) method to approximate the importance of a parameter.

Specifically, we rank each parameter w_i based on $\mathcal{E}^y_{w_i}$ and then prune γ of parameters out and leaving the rest as important parameters. We argue that the best pruning ratio γ_{best} should generate the most compact model while maintaining its performance. Thus, we iterate values from $\gamma_i = i$ for $i \in 0 \cdots 90$ with a step of 10 and record the associated performance. γ_{best} is selected based on the highest variation observed in the recorded performances. Moreover, instead of directly pruning γ_i parameters out, we prune gradually with γ_t with gradual pruning (Zhu and Gupta 2017):

$$\gamma_t = \gamma_{i+1} + (\gamma_i - \gamma_{i+1})(1 - \frac{t}{\Delta t})^3 \tag{4}$$

where γ_t increased from γ_i to γ_{i+1} with a step of Δt. After each pruning step involving γ_t, we will proceed by retraining the model for a single iteration, ensuring that its discriminatory capabilities for the task are preserved. In our experiment, we noticed that varying Δt from 10 to 100 had a negligible impact on the performance, thus we set $\Delta t = 10$. Once the γ_{best} is selected, we prune out γ_{best} of the total parameters and set them as unimportant parameters $\hat{\theta}_t$, which are used for further optimisation and decoupling, whereas the retained parameters are regarded as important parameters $\bar{\theta}_t$ which will remain fixed.

Instance-level Test-time Learning

After training, we have T sub-models, each containing a subset of parameters ($\bar{\theta}^*_t$=$\{\bar{\theta}_{1\cdots t-1} \cup \bar{\theta}_t\}$) that pose high importance to each class group. Since $\bar{\theta}_{1\cdots t-1}$ are activated during the optimisation of $\bar{\theta}_t$, thus both are included. During inference, to optimally balance the parameter importance across different classes, we use an instance-level test-time self-supervised learning method to generate aggregation weights (λ_t) for each sub-model $f_{\bar{\theta}^*_t}$, based on maximising prediction stability. This is inspired by (Zhang et al. 2022), which highlights the positive correlation between model expertise and prediction stability. However, they generate λ_t on a group level, *i.e.*, λ=$\{\lambda_t\}^T_{t=1}$. We argue that such coarse-grained λ can only partially reflect the model's stability across different classes. Therefore, in our approach, we apply aggregation on the instance level, *i.e.*, λ=$\{\{\lambda_t\}^T_{t=1}\}^U_{i=1}$, where U denotes the number of test samples.

Specifically, given an input test image x, we conduct two stochastic data augmentations to produce two views of x, denoted as x_1 and x_2. We then obtain the corresponding predictions $\tilde{y_1}$=$\sum^T_{t=1} \lambda_t f_{\bar{\theta}^*_t}(x_1)$ and $\tilde{y_2}$=$\sum^T_{t=1} \lambda_t f_{\bar{\theta}^*_t}(x_2)$. Our objective is to maximise the cosine similarity between the predictions from these two views using

$$\lambda = arg\max_{\lambda} \tilde{y_1}^T \tilde{y_2} \tag{5}$$

Note that λ=$[\lambda_1, \cdots, \lambda_t]$ are the only learnable hyper-parameters in these functions. By maximising the cosine similarity between the two predictions, the corresponding λ_t with respect to $f_{\bar{\theta}^*_t}$ (which demonstrates more stable predictions for samples from specialised classes) will be learned to increase. Consequently, these learned λ_t can effectively reflect the confidence of $f_{\bar{\theta}^*_t}$ in predicting an unseen sample. The higher the stability of predictions for a particular sub-model $f_{\bar{\theta}_t}$ on a given class, the more the corresponding aggregation weight λ_t will be emphasised during the inference process, leading to a more reliable and accurate overall prediction for that class.

Experiments

Datasets

ImageNet-LT (Liu et al. 2019) is a long-tailed version of ImageNet (Deng et al. 2009). It was generated by sampling a subset with the Pareto distribution using a power value $\alpha = 6$. It contains 115.8K images from 1,000 categories in which the class cardinality ranges from 5 to 1,280.

