Paper under review,
submitted the theNeurIPS'22 dataset and benchmark track

Meta-Album: Multi-domain Meta-Dataset for Few-Shot Image Classification

Ihsan Ullah[*], **Dustin Carrion**[*], **Sergio Escalera**[#%], **Isabelle Guyon**[*%], **Mike Huisman**[+], **Felix Mohr**[‡], **Jan N. van Rijn**[+], **Haozhe Sun**[*], **Joaquin Vanschoren**[§], **Phan Anh Vu**[*]
∗ LISN/CNRS/INRIA, Université Paris-Saclay, France.
Universitat de Barcelona, Spain. ‡ Universidad de La Sabana, Colombia.
+ LIACS, Leiden University, The Netherlands. % ChaLearn, USA.
§ TU/e Eindhoven University of Technology, The Netherlands.
https://meta-album.github.io/

Abstract

We introduce Meta-Album, an image classification meta-dataset designed to facilitate few-shot learning, transfer learning, meta-learning, among other tasks. It includes 40 open datasets, each having at least 20 classes with 40 examples per class, with verified licences. They stem from diverse domains, such as ecology (fauna and flora), manufacturing (textures, vehicles), human actions, and optical character recognition, featuring various image scales (microscopic, human scales, remote sensing). All datasets are preprocessed, annotated, and formatted uniformly, and come in 3 versions (Micro $\subset$ Mini $\subset$ Extended) to match users' computational resources. We showcase the utility of the first 30 datasets (to be released for NeurIPS 2022) on few-shot learning problems. The other 10 will be released shortly after. Meta-Album is already more diverse and larger (in number of datasets) than similar efforts, and we are committed to keep enlarging it via a series of meta-learning competitions. As competitions terminate, their test data are released, thus creating a rolling benchmark, available through OpenML.org. Our website https://meta-album.github.io/ contains the source code of challenge winning methods, baseline methods, data loaders, and instructions for contributing either new datasets or algorithms to our expandable meta-dataset.[1]

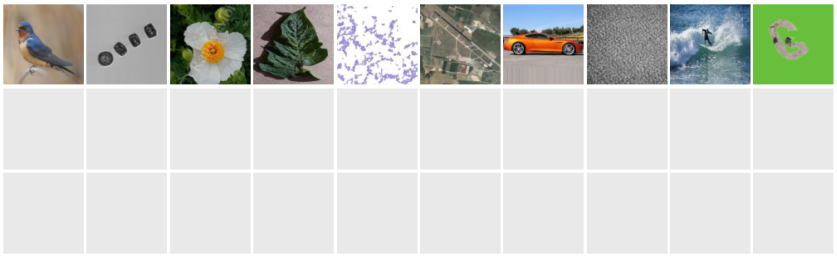


Figure 1: Meta-Album sample images. Each column represents one domain and each row one set. Domains are arranged in the same order as in Table 2.
Note: Only Set-0 is shown; Set-1 and Set-2 images will be revealed only in the final version of the paper. For reviewers, a separate paper copy is provided.

[1] All authors except for the first author are in alphabetical order of last name.

Preprint. Under review.

1 Introduction

1.1 Background

Machine learning has progressed rapidly in recent years and has enabled breakthroughs in various domains. The success of most machine learning techniques hinges on the availability of large amounts of data [47, 24], limiting their applicability in domains where only little data is available. Enabling machine learning algorithms to learn new tasks from only a few examples is studied within the field of *few-shot learning* [29, 42, 49]. Novel meta-learning algorithms have recently been proposed targeting few-shot learning, triggering a surge of popularity for such problems [3, 58, 17, 14]. Despite the popularity of the field, progress is held back by a lack of good, challenging, and computationally feasible meta-datasets, that enable us to accurately assess the generalization abilities of few-shot learning algorithms. To remedy this, we introduce **Meta-Album** (Figure 1), an extensible **multi-domain meta-dataset**, including (so far) 40 image classification datasets from 10 different domains: 30 of them will be released for NeurIPS 2022 and the remaining 10 in spring 2023. This is part of a long-term effort to create a publicly available and growing meta-dataset, in conjunction with a meta-learning challenge series (the 2021 and 2022 editions are both part of the NeurIPS competition program). As competitions terminate, older datasets get released, thus refreshing a rolling benchmark, made available through OpenML.org [54]. We check that all datasets are free for use in academic research and provide their original licenses.

Meta-Album was specifically designed to facilitate meta-learning research in the cross-domain few-shot setting, which is more realistic than commonly used evaluation protocols. Traditionally, few-shot learning algorithms (*e.g.,* [10, 45]) have been evaluated by taking an existing benchmark dataset from a particular "domain" (*e.g.,* handwriting recognition) with a large number of classes, and then breaking it down into smaller classification **tasks**, each including a *random* subset of classes (*e.g.,* a few specific characters). Algorithms are then tested for their ability to solve such tasks "quickly" from a small number of examples, after being trained on many other tasks. Typically the number of classes N and examples per class k are both *fixed* in what is known as an *N-way, k-shot* learning problem. While this setting has served research well, it is not very representative of practical real-world applications where tasks may come from various domains, include classes not drawn at random, but stemming from a class hierarchy, and include any number of classes and/or examples per class. By providing data from a wide variety of domains, including datasets with many classes and a minimum number of examples per class, and retaining class hierarchy annotations, Meta-Album enables benchmarking according to a variety of more realistic settings.

1.2 Related work

In this section, we review meta-datasets previously proposed to benchmark few-shot learning and meta-learning, as well as large-scale multi-class datasets, then contrast them with Meta-Album.

Single dataset benchmarks: Omniglot [22] is often used as a starting benchmark for few-shot and meta-learning. MiniImageNet [56] and Tiered-ImageNet [39] are adapted for few-shot image classification from ImageNet [41]. CIFAR-FS [1] and FC100 [31] are remodeled from CIFAR-100 [20] for few-shot settings.

Multi-dataset benchmarks: A recent trend is to assemble numerous datasets from different domains in the same benchmark. Visual Decathlon [38] gathers 10 diverse datasets. The focus is on finding a model with universal representation capacity for use in many tasks. VTAB (Visual Task Adaptation Benchmark) [61] assembles 19 image classification tasks across various domains. These tasks are grouped into 3 partitions: natural, specialized, and structured. Meta-Dataset [53] includes 10 image classification datasets from several application domains in one collection. Meta-Dataset also leverages the label hierarchy in ImageNet and Omniglot to organize the tasks.

