

# FungiTastic: A Multi-Modal Dataset and Benchmark for Image Categorization

Lukas Picek<sup>✉</sup>, Klára Janoušková<sup>✉</sup>, Vojtech Cermak<sup>✉</sup>, and Jiri Matas<sup>✉</sup>

<sup>✉</sup>University of West Bohemia & Inria, <sup>✉</sup>CTU in Prague

lukaspicek@gmail.com, {janoukl1, cermavo3, matas}@fel.cvut.cz



Figure 1. A **FungiTastic observation** includes one or more photos of an *observed* specimen with expert-verified taxon labels (some DNA sequenced) and occasionally also a microscopic image of its spores. Textual captions, observation metadata, geospatial data, and climatic time-series data are available for virtually all observations. For a subset ( $\sim 70k$  photos), we provide body part segmentation masks.

## Abstract

We introduce a new, challenging benchmark and a dataset, *FungiTastic*, based on fungal records continuously collected over a twenty-year span. The dataset is labelled and curated by experts and consists of about 350k multimodal observations of 6k fine-grained categories (species). The fungi observations include photographs and additional data, e.g., meteorological and climatic data, satellite images, and body part segmentation masks. *FungiTastic* is one of the few benchmarks that include a test set with DNA-sequenced ground truth of unprecedented label reliability. The benchmark is designed to support (i) standard closed-set classification, (ii) open-set classification, (iii) multi-modal classification, (iv) few-shot learning, (v) domain shift, and many more. We provide tailored baselines for many use cases, a multitude of ready-to-use pre-trained models on *HuggingFace*, and a framework for model training. The documentation and the baselines are available at [GitHub](#) and [Kaggle](#).

## 1. Introduction

Biological problems provide a natural, challenging setting for benchmarking image classification methods [49, 56, 65, 66]. Consider the following aspects inherently present in biological data. The species distribution is typically seasonal and constantly evolving under the influence of external

factors such as precipitation levels, temperature, and loss of habitat, exhibiting constant *domain shifts*. Species categorization is fine-grained, with high intra-class and low inter-class variance. The distribution is often long-tailed; only a few samples are available for rare species (*few-shot learning*). New species are being discovered, raising the need for the “unknown” class option (i.e., *open-set recognition*). Commonly, the set of classes has a hierarchical structure, and different misclassifications may have very different costs (i.e., *non-standard losses*). Think of mistaking a poisonous mushroom for an edible, potentially lethal, and an edible mushroom for a poisonous one, which at worst means an empty basket. Similarly, needlessly administering anti-venom after making a wrong decision about a harmless snake bite may be unpleasant, but its consequences are incomparable to not acting after a venomous bite.

Common benchmarks [15, 43, 65, 68] generate independent and identically distributed (i.i.d.) data by shuffling and randomly splitting it for training and evaluation. In real-world applications, i.i.d data are rare since training data are collected well before deployment and everything changes over time [70]. Moreover, they fail to address the above-mentioned aspects important in many instance of ML system deployment: robustness to distribution and domain shifts, ability to detect classes not represented in the training set, limited training data, and dealing with non-standard losses.

For benchmarking, it is crucial to ensure that methods are tested on data not indirectly “seen”, without know-

ing [24, 27], especially given the huge dataset used for training LLMs or VLMs and possibly covering the entirety of the internet at a certain point in time. Conveniently, many domains in nature are of interest to experts and the general public, who provide a continuous stream of new and annotated data [48, 61]. The public’s involvement introduces the problem of noisy training data; evaluating the robustness of this phenomenon is also of practical importance.

In the paper, we introduce **FungiTastic**, a multi-modal dataset and benchmark based on fungi observations<sup>1</sup>, which takes advantage of the favorable properties of natural data discussed above and shown in Figure 1. The fungi observations include photographs, satellite images, meteorological data, segmentation masks, textual captions, and location-related metadata. The *location metadata* enriches the observations with attributes such as the timestamp, GPS location, and information about the substrate and habitat.

By incorporating various modalities, the dataset supports a robust benchmark for multi-modal classification, enabling the development and evaluation of sophisticated machine-learning models under realistic and dynamic conditions.

**The key contributions of the FungiTastic benchmark are:**

- It addresses real-world challenges such as domain shifts, open-set problems, and few-shot classification, providing a realistic benchmark for developing robust ML models.
- The proposed benchmarks allow for addressing fundamental problems beyond standard image classification, such as novel-class discovery, few-shot classification, and evaluation with non-standard cost functions.
- It includes diverse data types, such as photographs, satellite images, bioclimatic time-series data, segmentation masks, contextual metadata (e.g., timestamp, camera metadata, location, substrate, and habitat), and image captions, offering a rich, multimodal benchmark.

## 2. Related Work

Classification of data originating in nature, including images of birds [6, 68], plants [21, 23], snakes [9, 47], fungi [49, 65], and insects [22, 44] has been widely used to benchmark machine learning algorithms, not just fine-grained visual categorization. The datasets were instrumental in focusing on fine-grained recognition and attracting attention to challenging natural problems.

However, the datasets are typically artificially sampled, solely image-based, and focused on traditional image classification. Most commonly used datasets are small by modern standards, with a limited number of categories, which restricts their usefulness for large-scale and highly diverse ap-

<sup>1</sup>A set of photographs and additional metadata describing one particular fungi specimen and surrounding environment. Usually, each photograph focuses on a different organ. For an example observation, see Figure 1

plications. Though performance being often saturated, reaching an accuracy of 85–95 % (rightmost column of Tab. 1), these datasets are still widely used in the community and have reached thousands of citations in the past few years. Many popular datasets also suffer from specific limitations that compromise their generalizability and robustness. Common issues include:

- **Lack of Multi-Modal Data:** Available datasets are predominantly image-based, with few offering auxiliary metadata like geographic or temporal context, which is essential for real-world applications where distribution changes and context is important.
- **Biases in Data Representation:** Many datasets exhibit regional and other biases [60], which can lead to biased models that do not perform well across different populations or environments. This lack of diversity can severely limit the usability of models trained on these datasets for global applications.
- **Single task focus:** While current ML applications require adaptability to tasks such as open-set classification, few-shot learning, and out-of-distribution detection, many of these datasets were not designed with these tasks in mind, limiting their usefulness for modern benchmarking.
- **Labeling Errors and Quality Control:** Label errors are prevalent in widely-used datasets [7, 64]. Mislabeling, especially in fine-grained categories, can reduce the reliability of these datasets as benchmarks and reduce the model’s ability to learn fine distinctions.

