Background Splitting: Finding Rare Classes in a Sea of Background

Ravi Teja Mullapudi[2,3]⋆ rmullapu@cs.cmu.edu
Fait Poms[1]⋆ fpoms@cs.stanford.edu
William R. Mark[3] billmark@google.com
Deva Ramanan[2] deva@cs.cmu.edu
Kayvon Fatahalian[1] kayvonf@cs.stanford.edu

Abstract

We focus on the problem of training deep image classification models for a small number of extremely rare categories. In this common, real-world scenario, almost all images belong to the background category in the dataset. We find that state-of-the-art approaches for training on imbalanced datasets do not produce accurate deep models in this regime. Our solution is to split the large, visually diverse background into many smaller, visually similar categories during training. We implement this idea by extending an image classification model with an additional auxiliary loss that learns to mimic the predictions of a pre-existing classification model on the training set. The auxiliary loss requires no additional human labels and regularizes feature learning in the shared network trunk by forcing the model to discriminate between auxiliary categories for all training set examples, including those belonging to the monolithic background of the main rare category classification task. To evaluate our method we contribute modified versions of the iNaturalist and Places365 datasets where only a small subset of rare category labels are available during training (all other images are labeled as background). By jointly learning to recognize both the selected rare categories and auxiliary categories, our approach yields models that perform 8.3 mAP points higher than state-of-the-art imbalanced learning baselines when 98.30% of the data is background, and up to 42.3 mAP points higher than fine-tuning baselines when 99.98% of the data is background.

1. Introduction

Image classification tends to be evaluated on manually curated datasets that are clean and balanced [32, 20]. Recent focus has shifted toward in-the-wild data that features long-tail distributions [39, 33, 43, 44, 47], but these datasets are still "artificial" in that every curated image is a positive example of some category [23, 23, 3, 3, 39]. In contrast, in many real-world settings, categories of interest are sufficiently rare that it is far more common for images to *not exhibit any*

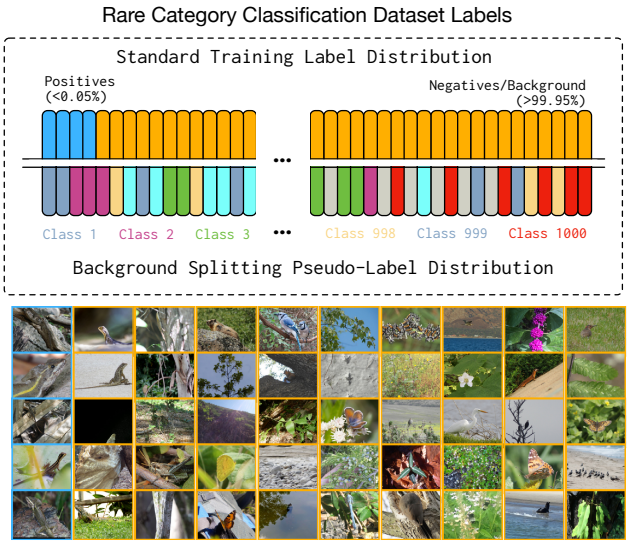

Figure 1. Top: Training classification models for rare categories is difficult because of the limited information content and extreme label imbalance provided by a large "background" category. During training we address these issues by *splitting the background category* into a large set of smaller categories according to pseudo-labels produced by an existing, pre-trained model. Bottom: positive (blue outline) and negative (orange outline) images for a fine-grained lizard category in iNaturalist. Even though the negatives all have the same label (background) they encompass a diverse set of hard negatives (hard to distinguish from the positives) and a vast number of easy negatives. Background splitting extracts value from all negatives by forcing the model to predict pseudo-labels for all images.

⋆Both authors contributed equally to this paper. [1]Stanford University [2]Carnegie Mellon University [3]Google Research

categories of interest. As a consequence, it is typically easy to collect a large number of easy negative instances (e.g., with weak- or semi-supervised methods [30, 5], or as a by-product of multi-label annotation [7], or with random sampling), but finding positives and hard negatives is difficult. For example, consider a newly collected dataset (e.g., obtained by an autonomous vehicle fleet) where a new category – an e-scooter – is annotated. This is a highly imbalanced binary image classification problem, as the vast majority of collected images will *not* contain an e-scooter (sparse positives). If additional categories are desired, the task becomes a multi-way image classification problem where the vast majority of images belong to the "background" category.

This work focuses on the problem of training accurate deep models for image classification of a small number of rare categories. In these scenarios, not only is there a small number of positives per category, but the overall number of positive examples for all categories is dominated by the number of background images (e.g., our experiments include cases where 99.98% of the dataset is background). We find that even state-of-the-art methods for deep imbalanced classification based on data sampling [13] or loss re-weighting [3] fail to produce accurate models in this extremely imbalanced setting.