Transfer learning and meta-learning benchmarks: VTAB + MD [8] attempts to unify common transfer and meta-learning datasets in a single benchmark. The authors also provide a comparison of popular meta- and transfer learning methods. BSCD-FSL (Broader Study of Cross-Domain Few-Shot Learning) [12] gathers 4 real-world tasks to compare few-shot, meta-learning and transfer learning methods. WILDS [18] is a benchmark of 10 datasets of various modalities (images, graphs, and text; 6 of them are image datasets), reflecting a diverse range of distribution shifts that naturally arise in real-world applications, and hence useful to evaluate meta-learning and transfer learning techniques.

Table 1: Feature comparison between Meta-Album and other large-scale or (meta-) datasets

Dataset/ Meta-Dataset	# of domains	# of datasets	# of images	min/max classes per domain	min/max images per class	size on disk	multi-domain	lightweight (<20GB)	uniform # of images per class	uniform image size	repeated extensions
Meta-Dataset	7	10	53,068,000	43/1,696	3/140,000	210 GB	✓	✗	✗	✗	✗
VTAB	3	19	2,244,000	2/397	40/1000	100 GB	✓	✗	✗	✗	✗
MS-COCO	1	1	328,000	80/80	9/10,777	44 GB	✗	✗	✗	✗	✗
Mini Imagenet	1	1	60,000	100/100	600/600	1 GB	✗	✓	✓	✓	✗
Omniglot	1	1	32,000	1623/1623	20/20	148 MB	✗	✓	✓	✓	✗
CUB-200	1	1	6,000	200/200	20/39	647 MB	✗	✓	✗	✗	✗
CIFAR-100	3	1	60,000	15/50	600/600	161 MB	✗	✓	✓	✓	✗
Meta-Album *Micro*	**10**	**40**	**32,000**	**19/20**	**40/40**	**380 MB**	✓	✓	✓	✓	✓
Meta-Album *Mini*	**10**	**40**	**220,950**	**19/706**	**40/40**	**3.9 GB**	✓	✓	✓	✓	✓
Meta-Album *Extended*	**10**	**40**	**1,583,624**	**19/706**	**1/187,384**	**15 GB**	✓	✓	✓	✓	✓

CTrL is a continual transfer-learning benchmark [55] including 7 commonly used datasets in image classification. In reinforcement learning, sets of simulation environments exist for meta- and transfer learning, such as Meta-World [60].

Outside few-shot and meta-learning: The AutoDL challenge [26] features a series of 66 datasets from numerous domains. These datasets cover a wide range of modalities: image, video, audio, text, and tabular. This competition focuses on finding a universal algorithm, which can solve many tasks without human supervision.

We compare Meta-Album with previous benchmarks/datasets in Table 1, and provide further details in Appendix F. Meta-Album covers a variety of domains, including ecology, manufacturing, textures, object classification, and character recognition, as well as a variety of scales: microscopic, macroscopic (human scale), or distant (remote sensing). While mostly re-purposing public datasets from heterogeneous sources to maximally vary recording conditions, we also introduce a few fresh datasets in OCR and ecology domains. Meta-Album comprises 3 different versions, Micro $\subset$ Mini $\subset$ Extended: **Micro** includes 20 classes and 40 images per class for ease of running sample code, **Mini** retains all original classes but also includes only 40 examples per class, while **Extended** includes all classes and examples. The variety of versions positions Meta-Album anywhere amongst small-scale datasets such as Omniglot [22], miniImageNet [56, 37] and CUB [57], which usually have at most 70,000 images in total and weigh at most a few GB, or very large-scale benchmarks such as Meta-dataset [53] and VTAB [8], which have more than 50 million images, weigh at least a few hundreds GB, and require high-end super-computer clusters. Its principal distinguishing feature is that it has, by far, the **largest number of domains and datasets**, collected in different conditions, and that it is designed to be **continually extended by either adding new domains or new datasets** in already existing domains, making it a tool of choice for cross-domain, domain-independent, and continual learning studies. Secondly, while other benchmarks usually provide only raw data, we **format all images uniformly as** 128×128 **pixel maps**, which has two benefits: reducing the storage/memory footprint and facilitating the benchmarking of methods independent of preprocessing steps. To that end, we optimized cropping and resizing to reduce dimensions as much as possible without degrading performance too much. In addition, Meta-Album includes datasets that have a **large number of classes** and class hierarchy annotations when available), with a **minimum number of classes and examples per class**: at least 20 classes (except one dataset having only 19 classes) with a minimum of 40 examples per class. This facilitates benchmark design, allowing us to vary the number of classes and the number of training examples per class over a large range of values. Finally and importantly, we selected datasets that are **not typically used in transfer-learning of meta-learning benchmarks** *e.g.,* for pre-training backbone networks, such as ImageNet (which is included *e.g.,* in Meta-Dataset), or for conducting other meta-learning or transfer-learning experiments, such as Omniglot, CIFAR-100, SVHN, or MNIST (which are included *e.g.,* in VTAB and CTrL). This avoids giving an unfair advantage to methods which were developed using such commonly used datasets.

1.3 Contributions and recommended use

In summary, the contributions of our work are the following.

- We provide a **new meta-dataset for few-shot learning and meta-learning** consisting of 40 uniformly formatted datasets from 10 domains, which facilitates research in cross-domain meta-learning as well as practical and realistic evaluation of few-shot algorithms.
- We provide 3 versions of each dataset : Micro, Mini, and Extended to **facilitate usage by researchers with access to different amounts of computational power**.
- We **uniformly pre-processed and formatted data**, but also provide **instructions to retrieve the corresponding raw data** on our website: https://meta-album.github.io.
- We stimulate **community-driven benchmarking**, in conjunction with our challenge series, by welcoming new contributors and providing software and instructions to create additional datasets for Meta-Album, with strict quality-control and review processes.
- We showcase our new meta-dataset by performing an **experimental evaluation** for a number of use cases, including transfer learning, few-shot meta-learning, and cross-domain few-shot meta-learning tasks, using a variety of algorithms, and we **open-source the code used**.

The recommended use of Meta-Album is to conduct fundamental research on machine learning algorithms and perform benchmarks, particularly in few-shot learning, meta-learning, continual learning, transfer learning, and image classification. Meta-Album is not recommended to create products, whether commercial or not, or to derive scientific findings outside benchmarking.

2 Meta-Album design and initial release

In this section, we explain the motivations behind the design of Meta-Album and present the 30 datasets included in the initial NeurIPS 2022 release. 10 more datasets are kept private (but available to the reviewers), and will be released in spring 2023.

