Long-Tailed Visual Recognition via Self-Heterogeneous Integration with Knowledge Excavation

Yan Jin[1,2] Mengke Li[3] Yang Lu[1,2*] Yiu-ming Cheung[4] Hanzi Wang[1,2]

[1] Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen, China

[2] Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, Xiamen, China

[3] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China

[4] Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China

jinyan7973@gmail.com limengke@gml.ac.cn {luyang, Hanzi.Wang}@xmu.edu.cn ymc@comp.hkbu.edu.hk

Abstract

Deep neural networks have made huge progress in the last few decades. However, as the real-world data often exhibits a long-tailed distribution, vanilla deep models tend to be heavily biased toward the majority classes. To address this problem, state-of-the-art methods usually adopt a mixture of experts (MoE) to focus on different parts of the long-tailed distribution. Experts in these methods are with the same model depth, which neglects the fact that different classes may have different preferences to be fit by models with different depths. To this end, we propose a novel MoE-based method called Self-Heterogeneous Integration with Knowledge Excavation (SHIKE). We first propose Depth-wise Knowledge Fusion (DKF) to fuse features between different shallow parts and the deep part in one network for each expert, which makes experts more diverse in terms of representation. Based on DKF, we further propose Dynamic Knowledge Transfer (DKT) to reduce the influence of the hardest negative class that has a non-negligible impact on the tail classes in our MoE framework. As a result, the classification accuracy of long-tailed data can be significantly improved, especially for the tail classes. SHIKE achieves the state-of-the-art performance of 56.3%, 60.3%, 75.4% and 41.9% on CIFAR100-LT (IF100), ImageNet-LT, iNaturalist 2018, and Places-LT, respectively. The source code is available at https://github.com/jinyan-06/SHIKE.

1. Introduction

Deep learning has made incredible progress in visual recognition tasks during the past few years. With well-designed models, e.g., ResNet [18], and Transformer [57], deep learning techniques have outperformed humans in many visual applications, like image classification [32], semantic segmentation [17, 42], and object detection [50, 52]. One key factor in the success of deep learning is the availability of large-scale datasets [13, 55, 70], which are usually manually constructed and annotated with balanced training samples for each class. However, in real-world applications, data typically follows a long-tailed distribution, where a small fraction of classes possess massive samples, but the others are with only a few samples [5, 12, 26, 41, 44]. Such imbalanced data distribution leads to a significant accuracy drop for deep learning models trained by empirical risk minimization (ERM) [56] as the model tends to be biased towards the head classes and ignore the tail classes to a great extent. Thus, the model's generalization ability on tail classes is severely degraded.

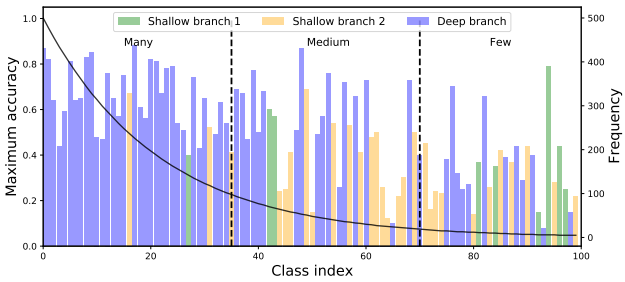


Figure 1. Comparison of test accuracy of a ResNet-32 model with two shallow branches and a deep branch. The model is jointly trained on CIFAR100-LT with an imbalance factor of 100. Only the highest accuracy among the three branches is shown for each class.

The most straightforward action for long-tailed recognition often focuses on re-balancing the learning process from either a data processing [3, 27, 46] or cost-sensitive perspective [12, 28, 71]. Recently, methods proposed for

*Corresponding author: Yang Lu, luyang@xmu.edu.cn

long-tailed data have drawn more attention to representation learning. For example, the decoupling strategy [27] is proposed to deal with the inferior representation caused by re-balancing methods. Contrastive learning [11,26] specializes in learning better and well-distributed representations. Among them, the methods that achieve state-of-the-art performance are usually based on a mixture of experts (MoE), also known as multi-experts. Some MoE-based methods prompt different experts to learn different parts of the long-tailed distribution (head, medium, tail) [4, 10, 39, 61], while some others were designed to reduce the overall model's prediction variance or uncertainty [33, 60].

Unlike traditional ensemble learning methods that adopt independent models for joint prediction, the MoE-based methods for long-tailed learning often adopt a multi-branch model architecture with shared shallow layers and exclusive deep layers. Thus, the features generated by different experts are actually from the model with the same depth, although the methods force them to be diverse from various perspectives. Recently, self-distillation [64] is proposed to enable shallow networks to have the ability to predict certain samples in the data distribution. This brings us to a new question: can we integrate the knowledge from shallow networks into some experts in MoE to fit the long-tailed data in a self-adaptive manner regardless of the number of samples? With this question, we conduct a quick experiment to reveal the preference of the deep neural network on different classes in long-tailed data. A ResNet-32 model with branches directly from shared layers is adopted. Each branch contains an independent classifier after feature alignment, and all classifiers are re-trained with balanced softmax cross entropy [49]. Fig. 1 shows the highest accuracy among the three branches for each class. We can clearly observe that shallow parts of the deep model are able to perform better on certain tail classes. This implies that different parts of the long-tailed distribution might accommodate the network differently according to the depth. Thus the shallow part of the deep model can provide more useful information for learning the long-tailed distribution.

