

# Danish Fungi 2020 - Not Just Another Image Recognition Dataset

Lukáš Picek  
University of West Bohemia  
picekl@kky.zcu.cz

Milan Šulc, Jiří Matas  
CTU in Prague  
sulcmila, matas@fel.cvut.cz

Thomas S. Jeppesen  
GBIF  
tsjeppesen@gbif.org

Jacob Heilmann-Clausen, Thomas Læssøe, Tobias Frøslev  
University of Copenhagen

jheilmann-clausen@smn.ku.dk, thomasl@bio.ku.dk, tobiasgf@sund.ku.dk

## Abstract

We introduce a novel fine-grained dataset and benchmark, the Danish Fungi 2020 (DF20). The dataset, constructed from observations submitted to the Danish Fungal Atlas, is unique in its taxonomy-accurate class labels, small number of errors, highly unbalanced long-tailed class distribution, rich observation metadata, and well-defined class hierarchy. DF20 has zero overlap with ImageNet, allowing unbiased comparison of models fine-tuned from publicly available ImageNet checkpoints. The proposed evaluation protocol enables testing the ability to improve classification using metadata – e.g. precise geographic location, habitat, and substrate, facilitates classifier calibration testing, and finally allows to study the impact of the device settings on the classification performance. Experiments using Convolutional Neural Networks (CNN) and the recent Vision Transformers (ViT) show that DF20 presents a challenging task. Interestingly, ViT achieves results superior to CNN baselines with 81.25% accuracy, reducing the CNN error by 13%. A baseline procedure for including metadata into the decision process improves the classification accuracy by more than 3.5 percentage points, reducing the error rate by 20%. The source code for all methods and experiments is available at <https://sites.google.com/view/danish-fungi-dataset>.

## 1. Introduction

Publicly available datasets and benchmarks accelerate machine learning research and allow for quantitative comparison of novel methods. In the area of deep learning and computer vision, the rapid progress over the past decade was, to a great extent, facilitated by the publication of large-scale image datasets. In the case of image recognition, the formation of the ImageNet [6] database and its



Figure 1. Selected images from the **DF20 dataset**. For visualisation, max-center-squares were extracted from the images.

usage in the ILSVRC<sup>1</sup> challenge [36], together with PASCAL VOC [9] and Caltech-256 [15] among others, helped start the CNN revolution. The same holds for the problem of fine-grained visual categorization (FGVC), where datasets and challenges like PlantCLEF [12, 13, 22], iNaturalist [46], CUB [48], and Oxford Flowers [33], have helped to develop and evaluate novel approaches to fine-grained domain adaptation [11], domain specific transfer learning [17], image retrieval [34, 38, 53], unsupervised visual representation, few-shot learning [49], transfer learning [17], prior-shift [39] and many others.

While the datasets have been extremely useful for the image recognition community, there are issues that limit their relevance to real-world applications. We mention several such problems. Flat class priors, common in

<sup>1</sup> The ImageNet Large Scale Visual Recognition Challenge.

research datasets, are rare in practice. Often, class prior distributions are the same in the training and test splits. This is a standard machine learning assumption that, nevertheless, is not valid if the collection of training data differs from the deployment of the trained system, which is not rare. Non-negligible percentage of noisy-labels restricts quality assessment [2], and, despite CNN’s surprising robustness to label noise [24], may influence the perceived relative merit of learning algorithms. Some commonly used datasets [6, 25, 33] are saturated in accuracy or close to the point, leaving limited space for improvement in future research [2]. Extremely large dataset sizes might discourage researchers that do not have access to massive computational resources as experiments have become time and hardware demanding.

With these observations in mind, we introduce **the DF20 dataset** with a number of unique characteristics. Its class labels are exceptionally accurate, annotated by domain experts. Estimates on a random subset show about 0.2% error; for the DNA sequenced private test set the labelling error is 0%. The minimal error levels allow highly accurate performance evaluation. With its zero overlap with ImageNet, it allows an unbiased comparison of models fine-tuned from publicly available ImageNet checkpoints.

The class frequencies in DF20 follow the natural species distribution, which is long-tailed. The frequencies change significantly within the calendar year, making the data suitable for testing the response of the classifier to differing long tailed distributions and changing class priors. The continuous data flow of collection over a long period starting from 1874 provide a ground for modelling and exploiting the temporal phenomena on different scales, e.g., month, season, year.

The visual data is accompanied with metadata for more than 99% of the image observations. The rich metadata include information about 20 attributes related to the environment, place, time and full taxonomy labels. The metadata enable to test the ability to improve classification accuracy using different metadata types – time, precise location, habitat, and substrate type, to perform hierarchical classification, evaluate fine-grained classification on different levels of granularity (taxonomic ranks), to test classifier calibration, and to model intra-metadata and metadata-visual appearance relationships. Moreover, EXIF metadata is available for many observations, which is useful, e.g., for studying the impact of the device settings on classification performance.