Contributions: We propose a surprisingly simple method to address learning challenges posed by severe background dominance: we *split the visually diverse, but monolithic background category* into many smaller, visually similar categories during training. However, rather than modify the main (rare category) classification loss of the model to identify more categories, our approach adds an auxiliary loss that forces the model to mimic the predictions of an existing, pre-trained image classification model *on all training examples*, the vast majority of which constitute "easy" background instances in the main rare category classification task. Jointly learning using the main classification loss and the auxiliary classification loss regularizes the rare category classification model by forcing it to discriminate between pseudo-categories for all training examples and helps reduce over-fitting to the small number of rare category positives. Most importantly, this solution transfers knowledge of an auxiliary classification task into the target model via distillation—it only requires access to the pre-trained model, *not access to any additional labels* beyond those used for the main classification task.

Benchmarking: Although background dominant scenarios are common in real-world applications of computer vision, no academic datasets directly target this important task. To evaluate our methods and facilitate future rare category learning research, we contribute modified label distributions for two existing large-vocabulary datasets: the highly-imbalanced species-classification iNaturalist dataset [39], and the scene-classification Places365 dataset [45]. In both cases we select only a small number of categories as targets of interest, and merge all other categories into a single, large background category. These new label sets model real-world training scenarios in that they contain a small number of rare categories, and a vast majority of images exhibit only background. (We will release these modified label distributions to the public.)

Analysis: We find that state-of-the-art techniques for imbalanced classification [13, 3] perform *worse* than standard fine-tuning in scenarios where the background category makes up more than 98% of the dataset. In contrast, background splitting outperforms prior art by 8.9 mAP points. In extreme cases of a single, rare foreground category (e.g., 99.98% background), background splitting yields mAP improvements from 10.6 to 52.9. We also evaluate the benefits of background splitting under different amounts of background dominance, and varying semantic overlap between the auxiliary task and main task categories.

2. Related Work

Category imbalance. Training deep classifiers on heavily imbalanced data is challenging because standard losses focus on majority categories (failing to learn minority categories) or over-fit to the few positive examples of minority categories. Most methods for learning under long-tail, imbalanced datasets [10] rely on techniques such as re-balancing, re-weighting/category-conditioned adjustments to final loss values, and category clustering. **Re-balancing methods** alter the training distribution to simulate traditional balanced training sets [3, 40]. Recent work showed re-balancing should be limited to a final stage of model training to encourage more general feature representations [13, 3, 42]. **Re-weighting/category-conditioned methods** adjust the loss attributed to a sample based on its category [46, 3, 13] or how hard the sample is. **Category clustering methods** use clustering of embeddings from training samples to improve transfer learning from head to tail categories [23] and to train submodels specialized to individual categories (avoiding imbalance) [47, 28, 17].

Instead of altering the training distribution through re-balancing or re-weighting, background splitting keeps the training distribution for the main loss fixed and instead adds an additional auxiliary loss that forces the model to make fine-grained distinctions among images in the background category (the model must perform a challenging task even for "easy" background images). This approach encourages the feature representation

to be more discriminative while still learning from the unaltered primary training distribution.

Large, diverse background. Background dominant scenarios are an important real-world case of imbalanced classification where instances of a large, diverse background category are more common than any foreground category. Prior work has modeled the background by adding an additional "N+1" category to represent the background [25, 24, 22] (which we use in our method, Sec. 4.2), modeling the background as a Gaussian distribution [27], or by training the model to predict low scores for background instances and then applying a threshold to filter them [6, 25]. Handling a large, diverse background in classification is related to the foreground-background class imbalance problem in object detection [26], where ideas such as rebalancing (hard-negative mining [36]) and re-weighting (focal loss [19]) are employed to improve object detection performance. We focus on background imbalance problem in the context of image classification, but our findings may also benefit object detection tasks as well.

Open-world, open-set, and out-of-distribution recognition. Unlike the background category setting (in which models are given access to training instances which are from the same distribution as the test instances), Open World, Open Set, or Out-of-distribution recognition datasets test on data from categories (Open World/Open Set) or entire datasets (Out-of-distribution) not seen during training [23, 1, 2, 6]. In this setting, the model is tasked with identifying "unknown unknowns" in the test distribution that are not present in the training distribution. Typically, these "unknown unknowns" are not a significant majority of the test distribution. In contrast, we are interested in "known unknowns" which are present in the training distribution, but make up a majority of the data.

Knowledge transfer and sharing. Our approach leverages knowledge contained in a pre-trained model to improve performance on the rare category classification task. Our approach modifies standard **transfer learning** methods [29, 31, 34, 15] by introducing an auxiliary distillation loss to improve model performance in background dominant scenarios. Thus our solution is similar in spirit to **multi-task learning** methods [4, 16], in particular those addressing the challenges of incremental learning [18, 37, 14, 35]. However, our approach uses supervision from an existing classifier to regularize training on the rare-category classification task. It does not aim to train a model that accurately classifies both prior and novel categories. Instead of transferring knowledge exclusively via fine tuning, we transfer knowledge using a combination of fine tuning and **knowledge distillation** [12, 41], where the distillation is performed using data from the new domain rather than data from the original domain. This approach to knowledge transfer has been shown to outperform fine tuning for some tasks [38, 21], and is also useful when the source and destination models are different [9].