Driven by the observation above, we propose a novel MoE-based method called Self-Heterogeneous Integration with Knowledge Excavation (SHIKE). SHIKE adopts an MoE-based model consisting of heterogeneous experts along with knowledge fusion and distillation. To fuse the knowledge diversely, we first introduce Depth-wise Knowledge Fusion (DKF) as a fundamental component to incorporate different intermediate features into deep features for each expert. The proposed DKF architecture can not only provide more informative features for experts but also optimize more directly to shallower layers of the networks by mutual distillation. In addition, we design Dynamic Knowledge Transfer (DKT) to address the problem of the hardest negatives during knowledge distillation between experts. DKT elects the non-target logits with large values to reform non-target predictions from all experts to one grand teacher, which can be used in distilling non-target predictions to suppress the hardest negative, especially for the tail classes. DKT can fully utilize the structure of MoE and diversity provided by DKF for better model optimization. In this paper, our contributions can be summarized as follow:

- We propose Depth-wise Knowledge Fusion (DKF) to encourage feature diversity in knowledge distillation among experts, which releases the potential of the MoE in long-tailed representation learning.
- We propose a novel knowledge distillation strategy DKT for MoE training to address the hardest negative problem for long-tailed data, which further exploits the diverse features fusing enabled by DKF.
- We outperform other state-of-the-art methods on four benchmarks by achieving performance 56.3%, 60.3%, 75.4% and 41.9% accuracy for CIFAR100-LT (IF100), ImageNet-LT, iNaturalist 2018, and Places-LT, respectively.

2. Related work

Long-tailed Visual Recognition. Long-tailed visual recognition aims at improving the accuracy of the tail classes with the least influence on the head classes. Resampling is the most common practice in early methods for long-tailed learning, which mainly focuses on balancing the data distribution during model training [6, 27, 43, 59, 63]. In terms of model optimization, re-weighting aims to rebalance classes in the way of adjusting loss value for different classes during training [15, 23, 40, 47, 49, 65, 71]. Data augmentation enables balanced training by means of either transferring the information from the head classes to the tail classes [29, 58] or generating data for the tail classes using prior [63] or estimated statistics [37].

Some other methods adopt logit adjustment to calibrate data distribution by post-hoc shifting the logit based on label frequencies in order to obtain a large margin between classes [23, 35, 36, 44, 48, 65]. As the re-balancing methods usually promote the accuracy of tail classes at the cost of harming the head classes, decoupled training [27] and contrastive learning methods [11, 72] are proposed to learn the better representation for long-tailed learning.

More recently, methods based on the mixture of experts (MoE) have been explored to improve performance by integrating more than one model in the learning framework. The basic idea is to make the experts focus on different parts of the long-tailed data. BBN [69] is proposed to use a two-branched to learn the long-tailed distribution and the balanced distribution simultaneously along with a smooth

transition between them. BAGS [39] and LFME [61] reduce the related imbalance ratio by dividing the long-tailed distribution into several sub-groups with several experts fitting on them. ACE [4] and ResLT [10] allow experts to be skilled at different parts of the long-tailed distribution and to complement each other. RIDE [60] and TLC [33] utilize several experts to learn the long-tailed distribution independently. Thus, the predictions of all experts are gradually integrated to reduce the overall model variance or uncertainty. NCL [34] adopts several complete networks to learn the long-tailed distribution individually and collaboratively along with self-supervised contrastive strategy [11].

Knowledge Distillation. Knowledge Distillation (KD) [22] is originally proposed to transfer knowledge from a large teacher model to a small student model by using soft labels output from the large model as targets. There are two main KD directions: logit-based KD and feature-based KD. Logit-based KD methods [7, 16, 30, 45, 67] directly use the output of the teacher model as supervision to guide the student model, while feature-based KD methods [1, 13, 20, 21, 25, 30, 51, 53, 54, 62] is designed to match the intermediate features between the teacher model and the student model. Recently, self-distillation has been proposed to utilize the idea of knowledge distillation for better and more efficient model optimization [64]. It treats the deep model as the teacher model to transfer knowledge directly to shallow layers viewed as student models.

3. Methodology

In this section, we introduce the proposed SHIKE in detail, which aims to enhance knowledge transfer in MoE. SHIKE contains two novel components named Depth-wise Knowledge Fusion (DKF) and Dynamic Knowledge Transfer (DKT). DKF aggregates knowledge diversely from experts with different model depths, and DKT is specifically designed for transferring knowledge among diverse experts. The overall structure is shown in Fig. 2.

3.1. Preliminaries

We denote $\{x_i, y_i\}_{i=1}^{N}$ as all data points in the training set, where each sample x_i has a corresponding label $y_i \in \{1, ..., C\}$. The size of the training set is $N = \sum_{c=1}^{C} n_c$, where n_c represents the number of training data in class c. Given a set of long-tailed data, the number of training data decreases according to the class indices, i.e., $n_1 > n_2 > ... > n_C$. Long-tailed learning aims to build a deep model on such long-tailed data by treating each class equally important.

