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@snm.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 Atlas of Danish Fungi, 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 80.45% accuracy and 0.743 macro F1 score, reducing the CNN error by 9% and 12% respectively. A simple procedure for including metadata into the decision process improves the classification accuracy by more than 2.95 percentage points, reducing the error rate by 15%. 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 [7] database and its usage in the

ILSVRC\(^1\) challenge [42], together with PASCAL VOC [10] 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 [13, 14, 24], iNaturalist [53], CUB [55], and Oxford Flowers [37], have helped to develop and evaluate novel approaches to fine-grained domain adaptation [12], domain specific transfer learning [18], image retrieval [39, 44, 60], unsupervised visual representation [30, 34], few-shot learning [56], transfer learning [18] and prior-shift [45]. While the datasets have been extremely useful for the image recognition community, there are issues that limit their

\(^1\) The ImageNet Large Scale Visual Recognition Challenge.
relevance to real-world applications. We mention several such problems. Uniform class distribution, common in 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. A non-negligible percentage of noisy labels restricts quality assessment [3], and, despite CNN’s surprising robustness to label noise [26], may influence the perceived relative merit of learning algorithms. Some commonly used datasets [7, 27, 37] are saturated in accuracy or close to the point, leaving limited space for improvement in future research [3]. Extremely large dataset sizes might discourage researchers that do not have access to massive computational resources as experiments have become time-intensive 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 – Mycologists with specialization on specific Families/Genera. 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 a long-tailed distribution, which is common in nature. 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 provides 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 includes information related to the environment, place, time and full taxonomy labels and enables testing 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 most 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 benchmark for different scenarios, including visual-based, metadata focused or classifier-calibration related research. Besides 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 [8]. The recent ViT achieves excellent results in fine-grained classification outperforming the state-of-the-art CNN classifiers. We show that ViT performs 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 80.45% and 75.85% accuracy on DF20 and DF20 - Mini, respectively. We propose a simple method for processing the habitat, substrate and time (month) metadata, showing that – even with the simple approach – utilizing the metadata increases the classification performance significantly. To support and accelerate future research on the DF20 we open source the code.

2. Related Work

This section overviews existing fine-grained image datasets, which, unlike datasets with visually distinct object classes [10, 28], are characterized by small inter-class differences and huge intra-class similarity. Currently, there exists a number of FGVC dataset with a focus on plants [14, 22, 24, 29, 37], animals [2, 25, 52, 55], cars [27, 48] or airplanes [36]. The dataset statistics are compared in Table 1.

Table 1. Overview of publicly available FGVC datasets, nature-related (middle section) and other (top), and the number of images and categories. Images with “clean” (accurate) labels only.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Classes</th>
<th>#Training</th>
<th>#Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGVC-Aircrafts</td>
<td>102</td>
<td>6,732</td>
<td>3,468</td>
</tr>
<tr>
<td>Standford Cars</td>
<td>196</td>
<td>8,144</td>
<td>8,041</td>
</tr>
<tr>
<td>VMMRdb</td>
<td>9,170</td>
<td>291,752</td>
<td>×</td>
</tr>
<tr>
<td>Oxford Flowers</td>
<td>102</td>
<td>1,020</td>
<td>7,169</td>
</tr>
<tr>
<td>Stanford Dogs</td>
<td>120</td>
<td>12,000</td>
<td>8,580</td>
</tr>
<tr>
<td>DogSnap</td>
<td>133</td>
<td>4,776</td>
<td>3,575</td>
</tr>
<tr>
<td>LeafSnap</td>
<td>185</td>
<td>30,866</td>
<td>×</td>
</tr>
<tr>
<td>CUB 200-2011</td>
<td>200</td>
<td>5,994</td>
<td>5,784</td>
</tr>
<tr>
<td>VegFru</td>
<td>292</td>
<td>29,200</td>
<td>116,931</td>
</tr>
<tr>
<td>Birdsnap</td>
<td>500</td>
<td>47,386</td>
<td>2,443</td>
</tr>
<tr>
<td>NABirds</td>
<td>555</td>
<td>48,562</td>
<td>×</td>
</tr>
<tr>
<td>SnakeCLEF 2021</td>
<td>772</td>
<td>386,006</td>
<td>23,673</td>
</tr>
<tr>
<td>PlantCLEF 2015</td>
<td>1,000</td>
<td>91,758</td>
<td>21,446</td>
</tr>
<tr>
<td>iNaturalist 2017</td>
<td>5,089</td>
<td>579,184</td>
<td>95,986</td>
</tr>
<tr>
<td>PlantCLEF 2017†</td>
<td>10,000</td>
<td>230,658</td>
<td>25,629</td>
</tr>
<tr>
<td><strong>DF20 - Mini</strong></td>
<td>182</td>
<td>32,753</td>
<td>3,640</td>
</tr>
<tr>
<td><strong>DF20</strong></td>
<td>1,604</td>
<td>266,344</td>
<td>29,594</td>
</tr>
</tbody>
</table>

