DataS^3: Dataset Subset Selection for Specialization

Anonymous CVPR submission

Paper ID *****

Abstract

In many real-world machine learning (ML) applications (e.g. detecting broken bones in x-ray images, detecting species in camera traps), in practice models need to perform well on specific deployments (e.g. a specific hospital, a specific national park) rather than the domain broadly. However, deployments often have imbalanced, unique data distributions. Discrepancy between the training distribution and the deployment distribution can lead to suboptimal performance, highlighting the need to select deployment-specialized subsets from the available training data. We formalize ***dataset subset selection for specialization (*** **DS3** ***)****: given a training set drawn from a general distribution and a (potentially unlabeled) query set drawn from the desired deployment-specific distribution, the goal is to select a subset of the training data that optimizes deployment performance.*

We introduce DATAS^3*; the first dataset and benchmark designed specifically for the* DS3 *problem.* DATAS^3 *encompasses diverse real-world application domains, each with a set of distinct deployments to specialize in. We conduct a comprehensive study evaluating algorithms from various families—including coresets, data filtering, and data curation—on* DATAS^3*, and find that general-distribution methods consistently fail on deployment-specific tasks. Additionally, we demonstrate the existence of manually curated (deployment-specific) expert subsets that outperform training on all available data by up to* 51.3%*. Our benchmark highlights the critical role of tailored dataset curation in enhancing performance and training efficiency on deployment-specific distributions, which we posit will only become more important as global, public datasets become available across domains and ML models are deployed in the real world.*

1. Background and Motivation

Machine learning models are typically trained on large datasets with the assumption that the training distribution closely matches the distribution of the deployment where the model will be applied. However, in real-world applications, deployment data distributions often diverge from general and/or global training set distributions [SRC24, TDS+20]. Selecting relevant data subsets aligned with specific deployments is crucial for maximizing in-field performance. The problem of *data subset selection for specialization* (DS3) is thus critical: given all available training data for a domain and a (small, usually unlabeled) query set that represents the desired deployment, the goal is to identify a subset of the training data, such that training the ML model on this subset maximises performance on the deployment distribution.

Real world example. Consider a wildlife ecologist who aims to build a classifier to detect the presence of invasive rodents in camera trap images collected at the Channel Islands. Existing data on invasive rodents in this context is limited, as they have been mostly eradicated by previous successful conservation action, thus training a classifier from scratch is likely to be unsuccessful. Oftentimes when faced with such challenges, a common approach has been to use a general pre-trained model (such as ViT or CLIP) and then finetune on all *relevant* camera trap data. But *what does "relevant data" mean?* Would using similar species data from other camera trap locations (perhaps on the mainland) improve performance, or introduce noise? What about including data from non-similar species at that location? While adding data to a training set can sometimes improve performance, it can also decrease individual subgroup performance in a biased way [CZPR23] and introduce spurious correlations that can enable models to learn potentially dangerous "shortcuts," resulting in biased predictions, shown across various deployments [GJM+20, BZOR+18, WLL+21, BWE+22a].

The Gap: General datasets vs. domain-specific needs. To address the gap between general models and specific deployment needs, we highlight the need for research emphasis on DS3: the development of methods that select optimal training data for deployment-specific model specialization. Currently, subset selection methods are evaluated on standard CIFAR10/100 [KH+09] and ImageNet [DDS+09] datasets, where test and validation sets have similar distribution to their training sets. Current benchmarks for data filtering [GIF+24] focus on generalization across many tasks, in contrast to specialization for a particular deployment. While these works are valuable, they do not capture, and thus enable progress on, the DS3 challenge.

CVPR #*****

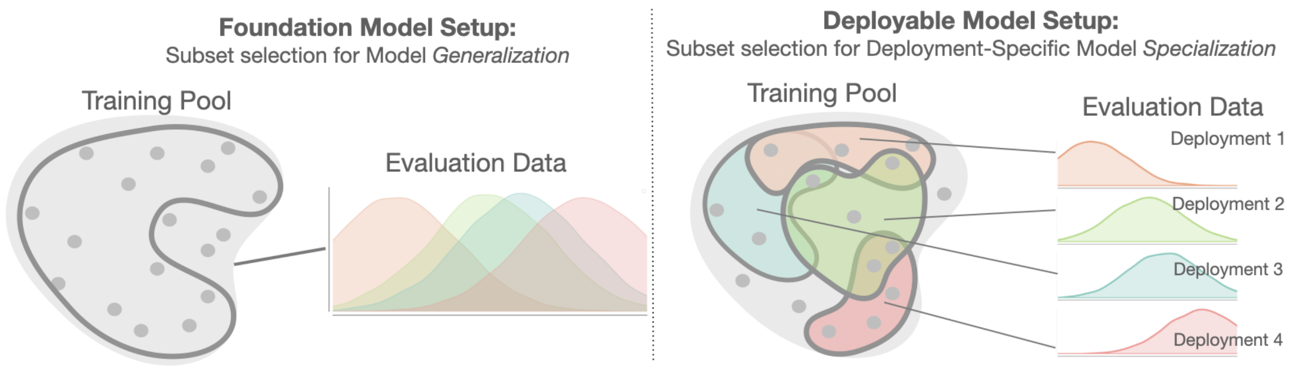

Figure 1. Foundation model training aims for broad generalization, by using all data available, usually from massive internet-scale datasets. In practice, we find these models are often suboptimal for specific deployments, which may exhibit different distributions over categories or data charateristics from the general training data pool. Dataset subset selection for specialization seeks to identify model training subsets closely aligned with the target deployment, achieving superior performance under the given distribution and attribute shifts.

Our contributions. We propose DATAS^3; a comprehensive benchmark to directly evaluate and compare deployment-specialization subset selection methods. Our key contributions are the following:

(i) DATAS^3: A DS3 benchmark of four datasets for evaluating algorithms on the DS3 problem. The datasets represent real-world scientific and engineering applications from different fields, including remote sensing for classifying religious buildings, camera traps to classify species, microscopy images to classify cell organelles, and vehicle footage to learn driving controllers. Each dataset includes multiple realistic deployments, e.g., the camera traps dataset aims to categorize species, and the deployments to specialize to are geographic locations where scientists want to analyze species, e.g., Central Africa or Southeast Asia.

