

# Impact of base dataset design on few-shot image classification

Othman Sbai<sup>1,2</sup>, Camille Couprie<sup>1</sup>, and Mathieu Aubry<sup>2</sup>

<sup>1</sup>Facebook AI Research, <sup>2</sup>LIGM (UMR 8049) - École des Ponts, UPE

**Abstract.** The quality and generality of deep image features is crucially determined by the data they have been trained on, but little is known about this often overlooked effect. In this paper, we systematically study the effect of variations in the training data by evaluating deep features trained on different image sets in a few-shot classification setting. The experimental protocol we define allows to explore key practical questions. What is the influence of the similarity between base and test classes? Given a fixed annotation budget, what is the optimal trade-off between the number of images per class and the number of classes? Given a fixed dataset, can features be improved by splitting or combining different classes? Should simple or diverse classes be annotated? In a wide range of experiments, we provide clear answers to these questions on the mini-ImageNet, ImageNet and CUB-200 benchmarks. We also show how the base dataset design can improve performance in few-shot classification more drastically than replacing a simple baseline by an advanced state of the art algorithm.

**Keywords:** Dataset labeling, few-shot classification, meta-learning, weakly-supervised learning

## 1 Introduction

Deep features can be trained on a base dataset and provide good descriptors on new images [39,31]. The importance of large scale image annotation for the base training is now fully recognized and many efforts are dedicated to creating very large scale datasets. However, little is known on the desirable properties of such dataset, even for standard image classification tasks. To evaluate the impact of the dataset on the quality of learned features, we propose an experimental protocol based on few-shot classification. In this setting, a first model is typically trained to extract features on a base training dataset, and in a second classification stage, features are used to label images of novel classes given only few exemplars. Beyond the interest of few-shot classification itself, our protocol is well suited to vary specific parameters in the base training set and answer specific questions about its design, such as the ones presented in Fig. 1.

We believe this work is the first to study, with a consistent approach, the importance of the similarity of training and test data, the suitable trade-off between the number of classes and the number of images per class, the possibility

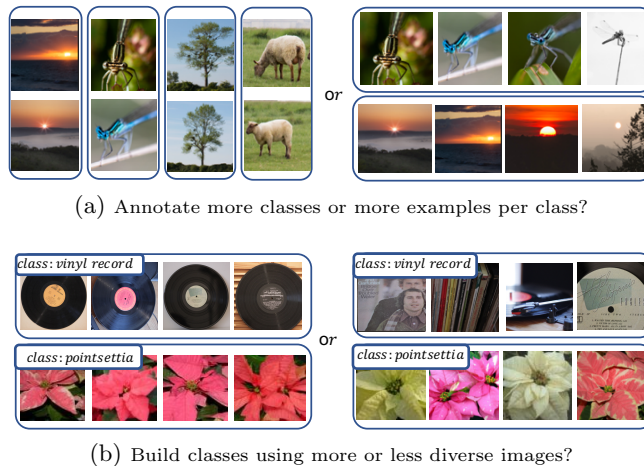


Fig. 1: How should we design the base training dataset and how will it influence the features? a) Many classes with few examples / few classes with many examples; b) Simple or diverse base training images.

of defining better labels for a given set of images, and the optimal diversity and complexity of the images and classes to annotate. Past studies have mostly focused on feature transfer between datasets and tasks [23,48]. The study most related to ours is likely [23], which asks the question “What makes ImageNet good for transfer learning?”. The authors present a variety of experiments on transferring features trained on ImageNet to SUN [47] and Pascal VOC classification and detection [11], as well as a one-shot experiment on ImageNet. However, using AlexNet fc7 features [26], and often relying on the WordNet hierarchy [13], the authors find that variations of the base training dataset do not significantly affect transfer performance, in particular for the balance between image-per-class and classes. This is in strong contrast with our results, which outline the importance of this trade-off in our setup. We believe this might partially be due to the importance of the effect of transfer between datasets, which overshadows the differences in the learned features. Our few-shot learning setting precisely allows to focus on the influence of the training data without considering the complex issues of domain or task transfer.

Our work also aims at outlining data collection strategies and research directions that might lead to new performance boosts. Indeed, several works [6,41] have recently stressed the limitations of performance improvements brought when training on larger datasets, obtained for example by aggregating datasets [41]. On the contrary, [15] shows performance can be improved using a “Selective Joint Fine-Tuning” strategy for transfer learning, selecting only images in the source dataset with low level feature similar to the target dataset and training jointly on both. Our results give insights on why it might happen, showing in particular that a limited number of images per class is often sufficient to

obtain good features. Code is available at [imagine.enpc.fr/~sbaio/fewshot\\_dataset\\_design](https://imagine.enpc.fr/~sbaio/fewshot_dataset_design).

*Contribution.* Our main contribution is an experimental protocol to systematically study the influence of the characteristics of the base training dataset on the resulting deep features for few-shot classification. It leads us to the following key conclusions:

- The similarity of the base training classes and the test classes has a crucial effect and standard datasets for few-shot learning consider only a very specific scenario.
- For a fixed annotation budget, the trade-off between the number of classes and the number of images per class has a major effect on the final performance. The best trade-off usually corresponds to much fewer images per class ( $\sim 60$ ) than collected in most datasets.
- If a dataset with a sub-optimal class number is already available, we demonstrate that a performance boost can be achieved by grouping or splitting classes. While oracle features work best, we show that class grouping can be achieved using self-supervised features.
- Class diversity and difficulty also have an independent influence, easier classes with lower than average diversity leading to better few-shot performances.