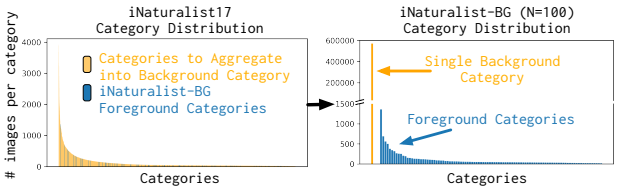

Figure 2. Left: distribution of images per category in the iNaturalist-2017 training set, sorted by category frequency. Although there is no "background" category in iNaturalist, we color categories based on whether they are placed in the background category in our modified dataset, iNaturalist-BG. Right: distribution of examples in iNaturalist-BG ($N=100$). The background category (yellow) contains 98.3% of the images in the dataset.

3. Large Background Datasets

To support image classification research in real-world, large-background settings, we created modified variants of the iNaturalist 2017 dataset and the Places365 dataset, which we call iNaturalist-BG and Places-BG. We selected iNaturalist 2017 because of its large (and imbalanced) category vocabulary (5089 categories, including many visually similar species), and because it is based on a real-world use case of identifying the world's flora and fauna. (iNaturalist was collected by thousands of real individuals interested in species identification, not by choosing a list of categories then collecting images by querying internet search engines [23, 3]). Places365 exhibits a more modestly sized vocabulary (365 categories), but contains content that differs significantly from iNaturalist, making the pair of datasets a good test of generalizability of rare category methods.

To construct our modified datasets, we consider a small number (N) of the original dataset's categories to be "labeled" and combine the remainder of the categories from the original dataset into a single background category. To study how the size of the background category affects model performance, we provide training and test set pairs for varying N. For example, for iNaturalist we create pairs for $N=5089$ (equivalent to the standard iNaturalist 2017 category distribution), 1100, 100, 10, and 1). Figure 2-right shows the distribution of images to categories for the iNaturalist-BG ($N=100$) training set. In total, the 100 chosen categories constitute only 1.7% of the dataset (98.3% background).

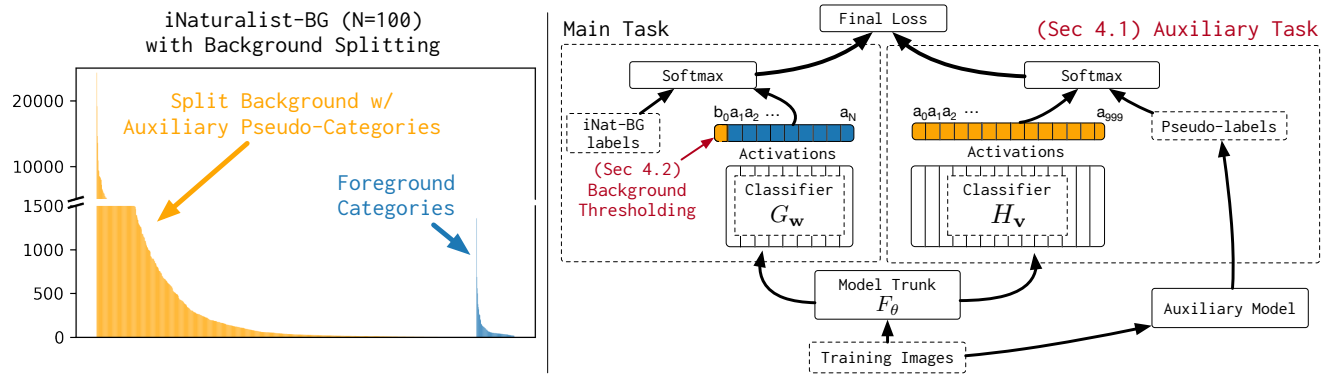

Figure 3. Left: distribution of images for the iNaturalist-BG ($N = 100$) dataset after partitioning into pseudo-categories determined by the output of a pre-trained auxiliary model (in this example, 1000 ImageNet categories). The largest pseudo-category is only 4.5% of the dataset, far smaller than the iNaturalist-BG "background" category (98.3%). Right: the training configuration for our method. We use two tasks, one supervised by labels from the target dataset (e.g., iNaturalist-BG) and one supervised by pseudo-labels from the auxiliary model. The main loss uses a fixed background logit to improve background classification.

Note that in all cases, models are trained on all images in the original dataset's training set (579,184 for iNaturalist, 1.8 million for Places), and tested on the dataset's full test set. (Images not belonging to the N categories of interest are labeled "background".)