2[www.sites.google.com/view/danish-fungi-dataset](www.sites.google.com/view/danish-fungi-dataset)
Many datasets are artificially constructed to have a flat class distribution. Some datasets [14, 15, 22, 27, 41, 48] use web scraped data that may contain out-of-domain images, wrong labels and image duplicates.

Fungi species have been covered by image classification datasets. In the FGVCx 2018 Fungi classification challenge\(^3\), a dataset sampled from the same source as DF20 was used. The challenge dataset was smaller, scrambled the species names, did not include taxonomic labels and did not contain any metadata. The latest edition\(^4\) of iNaturalist dataset covers 3 kingdoms. From the Fungi kingdom, it includes 90,048 images of 341 species from 210 genera. DF20 is more fine-grained in the Fungi kingdom, and it is thus more challenging, with more than 1,500 fungi species. Many of these visually similar but from different genera or families (Figure 2).

**Labels.** As species-level labels are essential for usage in real-world applications, the tedious labelling procedure often relies 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 [7, 25, 55]. The main drawback of this approach is related to poor domain knowledge of the annotators that results in a high number of noisy labels [52] – 4.4% in CUB 200-2011 and approximately 4.0% for fine-grained classes in ImageNet. To address this issue, more recent datasets use citizen-science platforms and their users – citizen scientists\(^5\) – to label data with high-quality annotations [41, 52, 53].

**ImageNet Overlap.** Strict separation of training and test data is a core machine learning principle and it is standard in the field of image recognition. Nevertheless, some datasets contain ImageNet images in their test set, and thus fine-tuning from ImageNet weights contradicts the separation principle. This is commonly overlooked and may lead to biased (inflated) test set accuracies. For instance, a number of publications [4, 31, 32, 39, 58, 59, 61, 62, 63] with high impact used ImageNet weights and performed the fine-tuning and testing with the CUB 200-2011 [55] dataset that overlaps with the ImageNet in 43 out of 5,794 images (0.75%).

**Metadata.** Besides images and class labels, image classification datasets often provide additional metadata, such as higher taxon labels [24, 41, 53], label hierarchy [7, 22, 36], object parts and attribute annotations [24, 52, 55], masks [55], location [24, 41], and time of observation [24]. The existence of such metadata enables the usage of these datasets in machine learning research beyond image classification. For example [1, 9, 35] use location context, and [16] use taxonomy labels.

3. **Atlas of Danish Fungi**

The Atlas of Danish Fungi (Svampeatlas) \([6, 11, 20]\) is supported by more than 3,300 volunteers who have contributed more than 950,000 content-checked observations of fungi, many with expert-validated class labels, submitted mostly since 2009.

The project has resulted in a vastly improved knowledge of fungi in Denmark\([20]\). More than 180 species belonging to Basidiomycota\(^6\) have been added to the list of known Danish species\([20]\), and several species that were considered extinct have been re-discovered\([21]\). Simultaneously, several search and assistance functions have been developed.

---

\(^3\)https://github.com/visipedia/fgvcx_fungi_comp
\(^4\)https://www.kaggle.com/c/inaturalist-2021
\(^5\) Domain specific nonprofessional enthusiasts - experts.

---

\(^6\) a group of fungi that produces their sexual spores (basidiospores) on a club-shaped spore-producing structure (basidium).
that present features relating to the individual species \[21\], making it much easier to include an understanding of endangered species in nature management and decision-making.