Table 1. **Resent and popular fine-grained classification datasets.** We list suitability for closed-set (C), open-set (OS), and few-shot (F) classification, segmentation (S), out-of-distribution (OOD) and multi-modal (M<sup>2</sup>) evaluation. Modalities, e.g., images (I), metadata (M), and masks (S), are available for training. The SOTA accuracy is limited to the classification task. For TaxaBench-8k, we report zero-shot performance.  $\forall = \{C, OS, FS, S, OOD, M^2\}$

Dataset	Classes	Images	Modals.			Tasks	SOTA <sup>†</sup>
			I	M	S		Accuracy
Oxford-Pets [46]	37	5k	✓	–	–	C	97.1 [19]
FGVC Aircraft [43]	102	10k	✓	–	–	C	95.4 [5]
Stanford Dogs [34]	120	20k	✓	–	–	C	97.3 [5]
Stanford Cars [36]	196	16k	✓	–	–	C	97.1 [38]
Species196 [26]	196	20k	✓	✓	–	C/M <sup>2</sup>	88.7 [26]
CUB-200-2011 [68]	200	12k	✓	✓	✓	C	93.1 [11]
NABirds [64]	555	49k	✓	–	–	C/F/M <sup>2</sup>	93.0 [16]
PlantNet300k [21]	1,081	275k	✓	–	–	C	92.4 [21]
DanishFungi2020 [49]	1,604	296k	✓	✓	–	C/M <sup>2</sup>	80.5 [49]
ImageNet-1k [15]	1,000	1.4m	✓	–	–	C/FS	92.4 [17]
TaxaBench-8k [56]	2225	9k	✓	✓	–	C/M <sup>2</sup>	37.5 [56]
iNaturalist [65]	5,089	675k	✓	–	–	C/FS	93.8 [58]
ImageNet-21k [52]	21,841	14m	✓	–	–	C/FS	88.3 [58]
Insect-1M [44]	34,212	1m	✓	✓	–	C/M <sup>2</sup>	–
(our) FungiTastic	2,829	620k	✓	✓	✓	$\forall$	75.3
(our) FungiTastic-Mini	215	68k	✓	✓	✓	$\forall$	74.8

### 3. The FungiTastic Benchmark

FungiTastic is built from fungi observations submitted to the Atlas of Danish Fungi before the end of 2023, which were labeled by taxon experts on a species level. In total, more than 350k observations consisting of 630k photographs collected over 20 years are used. Apart from the photographs, each observation includes additional observation data (see Figure 1) ranging from satellite images, meteorological data, and tabular metadata (e.g., timestamp, GPS location, and information about the substrate and habitat) to segmentation masks and toxicity status. The vast majority of observations got all of the attributes annotated. For details about the attribute description and its acquisition process, see Subsection 3.1. Since the data comes from a long-term conservation project, its seasonality and naturally shifting distribution make it suitable for time-based splitting. In this so-called **temporal division**, all data collected up to the end of 2021 is used for **training**, while data from 2022 and 2023 is reserved for **validation** and **testing**, respectively.

The FungiTastic benchmark is designed to go beyond standard closed-set classification and support a wide range of challenging machine learning tasks, including (i) open-set classification, (ii) few-shot learning, (iii) multi-modal learning, and (iv) domain shift evaluation. To facilitate these tasks, we provide several curated subsets, each tailored for specific experimental setups. A general overview of these subsets is provided below, with detailed statistics and further information in the Appendix (see Table 9).

**FungiTastic** is a general subset that includes around 346k observations of 4,507 species accompanied by a wide set of additional observation data. The FungiTastic has dedicated validation and test sets specifically designed for closed-set and open-set scenarios. While the closed-set validation and test sets only include species present in the training set, the open-set also includes observations with species observed only after 2022 (validation) and 2023 (test), i.e., species not available in the training set. All the species with no examples in the training set are labeled as "*unknown*". **Additionally, we include a DNA-based test set** of 725 species and 2,041 observations.

**FungiTastic-Mini (FungiTastic-M)** is a compact and challenging subset of the FungiTastic dataset designed primarily for prototyping and consisting of all observations belonging to 6 hand-picked genera (e.g., *Russula*, *Boletus*, *Amanita*, *Clitocybe*, *Agaricus*, and *Mycena*)<sup>2</sup>. This subset comprises 67,848 images (36,287 observations) of 253 species, greatly reducing the computational requirements for training. Exclusively, we include body part mask annotations.

<sup>2</sup>These genera produce fruiting bodies of the toadstool type, which include many visually similar species and are of significant interest to humans due to their common use in gastronomy.

**FungiTastic-FS** subset, FS for few-shot, is formed by species with less than 5 observations in the training set, which were removed from the main (FungiTastic) dataset. The subset contains 6,391 observations encompassing 12,015 images of a total of 2,427 species. As in the FungiTastic data, the split into validation and testing is done according to the year of acquisition.

#### 3.1. Additional Observation Data

This section provides an overview of the accompanying data available for virtually all user-submitted observations. For each type, we describe the data itself and, if needed, its acquisition process as well. Below, we describe (i) **tabular metadata**, which includes key environmental attributes and taxonomic information for nearly all observations, (ii) **remote sensing data** at fine-resolution geospatial scale for each observation site, (iii) **meteorological data**, which provides long-term climate variables, (iv) **body part segmentation masks** that delineate specific morphological features of fungi fruiting bodies, such as caps, gills, pores, rings, and stems, and (v) **image captions**. All that metadata is integral to advancing research combining visual, textual, environmental, and taxonomic information.

**Body part segmentation masks** of fungi fruiting bodies are essential for accurate identification and classification [13]. These morphological features provide crucial taxonomic information distinguishing some visually similar species. Therefore, we provide instance segmentation masks for all photographs in the FungiTastic-M. We consider various semantic categories such as *cap*, *gills*, *pores*, *ring*, *stem*, etc. These annotations (see Figure 2) are expected to drive advances in interpretable recognition methods [53] and evaluation [29], with masks also enabling instance segmentation for separate foreground and background modeling [8]. All segmentation mask annotations were semi-automatically generated in CVAT using the Segment Anything Model [35] and human supervision, i.e., annotators fixed all wrong masks.



