Long-Tailed Recognition Using Class-Balanced Experts

Saurabh Sharma[1](✉), Ning Yu[1,2], Mario Fritz[3], and Bernt Schiele[1]

[1] Max Planck Institute for Informatics, Saarland Informatics Campus,
Saarbrücken, Germany
{ssharma,ningyu,schiele}@mpi-inf.mpg.de
[2] University of Maryland, College Park, USA
[3] CISPA Helmholtz Center for Information Security, Saarland Informatics Campus,
Saarbrücken, Germany
fritz@cispa.saarland

Abstract. Deep learning enables impressive performance in image recognition using large-scale artificially-balanced datasets. However, real-world datasets exhibit highly class-imbalanced distributions, yielding two main challenges: relative imbalance amongst the classes and data scarcity for mediumshot or fewshot classes. In this work, we address the problem of long-tailed recognition wherein the training set is highly imbalanced and the test set is kept balanced. Differently from existing paradigms relying on data-resampling, cost-sensitive learning, online hard example mining, loss objective reshaping, and/or memory-based modeling, we propose an ensemble of class-balanced experts that combines the strength of diverse classifiers. Our ensemble of class-balanced experts reaches results close to state-of-the-art and an extended ensemble establishes a new state-of-the-art on two benchmarks for long-tailed recognition. We conduct extensive experiments to analyse the performance of the ensembles, and discover that in modern large-scale datasets, relative imbalance is a harder problem than data scarcity. The training and evaluation code is available at https://github.com/ssfootball04/class-balanced-experts.

1 Introduction

In the past decades, deep learning has boosted success in image recognition to a new level [14]. The availability of large-scale datasets with thousands of images in each class [4,47] has been a major factor in this revolution. However, these datasets are manually curated and artificially balanced, as opposed to real-world datasets that exhibit a highly skewed and class-imbalanced distribution in a long-tailed shape: a few common classes and many more rare classes. To address this practical challenge, in this work, we focus on the problem of long-tailed recognition, wherein datasets exhibit a natural power-law distribution [32], allowing us to assess model performance on four folds: *Manyshot* classes (≥100 samples), *Mediumshot* classes (20–100 samples), *Fewshot* classes (<20 samples), and *All* classes. Training data follows a highly class-imbalanced distribution, and testing data is balanced so that equally good performance over all classes is crucial [24].

Z. Akata et al. (Eds.): DAGM GCPR 2020, LNCS 12544, pp. 86–100, 2021.
https://doi.org/10.1007/978-3-030-71278-5_7

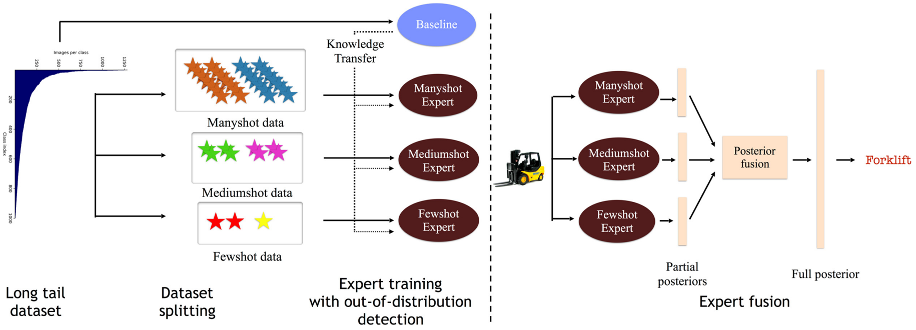


Fig. 1. Our pipeline for long-tailed recognition: an ensemble of experts trained on class-balanced subsets of *Manyshot*, *Mediumshot*, and *Fewshot* data. We *transfer knowledge* from *Manyshot* to *Mediumshot* and *Fewshot* classes by initialising experts with a *Baseline* model trained on all the data. Expert models classify samples outside their subset as out-of-distribution and output partial posteriors that are fused into a full posterior to obtain the final prediction.

The two main challenges for a long-tailed classification model are *relative imbalance* amongst the classes, and *data scarcity* or unobservable data modes [13]. Existing techniques for imbalanced classification have focused on data re-sampling [6,13] and cost-sensitive learning [3,23] to re-weigh the loss objective or counter *relative imbalance*, while techniques for fewshot learning have employed data augmentation [7,36,40,41], classifier weight prediction [9,28,29], or prototype-based non-parametric methods [24,30,33] to address *data scarcity*.

Unlike the aforementioned paradigms, we instead revisit the classic approach of ensemble of experts [17,19,44] and adapt it to long-tailed recognition. We first decompose the imbalanced classification problem into balanced classification problems by splitting the long-tailed training classes into balanced subsets. Then we train an expert on each balanced subset, so-called *Manyshot*, *Mediumshot*, or *Fewshot* data, with out-of-distribution detection for samples outside an expert's class-balanced subset. This explicitly tackles the issue of *relative imbalance*, and prevents competition between *Manyshot* and *Fewshot* classes during training.