While we focus most of our analysis on a single few-shot classification approach and architecture backbone, key experiments for other methods and architectures demonstrate the generality of our results.

## 2 Related work and classical few-shot benchmarks

### 2.1 Data selection and sampling

Training image selection is often tackled through the lens of **active learning** [7]. The goal of active learning is to select a subset of samples to label when training a model, while obtaining similar performance as in the case where the full dataset is annotated. A complete review of classical active learning approaches is beyond the scope of this work and can be found in [38]. A common strategy is to remove redundancy from datasets by designing acquisition functions (entropy, mutual information, and error count) [14,6] to better sample training data. Specifically, [6] introduces an “Adaptive Dataset Subsampling” approach designed to remove redundant samples in datasets. It predicts the uncertainty of ensemble of models to encourage the selection of samples with high “disagreement”. Another approach is to select samples close to the boundary decision of the model, which in the case of deep networks can be done using adversarial examples [10]. In [37], the authors adapt active learning strategies to batch training of neural networks and evaluate their method in a transfer learning setting. While these approaches select specific training samples based on their diversity or difficulty, they typically focus on performance on a fixed dataset and classes, and do not analyze performance of learned features on new classes as in our few-shot setting.

Related to active learning is the question of online **sampling strategies** to improve the training with fixed, large datasets [12,30,3,24]. For instance, the study of [3] on class imbalance highlights over-sampling or under-sampling strategies that are privileged in many works. [12] and [24] propose respectively reinforcement learning and importance sampling strategies to select the samples which lead to faster convergence for SGD.

The spirit of our work is more similar to studies that try to understand key properties of good training samples to **remove unnecessary samples** from large datasets. Focusing on the deep training process and inspired by active SVM learning approaches, [43] explored using the gradient magnitude as a measure of the importance of training images. However using this measure to select training examples leads to poor performances on CIFAR and ImageNet. [2] identifies redundancies in datasets such as ImageNet and CIFAR using agglomerative clustering [8]. Similar to us, they use features from a network pre-trained on the full dataset to compute an oracle similarity measure between the samples. However, their focus is to demonstrate that it is possible to slightly reduce the size of datasets (10%) without harming test performance, and they do not explore further the desirable properties of a training dataset.

## 2.2 Few-shot classification

The goal of few-shot image classification is to be able to classify images from novel classes using only a few labeled examples, relying on a large base dataset of annotated images from other classes. Among the many deep learning approaches, the pioneer Matching networks [42] and Prototypical networks [40] tackle the problem from a metric learning perspective. Both methods are meta-learning approaches, i.e. they train a model to learn from sampled classification episodes similar to those of evaluation. MatchingNet considers the cosine similarity to compute an attention over the support set, while ProtoNet employs an  $\ell_2$  between the query and the class mean of support features.

Recently, [5] revisited few-shot classification and showed that the simple, meta-learning free, Cosine Classifier baseline introduced in [17] performs better or on par with more sophisticated approaches. Notably, its results on the CUB and Mini-ImageNet benchmarks were close to the state-of-the-art [1,27]. Many more approaches have been proposed even more recently in this very active research area (e.g. [35,28]), including approaches relying on other self-supervised tasks (e.g. [16]) and semi-supervised approaches (e.g. [25,29,22]), but a complete review is outside the scope of this work, and exploration of novel methods orthogonal to our goal.

The choice of the base dataset remains indeed largely unexplored in previous studies, whereas we show that it has a huge impact on the performance, and different choices of base datasets might lead to different optimal approaches. The Meta-dataset [41] study is related to our work from the perspective of analyzing dataset impact on few-shot performance. However, it investigates the effect of meta-training hyper-parameters, while our study focuses on how the

base dataset design can improve few-shot classification performance. More recently, [49] investigates the same question of selecting base classes for few-shot learning, leading to a performance better than that of random choice, while highlighting the importance of base dataset selection in few-shot learning.

Since a Cosine Classifier (CC) with a Wide ResNet backbone is widely recognized as a strong baseline [17,16,5,45], we use it as reference, but also report results with two other classical algorithms, namely MatchingNet and ProtoNet.

The classical benchmarks for few-shot evaluation on which we build and evaluate are listed below. Note this is not an exhaustive review, but a selection of diverse datasets which are suited to our goals.

**Mini-ImageNet benchmark.** Mini-ImageNet is a common benchmark for few-shot learning of small resolution images [42,33]. It includes 600K images from 100 random classes sampled from the ImageNet-1K [9] dataset and downsampled to  $84 \times 84$  resolution. It has a standard split of base training, validation and test classes of 64, 16, and 20 classes respectively.

**ImageNet benchmark.** For high-resolution images, we consider the few-shot learning benchmark proposed by [19,46]. This benchmark splits the ImageNet-1K dataset into 389 base training, 300 validation and 311 novel classes. The base training set contains 497350 images.