Figure 2. **FungiTastic body part segmentation.** We consider five different categories, e.g., the cap, gills, stem, pores, and the ring.

**Multi-band remote sensing data** offer detailed and globally consistent environmental information at a fine resolution, making it a valuable resource for species categorization (i.e., identification) [54] and species distribution modeling [10, 50]. To allow testing the potential of such data and to facilitate easy use of geospatial data, we provide multi-band (e.g., R, G, B, NIR, elevation, and landcover) satellite patches with  $64 \times 64$  pixel resolution at 10m spatial resolution per pixel for (elevation and landcover are re-projected from 30m), centered on observation location. The data were extracted from rasters publicly available at [Ecodatacube](#), [ASTER](#), and [ESA WorldCover](#). The data are available in the form of torch tensors in a shape of  $[6 \times 64 \times 64]$ .

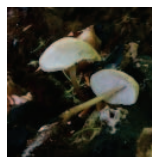


Figure 3. **Satellite RGB images** with  $64 \times 64$  resolution extracted from Sentinel-2A rasters available at [Ecodatacube](#).

**Meteorological data** and other climatic variables are vital assets for species identification and distribution modeling [4, 30]. In light of that, we provide 20 years of historical time-series monthly values of mean, min., and max. temperature and total precipitation for all observations (see Figure 8 in Appendix for example data). For each observation site, 20 years of data was extracted; for instance, an observation from 2000 includes data from 1980 to 2000. However, as the available climatic rasters only extend up to the year 2020, observations from 2020 to 2024 have missing values for those years not covered by existing data. In addition, we provide 19 bioclimatic variables (e.g., temp., seasonality, etc.) averaged over the period from 1981 to 2010. All data were extracted from [CHELSA](#) [32, 33].

**Image captions.** Recent advances in VLMs [1, 2, 37] have demonstrated strong performance across tasks such as image reasoning [1] and captioning [37] and shown that VLMs can effectively understand and reason about fine-grained details within images [39]. Building on these insights, we provide text descriptions for most photographs using the state-of-the-art open-source Malmo-7B VLM model [14]. We generate baseline captions (see Figure 4 and Figure 9, App.) with a prompt specifically designed to emphasize visual characteristics relevant to fungi identification, while avoiding unnecessary or potentially misleading details. The following prompt was used to guide the caption generation:

*“Describe the visual features of the fungi, such as their colour, shape, texture, and relative size. Focus on the fungi and their parts. Provide a detailed description of the visual features, but avoid speculations.”*



*Its stem is thick and light brown, with a hint of green at the base. The smaller mushroom on the right has a similar light brown cap, but its rim is more pronounced and has a white, almost translucent appearance. This gives it a delicate, lacy look. The stem of this mushroom is thinner and lighter in color compared to .....*

Figure 4. **Image caption sample.** For each photograph, we use a Malmo-7B [14] VLM to produce a realistic image caption with an exhaustive text description.

**Location-related metadata** is provided for approximately 99.9% of the observations and describes the location, time, taxonomy, and toxicity of the specimen, surrounding environment, and capturing device. See Table 2 for a detailed description of all available location-related metadata. While part of the metadata is usually provided by citizen scientists<sup>3</sup>, some attributes (e.g., elevation, land cover, and biogeographical) are crawled externally; all with potential to improve the classification accuracy and enable research on combining visual data with metadata.

Table 2. **List of available location-related metadata.** For virtually all observations (>99.9%), we provide data describing the surroundings or the specimen. Using such data for species identification allows to improve accuracy; see [16, 49].

Metadata	Description
<b>Observation date</b>	Date when the specimen was observed in yyyy-mm-dd format. Besides, three additional columns with pre-extracted <i>year</i> , <i>month</i> , and <i>day</i> values are provided.
<b>EXIF</b>	Camera attributes extracted from the image, e.g., metering mode, color space, device type, exposure, etc.
<b>Habitat</b>	The environment where the specimen was observed. Selected from 32 values such as <i>Mixed woodland</i> , <i>Deciduous woodland</i> , etc.
<b>Substrate</b>	The natural substance on which the specimen lives. Selected from 32 values such as <i>Bark</i> , <i>Soil</i> , <i>Stone</i> , etc.
<b>Taxonomic labels</b>	For each observation, we provide full taxonomic labels that include all ranks from species level up to kingdom. All are available in separate columns.
<b>Toxicity status</b>	Whether the species is poisonous or not as a binary value. Since non-edible species can cause serious health issues as well, we label them as poisonous.
<b>Location</b>	Latitude + longitude and coarser administrative divisions into regions, districts, and countries.
<b>Biogeographical zone</b>	One of the major biogeographical zones, e.g., <i>Atlantic</i> , <i>Continental</i> , <i>Alpine</i> , <i>Mediterranean</i> , and <i>Boreal</i> .
<b>Elevation</b>	Standardized height above the sea level.
<b>Landcover</b>	Land cover classification code with values like <i>savanna</i> , <i>barren</i> , etc. Taken from MODIS Terra+Aqua [20].

<sup>3</sup>A member of the public who actively participates in data collection, contributing valuable information to support professional scientists.

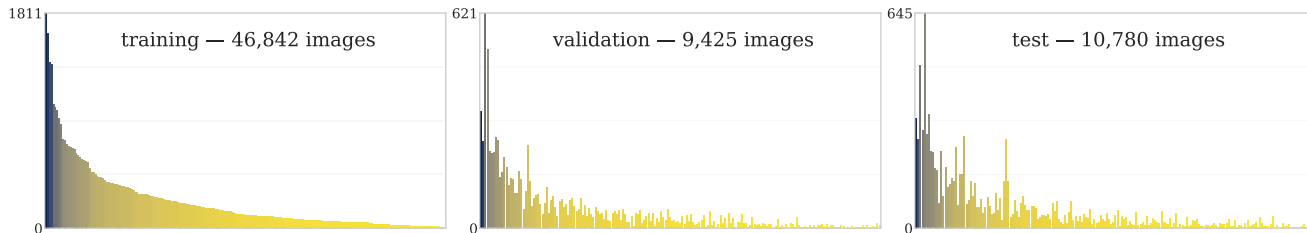


Figure 5. **Class distribution shift on the FungiTastic-M dataset.** The long-term data acquisition captures a phenomenon related to natural changes in species presence, i.e., class prior shift. Sorted in descending order based on their occurrence in the training set. The training set includes data from 2021 and before (215 species), the validation set from 2022 (196 species), and the test set from 2023 (193 species).

