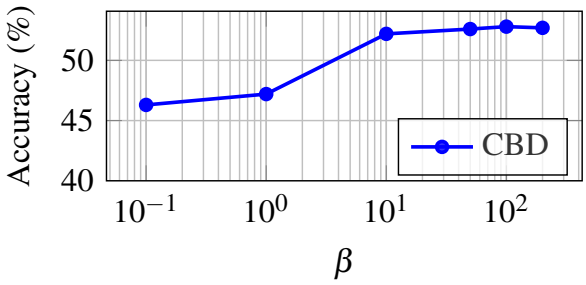

Figure 3: Impact of β scaling hyperparameter for CBD. Experiments are conducted with ResNet-50 on ImageNet-LT validation set with $\alpha = 0.4$.

A Appendix

A.1 Dataset details

We use three long-tailed datasets in our experiments, namely, ImageNet-LT [41], iNaturalist18 [27] and iNaturalist17[54]. ImageNet-LT is an artificially created subset of the original ImageNet dataset [13] where the classes follow a long-tailed distribution. It has 1000 classes and the number of training images per class varies from 5 to 1280. iNaturalist17 and iNaturalist18 training sets are long-tailed by nature. iNaturalist17 contains $5,089$ classes with the range of 9 to 3919 images per class. iNaturalist18 contains $8,142$ classes with the range of 2 to 1000 images per class. The validation and test sets for ImageNet-LT and iNaturalist18 are balanced. The validation set of iNaturalist17 set is more balanced than the training set.

A.2 Training details

Throughout our experiments, the networks are trained for 90 epochs on ImageNet-LT and 200 epochs on iNaturalist17 and iNaturalist18 in both stages, to make the results comparable with the existing work. We also report the performance with more epochs in Section 4.3. When training the model, we use a batch size of 256, learning rate of 0.2 which decays to 0 with cosine learning rate schedule [42], and SGD optimizer with momentum 0.9.

A.3 Impact of β

We first demonstrate the impact of β in Equation (2). This parameter scales the feature distillation loss in the loss objective. Figure 3 shows the accuracy for the ImageNet-LT validation set with different β. We see that the accuracy is relatively stable for $\beta \geq 10$. We set $\beta = 100$ for all of our experiments as this gives the highest accuracy for the validation set.

A.4 Baseline comparison on iNaturalist18

Table 4 reports the comparisons against the baselines on iNaturalist18. Similar to Table 2, we report the accuracy of many-shot, mid-shot, and few-shot classes separately, in addition to the overall accuracy for all classes. Our conclusions are similar to the baseline comparison on ImageNet-LT. When compared to other two-stage models, both CBD and CBD_{ENS} show significant improvements. *Teacher Ensemble* improves the accuracy of many-shot classes

iNaturalist18				
Method	Many-shot	Mid-shot	Few-shot	All
Standard - Instance	76.4	66.2	57.9	64.0
Standard - Class Bal.	59.3	65.2	62.6	63.6
Data Aug. - Instance	74.5	62.5	52.5	59.8
Data Aug. - Class Bal.	55.9	65.1	62.6	63.1
Standard - Fine-tuning	69.6	68.5	66.1	67.6
Standard - Classifier Re-Training	74.1	68.1	64.2	67.2
Data Aug. - Fine-tuning	69.5	68.1	65.3	67.1
Data Aug. - Classifier Re-Training	69.9	65.7	62.9	65.1
Teacher Ensemble	**81.9**	71.9	63.6	69.7
Ours - CBD	70.5	69.5	66.5	68.4
Ours - CBD_{ENS}	75.9	**74.7**	**71.5**	**73.6**

Table 4: **Baseline comparison.** Comprehensive evaluation on iNaturalist18 (validation set) with the ResNet-50 architecture. The accuracy for many-shot , mid-shot and few-shot classes are reported separately.

	ImageNet-LT	iNaturalist18
Instance - NCM	49.0	62.8
Fine-tuning - NCM	48.8	64.1
CBD- NCM	50.7	64.5
CBD_{ENS}- NCM	54.0	69.2

Table 5: **Evaluation with NCM.** Classification accuracy with the non-parametric Nearest Centroid Mean [33] classifier. ResNet-50 architecture is used for both datasets.

at the expense of mid-shot and few-shot classes. This confirms that our method is a better option as a two-stage or ensemble model.

A.5 Quality of the feature representation

We now investigate if our method indeed improves the feature representation. To this end, we measure the accuracy with a non-parametric classifier, *i.e.* *NCM* [19, 33]. NCM computes the classification vectors for each class by taking the mean of all vectors belonging to that class. Thus, the classification accuracy is directly related to the feature representation quality. Table 5 shows the classification performance with the NCM classifier. We test our method against *Instance*, which is the same feature representation used by cRT, and also *Fine-tuning*, which also updates the feature representation in the second stage. It is shown that CBD and CBD_{ENS} achieves higher accuracy, which validates our claims that the feature representation is also improved.