We suppose a deep neural network parameterized by θ contains M experts. Usually, the network architecture of MoE makes the first several layers shared for all experts and the last few layers exclusive for each expert. Without loss of generality, we take ResNet [19] as an example. We denote the shared layers of a ResNet model as S stages for M experts. Only the last stage is adopted as the exclusive parameter for each expert. For expert m, we denote the parameters of its exclusive stage as θ_{S+1}^{m}, which is then followed by a linear layer parameterized as φ^m. Given a data x, the intermediate features $\mathbf{f}_s$ from stage $s\,(1 \leq s \leq S)$ of the shared network are calculated by

$$\mathbf{f}_s = \theta_s \circ \cdots \circ \theta_2 \circ \theta_1(x). \tag{1}$$

The operation $\circ$ indicates function composition: $h \circ g(x) = h(g(x))$. After the shared network, the output logits generated by expert m are calculated by:

$$\mathbf{z}^m = \varphi^m(\mathbf{f}_{S+1}^m), \tag{2}$$

where $\mathbf{f}_{S+1}^m = \theta_{K+1}^m(\mathbf{f}_S)$ represents the exclusive features extracted by expert m, and $\mathbf{f}_S$ represents the features extracted by the last shared stage of the network. In this MoE framework, we denote exclusive features $\mathbf{f}_{S+1}^m$ as high-level features and their preceding features $\mathbf{f}_s, s = 1, ..., S$, as intermediate features. During training, the cross-entropy loss can be calculated for each expert after obtaining the softmax probabilities. For model inference, the logits are summed up among all experts for each class, and the class with the maximum one is regarded as the MoE model prediction.

3.2. Depth-Wise Knowledge Fusion

Knowledge distillation is a commonly adopted optimization strategy for MoE methods in long-tailed learning [34, 38, 60, 61]. However, these methods mainly focus on distillation from logits rather than intermediate features. In the field of knowledge distillation, most state-of-the-art methods mainly take a feature-based manner where intermediate feature plays an essential part [51, 62]. This is a lack of considering the intermediate features, especially in MoE-based methods where all experts often share the same part of the network. It has been shown in Fig. 1 that features from different depths of the network can provide comparative performance towards different parts of long-tailed distribution: deep features exhibit promising performance on the head classes while shallow features can be more effective for some tail classes.

Therefore, to fully utilize the intermediate features during knowledge distillation in an MoE framework, we propose Depth-wise Knowledge Fusion (DKF) to aggregate features from different depths of a shared network with the high-level features extracted from each expert. For simplification, we assume the number of experts M is less than or equal to the number of shared stages, namely $M \leq S$ so that each expert can utilize the intermediate features from a unique depth in the network. Then, we can assign M sets of intermediate features among $\mathbf{f}_s$, $s = 1, ..., S$ to each expert. As different intermediate features have different sizes, one expert cannot simply concatenate or multiply them with the

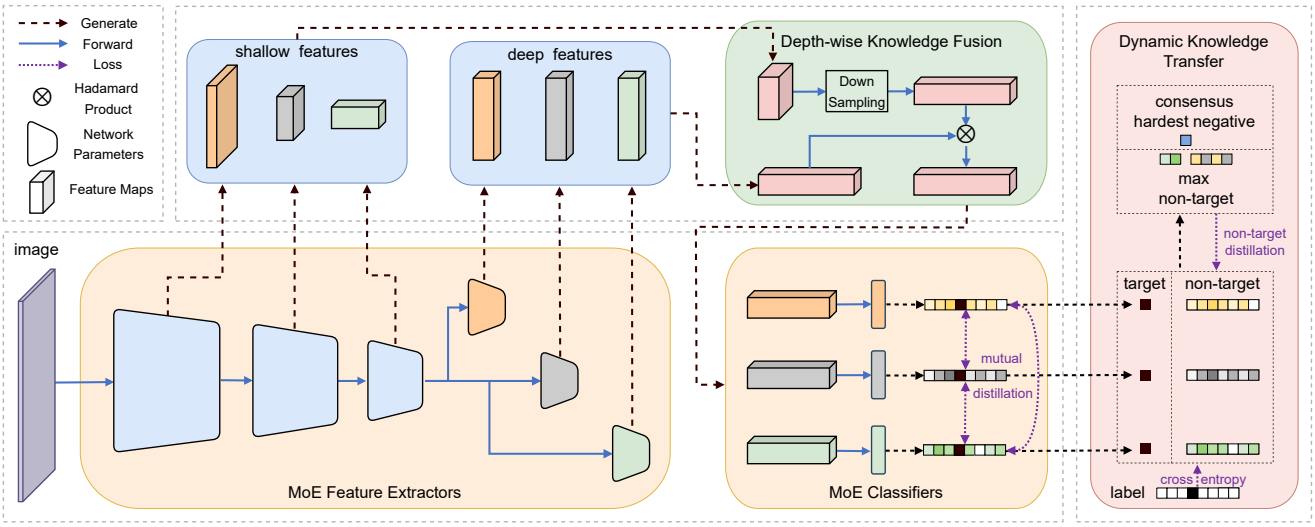


Figure 2. The structure of the proposed SHIKE. Each expert in MoE fuses the features from its own exclusive layers (deep features) and ones from the shared layers (intermediate features). The fused features are then used for mutual and dynamic knowledge distillation for better model optimization.

assigned features $\mathbf{f}_s$ directly. Therefore, we add several convolution layers for downsampling according to the depth of the intermediate features, achieving feature alignment between $\mathbf{f}_s$ and high-level features $\mathbf{f}_{S+1}^m$ extracted by expert m. Suppose the intermediate features after alignment are $\hat{\mathbf{f}}_s$. In DKF, we propose to fuse the intermediate features with high-level features by multiplication and then transform them into logits by φ^m:

$$\mathbf{z}^m = \varphi^m\left(\hat{\mathbf{f}}_s \otimes \mathbf{f}_{S+1}^m\right), \tag{3}$$

where $\otimes$ is the Hadamard product. As shown in Fig. 2, the intermediate features from a stage are assigned to an expert and aggregated with the high-level features of this expert.