**CUB benchmark.** For fine-grained classification, we experiment with the CUB-200-2011 dataset [44]. It contains 11,788 images from 200 classes, each class containing between 40 to 60 images. Following [21,5] we resize the images to  $84 \times 84$  pixels and use the standard splits in 100 base, 50 validation and 50 novel classes and use exactly the same evaluation protocol as for mini-ImageNet.

### 3 Base dataset design and evaluation for few-shot classification

In this section, we present the different components of our analysis. First, we explain in detail the main few-shot learning approach that we use to evaluate the influence of training data. Second, we present the large base dataset we use to sample training sets. Third, we discuss the different descriptors of images and classes that we consider, the different splitting and grouping strategies we use for dataset relabeling and the class selection methods we analyze. Finally we give details on architecture and training.

#### 3.1 Dataset evaluation using few-shot classification

Few-shot image classification aims at classifying test examples in novel categories using only a few annotated examples per category and typically relying on a larger base training set with annotated data for training categories. We use the simple but efficient nearest neighbor based approach, visualized in Fig. 2.

More precisely, we start by training a feature extractor  $f$  with a cosine classifier on base categories (Fig. 2 top). Then, we define a linear classifier for the

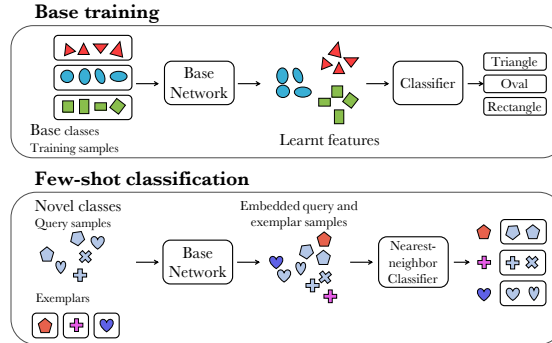


Fig. 2: Illustration of our few-shot learning framework. We train a feature extractor together with a classifier on base training classes. Then, we evaluate the few-shot classification performance of this learned feature extractor to classify novel unseen classes with few annotated examples using a nearest neighbor classifier.

novel classes as follows: if  $z_i$  for  $i = 1 \dots N$  are the labelled examples for a given novel class, we define the classifier weights  $w$  for this class as:

$$w = \frac{1}{N} \sum_{i=1}^N \frac{f(z_i)}{\|f(z_i)\|}. \quad (1)$$

In other words, we associate each test image to the novel class for which its average cosine similarity with the examples from this novel class is the highest. Previous work on few-shot learning focuses on algorithm design for improving the classifier defined on new labels. Instead, we explore the orthogonal dimension of base training dataset and compare the same baseline classifier using features trained on different base datasets.

### 3.2 A large base dataset, ImageNet-6K

To investigate a wide variety of base training datasets, we design the ImageNet-6K dataset from which we sample images and classes for our experiments. We require both a large number of classes and a large number of images per class, to allow very diverse image selections, class splittings or groupings. We define ImageNet-6K as the subset from the ImageNet-22K dataset [34,9] containing the largest 6K classes, excluding ImageNet-1K classes. Image duplicates are removed automatically as done in [36]. Each class has more than 900 images. For experiments on mini-ImageNet and CUB, we downsample the images to  $84 \times 84$ , and dub the resulting dataset MiniIN6K. For CUB experiments, to avoid training on classes corresponding to the CUB test set, we additionally look for the most similar images to each of the 2953 images of CUB test set using our oracle features (see Section 3.3), and completely remove the 296 classes they belong to. We denote this base dataset MiniIN6K\*.

### 3.3 Class definition and sampling strategies

**Image and class representation.** In most experiments, we represent images by what we call *oracle features*, i.e. features trained on our IN6k or miniIN6K datasets. These features can be expected to provide a good notion of distance between images, but can of course not be used in a practical scenario where no large annotated dataset is available. Each class is represented by its average feature as defined in Equation 1. This class representation can be used for examples to select training classes close or far from the test classes, or to group similar classes.

We also report results with several alternative representations and metrics. In particular, we experiment with *self-supervised features*, which could be computed on a new type of images from a non-annotated dataset. We tried using features from RotNet [18], DeepCluster [4], and MoCo [20] approaches, and obtained stronger results with MoCo features which we report in the paper. MoCo exploits the self-supervised feature clustering idea and builds a feature dictionary using a contrastive loss. As an additional baseline we report results using deep features with randomly initialized weights and updated batch normalization layers during 1 epoch of miniIN6k. Finally, similar to several prior works, we experiment using the WordNet [13] hierarchy to compute similarity between classes based on the shortest path that relates their synsets and on their respective depths.

**Defining new classes.** A natural question is whether for a fixed set of images, different labels could be used to train a better feature extractor.

Given a set of images, we propose to use existing class labels to define new classes by splitting or merging them. Using K-means to cluster images or classes would lead to unbalanced classes, we thus used different strategies for splitting and grouping, which we compare to K-means in the Sup. Mat:

- *Class splitting.* We iteratively split in half every class along the principal component computed over the features of the class images. We refer to this strategy as BPC (Bisection along Principal Component).
- *Class grouping.* To merge classes, we use a simple greedy algorithm which defines meta-classes by merging the two closest classes using their mean features, and repeat the same process for unprocessed classes recursively.

We display examples of resulting grouped and split classes in the Sup. Mat.