## 4. FungiTastic Benchmarks

The diversity and unique features of the FungiTastic dataset allow for the evaluation of various fundamental computer vision and machine learning problems. We present several benchmarks, each with its own evaluation protocol. This section provides a detailed description of each challenge and the corresponding evaluation metrics. Metrics are further defined in Appendix A.

**Closed-set classification:** The FungiTastic dataset is a challenging dataset with many visually similar species, a heavy long-tailed distribution, and considerable distribution shifts over time. Since the fine-grained closed-set classification methodology is well-defined, we follow the widely accepted standard, and we, apart from accuracy, use the macro-averaged F1-score ( $F_1^m$ ).

**Open-set classification:** In the Atlas of Danish Fungi (our data source), new species are continuously added to the database, including previously unreported species. This long-term ongoing data acquisition enables a yearly data split with a natural class distribution shift (see Figure 5), and many species in the test data are absent in the training set. We follow a widely accepted methodology, and we propose to use an AUC as the main metric. Besides, we calculate True Negative Rate at 95% True Positive Rate ( $TNR^{95}$ ) metric.

**Few-shot classification:** All the categories with less than five samples, usually uncommon and rare species, form the few-shot subset. Being capable of recognizing those is of high interest to the experts. Since the few-shot dataset has no severe class imbalance like the other FungiTastic subsets, the main metric is Top1 accuracy. The macro-averaged F1-score ( $F_1^m$ ) and Top3 total accuracy are also reported. This challenge does not have any “unknown” category.

**Chronological classification:** Each observation in the FungiTastic dataset has a timestamp, allowing the study of species distribution changes over time. Fungi distribution is

seasonal and influenced by weather, such as recent precipitation. New locations may be added over time, providing a real-world benchmark for domain adaptation methods, including online, continual, and test-time adaptation. The test dataset consists of fungi images ordered chronologically, meaning a model processing an observation at time  $t$  can access all observations with timestamps  $t' < t$ .

**Classification beyond 0–1 loss function:** Evaluation of classification networks is typically based on the 0–1 loss function, such as the mean accuracy, which also applies to the metrics defined for the previous challenges. This often falls short of the desired metric in practice since not all errors are equal. In this challenge, we define two practical scenarios: In the first scenario, confusing a poisonous species for an edible one (false positive edible mushroom) incurs a much higher cost than that of a false positive poisonous mushroom prediction. In the second scenario, the cost of not recognizing that an image belongs to a new species should be higher.

**Segmentation:** Acquiring human-annotated segmentation masks can be resource-intensive, yet segmentation is vital for advanced recognition and fine-grained classification methods [8, 53]. Accurate segmentation of fungal images supports these methods and enables automated analysis of species-specific morphological and environmental relationships and revelation of ecological and morphological patterns across locations. With its annotations, FungiTastic-M is built to accommodate semantic segmentation using the standard mean Intersection over Union (mIoU) metric and instance segmentation with the mean Average Precision (mAP) metric.

## 5. Baseline Experiments

In this section, we describe various weak and strong baselines based on state-of-the-art architectures and methods for four FungiTastic benchmarks. We report results for the closed-set, few-shot learning, and zero-shot segmentation, but other baselines will be provided later in the supplementary materials, documentation, or on the dataset website.

### 5.1. Closed-set Image Classification

We train a variety of state-of-the-art CNN architectures to establish some baselines for closed-set classification on the FungiTastic and FungiTastic-M. All selected architectures were optimized with Stochastic Gradient Descent with momentum set to 0.9, SeeSaw loss [69], a mini-batch size of 64, and Random Augment [12] with a magnitude of 0.2. The initial LR was set to 0.01 (except for ResNet and ResNeXt, with LR=0.1), and it was scheduled based on validation loss.

**Results:** Similarly to other fine-grained benchmarks, while the number of params, complexity of the model, and training time are more or less the same, the transformer-based architectures achieved considerably better performance on both FungiTastic and FungiTastic-M and two different input sizes (see Table 3 and Table 8 in Appendix). The best-performing model, BEiT-Base/p16 [3], achieved  $F_1^m$  just around 40%, which shows the severe difficulty.

Table 3. **Closed-set fine-grained classification on FungiTastic (FungiTastic) and FungiTastic-M.** A set of selected state-of-the-art Convolutional- (top section) and Transformer-based (bottom section) architectures evaluated on test sets. All reported metrics show the challenging nature of the dataset.

Architecture	FungiTastic-M – 224 <sup>2</sup>			FungiTastic – 224 <sup>2</sup>		
	Top1	Top3	F <sub>1</sub> <sup>m</sup>	Top1	Top3	F <sub>1</sub> <sup>m</sup>
ResNet-50 [25]	61.7	79.3	35.2	62.4	77.3	32.8
ResNeXt-50 [71]	62.3	79.6	36.0	63.6	78.3	33.8
EfficientNet-B3 [62]	61.9	79.2	36.0	64.8	79.4	34.7
EfficientNet-v2-B3 [63]	65.5	82.1	38.1	66.0	80.0	36.0
ConvNeXt-Base [42]	66.9	84.0	41.0	67.1	81.3	36.4
ViT-Base/p16 [18]	68.0	<u>84.9</u>	39.9	<u>69.7</u>	<u>82.8</u>	<u>38.6</u>
Swin-Base/p4w12 [41]	<b>69.2</b>	<b>85.0</b>	42.2	69.3	82.5	38.2
BEiT-Base/p16 [3]	<u>69.1</u>	84.6	<b>42.3</b>	<b>70.2</b>	<b>83.2</b>	<b>39.8</b>

### 5.2. Few-shot Image Classification

Three baseline methods are implemented. The first baseline is standard classifier training with the Cross-Entropy (CE) loss. The other two baselines are nearest-neighbor classification and centroid prototype classification based on deep embeddings extracted from large-scale pre-trained vision models, namely CLIP [51], BioCLIP [59] and DINOv2 [45].