To fully use the diverse features in DKF, we can apply knowledge distillation between any two experts to make them learn from each other. As each expert in MoE often has the same architecture located in the deepest position of the network, it can be guaranteed that each expert can play the role of either a teacher or a student. This enables mutual knowledge distillation between any two experts and provides a perfect opportunity for experts to aggregate different depths of knowledge:

$$\mathcal{L}_{mu} = \sum_{j=1}^{M}\sum_{k\neq j}^{M} KL(\mathbf{p}^j|\mathbf{p}^k), \tag{4}$$

where $\mathbf{p}^j$ and $\mathbf{p}^k$ represent the softmax probabilities of class j and k, respectively. This guarantees the knowledge transferring comprehensively between any two experts.

Feature fusion with mutual knowledge distillation in this way has two main advantages: (1) It dynamically fuses intermediate information from different depths of the network with semantic information from experts, which implicitly assigns different preferences of long-tailed distribution to experts without intuitive severance. (2) With more low-level information being aggregated, logit-based knowledge distillation can be more effective since each expert's output has more diversity corresponding to different depths of the model.

3.3. Dynamic Knowledge Transfer

The effectiveness of knowledge distillation largely relies on the non-target logits, i.e., the logits do not belong to the target class y, which provides similar semantic information in addition to the target logit. It is especially useful for long-tailed learning because the target logits of samples in the tail classes are usually relatively small during training such that the non-target logits can provide a comparable amount of information with the target label. With DKF, the non-target logits of different experts are more diverse because each expert can extract features with different semantic information due to different model depths. However, one circumstance during knowledge distillation may happen under long-tailed distribution that needs to be taken into consideration carefully. The model will be biased towards the head classes such that some samples in the tail classes will have high prediction confidence on the head classes, especially when they share similar semantic features. The non-target classes with high confidence logits are called hard negative classes [34,40]. It is dangerous to conduct knowledge distillation if the experts have a consensus on the hardest negative class, which may be a head class, for the tail class samples because the misleading information may be transferred.

Based on the analysis above, we propose Dynamic Knowledge Transfer (DKT) to address the issue of the hardest negative class during knowledge distillation with the proposed DKF. In addition to using the logits of all classes by the cross-entropy loss, DKT considers only non-target predictions from all experts and dynamically elects a teacher among them to handle the hardest negative class. For a sample x with label y, its corresponding output logits of expert m is $\mathbf{z}^m = [z_1^m, z_2^m, ..., z_C^m]$. Following [68], we first decouple the logits into a target logit z_y^m and non-target logits $[z_{I_1}^m, z_{I_2}^m, ..., z_{I_{C-1}}^m]$, where $\mathcal{I} = [I_1, I_2, ..., I_{C-1}]$ stores the index of non-target classes. After logits decoupling, we introduce the non-target set to a new knowledge distillation problem with $C-1$ classes. The average logits of all experts can be calculated for each non-target class $[\bar{z}_{I_1}, \bar{z}_{I_2}, ..., \bar{z}_{I_{C-1}}]$, where

$$\bar{z}_{I_i} = \frac{1}{M} \sum_{m=1}^{M} z_{I_i}^m, \tag{5}$$

for $i = 1, ..., C-1$. We can thus identify $\max_i\{\bar{z}_{I_i}\}$ among all non-target classes as the consensus hardest negative class, which is believed as the hardest negative class by joint prediction of MoE. To effectively suppress the logit of consensus hardest negative class through softmax suppression, a teacher who can comprehensively utilize the non-target knowledge is needed. Specifically, DKT chooses the maximum non-target logit among all experts denoted as $\hat{z}_{I_i}$:

$$\hat{z}_{I_i} = \max_{m=1,...,M} \{z_{I_i}^m\}, \tag{6}$$

for $i = 1, ..., C-1$. The large values of the maximum non-target logits can effectively suppress the value of the consensus hardest negative logit after softmax on the $C-1$ non-target classes. Taking advantage of the diversity in DKF, high values may appear on different non-target logits with different experts. Therefore, the maximum non-target logits can dynamically suppress the consensus hardest negative logit after softmax among $C-1$ non-target classes. Combining the consensus hardest negative with the maximum non-target logits, we can form a set of non-target logits called grand teacher:

$$z_{I_i}^{\mathcal{T}} = \begin{cases} \bar{z}_{I_i}, & i = \text{argmax}_j\{\bar{z}_{I_j}\}, \\ \hat{z}_{I_i}, & \text{otherwise}, \end{cases} \tag{7}$$

for $i = 1, ..., C-1$. Note that the grand teacher is only for suppressing the hardest negative class within non-target classes while the target class is not involved. After electing the grand teacher, non-target knowledge distillation is performed between it and each expert. The non-target probabilities for grand teacher and students are calculated by the following formulation:

$$\widetilde{p}_{I_i}^{\mathcal{T}} = \frac{\exp(z_{I_i}^{\mathcal{T}})}{\sum_{i=1}^{C-1} \exp(z_{I_i}^{\mathcal{T}})}, \quad \widetilde{p}_{I_i}^{m} = \frac{\exp(z_{I_i}^{m})}{\sum_{i=1}^{C-1} \exp(z_{I_i}^{m})}, \tag{8}$$

for $i = 1, ..., C-1$ and $m = 1, ..., M$. Therefore, the knowledge distillation for non-target logits among SHIKE's experts can be formulated as:

$$\mathcal{L}_{nt} = \sum_{m=1}^{M} KL(\widetilde{\mathbf{p}}^{\mathcal{T}} | \widetilde{\mathbf{p}}^{m}). \tag{9}$$

For a particular sample, the hardest negative of its corresponding outputs may vary not only among experts but also along the training process. DKT can dynamically choose the hardest negative among experts and reduce its probability without affecting the target logit.