Method	Many	Medium	Few	All
Single Model				
CE (baseline)	68.2	42.2	12.6	48.2
LWS	61.8	47.6	30.9	50.8
BSCE	64.1	48.2	33.4	52.3
MiSLAS	62.0	49.1	32.8	51.4
LADE	64.4	47.7	34.3	52.3
CMO	62.0	49.1	36.7	52.3
RSG	63.2	48.2	32.3	51.8
PaCo	64.8	55.9	39.1	57.0
TSC	63.5	49.7	30.4	52.4
GCL	-	-	-	54.9
CC-SAM	61.4	49.5	37.1	52.4
Multi-Expert Model				
RIDE	**67.0**	52.2	36.0	55.7
ACE	-	-	-	54.7
SHIKE	-	-	-	59.7
NCL	-	-	-	59.5
SADE	66.5	57.0	43.5	58.8
BalPoE	-	-	-	59.7
DO (ours)	**67.0**	**57.3**	**52.2**	**60.4**

Table 1: ImageNet-LT test accuracy (%) comparisons.

Method	Many	Medium	Few	All
Single Model				
CE (baseline)	78.8	68.3	55.4	64.5
LWS	71.0	69.8	68.8	69.5
BSCE	70.9	70.4	70.1	70.6
MiSLAS	71.5	69.7	70.7	71.7
LADE	68.7	70.2	69.3	68.9
CMO	68.8	70.0	72.3	70.9
PaCo	70.3	73.2	73.6	73.2
TSC	70.6	67.8	69.7	72.6
GCL	-	-	-	72.0
CC-SAM	65.4	70.9	72.2	70.9
Multi-Expert Model				
RIDE	70.0	71.7	71.8	71.5
ACE	-	-	-	72.9
SHIKE	-	-	-	75.4
NCL	-	-	-	74.9
SADE	75.5	**73.7**	75.1	74.5
BalPoE	-	-	-	73.5
DO (ours)	**77.1**	73.6	**75.6**	**75.8**

Table 2: iNaturalist2018 test accuracy (%) comparsions.

Method	Many	Medium	Few	All
Single Model				
CE (baseline)	46.2	27.5	12.7	31.4
BSCE	42.6	39.8	32.7	39.4
MiSLAS	41.6	39.3	27.5	37.6
LADE	42.6	39.4	32.3	39.2
CC-SAM	-	-	-	40.6
PaCo	36.1	**47.9**	35.3	41.2
Multi-Expert Model				
RIDE	43.1	41.0	33.0	40.3
SHIKE	43.6	39.2	**44.8**	41.9
NCL	-	-	-	41.8
SADE	40.4	43.2	36.8	40.9
DO (ours)	**43.7**	43.2	40.1	**42.8**

Table 3: Places-LT test accuracy (%) comparsions.

iNaturalist2018 (Van Horn et al. 2018) is a large-scale species classification dataset. It contains 8,142 classes which suffer from severe class imbalance issues with class cardinality ranging from 5 to 4,980.

CIFAR100-LT (Krizhevsky, Sutskever, and Hinton 2012) has 60,000 images, where 50,000 are used for training and 10,000 for validation. This work used a long-tailed version of CIFAR100 where the imbalance ratio (β) is manually selected using β=$\frac{N_{max}}{N_{min}}$ where N_{max} and N_{min} are the numbers of instances for the most and least frequent classes.

Places-LT (Liu et al. 2019) is a long-tailed version of the original Places-2 (Zhou et al. 2017), which contains 184.5K images which come from a total of 365 categories where the class cardinality ranges from 5 to 4,980.

Implementation Details

We use ResNet50 (He et al. 2016) for experiments on ImageNet-LT and iNaturalist2018, ResNet152 on Places-LT, and ResNet32 for experiments on CIFAR100-LT. For Stage1, we conduct training for 100 epochs and decay the learning rate by a cosine scheduler from 0.02 to 0 for ImageNet-LT, iNaturalist2018 and Places-LT, and 0.05 to 0 for CIFAR100-LT. For the remaining two stages, since we only fine-tune part of the model, we only train for 50 epochs and the learning rate is equal to 0.002 for ImageNet-LT, iNaturalist2018, and Places-LT, and 0.005 for CIFAR100-LT. All pieces of training are conducted with a batch size of 256. In all reported experiments, we use strong augmentations (Cubuk et al. 2020) that have demonstrated effectiveness in previous studies (Cui et al. 2021). All reported models are trained using 4 NVIDIA Tesla V100 GPUs.