(ii) Manually curated expert subsets of the training data for each deployment showing that selecting a well-curated subset can consistently outperform models trained on the entire dataset.

(iii) An extensive experimental study comparing current SOTA subset selection methods across the provided data pools and their respective deployments. After training a suite of baselines, our results clearly show that current subset selection methods fail on DS3, highlighting the need for our DATAS^3 benchmark.

2. Problem Statement

DS3 problem formulation. Let X be a ground set of data points, $T \subset X$ be a given *training set* drawn from a training (pool) distribution P_T over X, and let $Q \subset X$ be a *query set* drawn from the desired **deployment-specific distribution** P_Q over X. Given a model θ, the objective of **dataset subset selection for specialization (DS3)**, is to design an algorithm `SubsetSelection-ALG`, which takes T (the training set) and Q (the deployment representative query set) as input, and outputs a subset $S^* \subset T$ that minimizes the expected loss of θ trained on S^* over the desired deployment-specific distribution P_Q. More formally:

$$S^* = \underset{S \subset T}{\arg\min}\, \mathbb{E}_{q \sim P_Q} \left[\mathcal{L}(\theta(S), q) \right], \tag{1}$$

where $\theta(S)$ denotes the model trained on the subset $S \subset T$, and $\mathcal{L}(\theta(S), q)$ is the loss function evaluated on a single point q sampled from P_Q and the trained model $\theta(S)$. The term $\mathbb{E}_{q \sim P_Q}$ denotes the expected value over the distribution P_Q. Hence, the algorithm `SubsetSelection-ALG` outputs S^*, the subset of T that minimizes the expected loss of the entire desired deployment distribution P_Q. Notably, `SubsetSelection-ALG` can only access the desired deployment-specific distribution via the query set Q.

Is the query set annotated/labeled? This formalization can be divided into two cases: in the first, the query set Q is annotated with a set of labels, formally, Q is a set of $m > 0$ pairs $Q = \{(q_1, y_1), \cdots, (q_m, y_m)\}$, where for every $i \in [m]$, q_i is the ith feature vector describing the ith input, and y_i is it corresponding label/annotation. In this case the algorithm `SubsetSelection-ALG` has access to the set of labels $\{y_1, \cdots, y_n\}$. In the second scenario, no labels are provided for Q, meaning that the `SubsetSelection-ALG` does not have access to the set $\{y_1, \cdots, y_n\}$ and consequently $Q = \{q_1, \cdots, q_m\}$. Annotating Q for any specific deployment requires time, money, and expertise. Thus, DS3 progress without labels has a high potential for impact.

Is `SubsetSelection-ALG` model agnostic? Similarly, this formalization can be approached in two different ways: one where the computation of S^* depends on a given specific model θ, i.e., `SubsetSelection-ALG` is model dependent, and has access to the model θ we wish to train on. The more general, model-agnostic formulation aims to find S^* that performs well across all possible models, meaning that `SubsetSelection-ALG` has no access to θ.

3. Related Work

Traditional data subset selection approaches can be split into two main categories: 1) Data filtering or cleaning, which focuses on refining the dataset to enhance its quality [ZRG+22, RSR+20], and 2) Coresets for dataset subset selection, aimed at reducing training time by a computing a subset that effectively represents the larger training dataset [KSRI21, TZM+23].

Data filtering for better learning. Data pruning is widely used in NLP to clean noisy datasets [Ano23], often employing filtering and heuristics [BUSZ22]. Common methods include excluding texts with blocklisted words [RSR+20], removing duplicates [ZRG+22], filtering out non-English texts [RSR+20, RBC+22], and discarding short sentences [RSR+20, RBC+22]. Perplexity-based filtering removes high-perplexity samples considered unnatural and detrimental to performance [MRB+23, WLC+20, LSW+23]. Although simple filtering can enhance language models [PMH+23, RSR+20], their effectiveness varies, and some studies report no benefits [BBH+22, BSA+23], possibly due to their simplicity. [ZLX+24] showed that manually selecting a small subset satisfying quality and diversity improves alignment performance. **For vision tasks,** a smaller number of methods have been suggested for data filtering [SGS+23] to obtain better trainable subsets [SRM+22] through the use of model signals [MBR+22].

Coresets for efficient learning. Subset selection (hitherto referred to as coresets) is common for vision tasks. The goal is to compute a small subset from the training dataset, that approximates training on the full dataset, thus boosting the training process [BFL16, MEM+22]. Coresets proved to be useful in many applications such as regression [DDH+08, CDS20, TJF22, MMM+22, MJF19], clustering [HM04, Che09, HV20, JTMF20, CAGLS+22], low-rank approximation [CMM17, BDM+20, MJTF21], support vector machines (SVMs) [Cla10, TBFR21, MEM+22], and for compressing neural networks [BLG+22, LBL+19, TMM22]. For boosting the training of neural networks, [CYM+19] used proxy functions to select subsets of training data approximating the training process. Later [MBL20, MCL20] developed algorithms to estimate the full gradient of the deep neural network on the training data. These techniques were further refined by [KSR+21, KSRI21, PGD21, WPM+20]. Other methods require a neural network forward pass to get embeddings [SS18, SGS+22, KZCI21]. Notably, these methods rely on the properties of the models in training to select data. Later, [TZM+23] provided a method that does not require access to the trained model, but demands assumptions on the model and its complexity.

All these methods assume the training data well represents the test (deployment) data, as the case in known diverse, high-quality vision benchmarks (CIFAR10 and ImageNet). Thus, the aim was to approximate the training data via a subset (coresets) or enhance training (filtering) assuming that that the training and testing sets share the same distribution.

Active learning. There is a rich area of online active learning literature, which continually filters data while training [EDHG+20, WLY22, YLBG20, TND+22], requiring to query an annotator for more labeled data and oftentimes, rely on properties of the models in-training to select data. Here, we are interested in exploring data subselection prior to training and without knowledge of model weights.

Benchmarks. The works most related to ours are [GIF+24], [MBY+23] and [FXC+24]. DataComp [GIF+24] introduces a benchmark where the main challenge is to select the optimal data subset for pretraining *generalization*. It evaluates various data curation strategies using standardized CLIP training code, followed by zero-shot assessments on 38 downstream datasets. [MBY+23] has multiple benchmarks across domain-specific data sources, but is again aimed for generalization rather than specialization. [FXC+24] focuses on image-only models, which are smaller and easier to train to high accuracy.

