International Journal of Computer Vision
https://doi.org/10.1007/s11263-025-02450-2

Bamboo: Building Mega-Scale Vision Dataset Continually with Human–Machine Synergy

Yuanhan Zhang[1] · Qinghong Sun[2] · Yichun Zhou[3] · Zexin He[3] · Zhenfei Yin[4] · Kun Wang[2] · Lu Sheng[3] · Yu Qiao[5] · Jing Shao[5] · Ziwei Liu[1]

Received: 23 August 2023 / Accepted: 8 April 2025
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

Abstract

Large-scale datasets play a vital role in computer vision. But current datasets are annotated blindly without differentiation to samples, making the data collection inefficient and unscalable. The open question is how to build a mega-scale dataset actively. Although advanced active learning algorithms might be the answer, we experimentally found that they are lame in the realistic annotation scenario where out-of-distribution data is extensive. This work thus proposes a novel active learning framework for realistic dataset annotation. Equipped with this framework, we build a high-quality vision dataset—**Bamboo**, which consists of 69M image classification annotations with 119K categories and 28M object bounding box annotations with 809 categories. We organize these categories by a hierarchical taxonomy integrated from several knowledge bases. The classification annotations are four times larger than ImageNet22K, and that of detection is three times larger than Object365. Compared to ImageNet22K and Objects365, models pre-trained on Bamboo achieve superior performance among various downstream tasks (6.2% gains on classification and 2.1% gains on detection). We believe our active learning framework and Bamboo are essential for future work. Code and dataset are available at https://github.com/ZhangYuanhan-AI/Bamboo.

Keywords Vision Dataset · Human-Machine Synergy

1 Introduction

Large-scale pre-trained models, trained in supervised (Ghiasi et al., 2021; Kolesnikov et al., 2020; Yalniz et al., 2019) or unsupervised (He et al., 2020; Chen et al., 2020b; Caron et al., 2020) ways, are now essential for advanced computer vision. These pre-trained models (Radford et al., 2021) are versatile, useful for various tasks by adapting to different needs. The success of these models (Bommasani et al., 2021) depends heavily on access to large and varied datasets (Krizhevsky et al., 2009; Deng et al., 2009; Sun et al., 2017; Lin et al., 2014).

Creating a high-quality dataset involves careful selection from a pool of unlabeled data for annotation. Yet, public

Communicated by Gang Hua.

✉ Ziwei Liu
ziwei.liu@ntu.edu.sg

Yuanhan Zhang
yuanhan002@e.ntu.edu.sg

Qinghong Sun
sunqinghong@senseauto.com

Yichun Zhou
buaazyc@buaa.edu.cn

Zexin He
jacquesdeh@buaa.edu.cn

Zhenfei Yin
zhenfei.yin@sydney.edu.au

Kun Wang
wangkun@senseauto.com

Lu Sheng
lsheng@buaa.edu.cn

Yu Qiao
qiaoyu@pjlab.org.cn

Jing Shao
shaojing@pjlab.org.cn

[1] S-Lab, Nanyang Technological University, Singapore, Singapore

[2] SenseTime Research, Beijing, China

[3] Beihang University, Beijing, China

[4] The University of Sydney, Sydney, Australia

[5] Shanghai AI Laboratory, Shanghai, China

Published online: 13 May 2025

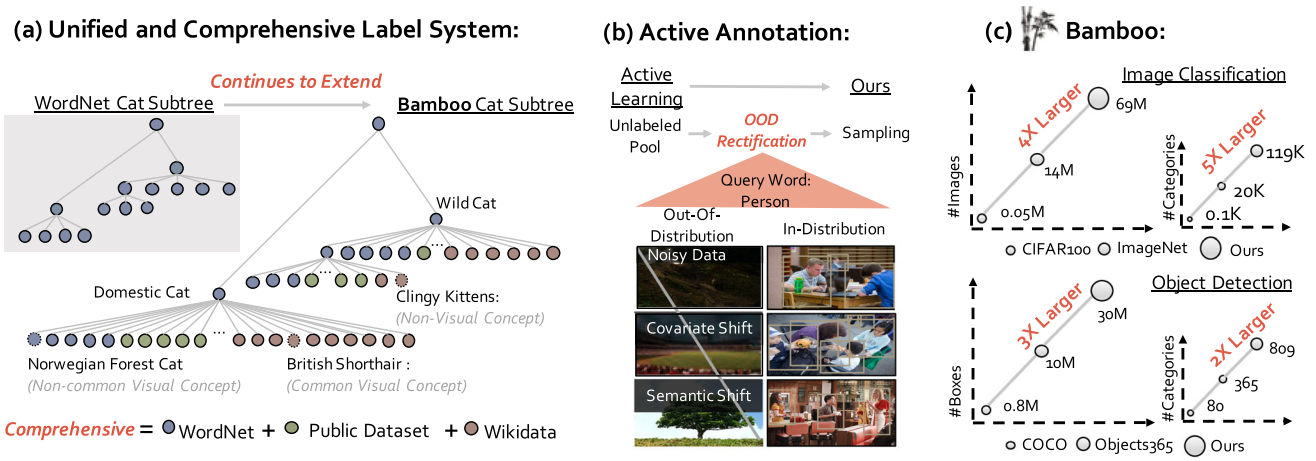

Fig. 1 The overview of *Bamboo* Dataset. Bamboo is a new mega-scale vision dataset built on a comprehensive label system with human–machine synergy. **a** Our label system continually extends from WordNet with our solutions. Concepts in the label system are distinguished as "common visual", "non-common visual" or "non-visual" concepts. **b** Raw data crawled by the query word *person* includes both the in-distribution (ID) data and out-of-distribution (OOD) data. OOD data implies noisy, covariate shift, and semantic shift data. Noisy data does not present useful semantic information. Covariate shift data implies semantic information, i.e. person. However, such semantic information is of poor quality, annotators thus hard to annotate. Semantic shift data also implies semantic information, i.e. tree. But the tree is not related to the query word person. *OOD rectification* mitigates the ineffectiveness of active learning through filtering OOD data. **c** Bamboo collects 69M classification annotation and 28M bounding box annotations

datasets often lack this selective approach, leading to significant waste in annotation efforts. Citovsky et al. (2021) found that only 70% of data in OpenImages (Kuznetsova et al., 2020) are as effective as the entire dataset. Active learning (AL) (Tong & Koller, 2001; Joshi et al., 2009; Settles, 2009; Yang et al., 2015; Gilad-Bachrach et al., 2005; Iglesias et al., 2011; Sznitman & Jedynak, 2010; Vezhnevets et al., 2012; Gal, 2016; Citovsky et al., 2021; Sener & Savarese, 2018; Roth & Small, 2006) has been extensively explored for selecting the most informative samples from an unlabeled data pool. However, we observe that some of these 'informative' samples may be out-of-distribution compared to the desired 'good' samples. For instance, in image classification, given the query label "person" and its web-crawled data, methods like ClusterMargin(Citovsky et al., 2021), Margin (Roth & Small, 2006), and CoreSet (Sener & Savarese, 2018) often choose informative yet out-of-distribution samples. These samples are discarded by annotators and are not used for model training because they are either noisy (not providing useful semantic information), indicate a covariate shift (misaligned but meaningful data), or show semantic shifts (relevant but inaccurate information, e.g., tree instead of person). In contrast, random sampling selects fewer informative samples but includes more data that is relevant to the distribution. When the annotated data from AL is 70% less than that from random sampling, the latter performs better, as the benefits of having more data outweigh those of having 'higher-quality' data (as shown in Fig. 1). To address this issue, we propose a new active learning framework that removes out-of-distribution data from the unlabeled pool before sampling. This ensures that the data chosen are not only informative but also relevant, leading to better performance than random sampling in enhancing the effectiveness of supervised learning models.

We aim to annotate a mega-scale classification and object detection dataset with our proposed active learning framework. First, we build a comprehensive label system for querying diverse data covering numerous semantics. Specifically, we form a unified label system with a hierarchical structure consisting of 304,048 categories. It integrates label systems from 19 latest public image classification datasets and five object detection datasets and also absorbs 170,586 new categories from knowledge bases, e.g. Wikidata (Vrandečić & Krötzsch, 2014). Then, we contribute **Bamboo Dataset**, a mega-scale and information-dense dataset for the pre-training of both classification and detection, which is active annotating by human–machine synergy. Bamboo has three appealing properties:

Comprehensive It consists of 69M image classification annotations and 28M object bounding box annotations, spanning over 119K visual categories. The scale of the label system and the annotated data are the largest among all the publicly available datasets. We illustrate the comparison of Bamboo and publicly available datasets in the Fig. 1c.

Information-dense We guarantee Bamboo is highly informative through the label system and the annotated data. The label system is constructed by thoroughly integrating public datasets and knowledge bases. Our active annotation pipeline

specifically selects the annotated data to reduce model uncertainty.