**The DF20 Benchmark.** To allow evaluation at any time, we have prepared a web-based public automatic benchmark<sup>2</sup> for different scenarios, including visual-based, metadata focused or classifier-calibration related research. Be-

sides the full benchmark, we introduce DF20 - mini, a small subset with roughly 1/10 of the data and species, for fast, low-energy friendly prototyping. DF20 - mini includes six well-known genera of fungi forming fruit-bodies of the toadstool type, and offers, surprisingly, an even more challenging problem than the full benchmark, while having a compact size.

We prepared a baseline performance evaluation, including the quantitative and qualitative analysis of the results for a number of well-known CNN and recent ViT architectures [7]. The recent ViT achieves excellent results in fine-grained classification outperforming the state-of-the-art CNN classifiers. We show that ViT performs way better on the FGVC domain, where attention to detail is needed, than in a common object recognition. We show that both the DF20 and DF20 - Mini benchmarks are far from saturated as the best performing model - ViT-Large/16-384 - achieved 81.25% and 75.32% accuracy on DF20 and DF20 - Mini, respectively. We propose a baseline solution for processing the habitat, substrate and time (month) metadata, showing that – even with the baseline approach – utilizing the metadata increases the classification performance significantly. To support and accelerate future research on the DF20 we open sourced the code through the public GitHub repository<sup>3</sup>.

## 2. Related Work

This section overviews existing fine-grained image datasets. Fine-grained datasets, unlike datasets with common object classes such as - dog, plant, aircraft, or vehicle [6, 9, 26], are characterized by classes with small-interclass differences and huge intra-class similarity. Currently, there exists a number of FGVC dataset with a focus on plants [13, 20, 22, 27, 33], animals [23, 45, 48], cars [25, 42] or airplanes [32]. They feature object classes within one meta-category,i.e., dog breeds, vehicle models and plant or animal species. The dataset statistics are compared in Table 1. Most of the datasets are artificially constructed to have a flat class distribution. Many datasets use web scraped data that may contain out-of-domain images and wrong labels.

**Labels.** As species-level labels are essential for usage in real-world applications, the tedious labelling procedure often rely on domain experts. With just a small number of experts and their limited time, the labelling process is frequently delegated to crowd-sourced annotation platforms such as the Amazon Mechanical Turk [6, 23, 48]. The main drawback of this approach is related to poor domain knowledge of the annotators that results in a high number of noisy labels [45] – 4.4% in CUB 200-2011 and approximately 4.0% for fine-grained classes in ImageNet. To ad-

<sup>2</sup> <https://aicrowd.com/challenges/danish-fungi>

<sup>3</sup> <https://github.com/picekl/DanishFungiDataset>

Name	# Classes	# Training	# Test
FGVC-Aircrafts [32]	102	6,732	3,468
Standford Cars [25]	196	8,144	8,041
VMMRdv [42]	9,170	291,752	×
Oxford Flowers [33]	102	1,020	7,169
Stanford Dogs [30]	120	12,000	8,580
DogSnap [30]	133	4,776	3,575
LeafSnap [27]	185	30,866	×
CUB 200-2011 [48]	200	5,994	5,784
VegFru [20]	292	29,200	116,931
NABirds [45]	555	48,562	×
PlantCLEF 2015 [22]	1,000	91,758	21,446
iNaturalist 2017 [46]	5,089	579,184	95,986
PlantCLEF 2017 <sup>†</sup> [13]	10,000	230,658	25,629
<b>DF20 - Mini</b>	182	32,753	3,640
<b>DF20</b>	1,604	248,466	27,608

Table 1. Overview of publicly available FGVC datasets, both nature-related (middle section) and other (top), and the number of images and categories. <sup>†</sup> Including only images with clean labels.

dress this issue, more recent datasets use citizen-science platforms and their users - citizen scientists<sup>4</sup> - to label data with high-quality annotations [45, 46].

**ImageNet Overlap.** Using different data for training and testing is a standard in the field of image recognition. Nevertheless, the data overlap, in fine-tuning, where ImageNet weights are primarily used, is commonly overlooked even though there is substantial overlap for multiple datasets [20, 30, 48]. Interestingly, a number of publications with high impact used ImageNet weights and performed the fine-tuning and testing with the CUB 200-2011 [48] dataset that overlaps with the ImageNet [3, 28, 29, 34, 51, 52, 54, 55, 56].

**Metadata.** Besides images and class labels, image classification datasets often provide additional metadata, such as higher taxon labels [22, 46], label hierarchy [6, 20, 32], object parts and attribute annotations [22, 45, 48], masks [48], GPS location [22], and time of observation [22]. The existence of such metadata enables the usage of these datasets in machine learning research beyond image classification. For example [31, 1, 8] use location context, and [14] use taxonomy labels.

### 3. Danish Fungal Atlas

The Danish Fungal Atlas (Svampeatlas) [5, 10, 19] involves more than 3,000 volunteers who have contributed with more than 900,000 quality-checked observations of fungi (many including images) since 2009. Data from pre-

<sup>4</sup> Domain specific nonprofessional enthusiasts - experts.

vious years has also been included.