Further, to use all available data for learning feature representations and to *transfer knowledge* from *Manyshot* to *Mediumshot* and *Fewshot* classes, we initialise the feature extractor of each expert using a *Baseline* model trained on the entire dataset. This simple and effective approach reaches close to state-of-the-art results without involving more complex models or sophisticated loss objectives. Moreover, the decomposition into class-balanced subsets allows us to analyse the upper bound on performance in each data regime. Specifically, our experiments with an *Oracle* upper bound allow us to bring *Fewshot* and *Mediumshot* accuracy on par with *Manyshot* accuracy, revealing that in modern large-scale datasets the data scarcity for *Mediumshot* and *Fewshot* classes can be effectively handled using knowledge transfer from *Manyshot* classes. Therefore, relative imbalance is a more severe problem.

We also leverage the flexibility and modularity of the ensemble framework to create larger and more diverse ensembles using existing solutions for long-tailed recognition. In particular, we involve the following methods in the solution space: (1) a *Baseline* model without any bells or whistles; (2) feature learning followed by classifier finetuning with uniform class sampling [31,41]; (3) data augmentation using feature generation networks [7,36,41]; and (4) knowledge transfer through prototype-based memory representation [24,30]. The extended ensemble consisting of all these models outperforms the current state-of-the-art on two benchmark datasets by a significant margin.

Our **contributions** in this work can be summarised as follows:

(1) We propose an effective and modular ensemble of experts framework for long-tailed recognition that decomposes the imbalanced classification problem into multiple balanced classification problems. Our framework utilises all available data for learning feature representations and transfers this knowledge from *Manyshot* to *Mediumshot* and *Fewshot* classes. The results of our ensemble of class-balanced experts are close to the state-of-the-art performance on two long-tailed benchmark datasets, ImageNet-LT and Places-LT [24].

(2) We enrich our ensemble with a diverse set of existing solutions for long-tailed recognition, namely data re-sampling, data augmentation using synthesised features, and prototype-based classification, and establish a new state-of-the-art for long-tailed recognition.

(3) We analyse the upper bound performance of our approach in the following manner: we assume Oracle access to the experts containing the ground truth classes of the test samples in their class-balanced subsets. We discover that *data scarcity* for rare classes is not a severe issue in modern large-scale datasets. Rather, *relative imbalance* is the main bottleneck.

2 Related Work

Imbalanced Classification and Long-Tailed Recognition. There is a long history of research in imbalanced classification [1,13,32], in binary and more generally multi-class classification problems. Classic problems that naturally encounter class imbalance are face attribute detection [18,26], object detection [23,48], and image defect detection [43]. Prior work on image classification [37,38] deals with long-tailed datasets, but only recently a benchmark for the problem on the ImageNet and Places dataset was proposed by [24]. They also propose splits for open-world classification, but in this work we only consider long-tailed recognition and we report the performance of our methods on the proposed ImageNet-LT and Places-LT. We summarise below the existing solutions for imbalanced classification and long-tailed recognition.

Data Re-sampling Heuristics and Cost-Sensitive Learning. These are classic ways to tackle long-tailed recognition. A more balanced data distribution is achieved by randomly over-sampling fewshot classes or randomly under-sampling of manyshot classes [6,13]. However, over-sampling suffers from over-fitting on fewshot classes while under-sampling cannot take full benefit of available data for generalization on manyshot classes. Other work has focused on hard example mining [5] or cost-sensitive learning [3,23] reasoned from class frequencies. Instead, to augment our ensemble of class-balanced experts, we use a uniform class sampling procedure in mini-batch training for finetuning the classifier after a representation learning phase, which has the advantage that all data is used to learn representations while decision boundary learning takes class imbalance into account. This has also been employed before in related zero-shot learning [41] and fewshot learning [31] work.

Synthetic Data Augmentation. This is a classic technique that synthesises features for minority classes based on feature space similarities [2,12]. More recently, generative models have been employed in zero-shot [7,40,41] and fewshot learning [36] literature to automatically generate images or feature embeddings for data-starved classes. In this work, we use the f-VAEGAN-D2 model from [41] that generates feature embeddings conditioned on available class embeddings using a VAE-GAN model, and integrate it into our ensemble of experts framework.

Prototype-Based Models and Knowledge Transfer. Prototype-based networks [30,33] maintain a memory module for all the classes such that each class is equally represented regardless of sample frequency. In particular, Liu et al. [24] learn prototype-based features on-the-fly to effectively transfer knowledge from manyshot classes to fewshot classes. We integrate their model into our ensemble due to its ability to perform consistently well across the entire class spectrum. Transfer learning [27] addresses data imbalance by transferring abundant features of manyshot classes to those of fewshot classes. Recent work includes transferring the intra-class variance [42] and transferring semantic deep features [24,46]. We instead transfer knowledge across the dataset by initialising our expert models with a baseline model pre-trained on the entire dataset.

Ensemble Learning. Ensemble methods are a well-studied topic in machine learning literature. In particular, a variety of ensemble-based methods using boosting [11,34], bagging [8,20], stacking [35], and evolutionary selection of classifiers [21] have been employed for imbalanced datasets. However, they all consider ensembles with the same kind of model and task. Our approach is related to the work of Hinton et al. [17] who train an ensemble of experts over disjoint semantically-close subsets of classes, thereby each expert deals with a different classification task. We instead train our experts on subsets of classes that are intrinsically balanced to counter relative imbalance and prevent competition between manyshot and fewshot classes during training. Moreover, we integrate a diverse set of models for long-tailed recognition into our ensemble of experts.