4. Method

When training a model for classifying a small set of rare categories (*main task*), we want to avoid overfitting to the small set of foreground positives and avoid solutions which predict all instances are part of the background category. We address these challenges by transforming the model optimization problem into a easier one with significantly lower distribution skew. We also describe how we combine our background splitting technique with the common softmax thresholding method introduced by Matan et al. [25].

4.1. Regularization via Auxiliary Loss

In addition to supervising the model with the main task's cross-entropy loss for $(N+1)$-way classification during training (the additional +1 category representing the background), we also add an *auxiliary task* by attaching a second classification head to the network supervised by pseudo-labels generated with using pre-trained model. (Figure 3-right, "Auxiliary Task"). Let $y \in \{0, \dots N\}$ be the main $(N+1)$-way classification task where 0 is treated as the background class. Given a training set with a large fraction of background examples, we assume access to an *auxiliary model* with K classes that provides pseudo-labels on all training pairs $\{(x_i, y_i)\}$:

$$\{(x_i, y_i, t_i)\} \qquad y_i \in \{0, \dots N\}, \quad t_i \in \{1, \dots K\}.$$

Figure 4. An auxiliary model pre-trained to perform ImageNet category classification groups semantically similar iNaturalist images together, resulting in pseudo-labels that split the large background category in iNaturalist-BG. Each row represents a category chosen at random from the auxiliary model categories (ImageNet categories). The image on the left is an example of the category taken from the auxiliary model training set (ImageNet). Images on the right are the auxiliary model's five most confident predictions for that category on the iNaturalist dataset. Notice that rows 1, 3, 4, and 5 contain visually similar images even though the auxiliary model category is different from the animals in the iNaturalist images.

where t_i is the auxiliary model's pseudo-label. We then learn a classifier with a multi-task loss:

$$\min_{\theta,\mathbf{w},\mathbf{v}} \sum_i \text{loss}\Big(y_i, G_{\mathbf{w}}(F_\theta(x_i))\Big) + \lambda_H \text{loss}\Big(t_i, H_{\mathbf{v}}(F_\theta(x_i))\Big)$$

where G and H refer to the main and auxiliary task classification heads, respectively, of a base network trunk F with shared features θ and task-specific features $\mathbf{w}$ and $\mathbf{v}$. λ_H is the weight for the auxiliary loss. The loss function is the standard softmax cross-entropy loss. Note that the pseudo-labels t_i are generated by evaluating the auxiliary model on all training data. *No additional*

human labeling effort is required beyond the labels y_i provided for the main rare category classification task. Figure 3-right visualizes the full training graph for our multi-task network, which defines task-specific linear weights $(\mathbf{w}, \mathbf{v})$ on the shared feature trunk θ. Note that at test time, performing the main rare category classification task requires only evaluating $G_{\mathbf{w}}(F_\theta(x))$ (for test sample x).

During training, the role of the auxiliary task is to effectively "split up" the large background category into a large number of visually coherent sub-categories. This reduces distribution skew and forces learning of robust features θ that discriminate between the many psuedo-categories, even for the large majority of training data that would otherwise serve as a "easy" background examples for the main classification task. Figure 3-left shows the results of splitting the iNaturalist-BG ($N = 100$) background category into K=1000 categories defined by an auxiliary model that classifies images into ImageNet categories. Although the background category is 98.3% of the training set in this case, the most frequently occurring auxiliary task category contains only 4.5% of the dataset.

Importantly, a direct mapping between auxiliary model categories and the content of the dataset being evaluated on is not necessary for visually coherent groupings to occur. Figure 4 illustrates that visual coherence of sub-categories that emerge when evaluating an ImageNet category classification model on the iNaturalist dataset. Rows 1, 3, 4, and 5 contain visually similar images, even though they do not contain the animal the auxiliary model category was trained for. (Row 1's category is "flamingo", but the birds at right are a different species of pink bird; Row 4's category is "European fire salamander" but the images to the right are of owls). In Section 5.3, we evaluate how different choices for the auxiliary task influence the extent auxiliary loss helps main task classification performance.

4.2. Background Thresholding

The straightforward approach to performing classification with a background category modifies the N-way classification problem into an $(N+1)$-way classification problem. However, linear (softmax) classification encourages examples that fall into the same category to have similar features that can be linearly separated from other classes. This may be problematic for the background class, which we expect to be very large and diverse in our setting (Figure 2-left). We address this issue by *choosing to not* learn a classifier for the background category. Instead, we adopt the method of Matan et al. [25], which assigns the background category (category 0) a fixed activation, as represented by b_0 in Figure 3-right. (See supplemental for full details of the modified loss.)