**Measuring class diversity and difficulty.** One of the questions we ask is whether class diversity impacts the trained features’ few-shot performance. We therefore analyze results by sampling classes more or less frequently according to their diversity and difficulty:

- *Class diversity.* We use the variance of the normalized features as a measure of class diversity. Classes with low feature variance consist of very similar looking objects or simple visual concepts while the ones with high feature variance represent abstract concepts or include very diverse images.
- *Class difficulty.* To measure the difficulty of a class, we use the validation accuracy of our oracle classifier.



### 3.4 Architecture and training details

We use different architectures and training methods in our experiments. Similar to previous works [45, 5], we employ WRN28-10, ResNet10, ResNet18 and Conv4 architectures. The ResNet architectures are adapted to handle  $84 \times 84$  images by replacing the first convolution with a kernel size of 3 and stride of 1 and removing the first max pooling layer. In addition to the cosine classifier described in Section 3.1, we experiment with the classical Prototypical Networks [40] and Matching Networks [42].

Since we compare different training datasets, we adapt the training schedule depending on the size of the training dataset and the method. For example on MiniIN-6k, we train Prototypical Networks and Matching Networks for 150k episodes, while when training on smaller size datasets we use 40k episodes as in [5]. We use fewer query images per class when training on classes with not enough images per class for Prototypical and Matching Networks.

When training a Cosine Classifier, we train using an SGD optimizer with momentum of 0.9 and weight decay of  $5.10^{-4}$  for 90 epochs starting with an initial learning rate of 0.05 and dividing it by 10 every 30 epochs. We also use a learning rate warmup for the first 6K iterations, that we found beneficial for stabilizing the training and limiting the variance of the results. For large datasets with more than  $10^6$  images, we use a batch size of 256 and 8 GPUs to speed up the training convergence, while for smaller datasets (most of our experiments are done using datasets of 38400 images, as in MiniIN training set), we use a batch size of 64 images and train on a single GPU. During training, we use a balanced class sampler that ensures sampled images come from a uniform distribution over the classes regardless of their number of images.

On the ImageNet benchmark, we use a ResNet-34 network and trained for 150K dividing the learning rate by 10 after 120K, 135K and 145K iterations using a batch size of 256 on 1 GPU.

Following common practices, during evaluation, we compute the average top-1 accuracy on 15 query examples over 10k episodes sampled from the test set on 5-way tasks for miniIN and CUB benchmarks, while we compute the top-5 accuracy on 6 query examples over 250-way tasks on the ImageNet benchmark.

## 4 Analysis

### 4.1 Importance of base data and its similarity to test data

We start by validating the importance of the base training dataset for the few-shot classification, both in terms of size and of the selection of classes. In Table 1, we report five shot results on the CUB and MiniIN datasets, the one shot results are available in the Sup. Mat. We write  $N$  the total number of images in the dataset and  $C$  the number of classes. Similar results on ImageNet benchmark can be read in the Sup. Mat. On the miniIN benchmark, we observe that our implementation of the strong CC baseline using a WRN backbone yields slightly better performance using miniIN base classes than the ones reported



	Base data Algo.	MiniIN test			CUB test		
		MiniIN $N=38400$ $C=64$	MiniIN6K Random $N=38400$ $C=64$	MiniIN6K $N\approx 7,1.10^6$ $C=6000$	CUB $N=5885$ $C=100$	MiniIN6K* Random $N=38400$ $C=64$	MiniIN6K* $N\approx 6,8.10^6$ $C=5704$
WRN	PN [40]	73.64 $\pm$ 0.84	70.26 $\pm$ 1.30	85.14 $\pm$ 0.28	87.84 $\pm$ 0.42	52.51 $\pm$ 1.57	68.62 $\pm$ 0.5
	MN [42]	69.19 $\pm$ 0.36	65.45 $\pm$ 1.87	82.12 $\pm$ 0.27	85.08 $\pm$ 0.62	46.32 $\pm$ 0.72	59.90 $\pm$ 0.45
	CC	<b>78.95<math>\pm</math>0.24</b>	<b>75.48<math>\pm</math>1.53</b>	<b>96.91<math>\pm</math>0.14</b>	<b>90.32<math>\pm</math>0.14</b>	<b>58.03<math>\pm</math>1.43</b>	<b>90.89<math>\pm</math>0.10</b>
Conv4	CC	65.99 $\pm$ 0.04	64.05 $\pm$ 0.75	74.56 $\pm$ 0.12	80.71 $\pm$ 0.15	56.44 $\pm$ 0.63	66.81 $\pm$ 0.30
ResNet10	CC	76.99 $\pm$ 0.07	74.17 $\pm$ 1.42	91.84 $\pm$ 0.06	89.07 $\pm$ 0.15	57.01 $\pm$ 1.44	82.20 $\pm$ 0.44
ResNet18	CC	78.29 $\pm$ 0.05	75.14 $\pm$ 1.58	93.36 $\pm$ 0.19	89.99 $\pm$ 0.07	56.64 $\pm$ 1.28	88.32 $\pm$ 0.23

Table 1: 5-shot, 5-way accuracy on MiniIN and CUB test sets using different base training data, algorithms and backbones. PN: Prototype Networks [40]. MN: Matching Networks [42]. CC: Cosine Classifier. WRN: Wide ResNet28-10. MiniIN6K (resp. MiniIN6K\*) Random: 600 images from 64 classes sampled randomly from MiniIN6K (resp. MiniIN6K\*). We evaluate the variances over 3 different runs.

in [16,27](76.59). We validate the consistency of our observations by varying algorithms and architectures using the codebase of [5].