2.1 Motivation

Meta-Album emerged from a sequence of few-shot meta-learning benchmarks, following the problem formulations described in Section 3.1. The first of these was the 2020 MetaDL-mini challenge, which was run in conjunction with AAAI 2021. It followed the "within domain few-shot learning" protocl, and algorithms were evaluated with small-scale public datasets (Omniglot and CIFAR-100). Subsequently, we designed a first version of Meta-Album, including 15 datasets, for a larger-scale "within domain few-shot learning" challenge (MetaDL @ NeurIPS 2021). Here, algorithms were meta-trained and meta-tested on tasks extracted from a single dataset at a time, and performances were averaged over 5 datasets, both in the feedback phase and the final evaluation phase, to obtain a more robust evaluation. The 5 extra datasets were provided for practice purposes. The results of this challenge (further detailed in Section 3) indicated that these tasks were well within reach of state-of-the-art methods. This motivated us to move to the "cross-domain few-shot learning" setting. The design of this new challenge (part of an official NeurIPS 2022 challenge) motivated us to grow Meta-Album to 30 datasets spanning multiple domains. We intend to continue growing Meta-Album and already have 10 more datasets lined up, in preparation of the next challenge. This will constitute a *rolling benchmark*: with each new challenge, previous feedback datasets are publicly released, previous final evaluation datasets become feedback datasets, and fresh datasets become final evaluation datasets.

Existing meta-datasets did not allow us to carry out our challenge program for several reasons: (1) they included datasets too familiar to the meta-learning community; (2) they did not include enough datasets to robustly evaluate participants (particularly in the cross-domain setting); (3) their datasets had a large variance in number of classes and examples per class, introducing bias in our experimental design. This required us to source new datasets. Furthermore, since these challenges include code submission, providing the same resources to all participants, we needed to limit computational resources. Therefore, we had to downscale images while taking care that this does not significantly degrade performance.

2.2 Data search

Many people were involved in the sourcing of all datasets and their preparation, and they are gratefully acknowledged in our acknowledgements. This collaboration followed precise instructions to identify datasets which are: (i) from the same domain; (ii) freely available for academic research; (iii) having at least 20 classes with at least 40 examples per class; (iv) with images of good enough quality by visual inspection and with no offensive material (we excluded "deprecated" datasets); (v) with baseline performance within a given range.

The last criterion was needed to ensure the success of our challenges, since tasks that are too easy or too hard do not allow us to separate challenge participants. For the purpose of designing Meta-Album, we defined a "domain" according to four characteristics: (1) application domain; (2) pattern recognition problem (texture or object classification); (3) scale: micro, human scale, or distant; (4) input channels. We ended up with 10 domains (see Table 2): Large animals, small animals, plants, plant diseases, microscopy, remote sensing, vehicles, manufacturing, human actions, optical character recognition (OCR). Data sources were very varied, mostly came from Internet searches, but we also produced our own OCR datasets and obtained novel donated data.

2.3 Data preparation

We performed several iterations of pre-processing, experiments, and analyses to prepare the datasets. This workflow included identifying and, when possible, correcting bias and artifacts (including artifacts we may have introduced by resizing and cropping images), and making sure that images are recognizable by human eye inspection.

As we work with datasets from diverse sources, each dataset requires a different pre-processing strategie, *e.g.,* the small animals datasets, plant-diseases datasets, manufacturing, and remote sensing datasets have images in different resolutions and orientations. However, usually, the object of interest lies in the middle of the image, which facilitated cropping images horizontally or vertically to get squared images. In some cases, like for the plankton dataset, cropping did not make sense, and we resorted to use image padding. In other cases, the area of interest was not necessarily centered, like for human action datasets, and we had to use a human face detector to locate the subject, then we cropped the upper body. For all datasets except for the optical character recognition datasets, which are synthetically generated directly to the correct dimension by OmniPrint [48], we resized the images to a 128×128 resolution using an anti-aliasing filter [2]. The pre-processed data was formatted in a Data format conserving as much meta-data as possible. More details about data preparation and formatting can be found in Appendix C.

2.4 Initial Meta-Album release

The initial release of Meta-Album consists of 3 datasets for each of 10 domains. Each dataset has 3 versions as explained in section 1. All datasets are annotated with class labels and other meta-data. All 30 datasets were chosen after careful and critical analysis during the data preparation and quality control steps as described in Appendix C. Table 2 provides statistics on the various versions; Figure 1 shows sample image from each dataset. More details about datasets and their meta-data are listed in Appendix A. License information for all datasets can be found in Appendix B. Meta-Album datasets will be used in the NeurIPS Cross-domain meta-learning Challenge 2022. After the competition is concluded, all 30 datasets will be available on the OpenML platform [54], then, later in spring 2023, 10 more datasets will be released, followed by other releases as our challenge program unfolds. Details about how to access Meta-Album datasets, contribute to the open meta-dataset, prepare new datasets with quality control, and submit these datasets for inclusion in Meta-Album can be found on the Meta-Album Website. This web page will also inform on software updates and revisions or new releases of our meta-dataset.

3 Use cases and baselines

This section illustrates how Meta-Album can be used for a variety of purposes. The code of all experiments is provided in our Github repository `https://github.com/ihsaan-ullah/meta-album`, and can serve as a basis to benchmark new algorithms against the baseline methods we investigate here. The problems investigated range from few-shot learning (for which Meta-Album was designed)

Table 2: Meta-Album: Datasets summary (*Mini versions*)
Note: 20 datasets' identities and statistics are hidden and will be revealed in the final version of the paper. For reviewers, a separate copy is provided.