The ρ_m in *medium*-enhanced sampling is set to 80% for ImageNet-LT, iNaturalist2018 and Places-LT, and 70% for CIFAR100-LT. To select γ_{best}, we iterate through ten possible values from 0% to 100%, with a step size of 10%. The γ_{best} for Stage 1 is set to 50% for all used datasets, whereas, for Stage 2, the γ_{best} is set to 80% for ImageNet-LT, Places-LT, and CIFAR100 (β=100) and 60% for iNaturalist2018 and CIFAR100 (β=50). We tune parameters on the validation set and report the test set results for ImageNet-LT. For the other datasets that only have train-val sets, the same validation set is used for tuning and benchmarking.

Comparison to the Prior Art

We compared DO with previous state-of-the-art methods. We show the results on ImageNet-LT (Tab. 1), iNaturalist2018 (Tab. 2), CIFAR100 (Tab. 4 and Tab. 5) and Places-LT (Tab. 3). For all compared methods, we report their performance with strong augmentation if used in their works. Specifically, we chose single model-based (*SE*) approaches (e.g., loss reweight (BSCE (Ren et al. 2020), LADE (Hong et al. 2021), GCL (Li, Cheung, and Lu 2022)), knowledge transfer (CMO (Park et al. 2022), RSG (Wang et al. 2021a)), decouple-based methods (MiSLAS (Zhong et al. 2021), WB (Alshammari et al. 2022)) and feature learning (PaCo (Cui et al. 2021),(Li et al. 2022b))) and multi-expert (*ME*) meth-

Method	Many	Medium	Few	All
Single Model				
CE (baseline)	66.8	37.4	15.5	45.6
BSCE	62.1	45.6	36.7	50.9
MiSLAS	61.8	48.9	33.9	51.5
WB	-	-	-	57.5
LADE	60.2	46.2	35.6	50.1
GCL	-	-	-	53.6
CC-SAM	-	-	-	53.9
Multi-Expert Model				
RIDE	**66.6**	46.2	30.3	51.7
NCL	-	-	-	58.2
SADE	61.5	50.2	45.0	53.9
BalPoE	-	-	-	**58.7**
DO (ours)	66.1	**56.1**	**52.2**	58.2

Table 4: CIFAR100-LT$_{\beta=50}$ test accuracy (%) comparsions.

Method	Many	Medium	Few	All
Single Model				
CE (baseline)	68.6	41.1	9.6	41.4
BSCE	64.1	48.2	33.4	50.8
MiSLAS	60.4	49.6	26.6	46.8
WB	**71.4**	**51.2**	35.3	53.8
LADE	58.7	45.8	29.8	45.6
GCL	-	-	-	48.7
CC-SAM	-	-	-	50.8
Multi-Expert Model				
RIDE	67.4	49.5	23.7	48.0
ACE	66.1	55.7	23.5	49.4
NCL	-	-	-	54.2
SADE	65.4	49.4	29.3	49.8
BalPoE	-	-	-	**54.7**
DO (ours)	67.5	47.8	**44.6**	53.8

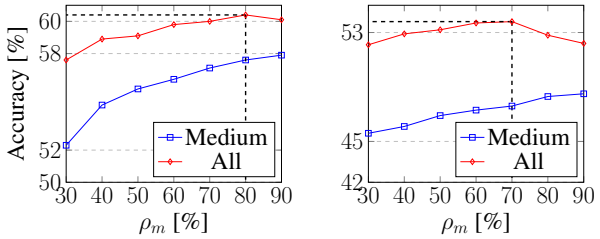
Table 5: CIFAR100-LT$_{\beta=100}$ test accuracy (%) comparsions.

ods (RIDE (Wang et al. 2020), ACE (Cai, Wang, and Hwang 2021), SADE (Zhang et al. 2022), NCL (Li et al. 2022a), SHIKE (Jin et al. 2023) and BalPoE (Aimar et al. 2023)).