Out-of-Distribution Detection and Confidence Calibration. Modern neural networks can function both as classification models and detectors for out-of-distribution examples [15]. Recent works focus on adding small perturbations in input space and applying temperature scaling [22], and adding loss terms to push out-of-distribution examples towards uniform confidence [16]. Related work on confidence calibration tries to fix overconfident predictions on in-distribution data using temperature scaling [10]. We instead focus on learning an ensemble of class-balanced experts for long-tailed recognition, where the problem of out-of-distribution detection arises when dealing with samples from outside an expert's subset, and jointly calibrate experts' confidences to fuse their posteriors.

3 Method

We propose an ensemble of experts for solving the problem of long-tailed recognition. We split the long-tailed dataset into (approximately) class-balanced subsets, and a separate classification model, or expert, is trained for each subset. Expert models identify samples belonging to classes outside their subset as out-of-distribution; therefore we train them to produce low confidence predictions on these samples. During inference, each classification model yields a partial posterior distribution for test samples, the ensemble of which is fused to form a complete posterior distribution. Our entire pipeline is depicted in Fig. 1. The modularity of our framework allows us to explictly address the problem of *relative imbalance*, and moreover analyse the upper bounds for performance in each data regime using Oracle access to experts containing ground truth classes of test samples in their class-balanced subsets.

3.1 Long-Tailed Recognition Using Class-Balanced Experts

The task of long-tailed visual recognition is as follows: given class-imbalanced training set $\mathcal{D}_{Train} = \{(x_i, y_i)\}_{i=1}^{n}$ and class-balanced validation set $\mathcal{D}_{Val}$ and class-balanced test set $\mathcal{D}_{Test}$, the objective is to maximise test accuracy on four folds, *Manyshot* classes (≥100 samples), *Mediumshot* classes (20–100 samples), *Fewshot* classes (<20 samples), and *All* classes. This is a hard problem, since any high performing model must deal with the two problems of relative imbalance and data scarcity.

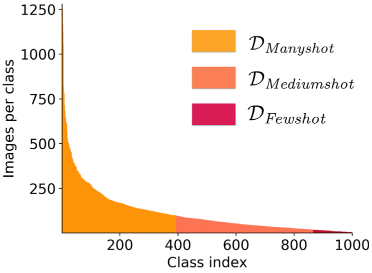


Fig. 2. Dataset splitting: We decompose ImageNet-LT into (relatively) class-balanced *Manyshot*, *Mediumshot*, and *Fewshot* data subsets.

Relative imbalance leads to biased classification boundaries wherein accuracy on fewshot samples is compromised in favor of manyshot samples that dominate the training objective. Data scarcity leads to representations that do not model unobserved data modes and is more severe. To tackle both these issues, we sort the class-imbalanced training set $\mathcal{D}_{Train}$ according to class frequencies and partition it into contiguous class-balanced subsets $\mathcal{D}_{Manyshot}$, $\mathcal{D}_{Mediumshot}$ and $\mathcal{D}_{Fewshot}$. This is visualised in Fig. 2.

For each subset, we train separate classification models or experts, that are initialised using a model pre-trained on the entire dataset. Consequently we obtain the expert models $\mathcal{E}_{Manyshot}$, $\mathcal{E}_{Mediumshot}$ and $\mathcal{E}_{Fewshot}$ corresponding to each class-balanced subset. The feature extractor part of each expert model $\mathcal{E}_{\_}$ is initialised using the *Baseline* model pre-trained on the entire training set $\mathcal{D}_{Train}$. This enables knowledge transfer from *Manyshot* to *Mediumshot* and *Fewshot* classes. In this work, the expert models $\mathcal{E}_{\_}$ and the *Baseline* model are deep fully convolutional neural networks with softmax classifiers.

3.2 Out-of-Distribution Detection for Experts

The expert models identify samples from classes outside their class-balanced subset as out-of-distribution or OOD for short, therefore we train them using an out-of-distribution detection strategy. Observe that this is a hard problem, since here OOD examples come from within the same distribution albeit from extra classes within the dataset, as opposed to standard out-of-distribution detection wherein OOD samples come from an entirely different dataset.

Training with Reject Class. We add a reject class to the softmax classifier of each expert. For instance, $\mathcal{E}_{Manyshot}$ treats samples from $\mathcal{D}_{Mediumshot} \cup \mathcal{D}_{Fewshot}$ as a single reject class. This introduces imbalance since the reject class has far more samples than any other class, therefore we undersample reject class samples appropriately during training. We correct for the statistical bias by incrementing its logit score by the log of the undersampling ratio. We note that samples in the reject class have very high variance and are therefore hard to fit.

3.3 Fusing Expert Posteriors

We consider various baseline strategies and propose a novel joint calibration module to fuse expert posteriors $\mathcal{E}_{\_}(x)$ into a complete posterior distribution. The final prediction and confidence scores are taken from this posterior, denoted as $q(x)$, using the argmax operation.