3.4. Overall Training Paradigm

SHIKE adopts a decoupled training scheme that optimizes the feature extractor and classifier separately, as it has been shown that class-balanced joint training strategies for long-tailed data may hurt representation learning [27, 69]. For feature extractor training, we adopt both mutual knowledge distillation loss $\mathcal{L}_{mu}$ in Eq. (4) by DKF and non-target knowledge distillation loss $\mathcal{L}_{nt}$ in Eq. (9) by DKT. Meanwhile, to preserve the information from the original distribution untouched for representation learning, vanilla cross-entropy loss $\mathcal{L}_{ce}$ is also applied for each expert. Therefore, the above three loss functions during the representation learning stage are assembled as a whole optimization objective:

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha \mathcal{L}_{nt} + \beta \mathcal{L}_{mu}, \tag{10}$$

where α and β are trade-off hyperparameters. For classifier training, the goal is to train a balanced classifier with the feature extractor frozen. We utilize the balanced softmax cross entropy (BSCE) [49] as the loss function L_{bsce} to simply optimize a new classifier for each expert:

$$\mathcal{L}_{bsce} = -\sum_{m=1}^{M} \log \left(\frac{n_y \exp(z^m)}{\sum_{j=1}^{C} n_j \exp(z_j^m)} \right). \tag{11}$$

Knowledge distillation is not considered in the stage of classifier re-training as it will encourage classifiers to be similar, which is harmful to experts to make joint predictions.

4. Experiments

4.1. Datasets

CIFAR-100-LT is a subset of the original CIFAR-100 [31] with long-tailed distribution. The imbalance factor of CIFAR-100-LT can be set at 50 or 100, where 100 means that the largest class contains 100 times of samples than the smallest class. The validation set remains the same as the original CIFAR-100 with 10,000 samples in total. **ImageNet-LT** is also a subset of the original ImageNet [13], which is first proposed in [41]. It has an imbalance factor of 256 following Pareto distribution with a power value of 0.6. With 1,000 classes in total, the training set and test set contain 115.8K samples and 50K samples, respectively. **iNaturalist 2018** [55] is a large-scale real-world dataset. It is

Method	Year	CIFAR100-LT	
		100	50
Single model			
Focal Loss [40]	2017	42.3	-
OLTR [41]	2019	43.4	-
LDAM-DRW [5]	2019	44.4	-
τ-norm [27]	2020	45.4	-
cRT [27]	2020	45.6	-
BALMS [49]	2020	50.7	-
LADE [23]	2021	45.4	50.5
GCL [36]	2022	48.7	53.6
Weight Balancing [2]	2022	53.6	57.7
Contrastive & Hybrid methods			
BALMS+BatchFormer [24]	2022	51.7	-
PaCo [11]	2021	51.9	56.0
PaCo+BatchFormer [24]	2022	52.4	-
BCL [72]	2022	52.0	56.6
MoE-based methods			
RIDE (3E) [60]	2021	48.3	-
ResLT (3E) [10]	2022	45.3	50.0
TLC (4E) [33]	2022	50.1	-
NCL (S) [34]	2022	53.3	56.8
NCL (3N) [34]	2022	54.2	58.2
Ours (3E)	-	**56.3**	**59.8**

Table 1. Comparison results on CIFAR100-LT with imbalance factor of 100 and 50.

Method	Year	ImageNet-LT		iNat
		R-50	RX-50	R-50
Single model				
OLTR [41]	2019	-	-	63.9
LDAM-DRW [5]	2019	-	-	68.0
cRT [27]	2020	47.3	49.6	65.2
τ-norm [27]	2020	46.7	49.4	65.6
BALMS [49]	2020	50.1	-	-
LA [44]	2021	-	-	66.4
CAM [66]	2021	-	-	70.9
GCL [36]	2022	54.9	-	72.0
Weight Balancing [2]	2022	-	53.9	70.2
Contrastive & Hybrid methods				
SSD [38]	2021	-	56.0	-
PaCo [11]	2021	57.0	58.2	73.2
BCL [72]	2022	56.0	-	71.8
RIDE+BF [24]	2022	55.7	-	74.1
BALMS+BF [24]	2022	51.1	-	-
MoE-based methods				
BBN [69]	2020	48.3	49.3	66.3
RIDE (3E) [60]	2021	55.4	56.8	72.6
ACE [4]	2021	54.7	56.6	-
NCL (S) [34]	2022	57.4	58.4	74.2
NCL (3N) [34]	2022	59.5	**60.5**	74.9
Ours (3E)	-	**59.7**	59.6	**75.4**

Table 2. Comparison results on ImageNet-LT and iNaturalist 2018 (iNat). R-50 and RX-50 are short for ResNet-50 and ResNeXt-50 (32x4d), respectively.

extremely imbalanced with 437.5K samples from 8,142 categories. **Places-LT** is created from the large-scale dataset Places [70] with 184.5K samples from 365 categories.