Standard deep classifiers are trained with the CE loss to output the class probabilities for each input sample. Nearest neighbors classification ( $k$ -NN) constructs a database of training image embeddings. At test time,  $k$  nearest neighbors are retrieved, and the classification decision is made based on the majority class of the nearest neighbors. Nearest-centroid-prototype classification constructs a prototype embedding for each class by aggregating the training data embeddings of the given class. The classification depends on the image embedding similarity to the class prototypes. These methods are inspired by prototype networks proposed in [57].

**Results:** While DINOv2 [45] embeddings greatly outperform CLIP [51] embeddings, BioCLIP [59] outperforms them both, highlighting the dominance of domain-specific models. Further, the centroid-prototype classification always outperforms the nearest-neighbor methods. Finally, the best standard classification models trained on the in-domain few-shot dataset underperform both DINOv2 and CLIP embeddings, which shows the power of methods tailored to the few-shot setup. For results summary, refer to Table 4.

Table 4. **Few shot classification on FungiTastic-Few-shot.** Pre-trained deep descriptors with the nearest centroid and 1-NN nearest neighbor classification (Left) and fully supervised (max 4 examples per class) classifier with cross-entropy-loss (Right). All pre-trained models are based on the ViT-B architecture, CLIP [51], and BioCLIP [59] with patch size 32 and DINOv2 [45] with patch size 16.

Model	Method	Top1	Top3	Architecture	Input	Top1	Top3
CLIP	1-NN	6.1	–	BEiT-B/p16	224×224	11.0	17.4
	centroid	7.2	13.0		384×384	11.4	18.4
DINOv2	1-NN	17.4	–	ConvNeXt-B	224×224	14.0	23.1
	centroid	17.9	27.8		384×384	15.4	23.6
BioCLIP	1-NN	18.8	–	ViT-Base/p16	224×224	13.9	21.5
	centroid	<b>21.8</b>	<b>32.6</b>		384×384	19.5	29.0

### 5.3. Experiments with Additional Metadata

We provide baseline experiments using tabular metadata (habitat, month, substrate) based on previous work [49]. Table 5 shows that all the attributes improve all the metrics. Individually, the addition of the habitat attribute results in the biggest gains in accuracy (2.3%), followed by substrate (1.2%) and month (0.9%). Overall, habitat was the most efficient way to improve performance. With the combination of Habitat, Substrate, and Month, we improved the EfficientNet-B3 model’s performance on FungiTastic-M by 3.62%, 3.42% and 7.46% in Top1, Top3, and F1, respectively, indicating the gains are mostly orthogonal. Using the MetaSubstrate instead of Substrate resulted in performance lower by 0.2%, 0.5%, and 0.3% in Top1, Top3, and F1, respectively.

Table 5. **Ablation on a combination of observation-related data.** Utilizing a simple yet effective approach based on previous work [49], we measure performance improvement using Habitat, Substrate, and Month and their combination. We also test how replacing Substrate variables with MetaSubstrate affects performance. Evaluated with EfficientNet-B3 on FungiTastic-M test set.

	Habitat	Month	Substrate	MetaSub.	Habitat	Month	Substrate	MetaSub.	Habitat	Month	Substrate	MetaSub.
	✓	–	–	–	✓	✓	✓	–	–	✓	✓	✓
	–	✓	–	–	✓	–	–	✓	✓	✓	✓	–
	–	–	✓	–	–	✓	–	✓	–	✓	–	–
	–	–	–	✓	–	–	✓	–	✓	–	–	✓
<b>Top1</b>	<b>+2.3</b>	+0.9	<u>+1.2</u>	+0.9	+3.1	+3.0	+2.7	+1.9	+1.6	+3.6	+3.3	+3.1
<b>F<sub>1</sub><sup>m</sup></b>	<b>+4.0</b>	+1.1	<u>+2.3</u>	+1.5	+6.0	+5.9	+5.1	+4.0	+3.2	+7.5	+6.8	+6.8
<b>Top3</b>	<b>+2.3</b>	+0.5	<u>+0.8</u>	+0.6	+2.7	+2.9	+2.6	+1.5	+1.1	+3.4	+3.1	+3.1

## 5.4. Segmentation

A zero-shot baseline for foreground-background binary segmentation of fungi is evaluated on the FungiTastic-M dataset. The method consists of two steps: 1. The GroundingDINO [40] (the ‘tiny’ version of the model) zero-shot object detection model is prompted with the text ‘mushroom’ and outputs a set of instance-level bounding boxes. 2. The bounding boxes from the first step are used as prompts for the SAM [35] segmentation model. All the experiments are conducted on images with the longest edge resized to 300 pixels while preserving the aspect ratio.

**Results:** The baseline method achieved an average per-image IoU of 89.36%. While the model exhibits strong zero-shot performance, it sometimes fails to detect mushrooms. These instances often involve small mushrooms, where a higher input resolution could enhance detection, and atypical mushrooms, such as very thin ones. Another common issue is SAM’s tendency to miss mushroom stems. The results for the simplified foreground-background segmentation task underline the need for further development of domain-specific models. Qualitative results, including random images and examples where the segmentation performs best and worst, are reported in Figure 6.