5. Evaluation

We perform an in-depth analysis of background splitting on the iNaturalist-BG dataset by comparing its performance to strong baselines in multiple different regimes of background dominance. We also provide diagnostic experiments that explore the sensitivity of background splitting to different choices of auxiliary task and show the necessity of jointly optimizing the main task loss with the auxiliary loss. Finally, we also demonstrate the benefits of background splitting on Places-BG, suggesting applicability to a wide range of classification tasks.

5.1. Experimental Setup

Evaluation metrics. Traditional metrics for image classification include top-1 or top-5 error, but these metrics can be manipulated by constructing models that favor the background. (Always predicting background yields a top-1 accuracy of 98.3% on iNaturalist-BG N=100.) Instead, we propose two evaluation protocols for datasets with N category labels plus a large background category. The first protocol is motivated by top-1 error: algorithms must report a single $(N+1)$-way label for each test sample. It computes the F1 accuracy (harmonic mean of precision and recall) for each of the N classes and reports their mean. The second protocol, motivated by those used for large-background tasks such as object detection [8], recasts the problem as N retrieval tasks corresponding to each foreground category. For each category, algorithms must return a *confidence* for each test sample. These are ranked to produce N precision-recall curves, which in turn are summarized by their average area underneath (mAP).

For iNaturalist-BG, we evaluate using label sets that range from no background ($N = 5089$) to increasing levels of background dominance (N=1100, 100, 10, 1). Even though the foreground category distribution varies with N, we construct training/set pairs so that evaluation metrics remain comparable across N. Specifically, we evaluate performance using a fixed set of 100 categories. When $N > 100$, we only compute mean F1 or mAP from the selected 100 categories. The $N = 10$ and $N = 1$ setups do not contain all 100 categories, so we provide 10 test set pairs for $N = 10$ (spanning all 100 test categories) and average performance across these subsets. Computational constrains required us to limit evaluation to 10 pairs for $N = 1$. (Evaluation averages over only 10 of the 100 test categories, so $N = 1$ results are not directly comparable to results for other N.)

Model and training details. In our background splitting configurations (referred to as BG-SPLIT) we use the ResNet-50 architecture [11] (initialized via ImageNet pre-training) in all experiments. We set the background threshold value to $b_0 = 0.1$, the weight on the auxiliary loss to $\lambda_G = 0.1$, and batch size to 1024. We find that large batch sizes are crucial for training models when the background category is dominant, and provide an additional evaluation of the effect of batch size on different methods in the supplement.

Unless otherwise stated, we use a standard ResNet50 model trained on the 1000-category ImageNet [32] dataset, as an auxiliary model. We evaluate alternative choices for auxiliary models in Sec. 5.3.

Baselines. We compare our method to standard fine-tuning with cross entropy loss (denoted as FT in the rest of the paper), as well as two state-of-the-art baselines for training on the iNaturalist dataset: LWS [13] which performs sample re-balancing; and LDAM [3], which performs loss re-weighting. LDAM and LWS have been shown to perform better than a broad set of other imbalanced learning methods, including recent methods from the object detection literature such as focal loss [19]. LWS trains in two stages: first training a standard cross-entropy model with uniform sampling and no re-weighting for 90 epochs; then freezing all layers but the final classification layer and retraining for 15 epochs using class-balanced sampling. LDAM also uses a two-stage training approach: first training a classification model using uniform sampling and weighting samples with a Label-Distribution Aware Margin (LDAM) loss for 60 epochs; and then continue training the same model by re-weighting the loss for individual examples based upon their frequency for 30 epochs. We train these baseline models using the official code provided by each method .

We tuned the learning rate and batch size for FT via hyperparameter search. (We use a batch size 512, see supplement for motivation of this large batch size). For LWS when $N = 5089$ we train the base representation model for the first stage using published hyperparameters [13]. When $N < 5089$ we use the same hyperparameters as FT. We report results for the best-performing second-stage method (*Tau Norm* for $N = 100$ and 1100, *LWS* for $N = 5089$).

We refer the reader to the supplement for comparison to further baselines, such as background downsampling and focal loss [19].

LWS: https://github.com/facebookresearch/classifier-balancing

LDAM: https://github.com/kaidic/LDAM-DRW

5.2. Comparison with Baselines

iNaturalist-BG comparison. Table 1 compares the mAP and F1 scores of models trained using BG-SPLIT against all baselines on iNaturalist-BG. LDAM and LWS perform slightly better than BG-SPLIT in the N=5089 case they were designed for (original iNaturalist setting, no background), however as background size increases (N=1100, 78.0% background) and (N=100, 98.3% background), both baselines degrade rapidly, *performing worse* than the FT baseline, even when using the best hyperparameter configurations found for each N. In the N=100 configuration BG-SPLIT outperforms the best results of LDAM and LWS by 11.7 (mAP) and 4.1 (F1) points. We observe that class-balanced sampling (LWS) and loss re-weighting based on class frequency (LDAM) cause accuracy on the dominant background category to decrease, yielding a significant increase in false positives for foreground categories. This results in a minor increase in foreground category recall, but a large reduction in precision and overall worse performance for LDAM and LWS in the sparse positive setting. Since the performance of LDAM and LWS further degrades with decreasing N, we do not evaluate these methods for $N = 10$ and $N = 1$ to save experimental costs.