4.2. Implementation Details

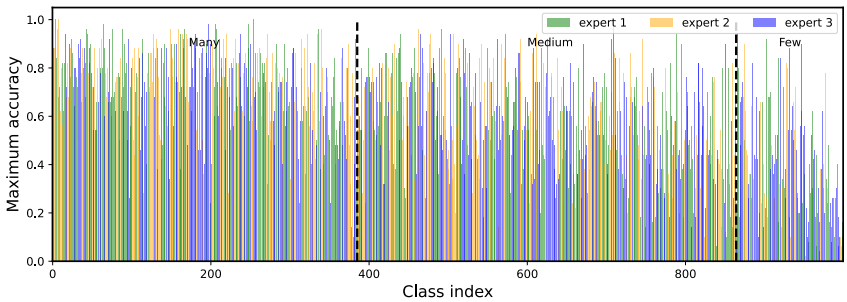
For CIFAR100-LT, we use ResNet-32 as the backbone. AutoAugment [8] and Cutout [14] are adopted by following [11, 49]. For ImageNet-LT, ResNet-50 and ResNeXt-50 (32x4d) are adopted. Similarly, we use ResNet-50 and ResNet-152 for iNaturalist 2018 and Places-LT. The above four datasets are trained with learning rates of 0.05, 0.2, 0.025, and 0.02, respectively. All models are trained for 180 epochs except Places-LT with 30 epochs of fine-tuning as it utilizes a pre-trained model. If not specified, we adopt SGD optimizer with momentum 0.9, cosine schedule of decaying to 0, and weight decay of 5e-4 for all experiments. RandAugment [9] is used for ImageNet-LT and iNaturalist-2018 by following [34]. Also, during the classifier training phase, the cosine learning rate scheduler restarts, and classifiers of experts are trained for 20 more epochs.

4.3. Main Results

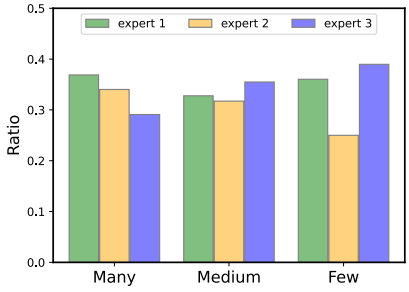
All comparison results with other state-of-the-art methods for long-tailed learning are presented in Tab. 1-3. We compare the proposed SHIKE with single model methods, contrastive methods, hybrid methods and MoE-based methods. All of them are proposed for long-tailed data. Specific settings for MoE-based models are marked in parentheses: S for the single model, N for the full network, and E for the expert. The best results are presented in **bold**.

CIFAR100-LT The comparison results on CIFAR100-LT with the imbalance factor of 100 and 50 are shown in Tab. 1. Note that the adopted ResNet-32 model has only three stages, which provide two shared stages to generate intermediate features for fusion and one exclusive stage for each expert. In this case, we let two experts fuse with the same shallowest intermediate features. We group the existing methods based on their types. SHIKE achieves better performance within or out of MoE-based methods. For example, it achieves 2.1% and 1.6% improvements over the second-best method NCL (3N) on the imbalance factor of 100 and 50, respectively. Besides, it is worth noting that in the previous state-of-the-art methods, NCL adopts three complete networks as its experts, while SHIKE only utilizes experts consisting of the last stage of the ResNet along with one or two downsampling layers. Another advantage of the proposed SHIKE is we only train it for 200 epochs, which is less than the method following a contrastive learning strategy that requires more epochs.

ImageNet-LT and iNaturalist 2018. We report overall Top-1 accuracy for ImageNet-LT and iNaturalist 2018 in Tab. 2. For a fair comparison, we use SHIKE to train MoEs 200 epochs for both datasets. For ImageNet-LT, SHIKE can perform better than all the competing meth-



(a) Maximum class accuracy among experts.



(b) Ratio of classes with the highest accuracy.

Figure 3. Preferences of different experts in SHIKE. (a) The highest accuracy among experts is shown for each class on the test set of ImageNet-LT. (b) We calculate the ratio of class numbers that each expert is most skilled at within three divisions. The experiment is conducted with ResNet-50 and the number of experts is set to 3.

Method	Year	Places-LT			
		Many	Med	Few	All
Single model					
Focal Loss [40]	2017	41.1	34.8	22.4	34.6
OLTR [41]	2019	**44.7**	37.0	25.3	35.9
NCM [27]	2020	40.4	37.1	27.3	36.4
cRT [27]	2020	42.0	37.6	24.9	36.7
τ-norm [27]	2020	37.8	40.7	31.8	37.9
LWS [27]	2020	40.6	39.1	28.6	37.6
BALMS [49]	2020	41.2	39.8	31.6	38.7
LADE [23]	2021	42.8	39.0	31.2	38.8
DisAlign [65]	2021	40.4	42.4	30.1	39.3
GCL [36]	2022	-	-	-	40.6
Contrastive & Hybrid methods					
LDAM+RSG [58]	2021	41.9	41.4	32.0	39.3
PaCo [11]	2021	37.5	**47.2**	33.9	41.2
MoE-based methods					
LFME [61]	2020	39.3	39.6	24.2	36.2
NCL (S) [34]	2022	-	-	-	41.5
NCL (3N) [34]	2022	-	-	-	41.8
Ours (3E)	-	43.6	39.2	**44.8**	**41.9**

Table 3. Comparison results on Places-LT. The results are shown by different class divisions (Many, Medium, and Few) as well as the overall accuracy (All).

ods, including NCL [34]. It is worth noting that NCL trains its MoE for 400 epochs with contrastive training strategies. Moreover, NCL adopts three whole networks as experts, consuming more computational overhead for training. Our method is less computationally expensive but still achieves comparable or even better performance than NCL. We also conduct an experiment with 400 epochs for ResNet-50 on ImageNet-LT, achieving a performance of 60.3%.