Continual Our label system keeps the dataset growing with the automatic concept linking strategies. We can constantly absorb new categories in the real world and integrate them into our label system. Moreover, leveraging the ever-increasing internet data, our active annotation pipeline will steadily sustainably expand the Bamboo dataset size.

Extensive experiments demonstrate that Bamboo dataset is an effective pre-training source. The Bamboo pre-trained model significantly outperforms CLIP ViT B/16 (Radford et al., 2021) pre-trained model with 6.2% gain (85.6% vs 91.8%) on classification, and outperforms Objects365 (Shao et al., 2019) pre-trained model with 2.1% gain (14.7% vs 12.6%) on CityPersons (Zhang et al., 2017). In addition, we provide valuable observations regarding large-scale pre-training from over 1000 experiments. We hope the Bamboo dataset and these observations will pave the way for developing more general and effective vision models.

2 Related Works

Learning of Representation at Scale Representation learning has advanced thanks to improvements in various learning paradigms and large-scale datasets. Supervised learning models (He et al., 2016; Chollet, 2017) leverage label information to supervise representation learning, achieving excellent performance among various downstream tasks. To avoid the need for annotations that require a tremendous amount of human and labeling cost, weakly-supervised and self-supervised pre-training methods have been proposed. Self-supervised methods (Caron et al., 2018; Misra & van der Maaten, 2020; He et al., 2020; Chen et al., 2020a; Caron et al., 2020; Zhang et al., 2016; Larsson et al., 2016; Wu et al., 2018; Dosovitskiy et al., 2015; Li et al., 2020) with contrastive learning have shown that unsupervised pre-training produces features surpassing the supervised feature representations on many downstream tasks (Krizhevsky et al., 2009; Nilsback & Zisserman, 2006; Parkhi et al., 2012; Xiao et al., 2016; Krause et al., 2013). Large weakly-supervised datasets, such as Instagram hashtags (Mahajan et al., 2018) and JFT (Sun et al., 2017), helps model (Yalniz et al., 2019; Zoph et al., 2020; Xie et al., 2020) achieve significant gains on downstream tasks. In addition, CLIP (Radford et al., 2021) pre-train models on both the image signal and text signal, achieving good performance for the zero-shot evaluation. Our study is part of a larger body of work on training models on sizeable supervised image datasets. As the labeling cost that hurdles the supervised learning dataset is becoming increasingly significant, we systematically investigate how to collect, annotate and build a mega-scale dataset efficiently, actively and continually.

Active Learning Active learning (AL) aims at finding the minimum number of labeled images to have a supervised learning algorithm reach a certain performance (Tong & Koller, 2001; Joshi et al., 2009; Settles, 2009; Yang et al., 2015; Gilad-Bachrach et al., 2005; Iglesias et al., 2011; Sznitman & Jedynak, 2010; Vezhnevets et al., 2012), [27]. The main component in an active learning loop is sampling strategies. The existing AL research is conducted on the curated datasets. Each data point in the labeled and unlabeled pool of these datasets is valid, i.e. available for labeling. However, curated datasets can hardly imitate the annotation in realistic scenarios where out-of-distribution data that is unavailable for labeling appears on a large scale in the unlabeled pool. From our experiments, we find the existing AL methods lag in realistic scenarios. Therefore, we propose a novel active annotation pipeline to improve the performance of AL methods in realistic scenarios.

3 Label System Construction

In this section, we briefly introduce how to build a comprehensive label system. The number of concepts decides the data amount upper bound—we crawl data based on querying these labels. Starting from WordNet (Miller, 1998), we enrich its concepts from another two concept resources (Sect. 3.1) through three designed linking strategies (Sect. 3.2).

3.1 Concepts Resources

WordNet WordNet is a lexical database of semantic relations between concepts in more than 200 languages. Each meaningful concept in WordNet, possibly described by multiple words or phrases, is called a "synset". Referring to ImageNet22K (Deng et al., 2009), we only use the Noun words of WordNet. WordNet is the foundation of our label system.

Public Datasets We collect 27 public datasets, including ImageNet22K (Deng et al., 2009), , iNaturalist (Van Horn et al., 2018), Herbarium2021,[1] Danish Fungi 2020 (Picek et al., 2021), iWildCam2020 (Beery et al., 2021), TsinghuaDogs (Zou et al., 2020), Places (Zhou et al., 2017), FoodX-251 (Kaur et al., 2019), CompCars (Yang et al., 2015), COCO (Lin et al., 2014), Objects365 (Shao et al., 2019), OpenImages (Kuznetsova et al., 2020), SUN397 (Xiao et al., 2016), Caltech101 (Fei-Fei et al., 2004), CUB (Welinder et al., 2010), FGVC-Aircraft (Maji et al., 2013), STL10 (Coates et al., 2011), EuroSAT (Helber et al., 2019), DTD (Cimpoi et al., 2014), AID (Xia et al., 2017), Places365 (Zhou et al., 2017), Pets (Parkhi et al., 2012), StandfordCars (Krause et al., 2013), Flowers102 (Nilsback & Zisserman, 2006),

[1] https://www.kaggle.com/c/herbarium-2021-fgvc8/overview.

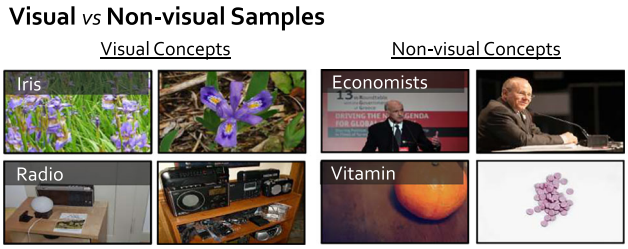

Fig. 2 The illustration of visual and non-visual concept. *Vitamin* do not share any common semantic information. *Economists* implies common semantic information—Man—but economists are not necessarily men

Cifar10 (Krizhevsky et al., 2009), Cifar100 (Krizhevsky et al., 2009), Food-101 (Bossard et al., 2014).

Wikidata Wikidata (Vrandečić & Krötzsch, 2014) contains a large number of concepts, such as different kinds of foods, animals, and locations. As the number of concepts in Wikidata continues to grow, so far, we have included 170,586 concepts from it. These concepts are the leaf nodes in their taxonomy.

3.2 Concepts Integration

WordNet is a lexical graph whose concepts imply semantic relation. For example, the father node of "British Shorthair" is "Domestic Cat". How to integrate concepts from public datasets and Wikidata into this WordNet is an open question. We propose three parallel solutions to integrate these categories into WordNet in this work.

Solution 1: Leveraging on the *subclassOf* The taxonomy of Wikidata is contributed by adding the "subclassOf" that is related to the hypernyms relationship in the taxonomy of WordNet. Referred to Tanon et al. (2020), we link Wikidata leaf node concepts to the WordNet by leveraging the "subclassOf".

Solution 2: Parsing the Concept Referred to the previous work Fabian et al. (2007), we can also link the concept to the WordNet through word parsing. For example, for the concept *Sumatran Orangutan*, we parse this concept (Honnibal & Montani, 2017) and get its head compound "Orangutan". In this way, we add *Sumatran Orangutan* as the new hyponym of the "Orangutan" if "Orangutan" exists in WordNet.

Solution 3: Linking to the Closed Synset We calculate the word embedding of both the synsets and given concepts through Spacy (Honnibal & Montani, 2017). If a given concept cannot be linked to WordNet, we add this category to the hyponym of its nearest cosine distance synset.

3.3 Concepts Tagging

Visuality Yang et al has mentioned the non-visual category problem in their work Yang et al. (2020). We illustrate visual and non-visual words in Fig. 2. To mitigate this problem, we conduct visual concept tagging for our build label system. Specifically, a concept is non-visual if three out of five annotators think this word is less concrete, and its sample images can rarely imply a common semantic meaning. We illustrate the concept tagging in Fig. 3a.

Commonality Based on the visual concepts, we further conduct common concept annotation for all visual concepts. Referred to COCO (Lin et al., 2014), "common concept" refers to entry-level categories that are commonly used by humans when describing objects (e.g. dog, chair, person). Specifically, a concept is positive only if it receives at least three-fifths of the votes. Based on the proposed annotation method, we retain 809 common concepts for the annotation of object detection.

4 Active Dataset Construction—Bamboo

Equipped with the unified and comprehensive label system, we start to construct Bamboo actively. In this section. We first introduce the active learning pipeline for building Bamboo in Sec 4.1. We summarize this pipeline in Algorithm 1. Then in Sect. 4.2, we discuss the superiority of our newly proposed active learning methods—we are the **first** time beat the random sampling in selecting the most valuable data for data pre-training[2].