Figure 4 presents qualitative examples that demonstrate the quality of the feature representation produced by CBD, in comparison to the features of *Instance*, *i.e.*, the same representation used in *cRT* method. For each test image, the top-5 nearest neighbor training images are depicted. CBD improves the challenging cases by learning a much improved feature representation for tail classes (*white-sided dolphin*). Features of dolphins and whales seem to be collapsed together with the feature representation used in *Instance*. Our method

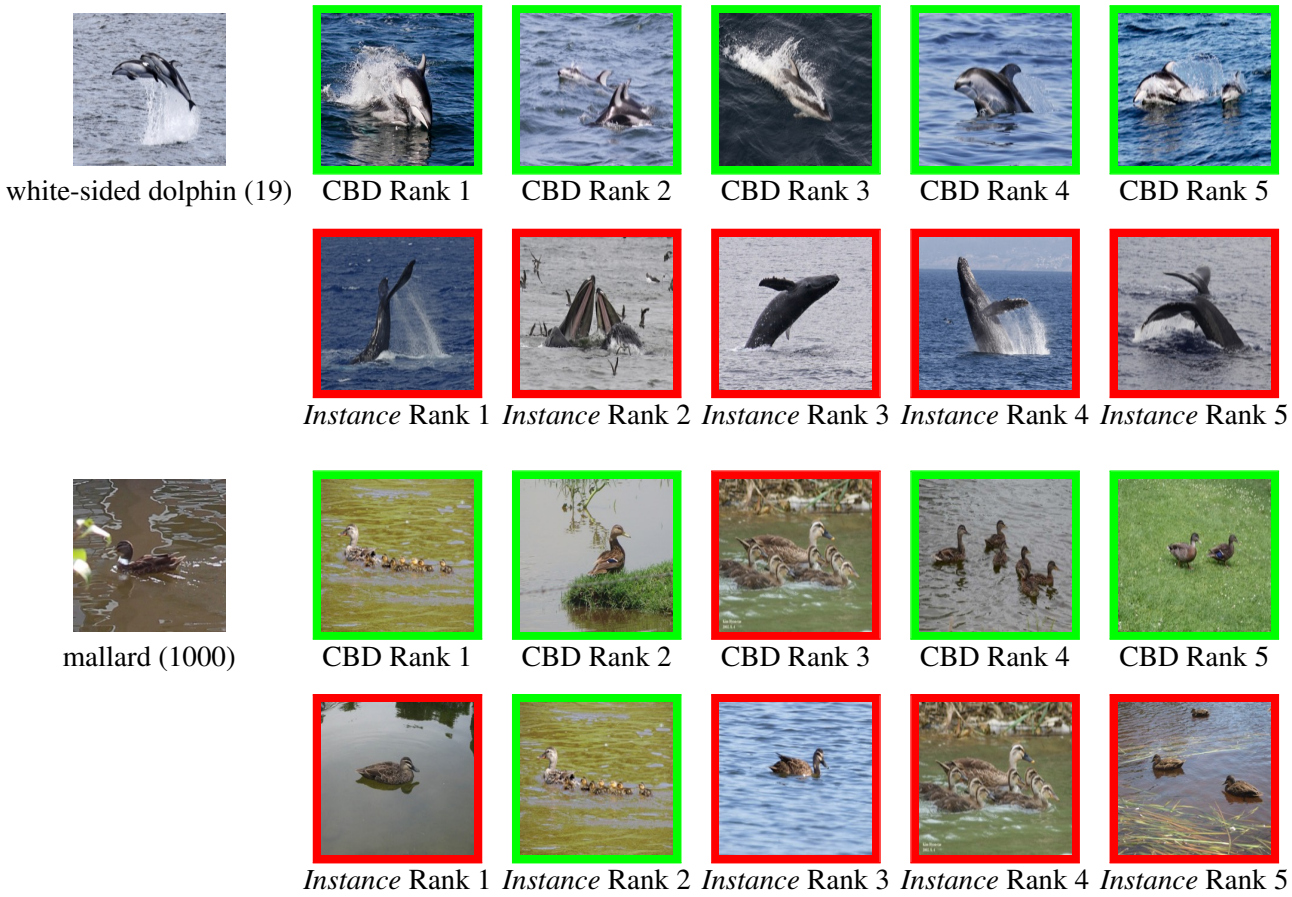

Figure 4: **Qualitative results.** Test images from the iNaturalist18 dataset are depicted on the left, along with their class labels and associated number of training images. For each test image, we show its nearest neighbors from the training set using the feature representations of CBD (top) and *Instance sampling* (bottom). Green and red boundaries denote training images from the same or different classes, respectively.