KL-Divergence Minimisation. We find the full posterior distribution for each sample, by minimising its KL-divergence with all the partial posterior distributions predicted by the experts [17], that is,

$$\min_{q(x)} \sum_{\mathcal{E}_{\_}} KL(\mathcal{E}_{\_}(x)||q(x))$$

where $q(x)$ is parameterised using logits z and a softmax function as $q(x) = softmax(z)$. Note that probabilities corresponding to out-of-distribution classes for the expert $\mathcal{E}_{\_}$ are summed up into one probability score in $q(x)$ to align the two distributions.

Soft-Voting. We find the full posterior by summing up the partial posteriors directly and normalising the sum to 1,

$$q(x) = \frac{\sum_{\mathcal{E}_{\_}} g(\mathcal{E}_{\_}(x))}{\sum_{\mathcal{E}_{\_}} \mathbb{1}}$$

Here $g(.)$ is a function that converts an expert's partial posterior into a full posterior. Since experts are trained with a reject class, $g(.)$ averages reject class probability score across out-of-distribution classes corresponding to expert $\mathcal{E}_{\_}$.

Expert Selection. We train a 3-way classifier on the validation set, taking the partial posterior vectors $\mathcal{E}_{\_}(x)$ of each expert $\mathcal{E}_{\_}$ as input, to predict for a sample x the expert model $\mathcal{E}_{\_}$ that contain's the sample's ground truth class in its class-balanced subset. Thus, for instance, the classifier learns to predict that a manyshot sample lies in the class-balanced subset of the manyshot expert $\mathcal{E}_{Manyshot}$. The full posterior $q(x)$ is then given by $g(\mathcal{E}_{\_}(x))$ for the predicted expert $\mathcal{E}_{\_}$, where $g(.)$ is defined similarly as before.

Model Stacking. We train a single layer linear softmax classifier to predict the full posterior q(x) from the partial posterior vectors $\mathcal{E}_{\_}(x)$ of each expert $\mathcal{E}_{\_}$. The vectors $\mathcal{E}_{\_}(x)$ are concatenated to form a feature embedding for the softmax classifier which is trained by optimising the cross entropy loss on the validation set. This is a standard way for ensemble fusion known as model stacking [39].

Joint Calibration. We calibrate the partial posteriors $\mathcal{E}_{\_}(x)$ by learning scaling and shift parameters before adding up the posteriors similarly to soft-voting,

$$q(x) = \frac{\sum_{\mathcal{E}_{\_}} g(\sigma_{SM}(w_{\mathcal{E}_{\_}} \odot z_{\mathcal{E}_{\_}}(x) + b_{\mathcal{E}_{\_}}))}{\mathbb{Z}}$$

where σ_{SM} denotes the softmax operation, $w_{\mathcal{E}_{\_}}$ and $b_{\mathcal{E}_{\_}}$ are scale and shift parameters respectively, $z_{\mathcal{E}_{\_}}(x)$ denotes the logit scores of expert $\mathcal{E}_{\_}$ for sample x, $\odot$ denotes elementwise multiplication of two vectors, $\mathbb{Z}$ is a normalisation factor, and $g(.)$ is defined as before. We learn scale and shift parameters by minimising the cross entropy loss on the validation set. This module effectively learns the right alignment for experts' partial posteriors before performing soft-voting.

4 Experiments

Datasets. We use the object-centric ImageNet-LT and scene-centric Places-LT datasets for long-tailed recognition, released by Liu et al. [24]. The training set statistics are depicted in Table 1. ImageNet-LT has an imbalanced training set with 115,846 images for 1,000 classes from ImageNet-1K [4].

Table 1. Statistics for training sets in ImageNet-LT and Places-LT.

Datasets	Attributes	Many	Medium	Few	All
ImageNet-LT	Classes	391	473	136	1,000
	Samples	89,293	24,910	1,643	115,846
Places-LT	Classes	132	162	71	365
	Samples	52,862	8,834	804	62,500

The class frequencies follow a natural power-law distribution [32] with a maximum number of 1,280 images per class and a minimum number of 5 images per class. The validation and testing sets are balanced and contain 20 and 50 images per class respectively. Places-LT has an imbalanced training set with 62,500 images for 365 classes from Places-2 [47]. The class frequencies follow a natural power-law distribution [32] with a maximum number of 4,980 images per class and a minimum number of 5 images per class. The validation and testing sets are balanced and contain 20 and 100 images per class respectively.

Evaluation Metrics. We report average top-1 accuracy across the four folds, *Manyshot* classes ($\geq$100 samples), *Mediumshot* classes (20–100 samples), *Fewshot* classes (<20 samples), and *All* classes. Since the test set is balanced across all classes, the average accuracy and mean precision coincide. These four metrics are important for fine-grained evaluation since high accuracy on *All* classes does not imply high accuracy on *Fewshot* classes or *Mediumshot* classes.