Places-LT. As previous works consider the pre-trained backbone of ResNet-152 as a whole, it is troublesome for SHIKE to implement the model. To demonstrate the effectiveness of the proposed SHIKE, we keep the shared part of the model fixed after loading the pre-trained parameters. Then, we only fine-tune the exclusive part within experts and its corresponding downsampling layers, which

MoE	DKF	$\mathcal{L}_{mu}$	$\mathcal{L}_{nt}$	Acc
				50.04
✓				54.23
✓	✓			55.26
✓	✓	✓		55.59
✓	✓		✓	55.79
✓		✓	✓	54.82
✓	✓	✓	✓	**56.34**

Table 4. Ablation study on the effects of different components in the proposed SHIKE. The experiment is conducted on CIFAR100-LT with an imbalance factor of 100.

are crucial for the proposed component DKF. The comparison results are listed in Tab. 3. This experiment shows that intermediate features from the pre-trained model can also help to boost the overall performance on long-tailed visual recognition. To further reveal the effectiveness of SHIKE, we report the accuracy on three divisions of the classes, namely many-shot classes (>100 training samples), medium-shot classes (20∼100 training samples) and few-shot classes (<20 training samples). In addition to achieving slightly higher performance than the state-of-the-art, we find it more fascinating that SHIKE achieves 44.8% accuracy on the few-shot classes, which is more than 10% higher than the runner-up 33.9% achieved by PaCo [11]. As a result, SHIKE has a more balanced test performance compared to the contrastive learning methods.

4.4. Ablations and Model Validation

Ablation Studies on Components of SHIKE. Shown in Tab. 4, the ablation study is conducted on CIFAR100-LT with an imbalance factor of 100. Key components of SHIKE are further subdivided into MoE, DKF, and two loss functions $\mathcal{L}_{mu}$ and $\mathcal{L}_{nt}$. MoE means whether the method uses three experts or a single plain ResNet-32. All experiments are conducted following a decoupled training scheme, and accuracy is calculated on the balanced test set. The accuracy for a single plain model without any component is 50.04%, which acts as a baseline for ablation. When applying the MoE architecture, the accuracy is signif-

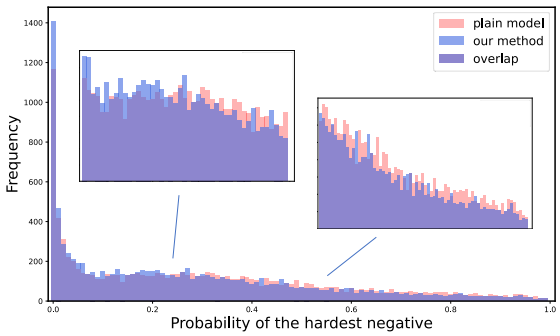


Figure 4. Probability distribution of the hardest negative class for models trained on CIFAR100-LT with an imbalance factor of 100. The result is counted on the test set of CIFAR100-LT to show the effectiveness of suppressing the hardest negative class by DKT.

icantly boosted to 54.23%, which validates the effectiveness of experts' ensemble toward long-tailed learning. When applying DKF to MoE, the performance can be boosted by around 1%. Based on DKF, the mutual knowledge distillation loss $\mathcal{L}_{mu}$ and the non-target knowledge distillation $\mathcal{L}_{mt}$ (DKT) can further improve the accuracy by 0.33% and 0.53%, respectively. When all the components are applied, we achieve the highest accuracy of 56.34%, which is around 2% higher than the plain MoE model. An interesting observation is that if we do not apply DKF, all the other components can only bring 0.59% improvement. This proves that the DKF is the fundamental architecture in SHIKE, which assigns more meaningful features to experts in MoE for further knowledge distillation and model optimization.

Evaluation on the Hardest Negatives. The ablation study in Tab. 4 has shown the effectiveness of DKT in terms of accuracy. To further validate how DKT suppresses the hardest negatives, we conduct an evaluation based on the test set. A single model and the proposed SHIKE are utilized, and both model's classifiers are trained with BSCE along with keeping the feature extractor fixed. As shown in Fig. 4, we can see that the number of the hardest negatives with large values is reduced, which indicates that SHIKE can effectively alleviate the influence of hardest negatives during knowledge distillation in MoE.

Evaluation on the Preference of Experts. To show the preferences of experts for long-tailed distribution, an experiment is conducted on the test set of ImageNet-LT, and the experts are evenly assigned with three different levels of intermediate features from shallow to deep accordingly. Fig. 3 (a) shows the highest accuracy among experts for each class. It can be observed that the diversity among the three experts is quite high, where each expert performs well in different classes that are distributed from head to tail of the distribution. We also calculate the ratio of classes with the highest accuracy among experts. From Fig. 3(b), we can see that expert 1 and 3 with the shallowest and the deepest intermediate features performs better on few-shot division. While expert 2, which is assigned by the middle intermediate features, follows a normal accuracy distribution which is consistent with long-tailed distribution. Moreover, for classes in the many-shot division, the shallowest feature is superior in helping expert to achieve higher performance. In all, experts with different intermediate features appear to have different preferences for long-tailed distribution, which makes experts skilled at different parts of the distribution.

E	Depth arrangement		
1	A 54.10	B 54.75	C **55.84**
2	A B 57.39	B C **58.79**	A C 58.62
3		A B C **59.72**	
4	A B C A 59.83	A B C B **59.90**	A B C C 59.77

Table 5. Ablation study on the effects of expert number in the proposed SHIKE. The experiment is conducted on ImageNet-LT.