Domain ID	Domain Name	Dataset ID	Set #	Dataset Name	# Categories	# Images	Original source
		BRD	0	Birds	315	12,600	Birds 400 [36]
LR_AM	Large Animals	*LR_AM_2*	1	Dataset 2	XX	YY	
		LR_AM_3	2	Dataset 3	XX	YY	
		PLK	0	Plankton	86	3,440	WHOI [46]
SM_AM	Small Animals	*SM_AM_2*	1	Dataset 2	XX	YY	
		SM_AM_3	2	Dataset 3	XX	YY	
		FLW	0	Flowers	102	4,080	Flowers [30]
PLT	Plants	*PLT_2*	1	Dataset 2	XX	YY	
		PLT_3	2	Dataset 3	XX	YY	
		PLT_VIL	0	PlantVillage	38	1,520	PlantVillage [15, 32]
PLT_DIS	Plant Diseases	*PLT_DIS_2*	1	Dataset 2	XX	YY	
		PLT_DIS_3	2	Dataset 3	XX	YY	
		BCT	0	Bacteria	33	1,320	DiBas [63]
MCR	Microscopy	*MCR_2*	1	Dataset 2	XX	YY	
		MCR_3	2	Dataset 3	XX	YY	
		RESISC	0	RESISC	45	1,800	RESISC45 [6]
REM_SEN	Remote Sensing	*REM_SEN_2*	1	Dataset 2	XX	YY	
		REM_SEN_3	2	Dataset 3	XX	YY	
		CRS	0	Cars	196	7,840	Cars [19]
VCL	Vehicles	*VCL_2*	1	Dataset 2	XX	YY	
		VCL_3	2	Dataset 3	XX	YY	
		TEX	0	Textures	64	2,560	KTH-TIPS [11, 28] Kylberg [21] UIUC [23]
MNF	Manufacturing	*MNF_2*	1	Dataset 2	XX	YY	
		MNF_3	2	Dataset 3	XX	YY	
		SPT	0	100 Sports	73	2,920	100 Sports [35]
HUM_ACT	Human Actions	*HUM_ACT_2*	1	Dataset 2	XX	YY	
		HUM_ACT_3	2	Dataset 3	XX	YY	
		MD_MIX	0	OmniPrint-MD-mix	706	28,240	OmniPrint [48]
OCR	Optical Char. Recog.	*OCR_2*	1	Dataset 3	XX	YY	
		OCR_3	2	Dataset 2	XX	YY	

to multi-class image classification, transfer learning, hierarchical classification, and continual learning. Because of lack of space, we only report few-shot learning experiments.

3.1 Problem setting

In this paper we focus on **few-shot image classification**, where the goal is to **learn to perform new classification tasks from a limited number of examples**. Here, every task $\mathcal{T}_j = (\mathcal{D}^{train}_{\mathcal{T}_j}, \mathcal{D}^{test}_{\mathcal{T}_j})$ consists of a *support set* $\mathcal{D}^{train}_{\mathcal{T}_j}$ with training examples and a *query set* $\mathcal{D}^{test}_{\mathcal{T}_j}$ with test examples[2]. In N-way k-shot classification, we require that every *support set* contain exactly N classes with k examples per class ($kN = |\mathcal{D}^{train}_{\mathcal{T}_j}|$). Another requirement is that the classes in the query set must occur in the support set.

Few-shot learning does not necessarily require meta-learning. As in other "regular" learning problems, a *learner*, having available a set of training examples $\mathcal{D}^{train}_{\mathcal{T}_j}$ for a given task, can just return a *trained model* (classifier). But meta-learning is frequently used to enhance few-shot learning.

In a meta-learning problem, a *meta-learner*, having available a set of m training *tasks* $\mathcal{M}^{train}_{\mathcal{D}} = \{T_j\}_{j=1}^{m}$, returns a meta-trained *learner*. In order to develop a meta-trained few-shot *learner*, available data organized in tasks $\mathcal{M}_{\mathcal{D}}$ (coming either from one or multiple datasets) are split into three "meta-splits" containing *disjoint sets of classes*: *meta-training* split $\mathcal{M}^{train}_{\mathcal{D}}$, *meta-validation* split $\mathcal{M}^{valid}_{\mathcal{D}}$, and *meta-testing* split $\mathcal{M}^{test}_{\mathcal{D}}$. The *learner* is meta-trained with $\mathcal{M}^{train}_{\mathcal{D}}$. During meta-training, the *learner* is evaluated with $\mathcal{M}^{valid}_{\mathcal{D}}$ every few meta-training cycles, to monitor progress. The final product of meta-training when the time budget has elapsed, is the *learner* with highest performance on $\mathcal{M}^{valid}_{\mathcal{D}}$ tasks. It is then evaluated on tasks from $\mathcal{M}^{test}_{\mathcal{D}}$.

Within the realm of few-shot learning, we distinguish two cases. **Within domain few-shot learning** refers to the problem where data from the meta-validation and meta-test splits come from the same domain as meta-training data. Here, domain refers to one single dataset of Meta-Album

[2]The nomenclature *support set* instead of *training set*, and *query set* instead of *test set* is common in the meta-learning literature. It highlights the fact that, when meta-training on *tasks = {support set, query set}*, we are not "training on test data", which is a "no no" in machine learning. Meta-test data also includes pairs of support and query sets, from which the ground truth of query set samples is hidden from the classifier.

$\mathcal{D}_i,\ i \in \{1, \ldots, 30\}$. We enforce that $\mathcal{D}_i$ is partitioned into $\mathcal{M}_{\mathcal{D}_i}^{train}$, $\mathcal{M}_{\mathcal{D}_i}^{valid}$, and $\mathcal{M}_{\mathcal{D}_i}^{test}$, using three disjoint sets of classes. In this setting, the goal of *learners* is to learn tasks including classes coming from the same original domain/dataset. If the *learner* has been meta-trained, **test tasks include new classes unseen during meta-training**. **Cross-domain few-shot learning**, in contrast, is a setting for which meta-split is performed at *dataset level* instead of *class level*. For example, in the experiments of subsection 3.2, we use one dataset from every Meta-Album domain to form meta-training, meta-valid, and meta-test splits: $\mathcal{M}_{\mathcal{D}}^{train} = \{\mathcal{D}_1, \ldots, \mathcal{D}_{10}\}$, $\mathcal{M}_{\mathcal{D}}^{valid} = \{\mathcal{D}_{11}, \ldots, \mathcal{D}_{20}\}$, and $\mathcal{M}_{\mathcal{D}}^{test} = \{\mathcal{D}_{21}, \ldots, \mathcal{D}_{30}\}$. In this approach, inside each $\mathcal{M}_{\mathcal{D}}$ there are no overlapping domains, *i.e.*, each $\mathcal{M}_{\mathcal{D}}$ has one dataset per domain. The goal here is to learn tasks sampled from various datasets. If the *learner* has been meta-trained, **test tasks come from new datasets unseen during meta-training**.

We also distinguish between **fixed N-way k-shot** evaluations and **any-way any-shot** evaluations. The former requires fixing the value of N and k for the entire benchmark. The latter requires randomly choosing N and k for each task, within pre-defined ranges. Meta-Album allows us to choose $N \in [2, 20]$ and $k \in [1, 20]$.