Implementation Details. For the *Baseline* model, we take a Resnet-10 backbone for ImageNet-LT, following [24]. We initialise the model with Gaussian weights, use an initial learning rate of 0.2, and train for 100 epochs with a cosine learning rate schedule [25]. For Places-LT, we start with an ImageNet pre-trained Resnet-152 model, and finetune it with 0.01 learning rate for the first 30 epochs followed by 0.1 exponential decay in every 10 epochs. To train expert models, we initialise the feature extractor of each expert $\mathcal{E}_{\_}$ from the *Baseline* model, and finetune it on its class-balanced subset. For $\mathcal{E}_{Mediumshot}$ and $\mathcal{E}_{Fewshot}$, we freeze the lower layers of the feature extractor and only learn the top few layers. The number of learnable layers is a hyperparameter that is fixed by measuring performance on the validation set. To train experts with the reject class, we fix the undersampling ratio for samples from the reject class by measuring performance on the validation set. Note that the classifier for each expert $\mathcal{E}_{\_}$ is smaller than the *Baseline* model; it equals the number of classes in the expert's class-balanced subset, plus an additional reject class.

4.1 Oracle Performance

To estimate the upper bound of our approach, we consider the performance with *Oracle* access to expert selection information, that is, with apriori knowledge of the expert $\mathcal{E}_{\_}$ that contains the ground-truth class of a test sample in its class-balanced subset. The results are depicted in Table 2 and Table 3. The *Oracle* outperforms the *Baseline* by a significant margin on *Mediumshot*, *Fewshot* and *All* accuracy. Moreover, it is significantly interesting to note that the Oracle accuracies on *Mediumshot* and *Fewshot* classes are on par with *Manyshot* accuracy. This illustrates that performance drops on *Mediumshot* and *Fewshot* classes result from *relative*

Table 2. Performance of Oracle vs Baseline on ImageNet-LT.

Method	Many	Medium	Few	All
Baseline	**54.3**	26.2	5.8	34.4
Experts (Oracle)	54.2	**43.3**	**45.7**	**47.9**

Table 3. Performance of Oracle vs Baseline on Places-LT.

Method	Many	Medium	Few	All
Baseline	45.4	25.6	9.0	29.5
Experts (Oracle)	**47.3**	**46.1**	**46.5**	**46.6**

imbalance rather than *data scarcity*. Therefore, in principle, it is possible for a classification model to match *Fewshot* and *Mediumshot* accuracy with *Manyshot* accuracy in modern large-scale datasets. It is also interesting to see that the *Manyshot* accuracy does not improve much by using an *Oracle*, suggesting that *Manyshot* accuracy is already saturated in the *Baseline* model.

4.2 Effect of Joint Calibration Module

We apply the methods outlined in Sect. 3.3 for fusing expert posteriors and compare their performance on ImageNet-LT and Places-LT. The results are depicted in Table 4 and Table 5. KL-div minimisation and Soft-voting yield the highest *Fewshot* accuracy, however *All* accuracy is much lower than the other methods. Expert selection and Stacking are better than KL-div minimisation and Soft-voting on *Manyshot*, *Mediumshot* and *All* accuracy, but worse on *Fewshot* accuracy. The Joint-calibration module obtains the best *Manyshot*, *Mediumshot* and *All* accuracy, even though *Fewshot* accuracy suffers.

Table 4. Effect of joint calibration module for ImageNet-LT.

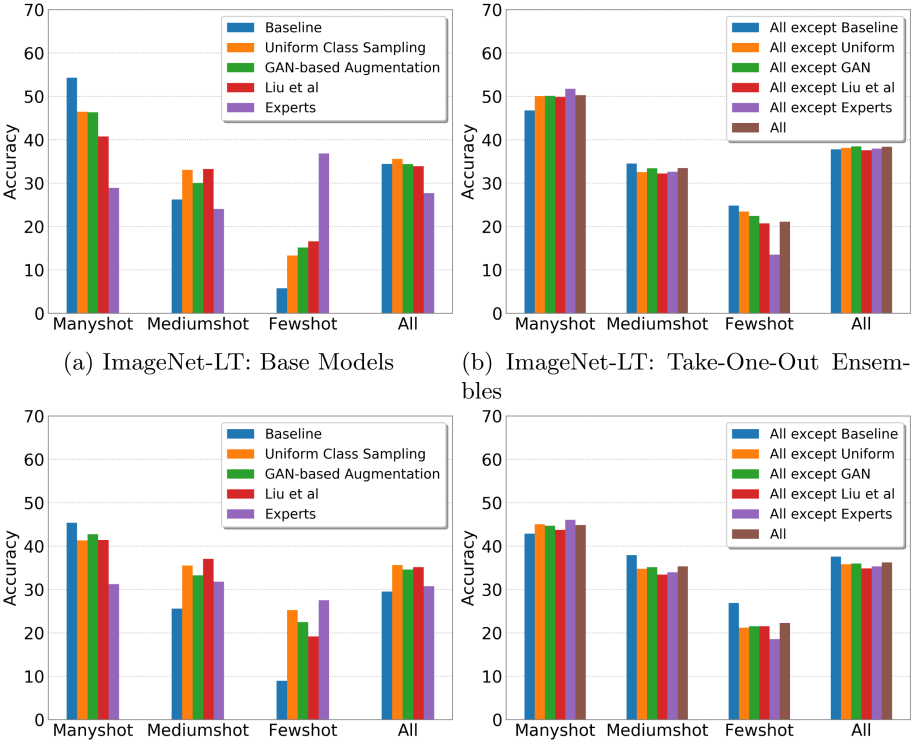
Module	Many	Medium	Few	All
KL-div min	25.3	20.5	**39.1**	21.9
Soft-voting	26.3	21.3	38.9	25.6
Expert selection	38.3	32.6	17.2	32.8
Stacking	28.1	27.5	33.8	28.6
Joint calibration	**43.2**	**34.3**	18.9	**35.7**

Table 5. Effect of joint calibration module for Places-LT.