Algorithm 1: Outline of AL Framework

input : Raw unlabeled pool $\mathcal{P}$; Number of active learning rounds T; Neural network f;
$\mathcal{L}_0 \leftarrow$ Annotating a few data from $\mathcal{P}$ and adding all inherited data as cold start; $\mathcal{U}_0 \leftarrow \mathcal{P} - \mathcal{L}_0$;
Initializing model f_0 based on $\mathcal{L}_0$;
for $r \leftarrow 1$ **to** T **do**
  $\mathcal{P}_r \leftarrow$ Rectifying $\mathcal{U}_{r-1}$ using f_{r-1};
  $\mathcal{U}_r \leftarrow$ Sampling in $\mathcal{P}_r$ using f_{r-1};
  $\mathcal{L}_r \leftarrow$ Annotating valid data from $\mathcal{U}_r$;
  $\mathcal{U}_r \leftarrow \mathcal{U}_{r-1} - \mathcal{U}_r$;
  $\mathcal{L}_r \leftarrow \mathcal{L}_{r-1} \cup \mathcal{L}_r$;
  Training f_r on $\mathcal{L}_r$;
end

4.1 Active Learning Framework

4.1.1 Building Unlabeled Data Pool

For image classification, one query word has one visual concept mentioned in Sect. 3.3. For object detection, one query has two concepts, i.e., common concept + scene semantic word or common concept + common concept. For example,

[2] This step is not included in the current active learning research.

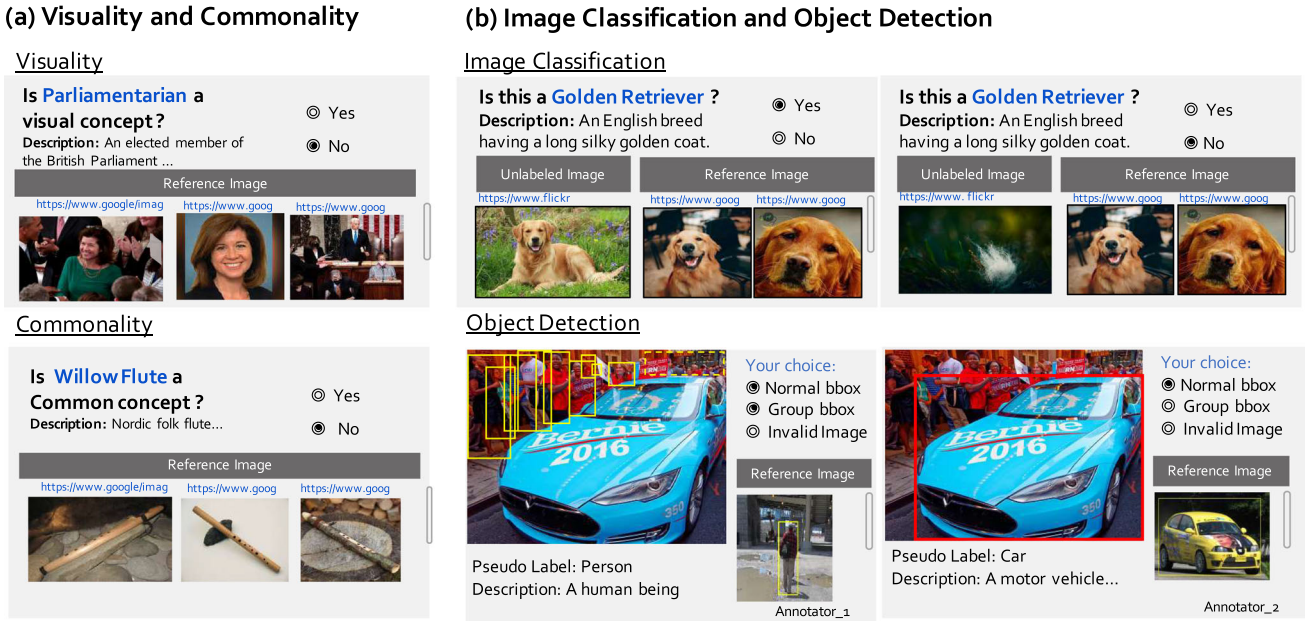

Fig. 3 User interfaces for concept tagging and annotation. **a** The meta information of the concept tagging consists of tags, descriptions, and reference images. **b** Interface for image classification and object detection. For the object detection task. The image is assigned to different annotators based on its multiple pseudo labels. In addition, annotators should choose the attribute of the bounding box. The box that covers more than five bounding boxes of the same category, which heavily occlude each other, should be marked as Group bbox. Other clear bounding box are supposed to be marked as Normal bbox option

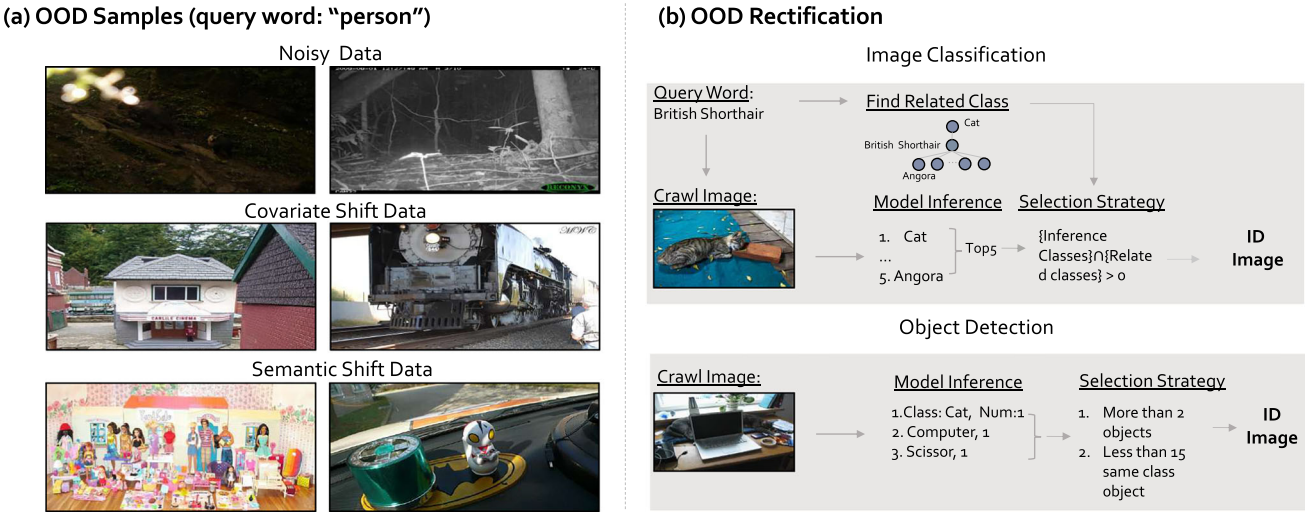

Fig. 4 **a** The illustration of out-of-distribution (OOD) data in realistic scenarios. Mainly, three types of OOD data exist in the unlabeled data pool, including noisy data, covariate shift data (i.e., OOD samples from a different domain), and semantic shift data (i.e., OOD samples are drawn from different classes). **b** The illustration of OOD rectification. OOD rectification filters OOD data in the unlabeled data pool, which is crucial for active learning

dog + *street* or *dog* + *ball*. To further enrich the searching results, any given query word can be converted to its synonyms or its Chinese, Spanish, Dutch and Italian version for querying. Totally, we obtain a 170M unlabeled pool for classification and a 200M unlabeled pool for detection.

4.1.2 Cold Start

Cold start is the very first step for active learning. The labeled data pool $\mathcal{L}(0)$ to initialize the model $\phi(0)$ for the cold start phase include two parts as follows.

Public Dataset As mentioned in Sect. 3.1, we use 24 datasets as concept resources, including 19 image classification datasets and 5 object detection datasets. Refereed to the evaluation suite of popular transfer learning study (Kornblith et al., 2019; Zhai et al., 2019; He et al., 2020), we select 12 datasets for downstream evaluation. We include the annotation of the other 12 datasets—9 image classification datasets and 3 object detection datasets. In total, we inherit 27,848,477 classification annotations and 21,983,223 object bounding box annotations from those 12 datasets.

New Annotated Data For concepts not included in public datasets, we sample images from unlabeled pool Θ and annotate data for them until they have 50 annotated data.

4.1.3 OOD Rectification

Image Classification In this step, we rectify the latest unlabeled data pool $\mathcal{U}_{r-1}^{\text{Cls}}$. As shown in Fig 4b, in each round r, we firstly utilize f_{r-1}^{Cls} trained on $\mathcal{L}_{r-1}^{\text{Cls}}$ to acquire predictions of each image in $\mathcal{U}_{r-1}^{\text{Cls}}$. We infer an image is out-of-distribution if its top-5 predicted categories do not overlap with its related categories. Specifically, we define the related categories of an image as the sub-population of hypernyms of its query word. If an image is not out-of-distribution, we add it into $\mathcal{P}_{r-1}^{\text{Cls}}$ for further data sampling. In Sect. 4.2, we empirically observe that OOD rectification is essential for AL to function in realistic scenarios.