3.2 Experiments

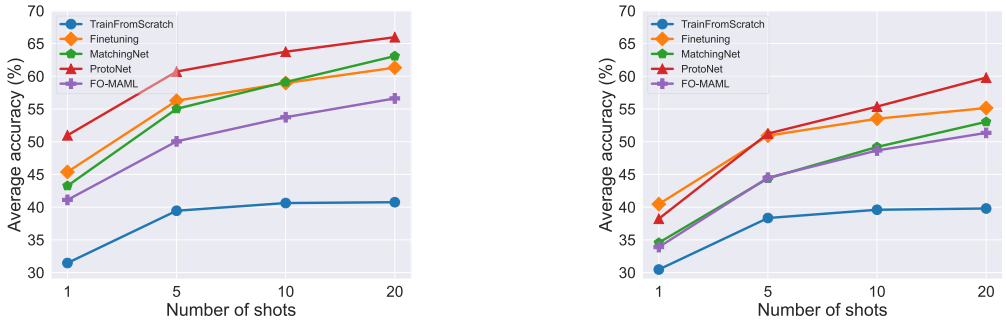
The first motivational use of Meta-Album has been the NeurIPS 2021 MetaDL challenge. This was a meta-learning challenge with code submission, aiming at evaluating **few-shot learning methods in the within domain setting**, as described in subsection 3.1. The evaluation was carried out with 600 tasks in the **5-way 5-shot setting**, using a subset of Meta-Album (Feedback phase: SM_AM.PLK, MDN.MLD, MNF.TEX_DTD, REM_SEN.RSICB, OCR.MD_MIX. Final test phase: SM_AM.INS, PLT_DIS.PLT_VIL, MNF.TEX, REM_SEN.RESISC, OCR.MD_5_BIS). The solutions of the top participants have been open-sourced. In a paper, authored collaboratively between the competition organizers and the top-ranked participants [9], we analyse the results of the competition. The lessons learned include that learning good representations is essential for effective transfer learning. The winner's solution MetaDelta++ [5], based on a combination of pre-trained backbone networks, performed best on all final 5 test phase datasets, with impressive accuracies (0.98, 0.94, 0.99, 0.92, 0.94). This indicates that, in future challenges, we are ready to tackle harder tasks, and motivated us to move to **cross-domain few-shot learning**, in the **any-way any-shot setting** for the NeurIPS 2022 challenge. Fine-tuning backbones on meta-training data turned out to be important, though there are indications that off-the-shelf backbones pre-trained with self-supervised learning on massive datasets might become the way of the future, essentially making meta-learning unnecessary for image classification problems. Thus, meta-learning should be benchmarked in **de novo training conditions**, in the future, to prepare for scenarios (in other domains) in which such backbones are not available. The NeurIPS 2022 challenge encourages *de novo* training in a dedicated league.

Difficulty of cross-domain few-shot learning

To evaluate the gap in difficulty between "within domain" and "cross-domain" few-shot learning problems (subsection 3.1), we carried out first experiments in the 5-ways [1, 5, 10, 20]-shot setting. For all experiments, we use Meta-Album Mini, single PNY GeForce RTX 2080TI GPUs with 11GB of VRAM or a single NVIDIA V100 with 16GB of VRAM. Each experimental run took at most 24 hours on the former GPU (for details, please see Appendix D and Appendix E).

We investigated the few-shot learning performance of popular meta-learning methods: MAML [10], Matching networks [56], and Prototypical networks [45]. We compared them against two baseline methods: TrainFromScratch (learning every task starting from a random initialization at meta-test time, *i.e.,* no meta-learning) and FineTuning, which is pre-trained on the classification problem arising from concatenating all meta-training classes and corresponding data and only fine-tunes the last layer at meta-test time. All techniques use a ResNet-18 backbone [13] and are trained from scratch on Meta-Album (not using any pre-trained feature extractors) using the best-reported hyperparameters by the original authors on 5-way 5-shot miniImageNet (*i.e.,* for FineTuning the backbone is pre-trained with Meta-Album meta-training data only).

For a given dataset, all meta-learning techniques are meta-trained on 60,000 tasks. However, the backbone used for FineTuning is meta-trained (pre-trained) on 60,000 randomly sampled batches of size 16. The performance of trainers is validated every 2,500 tasks (or batches in case of pre-training the FineTuning backbone). The query set for every task contains 16 examples per class, following



(a) Within domain (b) Cross-domain

Figure 2: **Comparison of "within domain" and "cross-domain" few-shot learning.** We plot 5-way [1, 5, 10, 20]-shot learning meta-test mean task accuracy, averaged over 1800 tasks drawn from the 30 released Meta-Album datasets. Corresponding 95% confidence intervals are within the size of the symbols.

[5]. The learning algorithm with the best validation performance is evaluated on 600 meta-test tasks randomly sampled from the meta-testing split, which has information from unseen classes during training and validation. We average the results over 3 runs with different random seeds. Error bars are 95% confidence intervals of the mean over all meta-test tasks in all runs (1800 tasks per dataset).

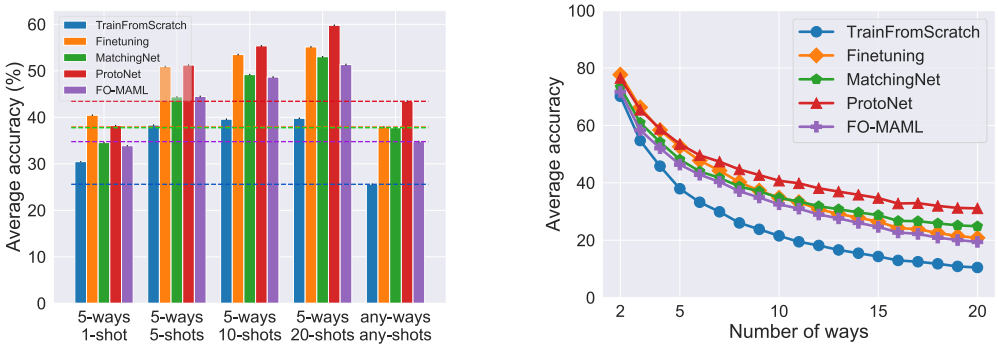
Results are shown in Figure 2. A first observation is that Prototypical Networks (ProtoNet) dominate other algorithms (both within domain and cross-domain) and that the ranking of algorithms does not significantly change with the number of shots. However, the exception is FineTuning for 1-shot learning in the cross-domain configuration, which outperforms ProtoNet by a small margin. Moreover, we observe that FineTuning outperforms MAML and Matching Networks (the other episodic meta-learning algorithm we tried), corroborating findings showing that finetuning yields excellent few-shot learning performance without using episodic meta-learning [51, 16, 5, 52]. We also see that the naive baseline TrainFromScratch yields the worst performance, indicating that meta-learning actually helps transfer knowledge to new tasks. Furthermore, we observe that the performances improve with the number of shots (training examples per class). Lastly, the details provided in Appendix D and Appendix E show that FineTuning is the fastest method at training time while ProtoNet and MatchingNet are the fastest methods at inference time with less than 1 second per task.