DO outperforms the existing *SE* methods; for example, it achieves 3.4% and 2.6% improvements over PaCo on ImageNet-LT and iNaturalist2018, respectively. Moreover, DO is only trained for 200 epochs, which is much less than the 400-epoch training used in contrastive learning-based methods. Compared with *ME* methods, DO demonstrates state-of-the-art performance on three particularly challenging datasets (60.4% on ImageNet-LT, 75.8% on iNaturalist2018, and 42.8% Places-LT) and achieves competitive results on the CIFAR100-LT dataset (58.2% β=50 and 53.8% β=100) However, most *ME* methods use three or more complete networks as their experts during inference, whereas our DO employs sub-models that utilise only a subset of the entire model parameters, which not only reduces computational overhead but also achieves comparable or even superior performance.

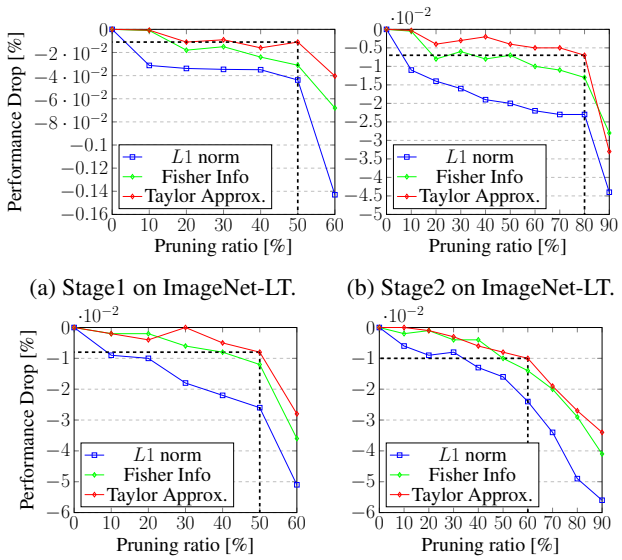
Discussion & Ablations

Sensitivity analysis of ρ_m and γ. ρ_m controls the degree of how much the model concentrates on the *medium* group learning. Fig. 3 shows that increasing ρ_m improves *medium* group performance as more classes from the *medium* group can be sampled. However, improving ρ_m also suppresses the learning in the other two groups. Thus, the degree of improvement becomes marginal with large ρ_m.



(a) ρ_m on ImageNet-LT. (b) ρ_m on CIFAR100-LT.

Figure 3: Senstivity analysis of medium enhanced sampling ratio ρ_m.



(a) Stage1 on ImageNet-LT. (b) Stage2 on ImageNet-LT.

(c) Stage1 on iNaturalist2018. (d) Stage2 on iNaturalist2018.

Figure 4: Performance Drop vs. different pruning ratio with different pruning methods. γ_{best} is selected where the maximum variation in performance drop is observed.

We show the performance changes regarding different pruning ratios in different stages in Fig. 4. When the pruning ratio is small, the performance drop is also minor. This implies that the model might be over-parameterised, and removing a partition of parameters will not significantly impact the performance. Moreover, a marked drop in performance is observed with a larger pruning ratio, e.g., 50% in Stage 1 and 80% in Stage 2 for ImageNet-LT. Thus, we set this value as γ_{best} as it produces the most compact model with the lowest performance drop. In addition, we compare different parameter importance estimation methods: $L1$ norm (Hung et al. 2019), Taylor approximation (Xia et al. 2020), and Fisher information (Liu et al. 2021b). Among them, the $L1$ norm has the largest performance drop. In con-

	Many	Medium		Few	Avg
2-stage	66.5	47.8		50.4	50.3
3-stage	67.0	57.3		52.2	60.4
	Many	Medium-top	Medium-few	Few	Avg
4-stage	66.0	59.8	56.4	53.0	60.8

Table 6: ImageNet-LT test accuracy (%) comparison with different numbers of stages.