Object Detection Similar to classification, we acquire proposal predictions of each image in $\mathcal{U}_{r-1}^{\text{Det}}$ by f_{r-1}^{Det}. On the one hand, we filter out the image with less than two proposals. Such images might be noisy data or semantic shift data. On the other hand, we filter out the image with more than 15 identical semantic proposals since such image might be the covariate shift data. As shown in Fig 4b, the remaining in-distribution data forms $\mathcal{P}_{r-1}^{\text{Det}}$ for the data sampling.

4.1.4 Data Sampling

In this step, we use ClusterMargin (Citovsky et al., 2021), which considers both the uncertainty and diversity in data, to select the most valuable data from the latest rectified data pool $\mathcal{P}_r$ for annotation.

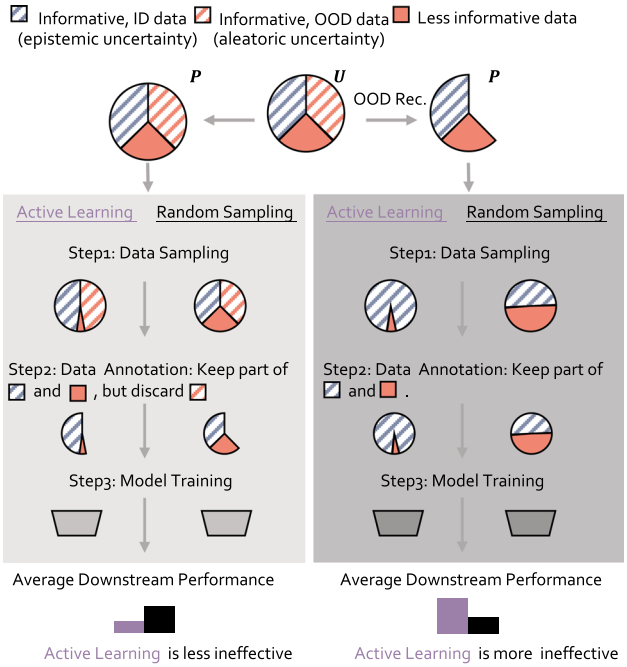

Fig. 5 The Illustration of how our OOD rectification step helps active learning performs better in realistic scenarios

4.1.5 Data Annotation

Finally, we rely on an online platform to annotate valid data—its querying word accurately describes the semantic meaning of this data—in $\mathcal{U}_r$, forming the labeled data set $\mathcal{L}_r$. We illustrate our online platform in Fig. 3, and introduce the details of annotations as follows.

Image Classification To provide a comprehensive definition of each category, we construct reference images that are collected by querying Google image search and Wikipedia (Estimation Lemma, 2010). For each image in $\mathcal{U}_r$, its meta-information has two parts: the query word of this image and the reference images of the query word. We then ask the five annotators whether this image conforms to its meta information. An image is annotated and added into $\mathcal{L}_r$—valid data—only if at least 3 out of 5 annotators give the positive answer to the question as mentioned above.

Object Detection Following Objects365 (Shao et al., 2019), one annotator is responsible for annotating a specific category, which improves the annotation efficiency and quality. Similar to OpenImages (Kuznetsova et al., 2020), meta information of an image includes not only its reference images but also its pseudo labels that include (i) the query words of this image. (ii) the category predictions of available detection models. iii) re-labeling predictions (Yun et al., 2021) of the latest trained classification model f_C.

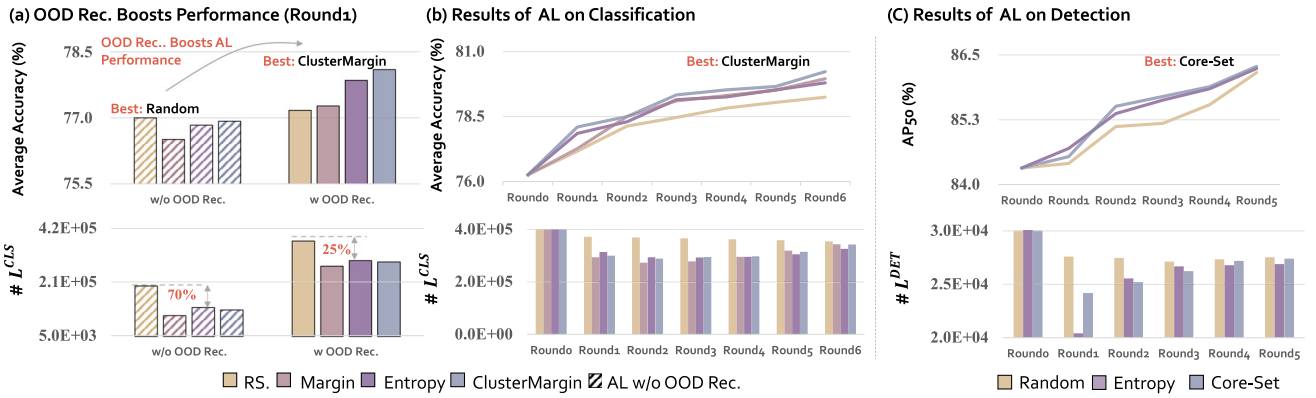

Fig. 6 The study of active annotation in Bamboo. **a** current AL methods struggle in realistic scenarios. Random sampling achieves better performance than each AL method. *OOD Rectification* boosts all AL methods to outperform random sampling. AL methods are still more helpful for model training with less valid data. It implies that the valid data that AL methods selected are much more informative. **b**, **c** in both classification and detection tasks, AL methods (ClusterMargin and Core-Set) that consider both the uncertainty and diversity select the most valuable data for model training. $\mathcal{L}^{\text{Cls}}$ refers annotated valid data from a given AL batch. Average accuracy denotes the average performance of models on the downstream datasets

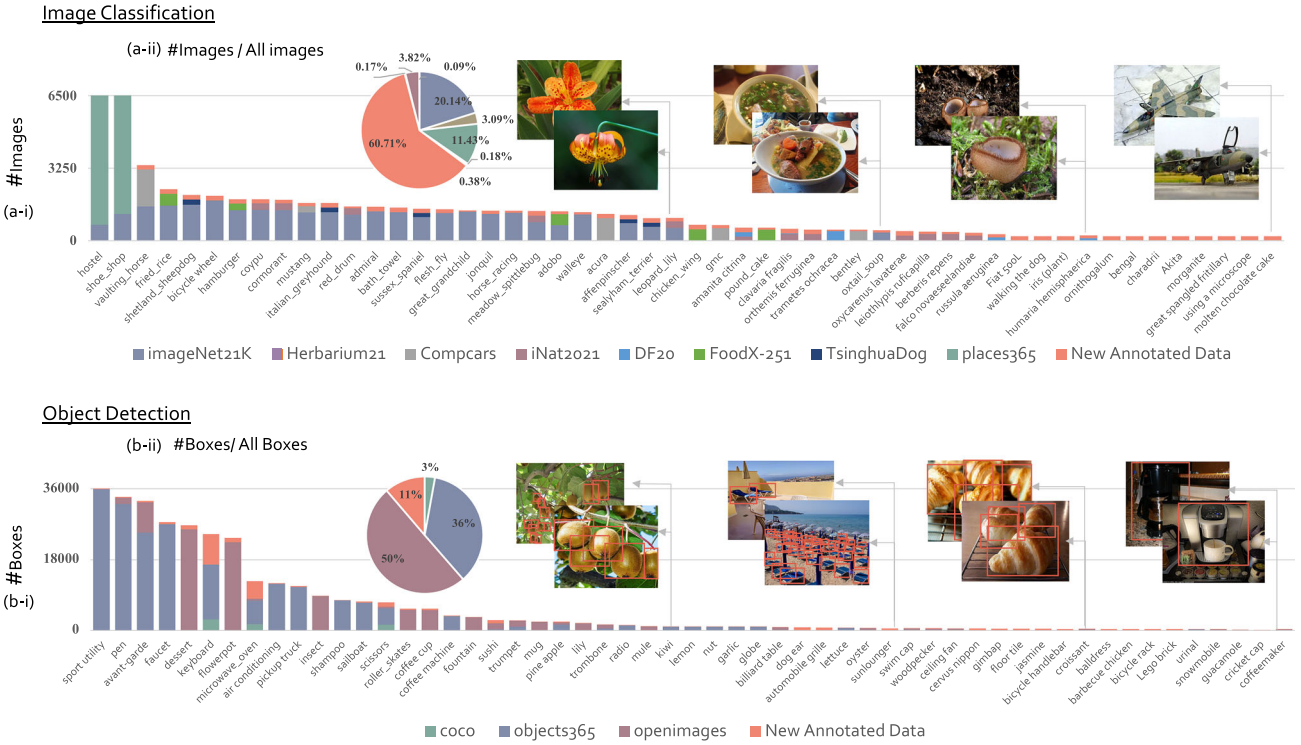

Fig. 7 Sorted distribution of image number per category in the Bamboo. **a-i** Bamboo-CLS contains 68,884,828 images spread across 119,035 categories. Category names are shown for every 250 intervals. Bamboo-CLS includes some fine-grained concepts that not be included in the current public datasets, such as *Folland Midge*. **a-ii** The new classification annotated data accounts for 60.71% of images in Bamboo. **b-i** Bamboo-DET contains 3,104,012 images across 809 categories. Category names are shown for every 16 intervals. **b-ii** The new detection annotated data accounts for 11% of images in Bamboo

4.2 Studies on Active Annotation

In academic active learning (AL) works Citovsky et al. (2021), Huang et al. (2021), researchers conduct data sampling on the leave-out "unlabeled" data pool that are separated from a curated dataset, e.g. ImageNet (Citovsky et al., 2021) and CIFAR10 (Roth & Small, 2006). All the data in this "unlabeled" data pool is strictly valid.[3] However, in realistic annotation scenarios, the real unlabeled data pool is composed of valid data and invalid data that is mostly out-of-distribution data, as shown in Fig. 4a. Therefore, can AL methods are effective when the invalid data is in the unlabeled data pool is an open question. And we found that:

Current Active Learning Methods are Ineffective for Sampling Valuable Data in the Real unlabeled data pool.