The value of BG-SPLIT increases as background dominance is further increased. In the extreme case of N=1 (98.98% background), BG-SPLIT *beats* FT *by 42.3 (mAP) and 40.9 (F1) points.* We note that the extreme background dominance in the N=1 configuration is typical of many real-world applications. The overall performance of BG-SPLIT decays gracefully with decreasing N without modifying key hyperparameters (e.g., learning rate, batch size, b_0, λ_G). Hyperparameter robustness across different levels of background imbalance is an attractive property of our method which makes it easy to use in practice.

Places-BG comparison. Table 2 compares BG-SPLIT to FT on two configurations of Places-BG: N=10 (97.2% background) and N=1 (99.7% background). Despite Places-BG containing notably different categories and image content than iNaturalist-BG, the overall trends on Places-BG are similar. The benefit of BG-SPLIT is significant and increases with increasing background dominance. However, for a given N, the benefits of BG-SPLIT are lower on Places-BG than for iNaturalist-BG because Places-BG contains a smaller number of categories with higher per-category frequency, making the foreground category classification problem easier (helping the FT baseline). Note that N=10 background frequency on Places-BG is similar to that of N=100 on iNaturalist-BG.

Model Performance: iNaturalist-BG

	$N=1$ (99.98%)		$N=10$ (99.83%)		$N=100$ (98.30%)		$N=1100$ (77.95%)		$N=5089$ (0%)	
	mAP	F1	mAP	F1	mAP	F1	mAP	F1	mAP	F1
FT	10.6	10.8	9.3	8.9	38.3	38.4	50.9	**44.8**	57.7	49.9
LDAM	-	-	-	-	24.7	21.1	41.6	37.1	57.4	55.1
LWS	-	-	-	-	35.5	36.6	42.5	37.3	**60.0**	**56.9**
BG-SPLIT	**52.9**	**51.7**	**44.6**	**39.4**	**47.2**	**40.7**	**51.4**	44.7	59.9	52.5

Table 1. **BG-SPLIT outperforms all baselines when the background frequency exceeds 98% on iNaturalist-BG.** Comparison of mAP and F1 scores of BG-SPLIT to baselines on the iNaturalist-BG dataset (percentages indicate background frequency). BG-SPLIT is 8.3 mAP points more accurate than state-of-the-art baselines in the $N=100$ setting, and *42.3 points higher than fine tuning* (FT) in the extremely imbalanced $N=1$ setting that is typical of many real-world image classification scenarios. Prior state-of-the-art baselines for imbalanced training (LDAM [3], LWS [13]) perform worse than FT in background dominated settings. ($N=1$ results are not directly comparable to other choices of N, see Section 5.1).

Model Performance: Places-BG

	$N=1$ (99.72%)		$N=10$ (97.22%)	
	mAP	F1	mAP	F1
FT	44.0	41.3	56.8	51.4
BG-SPLIT	**53.8**	**48.7**	**61.0**	**56.6**

Table 2. **BG-SPLIT outperforms FT on Places-BG.** The benefit of BG-SPLIT over FT increases with increasing background size (percentages indicate background frequency). Overall, the magnitude of benefit of BG-SPLIT is less on Places-BG than on iNaturalist-BG (Table 1) because categories in Places-BG are on average over 13× more frequent than iNaturalist-BGcategories (0.27% vs. 0.02%). Places-BG categories are also less fine-grained than those in iNaturalist-BG, making the background easier to discriminate.

5.3. Sensitivity to Choice of Auxiliary Task

A goal of background splitting's auxiliary task is to leverage the vast number of background examples to learn useful representations that aid the main classification task. To understand the sensitivity of background splitting's benefits to auxiliary tasks that exhibit varying degrees of similarity to the main classification task, we trained BG-SPLIT models on iNaturalist-BG with the following auxiliary tasks: pseudo-labeling via a pre-trained ImageNet classification model (IMAGENET/CLASSIFIER), pseudo-labeling via pre-trained image classification model for the Places365 dataset[45] (PLACES/CLASSIFIER), performing approximate k-means clustering on features generated by the pre-trained ImageNet model (IMAGENET/CLUSTER-1K, IMAGENET/CLUSTER-5K), and assigning 1000 pseudo-categories to training images at random (RANDOM-1K).