Module	Many	Medium	Few	All
KL-div min	30.2	31.7	**28.9**	30.4
Soft-voting	30.0	31.8	28.9	30.6
Expert selection	32.6	31.8	24.5	30.7
Stacking	28.2	36.0	26.2	31.3
Joint calibration	**37.2**	**35.3**	26.3	**34.2**

4.3 Diverse Ensembles with Experts

In this section, we extend our ensemble using existing long-tailed recognition solutions and analyse the performance of various combinations of models in the ensemble. We experiment with the following models: (i) The *Baseline* model, (ii) The three expert models, $\mathcal{E}_{Manyshot}$, $\mathcal{E}_{Mediumshot}$ and $\mathcal{E}_{Fewshot}$ fused using Soft-voting, collectively referred to as *Experts*, (iii) Classifier finetuning with uniform class sampling,wherein we freeze the feature extractor of the *Baseline* model and



(a) ImageNet-LT: Base Models

(b) ImageNet-LT: Take-One-Out Ensembles

(c) Places-LT: Base Models

(d) Places-LT: Take-One-Out Ensembles

Fig. 3. From L-R: Performance of - Base Models, and Take-One-Out ensembles. All results are evaluated on the testing set. Top and bottom rows correspond to ImageNet-LT and Places-LT respectively. Best viewed in color with zoom.

finetune the classifier with uniform class sampling. This is referred to as *Uniform class sampling* or *Uniform*, (iv) Data augmentation for *Mediumshot* and *Fewshot* classes using a conditional generative model from class embeddings to feature embeddings, denoted as *GAN based augmentation* or simply *GAN*, (v) Knowledge transfer from *Manyshot* to *Fewshot* classes using a learned convex combination of class prototypes from [24], denoted as *Liu et al.*. The performances of these base models are depicted in Fig. 3a and Fig. 3c. Notice how the performance of the *Baseline* model degrades from *Manyshot* to *Mediumshot* to *Fewshot* accuracy. The *Expert* models give the highest accuracy on the *Fewshot* classes, but are worse on *Manyshot* accuracy.

We combine all these models into a single ensemble, take one model out and see the effect on the performance. To keep the analysis simple, we use Soft-voting for fusing posteriors from all the models, since it doesn't involve learning additional parameters. This ablation is depicted in Fig. 3b and Fig. 3d. As expected, the diverse ensembles give higher *All* accuracy than the base models. Taking

Experts out causes performance drop on *Mediumshot*, *Fewshot* and *All* accuracy, and increase in accuracy on *Manyshot* classes. This suggests that the *Experts* are important in the ensemble for high *Mediumshot* and *Fewshot* accuracy. On the other hand, taking the *Baseline* model out of the ensemble causes an increase in *Fewshot* accuracy while *Manyshot* accuracy drops. The ablation also reveals the inherent trade-off between *Manyshot* and *Fewshot* accuracy; an appropriate combination of models can tilt accuracy in favor of *Manyshot* or *Fewshot* classes.

4.4 Comparison to the State-of-the-Art

We now compare our ensemble of class-balanced experts and the diverse ensemble described in the previous section to the state-of-the-art on the test set of ImageNet-LT and Places-LT. All ensemble combinations use the joint calibration module to fuse model posteriors as it gives us the highest average accuracy. The results are depicted in Table 6 and Table 7. We observe that Ours (Experts) gives us close to state-of-the-art results, and Ours (All) establishes a new state-of-the-art on both the benchmark datasets. This validates our hypothesis that an ensemble of class-balanced expert models is a simple and effective strategy for dealing with long-tailed datasets.

Table 6. Results on ImageNet-LT, using backbone Resnet-10. *Results obtained from the author's code. ‡Results taken directly from [24].

Methods	Many	Medium	Few	All
Lifted Loss‡ [26]	35.8	30.4	17.9	30.8
Focal Loss‡ [23]	36.4	29.9	16	30.5
Range Loss‡ [45]	35.8	30.3	17.6	30.7
FSLwF‡ [9]	40.9	22.1	15	28.4
Liu et al.‡ [24]	43.2	35.1	18.5	35.6
Baseline	**54.3**	26.2	5.7	34.4
Uniform	46.5	33.0	13.3	35.6
GAN	46.4	30.0	15.2	34.4
Liu et al.* [24]	40.8	33.3	16.6	33.9
Ours (*Experts*)	43.2	34.3	18.9	35.7
Ours (*All*)	48.2	**37.0**	**21.5**	**39.2**