As shown in Fig. 6a, we illustrate the number of $\mathcal{L}_1$. We observe that AL sampling would retain fewer data in $\mathcal{L}_1$ than random sampling. For example, Entropy Sampling selects 70% less data than random sampling, resulting in worse downstream performance.

The Devils are in Uncertainty Modeling As discussed in Kendall and Gal (2017), D'souza et al. (2021), there are mainly two types of uncertainty for the deep models: *Aleatoric* and *Epistemic*. Both uncertainties are informative, but the aleatoric uncertainty is the out-of-distribution data, and the epistemic uncertainty is the in-distribution data. Considering $\mathcal{U}_0$ where aleatoric-uncertain data, epistemic-uncertain data, and other less-informative data exist, when $\mathcal{P}(1) \leftarrow \mathcal{U}(0)$, $\mathcal{P}(1)$ under AL sampling would have more aleatoric-uncertain data than that under random sampling, as AL methods tend to select uncertain data. Eventually, $\mathcal{L}(1)$ under AL sampling should has less data than that under random sampling as aleatoric uncertain data is invalid for annotators. We illustrate this phenomenon in Fig. 5 left. As shown in Fig. 6a, with much less $\mathcal{L}(1)$, AL methods' performances are hence worse than RS.

OOD Rec. Boosts AL Performance When $\mathcal{P}_r \leftarrow$ Rectifying $\mathcal{U}_{r-1}$ using f_{r-1} (our active learning framework), our proposed OOD rectification filters out the aleatoric uncertain data in $\mathcal{U}(0)$. Therefore, $\mathcal{P}(1)$ is only comprised of epistemic-uncertain data—which is informative—and other less-informative data. Since AL methods would select more epistemic uncertain data in $\mathcal{U}(1)$ than random sampling, they eventually perform better. We illustrate how OOD rectification helps active learning performs better in realistic scenarios in Fig. 5 right. As shown in Fig. 6(b,c), with OOD rectification, in both classification and detection tasks, AL methods (ClusterMargin and Core-Set) that consider both the uncertainty and diversity select the most valuable data for model training.

[3] Annotator had deleted invalid data as dataset established.

5 Dataset Statistics

As shown in Fig. 7, we illustrate the sorted distribution of image numbers per category in the Bamboo. Generally, we emphasize that the new annotated data in the Bamboo-CLS and Bamboo-DET are a powerful complement to the current public datasets—This data mostly belongs to tail classes of public datasets and new classes. In the following, we briefly describe the data statistics of Bamboo.

Image Classification (Bamboo-CLS) Bamboo-CLS has 68,884,828 images spread across 119,035 categories. 42,648,217 out of 68,884,828 images are newly annotated, which is twice of ImageNet22K. In addition, 20,000 out of 119,035 categories are from Wikidata. These categories mainly are fine-grained concepts, such as *Folland Midge* (one type of fighter) and *hemaria hemishphaerica* (a species of fungi). To our knowledge, Bamboo-CLS is the largest clean image dataset available to the vision research community, in terms of the total number of images and categories.

Object Detection (Bamboo-DET) Bamboo-DET has 3,104,012 images across 809 categories. Specifically, 557,457 images are newly annotated, and 150 concepts are from the Wikidata. In addition, we provide the statistics on instances per image of Bamboo-DET. As shown in Fig. 8, Our newly annotated data has 8.3 instances (on average) per image, which is dense than existing datasets, i.e. MS-COCO, Object-365, and OpenImages.

6 Experiments

6.1 Experimental Setups

6.1.1 Upstream Pre-training

Hyper-parameter We train the models on 64 A100 GPUs for image classification, with a total batch size of 8192. We employ an AdamW (Loshchilov & Hutter, 2017) optimizer of 30 epochs using a cosine decay scheduler with two epochs of linear warm-up. The ResNet-50 backbone is initialized from the model officially offered by PyTorch. The ViT B/16 backbone is initialized from ImageNet1K model provided by timm.[4] The weight decay, and warm-up learning rate are 2×10^{-8}, 10^{-6}, and 2×10^{-2}.

Datasets Beyond the new annotated data, we include ImageNet22K (Deng et al., 2009), INaturalist2021 (Van Horn et al., 2018), Herbarium2021,[5] Danish Fungi 2020 (Picek et al., 2021), iWildCam2020 (Beery et al., 2021), TsinghuaDogs (Zou et al., 2020), Places (Zhou et al., 2017), FoodX-251 (Kaur et

[4] https://github.com/rwightman/pytorch-image-models/tree/master/timm.

[5] https://www.kaggle.com/c/herbarium-2021-fgvc8/overview

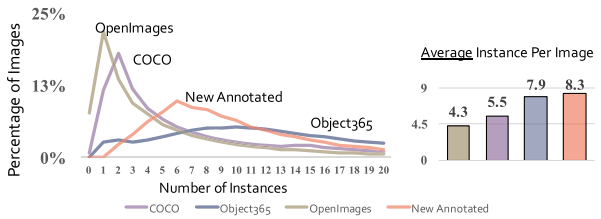

Fig. 8 The statistics of the number of bounding boxes per image. Quantitatively, our new annotated data has 8.3 instances (on average) per image, which is more dense compared with the other datasets like COCO and OpenImages

al., 2019), CompCars (Yang et al., 2015) in the upstream classification dataset training. We train the models on 48 A100 GPUs for detection, with a total batch size of 384, a total learning rate of 0.4, SGD optimizer of momentum 0.9, and a weight decay of 0.0001. We use the MultiStep learning rate scheduler with the decay rate of 0.1 on [16, 22] epochs and train for 26 epochs in total. We also applied the warm-up learning rate of 0.0004 for 1 epoch. We used Cross-Entropy-Loss for categorization and Smoothed-L1-Loss for bounding box regression. Beyond the new annotated data, we include COCO (Lin et al., 2014), Objects365 (Shao et al., 2019) and OpenImages (Kuznetsova et al., 2020) in the upstream object detection dataset training.

6.1.2 Downstream Evaluation

Datasets In the following sections, we adopt the downstream datasets that are widely used in the transfer learning study (Kornblith et al., 2019; Zhai et al., 2019; He et al., 2020). For models pre-trained on the image classification datasets, we use CIFAR10 (Krizhevsky et al., 2009), CIFAR100 (Krizhevsky et al., 2009), OxfordFlower (Nilsback & Zisserman, 2006), Food101 (Bossard et al., 2014), Caltech101 (Fei-Fei et al., 2004), OxfordPets (Parkhi et al., 2012), DTD (Cimpoi et al., 2014), StanfordCars (Krause et al., 2013), FGVC-Aircraft (Maji et al., 2013), SUN397 (Xiao et al., 2016), ImageNet1K (Russakovsky et al., 2015) as the downstream evaluation datasets. As for the object detection task, we select PASCAL VOC (Everingham et al., 2010) and CityPersons (Zhang et al., 2017) as the downstream evaluation datasets. These datasets cover a wide range of image domains. The number of images in each dataset ranges from 2000 to 80,000, and the number of classes in each dataset ranges from 10 to 8000.

Evaluation Protocol For the classification task, we use image features taken from the penultimate layer of each model, ignoring any classification layer provided. We train a logistic regression classifier for the linear probe evaluation setting. We finetune the entire model loaded with its backbone and FPN weights for the detection task. We only report the evaluation performance of models on downstream datasets. We finetune the model on 8 1080-Ti GPUs for detection, with the batch size of 16, SGD optimizer of momentum 0.9, and weight decay 0.0001 by loading the weights of backbone and FPN. We conduct a grid search on learning rate among $[5 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}]$. The learning rate is decayed by 0.1 at 16 and 18 and stopped training at 19 epochs.

6.2 Power of Bamboo as Pre-Training

6.2.1 Main Results

Information-Dense Annotations Matter As shown in Table 2, ResNet-50 (RN50) pre-trained on CLIP (400M) or IG-1B (1B) achieves better downstream task performance than BiT pre-trained on ImageNet1K (IN1K) (Russakovsky et al., 2015). However, compared to RN50 pre-trained on Bamboo, CLIP-RN50 or RN50 pre-trained on IG-1B achieves inferior performance.