The project has resulted in a vastly improved knowledge of Denmark’s fungi [19]. More than 180 species belonging to *Basidiomycota*<sup>5</sup> have been added to the list of known Danish species, and several species that were considered extinct have been re-discovered. Simultaneously, several search and assistance functions have been developed that present features relating to the individual species, making it much easier to include an understanding of endangered species in nature management and decision-making.

All validated records are published to the Global Biodiversity Information Facility (GBIF) every week, since 2017. The database includes more than 372,154 observation with images till this day.

### 4. Annotation Process

The Danish Fungal Atlas uses an interactive labelling method for all submitted observations. When a user submits a fungal sighting (record) at species level, a “reliability score” (1–100) is calculated based on following factors:

- Species rarity, *i.e.* its relative frequency in the Atlas.
- The geographical distribution of the species.
- Phenology of the species, its seasonality.
- User’s historical species-level proposal precision.
- As above, within the proposal’s higher taxon rank.

Subsequently, other users may agree with the proposed species identity, increasing the identification score following the same principles, or proposing alternative identification for non-committal suggestions. Once the submission reaches a score of 80, the label (identification) is internally approved. Simultaneously, a small group of taxonomic experts (validators) monitor most of the observation on their own. These have the power to approve or reject species identifications regardless of the score in the interactive validation. This can be relevant for discoveries of new species, rare species, and records of species where experience or sequencing of genetic material (DNA) is required for safe identification. Since 2019, the Danish Fungal Atlas’ observation identification is simplified thanks to an image recognition system [40].

### 5. Dataset Description

The **Danish Fungi 2020** (DF20) dataset contains image observations from the Danish Fungal Atlas belonging to species with more than 30 images. The dataset represents real observations dated from 1874 to the end of 2020, coming from more than 30 countries, and including samples from all seasons. It consists of 295,938 images belonging to 1,604 species from the *Fungi*, *Protozoa* and *Chromista* kingdoms. While the DF20 dataset uses approximately 80%

<sup>5</sup> a group of fungi that produces their sexual spores (basidiospores) on a club-shaped spore-producing structure (basidium).

	# Images	# Species	# Genera	# Family
Svampeatlas	372,154	5,923	1,417	368
DF20	295,938	1,604	566	190
DF20 - Mini	36,393	182	6	6

Table 2. Number of images, species, genera and families in the Danish Fungal Atlas and their subsets DF20 and DF20 - Mini.

of the Svampeatlas database images, it represent just 27% of the species. For a quantitative summary of the data selection, see Table 2. The most frequent species – *Trametes versicolor* – is represented by 1,913 images and the least present with 31.

Additionally, we hand-picked a subset of 36,393 images belonging to 182 species from 6 genera with a similar visual appearance. This compact dataset, **DF20 - Mini**, introduces a challenging fine-grained recognition task, while allowing to decrease the necessary training times and hardware requirements. As species in the same genus are most likely to be confused, we used six well-known genera of fungi forming fruit-bodies of the toadstool type (*Russula*, *Boletus*, *Amanita*, *Clitocybe*, *Agaricus* and *Mycena*) for the construction of the DF20 - Mini. The most frequent species in the DF20 - Mini dataset – *Mycena galericulata* – has 1,221 images, the minimum is again 31 images per class.

The DF20 and DF20 - Mini datasets were randomly split into the provided training and (public) test sets, where the training set contains [90%] of images of each species.

## 5.1. Metadata

Unlike most computer vision datasets, DF20 and DF20 - Mini include rich metadata acquired by citizen-scientists in the field while recording the observations. For clarity, we decided to only publish the frequently filled-in and relevant 21 categories with the dataset, although the Danish Fungal Atlas contains up to 255 data-points for each observation. We see a promising research direction in combining visual data with metadata like timestamp, location at multiple scales, substrate, habitat, full taxonomy labels and camera device settings. For detailed description see Table 3.

**Substrate.** Substrates on which fungi live and fruit are an important source of information that helps differentiate similarly looking species. Each species or genus has its preferable substrate, and it is rare to find it on other substrates. For example, *Trametes* occurs only on wood and *Russula* on soil. As such metadata can improve the final categorization capability, we provide one of 32 substrate types for more than 99% of images. We differentiate wood of living trees, dead wood, soil, bark, stone, fruits and others.

**Habitat.** While substrate denotes the spots, the habitat indicates the more overall environment where fungi grow and hence is vital for fungal recognition. It is well known