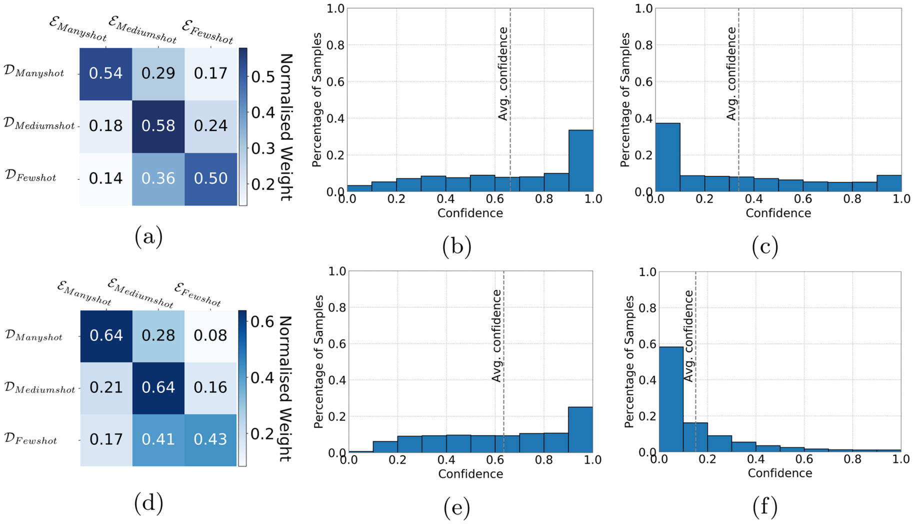
Table 7. Results on Places-LT, using backbone Resnet-152. *Results obtained from the author's code. ‡Results taken directly from [24].

Methods	Many	Medium	Fews	All
Lifted Loss‡ [26]	41.1	35.4	24.0	35.2
Focal Loss‡ [23]	41.1	34.8	22.4	34.6
Range Loss‡ [45]	41.1	35.4	23.2	35.1
FSLwF‡ [9]	43.9	29.9	**29.5**	34.9
Liu et al.‡ [24]	44.7	37.0	25.3	35.9
Baseline	**45.4**	25.6	9.0	29.5
Uniform	41.3	35.5	25.2	35.6
GAN	42.7	33.3	22.5	34.6
Liu et al.* [24]	41.4	37.1	19.2	35.2
Ours (*Experts*)	37.2	35.3	26.3	34.2
Ours (*All*)	43.6	**39.9**	27.7	**38.9**

4.5 Discussion

There is significant difference between the results depicted in Table 2 and Table 3, and Table 6 and Table 7. This shows that the various strategies used for fusing expert posteriors are sub-optimal. To analyse the underlying cause, we take our ensemble of class-balanced experts and plot a confusion matrix, each entry showing the percentage of samples from dataset $\mathcal{D}\_$ that are classified by expert model

$\mathcal{E}_{\_}$. For the preliminary analysis we use Soft-voting for fusing expert posteriors. Figure 4a shows the result for Places-LT. The plot shows there is significant confusion amongst experts; experts aren't selected optimally for classes to which a test sample belongs. We term this phenomenon as *Expert collision*.



(a) (b) (c)

(d) (e) (f)

Fig. 4. Top (bottom): Before (after) joint calibration. L-R: Expert confusion matrix, confidence histograms of $\mathcal{E}_{Manyshot}$ for samples it correctly classifies in $\mathcal{E}_{Manyshot}$, and $\mathcal{E}_{Fewshot}$ for the same samples. All results on Places-LT. Joint calibration aligns experts' confidences and decreases *expert collision*.

We further consider each expert's confidence in its predictions. We take the confidence or the maximum softmax probability (MSP) from the expert posteriors and plot confidence histograms. We do this for $\mathcal{E}_{Manyshot}$ on its class-balanced subset $\mathcal{D}_{Manyshot}$, for samples from the test set it correctly classifies, and for $\mathcal{E}_{Fewshot}$ on the same test samples from $\mathcal{D}_{Manyshot}$. This is depicted in Fig. 4b and Fig. 4c. The plots show that $\mathcal{E}_{Manyshot}$ has high confidence predictions while $\mathcal{E}_{Fewshot}$ has low confidence predictions on these samples. However, to avoid *Expert collision* both the confidence histograms should have a reasonable margin in between and not overlap. Figure 4d and Fig. 4e, 4f respectively show the confusion matrix and confidence histograms after joint calibration. It's essential to align confidences of the three experts correctly, and this is precisely what *joint calibration* does by learning scale and shift parameters for each class.

5 Conclusion

This article presented an ensemble of class-balanced experts framework for long-tailed recognition. Our effective and modular strategy explicitly tackles *relative imbalance* without resorting to complex models or sophisticated loss objectives. We decompose the imbalanced classification problem into balanced classification problems that are more tractable, and train separate expert models for *Manyshot*, *Mediumshot* and *Fewshot* subsets of the data with a reject class for samples lying outside an expert's class-balanced subset. We scale and shift experts' partial posteriors to jointly calibrate experts' predictions, and our ensemble of class-balanced experts reaches close to state-of-the-art performance on two long-tailed benchmarks. We also extend our ensemble with diverse existing solutions for long-tailed recognition and establish a new state-of-the-art on the two benchmark datasets. Moreover, our experiments with an Oracle upper bound reveal that performance drops on *Mediumshot* accuracy and *Fewshot* accuracy are caused by *relative imbalance* and not *data scarcity* for rare classes. Therefore, it is possible to bring *Mediumshot* and *Fewshot* accuracy on par with *Manyshot* accuracy by remedying *relative imbalance* in modern large-scale datasets, which motivates further research in this direction.