It indicates that the amount of informative-dense annotations instead of the sheer number of annotations is much more essential for model pre-training. Compared to CLIP, which leverages the vast amount of image-text pairs on the web for pre-training, our Bamboo presents an active and continual framework that collects and annotates fully-supervised samples in a highly scalable manner.

Comprehensive Label System Helps As shown in Table 2, most methods pre-trained on IN1K, IG-1B, or WIT achieve more than 90% accuracy on the OxfordPets and OxfordFlower. But they only achieve less than 80% accuracy on the StanfordCars and FGVC-Aircraft. It indicates that these pre-trained datasets might include more semantic concepts related to OxfordPets and OxfordFlower. Our BambooTX spreads a large spectrum of concepts. Notably, it includes much more concepts that are neglected in the current public and nonpublic datasets. As a result, models pre-trained on Bamboo achieve much better performance than other methods. Beyond general object detection, it is also important to validate the generalization ability on specific object detection problems like pedestrian detection.

Bamboo is an Effective Pre-Training Source Compared to other methods, Bamboo achieves the best performance among downstream tasks on average. As shown in Table 2, ViT B/16 pretrained on Bamboo outperforms CLIP with 6.2 points gain. It indicates that our annotation is much more informative and hence more helpful for the model pre-training. In addition, Table 3 presents that ResNet-50 with FPN pretrained on Bamboo outperforms Objects365 with 1.1 points gain on PASCAL VOC and 2.1 points gain on CityPersons.

Table 1 Summary of Bamboo

Datasets	Concepts	Images	Boxes	Anno
YFCC-100M (Thomee et al., 2016)	–	100M	–	No
ImageNet22K (Deng et al., 2009)	22K	14M	–	Yes
Bamboo-CLS	**119K**	**69M**	–	Yes
COCO (Lin et al., 2014)	80	118K	1M	Yes
Objects365 (Shao et al., 2019)	365	609K	10M	Yes
OpenImages (Kuznetsova et al., 2020)	600	2M	14M	Partial
Bamboo-DET	**809**	**3M**	**27M**	Yes

Bamboo is the largest fully annotated vision dataset available to the general research community, in terms of the total number of images, the number of concepts, and the number of bounding boxes (for object detection task)

6.2.2 Further Analysis

The Influence of Similar Semantic Proposals The total annotation cost for the object detection task depends on the number of proposals. Images with dense proposals are more expensive than sparse ones. Based on our observation, many proposals with similar semantics tend to form a group in a single image. To evaluate their effectiveness, we conduct the following experiments on Objects365 (Shao et al., 2019) dataset.

Firstly, we define an image as a crowded image if it contains at least one category with more than 15 proposals. By removing all 27K crowded images from the full Objects365 dataset, we denote the remaining part as Objects365-sparse. Keeping the number of proposals the same as Objects365-sparse, we randomly removed 90K images from the full Objects365 dataset and marked the remaining part as Objects365-random. Furthermore, keeping the total object amount the same as Objects365-sparse, we randomly removed 101K non-crowded images from the full Objects365 dataset and denoted the remaining part as Objects365-dense.

Given the same annotation budget, we find that choosing to label non-crowded images yields better results for pre-training performance. Therefore, as mentioned in Sect. 3.3.2 of the main paper, we filter out covariate shift data in the OOD rectification step.

Finetuning Transfer We compared our model pre-trained on Bamboo to various with the ResNet-50 backbone. We present the finetuning transfer performance of the models pre-trained on Bamboo. The finetuning strategy among each downstream task is followed by the SimCLR (Chen et al., 2020a). Table 5 shows the comparison. Bamboo model achieves a 1.3% average accuracy gain compared to BiT-M pre-trained on the current largest public classification dataset: ImageNet22K. It indicates a larger, carefully annotated dataset can continually improve the performance of models. Besides, Bamboo model achieves a 0.5% average accuracy gain compared to SWSL, pre-trained on the IG-1B with 1B weakly supervised hashtags. Bamboo is 20 times smaller than IG-1B, which indicates that the amount of informative-dense annotations instead of the sheer number of weak annotations is much more essential for model pre-training.

Few-Shot Linear-probing Transfer We present the fero-shot transfer performance of the models pre-trained on Bamboo. We compared our model pre-trained on Bamboo to CLIP models with the same backbone.

Table 4 shows the comparison. We can indicate that Bamboo model conclusively outperforms CLIP model with the same backbone: RN50. Specifically, Bamboo model achieves a 6% average accuracy gain. On the FGVC-Aircraft, Bamboo model achieves 87.2%, despite having never seen any training images from this dataset. Bamboo includes all the concepts in the downstream tasks. However, we conduct data overlap analysis of Bamboo in Sect. 7, ensuring Bamboo rarely includes downstream data.

Robustness to Natural Distribution Shift We conduct experiments on the ObjectNet (Barbu et al., 2019) to compare Bamboo models with other models when evaluated on the data with controls for rotation, background, and viewpoint. ObjectNet is a dataset collected in the real world, where multiple objects are always present. There are 313 object classes in total, with 113 overlapping with ImageNet1K. We follow the literature (Kolesnikov et al., 2020; Radford et al., 2021) and evaluate our models on those 113 classes.

As shown above, we compare Bamboo models with the state-of-the-art model with the same backbone. Specifically, ResNet-50 pre-trained on Bamboo achieves 1.2% gains compared with ResNet-50 pre-trained on JFT-300M. ViT B/16 pre-trained on Bamboo achieves 3.2% gains compared with ViT B/16 pre-trained on Anno-1.3B. Even though JFT-300M and Anno-1.3B are much larger than Bamboo, the informative data in Bamboo is more helpful for pre-trained models in real scenarios.

Table 2 Downstream classification tasks performance among different pre-training methods

Method	Data	Annotation	Model	Paradigm	CIFAR10	CIFAR100	Food101	Pets	Flowers	SUN397	Cars	DTD	Caltech101	Aircraft	IN1K	AVG↑
SwAV (Caron et al., 2020)	IN1K	1.2M	RN50	Self	92.5	76.6	76.4	88.0	93.0	65.5	60.5	78.1	91.0	56.0	66.9	76.8
DINO (Caron et al., 2021)	IN1K	1.2M	RN50	Self	93.7	79.2	77.2	89.2	96.2	66.0	68.3	77.6	92.3	63.1	83.3	79.8
SWSL (Yalniz et al., 2019)	IG-1B	1B	RN50	Semi	94.7	79.5	79.1	94.4	94.6	67.8	65.9	77.8	96.1	58.4	81.2	80.9
WSL (Mahajan et al., 2018)	IG-1B	1B	RX101	Weak	95.0	78.2	83.5	**95.5**	90.8	67.9	72.3	75.3	93.3	53.9	83.3	81.0
CLIP (Radford et al., 2021)	*WIT*	*400M*	*RN50*	*Lang.*	*88.7*	*70.3*	*86.4*	*88.2*	*96.1*	*73.3*	*78.3*	*76.4*	*89.6*	*49.1*	*73.3*	*79.1*
CLIP (Radford et al., 2021)	*WIT*	*400M*	*B/16*	*Lang.*	*96.2*	*83.1*	*92.8*	*93.1*	*98.1*	*78.4*	*86.7*	*79.2*	*94.7*	*59.5*	*80.2*	*85.6*
BiT (Kolesnikov et al., 2020)	IN1K	1.2M	RN50	Sup	91.7	74.8	72.5	92.3	92.0	61.1	53.5	72.4	91.2	52.5	75.2	73.6
BiT (Kolesnikov et al., 2020)	IN22K	14M	RN50	Sup	94.9	82.2	83.3	91.5	99.4	69.9	59.0	77.3	93.9	55.6	76.7	80.3
RN50	Bamboo	69M	RN50	Sup	93.9	81.2	85.3	92.0	99.4	72.2	91.1	76.5	93.2	84.0	77.2	86.0
B/16	Bamboo	69M	B/16	Sup	**98.2**	**90.2**	**92.9**	95.1	**99.8**	**79.0**	**93.3**	**81.2**	**97.0**	**88.1**	**83.6**	**91.8**

Bamboo achieves the state-of-the-arts linear probe performance on the downstream tasks. Lang. indicates image-text pair. Bamboo here refers to the Bamboo-CLS. Pets indicates OxfordPets. Flowers indicates OxfordFlower. Cars indicates StanfordCars. Aircraft indicates FGVC-Aircraft. IN1K indicates ImageNet1K. Results reported by the author are marked in italic. We mainly compare with the methods conducted on supervised learning. Other performance of current methods are also presented

Table 3 Comparisons of downstream detection tasks performance

Data	Anno.	VOC AP50 ↑	CITY MR ↓	COCO mmAP ↑
COCO (Shao et al., 2019)	1M	85.1	16.2	–
OpenImages (Shao et al., 2019)	14M	82.4	16.8	37.4
Objects365	10M	86.4	14.7	39.3
Bamboo	27M	**87.5**	**12.6**	**43.9**

Pre-trained model on Bamboo achieves significant performance gain. Bamboo here refers to the Bamboo-DET. VOC means the PASCAL VOC dataset (Everingham et al., 2010). CITY. means the CityPersons dataset (Zhang et al., 2017)

7 Social Impact

The proposed Bamboo dataset and pre-training model shows the capacity and generalization of learned image representation which could benefit many applications of computer vision. However, our data usage might bring several risks, such as data overlapping, privacy, and inappropriate content. We discuss these risks and their mitigation strategies as follows.