Attribute Description	
<b>EventDate</b>	Date of observation.
<b>IdentifiedBy</b>	Name of the user that identified the specimen.
<b>EXIF</b>	Camera device attributes extracted from the image, e.g., metering mode, color space, device type, exposure time, and shutter speed.
<b>Habitat</b>	The environment where the specimen was observed. Selected from 32 values such as Mixed woodland, Deciduous woodland etc.
<b>Substrate</b>	The natural substance on which the specimen lives. Selected from 32 values such as Bark of a living tree, Soil, Stone, etc.
<b>Scientific</b>	Lowest taxonomic rank including specific Epithets. 1,604 unique values present.
<b>Species</b>	1th taxon rank. 1,578 unique values present.
<b>Genus</b>	2nd taxon rank. 566 unique values present.
<b>Family</b>	3rd taxon rank. 190 unique values present.
<b>Order</b>	4th taxon rank. 66 unique values present.
<b>Class</b>	5th taxon rank. 23 unique values present.
<b>Phylum</b>	6th taxon rank. 5 unique values present.
<b>Kingdom</b>	7th taxon rank. 3 unique values present.
<b>CountryCode</b>	ISO 3166-1 alpha-2 code (DK, AT, etc.) of the observation. The dataset covers 30 countries.
<b>Locality</b>	More precise location information. Mostly smaller than a district, e.g. part of a city or a specific forest. 9003 values present.
<b>Level1Gid</b>	ID of a Country region related to the specimen observation, 115 regions are listed.
<b>Level2Gid</b>	ID of a district region related to the specimen observation, 317 districts are listed.
<b>Latitude</b>	A decimal GPS coordinate.
<b>Longitude</b>	A decimal GPS coordinate.
<b>GPSUncert</b>	GPS coordinates uncertainty in meters.

Table 3. Description of the provided metadata (observation attributes). For almost all images, a detailed information about taxonomy, location, time, habitat and substrate type is included.

that some species occur in deciduous forests rather than in conifer forests or plantations, while others grow in farmland. For a deeper understanding of such relation, we include the information about the habitat for approximately 99.5% of observations.

**Location.** Fungi are highly location dependent with different species distributions across continents, states, regions or even districts. To support studies on better understanding where Fungi species lives, we include multi-level location information. Starting from GPS coordinates with included uncertainty, we further extracted information about the country, region and district.

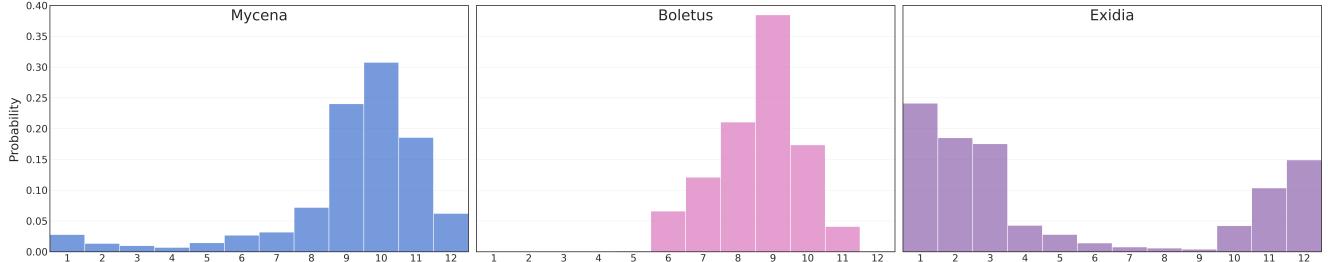


Figure 2. Monthly distribution of observations in the DF20 dataset for genera *Mycena*, *Boletus*, and *Agaricus*. The differences imply that the class prior distribution varies significantly over time.

**Time-Stamp.** Time of observation is essential for fungi classification in the wild as fruitbodies’ presence depends on seasonality or even (but rarely) the time in a day. Considering the existence of such dependency, integrating information about time into the classification should also improve fungal recognition. In Figure 2 we show the probability of three genera being observed in different months of the calendar year. Brief inspection shows that there is almost zero probability to spot a *Boletus* in January but still a small chance to find a *Mycena*. In contrast to *Boletus*, *Exidia* occurs mostly during the cold months.

**EXIF data.** Since the camera device and its settings affect the resulting image, the image classification models may be biased towards certain (e.g. more common) device attributes. To allow a deeper study of such phenomena, we include the EXIF data for approximately 84% of images, where the EXIF information was available in the Danish Fungal Atlas. The included attributes, the number of unique values in the dataset and the proportion of images with the attributes present are summarized in Table 4.

Attribute	Coverage [%]	# Values
White Balance	79.99	2
Color Space	84.38	3
Metering Mode	78.23	10
Metering Mode	81.04	13
Compressed Bits Per Pixel	37.75	88
Aperture Value	46.63	297
Device	80.31	688
Focal Length	80.12	1,580
Exposure Time	80.12	4,594
Shutter Speed Value	46.78	7,079

Table 4. Device settings extracted from the original images’ EXIF in the Danish Fungal Atlas, with the proportion of images where the attributes are present (Coverage), and the number of unique values of the attribute in the dataset.

## 6. Experiments

To establish a baseline performance on the DF20 and DF20-Mini datasets, we performed multiple experiments. First, we train a wide variety of well known CNN architectures such as ResNets [18], Efficient-Nets [43], MobileNet [37], Inception networks [41] and SE-ResNeXt-101-32x4d that extends the ResNet-101 by cardinality [50] and Squeeze and Excite blocks [21]. Second, the EfficientNet-B0, EfficientNet-B3, and SE-ResNeXt-101-32x4d are compared with Vision Transformer architectures ViT-Large/16 and ViT-Base/16 [7]. Finally, the impact of different metadata and their combinations on both the CNN and the ViT final prediction performance is evaluated.