Expert-validated Svampeatlas records are published in the Global Biodiversity Information Facility (GBIF), weekly, since 2017. As of end of July 2021, GBIF included 438,872 such images \[38\].

### 3.1. Annotation Process

The Atlas of Danish Fungi uses an interactive labelling procedure 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. Since 2019, the Atlas of Danish Fungi observation identification has been streamlined thanks to an image recognition system \[46\].

### 4. Dataset Description

The Danish Fungi 2020 (DF20) dataset contains image observations from the Atlas of Danish Fungi belonging to species with more than 30 images. The data are observations collected before the end of 2020 \(^7\), originating from 30 countries, and including samples from all seasons. It consists of 295,938 images belonging to 1,604 species mainly from the Fungi kingdom with a few visually similar species (See Figure 3) from Protozoa (1.7% classes / 1.1% images) and Chromista kingdoms (0.06% classes / 0.03% images) kingdoms, covering 566 genera and 190 families. The most frequent species – Trametes versicolor – is represented by 1,913 images and the least present with 31.

Additionally, we hand-picked 6 genera with a similar visual appearance, containing 36,393 images belonging to 182 species. 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 chose all species from six commonly known genera of fungi forming fruit-bodies of the toadstool type with a large number of species: (Russula, Boleteus, Amanita, Clitocybe, Agaricus and Mycenae) for the construction of the DF20 - Mini. The most frequent species in the DF20 - Mini dataset – Mycena galericulata – has 1,221 images, the rarest have 31 samples. For a quantitative summary of the data selection, see Table 2.

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

### 4.1. Metadata

Unlike most computer vision datasets, DF20 and DF20 - Mini include rich metadata acquired by citizen-scientists in the field while recording the observations. 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 a detailed list see Table 4.

**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 uncommon to find it on other substrates. For example, Trametes occurs only on wood and Russula on soil. As such metadata is crucial for the final categorization capability. We provide 32 substrate types, \( \text{e.g.,} \) wood of living trees, dead wood, soil, bark, stone, fruits, mosses 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 that some species occur in deciduous forests rather than in

\(^7\)Including 3 preserved specimens collected in 1874, 1882, and 1887, recently photographed.
conifer forests or plantations, while others grow in farmland. For a deeper understanding of such relation, we include the information about the 32 different habitats.

**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 4 we show the probability of three genera (Mycena, Boletus and Exidia) being observed in different months of the calendar year and extracted from the DF20 dataset. 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.** 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 Atlas of Danish Fungi. The included attributes, the number of unique values in the dataset and the proportion of images with the attributes present are summarized in Table 3.

**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 lives, we include multi-level location information. Starting from latitude and longitude and their uncertainty, we further extracted information about the country, region, district, and geographic localities. We cover data from 30 countries and 9,003 geographic localities.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Coverage [%]</th>
<th># Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compressed Bits Per Pixel</td>
<td>37.75</td>
<td>88</td>
</tr>
<tr>
<td>Aperture Value</td>
<td>46.63</td>
<td>297</td>
</tr>
<tr>
<td>Shutter Speed Value</td>
<td>46.78</td>
<td>7,079</td>
</tr>
<tr>
<td>Metering Mode</td>
<td>78.23</td>
<td>10</td>
</tr>
<tr>
<td>White Balance</td>
<td>79.99</td>
<td>2</td>
</tr>
<tr>
<td>Exposure Time</td>
<td>80.12</td>
<td>4,594</td>
</tr>
<tr>
<td>Focal Length</td>
<td>80.13</td>
<td>1,580</td>
</tr>
<tr>
<td>Device</td>
<td>80.31</td>
<td>688</td>
</tr>
<tr>
<td>SceneCaptureType</td>
<td>81.04</td>
<td>13</td>
</tr>
<tr>
<td>Color Space</td>
<td>84.38</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 3. Device settings extracted from the original image EXIFs in the Atlas of Danish Fungi, with the proportion of images where the attributes are present (Coverage), and the number of unique values of the attribute in the dataset.