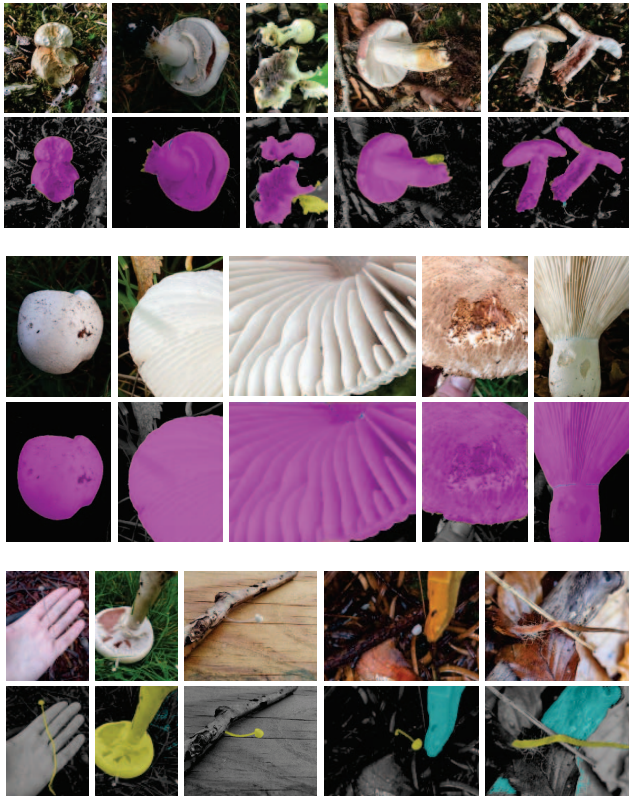


Figure 6. **Zero-shot Fungi segmentations on FungiTastic-M benchmark.** Random samples (top section), best IoU samples (mid), and worst IoU samples (bottom). Highlighted pixels correspond to: true positives, false positives, and false negatives.

## 5.5. Open-set Image Classification

While constructing baselines for the FungiTastic open-set benchmark, we approached open-set classification as a binary decision-making problem, where the model determines whether a new image belongs to a known class or a novel class. This method serves as an initial step in the classification pipeline, deciding if a closed-set classifier is suitable for recognizing a given sample. We evaluate several approaches for open-set classification:

- Maximum Softmax Probability [28] (MSP): Uses the highest probability from softmax output from a closed-set classifier as the open-set score.
- Maximum Logit Score [67] (MLS): Uses the highest logit output from a closed-set classifier as the open-set score.
- Nearest Mean Score (NM): Computes the mean embedding for each class, then calculates the Euclidean distance between an image embedding and the nearest class mean.

We use features and logits from the BEiT-Base/p16 closed-set classifier baseline, trained on the full dataset (i.e., FungiTastic) at a  $384 \times 384$  resolution. To explore the potential of general pre-trained representation, we compare the fully-supervised model with generic features from a pre-trained DINOv2 model [45]. Using DINOv2 features, we train a simple linear layer to obtain MSP and MLS scores. Note that BEiT represents the best model from the closed-set classification baseline.

**Results:** The MLS method achieved the best open-set classification performance on both backbones. With BEiT-Base/p16, MLS achieves a  $\text{TNR}^{95}$  of 27.7% and an AUC of 83.9%, which are the highest AUC across all methods. DINOv2, in contrast, achieves the best  $\text{TNR}^{95}$  with a score of 36.9% using the MLS method, though its AUC is slightly lower at 74.5%. The MSP method also performs well with DINOv2, reaching a  $\text{TNR}^{95}$  of 32.5% and an AUC of 82.4%. However, the NM method, which relies on feature embeddings rather than classifier outputs, significantly underperforms in both metrics. See Table 6 for more details.

Table 6. **Open-set classification baselines.** Overall, the results are inconclusive and highly metric-dependent. The MLS (*Max. Logit*) method with the BEiT-Base/p16 backbone yields the highest AUC (83.9%), while the DINOv2 backbone with MLS achieves the highest  $\text{TNR}^{95}$  (36.9%). The NM (*Nearest Mean*) method consistently underperforms in both metrics across both backbones. For the AUC metric, MLS with a BEiT backbone (fine-tuned on the FungiTastic closed-set dataset) outperforms other approaches. However, both MSP (*Max. Softmax*) and MLS using DINOv2 linear layer are better when  $\text{TNR}^{95}$  performance is considered.

Backbone	Nearest Mean		Max. Logit		Max. Softmax	
	$\text{TNR}^{95}$	AUC	$\text{TNR}^{95}$	AUC	$\text{TNR}^{95}$	AUC
BEiT-Base/p16	23.2	73.9	27.7	83.9	25.3	79.8
DINOv2	12.1	69.2	36.9	74.5	32.5	82.4

## 5.6. Vision-Language Fusion

To evaluate the relevance of the available textual data (i.e., photograph captions) for species classification, we provide baselines that use a sequence classification variant of the lightweight DistilBERT [55] model trained as a classifier on textual descriptions only. The model was trained for 10 epochs using the standard cross-entropy loss, with logits obtained from a classification head applied to the pooled features of the class token in DistilBERT. For evaluation, we use text descriptions generated for the images in the test set.

**Results:** The DistilBERT classifier achieves a Top1 accuracy of 31.2% on FungiTastic-M and 24.1% on the full benchmark; significantly lower than fully supervised BEiT classifiers. However, this is still a strong result, given that it relies solely on textual descriptions. A simple ensemble that averages logits from the image and text classifiers shows potential for improved accuracy, indicating that the two methods are complementary. VLM-based descriptions capture useful details often missed by the image model. Further analysis shows the ensemble improves performance mainly on common categories, while the image classifier performs better on rare ones. This trade-off likely accounts for the drop in  $F_1^m$  and overall accuracy on the full benchmark. For more details, see Table 7 and Figure 7.

Table 7. **Vision-Language fusion performance.** DistilBERT uses text descriptions of images for species classification. Fusion method predictions are the mean of DistilBERT and BEiT logits.

Architectures	FungiTastic-M – 224 <sup>2</sup>			FungiTastic – 224 <sup>2</sup>		
	Top1	Top3	$F_1^m$	Top1	Top3	$F_1^m$
DistilBERT	31.2	50.2	11.5	24.1	39.1	8.8
BEiT-Base/p16	<u>67.3</u>	<u>83.3</u>	<u>40.5</u>	<u>70.2</u>	<u>83.2</u>	<u>41.1</u>
Fusion	<b>67.7</b>	<b>83.8</b>	<u>39.8</u>	<u>69.0</u>	<u>82.6</u>	<u>40.0</u>

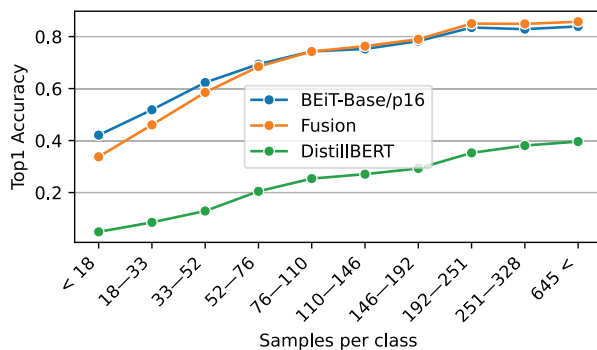


Figure 7. **Vision-Language fusion – accuracy dependence on class frequency.** Like the vision model (BEiT-Base/p16), the language model (DistilBERT) struggles with infrequent classes. Fusion improves accuracy mainly for species with over 100 samples. The test set is binned by class frequency into deciles (x-axis).