Our benchmark. In contrast to these benchmarks, DATAS^3 is specifically designed to evaluate subset selection methods for *deployment-specific specialization*, rather than generalization, where the training and testing (deployment) data exhibit distributional shifts.

4. The DATAS^3 Benchmark

Benchmark design. Traditional dataset subset selection methods often aim to build a maximally generalizable model with the least amount of training data. In contrast, *The goal of our benchmark is to identify the optimal subset for a specific deployment.* Our benchmark includes four application-domain datasets, and defines multiple *distinct deployments* for each. Given a small query set Q from a deployment, the objective is to select a subset of the training data that achieves optimal performance on the given deployment of Q. Subset selection methods are evaluated on four held-out test sets, each corresponding to a specific deployment.

Datasets. Our benchmark includes four datasets, each capturing a unique and diverse application of ML: Auto Arborist for street-level tree classification [BWE+22b], iWildCam for camera trap species identification [BACB21], GeoDE for diverse object classification [RLZ+23], and NuScenes for driving footage steering regression (autonomous-driving) [CBL+20]. Each of these datasets inherently represents many of the real-world challenges that make dataset subset selection a deployment-specific problem, including covariate shifts, subpopulation shifts, and long-tailed distributions. For each dataset, we provide an expert-knowledge-guided subset that demonstrates the usefulness of dataset subselection, with improvement over using all the training data. These subsets were created using expert-driven knowledge with access to information that benchmark users are

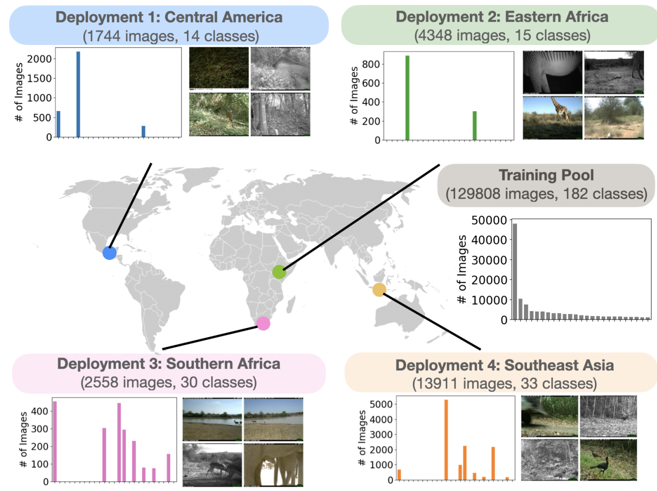

Figure 2. **The iWildCam dataset** is long-tailed, with each plots of the class counts showing significant label distribution shift per deployment. Additionally, there are major axes of variation between deployments in images, ranging from background, species of interest, night/day captures, camera type, and more.

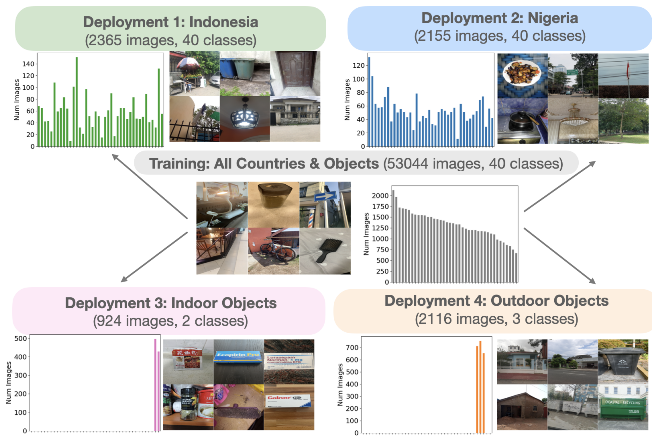

Figure 3. **The GeoDE dataset** has both covariate and distribution shift present. Deployments 1 and 2 have strong covariate shift, with the data being collected from specific countries, whereas deployments 3 and 4 have strong label shift, focusing on specific categories of objects. The training pool contains data from all countries and object types, meaning that subset selection is uniquely helpful for specialization.

not provided (e.g. metadata, GPS location, region, etc). In what follows, we discuss each dataset and its corresponding deployments. Additional details about each dataset can be found in Appendix A.

4.1. iWildCam

Background: Animal populations have declined by 68% on average since 1970 [Sta20]. To monitor this biodiversity loss, ecologists deploy camera traps—motion-activated cameras placed in the wild [WGK17]—and process the data with machine learning models [NMB+19, BMY19]. However, variations in illumination, camera angle, background, vegetation, color, and animal frequencies across different locations cause these models to generalize poorly to new deployments. To specialize models for specific locations, selecting appropriate data subsets for deployment-specific (in this case location) specialization becomes essential.
Problem Setting: To study this problem, we use the iWildCam 2020 dataset. The task is multi-class species classification. Concretely, the input x is a photo taken by a camera trap, the label y is one of 182 different animal species, and the deployment d is an integer that identifies the camera trap that took the photo. Performance is measured by classification overall accuracy for species identification.
Data: The dataset comprises $203,029$ images from 323 different camera traps spread across multiple countries in different parts of the world. The original camera trap data comes from the Wildlife Conservation Society (link). These images tend to be taken in short bursts following the motion-activation of a camera trap, so the images can be additionally grouped into sequences of images from the same burst, though our baseline models do not exploit this information, and our evaluation metric treats each image individually. However, a grouped sequence is in the same split of the data (train, test, query) in order to avoid model memorization. Each image is associated with the following metadata: camera trap ID, sequence ID, and datetime.
Deployments: Our deployments were defined to be split across camera trap locations to simulate the common scenario of researchers setting up new cameras within a region, with poor model generalization on the new cameras [WGK17]. Our train/test split was done randomly across the 200 locations, with the four downstream test tasks created by clustered by the latitude and longitude of camera GPS location in 4 deployments: (1) Central America, (2) Eastern Africa, (3) Southern Africa, and (4) Southeast Asia. Similar to most other camera trap datasets, iWildCam has significant long-tailed label distributions, with variation in species and backgrounds between locations, as can be seen in Figure 2.
Expert Knowledge Subset: Expert subsets were generated by selecting the camera locations within the training pool within the location "cluster" as the deployment (eg. the relevant geographic area) and then sampling the training data to closely match the species distributions in the deployment.