For cross-domain few-shot learning, as can be expected, the accuracy is lower since the problem is more complex. However, it does not dramatically decrease compared to within domain few-shot learning, which let us hope that such new problem is within reach of today's state of the art. This gives rise to new opportunities for improvement in this more complicated and more realistic setting.

Difficulty of "any"-way "any"-shot learning

Moving to yet more realistic and harder tasks, we also investigated the performance in the "any"-way "any"-shot setting, where tasks at meta-test time include a varying number of classes between 2 to 20 and a varying number of examples per class between 1 to 20. For example, at meta-test and meta-validation time, some carved out tasks might be as follows: **Test task 1:** 5-way 1-shot task from Dataset 9; **Test task 2:** 3-way 15-shots task from Dataset 3; **Test task 3:** 12-way 4-shots task from Dataset 8; etc. However, at meta-training time, we kept the number of classes constant (specifically, we used 5-way any-shot tasks). This facilitates using off-the-shelf meta-learning techniques. All other experimental conditions (hyper-parameters, computational resources) are the same as in the previous section.

In Figure 3a we can observe that the complexity of the any-way any-shot setting is similar to the 5-way 1-shot setting. Nevertheless, the meta-learning approaches (ProtoNet, MatchingNet, MAML) adapt better to this novel setting since their performance is better than the one achieved in the 5-way 1-shot setting, while the performance of FineTuning and TrainFromScractch is worse compared to the same setting. Additionally, the results presented in Figure 3b and Appendix Appendix E show that the dominant difficulty factor in any-way any-shot learning is the variability in the number of



(a) Difficulty of fixed and variable # of ways and shots

(b) Influence of # of ways

Figure 3: **Comparison of "cross-domain" few-shot learning using fixed and variable number of ways and shots, and influence of number of ways on performance.** We plot few-shot learning meta-test mean task accuracy, averaged over test tasks drawn from the 30 released Meta-Album datasets. Corresponding 95% confidence intervals are almost imperceptible as they are around ± 0.15.

ways since as it can be seen, the performance of the evaluated methods is highly affected by the increment in this number. This is supported by the fact that the absolute Pearson correlation between the number of ways and the test accuracy is larger (r=-0.55, $p<0.05$) than the correlation between the number of shots and the accuracy (r=0.1, $p<0.05$). Therefore, we anticipate that this new setting of any-way any-shot learning will deliver new interesting results in the upcoming challenge.

4 Discussion and conclusion

We introduce Meta-Album, a new meta-dataset for few-shot image classification, which is both practical and extensive: it includes many datasets from a wide variety of domains, all preprocessed to allow training according to different settings on commodity GPUs. It is especially amenable to evaluating meta-learning and transfer learning techniques. It can also be used for hierarchical classification as well as domain adaptation, due to the presence of overlapping classes between datasets, and continual learning, where algorithm are progressively trained across datasets.

We evaluate the utility of Meta-Album using a range of few-shot learning experiments. Our findings include that Prototypical Networks and FineTuning baseline methods perform quite well. This corroborates the results of the NeurIPS'21 challenge, in which the winners capitalized on the use of pre-trained backbones, obtaining results in the high 90% classification accuracy in the "within domain" 5-way 5-shot setting. Meta-Album will further challenge them and other participants by being considerably larger and by mixing tasks from multiple domains, in [2-20]-ways [1-20]-shots settings. Furthermore, Meta-Album allows *de novo* training in a dedicated league. We tested and compared this new framework to that of previous challenges and demonstrated an increased difficulty on all our baseline methods.

In preparing the datasets we identified several types of biases, including correlations between class labels and nuisance variables (*e.g.,* background, luminosity, contrast, color spectrum, position and orientation of objects). In this first release, we avoided correcting such biases, to avoid introducing yet more bias, and opted to homogenize the datasets by shuffling the examples. We documented our findings to facilitate the creation of challenges that study the problem of bias, in which the (meta-)training data and (meta-)test data will have distribution shifts.

In future work, we want to avoid that tasks randomly draw subsets of classes, since real-world tasks often include classes that are part of super-classes (for example insects from the Coleoptera order resemble more one another than say butterflies). This requires datasets in which class hierarchies are provided, and we only have a few of those. Further work also include replacing the "cross-domain" setting by more difficult "domain independent" settings, in which meta-training and meta-testing are performed on different domains.

Acknowledgments and Disclosure of Funding

We gratefully acknowledge the data owners/creators:
Note: Some data owners/creators are hidden from this section to keep the datasets secret. These acknowledgements will be shown in the final version of the paper. For reviewers, a separate copy is provided.

LR_AM.BRD: Gerald Piosenka, *Scottsdale, Arizona, United States*;

SM_AM.PLK: Heidi M. Sosik, Emily E. Peacock, Emily F. Brownlee and Eric Orenstein from *Woods Hole Oceanographic Institution, United States*;

PLT.FLW: Maria-Elena Nilsback and Andrew Zisserman from *University of Oxford, England*;

PLT_DIS.PLT_VIL: Sharada Mohanty, David Hughes, and Marcel Salathé, from *EPFL Switzerland* and *Penn State University*, J. Arun Pandian and G. Geetharamani, from *Department of Mathematics, University College of Engineering, Anna University - BIT Campus and Department of Computer Science and Engineering, M.A.M. College of Engineering and Technology, Tiruchirappalli, India*;

MCR.BCT: Bartosz Zieliński, Anna Plichta, Krzysztof Misztal, Przemysław Spurek, Monika Brzychczy-Włoch and Dorota Ochońska from *Uniwersytet Jagielloński*;

REM_SEN.RESISC: Gong Cheng, Junwei Han, and Xiaoqiang Lu from *Northwestern Polytechnical University, Xi'an, China*;

VCL.CRS: Jonathan Krause, Michael Stark, Jia Deng and Li Fei-Fei from *Stanford University*;

MNF.TEX: Eric Hayman, Barbara Caputo, Mario Fritz, P. Mallikarjuna and Alireza Tavakoli Targhi from *KTH Royal Institute of Technology in Stockholm* (for KTH TIPS and KTH TIPS 2); Gustaf Kylberg from *Uppsala University, Sweden* (for Kylberg Texture); Jean Ponce, Svetlana Lazebnik and Cordelia Schmid from *University of Illinois Urbana-Champaign* (for UIUC Textures);

HUM_ACT.SPT: Gerald Piosenka, *Scottsdale, Arizona, United States*;

OCR.MD_MIX: Generated by Haozhe Sun (co-author).