## 6. Conclusion

In this work, we introduced the FungiTastic, a comprehensive and multi-modal dataset and benchmark. The dataset includes a variety of data types, such as photographs, satellite images, climatic data, segmentation masks, and observation metadata. FungiTastic has many interesting features, which make it attractive to the broad ML community. With its data sampling spanning 20 years, precise labels, rich metadata, long-tailed distribution, distribution shifts over time, the visual similarity between the categories, and multimodal nature, it is a unique addition to the existing benchmarks.

In the provided baseline experiments, we demonstrate how challenging the FungiTastic Benchmarks are. Even state-of-the-art architectures and methods yield modest F-scores of 39.8% in closed-set classification and 9.1% in few-shot learning, highlighting the dataset’s challenging nature compared to traditional benchmarks such as CUB-200-2011, Stanford Cars, and FGVC Aircraft. The proposed zero-shot baseline for the simplest segmentation task, binary segmentation of fungi fruiting body, achieved an average IoU of 89.36%, which still shows the potential for improvement in fine-grained visual segmentation of fungi. The open-set baselines show that discovering novel classes remains a difficult task, demanding new techniques tailored to fine-grained recognition. Additionally, results with non-domain-specific vision-language models reveal a surprisingly strong performance of such models. The fusion experiments of VLMs with supervised models confirm the challenge of accurately classifying rare species in highly imbalanced datasets.

**Limitations** lie in the data collection process, which affects the overall distribution. Most of the data comes from Denmark, and bias is further introduced through "random" sampling. Therefore, some species are more common in frequently sampled areas or are favored by collectors. Some recent observations also miss metadata, which can reduce the effectiveness of classification methods that rely on it.

**Future work** includes setting up and running future challenges [31], expanding baseline models, adding new test sets, and exploring extra data like traits and species descriptions to enhance multi-modal performance.

## Acknowledgement

This research was supported by the Technology Agency of the Czech Republic, project No. SS73020004. We extend our sincere gratitude to the mycologists from the Danish Mycological Society, particularly Jacob Heilmann-Clausen, Thomas Læssøe, Thomas Stjernegaard Jeppesen, Tobias Guldberg Frøslev, Ulrik Søchting, and Jens Henrik Petersen, for their contributions and expertise. We also thank the dedicated citizen scientists whose data and efforts have been instrumental to this project. Your support and collaboration have greatly enriched our work and made this research possible. Thank you for your commitment to advancing ecological understanding and conservation.

## References

- [1] OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and et al. Gpt-4 technical report. 2023. 4
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 4
- [3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 6
- [4] Linda J Beaumont, Lesley Hughes, and Michael Poulsen. Predicting species distributions: use of climatic parameters in bioclim and its impact on predictions of species' current and future distributions. *Ecological modelling*, 186(2):251–270, 2005. 4
- [5] Asish Bera, Zachary Wharton, Yonghuai Liu, Nik Bessis, and Ardhendu Behera. Sr-gnn: Spatial relation-aware graph neural network for fine-grained image categorization. *IEEE Transactions on Image Processing*, 31:6017–6031, 2022. 2
- [6] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2011–2018, 2014. 2
- [7] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xi-aohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 2
- [8] Gaurav Bhatt, Deepayan Das, Leonid Sigal, and Vineeth N Balasubramanian. Mitigating the effect of incidental correlations on part-based learning. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 5
- [9] Isabelle Bolon, Lukáš Pícek, Andrew M Durso, Gabriel Alcoba, François Chappuis, and Rafael Ruiz de Castañeda. An artificial intelligence model to identify snakes from across the world: Opportunities and challenges for global health and herpetology. *PLoS neglected tropical diseases*, 16(8): e0010647, 2022. 2
- [10] Christophe Botella, Alexis Joly, Pierre Bonnet, Pascal Monestiez, and François Munoz. A deep learning approach to species distribution modelling. *Multimedia tools and applications for environmental & biodiversity informatics*, pages 169–199, 2018. 4
- [11] Po-Yung Chou, Yu-Yung Kao, and Cheng-Hung Lin. Fine-grained visual classification with high-temperature refinement and background suppression. *arXiv preprint arXiv:2303.06442*, 2023. 2
- [12] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 6
- [13] Jim W Deacon. *Fungal biology*. John Wiley & Sons, 2013. 3
- [14] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 4
- [15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2
- [16] Qishuai Diao, Yi Jiang, Bin Wen, Jia Sun, and Zehuan Yuan. Metaformer: A unified meta framework for fine-grained recognition. *arXiv preprint arXiv:2203.02751*, 2022. 2, 4
- [17] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, Nenghai Yu, and Baining Guo. Peco: Perceptual codebook for bert pre-training of vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 552–560, 2023. 2
- [18] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6
- [19] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 2
- [20] Mark A Friedl, Damien Sulla-Menashe, Bin Tan, Annemarie Schneider, Navin Ramankutty, Adam Sibley, and Xiaoman Huang. Modis collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote sensing of Environment*, 114(1):168–182, 2010. 4
- [21] Camille Garcin, Alexis Joly, Pierre Bonnet, Jean-Christophe Lombardo, Antoine Affouard, Mathias Chouet, Maximilien Servajean, Titouan Lorieul, and Joseph Salmon. Pl@ ntnet-300k: a plant image dataset with high label ambiguity and a long-tailed distribution. In *NeurIPS 2021-35th Conference on Neural Information Processing Systems*, 2021. 2
- [22] Zahra Gharaee, ZeMing Gong, Nicholas Pellegrino, Iuliia Zarubiieva, Joakim Bruslund Haurum, Scott Lowe, Jaclyn McKeown, Chris Ho, Joschka McLeod, Yi-Yun Wei, et al. A step towards worldwide biodiversity assessment: The bioscan-1m insect dataset. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [23] Herve Goeau, Pierre Bonnet, and Alexis Joly. Plant identification based on noisy web data: the amazing performance of deep learning (lifeclef 2017). *CEUR Workshop Proceedings*, 2017. 2
- [24] Ian Goodfellow. *Deep learning*. MIT press, 2016. 2
- [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [26] Wei He, Kai Han, Ying Nie, Chengcheng Wang, and Yunhe Wang. Species196: A one-million semi-supervised dataset for fine-grained species recognition. *Advances in Neural Information Processing Systems*, 36, 2024. 2