4.2. GeoDE

Motivation. Object classification datasets are often constructed by scraping images from the web but contain geographical biases [SHB+17]. Instead of scraping images from the web, GeoDE [RLZ+23] crowdsources a dataset that is roughly balanced across 40 different objects and six world regions, showing that common objects (stoves, bicycles, etc),

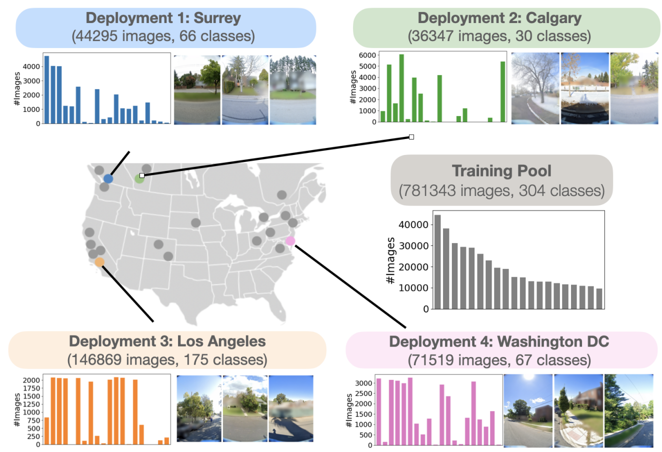

Figure 4. **The Auto Arborist dataset** is long-tailed, with each deployment having significant label distribution shift. Additionally, there are major axes of variation between deployments in images, coming from factors described in Sec. 4.3

vary in appearance across the world. Accordingly, specializing models to different regions becomes useful when the objects have strong covariate shift.

Problem setting. GeoDE is a diverse dataset comprising 40 different objects collected from 6 world regions. The associated task is multiclass classification, where the goal is to predict the object depicted in each image.

Data. GeoDE contains 61,490 images, each labelled with both a region (Africa, Americas, East Asia, Europe, Southeast Asia, West Asia), and an object (examples: bag, backyard, toothpaste, etc.). The dataset was crowdsourced, with participants submitting photographs for each object, which were then assessed for quality.

Deployments: We propose 4 different deployments: (1) objects in Indonesia, (2) objects in Nigeria, (3) indoor objects, and (4) outdoor objects, as shown in Figure 3. Nigeria and Indonesia were selected as the two countries with the poorest performance, and the indoor/outdoor deployment tasks were selected for enabling model specialization. The training dataset includes images from all countries, and the test data contains only images from Nigeria and Indonesia.

Expert Knowledge Subset: Expert subsets were generated by selecting data from the relevant countries/categories in the training data.

4.3. Auto Arborist

Motivation: Ecological imagery for environmental monitoring and Earth observation provides policymakers with critical, data-driven insights to support climate adaptation [BLF+16]. Automated tree classification, for instance, offers substantial benefits for humanitarian aid, disaster relief, forestry, agriculture, and urban planning, supporting applications in city planning, resource management, and environmental monitoring.

Problem Setting: Automated tree classification in street-level imagery is inherently difficult and is associated with fundamental challenges such as:

- **Noisy labels.** Images are commonly mislabeled: genus classification is difficult and requires specialized expertise, GPS localization from the ground can be in error, there are often multiple trees within a single image with only a single label, and temporal inconsistencies can occur as trees are not imaged and labeled at the same time.
- **Non-IID data.** Geospatial data also breaks the typical deep learning assumption that data will be independent and identically distributed (IID) spatially close examples often contain correlations. For example, trees are often planted in groups (e.g. a row of cherry trees along the same street).
- **Fine-grained and long-tailed class distribution.** Tree classification is fine-grained, with only subtle differences between many genera, and the distribution of trees is long-tailed. These characteristics tend to skew classification models towards predicting predominant classes.
- **Geospatial distribution shift.** Finally, this dataset contains significant covariate and subpopulation distribution shift due to variations in weather, differences in urban planning specific to each city, and temporal changes at different locations.

Data: The Auto Arborist dataset is a multi-view, fine-grained visual tree categorization dataset containing images of over 1 million public zone trees from 300 genus-level categories across 23 major cities in the US and Canada (We note that the dataset represents only a portion of the total tree population). Specifically, each tree record in the dataset is associated with a street-level and aerial image. For our benchmark, we focus on the street-level images.

Deployments: Deployments in Auto Arborist correspond to the development models for use by individual cities. The deployment cities of (1) Surrey with 66 distinct tree genus classes, (2) Calgary with 30 classes, (3) Los Angeles with 175 classes, and (4) Washington DC with 67 classes were chosen due to their diverse climates, species distributions, and urban structures, as seen in Figure 4. Historical development patterns also significantly change from city to city with many trees being planted intentionally through city planning not by random chance, causing city-specific signals for tree genera. Moreover, the number of classes and distribution of classes (long-tailedness) also varies significantly. City-specific models allow us to capture the unique features of each city's tree population, optimizing the model to perform best in the environment it will be deployed. As cities increasingly rely on data-driven methods for urban planning and environmental monitoring, having tailored models ensures higher accuracy and utility, especially for cities with limited resources for ground surveys.

Expert Knowledge Subset Expert-driven subsets were generated by selecting data from the nearest cities in the training pool, and then sampling the training data to closely match the tree genus/class distributions in the deployment set.

4.4. NuScenes

Motivation: End-to-end autonomous driving systems streamline vehicle control by directly mapping sensory inputs, such as images, to control outputs like steering angles [WMX+24]. This approach eliminates traditional, multi-step processing pipelines, enabling real-time adaptation to complex environments. By integrating perception and control, these systems enhance efficiency, responsiveness, and adaptability, crucial for safe autonomous navigation. The idea is that adapting these systems to specialize in particular streets or environments is made easier as a single model encompasses the full system. Thus, training this model to specialize in a specific environment brings advantages, capturing detailed local road layouts, typical traffic patterns, area-specific obstacles, and more.