Our first finding is that using the whole miniIN-6K dataset for the base training boosts the performance on miniIN by a very large amount, 20% and 18% for 1-shot and 5-shot classification respectively, compared to training on 64 miniIN base classes. Training on IN-6K images also results in a large 10% boost in 5-shot top-5 accuracy on ImageNet benchmark. Another interesting result is that sampling random datasets of 64 classes and 600 images per class leads to a 5-shot performance of 75.48% on MiniIN clearly below the one using the base classes from miniIN 78.95%. A similar observation can be made for different backbones (Conv4, ResNets) and algorithms tested (ProtoNet, MatchingNets), as well as on the ImageNet benchmark. A natural explanation for these differences is that the base training classes from the benchmarks are correlated to the test classes.

To validate this hypothesis, we selected a varying number of base training classes from miniIN-6K closest and farthest to miniIN test classes using either oracle features, MoCo features, or the WordNet hierarchy, and report the results of training using a cosine classifier with WRN architecture in Fig. 3a. A similar experiment on CUB is shown in the Sup. Mat. We use 900 random images for each class. While all features used for class selection yield similarly superior results for closest class selection and worst results for farthest class selection, we observe that using oracle features leads to larger differences than using MoCo features and Wordnet hierarchy. In Fig. 3b, we study the influence of the architecture and training method on the previously observed importance of class similarity to test classes. Similar gaps can be observed in all cases. Note however that for ProtoNet, MatchingNet and smaller backbones with CC, the best performance is not obtained with the largest number of classes.

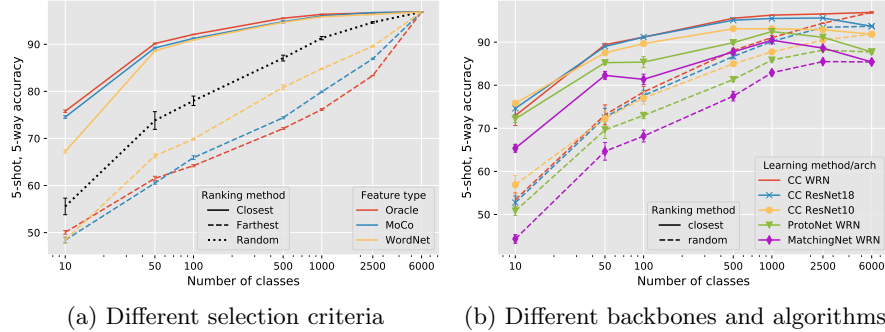


Fig. 3: Five-shot accuracy on miniIN when sampling classes from miniIN-6K randomly or closest/farthest to the miniIN test set using 900 images per class. (a) Comparison between different class selection criteria for selecting classes closest or farthest from the test classes. (b) Comparison of results with different algorithms and backbones using oracle features to select closest classes.

While these findings themselves are not surprising, the amplitude of performance variations demonstrates the importance of studying the influence of training data and strategies for data selection, especially considering that most advanced few-shot learning strategies only increase performance by a few points compared to strong nearest neighbor based baselines such as CC [5,32].

#### 4.2 Effect of the number of classes for a fixed number of annotations

An important practical question when building a base training dataset is the number of classes and images to annotate, since the constraint is often the cost of the annotation process. We thus consider a fixed number of annotated images and explore the effect of the trade-off between the number of images per class and the number of classes. In Fig. 4, we visualize the 5-shot performance resulting from this trade-off in the base training classes on the miniIN and CUB benchmarks. In all cases, we select the classes and images randomly from our miniIN6K and miniIN6k\* dataset respectively, and plot the variance over 3 runs.

First, in Fig. 4 (a,b) we compare the trade-off for different numbers of annotated images. We sample randomly datasets of 38400 or 3840 images with different number of classes and the same number of image in each class. We also indicate the performance with the standard benchmarks base dataset and the full miniIN6K data. The same graph on ImageNet benchmark can be seen in the Sup. Mat using 50k and 500k images datasets.

As expected, the performance decreases when too few classes or too few images per classes are available. Interestingly, on the miniIN test benchmark (Fig. 4a) the best performance is obtained around 384 classes and 100 images per class with a clear boost (around 5%) over the performance using 600 images for 64 classes which is the trade-off chosen in the miniIN benchmark, we observe

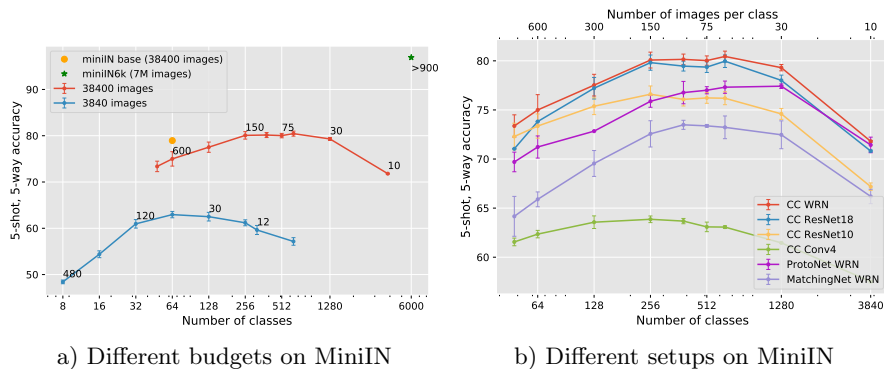


Fig. 4: Trade-off between the number of classes and images per class for a fixed image budget. In (a) we show the trade-off for different dataset sizes and points are annotated with the corresponding number of images per class. In (b) we consider a total budget of 38400 annotated images and show the trade-off for different architectures and methods. The top scale shows the number of images per class and the bottom scale the number of classes.

that the best trade-off is very different on the CUB benchmark, corresponding to more classes and very few images per class. We believe this is due to the fine-grained nature of the dataset.