We acknowledge the efforts of Maria Belen Guaranda Cabezas, Jilin He, Felix Heron, Gabriel Lauzzana, Romain Mussard, and Manh Hung Nguyen for datasets and datasheets preparation. We also received useful input from many members of the TAU team of the LISN laboratory, Wei Wei Tu from 4Paradigm Inc, China, and the MetaDL technical crew: Adrian El Baz, Zhengying Liu, Adrien Pavao, Jennifer (Yuxuan) He, Yui Man Lui, Sébastien Treguer, Benjia Zhou, and Jun Wan, who participates in identifying datasets and contributed to discussions.

This work was supported by ChaLearn, the ANR (Agence Nationale de la Recherche, National Agency for Research) under AI chair of excellence HUMANIA, grant number ANR-19-CHIA-0022 and Labex Digicosme project ANR11LABEX0045DIGICOSME operated by ANR as part of the program Investissement d'Avenir Idex Paris Saclay (ANR11IDEX000302). In addition, some experiments were performed using the compute resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University. This research was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

[1] L. Bertinetto, J. F. Henriques, P. Torr, and A. Vedaldi. "Meta-learning with differentiable closed-form solvers". In: *International Conference on Learning Representations*. 2019.

[2] G. Bradski. "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools* (2000).

[3] P. Brazdil, J. N. van Rijn, C. Soares, and J. Vanschoren. *Metalearning: Applications to Automated Machine Learning and Data Mining*. 2nd. Springer, 2022.

[4] G. J. Burghouts and J.-M. Geusebroek. "Material-Specific Adaptation of Color Invariant Features". In: *Pattern Recogn. Lett.* 30.3 (2009), pp. 306–313.

[5] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang. "A Closer Look at Few-shot Classification". In: *International Conference on Learning Representations*. ICLR'19. 2019.

[6] G. Cheng, J. Han, and X. Lu. "Remote Sensing Image Scene Classification: Benchmark and State of the Art". In: *Proceedings of the IEEE* 105.10 (2017), pp. 1865–1883.

[7] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. *Describing Textures in the Wild*. 2013. arXiv: 1311.3618 [cs.CV].

[8] V. Dumoulin, N. Houlsby, U. Evci, X. Zhai, R. Goroshin, S. Gelly, and H. Larochelle. "Comparing Transfer and Meta Learning Approaches on a Unified Few-Shot Classification Benchmark". In: *arXiv preprint arXiv:2104.02638* (2021).

[9] A. El Baz et al. "Lessons learned from the NeurIPS 2021 MetaDL challenge: Backbone fine-tuning without episodic meta-learning dominates for few-shot learning image classification". In: *NeurIPS 2021 Competition and Demonstration Track*. On-line, United States, 2021.

[10] C. Finn, P. Abbeel, and S. Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. 2017, pp. 1126–1135.

[11] M. Fritz, E. Hayman, B. Caputo, and J. Eklundh. *THE KTH-TIPS database*. https://www.csc.kth.se/cvap/databases/kth-tips/index.html. 2004.

[12] Y. Guo, N. C. Codella, L. Karlinsky, J. V. Codella, J. R. Smith, K. Saenko, T. Rosing, and R. Feris. *A Broader Study of Cross-Domain Few-Shot Learning*. 2020. arXiv: 1912.07200 [cs.CV].

[13] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[14] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey. "Meta-learning in neural networks: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[15] D. P. Hughes and M. Salathé. "An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing". In: *CoRR* abs/1511.08060 (2015). arXiv: 1511.08060.

[16] M. Huisman, J. N. van Rijn, and A. Plaat. "A Preliminary Study on the Feature Representations of Transfer Learning and Gradient-Based Meta-Learning Techniques". In: *Fifth Workshop on Meta-Learning at the Conference on Neural Information Processing Systems*. 2021.

[17] M. Huisman, J. N. van Rijn, and A. Plaat. "A survey of deep meta-learning". In: *Artificial Intelligence Review* (2021).

[18] P. W. Koh et al. "WILDS: A Benchmark of in-the-Wild Distribution Shifts". In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. 2021, pp. 5637–5664.

[19] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. "3D Object Representations for Fine-Grained Categorization". In: *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*. Sydney, Australia, 2013.

[20] A. Krizhevsky, G. Hinton, et al. *Learning multiple layers of features from tiny images*. Tech. rep. University of Toronto, 2009.

[21] G. Kylberg. *The Kylberg Texture Dataset v. 1.0*. External report (Blue series) 35. Centre for Image Analysis, Swedish University of Agricultural Sciences and Uppsala University, Uppsala, Sweden, 2011.

[22] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. "Human-level concept learning through probabilistic program induction". In: *Science* 350.6266 (2015), pp. 1332–1338. eprint: https://science.sciencemag.org/content/350/6266/1332.full.pdf.

[23] S. Lazebnik, C. Schmid, and J. Ponce. "A sparse texture representation using local affine regions". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.8 (2005), pp. 1265–1278.

[24] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444.

[25] H. Li, X. Dou, C. Tao, Z. Wu, J. Chen, J. Peng, M. Deng, and L. Zhao. "RSI-CB: A Large-Scale Remote Sensing Image Classification Benchmark Using Crowdsourced Data". In: *Sensors* 20.6 (2020), p. 1594.

[26] Z. Liu et al. "Winning solutions and post-challenge analyses of the ChaLearn AutoDL challenge 2019". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), p. 17.

[27] Y. Long, Y. Gong, Z. Xiao, and Q. Liu. "Accurate object localization in remote sensing images based on convolutional neural networks". In: *IEEE Transactions on Geoscience and Remote Sensing* 55.5 (2017), pp. 2486–2498.

[28] P. Mallikarjuna, A. T. Targhi, M. Fritz, E. Hayman, B. Caputo, and J. Eklundh. *THE KTH-TIPS 2 database*. https://www.csc.kth.se/cvap/databases/kth-tips/index.html. 2006.

[29] D. K. Naik and R. J. Mammone. "Meta-neural networks that learn by learning". In: *International Joint Conference on Neural Networks*. Vol. 1. IJCNN'92. IEEE. 1992, pp. 437–442.

[30] M. Nilsback and A. Zisserman. "Automated Flower Classification over a Large Number of Classes". In: *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*. 2008, pp. 722–729.