### 6.1. Setup

In this section, we describe the full training and evaluation procedure, including the training strategy and image augmentations.

**Training Strategy.** All architectures were initialized from publicly available ImageNet-1k pre-trained checkpoints and further fine-tuned with the same strategy for 100 epochs with the PyTorch framework [35] within the 20.12 NGC deep learning framework Docker container. All neural networks were optimized by Stochastic Gradient Descent with momentum set to 0.9. The start Learning Rate (LR) was set to 0.01 and was further decreased with a specific adaptive learning rate schedule strategy - if the validation loss is not reduced from one epoch to another, reduce LR by 10%. To have the same effective mini-batch size of 64 for all architectures, we accumulated gradients from smaller mini-batches accordingly, where needed.

**Augmentations.** For training, we utilized several augmentation techniques from the Albumentations library [4]. More specifically, we used: random horizontal flip with 50% probability, random vertical flip with 50% probability, random resized crop with a scale of 0.8 - 1.0, random brightness / contrast adjustments with 20% probability, and mean and std. dev. normalization. All images were resized to the required network input size: For the CNN perfor-

mance experiment, inputs of size  $299 \times 299$  were used. In the case of the CNN vs ViT experiment, we used two different resolutions,  $224 \times 224$  and  $384 \times 384$ , to match the input resolutions of the pre-trained models.

**Test-time.** While testing, we avoided any extensive techniques such as ensembles, centre-cropping, prior weighting, etc. Only the resize and normalization operations were used to pre-process the data. The impact of test-time augmentation methods on the final performance can be studied in the future.

## 6.2. Metadata Use

We propose a simple baseline for the use of metadata. For a given type of metadata, e.g. month of year, substrate type or habitat, we adopt the following assumption:

$$P(I|S) = P(I|S, D), \quad (1)$$

i.e. that the visual appearance of a species does not depend on the metadata. This does not mean that the posterior probability of a species given an image is independent of metadata  $D$ . The dependence under the assumption is via the conditional probability  $P(S|D)$ . Note that this is the only possible approach not requiring to retrain or fine-tune the model using the metadata;<sup>6</sup> the model trained without metadata has no information about visual appearance changes of a species as a function of  $D$ . Moreover, this assumption is applicable for situations where the classifier has to be treated as a black box when the user has no access to the code or has no possibility to retrain the model.

A few lines of algebraic manipulation prove that under assumption Eq. (1), the class posterior given the image  $I$  and metadata  $D$  is easily obtained:

$$\begin{aligned} P(S|I, D) &= P(S|I) \frac{P(S|D)}{P(S)} \frac{P(I)}{P(I|D)} \\ &\propto P(S|I) \frac{p(S|D)}{p(S)}, \end{aligned} \quad (2)$$

where  $P(S)$  is the class prior in the training set. The conditional probability  $P(S|D)$  is estimated as the relative frequency of species  $S$  with metadata  $D$  in the training set.

To utilize several types of metadata at once, e.g. month and habitat, we combine the posteriors assuming statistical independence:

$$P(S|D_1, D_2) \propto \frac{P(S|D_1)P(S|D_2)}{P(S)}. \quad (3)$$

This is a simple, baseline assumption. Direct estimation of  $P(S|D_1, D_2)$ , e.g. as relative frequencies, is another possibility. The D20 benchmark has thus the potential to be

<sup>6</sup> More precisely "not requiring modelling the dependence of visual appearance and the metadata". The modelling need not necessarily mean training or fine-tuning the species classifier.

a fertile ground for evaluation of intra-metadata, as well as visual-metadata, dependencies.

The approach of Eq. (2) needs the output of the species classifier to be turned into an estimator of  $P(S|I)$ . The process is called *calibration* in the literature [16, 44]. CNNs typically need re-calibration as their estimates of  $\max P(S|I)$  are overconfident. The proposed benchmark may be used, in the context of exploiting metadata, to evaluate and compare classifier calibration techniques.

## 6.3. Metrics

Besides commonly used metrics, Top1 and Top3 Accuracy, we measured the macro-averaged  $F_1$  score,  $F_1^m$ , which is not biased by class frequencies and is more suitable for the long-tailed class distributions observed in the nature and FGVC datasets. Interestingly, the  $F_1^m$  metric is not used as a validation metric in any dataset mentioned in Table 1. The  $F_1^m$  is defined as the mean of class-wise  $F_1$  scores:

$$F_1^m = \frac{1}{N} \sum_{s=0}^N F_{1_s}, \quad (4)$$

where  $N$  represents the number of classes and  $S$  is the species index. Then the  $F_1$  score for each class is calculated as a harmonic mean of the class precision  $P_s$  and recall  $R_s$ :

$$F_{1_s} = 2 \times \frac{P_s \times R_s}{P_s + R_s}, \quad (5)$$

$$P_s = \frac{TP_s}{TP_s + FP_s}, \quad R_s = \frac{TP_s}{TP_s + FN_s} \quad (6)$$