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

Table 4. Description of the provided metadata (observation attributes). For almost all images, a detailed information about taxonomy, location, time, habitat and substrate type is included.
5. 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 Inception networks [47], ResNets [19], MobileNet [43], SE-ResNeXt-101-32x4d [23, 57], EfficientNets [49], and EfficientNetV2-L [50]. Second, EfficientNets and SE-ResNeXt-101-32x4d are compared with Vision Transformer architectures ViT-Base/16 and ViTLarge/16 [8]. Finally, the impact of different metadata and their combinations on both the CNN and the ViT final prediction performance is evaluated.

5.1. Setup

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

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 [40] within the 21.05 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. 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 [5]. 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 performance experiment, inputs of size 299×299 were used. In the case of the CNN vs ViT experiment, we used two different resolutions, 224×224 and 384×384, to match the input resolutions of the pre-trained models.

Test-time. While testing, we avoided any additional 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.

5.2. Metadata Use

We propose a simple method for the use of metadata to improve the categorization performance – similar to spatiotemporal prior used in [2]. For a given type of metadata (D) and image (I), we adopt the following assumption for the likelihood of an image observation:

\[ P(I|S) = P(I|S, D), \]

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.

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:

\[ P(S|D) = P(S|D) \frac{P(S|D) P(I)}{P(I|D) P(I)} \]

\[ \propto P(S|D) \frac{P(S|D)}{P(S)}, \]

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

While we know this assumption is not always true in practice, since metadata like substrate or time in fact do impact the image background as well as the appearance of the specimen, this is the only possible approach not requiring modelling the dependence of visual appearance and the metadata. 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 without the possibility to retrain the model. Even this simplistic model based on an unrealistic assumption reduces error rates, see Table 7.

With multiple metadata at once, e.g., substrate and habitat or substrate and month, we combine the posteriors assuming statistical independence:

\[ P(S|D_1, D_2) \propto \frac{P(S|D_1) P(S|D_2)}{P(S)}, \]

This is a simple, baseline assumption, which again may not always be valid for related meta-data. 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 a fertile ground for evaluation of intra-metadata, as well as visual-metadata, dependencies.

The approach of Eq. (2) needs a probabilistic classifier to serve as an estimator of \( P(S|I) \). In our experiments, we use the outputs of the softmax layer. Note that for CNNs, the estimates of \( \max P(S|I) \) are typically overconfident,
and the quality of the estimator can be improved by a process called calibration in the literature [17, 51]. The proposed benchmark may be used, in the context of exploiting metadata, to evaluate and compare classifier calibration techniques.

5.3. Metrics

Besides commonly used metrics, Top1 and Top3 accuracy, we measured the macro-averaged F1 score, $F^n_{1}$, which is not biased by class frequencies and is more suitable for the long-tailed class distributions observed in nature. Interestingly, even though the performance across the whole taxonomy is highly demanded in nature-related applications, most existing benchmarks are only using accuracy as the score measure. Considering that the datasets are highly imbalanced with long-tail distribution, learning procedure may ignore the least present species. Additionally, usage of $F^n_{1}$ allows to easily assign a cost value to both types of error ($fp$ and $fn$) 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.

The $F^n_{1}$ is defined as the mean of class-wise F1 scores:

$$F^n_{1} = \frac{1}{N} \sum_{S=1}^{N} F_{1S},$$

(4)

where $N$ represents the number of classes and $S$ is the species index. Than the F1 score for each class is calculated as a harmonic mean of the class precision $P_S$ and recall $R_S$:

$$F_{1S} = \frac{2 \times P_S \times R_S}{P_S + R_S},$$

(5)