References

1. Bengio, S.: The battle against the long tail. In: Workshop on Big Data and Statistical Machine Learning (2015)
2. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. JAIR **16**, 321–357 (2002)
3. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: CVPR (2019)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR (2009)
5. Dong, Q., Gong, S., Zhu, X.: Class rectification hard mining for imbalanced deep learning. In: ICCCV (2017)
6. Estabrooks, A., Jo, T., Japkowicz, N.: A multiple resampling method for learning from imbalanced data sets. Comput. Intell. **20**, 18–36 (2004)
7. Felix, R., Vijay Kumar, B.G., Reid, I., Carneiro, G.: Multi-modal cycle-consistent generalized zero-shot learning. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11210, pp. 21–37. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_2
8. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **42**, 463–484 (2011)
9. Gidaris, S., Komodakis, N.: Dynamic few-shot visual learning without forgetting. In: CVPR (2018)
10. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 1321–1330. JMLR. org (2017)
11. Guo, H., Viktor, H.L.: Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. KDD Explor. Newslett. **6**, 30–39 (2004)
12. Han, H., Wang, W.Y., Mao, B.H.: Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: ICIC (2005)
13. He, H., Garcia, E.A.: Learning from imbalanced data. TKDE **21**, 1263–1284 (2009)

14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
15. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: Proceedings of International Conference on Learning Representations (2017)
16. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. In: Proceedings of the International Conference on Learning Representations (2019)
17. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
18. Huang, C., Li, Y., Chen, C.L., Tang, X.: Deep imbalanced learning for face recognition and attribute prediction. TPAMI **42**, 2781–2794 (2019)
19. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Comput. **3**(1), 79–87 (1991)
20. Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A.: Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **41**, 552–568 (2010)
21. Krawczyk, B., Woźniak, M., Schaefer, G.: Cost-sensitive decision tree ensembles for effective imbalanced classification. Appl. Soft Comput. **14**, 554–562 (2014)
22. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: 6th International Conference on Learning Representations, ICLR 2018 (2018)
23. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017)
24. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: CVPR (2019)
25. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
26. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: CVPR (2016)
27. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1717–1724 (2014)
28. Qi, H., Brown, M., Lowe, D.G.: Low-shot learning with imprinted weights. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5822–5830 (2018)
29. Qiao, S., Liu, C., Shen, W., Yuille, A.L.: Few-shot image recognition by predicting parameters from activations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7229–7238 (2018)
30. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NeurIPS (2017)
31. Sun, Q., Liu, Y., Chua, T.S., Schiele, B.: Meta-transfer learning for few-shot learning. In: CVPR (2019)
32. Van Horn, G., Perona, P.: The devil is in the tails: fine-grained classification in the wild. arXiv preprint arXiv:1709.01450 (2017)
33. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: NIPS (2016)
34. Wang, B.X., Japkowicz, N.: Boosting support vector machines for imbalanced data sets. Knowl. Inf. Syst. **25**, 1–20 (2010)

35. Wang, S., Yao, X.: Diversity analysis on imbalanced data sets by using ensemble models. In: CIDM (2009)
36. Wang, Y.X., Girshick, R., Hebert, M., Hariharan, B.: Low-shot learning from imaginary data. In: CVPR (2018)
37. Wang, Y.-X., Hebert, M.: Learning to learn: model regression networks for easy small sample learning. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 616–634. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_37
38. Wang, Y.X., Ramanan, D., Hebert, M.: Learning to model the tail. In: NeurIPS (2017)
39. Wolpert, D.H.: Stacked generalization. Neural Netw. **5**, 241–259 (1992)
40. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. In: CVPR (2018)
41. Xian, Y., Sharma, S., Schiele, B., Akata, Z.: f-vaegan-d2: a feature generating framework for any-shot learning. In: CVPR (2019)
42. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Feature transfer learning for face recognition with under-represented data. In: CVPR (2019)
43. Yu, N., Shen, X., Lin, Z., Mech, R., Barnes, C.: Learning to detect multiple photographic defects. In: WACV (2018)
44. Yuksel, S.E., Wilson, J.N., Gader, P.D.: Twenty years of mixture of experts. IEEE Trans. Neural Netw. Learn. Syst. **23**(8), 1177–1193 (2012)
45. Zhang, X., Fang, Z., Wen, Y., Li, Z., Qiao, Y.: Range loss for deep face recognition with long-tailed training data. In: ICCV (2017)
46. Zhong, Y., et al.: Unequal-training for deep face recognition with long-tailed noisy data. In: CVPR (2019)
47. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: a 10 million image database for scene recognition. TPAMI **40**, 1452–1464 (2017)
48. Zhu, X., Anguelov, D., Ramanan, D.: Capturing long-tail distributions of object subcategories. In: CVPR (2014)