Second, in Fig. 4 (b), we study the consistency of these findings for different architectures and few-shot algorithms with a 38400 annotated images budget. While the trade-off depends on the architecture and method, there is always a strong effect, and the optimum tends to correspond to much fewer images per class than in standard benchmarks. For example, the best performance with ProtoNet and MatchingNet on the miniIN benchmark is obtained with as few as 30 images per class. This is interesting since it shows that the ranking of different few-shot approaches may depend on the trade-off between number of base images and classes selected in the benchmark.

The importance of this balance, and the fact that it does not necessarily correspond to the one used in the standard datasets is also important if one wants to pre-train features with limited resources. Indeed, better features can be obtained by using more classes and less images per class compared to using all available images for the classes with the largest number of images as is often done, with the idea to avoid over-fitting. Again, the boost observed for few-shot classification performance is very important compared to the ones provided by many advanced few-shot learning approaches.

### 4.3 Redefining classes

There are two possible explanations for the improvement provided by the increased number of classes for a fixed number of annotated images discussed in

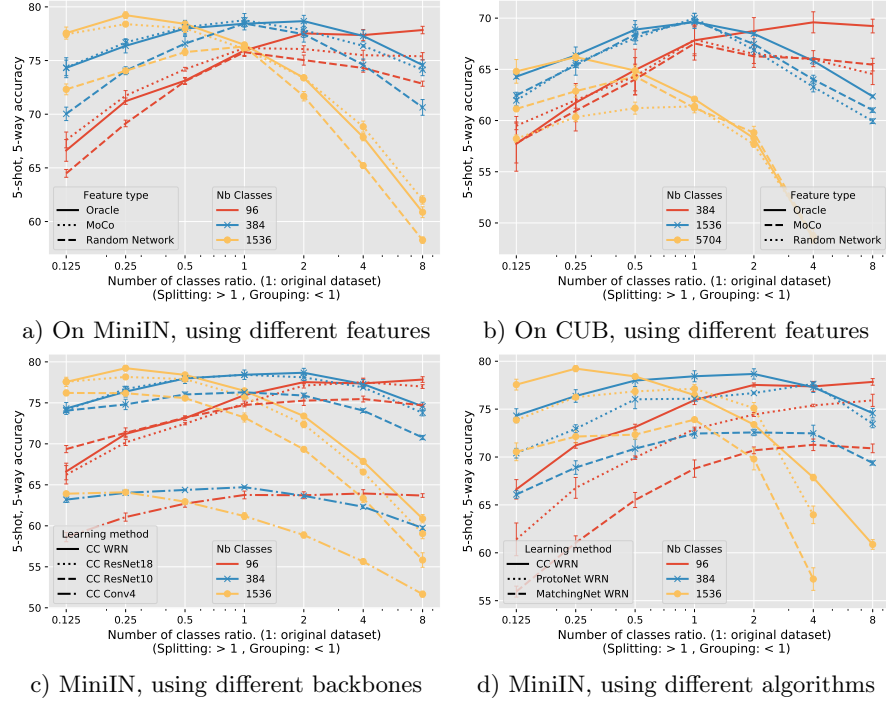


Fig. 5: Impact of class grouping or splitting on few-shot accuracy on miniIN and CUB depending on the initial number of classes. Starting from different number of classes  $C$ , we group similar classes together into meta-classes or split them into sub-classes to obtain  $\alpha \times C$  ones.  $\alpha \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8\}$  is the x-axis. Experiments in a) and b) use CC WRN setup.

the previous paragraph. The first one is that the images sampled from more random classes cover better the space of natural images, and thus provide images more likely similar to the test images. The second one is that learning a classifier with more classes is itself beneficial to the quality of the features. To investigate whether for fixed data, increasing the number of classes can boost performances, we relabel images inside each class as described in Section 3.3.