Stage Order	Many→ Few→ Med			Many→ Med→ Few		
	Many	Med	Few	Many	Med	Few
Stage1-CPO	68.2	42.2	12.6	68.2	42.2	12.6
Stage1-TPD	67.7	40.5	10.1	67.7	40.5	10.1
Stage2-CPO	24.4	42.3	51.8	31.5	57.6	14.4
Stage2-TPD	32.7	47.1	51.7	45.0	57.4	29.5
Stage3-CPO	52.2	53.3	33.9	25.9	42.1	54.5
Overall	67.1	53.6	51.0	67.0	57.3	52.2

Table 7: ImageNet-LT test accuracy (%) comparison with different stage orders.

Agg_{int}	Agg_{gp}	Agg_{avg}	S_{uni}	S_{med}	$\mathcal{L}_{comp}$	$\mathcal{L}_{ce}$	Acc
			✓			✓	52.4
	✓		✓			✓	53.6
✓			✓			✓	55.8
✓				✓		✓	57.6
		✓		✓	✓		57.9
✓				✓	✓		60.4

Table 8: Ablation studies on ImageNet-LT. "$Agg_{int/gp/avg}$": the instance-level test-time aggregation, group-level test-time aggregation, and average aggregation. "$S_{uni/med}$": uniform or enhanced sampling for *medium* group and "$L_{comp/ce}$": the compensation or cross-entropy loss.

	Many	Medium	Few
Stage 1 Params (λ_1)	0.82	0.21	0.09
Stage 2 Params (λ_2)	0.16	0.73	0.18
Stage 3 Params (λ_3)	0.02	0.10	0.69

Table 9: The averaged λ_t for instances of different groups.

trast, the two gradient-based methods yield more reliable estimations. While the Fisher estimation assumes that the importance of all neurons is strictly positive (which is only sometimes true, as indicated in (Molchanov et al. 2019)), the Taylor approximation explicitly estimates changes in the loss and proves to be a better importance estimator.

Influence of stage selection. In Tab. 6, we list two configurations of the proposed framework: two-stage and four-stage. The former one only conducts *many* and *few* group learning. It achieves good performance for *many* and *few* groups, but the *medium* group still requires further improvements. For the four-stage configuration, we evenly divide the *medium* group into *Medium*-top (100∼50 samples) and *Medium*-low (50∼20 samples), respectively. This requires longer training but only brings marginal enhancement.

The stage order also matters, as shown in Tab. 7. Besides the default setting, we conducted a *few* group first learning, which significantly improves the *medium* group accuracy. In contrast, our default setting shows further improvements in both *medium* and *few* groups. This might be attributable to two reasons. Firstly, Stage 2 always has more free parameters, which can improve *medium* group performance. Secondly, the parameters learned from *many* and *medium* classes embed much information, which may facilitate the learning on the *few* classes.

Effectiveness of *Medium*-preferred sampling strategy and $\mathcal{L}_{comp}$ As shown in Tab. 8, replacing the uniform sampling with the enhanced sampling strategy increases performance from 55.8% to 57.6%, indicating that enhancing the *medium* group learning is useful. Furthermore, changing the loss function from cross-entropy to compensation loss is important as it encourages each parameter group to focus on learning complementary information. The best performance is achieved by combining all proposed components.

Effectiveness of instance-level test-time Learning. As shown in Tab. 8, test-time learning is beneficial for performance improvement. Instance-level test-time learning strategy can approximately increase the performance by $2\%\sim3\%$, compared to the group-level aggregation strategy (Zhang et al. 2022). Moreover, results in Tab. 9 show our instance-level test-time learning strategy learns suitable weights for sub-models with different expertise. For the sub-model with weights learned from Stage 1, λ_1 for the *many* groups is higher, while for the sub-model with weights learned from Stage 3, λ_3 for the *few* groups is higher. To better understand the trade-off between performance and computational costs, in Tab. 10, we compare the test-time cost with PaCo (a method without test-time learning) and SADE (a method with test-time learning). The results show that our instance-level test-time learning incurs higher computational overhead but leads to substantially superior results. To address the challenge of high computational costs, we propose an alternative approach where we replace test-time learning (TTL) by averaging outputs from all sub-models ($\text{Ours}_{\text{wo TTL}}$). This version still delivers competitive performance while significantly speeding up the inference process. This outcome underscores the efficacy of balancing weight importance in mitigating the long-tailed issue.