In multi-class classification, the True Positive ( $TP$ ) represents the number of correct Top1 predictions, False Positive ( $FP$ ) how many times was a specific class predicted instead of the  $TP$ , and False Negative ( $FN$ ). Differently from the mean Accuracy,  $F_1^m$ , through precision and recall, it allows to easily assign a cost value to both types of error ( $FP/FN$ ) for each label and to measure more task-relevant performance. For example, in fungi recognition, mistaking a poisonous mushroom for the edible one is a more significant problem than the opposite. Interestingly, even though the performance across the whole taxonomy in nature-related FGVC datasets is highly demanded, most existing datasets are only using accuracy as the score measure. Considering that the datasets are highly imbalanced with so-called long-tail distribution, CNNs may ignore the least present species.

## 6.4. Results

In this section, we compare the performance of the well known CNN based models and ViT models in terms of Top1 and Top3 accuracy, and the newly included  $F_1^m$  metric. Additionally, we discuss the impact of the metadata on the classification performance.

	<b>Top1</b> [%]	<b>Top3</b> [%]	<b>F<sub>1</sub><sup>m</sup></b>	<b>Top1</b> [%]	<b>Top3</b> [%]	<b>F<sub>1</sub><sup>m</sup></b>
MobileNet-V2	65.74	83.65	0.546	69.51	84.55	0.602
ResNet-18	63.24	82.23	0.526	67.21	82.71	0.580
ResNet-34	63.60	81.68	0.522	69.92	84.72	0.605
ResNet-50	69.26	85.03	0.590	73.15	87.03	0.643
EfficientNet-B0	69.12	85.66	0.579	73.63	87.51	0.652
EfficientNet-B1	69.23	85.38	0.592	74.11	87.62	0.658
EfficientNet-B3	70.05	85.27	0.595	74.73	88.01	0.662
EfficientNet-B5	66.87	84.04	0.560	73.07	86.91	0.636
Inception-V3	65.30	82.83	0.530	71.45	85.64	0.622
InceptionResnet-V2	67.42	83.60	0.559	72.68	86.37	0.629
Inception-V4	67.50	83.63	0.572	74.19	87.63	0.655
SE-ResNeXt-101-32x4d	<b>72.39</b>	<b>86.57</b>	<b>0.635</b>	<b>76.73</b>	<b>89.09</b>	<b>0.691</b>
<b>Danish Fungi 2020 - Mini</b>			<b>Danish Fungi 2020</b>			

Table 5. Classification performance of selected CNN architectures on DF20 and DF20 - Mini. All networks share the settings described in Section 6.1 and were trained on  $299 \times 299$  images. The top results –  $F_1^m$ , see Eq. (4), equal to 0.635 / 0.691 and Top1 to 72.39% / 76.73% – are far from saturated. The datasets are challenging for the state-of-the-art CNN based classifiers.

**Convolutional Neural Networks.** Comparing well known CNN architectures on DF20 and DF20 - Mini, we can see a similar behaviour as on other datasets [6, 46, 48]. The EfficientNet models achieve excellent results. However, the high dependence of EfficientNet-B5 on the input size makes it worst than EfficientNet-B0 in the scenario with the  $299 \times 299$  image resolution. Similarly, the compact MobileNet-V2 performed better than ResNet-18, ResNet-34, and even better than Inception-V3. The best performing model on both datasets was SE-ResNeXt-101 with 0.635  $F_1^m$  score on DF20 - Mini and 0.691  $F_1^m$  score on DF20. A more detailed comparison of the achieved scores (Top1, Top3, and  $F_1^m$ ) for each model are summarized in Table 5.

**Vision Transformers.** The recently introduced ViT [7] showed excellent performance in common object classification compared to state-of-the-art convolutional networks. Appart from the CNNs, the ViT is not using convolutions but interprets an image as a sequence of patches and process it by a standard Transformer encoder as used in natural language processing [47]. To evaluate its performance for transfer-learning in the FGVC domain, we compare two ViT architectures - ViT-Base/16 and ViT-Large/16 - against the well performing CNN models - EfficientNet-B0, EfficientNet-B3 and SE-ResNeXt-101. As ImageNet pre-trained ViT models were available just for input sizes of  $224 \times 224$  and  $384 \times 384$ , we trained all networks on these resolutions while following the training setup fully described in subsection 6.1. Differently from the performance validation of Dosovitskiy et al. [7] on ImageNet, in our evaluation on DF20, ViTs ourperform state-of-the-art CNNs by a large margin. The best performing ViT model achieved an impressive 0.664  $F_1^m$  score while outperforming the SE-