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MiniBatchKMeans.html

Aux Task Ablations: iNaturalist-BG (N=100)

Auxiliary Task		BG-SPLIT	
Source Dataset	Label Method	mAP	F1
— *No auxiliary loss* —		41.1	37.7
None	Random-1K	37.2	35.0
Places365	Classifier	39.3	34.1
ImageNet	Cluster-1K	45.9	39.0
ImageNet	Cluster-5K	45.9	**41.0**
ImageNet	Classifier	**47.2**	40.7

Table 3. **Choice of auxiliary task significantly impacts BG-SPLIT model performance.** Due to similarity between ImageNet classification and iNaturalist-BG classification tasks, pseudo-labels from ImageNet-based auxiliary tasks increase the performance of models trained using BG-SPLIT on iNaturalist-BG, even when categories are machine-defined (IMAGENET/CLUSTER-1K, IMAGENET/CLUSTER-5K). Auxiliary tasks based on random pseudo-labels (RANDOM-1K) or significantly different classification tasks (PLACES/CLASSIFIER) are destructive to model training for iNaturalist-BG. Note that the *no auxiliary loss* configuration is similar to that of FT (N=100) in Table 1, but with two differences for consistency within this table: *no auxiliary loss* uses batch size 1024 instead of 512, and uses the background thresholding loss modification.

Table 3 summarizes results for iNaturalist-BG ($N=100$), where all methods are trained using a batch size of 1024 and include the background thresholding loss modification. RANDOM-1K, which splits the background into equal-sized categories, but yields categories that lack visual or semantic coherence, *degrades performance* compared to a training configuration that does not employ any auxiliary task (it is destructive to feature learning).

Feature Learning Ablations: iNaturalist-BG (N=100)

Method	Step 1 loss	Freeze/FT	Step 2 loss(es)	mAP
Linear model on ImageNet features	–	Freeze	Main	25.5
Decoupled aux/main	Aux	Freeze	Main	28.3
BG-SPLIT	–	Fine tune	Main & Aux	47.2

Table 4. **Combining sparse foreground category labels with pseudo-labels is critical to feature learning.** Feature learning using only auxiliary task loss (row 2) provides improvement over initial ImageNet features (row 1). However a model trained using BG-SPLIT (row 3) is over 18 mAP points higher.

PLACES/CLASSIFIER also degrades performance because scene-specific classification on Places365 is a significantly different task than species classification. (Table 2 showed ImageNet classification is a beneficial auxiliary task for Places-BG but Places365 classification is a destructive auxiliary task for iNaturalist-BG classification.) However, since the ImageNet dataset has a large number of categories (1000), including some with semantic category overlap with iNaturalist (e.g., many animal categories), there is substantial improvement from auxiliary tasks based on ImageNet models. For example, even though IMAGENET/CLUSTER-1K and IMAGENET/CLUSTER-5K generate pseudo-labels for machine-generated categories, these auxiliary tasks significantly improve final model performance. Although IMAGENET/CLASSIFIER yields the highest mAP gain, the fine-scale clusters of IMAGENET/CLUSTER-5K are the pseudo-labels that produce the highest F1 gain.

5.4. Value of Joint Training

Can the auxiliary loss alone produce an effective feature representation without target foreground category labels? If so, one could use the auxiliary loss for dataset pre-training, and leverage the resulting representations to rapidly learn a model for novel foreground categories. To isolate the value of background splitting's joint training approach vs. supervision from main task and pseudo-labels alone, we conducted an ablation experiment where we separated the aux loss training and main loss training into separate phases, and froze the features after the first phase.

Table 4 compares the performance of the resulting models on iNaturalist-BG (N=100). Even though pre-training using only the auxiliary loss (row 2) keeps 98.3% of the labels the same as the full BG-SPLIT solution (row 3), the BG-SPLIT method is over *over 18 mAP points higher* than this decoupled method. These results suggest that feature learning via joint training from sparse positive examples and abundant pseudo-labels from the auxiliary task is critical.

BG-SPLIT Component Analysis: iNaturalist-BG (N=100)

None (FT)		Aux loss only		BG thresh only		Both (BG-SPLIT)	
mAP	F1	mAP	F1	mAP	F1	mAP	F1
36.0	35.6	46.0	40.4	41.1	37.7	47.2	40.7

Table 5. **Most of the benefits of the full BG-SPLIT solution are due to the use of an auxiliary loss.** However, use of auxiliary loss and background logic clamping are complementary techniques (additive benefits), suggesting auxiliary losses could be combined with other techniques for learning on imbalanced data.

5.5. Component Analysis

The full BG-SPLIT solution has two new components: an auxiliary loss and the background thresholding modification to the main task loss. Table 5 isolates the relative gains due to each of these components on iNaturalist-BG (N=100). The majority of the benefit of BG-SPLIT comes from the use of the auxiliary loss (10 mAP points "aux loss only"). Background logit thresholding in isolation yields 1.1 mAP points ("bg thresh only"). These techniques are complementary techniques (additive effects in the final solution), suggesting that the benefits of using an auxiliary loss could be combined with other state-of-the-art techniques for imbalanced learning such as LDAM.