Conclusion

We proposed a novel parameter decoupled and optimisation framework for long-tailed visual recognition in this work. The proposed framework optimally balances the parameter importance across all classes by decoupling the model parameters into different groups in which each is optimised for a separate class partition. Extensive experiments have demonstrated the effectiveness of the proposed framework.

	PaCo	SADE	$\text{Ours}_{\text{wo TTL}}$	Ours
Per-epoch Time (s)	90	253	95	268
Acc (%)	57.0	58.5	57.9	60.4

Table 10: Evaluation of run-time cost.

Acknowledgments

This work is supported in part by the Natural Science Foundation of China under Grant No. U20B2052, 61936011, in part by the Okawa Foundation Research Award.

References

Aimar, E. S.; Jonnarth, A.; Felsberg, M.; and Kuhlmann, M. 2023. Balanced Product of Calibrated Experts for Long-Tailed Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19967–19977.

Alshammari, S.; Wang, Y.-X.; Ramanan, D.; and Kong, S. 2022. Long-tailed recognition via weight balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6897–6907.

Cai, J.; Wang, Y.; and Hwang, J.-N. 2021. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 112–121.

Chaudhry, A.; Gordo, A.; Dokania, P.; Torr, P.; and Lopez-Paz, D. 2021. Using hindsight to anchor past knowledge in continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 6993–7001.

Chu, P.; Bian, X.; Liu, S.; and Ling, H. 2020. Feature space augmentation for long-tailed data. In *European Conference on Computer Vision*, 694–710. Springer.

Cong, C.; Yang, Y.; Liu, S.; Pagnucco, M.; Di Ieva, A.; Berkovsky, S.; and Song, Y. 2022a. Adaptive Unified Contrastive Learning for Imbalanced Classification. In *Machine Learning in Medical Imaging: 13th International Workshop, MLMI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*, 348–357. Springer.

Cong, C.; Yang, Y.; Liu, S.; Pagnucco, M.; and Song, Y. 2022b. Imbalanced Histopathology Image Classification Using Deep Feature Graph Attention Network. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 1–4.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, 702–703.

Cui, J.; Zhong, Z.; Liu, S.; Yu, B.; and Jia, J. 2021. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 715–724.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. Ieee.

Denil, M.; Shakibi, B.; Dinh, L.; Ranzato, M.; and De Freitas, N. 2013. Predicting parameters in deep learning. *Advances in Neural Information Processing Systems*, 26.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hong, Y.; Han, S.; Choi, K.; Seo, S.; Kim, B.; and Chang, B. 2021. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6626–6636.

Hung, C.-Y.; Tu, C.-H.; Wu, C.-E.; Chen, C.-H.; Chan, Y.-M.; and Chen, C.-S. 2019. Compacting, picking and growing for unforgetting continual learning. *Advances in Neural Information Processing Systems*, 32.

Jin, H.; and Kim, E. 2022. Helpful or Harmful: Inter-task Association in Continual Learning. In *European Conference on Computer Vision*, 519–535. Springer.

Jin, Y.; Li, M.; Lu, Y.; Cheung, Y.-m.; and Wang, H. 2023. Long-Tailed Visual Recognition via Self-Heterogeneous Integration with Knowledge Excavation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23695–23704.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2019. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25.

Li, J.; Tan, Z.; Wan, J.; Lei, Z.; and Guo, G. 2022a. Nested Collaborative Learning for Long-Tailed Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6949–6958.

Li, M.; Cheung, Y.-m.; and Lu, Y. 2022. Long-tailed visual recognition via gaussian clouded logit adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6929–6938.

Li, S.; Gong, K.; Liu, C. H.; Wang, Y.; Qiao, F.; and Cheng, X. 2021. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5212–5221.

Li, T.; Cao, P.; Yuan, Y.; Fan, L.; Yang, Y.; Feris, R. S.; Indyk, P.; and Katabi, D. 2022b. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6918–6928.

Li, T.; Wang, L.; and Wu, G. 2021. Self supervision to distillation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 630–639.