		<b>Top1</b> [%]	<b>Top3</b> [%]	<b>F<sub>1</sub><sup>m</sup></b>	
DF20 - Mini	EfficientNet-B0	65.66	83.35	0.531	224 × 224
	EfficientNet-B3	66.90	83.49	0.537	
	SE-ResNeXt-101	69.48	85.58	0.593	
	ViT-Base/16	69.37	<b>86.54</b>	0.589	
	ViT-Large/16	<b>70.71</b>	86.51	<b>0.599</b>	
DF20	EfficientNet-B0	70.38	85.18	0.613	384 × 384
	ViT-Base/16	73.45	87.15	0.658	
	ViT-Large/16	<b>75.32</b>	<b>88.12</b>	<b>0.679</b>	
DF20 - Mini	EfficientNet-B0	70.22	85.69	0.596	
	EfficientNet-B3	72.09	87.17	0.624	
	SE-ResNeXt-101	72.34	87.53	0.631	
	ViT-Base/16	74.84	88.74	0.655	
	ViT-Large/16	<b>75.96</b>	<b>89.37</b>	<b>0.664</b>	
DF20	EfficientNet-B0	75.27	88.65	0.670	
	ViT-Base/16	79.40	90.93	0.724	
	ViT-Large/16	<b>81.25</b>	<b>91.93</b>	<b>0.747</b>	

Table 6. Classification results of selected CNN and ViT architectures on DF20 and DF20 - Mini dataset for two input resolutions.

ResNeXt-101 by a significant margin of 0.033 in  $F_1^m$ , and 3.62% of Top1 Accuracy on the images with  $384 \times 384$  input size. In the case of the  $224 \times 224$ , we see a smaller margin of 1.29% in Top1 Accuracy and 0.006 in the  $F_1^m$  score. All the results are shown in Table 6.

**Importance of the metadata.** Inspired by the common practice in mycology, we set up an experiment to show the importance of metadata for *Fungus* species identification. Using the approach described in Section 6.2, we improved

performance in all measured metrics by a significant margin. We measured the performance improvement with all metadata types and their combinations. Overall, habitat was most efficient in improving the performance. More precisely, we improved the ViT-Base/16 model’s performance on DF20 by 3.58%, 3.05% and 0.062 in Top1, Top3 and  $F_1^m$ , respectively. Similarly to DF20, we measured the relative performance gain of 2.88%, 1.65% and 0.047 in the Top1, Top3 and  $F_1^m$  metrics respectively on DF20 - Mini. Detailed evaluation of the performance gain using different observation metadata and their combinations is shown in Table 7.

**DF20 vs DF20 - Mini.** The performance evaluation with selected CNN and ViT architectures showed that even with a smaller number of classes and one-tenth of the data, DF20 - Mini as a compact subset of DF20 offers an even more challenging problem for state-of-the-art architectures while being less time and hardware demanding.

## 7. Conclusion

This paper introduced a novel fine-grained dataset and classification benchmark, the Danish Fungi 2020, and its subset, Danish Fungi 2020 - Mini. The dataset was constructed from data submitted to the Danish Fungal Atlas and includes 295,938 photographs of 1,604 species - mainly from the Fungi kingdom - together with full taxonomic labels, rich metadata, compact size and severe difficulty, and the same training and test set species distribution.

The quantitative and qualitative analysis of CNNs and ViTs shows superior performance of the ViT in fine-grained

	H	M	S	Top1 [%]	Top3 [%]	$F_1^m$
Danish Fungi 2020	×	×	×	73.45	87.15	0.658
	✓	✗	✗	+2.00	+1.42	+0.036
	✗	✓	✗	+1.37	+1.23	+0.024
	✗	✗	✓	+0.98	+0.96	+0.016
	✗	✓	✓	+2.30	+2.10	+0.039
	✓	✗	✓	+2.92	+2.41	+0.051
	✓	✓	✗	+3.16	+2.50	+0.056
	✓	✓	✓	<b>+3.58</b>	<b>+3.05</b>	<b>+0.062</b>
Danish Fungi 20 Mini	✗	✗	✗	69.37	86.54	0.589
	✓	✗	✗	+1.70	+1.10	+0.029
	✗	✓	✗	+0.77	+0.19	+0.011
	✗	✗	✓	+0.85	+0.69	+0.014
	✗	✓	✓	+1.29	+0.80	+0.020
	✓	✗	✓	+2.75	+2.01	+0.043
	✓	✓	✗	+2.20	+1.24	+0.037
	✓	✓	✓	<b>+2.88</b>	<b>+1.65</b>	<b>+0.047</b>

Table 7. Performance gains from *Fungus* observation metadata: H - Habitat, S - Substrate, M - Month, and their combinations, on DF20 and DF20 - Mini. ViT-Base/16 with image size 224 × 224.

classification. We present the baselines for processing the habitat, substrate and time (month) metadata. We show that – even with the baseline approach – utilizing the metadata increases the classification performance significantly. We provide the code and trained model check-points to all our baselines. A publicly available benchmark allows for an online comparison of state-of-the-art results for both image-only and image+metadata submissions. With the precise annotation and rich metadata, we would like to encourage further research in other areas of computer vision and machine learning, beyond fine-grained visual categorization. The benchmark may help research in classifier calibration, loss functions, validation metrics, taxonomy/hierarchical learning, device dependency or time series based species prediction. For example, the standard loss function focusing on recognition accuracy ignores the practically important cost of predicting a species with high toxicity.