6. Conclusion

State-of-the-art classification methods for handling imbalanced data do not perform well in the presence of a large and diverse background. In response we contribute a new approach that reduces the imbalance by jointly training the main classification task along with an auxiliary classification task that involves categories that are better balanced and less rare. This approach improves on baselines by over 42 mAP points in situations of significant background dominance, making it feasible to train useful binary classification models for rare categories without additional human-provided labels. We also contribute the iNaturalist-BG and Places-BG datasets, which we hope will encourage further research in this important training regime.

References

[1] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1893–1902, 2015. 3

[2] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016. 3

[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1565–1576, 2019. 1, 2, 3, 6, 7

[4] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997. 3

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. 2

[6] Akshay Raj Dhamija, Manuel Günther, and Terrance Boult. Reducing network agnostophobia. In *Advances in Neural Information Processing Systems*, pages 9157–9168, 2018. 3

[7] Deng et.al. Scalable multi-label annotation. In *SIGCHI*, pages 3099–3102, 2014. 2

[8] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 5

[9] Rohit Girdhar, Du Tran, Lorenzo Torresani, and Deva Ramanan. Distinit: Learning video representations without a single labeled video, 2019. 3

[10] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 3

[13] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. 2, 6, 7

[14] Marvin Klingner, Andreas Bär, Philipp Donn, and Tim Fingscheidt. Class-incremental learning for semantic segmentation re-using neither old data nor old labels. *CoRR*, abs/2005.06050, 2020. 3

[15] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? *CoRR*, abs/1805.08974, 2018. 3

[16] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 2, 2013. 3

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10991–11000, 2020. 2

[18] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 3

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3, 6

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[21] Qing Liu, Lingxi Xie, Huiyu Wang, and Alan Yuille. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval, 2019. 3

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. *Lecture Notes in Computer Science*, page 21–37, 2016. 3

[23] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3

[24] Martin Loncaric. Handling "background" classes in machine learning. https://thehive.ai/insights/handling-background-classes-in-machine-learning, Jul 2018. Accessed 2020-05-31. 3

[25] Ofer Matan, RK Kiang, CE Stenard, B Boser, JS Denker, D Henderson, RE Howard, W Hubbard, LD Jackel, and Yann Lecun. Handwritten character recognition using neural network architectures. In *Proceedings of the 4th US Postal Service Advanced Technology Conference, Washington DC, November 1990*, 1990. 3, 4, 5

[26] Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3

[27] Margarita Osadchy, Daniel Keren, and Bella Fadida-Specktor. Hybrid classifiers for object classification with a rich background. In *European Conference on Computer Vision*, pages 284–297. Springer, 2012. 3

[28] Wanli Ouyang, Xiaogang Wang, Cong Zhang, and Xiaokang Yang. Factors in finetuning deep model for object detection with long-tail distribution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 864–873, 2016. 2

[29] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009. 3

[30] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access, 2017. 2

[31] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *CoRR*, abs/1403.6382, 2014. 3

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1, 6

[33] Ruslan Salakhutdinov, Antonio Torralba, and Josh Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR 2011*, pages 1481–1488. IEEE, 2011. 1

[34] L. Shao, F. Zhu, and X. Li. Transfer learning for visual categorization: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 26(5):1019–1034, 2015. 3

[35] K. Shmelkov, C. Schmid, and K. Alahari. Incremental learning of object detectors without catastrophic forgetting. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[36] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016. 3

[37] Daniel L Silver and Robert E Mercer. The task rehearsal method of life-long learning: Overcoming impoverished data. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 90–101. Springer, 2002. 3

[38] Jong-Chyi Su and Subhransu Maji. Adapting models to signal degradation using distillation. In *Proc. British Machine Vision Conference*, 2017. 3

[39] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018. 1, 2

[40] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017. 2

[41] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *arXiv preprint arXiv:2004.05937*, 2020. 3

[42] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Jun Hao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. Classification calibration for long-tail instance segmentation. *arXiv preprint arXiv:1910.13081*, 2019. 2

[43] Yuxiong Wang and Martial Hebert. Learning to learn: Model regression networks for easy small sample learning. In *Proceedings of European Conference on Computer Vision (ECCV)*, October 2016. 1

[44] Yuxiong Wang, Deva Kannan Ramanan, and Martial Hebert. Learning to model the tail. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*, page 7032 – 7042, December 2017. 1

[45] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2, 7

[46] Wei Zhu, Haofu Liao, Wenbin Li, Weijian Li, and Jiebo Luo. Alleviating the incompatibility between cross entropy loss and episode training for few-shot skin disease classification. *arXiv preprint arXiv:2004.09694*, 2020. 2

[47] Xiangxin Zhu, Dragomir Anguelov, and Deva Ramanan. Capturing long-tail distributions of object subcategories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 915–922, 2014. 1, 2