Data Overlapping A concern with pre-training on an extensive dataset is unintentional overlap with downstream evaluation (Radford et al., 2021). To enable a meaningful test of generalization, we identify and remove all duplicates among upstream data. Specifically, we utilize Difference Hash (DHash) (Ben, 2017) to present the information of each image. We calculate the hash-code of each downstream image and each crawled image, and two images with the same hash-code are regarded as similar ones. Then, we filter out the crawled images that are similar to downstream images. Based on the above method, we discard 122,939 images for classification and 1046 images for detection from the unlabeled pool.

Copyright We crawl only the data under the Creative Commons license (CC-BY) for the Bamboo-DET. This license allows free use, redistribution, and adaptation for non-commercial purposes. For the Bamboo-CLS data, 30% of data is under the CC-BY license because of its large volume of data. For Bamboo-CLS data that is not under the CC-BY license, referred to LAION-400M (Schuhmann et al., 2021) and Conceptual 12M (Changpinyo et al., 2021), we only present the lists of URLs to this data without redistribution. We build the meta file as follow.

[image_url] [class_index]

Referred to *Authors Guild, Inc. v. Google Inc.* (Campbell, 2016), training data on the copyrighted works might be considered as transformative uses and was thus might be regarded as *Fair Use*.[6] In addition, referred to *Article 30-4 of the new Copyright Act* (act, 2010), there are no restrictions on the subject, purpose, and method of data analysis, and there is no obligation to compensate the copyright holder. However, we admit that using copyright material as training data is still a controversial issue in Artificial Intelligence, and we would no doubt follow the newest law worldwide. Bamboo is specifically open for non-commercial research and/or educational purposes to respect the copyright law. For researchers and educators who wish to use copyrighted images for that purpose, training or benchmarking models with copyrighted works would be qualified as *transformative* uses and thus not infringe copyright law in the U.S. Nevertheless, the users must strictly follow the Flickr Terms of Use.[7] And the users of these images accept full responsibility for the use of the image.

Problematic Content The inappropriate contents such as drugs, nudity, and other offensive content exist in the web data. we ask annotators to discard such images instead of conducting annotation.

Privacy To mitigate privacy issues with public visual datasets, researchers have attempted to obfuscate private information before publishing the data (Frome et al., 2009; Yang et al., 2021). We plan to follow this line of work to blur faces, and license plates in our new annotated data. In addition, if the original picture found at the URL present on the Bamboo on the record states users' names, phone numbers, or any personal information, users can request a takedown of this image.

Bias The images were crawled from Flickr, thus inheriting all the biases of that website. The usage of user-generated data might bring the risk of bias. We plan to tackle this problem by balancing various categories.

8 Conclusion

In our work, with a human–machine synergy, we actively and continually build a mega-scale and information-dense dataset, namely Bamboo. Bamboo is the largest clean image dataset available to the vision research community,

[6] https://www.copyright.gov/fair-use/index.html.

[7] https://www.flickr.com/help/terms/api.

Table 4 Comparisons of few-shot downstream classification tasks performance among different pre-training methods

Method	Data	Annotation	Model	Paradigm	CIFAR10	CIFAR100	Food101	Pets	Flowers	SUN397	Cars	DTD	Caltech101	Aircraft	IN1K	AVG↑
CLIP (Radford et al., 2021)	*WIT*	*400M*	*RN50*	*Lang.*	*91.6*	*68.7*	*89.2*	*88.9*	*70.4*	*65.2*	*65.6*	*46*	*89.3*	*27.1*	*68.6*	*70.0*
RN50	Bamboo	69M	RN50	Sup	93.8	67.7	81.6	74.3	87.3	58.7	63.0	51.1	88.4	87.2	82.5	76.0

Bamboo achieves the state-of-the-arts linear probe performance on the downstream tasks. Lang. indicates image-text pair. Bamboo here refers to the Bamboo-CLS. Pets indicates OxfordPets. Flowers indicates OxfordFlower. Cars indicates StanfordCars. Aircraft indicates FGVC-Aircraft. IN1K indicates ImageNet1K. Results reported by the author are marked in italic. We mainly compare with the methods conducted on supervised learning. Other performance of current methods are also presented

Table 5 Comparisons of fine-tuning downstream classification tasks performance among different pre-training methods

Method	Data	Annotation	Model	Paradigm	CIFAR10	CIFAR100	Food101	Pets	Flowers	SUN397	Cars	DTD	Caltech101	Aircraft	IN1K	AVG↑
DINO	IN1K	1.2M	RN50	Self	97.1	84.0	86.3	90.0	96.1	65.2	84.6	77.6	91.4	81.8	66.5	83.7
SWAV	IN1K	1.2M	RN50	Self	97.2	84.2	86.0	90.3	95.7	64.4	83.9	77.2	91.7	81.2	66.9	83.5
SWSL	IG-1B	1B	RN50	Semi	97.0	86.5	87.3	94.4	97.0	66.0	88.5	78.3	93.8	84.0	81.7	86.8
BiT-S	IN1K	1.2M	RN50	Sup	97.0	85.0	85.7	92.8	95.0	60.3	87.5	74.7	92.0	83.8	75.2	84.5
BiT-M	IN22K	14M	RN50	Sup	97.6	86.2	87.9	91.5	98.1	64.2	88.2	78.4	92.9	84.3	76.7	86.0
RN50	Bamboo	69M	RN50	Sup	97.3	87.0	87.5	92.0	99.4	72.2	91.4	77.1	93.9	85.9	77.1	87.3

Bamboo achieves the state-of-the-arts fine-tuning performance on the downstream tasks

Dataset	Images	Proposals	VOC AP50 ↑
Objects365-sparse	581K	8.2M	86.3
Objects365-random	519K	8.2M	85.8
Objects365-dense	508K	8.2M	85.1

Method	Data	Model	Para.	ObjectNet ↑
BiT-L (Kolesnikov et al., 2020)	JFT-300M	RN50	Weak.	37.6
ANN-1.3B (Beal et al., 2021)	ANN-1.3B	B/16	Weak.	50.7
RN50	Bamboo	RN50	Sup.	**38.8**
B/16	Bamboo	B/16	Sup.	**53.9**

in terms of the total number of images and the number of categories, for classification and detection tasks. Our key insight is that a unified and visually-oriented label system is crucial for model pre-training, and rectifying OOD samples is indispensable for AL to function in realistic scenarios. We have demonstrated the effectiveness of Bamboo as a better pre-training dataset for various downstream tasks and provided several valuable observations.

Acknowledgements This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20221-0012, MOE-T2EP20223-0002), and under the RIE2020 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

Data Availability All data used in this paper are publicly available and can be accessed through this Github repository: https://github.com/ZhangYuanhan-AI/Bamboo.

References

The act partially amending the copyright act (2010) (act no. 52 of 2021; enacted May 2021).

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., & Katz, B. (2019). ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models.

Beal, J., Wu, H.-Y., Park, D. H., Zhai, A., & Kislyuk, D. (2021). Billion-scale pretraining with vision transformers for multi-task visual representations. arXiv preprint arXiv:2108.05887

Beery, S., Agarwal, A., Cole, E., & Birodkar, V. (2021). The iwildcam 2021 competition dataset. arXiv preprint arXiv:2105.03494

Ben, H. (2017). Duplicate image detection with perceptual hashing in python. https://benhoyt.com/writings/duplicate-image-detection/#difference-hash-dhash

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et a. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258

Bossard, L., Guillaumin, M., & Van Gool, L. (2014). Food-101—Mining discriminative components with random forests. In *ECCV*.

Campbell, V. (2016). Authors Guild v. Google, Inc. *DePaul Journal of Art, Technology & Intellectual Property Law, 27*, 59.

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *ECCV* (pp. 132–149).

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294

Changpinyo, S., Sharma, P., Ding, N., & Soricut, R. (2021). Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *ICML* (pp. 1597–1607). PMLR.

Chen, X., Fan, H., Girshick, R., & He, K. (2020). Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *CVPR* (pp. 51–1258).

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S. & Vedaldi, A. (2014). Describing textures in the wild. In *CVPR*.