Problem Setting: We explore vision-based control for self-driving across diverse environments (e.g., different city areas) and driving scenarios (e.g., pedestrians crossing, construction zones), formulated as a regression task. The model's goal is to predict a single scalar value representing the car's steering angle. Performance is evaluated in an open-loop manner using metrics like mean squared error.

Data: This dataset includes $88,461$ images from the NuScenes dataset, subsampled from the image sweeps at a rate of 2. The images were captured from a video stream recorded while driving a car. Each image is paired with a steering angle control from the CAN bus, synchronized with the sensor timestamps of both the camera and CAN bus data. To label each image with the correct steering angle, we apply 1D interpolation to create a continuous function of the steering angle and query it based on the camera's timestamp. The steering angle, measured in radians, ranges from -7.7 to 6.3, with 0 indicating straight driving, positive values indicating left turns, and negative values indicating right turns. To ensure alignment between images and steering control data, samples with vehicle velocities below 1 m/s are removed.

Deployments: Deployments are organized by the geographic locations where the data was collected, including (1) Boston Seaport, (2) Singapore Holland Village, (3) Singapore One-North, and (4) Singapore Queenstown. While all tasks are based on expert demonstrations of driving and general driving behaviors, each location presents varying environmental features—such as vegetation, road types, roadside infrastructure, and weather—as well as differences in driving style and road regulations. Train/test splits are randomly sampled within each deployment.

Expert knowledge subset: Expert subsets were generated by selecting data from the relevant areas in the training data.

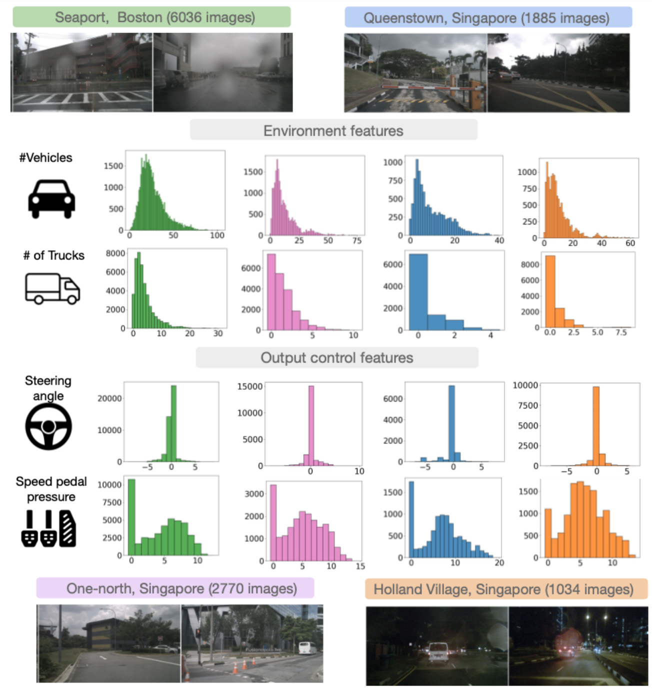

Figure 5. **The NuScenes dataset.** The driving area influences environmental features, which, in turn, impact the control outputs. In Boston's Seaport, trucks are common on the roads, unlike in Queenstown or Holland Village in Singapore, where trucks are less prevalent. Similarly, the number of vehicles requiring the car's attention varies by area. Speed control outputs may be higher in Queenstown and One-North, Singapore, due to the nature of the streets, compared to Boston's Seaport or Holland Village, where speeds tend to be lower. Steering angles are also affected by location-specific road layouts; for instance, Queenstown features more frequent sharp right curves, which are less common in other areas.

4.5. Benchmark Pipeline

Within our benchmark, each dataset has a two-step process for evaluation: (i) Given a small query set representing the deployment data, curate a subset of data from the training for a specific deployment. (ii) Finetune/train a fixed model on the chosen subset from the training pool and evaluate on the deployment (test) set. For each dataset, we fix the training procedure for all subsets of data, fixing model architecture, optimizers, and loss functions. We run a small hyperparameter sweep for each training subset across batch sizes $\{32, 64, 128\}$ and learning rates $\{0.01, 0.001, 0.0001\}$ for each deployment. For all datasets, we use ResNet50 [HZRS15] for full-finetuning and a ViT [DBK+20] for linear probes of the training subsets, chosen for efficiency due to the number of baselines. Full details are in Appendix B.

4.6. Metrics

Participants are evaluated across 12 deployments from 4 datasets, as outlined in Section 4. For the classification task datasets of GeoDE, Auto Arborist, and iWildCam, we report accuracy for each deployment, and for the regression task dataste NuScenes, we report mean squared error. For each deployment, we evaluate participants of the benchmark on overall accuracy and size of the training subset; the less data used the better while balancing optimal performance. We also report precision and recall in Appendix C.

5. Baselines

We compare performance of coreset/data filtering algorithms for dataset subselection across our benchmark, across different scenarios: (a) access to an unlabeled query set, and (b) access to a labeled query set. We also curate a third category, (c), which leverages domain expertise to generate expert-selected subsets, in order to demonstrate the existence of better-than-all subsets for these deployments.

Non-subset comparisons:
No filtering: Performance of a model trained on the entire training pool, without any filtering.
Query Sets: As a comparison, we also include performance of a model trained directly on the labeled query set for each deployment. Note that this would require access to query labels, which are not always available. When labels are available, performance of models trained on the small query sets are often poor, hence the value of learning from larger-scale general-pool data. As a logistical point, none of the baselines we show in our results train on query set data.

Expert-Driven Subsets: We contribute curated, "expert-driven" subsets using domain knowledge and/or metadata. We find these knowledge-guided subsets often outperform using all samples in the training pool (no filtering). The creation of these subsets is described in Sec. 4.

Unlabeled-query baselines:
Image-alignment (Image-Align): We take the cosine similarity between the training and query embedding space, using examples that exceed a threshold for at least x samples, where x is a hyperparameter chosen from {1,10,100}.
Nearest neighbors features (Near-Nbors): To better align our method with the downstream deployment, we explore using examples whose embedding space overlaps with the query set of data. To do so, we cluster image embeddings extracted by an OpenAI ViT model for each image into 1000 clusters using Faiss [JDJ19]. Then, we find the nearest neighbor clusters for every query set example and keep the training cluster closest to each query set cluster. This method was inspired by the similar DataComp baseline [GIF+24].