Method	ResNet-50				ResNet-152			
	Many-shot	Mid-shot	Few-shot	All	Many-shot	Mid-shot	Few-shot	All
LWS [63]	57.1	45.2	29.3	47.7	60.6	47.8	31.4	50.5
cRT [63]	58.8	44.0	26.1	47.3	61.8	46.8	28.4	50.1
cRT+SSP [61]	-	-	-	51.3	-	-	-	-
Logit Adj. [42]	-	-	-	51.1	-	-	-	52.1
ELF(CE) [15]	60.7	45.5	27.7	48.9	-	-	-	-
ELF(LDAM) [15]	64.3	47.9	31.4	52.0	-	-	-	-
Ours - CBD	65.2	48.0	25.9	51.6	68.1	50.1	27.1	53.9
Ours - CBD_{ENS}	**68.5**	**52.7**	**29.2**	**55.6**	**71.2**	**54.5**	**30.7**	**57.7**

Table 6: **ImageNet-LT state-of-the-art comparison.** Comparison of CBD variants against the state of the art with ResNet-50 and ResNet-152.

obtains improvements even for head classes, such as *mallard*, where our learned feature can more easily distinguish similar species based on very detailed information.

A.6 Upper-bound performance

Even though our method brings significant improvements over other baselines, we investigate whether there is further room for improvement. To illustrate the headroom for improvement in terms of feature learning, we propose an experiment where we assume to have the *optimal* feature representation for a long-tailed recognition problem. First, we train a ResNet-50 model $\phi^*_{\theta,W}$ on the full ImageNet [13] dataset. We remove the classifier from this model, only keeping the feature extractor f^*_θ. This feature extractor produces the optimal feature representation for the ImageNet-LT [41] dataset, in the sense that ImageNet-LT is a subset of ImageNet, where the classes follow a long-tailed distribution. By training f^*_θ on the *full* ImageNet, we essentially learn the best possible feature extractor for ImageNet-LT for a given architecture. We now fix f^*_θ and only train the classifier g_W on ImageNet-LT with class-balanced sampling. The resulting model achieves 73.5% top-1 accuracy, which can be seen as an upper bound on ImageNet-LT with ResNet-50. On the other hand, CBD_{ENS} achieves 55.6% top-1 accuracy. This suggests that there is still a lot of room for improvement on long-tailed datasets.

A.7 Comprehensive comparison with State of the Art

Table 3 shows the overall accuracy for the ImageNet-LT, iNaturalist18, and iNaturalist17 datasets. We now show results separately for *low-shot*, *mid-shot* and *many-shot* classes on Tables 6, 7 and 8. Note that the other methods [10, 12, 30, 65] do not report such detailed results on iNaturalist17, thus we could not include them in the comparison. Thus we compare our method with emphStandard Instance, our implementation of classifier re-training [33]. Our method also outperforms the prior work for each class split (many-shot, mid-shot, few-shot) in all scenarios. We do not sacrifice the accuracy of *many-shot* classes to increase the overall accuracy; we achieve a higher overall accuracy by improving *many-shot* accuracy as well as the accuracy of other groups.

Method	ResNet-50				ResNet-101			
	Many-shot	Mid-shot	Few-shot	All	Many-shot	Mid-shot	Few-shot	All
LWS [33]	71.0	69.8	68.8	69.5	73.9	70.4	67.8	69.7
cRT [33]	73.2	68.8	66.1	68.2	71.5	71.3	69.7	70.7
cRT+SSP [61]	-	-	-	68.1	-	-	-	-
Logit Adj. [42]	-	-	-	68.4	-	-	-	70.8
ELF(CE) [13]	67.4	66.3	65.1	66.0	-	-	-	-
ELF(LDAM) [13]	72.7	70.4	68.3	69.8	-	-	-	-
BBN [63]	-	-	-	69.6	-	-	-	-
Ours - CBD	70.5	69.5	66.5	68.4	74.2	71.5	68.3	70.5
Ours - CBD_{ENS}	**75.9**	**74.7**	**71.5**	**73.6**	**77.9**	**76.5**	**73.2**	**75.3**

Table 7: **iNaturalist18 state-of-the-art comparison.** Comparison of CBD variants against the state of the art with ResNet-50 and ResNet-101.

ResNet-50				
	Many-shot	Mid-shot	Few-shot	All
Standard Instance	74.2	55.8	42.9	62.5
Classifier Re-training	71.7	59.3	53.7	63.9
Ours - CBD	70.8	61.0	56.0	64.6
Ours - CBD_{ENS}	**74.3**	**66.4**	**62.0**	**69.3**
ResNet-101				
	Many-shot	Mid-shot	Few-shot	All
Standard Instance	75.2	57.7	45.7	64.1
Classifier Re-training	72.9	60.6	54.9	65.2
Ours - CBD	73.3	62.5	57.9	66.5
Ours - CBD_{ENS}	**76.8**	**68.1**	**63.2**	**71.3**

Table 8: **iNaturalist17 comprehensive comparison.** Comparison of CBD variants against the other methods with ResNet-50 and ResNet-101. The accuracy for many-shot ($>$ 100 images), mid-shot (20-100 images) and few-shot ($<$ 20 images) classes are reported separately.