Citovsky, G., DeSalvo, G., Gentile, C., Karydas, L., Rajagopalan, A., Rostamizadeh, A., & Kumar, S. (2021). Batch active learning at scale. arXiv preprint arXiv:2107.14263

Coates, A., Ng, A., & Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 215–223). JMLR workshop and conference proceedings.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *CVPR* (pp. 248–255). IEEE.

Dosovitskiy, A., Fischer, P., Springenberg, J. T., Riedmiller, M., & Brox, T. (2015). Discriminative unsupervised feature learning with exemplar convolutional neural networks. *TPAMI, 38*(9), 1734–1747.

D'souza, D., Nussbaum, Z., Agarwal, C., & Hooker, S. (2021). A tale of two long tails. arXiv preprint arXiv:2107.13098

Estimation Lemma. (2010). Estimation Lemma—Wikipedia, the free encyclopedia. [Online; accessed September 29, 2012].

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The Pascal visual object classes (VOC) challenge. *IJCV, 88*(2), 303–338.

Fabian, M. S., Gjergji, K., Gerhard, W., et al.. (2007). Yago: A core of semantic knowledge unifying wordnet and Wikipedia. In *WWW* (pp. 697–706).

Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *CVPR workshop* (p. 178). IEEE.

Frome, A., Cheung, G., Abdulkader, A., Zennaro, M., Wu, B., Bissacco, A., Adam, H., Neven, H., & Vincent, L. (2009). Large-scale privacy protection in google street view. In *ICCV* (pp. 2373–2380). IEEE.

Gal, Y. (2016). Uncertainty in deep learning.

Ghiasi, G., Zoph, B., Cubuk, E. D., Le, Q. V., & Lin, T.-Y. (2021). Multi-task self-training for learning general representations. arXiv preprint arXiv:2108.11353

Gilad-Bachrach, R., Navot, A., & Tishby, N. (2005). Query by committee made real. In *Advances in neural information processing systems*, 18.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *CVPR* (pp. 9729–9738).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR* (pp. 770–778).

Helber, P., Bischke, B., Dengel, A., & Borth, D. (2019). EuroSat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12*(7), 2217–2226.

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (to appear).

Huang, S., Wang, T., Xiong, H., Huan, J., & Dou, D. (2021). Semi-supervised active learning with temporal output discrepancy. In *ICCV* (pp. 3447–3456).

Iglesias, J. E., Konukoglu, E., Montillo, A., Tu, Z., & Criminisi, A. (2011). Combining generative and discriminative models for semantic segmentation of CT scans via active learning. In *Biennial international conference on information processing in medical imaging* (pp. 25–36). Springer.

Joshi, A. J., Porikli, F., & Papanikolopoulos, N. (2009). Multi-class active learning for image classification. In *CVPR* (pp. 2372–2379). IEEE.

Kaur, P., Sikka, K., Wang, W., Belongie, S., & Divakaran, A. (2019) Foodx-251: A dataset for fine-grained food classification. arXiv preprint arXiv:1907.06167

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? arXiv preprint arXiv:1703.04977

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., & Houlsby, N. (2020). Big transfer (bit): General visual representation learning. In *ECCV* (pp. 491–507). Springer.

Kornblith, S., Shlens, J., & Le, Q. V. (2019). Do better ImageNet models transfer better? In *CVPR* (pp. 2661–2671).

Krause, J., Stark, M., Deng, J., & Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al. (2020). The open images dataset v4. *IJCV, 128*(7), 1956–1981.

Larsson, G., Maire, M., & Shakhnarovich, G. (2016). Learning representations for automatic colorization. In *ECCV* (pp. 577–593). Springer.

Li, J., Zhou, P., Xiong, C., Socher, R., & Hoi, S. C. H. (2020). Prototypical contrastive learning of unsupervised representations. arXiv preprint arXiv:2005.04966

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *ECCV* (pp. 740–755). Springer.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., & Van Der Maaten, L. (2018). Exploring the limits of weakly supervised pretraining. In *ECCV* (pp. 181–196).

Maji, S., Kannala, J., Rahtu, E., Blaschko, M., & Vedaldi, A. (2013). Fine-grained visual classification of aircraft. Technical report.

Miller, G. A. (1998). *WordNet: An electronic lexical database*. MIT Press.

Misra, I., & van der Maaten, L. (2020). Self-supervised learning of pretext-invariant representations. In *CVPR* (pp. 6707–6717).

Nilsback, M.-E., & Zisserman, A. (2006). A visual vocabulary for flower classification. In *CVPR* (Vol. 2, pp. 1447–1454). IEEE.

Parkhi, O. M., Vedaldi, A., Zisserman, A., & Jawahar, C. V. (2012). Cats and dogs. In *CVPR* (pp. 3498–3505). IEEE.

Picek, L., Šulc, M., Matas, J., Heilmann-Clausen, J., Jeppesen, T. S., Læssøe, T., & Frøslev, T. (2021). Danish fungi 2020–Not just another image recognition dataset. arXiv preprint arXiv:2103.10107

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020

Roth, D., & Small, K. (2006). Margin-based active learning for structured output spaces. In *European conference on machine learning* (pp. 413–424). Springer.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). ImageNet large scale visual recognition challenge. *IJCV, 115*(3), 211–252.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., & Komatsuzaki, A. (2021). Laion-400m: Open dataset of clip-filtered 400 million image-text pairs.

Sener, O., & Savarese, S. (2018). Active learning for convolutional neural networks: A core-set approach. In *ICLR*. OpenReview.net.

Settles, B. (2009). Active learning literature survey.

Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., & Sun, J. (2019). Objects365: A large-scale, high-quality dataset for object detection. In *ICCV* (pp. 8430–8439).

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV* (pp. 843–852).

Sznitman, R., & Jedynak, B. (2010). Active testing for face detection and localization. *TPAMI, 32*(10), 1914–1920.

Tanon, T. P., Weikum, G., & Suchanek, F. M. (2020). Yago 4: A reasonable knowledge base. *The Semantic Web, 12123*, 583–596.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., & Li, L.-J. (2016). Yfcc100m: The new data in multimedia research. *Communications of the ACM, 59*(2), 64–73.

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research, 2*(Nov), 45–66.

Van Horn, G., Aodha, O. M., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., & Belongie, S. (2018). The inaturalist species classification and detection dataset. In *CVPR* (pp. 8769–8778).

Vezhnevets, A., Ferrari, V., & Buhmann, J. M. (2012). Weakly supervised structured output learning for semantic segmentation. In *CVPR* (pp. 845–852). IEEE.

Vrandečić, D., & Krötzsch, M. (2014). WikiData: A free collaborative knowledgebase. *Communications of the ACM, 57*(10), 78–85.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., & Perona, P. (2010). Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology.

Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *CVPR* (pp. 3733–3742).

Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., & Lu, X. (2017). Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing, 55*(7), 3965–3981.

Xiao, J., Ehinger, K. A., Hays, J., Torralba, A., & Oliva, A. (2016). Sun database: Exploring a large collection of scene categories. *IJCV, 119*(1), 3–22.

Xie, Q., Luong, M.-T., Hovy, E., & Le, Q. V. (2020). Self-training with noisy student improves ImageNet classification. In *CVPR* (pp. 10687–10698).

Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M., & Mahajan, D. (2019). Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546.

Yang, K., Qinami, K., Fei-Fei, L., Deng, J., & Russakovsky, O. (2020). Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 547–558).

Yang, K., Yau, J., Fei-Fei, L., Deng, J., & Russakovsky, O. (2021). A study of face obfuscation in ImageNet. arXiv preprint arXiv:2103.06191

Yang, L., Luo, P., Loy, C. C., & Tang, X. (2015). A large-scale car dataset for fine-grained categorization and verification. In *ICCV* (pp. 3973–3981).

Yang, Y., Ma, Z., Nie, F., Chang, X., & Hauptmann, A. G. (2015). Multi-class active learning by uncertainty sampling with diversity maximization. *IJCV, 113*(2), 113–127.

Yun, S., Oh, S. J., Heo, B., Han, D., Choe, J., & Chun, S. (2021). Re-labeling ImageNet: From single to multi-labels, from global to localized labels. arXiv preprint arXiv:2101.05022

Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A. S., Neumann, M., Dosovitskiy, A., et al. (2019). A large-scale study of representation learning with the visual task adaptation benchmark. arXiv preprint arXiv:1910.04867

Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In *ECCV* (pp. 649–666). Springer.

Zhang, S., Benenson, R., & Schiele, B. (2017). Citypersons: A diverse dataset for pedestrian detection. In *CVPR*, July 2017.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. In *TPAMI*.

Zoph, B., Ghiasi, G., Lin, T.-Y., Cui, Y., Liu, H., Cubuk, E. D., & Le, Q. V. (2020). Rethinking pre-training and self-training. arXiv preprint arXiv:2006.06882

Zou, D.-N., Zhang, S.-H., Mu, T.-J., & Zhang, M. (2020). A new dataset of dog breed images and a benchmark for fine-grained classification. *Computational Visual Media*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.