[31] B. N. Oreshkin, P. R. Lopez, and A. Lacoste. "TADAM: Task dependent adaptive metric for improved few-shot learning". In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. 2018, pp. 719–729.

[32] J. A. Pandian and G. Geetharamani. "Data for: Identification of Plant Leaf Diseases Using a 9-layer Deep Convolutional Neural Network". In: *Mendeley Data, V1* (2019).

[33] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[34] P. Pérez, M. Gangnet, and A. Blake. "Poisson Image Editing". In: *ACM Trans. Graph.* 22.3 (2003), pp. 313–318.

[35] G. Piosenka. *100 Sports Image Classification*. https://www.kaggle.com/datasets/gpiosenka/sports-classification.

[36] G. Piosenka. *BIRDS 400 - SPECIES IMAGE CLASSIFICATION*. https://www.kaggle.com/datasets/gpiosenka/100-bird-species.

[37] S. Ravi and H. Larochelle. "Optimization as a Model for Few-Shot Learning". In: *International Conference on Learning Representations*. ICLR'17. 2017.

[38] S.-A. Rebuffi, H. Bilen, and A. Vedaldi. "Learning multiple visual domains with residual adapters". In: *arXiv preprint arXiv:1705.08045* (2017).

[39] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J. B. Tenenbaum, H. Larochelle, and R. S. Zemel. "Meta-learning for semi-supervised few-shot classification". In: *arXiv preprint arXiv:1803.00676* (2018).

[40] S. Roopashree and J. Anitha. *Medicinal Leaf Dataset*. https://data.mendeley.com/datasets/nnytj2v3n5/1. Version 1. 2020.

[41] O. Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252.

[42] J. Schmidhuber. "Evolutionary Principles in Self-Referential Learning". Diploma Thesis. Technische Universität München, 1987.

[43] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[44] H. Serret, N. Deguines, Y. Jang, G. Lois, and R. Julliard. "Data quality and participant engagement in citizen science: comparing two approaches for monitoring pollinators in France and South Korea". In: *Citizen Science: Theory and Practice* 4.1 (2019), p. 22.

[45] J. Snell, K. Swersky, and R. Zemel. "Prototypical Networks for Few-shot Learning". In: *Advances in Neural Information Processing Systems*. Vol. 30. NIPS'17. 2017.

[46] H. M. Sosik, E. E. Peacock, and E. F. Brownlee. *Annotated Plankton Images - Data Set for Developing and Evaluating Classification Methods*. https://hdl.handle.net/10.1575/1912/7341. 2015.

[47] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. "Revisiting unreasonable effectiveness of data in deep learning era". In: *Proceedings of the IEEE International Conference on Computer Vision*. ICCV'17. 2017, pp. 843–852.

[48] H. Sun, W.-W. Tu, and I. M. Guyon. "OmniPrint: A Configurable Printed Character Synthesizer". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2021.

[49] S. Thrun. "Lifelong Learning Algorithms". In: *Learning to learn*. 1998, pp. 181–209.

[50] P. J. Thul, L. Åkesson, M. Wiking, D. Mahdessian, A. Geladaki, H. Ait Blal, T. Alm, A. Asplund, L. Björk, L. M. Breckels, et al. "A subcellular map of the human proteome". In: *Science* 356.6340 (2017), eaal3321.

[51] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola. "Rethinking few-shot image classification: a good embedding is all you need?" In: *European Conference on Computer Vision*. Springer. 2020, pp. 266–282.

[52] E. Triantafillou, V. Dumoulin, H. Larochelle, and R. Zemel. *Learning Flexible Classifiers with Shot-{CON}ditional Episodic ({SCONE}) Training*. https://openreview.net/forum?id=0MjC3uMthAb. 2021.

[53] E. Triantafillou et al. "Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples". In: *International Conference on Learning Representations*. 2020.

[54] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo. "OpenML: networked science in machine learning". In: *SIGKDD Explorations* 15.2 (2013), pp. 49–60.

[55] T. Veniat, L. Denoyer, and M. Ranzato. "Efficient Continual Learning with Modular Networks and Task-Driven Priors". In: *9th International Conference on Learning Representations, ICLR 2021*. 2021.

[56] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. "Matching networks for one shot learning". In: *Advances in Neural Information Processing Systems* 29 (2016), pp. 3630–3638.

[57] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. *The Caltech-UCSD Birds-200-2011 Dataset*. Tech. rep. CNS-TR-2011-001. California Institute of Technology, 2011.

[58] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. "Generalizing from a Few Examples: A Survey on Few-Shot Learning". In: *ACM Comput. Surv.* 53.3 (2020).

[59] Z. Xiao, Y. Long, D. Li, C. Wei, G. Tang, and J. Liu. "High-resolution remote sensing image retrieval based on CNNs from a dimensional perspective". In: *Remote Sensing* 9.7 (2017), p. 725.

[60] T. Yu, D. Quillen, Z. He, R. Julian, A. Narayan, H. Shively, A. Bellathur, K. Hausman, C. Finn, and S. Levine. "Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning". In: *Proceedings of the 3rd Conference on Robotic Learning*. 2021. arXiv: 1910.10897 [cs.LG].

[61] X. Zhai et al. *A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark*. 2020. arXiv: 1910.04867 [cs.CV].

[62] Q. Zhang and S. Zhu. "Visual Interpretability for Deep Learning: a Survey". In: *CoRR* abs/1802.00614 (2018). arXiv: 1802.00614.

[63] B. Zieliński, A. Plichta, K. Misztal, P. Spurek, M. Brzychczy-Włoch, and D. Ochońska. "Deep learning approach to bacterial colony classification". In: *PLOS ONE* 12.9 (2017), pp. 1–14.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See section 4
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] There are no negative societal impacts. Rather, this meta-dataset can foster progress in the fields of few-shot learning and meta-learning. We have added "recommended use" in Section 1.3.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code and instructions are publicly released on the Meta-Album GitHub repository (`https://github.com/ihsaan-ullah/meta-album`)
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
   [Yes] See section 3.2. Additional details can be found in the Meta-Album GitHub repository.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure 2, Figure 3, Appendix D, Appendix E
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] For used resources, see section 3.2, and for detailed running times see Appendix D, Appendix E.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators?
   [Yes] See subsection 3.2, Appendix A and datasheets for dataset
   (b) Did you mention the license of the assets?
   [Yes] See Appendix B
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We provide software for data formatting, data quality control and conversion on the Github repository `https://github.com/ihsaan-ullah/meta-album`
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, we provide details in Appendix B.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] None of the datasets allow for personal identification
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]