Labeled-query baselines:
CLIP score filtering (CLIP-score): We also experiment with CLIP score filtering, using examples that exceed a threshold for cosine similarity between CLIP image and text similarity. Text for each image was created with manual captioning (e.g. for iWildCam, *"This is a camera trap image of a lion taken at time 10-2-2016 at 04:26:13 in Nigeria"*). We select the subset that exceeds a threshold of CLIP-score similarity, with the threshold calculated for subsets that make up 25%, 50%, 75%, and 90% of the dataset.
Matching relative frequency (Match-Dist): We explore having access to the relative frequency of each label in the downstream deployment. For example, a domain expert at a national park might know the relative frequency of species (deployment-specific domain knowledge) that we can utilize for dataset subset selection. We create subsets by sampling 25%, 50%, 75%, and 90% of the training pool to match the label distribution of the deployment.
Matching labels (Match-Label): Similarly, a domain expert may know the classes present in the downstream deployment. For example, a domain expert at a national park might know the species present (deployment-specific domain knowledge) that we can utilize for dataset subset selection. For these subsets, we simply remove the classes present in the training pool that are not present in the testing pool.

6. Results and discussion

Well chosen subsets outperform training on all data. The expert-driven subsets in Table 1 show that deployment-specific well-chosen subsets of the data can significantly outperform models trained on all the data, with improvements in deployment accuracy up to 3.6% for GeoDE, 11.9% for iWildCam, 51.3% for Auto Arborist, and a 0.03 reduction in MSE for NuScenes. Even when the expert subsets underperform all training data, as in NuScenes Deployment 2, there exist subsets from other baselines that outperform using all the data. Due to the extreme long-tailed nature and significant label distribution shift between the training pool and deployments of iWildCam and Auto Arborist, well-chosen subsets improve performance significantly. This indicates that using "irrelevent" data from the training pool is actively harmful to performance for specialized deployments, compared to a closer in-distribution subset. As an example, the iWildCam's training pool contains many Thompson's Gazelle, but only Deployment 2 has Thompson's Gazelle present. Accordingly, Deployments 1, 3, and 4 had more improvement between all data and expert subsets than Deployment 2 since the former had a greater label distribution shift from the training pool.

There is a need for unsupervised methods for dataset subselection. While the expert-driven subsets in Table 1 demonstrate that a well-chosen subset *does exist* for all deployments, finding this subset without expert knowledge is still an open problem. While some of our baselines require access to query labels, this requirement can in many cases be unrealistic in the deployable ML setting (labels can be expensive or difficult to collect). The two unsupervised baselines, the nearest neighbors and image alignment methods, do not perform optimally on the deployments, often under-

Dataset	Deploy #	Non subset		Expert subset	Unlabeled query set		Labeled query set		
		Query-set	All-data		Image-Align	Near-Nbors	CLIP-score	Match-Label	Match-Dist
GeoDE (Acc)	Deploy 1	0.872	0.885	**0.921**	0.879	0.88	0.887	0.882	0.886
	Deploy 2	0.450	0.890	**0.910**	0.897	0.892	0.899	0.900	0.882
	Deploy 3	0.950	0.821	**0.85**	0.845	0.760	0.838	0.83	0.879
	Deploy 4	0.827	0.829	**0.845**	0.791	0.783	0.828	0.841	0.830
iWildCam (Acc)	Deploy 1	0.703	0.655	0.650	0.555	0.502	0.503	0.740	**0.743**
	Deploy 2	0.780	0.341	0.346	0.438	0.469	0.463	0.350	**0.490**
	Deploy 3	0.438	0.716	0.745	0.537	0.450	0.420	0.723	**0.750**
	Deploy 4	0.463	0.660	0.670	0.599	0.600	0.290	0.687	**0.741**
Auto Arborist(Acc)	Deploy 1	0.159	0.348	**0.861**	0.382	0.392	0.380	0.665	0.740
	Deploy 2	0.197	0.483	**0.859**	0.114	0.141	0.137	0.650	0.560
	Deploy 3	0.124	0.157	**0.382**	0.159	0.099	0.167	0.159	0.234
	Deploy 4	0.119	0.135	**0.392**	0.102	0.108	0.106	0.102	0.230
NuScenes (MSE)	Deploy 1	0.063	0.050	**0.029**	0.040	0.040	0.073	-	-
	Deploy 2	0.070	0.021	0.049	0.147	0.042	**0.032**	-	-
	Deploy 3	0.089	0.068	**0.038**	0.049	0.125	0.071	-	-
	Deploy 4	0.123	0.048	**0.039**	0.086	0.389	0.050	-	-

Table 1. Best-performing subsets across hyperparameters for baseline methods across all datasets and deployments (abbreviated as Deploy) for ResNet50 full-finetuning. Overall accuracy is reported for the classification tasks of GeoDE, iWildCam, and Auto Arborist (greater is better) and MSE is reported for the regression task of NuScenes (smaller is better). Match-Dist and Match-Label are not applicable for NuScenes, as it is a regression task and does not have clear classes/labels for these methods. Baselines are distinguished from one another by their access to information, with each baseline having access to expert knowledge, or a labeled/unlabeled query set. We do not report the random baseline in this table, but demonstrate results in Appendix C as it mainly refers to subset size. Well-chosen subsets outperform using all the training data in each deployment, indicated in bold.

performing using all the training data. Our benchmark opens up the line of research for potential unsupervised methods for this data subselection process.

Training on more data has diminishing returns. For all deployments, we see that we can achieve near-optimal performance with subsets of the data. Appendix C shows that even 25% of the data can perform near-optimally in some cases, with little performance loss with 50% of the data. Additionally, while not a realistic deployment scenario, the "lower bound" of training on the small query set (results in Table 1) performs close to optimally in several deployments (this is expected, since the query sets are in distribution with the deployments). However, this again indicates that having a small relevant subset of data is most useful. Overall, these results demonstrate that greater efficiency for training specialised ML models is possible, potentially reducing computational and data storage burdens in deployable settings. We hypothesize this is because many deployments have significant distribution shift from the training pool, so as the data added gets farther from the deployment distribution, it becomes less relevant for optimal performance.