In Figure 5, we compare the effect of grouping and splitting classes on three dataset configurations sampled from miniIN-6K and miniIN6K\*, with a total number of images of 38400 with different number of classes  $C \in \{96, 384, 1536\}$  for miniIN and  $C \in \{384, 1536, 5704\}$  for CUB. Given images originally labeled with  $C$  classes, we relabel images of each class to obtain  $\alpha \times C$  sub-classes. The x-axes represent the class ratio  $\alpha \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8\}$ . For class ratios lower than 1, we group classes using our greedy iterative grouping, while for ratios  $\alpha$  greater than 1, we split classes using our BCP method. In Fig 5 (a,b), we show three possible behaviors on miniIN and CUB when using our oracle features: (i) if the number of initial classes is higher than the optimal trade-off, grouping

is beneficial and splitting hurts performances (yellow curves); (ii) if the number of initial classes is the optimal one, both splitting and grouping decrease performances (blue curves); (iii) if the number of initial classes is smaller than the optimal tradeoff, splitting is beneficial and grouping hurts performance (red curves). This is a strong result, since it shows there is potential to improve performances with a fixed training dataset by redefining new classes. This can be done for grouping using the self-supervised MoCo features. However, we found it is not sufficient to split classes in a way that improves performances. Using random features on the contrary does not lead to any significant improvements. Fig. 5c confirms the consistency of results with various architecture on miniIN benchmark. Fig. 5d compares these results to the ones obtained with ProtoNet and MatchingNet. Interestingly, we see that since the trade-off for these methods is with much fewer images per class, class splitting increases performances in all the scenarios we considered.

These results outline the need to adapt not only the base training images but also the base training granularity to the target few-shot task and algorithm. They also clearly demonstrate that the performance improvements we observe compared to standard trade-offs by using more classes and less images per class is not only due to the fact that the training data is more diverse, but also to the fact that training a classifier with more classes leads to improved features for few-shot classification.

#### 4.4 Selecting classes based on their diversity or difficulty

After observing in Sec. 4.1 the importance of the similarity between base training classes and the test classes, we now study whether the diversity of the base classes or their difficulty is also an important factor. To this end, we compute the measures described in Sec. 3.3 for every miniIN-6K class and rank them by increasing order. Then, we split the ranked classes into 10 bins of similar diversity or validation accuracy. The classes in the obtained bins are correlated to the test classes and thus introduces a bias in the performance due to this similarity instead of the diversity or difficulty we want to study (see the Sup. Mat, showing the similarity of classes in each bin to the test classes). To avoid this sampling bias, we associate to each class its distance to test classes, and sample base classes in each bin only in a small range of similarities, so that the average distance to the test classes is constant over all bins. In Fig. 6 we show the performances obtained by sampling using this strategy 64 classes and 600 images per class for a total of 38400 images in each bin. The performances obtained are shown on miniIN and CUB in Fig. 6a, 6b both using random sampling from the bin and using sampling with distance filtering as explained before. It can be seen that the effect of distance filtering is very strong, decreasing significantly the range of performance variation especially on the CUB dataset, however the difference in performance is still significant, around 5% in all experiments. Both for CUB and miniIN, moderate class diversity - avoiding both the most and least diverse classes - seems beneficial, while using the most difficult classes seems to harm performances.

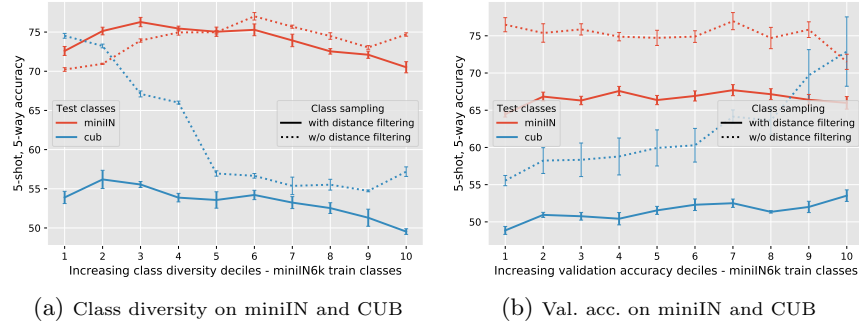


Fig. 6: Impact of **class selection using class diversity and validation accuracy** on few-shot accuracy on miniIN and CUB. For training, we rank the classes of miniIN-6K in increasing feature variance or validation accuracy and split them into 10 bins from which we sample  $C = 64$  classes that we use for base training. Fig. 6a, 6b show the importance of selecting classes in each bin while considering their distance to test classes to disentangle both selection effects.

## 5 Conclusion

Our empirical study outlines the key importance of the base training data in few-shot learning scenarios, with seemingly minor modifications of the base data resulting in large changes in performance, and carefully selected data leading to much better accuracy. We also show that few-shot performance can be improved by automatically relabelling an initial dataset by merging or splitting classes. We hope the analysis and insights that we present will:

1. impact dataset design for practical applications, e.g. given a fixed number of images to label, one should prioritize a large number of different classes and potentially use class grouping strategies using self-supervised features. In addition to base classes similar to test data, one should also prioritize simple classes, with moderate diversity.
2. lead to new evaluations of few-shot learning algorithm, considering explicitly the influence of the base data training in the results: the current miniIN setting of 64 classes and 600 images per class is far from optimal for several approaches. Furthermore, the optimal trade-off between number of classes and number of images per class is different for different few-shot algorithms, suggesting taking into account different base data distributions in future few-shot evaluation benchmarks.
3. inspire advances in few-shot learning, e.g. the design of practical approaches to adapt base training data automatically and efficiently to target few-shot tasks.

*Acknowledgements:* This work was supported in part by ANR project EnHerit ANR-17-CE23-0008, project Rapid Tabasco. We thank Maxime Oquab, Diane Bouchacourt and Alexei Efros for helpful discussions and feedback.