Evaluation on the Number of Experts. To show the influence of the number of experts, we conduct an experiment on ImageNet-LT by varying the number of experts from 1 to 4 and different depth combinations. As shown in Tab. 5, letters from A to C represent depths from shallow to deep. The overall accuracy of the ensemble generally rises along with the increasing number of experts. Moreover, it shows that as the number of experts grows, architectures with more heterogeneous experts promote more.

5. Conclusion

We have proposed a Self-Heterogeneous Integration with Knowledge Excavation (SHIKE) for long-tailed visual recognition. The proposed SHIKE consists of Depth-wise Knowledge Fusion (DKF) and Dynamic Knowledge Transfer (DKT). DKF fuses the depth-wise intermediate features with high-level features and thereby provides more informative features for experts to accommodate the long-tailed distribution. DKT exploits the non-target knowledge among diversified experts to reduce the hardest negative for representation learning, which can further improve the performance on the tail classes. Extensive experiments have been conducted on four benchmarks and SHIKE achieved excellent performance compared to state-of-the-art counterparts.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grants 62002302 and U21A20514, the FuXiaQuan National Independent Innovation Demonstration Zone Collaborative Innovation Platform under Grant 3502ZCQXT2022008, NSFC/Research Grants Council (RGC) Joint Research Scheme under Grant N_HKBU214/21, the General Research Fund of RGC under Grants 12201321 and 12202622, the Natural Science Foundation of Fujian Province under Grant 2020J01005.

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *CVPR*, 2019. 3

[2] Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *CVPR*, 2022. 6

[3] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. 1

[4] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *ICCV*, 2021. 2, 3, 6

[5] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *NeurIPS*, 2019. 1, 6

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *JAIR*, 16:321–357, 2002. 2

[7] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *ICCV*, 2019. 3

[8] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, 2019. 6

[9] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, 2020. 6

[10] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *IEEE TPAMI*, 45(3):3695–3706, 2023. 2, 3, 6

[11] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *ICCV*, 2021. 2, 3, 6, 7

[12] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 1

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 3, 5

[14] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 6

[15] Charles Elkan. The foundations of cost-sensitive learning. In *IJCAI*, 2001. 2

[16] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, 2018. 3

[17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[20] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, 2019. 3

[21] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *AAAI*, 2019. 3

[22] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[23] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *CVPR*, 2021. 2, 6, 7

[24] Zhi Hou, Baosheng Yu, and Dacheng Tao. Batchformer: Learning to explore sample relationships for robust representation learning. In *CVPR*, 2022. 6

[25] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017. 3

[26] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *ICLR*, 2020. 1, 2

[27] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020. 1, 2, 5, 6, 7

[28] Salman H. Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A. Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE TNNLS*, 29(8):3573–3587, 2018. 1

[29] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *CVPR*, 2020. 2

[30] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *NeurIPS*, 2018. 3

[31] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

[33] Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang. Trustworthy long-tailed classification. In *CVPR*, 2022. 2, 3, 6

[34] Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested collaborative learning for long-tailed visual recognition. In *CVPR*, 2022. 3, 4, 6, 7

[35] Mengke Li, Yiu-ming Cheung, and Zhikai Hu. Key point sensitive loss for long-tailed visual recognition. *IEEE TPAMI*, 45(4):4812–4825, 2022. 2

[36] Mengke Li, Yiu-ming Cheung, and Yang Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *CVPR*, 2022. 2, 6, 7

[37] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *CVPR*, 2021. 2

[38] Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition. In *ICCV*, 2021. 3, 6

[39] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *CVPR*, 2020. 2, 3

[40] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2, 4, 6, 7

[41] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019. 1, 5, 6, 7

[42] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[43] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 2

[44] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021. 1, 2, 6

[45] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, 2020. 3

[46] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *CVPR*, 2022. 1

[47] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *ICCV*, 2021. 2

[48] Foster Provost. Machine learning from imbalanced data sets 101. In *AAAI Workshops*, 2000. 2

[49] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *NeurIPS*, 2020. 2, 5, 6, 7

[50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015. 1

[51] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2014. 3

[52] Christian Szegedy, Alexander Toshev, and Dumitru Erhan. Deep neural networks for object detection. *NeurIPS*, 2013. 1

[53] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019. 3

[54] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *ICCV*, 2019. 3

[55] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018. 1, 5

[56] Vladimir Vapnik. Principles of risk minimization for learning theory. *NeurIPS*, 1991. 1

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 1

[58] Jianfeng Wang, Thomas Lukasiewicz, Xiaolin Hu, Jianfei Cai, and Zhenghua Xu. Rsg: A simple but effective module for learning imbalanced datasets. In *CVPR*, 2021. 2, 7

[59] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *ECCV*, 2020. 2

[60] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*, 2021. 2, 3, 6

[61] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *ECCV*, 2020. 2, 3, 7

[62] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2016. 3

[63] Yuhang Zang, Chen Huang, and Chen Change Loy. Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. In *ICCV*, 2021. 2

[64] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *IEEE TPAMI*, 44(8):4388–4403, 2021. 2, 3

[65] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, 2021. 2, 7

[66] Yongshun Zhang, Xiu-Shen Wei, Boyan Zhou, and Jianxin Wu. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *AAAI*, 2021. 6

[67] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, 2018. 3

[68] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, 2022. 5

[69] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, 2020. 2, 5, 6

[70] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 40(6):1452–1464, 2017. 1, 6

[71] Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE TKDE*, 18(1):63–77, 2005. 1, 2

[72] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *CVPR*, 2022. 2, 6