7. Conclusions

We present DATAS^3, a benchmark to explore model specialization via dataset subselection for scientific and engineering domains, and provide: **(1)** a test suite for the problem across 4 ML application domains, each represented by a dataset containing a general training data pool and 4 distinct deployment scenarios **(2)** expert- and knowledge-guided subsets for each deployment which outperform training on all data, sometimes by a significant margin, demonstrating the value of specialized training data curation **(3)** an extensive experimental study highlighting that current methods for subset selection, designed for generalization instead of specialization, do not perform well on DATAS^3.

We find that there does not currently exist a winning method that performs well across multiple domains/datasets, posing an open challenge to the research community. While well-performing subsets exist via expert-driven knowledge, models without access to labeled query sets systematically underperform. We also find that certain datasets are more challenging than others– perhaps different subselection methods are necessary for different domains or types of shifts.

Limitations and future work Due to computational constraints, hyperparameter searches were restricted to learning rate and batch size. Additionally, while smaller-scale models were used (ResNet50) for full finetuning with a larger model only used for linear probes, future work could explore larger model finetuning and its effects.

Additionally, model specialization for deployments isn't limited to the domains we include in our benchmark. We plan to expand this benchmark to capture more scientific domains with similar needs, including the tasks of histopathology disease prediction, medical eICU record mortality prediction, satellite imagery for crop type classification, and astrophysics galaxy classification.

References

[Ano23] Anonymous. When less is more: Investigating data pruning for pretraining LLMs at scale. In *NeurIPS Workshop on Attributing Model Behavior at Scale*, 2023. 3

[BACB21] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset, 2021. 3

[BBH+22] Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin, May 2022. Association for Computational Linguistics. 3

[BDM+20] Vladimir Braverman, Petros Drineas, Cameron Musco, Christopher Musco, Jalaj Upadhyay, David P. Woodruff, and Samson Zhou. Near optimal linear algebra in the online and sliding window models. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS*, pages 517–528, 2020. 3

[BFL16] Vladimir Braverman, Dan Feldman, and Harry Lang. New frameworks for offline and streaming coreset constructions. *arXiv preprint arXiv:1612.00889*, 2016. 3

[BLF+16] Leslie A. Brandt, Abigail Derby Lewis, Robert T. Fahey, Lydia Scott, Lindsay E. Darling, and Christopher W. Swanston. A framework for adapting urban forests to climate change. *Environmental Science & Policy*, 66:393–402, 2016. 5

[BLG+22] Cenk Baykal, Lucas Liebenwein, Igor Gilitschenski, Dan Feldman, and Daniela Rus. Sensitivity-informed provable pruning of neural networks. *SIAM J. Math. Data Sci.*, 4(1):26–45, 2022. 3

[BMY19] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review, 2019. 4

[BSA+23] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023. 3

[BUSZ22] Fred Bane, Celia Soler Uguet, Wiktor Stribiżew, and Anna Zaretskaya. A comparison of data filtering methods for neural machine translation. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 313–325, Orlando, USA, September 2022. Association for Machine Translation in the Americas. 3

[BWE+22a] Sara Beery, Guanhang Wu, Trevor Edwards, Filip Pavetic, Bohdan S. Majewski, Shreyasee Mukherjee, Stanley Chan, John Morgan, Vivek Rathod, and Jonathan Huang. The auto arborist dataset: A large-scale benchmark for multiview urban forest monitoring under domain shift. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21262–21275, 2022. 1

[BWE+22b] Sara Meghan Beery, Guanhang Wu, Trevor Edwards, Filip Pavetić, Bo Majewski, Shreyasee Mukherjee, Stan Chan, John Morgan, Vivek Mansing Rathod, and Jonathan Chung-kuan Huang. The auto-arborist dataset: A large-scale benchmark for generalizable, multimodal urban forest monitoring. 2022. 3

[BZOR+18] Marcus A. Badgeley, John R. Zech, Luke Oakden-Rayner, Benjamin Scott Glicksberg, Manway Liu, William Gale, Michael V. McConnell, Bethany Percha, Thomas M. Snyder, and Joel T. Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digital Medicine*, 2, 2018. 1

[CAGLS+22] Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, Chris Schwiegelshohn, and Omar Ali Sheikh-Omar. Improved coresets for euclidean k-means. *Advances in Neural Information Processing Systems*, 35:2679–2694, 2022. 3

[CBL+20] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 3

[CDS20] Rachit Chhaya, Anirban Dasgupta, and Supratim Shit. On coresets for regularized regression. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, 2020. 3

[Che09] Ke Chen. On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM J. Comput.*, 39(3):923–947, 2009. 3

[Cla10] Kenneth L. Clarkson. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Trans. Algorithms*, 6(4):63:1–63:30, 2010. 3

[CMM17] Michael B. Cohen, Cameron Musco, and Christopher Musco. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 1758–1777, 2017. 3

[CYM+19] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*, 2019. 3

[CZPR23] Rhys Compton, Lily H. Zhang, Aahlad Manas Puli, and Rajesh Ranganath. When more is less: Incorporating additional datasets can hurt performance by introducing spurious correlations. *ArXiv*, abs/2308.04431, 2023. 1

[DBK+20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 6

[DDH+08] Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W. Mahoney. Sampling algorithms and coresets for l_p regression. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 932–941, 2008. 3

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[EDHG+20] Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online, November 2020. Association for Computational Linguistics. 3

[FXC+24] Benjamin Feuer, Jiawei Xu, Niv Cohen, Patrick Yubeaton, Govind Mittal, and Chinmay Hegde. Select: A large-scale benchmark of data curation strategies for image classification. *arXiv preprint arXiv:2410.05057*, 2024. 3

[GIF+24] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3, 7

[GJM+20] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665 – 673, 2020. 1

[HM04] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 291–300, 2004. 3

[HV20] Lingxiao Huang and Nisheeth K. Vishnoi. Coresets for clustering in euclidean spaces: importance sampling is nearly optimal. In *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC*, pages 1416–1429, 2020. 3

[HZRS15] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 6

[JDJ19] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 7

[JTMF20] Ibrahim Jubran, Murad Tukan, Alaa Maalouf, and Dan Feldman. Sets clustering. In *International Conference on Machine Learning*, pages 4994–5005. PMLR, 2020. 3