Liu, H.; HaoChen, J. Z.; Gaidon, A.; and Ma, T. 2021a. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*.

Liu, J.; Li, W.; and Sun, Y. 2022. Memory-based jitter: Improving visual recognition on long-tailed data with diversity in memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1720–1728.

Liu, L.; Zhang, S.; Kuang, Z.; Zhou, A.; Xue, J.-H.; Wang, X.; Chen, Y.; Yang, W.; Liao, Q.; and Zhang, W. 2021b. Group fisher pruning for practical network compression. In *International Conference on Machine Learning*, 7021–7032. PMLR.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2537–2546.

Mai, Z.; Li, R.; Jeong, J.; Quispe, D.; Kim, H.; and Sanner, S. 2022. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469: 28–51.

Mallya, A.; Davis, D.; and Lazebnik, S. 2018. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European Conference on Computer Vision*, 67–82.

Marrakchi, Y.; Makansi, O.; and Brox, T. 2021. Fighting class imbalance with contrastive learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 466–476. Springer.

Molchanov, P.; Mallya, A.; Tyree, S.; Frosio, I.; and Kautz, J. 2019. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11264–11272.

Ostapenko, O.; Puscas, M.; Klein, T.; Jahnichen, P.; and Nabi, M. 2019. Learning to remember: A synaptic plasticity driven framework for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11321–11329.

Park, S.; Hong, Y.; Heo, B.; Yun, S.; and Choi, J. Y. 2022. The Majority Can Help The Minority: Context-rich Minority Oversampling for Long-tailed Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6887–6896.

Pham, Q.; Liu, C.; and Hoi, S. 2021. Dualnet: Continual learning, fast and slow. *Advances in Neural Information Processing Systems*, 34: 16131–16144.

Ren, J.; Yu, C.; Ma, X.; Zhao, H.; Yi, S.; et al. 2020. Balanced meta-softmax for long-tailed visual recognition. *Advances in Neural Information Processing Systems*, 33: 4175–4186.

Shin, H.; Lee, J. K.; Kim, J.; and Kim, J. 2017. Continual learning with deep generative replay. *Advances in Neural Information Processing Systems*, 30.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8769–8778.

Wang, J.; Lukasiewicz, T.; Hu, X.; Cai, J.; and Xu, Z. 2021a. Rsg: A simple but effective module for learning imbalanced datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3784–3793.

Wang, J.; Zhang, W.; Zang, Y.; Cao, Y.; Pang, J.; Gong, T.; Chen, K.; Liu, Z.; Loy, C. C.; and Lin, D. 2021b. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9695–9704.

Wang, T.; Zhu, Y.; Zhao, C.; Zeng, W.; Wang, J.; and Tang, M. 2021c. Adaptive class suppression loss for long-tail object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3103–3112.

Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; and Yu, S. X. 2020. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*.

Wu, Y.; Chen, Y.; Wang, L.; Ye, Y.; Liu, Z.; Guo, Y.; and Fu, Y. 2019. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 374–382.

Xia, X.; Liu, T.; Han, B.; Gong, C.; Wang, N.; Ge, Z.; and Chang, Y. 2020. Robust early-learning: Hindering the memorization of noisy labels. In *International Conference on Learning Representations*.

Yurtsever, E.; Lambert, J.; Carballo, A.; and Takeda, K. 2020. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8: 58443–58469.

Zhang, S.; Li, Z.; Yan, S.; He, X.; and Sun, J. 2021a. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2361–2370.

Zhang, X.; Wu, Z.; Weng, Z.; Fu, H.; Chen, J.; Jiang, Y.-G.; and Davis, L. S. 2021b. Videolt: Large-scale long-tailed video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7960–7969.

Zhang, Y.; Hooi, B.; Hong, L.; and Feng, J. 2022. Self-Supervised Aggregation of Diverse Experts for Test-Agnostic Long-Tailed Recognition. In *Advances in Neural Information Processing Systems*.

Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; and Feng, J. 2021c. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*.

Zhong, Z.; Cui, J.; Liu, S.; and Jia, J. 2021. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16489–16498.

Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9719–9728.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6): 1452–1464.

Zhu, M.; and Gupta, S. 2017. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*.