- [27] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 2
- [28] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. 7
- [29] Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Funnybirds: A synthetic vision dataset for a part-based analysis of explainable ai methods. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3981–3991, 2023. 3
- [30] Robert J Hijmans and Catherine H Graham. The ability of climate envelope models to predict the effect of climate change on species distributions. *Global change biology*, 12(12):2272–2281, 2006. 4
- [31] Alexis Joly, Lukáš Pícek, Stefan Kahl, Hervé Goëau, Lukáš Adam, Christophe Botella, Maximilien Servajean, Diego Marcos, Cesar Leblanc, Théo Larcher, Jiří Matas, Klára Janoušková, Vojtěch Čermák, Kostas Papafitsoros, Robert Planqué, Willem-Pier Vellinga, Holger Klinck, Tom Denton, Pierre Bonnet, and Henning Müller. Lifeclef 2025 teaser: Challenges on species presence prediction and identification, and individual animal identification. In *Advances in Information Retrieval*, pages 373–381, Cham, 2025. Springer Nature Switzerland. 8
- [32] DN Karger, O Conrad, J Böhner, T Kawohl, H Kreft, RW Soria-Auza, NE Zimmermann, HP Linder, and M Kessler. Climatologies at high resolution for the earth’s land surface areas. *sci. data* 4, 170122, 2017. 4
- [33] Dirk Nikolaus Karger, Olaf Conrad, Jürgen Böhner, Tobias Kawohl, Holger Kreft, Rodrigo Wilber Soria-Auza, Niklaus E Zimmermann, H Peter Linder, and Michael Kessler. Climatologies at high resolution for the earth’s land surface areas. *Scientific data*, 4(1):1–20, 2017. 4
- [34] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*. Citeseer, 2011. 2
- [35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 3, 7
- [36] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 2
- [37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 4
- [38] Dichao Liu, Longjiao Zhao, Yu Wang, and Jien Kato. Learn from each other to classify better: Cross-layer mutual attention learning for fine-grained visual classification. *Pattern Recognition*, 140:109550, 2023. 2
- [39] Mingxuan Liu, Subhankar Roy, Wenjing Li, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Democratizing fine-grained visual recognition with large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 4
- [40] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 7
- [41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6
- [42] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 6
- [43] Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1, 2
- [44] Hoang-Quan Nguyen, Thanh-Dat Truong, Xuan Bac Nguyen, Ashley Dowling, Xin Li, and Khoa Luu. Insect-foundation: A foundation model and large-scale 1m dataset for visual insect understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21945–21955, 2024. 2
- [45] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6, 7
- [46] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 2
- [47] Lukáš Pícek, Marek Hruží, Andrew M Durso, and Isabelle Bolon. Overview of snakeclef 2022: Automated snake species identification on a global scale. 2022. 2
- [48] Lukáš Pícek, Milan Šulc, Jiří Matas, Jacob Heilmann-Clausen, Thomas S Jeppesen, and Emil Lind. Automatic fungi recognition: deep learning meets mycology. *Sensors*, 22(2):633, 2022. 2
- [49] Lukáš Pícek, Milan Šulc, Jiří Matas, Thomas S Jeppesen, Jacob Heilmann-Clausen, Thomas Læssøe, and Tobias Frøslev. Danish fungi 2020—not just another image recognition dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1525–1535, 2022. 1, 2, 4, 6
- [50] Lukas Pícek, Christophe Botella, Maximilien Servajean, César Leblanc, Rémi Palard, Théo Larcher, Benjamin Deneu, Diego Marcos, Pierre Bonnet, and Alexis Joly. Geoplant: Spatial plant species prediction dataset. *arXiv preprint arXiv:2408.13928*, 2024. 4

- [51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6
- [52] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 2
- [53] Mattia Rigotti, Christoph Miksovics, Ioana Giurgiu, Thomas Gschwind, and Paolo Scotton. Attention-based interpretability with concept transformers. In *International conference on learning representations*, 2021. 3, 5
- [54] Duccio Rocchini, Doreen S Boyd, Jean-Baptiste Féret, Giles M Foody, Kate S He, Angela Lausch, Harini Nagendra, Martin Wegmann, and Nathalie Pettorelli. Satellite remote sensing to monitor species diversity: Potential and pitfalls. *Remote Sensing in Ecology and Conservation*, 2(1):25–36, 2016. 4
- [55] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019. 8
- [56] Sri Kumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. Taxabind: A unified embedding space for ecological applications. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1765–1774. IEEE, 2025. 1, 2
- [57] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 6
- [58] Siddharth Srivastava and Gaurav Sharma. Omnivec: Learning robust representations with cross modal sharing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1236–1248, 2024. 2
- [59] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19412–19424, 2024. 6
- [60] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European conference on computer vision (ECCV)*, pages 498–512, 2018. 2
- [61] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific data*, 2(1): 1–14, 2015. 2
- [62] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6
- [63] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 6
- [64] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 595–604, 2015. 2
- [65] Grant Van Horn, Oisín Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 1, 2
- [66] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisín Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893, 2021. 1
- [67] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations*, 2022. 7
- [68] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1, 2
- [69] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9695–9704, 2021. 6
- [70] Wikipedia. Heraclitus — Wikipedia, the free encyclopedia. <http://en.wikipedia.org/w/index.php?title=Heraclitus&oldid=1227413074>, 2024. 1
- [71] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 6