[KH+09] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009. 1

[KSR+21] KrishnaTeja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, Abir De, and Rishabh K. Iyer. GRAD-MATCH: gradient matching based data subset selection for efficient deep model training. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, pages 5464–5474, 2021. 3

[KSRI21] KrishnaTeja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh K. Iyer. GLISTER: generalization based data subset selection for efficient and robust learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, 2021. 3

[KZCI21] Krishnateja Killamsetty, Xujiang Zhao, Feng Chen, and Rishabh Iyer. Retrieve: Coreset selection for efficient and robust semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[LBL+19] Lucas Liebenwein, Cenk Baykal, Harry Lang, Dan Feldman, and Daniela Rus. Provable filter pruning for efficient neural networks. In *International Conference on Learning Representations*, 2019. 3

[LSW+23] Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. The bigscience roots corpus: A 1.6tb composite multilingual dataset, 2023. 3

[MBL20] Baharan Mirzasoleiman, Jeff A. Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, pages 6950–6960, 2020. 3

[MBR+22] Sören Mindermann, Jan Brauner, Muhammed Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt, 2022. 3

[MBY+23] Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Smriti Raje, Max Bartolo, Sabri Eyuboglu, Amirata Ghorbani, Emmett Goodman, Oana Inel, Tariq Kane, Christine R. Kirkpatrick, Tzu-Sheng Kuo, Jonas Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Lilith Bat-Leah, Ce Zhang, James Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi. Dataperf: Benchmarks for data-centric ai development, 2023. 3

[MCL20] Baharan Mirzasoleiman, Kaidi Cao, and Jure Leskovec. Coresets for robust training of deep neural networks against noisy labels. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2020. 3

[MEM+22] Alaa Maalouf, Gilad Eini, Ben Mussay, Dan Feldman, and Margarita Osadchy. A unified approach to coreset learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 3

[MJF19] Alaa Maalouf, Ibrahim Jubran, and Dan Feldman. Fast and accurate least-mean-squares solvers. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 8307–8318, 2019. 3

[MJTF21] Alaa Maalouf, Ibrahim Jubran, Murad Tukan, and Dan Feldman. Coresets for the average case error for finite query sets. *Sensors*, 21(19):6689, 2021. 3

[MMM+22] Raphael A. Meyer, Cameron Musco, Christopher Musco, David P. Woodruff, and Samson Zhou. Fast regression for structured inputs. In *The Tenth International Conference on Learning Representations, ICLR*, 2022. 3

[MRB+23] Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models, 2023. 3

[NMB+19] Mohammad Sadegh Norouzzadeh, Dan Morris, Sara Beery, Neel Joshi, Nebojsa Jojic, and Jeff Clune. A deep active learning system for species identification and counting in camera trap images, 2019. 4

[PGD21] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2021. 3

[PMH+23] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023. 3

[RBC+22] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis and insights from training gopher, 2022. 3

[RLZ+23] Vikram V. Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron B. Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition, 2023. 3, 4

[RSR+20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020. 3

[SGS+22] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *arXiv*, 2022. 3

[SGS+23] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning, 2023. 3

[SHB+17] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv: Machine Learning*, 2017. 4

[SRC24] Judy Hanwen Shen, Inioluwa Deborah Raji, and Irene Y. Chen. The data addition dilemma, 2024. 1

[SRM+22] Shoaib Ahmed Siddiqui, Nitarshan Rajkumar, Tegan Maharaj, David Krueger, and Sara Hooker. Metadata archaeology: Unearthing data subsets by leveraging training dynamics, 2022. 3

[SS18] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[Sta20] Francis Staub. Living planet report 2020: Bending the curve of biodiversity loss, Sep 2020. 4

[TBFR21] Murad Tukan, Cenk Baykal, Dan Feldman, and Daniela Rus. On coresets for support vector machines. *Theor. Comput. Sci.*, 890:171–191, 2021. 3

[TDS+20] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *ArXiv*, abs/2007.00644, 2020. 1

[TJF22] Elad Tolochinsky, Ibrahim Jubran, and Dan Feldman. Generic coreset for scalable learning of monotonic kernels: Logistic regression, sigmoid and more. In *International Conference on Machine Learning, ICML*, 2022. 3

[TMM22] Murad Tukan, Loay Mualem, and Alaa Maalouf. Pruning neural networks via coresets and convex geometry: Towards no assumptions. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022. 3

[TND+22] Alex Tamkin, Dat Nguyen, Salil Deshpande, Jesse Mu, and Noah Goodman. Active learning helps pretrained models learn the intended task, 2022. 3

[TZM+23] Murad Tukan, Samson Zhou, Alaa Maalouf, Daniela Rus, Vladimir Braverman, and Dan Feldman. Provable data subset selection for efficient neural networks training. In *International Conference on Machine Learning*, pages 34533–34555. PMLR, 2023. 3

[WGK17] Oliver Wearn and Paul Glover-Kapfer. Camera-trapping for conservation: a guide to best-practices, 10 2017. 4

[WLC+20] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, May 2020. European Language Resources Association. 3

[WLL+21] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35:8052–8072, 2021. 1

[WLY22] Xudong Wang, Long Lian, and Stella X Yu. Unsupervised selective labeling for more effective semi-supervised learning. In *European Conference on Computer Vision*, pages 427–445. Springer, 2022. 3

[WMX+24] Tsun-Hsuan Wang, Alaa Maalouf, Wei Xiao, Yutong Ban, Alexander Amini, Guy Rosman, Sertac Karaman, and Daniela Rus. Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6687–6694. IEEE, 2024. 6

[WPM+20] Xinyi Wang, Hieu Pham, Paul Michel, Antonios Anastasopoulos, Jaime Carbonell, and Graham Neubig. Optimizing data usage via differentiable rewards. In *International Conference on Machine Learning (ICML)*, 2020. 3

[YLBG20] Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online, November 2020. Association for Computational Linguistics. 3

[ZLX+24] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[ZRG+22] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. 3

CVPR
#*****

A. Additional Dataset Details

B. Additional Training Details

C. Additional Results

C.1. Efficiency

C.2. Linear Probing