## References

- [1] Suchet Bargoti and James Underwood. Image classification with orchard metadata. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5164–5170. IEEE, 2016. 3
- [2] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 2
- [3] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014. 3
- [4] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. 5
- [5] Danish Mycological Society. <https://svampe.databasen.org>, 2009. 3
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1, 2, 3, 7
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 5, 7
- [8] Jeffrey S Ellen, Casey A Graff, and Mark D Ohman. Improving plankton image classification using context metadata. *Limnology and Oceanography: Methods*, 17(8):439–461, 2019. 3
- [9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 1, 2
- [10] Tobias Guldborg Frøslev, Jacob Heilmann-Clausen, Christian Lange, Thomas Læssøe, Jens Henrik Petersen, Ulrik

- Søchting, Thomas Stjernegaard Jeppesen, and Jan Vesterholt. Danish mycological society, fungal records database, 2019. 3
- [11] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1349–1358, 2017. 1
- [12] Hervé Goëau, Pierre Bonnet, and Alexis Joly. Plant identification in an open-world (lifeclef 2016). In *CLEF working notes 2016*, 2016. 1
- [13] Herve Goeau, Pierre Bonnet, and Alexis Joly. Plant identification based on noisy web data: the amazing performance of deep learning (lifeclef 2017). 2017. 1, 2, 3
- [14] Wonjoon Goo, Juyong Kim, Gunhee Kim, and Sung Ju Hwang. Taxonomy-regularized semantic deep convolutional neural networks. In *European Conference on Computer Vision*, pages 86–101. Springer, 2016. 3
- [15] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 1
- [16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 6
- [17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1
- [18] 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. 5
- [19] Jacob Heilmann-Clausen, Hans Henrik Bruun, Rasmus Ejrnæs, Tobias Guldberg Frøslev, Thomas Læssøe, and Jens H. Petersen. How citizen science boosted primary knowledge on fungal biodiversity in denmark. *Biological Conservation*, 237:366 – 372, 2019. 3
- [20] Saihui Hou, Yushan Feng, and Zilei Wang. Vegfru: A domain-specific dataset for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 3
- [21] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5
- [22] Alexis Joly, Hervé Goëau, Hervé Glotin, Concetto Spampinato, Pierre Bonnet, Willem-Pier Vellinga, Robert Planqué, Andreas Rauber, Simone Palazzo, Bob Fisher, et al. Lifeclef 2015: multimedia life species identification challenges. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 462–483. Springer, 2015. 1, 2, 3
- [23] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization : Stanford dogs. 2012. 2
- [24] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *European Conference on Computer Vision*, pages 301–320. Springer, 2016. 2
- [25] Jonathan Krause, Michael Stark, Jia Deng, and Fei Fei Li. 3d object representations for fine-grained categorization. 11 2013. 2, 3
- [26] Alex Krizhevsky. Convolutional deep belief networks on cifar-10, 2010. 2
- [27] Neeraj Kumar, Peter N. Belhumeur, Arijit Biswas, David W. Jacobs, W. John Kress, Ida Lopez, and João V. B. Soares. Leafsnap: A computer vision system for automatic plant species identification. In *The 12th European Conference on Computer Vision (ECCV)*, October 2012. 2, 3
- [28] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 3
- [29] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015. 3
- [30] Jiongxin Liu, Angjoo Kanazawa, David Jacobs, and Peter Belhumeur. Dog breed classification using part localization. In *European conference on computer vision*, pages 172–185. Springer, 2012. 3
- [31] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3
- [32] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 2, 3
- [33] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. 1, 2, 3
- [34] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016. 1, 3
- [35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5
- [36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1
- [37] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted

- residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 5
- [38] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1857–1865, 2016. 1
- [39] Milan Sulc and Jiri Matas. Improving cnn classifiers by estimating test-time priors. *arXiv preprint arXiv:1805.08235v2*, 2019. 1
- [40] Milan Sulc, Lukas Picek, Jiri Matas, Thomas Jeppesen, and Jacob Heilmann-Clausen. Fungi recognition: A practical use case. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 3
- [41] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 5
- [42] Faezeh Tafazzoli, Hichem Frigui, and Keishin Nishiyama. A large and diverse dataset for improved vehicle make and model recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2017. 2, 3
- [43] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019. 5
- [44] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR, 2019. 6
- [45] 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, 3
- [46] Grant Van Horn, Oisin 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, 3, 7
- [47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 7
- [48] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1, 2, 3, 7
- [49] Davis Wertheimer and Bharath Hariharan. Few-shot learning with localization in realistic settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6558–6567, 2019. 1
- [50] 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. 5
- [51] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 3
- [52] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017. 3
- [53] Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning. In *Proceedings of the British machine vision conference*, 2019. 1
- [54] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8827–8836, 2018. 3
- [55] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking deep filter responses for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1134–1142, 2016. 3
- [56] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762, 2017. 3