## References

1. Antoniou, A., Storkey, A.J.: Learning to learn via self-critique. *NeurIPS* (2019)
2. Birodkar, V., Mobahi, H., Bengio, S.: Semantic redundancies in image-classification datasets: The 10% you don’t need. *ArXiv preprint 1901.11409* (2019)
3. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. *ArXiv preprint 1710.05381* (2017)
4. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: *ECCV* (2018)
5. Chen, W., Liu, Y., Kira, Z., Wang, Y.F., Huang, J.: A closer look at few-shot classification. *ICLR* (2019)
6. Chitta, K., Alvarez, J.M., Haussmann, E., Farabet, C.: Less is more: An exploration of data redundancy with active dataset subsampling. *ArXiv preprint 1905.12737* (2019)
7. Cohn, D., Ladner, R., Waibel, A.: Improving generalization with active learning. In: *Machine Learning*. pp. 201–221 (1994)
8. Defays, D.: An efficient algorithm for a complete link method. *The Computer Journal* **20**(4), 364–366 (1977)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *CVPR* (2009)
10. Ducoffe, M., Precioso, F.: Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841* (2018)
11. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. *International journal of computer vision* **88**(2), 303–338 (2010)
12. Fan, Y., Tian, F., Qin, T., Liu, T.Y.: Neural data filter for bootstrapping stochastic gradient descent (2016)
13. Fellbaum, C.: Wordnet: An electronic lexical database and some of its applications (1998)
14. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: *ICML* (2017)
15. Ge, W., Yu, Y.: Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In: *CVPR* (2017)
16. Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., Cord, M.: Boosting few-shot visual learning with self-supervision. *ICCV* (2019)
17. Gidaris, S., Komodakis, N.: Dynamic few-shot visual learning without forgetting. In: *CVPR* (2018)
18. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. *CVPR* (2019)
19. Hariharan, B., Girshick, R.: Low-shot visual recognition by shrinking and hallucinating features. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 3018–3027 (2017)
20. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. *ArXiv preprint 1911.05722* (2019)
21. Hilliard, N., Phillips, L., Howland, S., Yankov, A., Corley, C.D., Hodas, N.O.: Few-shot learning with metric-agnostic conditional embeddings. *arXiv preprint 1802.04376* (2018)
22. Hu, S.X., Moreno, P., Xiao, Y., Shen, X., Obozinski, G., Lawrence, N., Damianou, A.: Empirical bayes transductive meta-learning with synthetic gradients. In: *ICLR* (2019)



23. Huh, M., Agrawal, P., Efros, A.A.: What makes imagenet good for transfer learning? NeurIPS LSCVS 2016 Workshop (2016)
24. Katharopoulos, A., Fleuret, F.: Not all samples are created equal: Deep learning with importance sampling. arXiv preprint arXiv:1803.00942 (2018)
25. Kim, J., Kim, T., Kim, S., Yoo, C.D.: Edge-labeling graph neural network for few-shot learning. In: CVPR (2019)
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NeurIPS (2012)
27. Lee, K., Maji, S., Ravichandran, A., Soatto, S.: Meta-learning with differentiable convex optimization. In: CVPR (2019)
28. Li, H., Eigen, D., Dodge, S., Zeiler, M., Wang, X.: Finding task-relevant features for few-shot learning by category traversal. In: CVPR (2019)
29. Liu, Y., Lee, J., Park, M., Kim, S., Yang, E., Hwang, S.J., Yang, Y.: Learning to propagate labels: Transductive propagation network for few-shot learning. In: ICLR (2019)
30. London, B.: A pac-bayesian analysis of randomized learning with application to stochastic gradient descent. NeurIPS (2017)
31. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR (2014)
32. Qiao, S., Liu, C., Shen, W., Yuille, A.L.: Few-shot image recognition by predicting parameters from activations. In: CVPR (2018)
33. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: ICLR (2017)
34. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) (2015)
35. Rusu, A.A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., Hadsell, R.: Meta-learning with latent embedding optimization. In: ICLR (2019)
36. Sablayrolles, A., Douze, M., Schmid, C., Jégou, H.: Déjà vu: an empirical evaluation of the memorization properties of convnets. ArXiv preprint 1809.06396 (2018)
37. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: ICLR (2017)
38. Settles, B.: Active learning literature survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences (2009)
39. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: CVPR workshops (2014)
40. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NeurIPS (2017)
41. Triantafillou, E., Zhu, T., Dumoulin, V., Lamblin, P., Xu, K., Goroshin, R., Gelada, C., Swersky, K., Manzagol, P., Larochelle, H.: Meta-dataset: A dataset of datasets for learning to learn from few examples. ICLR (2020)
42. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: NeurIPS (2016)
43. Vodrahalli, K., Li, K., Malik, J.: Are all training examples created equal? an empirical study. ArXiv preprint 1811.12569 (2018)
44. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD birds-200-2011 dataset (2011)
45. Wang, Y., Chao, W.L., Weinberger, K.Q., van der Maaten, L.: Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. ArXiv preprint 1911.04623

46. Wang, Y.X., Girshick, R., Hebert, M., Hariharan, B.: Low-shot learning from imaginary data. In: CVPR (2018)
47. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR (2010)
48. Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: CVPR (2018)
49. Zhou, L., Cui, P., Jia, X., Yang, S., Tian, Q.: Learning to select base classes for few-shot classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4624–4633 (2020)