<table>
<thead>
<tr>
<th>Top1</th>
<th>Top3</th>
<th>$F^n_{1}$</th>
<th>Top1</th>
<th>Top3</th>
<th>$F^n_{1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MobileNet-V2</td>
<td>65.58</td>
<td>83.65</td>
<td>0.550</td>
<td>69.77</td>
<td>85.01</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>62.91</td>
<td>81.65</td>
<td>0.514</td>
<td>67.13</td>
<td>82.65</td>
</tr>
<tr>
<td>ResNet-34</td>
<td>65.63</td>
<td>83.52</td>
<td>0.559</td>
<td>69.81</td>
<td>84.76</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>68.49</td>
<td>85.22</td>
<td>0.587</td>
<td>73.49</td>
<td>87.13</td>
</tr>
<tr>
<td>EfficientNet-B0</td>
<td>67.94</td>
<td>85.71</td>
<td>0.567</td>
<td>73.65</td>
<td>87.55</td>
</tr>
<tr>
<td>EfficientNet-B1</td>
<td>68.35</td>
<td>84.67</td>
<td>0.572</td>
<td>74.08</td>
<td>87.68</td>
</tr>
<tr>
<td>EfficientNet-B3</td>
<td>69.59</td>
<td>85.55</td>
<td>0.590</td>
<td>75.69</td>
<td>88.72</td>
</tr>
<tr>
<td>Inception-V3</td>
<td>65.91</td>
<td>82.97</td>
<td>0.535</td>
<td>72.10</td>
<td>86.58</td>
</tr>
<tr>
<td>Inception-ResNet-V2</td>
<td>64.67</td>
<td>81.42</td>
<td>0.542</td>
<td>74.01</td>
<td>87.49</td>
</tr>
<tr>
<td>Inception-V4</td>
<td>67.45</td>
<td>82.78</td>
<td>0.560</td>
<td>73.00</td>
<td>88.87</td>
</tr>
<tr>
<td>EfficientNetV2-L</td>
<td>70.77</td>
<td>86.48</td>
<td>0.595</td>
<td>77.43</td>
<td>89.65</td>
</tr>
<tr>
<td>SE-ResNeXt-101</td>
<td>72.23</td>
<td>87.28</td>
<td>0.620</td>
<td>77.13</td>
<td>89.48</td>
</tr>
</tbody>
</table>

Table 5. Classification performance of selected CNN architectures on DF20- Mini and DF20. All networks share the settings described in Section 5.1 and were trained on 299×299 images. The top results – $F^n_{1}$, see Eq. (4), equal to 0.620 / 0.693 and Top1 to 72.23% / 77.43% – are far from saturated. The datasets are challenging for the state-of-the-art CNN classifiers.

$$P_S = \frac{tp_S}{tp_S + fp_S}, \quad R_S = \frac{tp_S}{tp_S + fn_S}.$$  

(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 ($fn$) how many images of class $S$ have been misclassified.

5.4. 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 $F^n_{1}$ metric. Additionally, we discuss the impact of the metadata on the classification performance and differences in performance between CNNs and ViTs.

Convolutional Neural Networks. Comparing well known CNN architectures on DF20 and DF20-Mini, we can see a similar behaviour as on other datasets [7,53, 55]. The best performing model in the $F^n_{1}$ score was SE-ResNeXt-101 with 0.620 $F^n_{1}$ score on DF20-Mini and 0.693 $F^n_{1}$ score on DF20. EfficientNetV2-L achieved slightly better accuracy of 77.43% on the DF20 dataset. A detailed comparison of the achieved scores (Top1, Top3, and $F^n_{1}$) for each model are summarized in Table 5.

Vision Transformers. The recently introduced ViT [8] showed excellent performance in object classification compared to state-of-the-art convolutional networks. Unlike CNNs, the ViT is not using convolutions, but interprets an image as a sequence of patches and processes it by a standard Transformer encoder as used in natural language processing [54]. 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-
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 5.2, 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. With the combination of habitat, substrate and month, we improved the ViT-Large/16 model’s performance on DF20 by 2.95%, 1.92% and 0.053 in Top1, Top3 and Fm score, respectively, and the performance of the ViT-Base/16 model by 3.81%, 2.84% and 0.070 in Top1, Top3 and Fm. Detailed evaluation of the performance gain using different observation metadata and their combinations is shown in Table 6.

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.

6. 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 Atlas of Danish Fungi and labeled by mycologists. It 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 classification. We present the baselines for processing the habitat, substrate and time (month) metadata. We show that – even with the simple method from Section 5.2 – utilizing the metadata increases the classification performance significantly. We provide the code and trained model checkpoints to all our baselines. A publicly available web benchmark allows – through CSV submission file – for an on-line 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 research in other areas of computer vision and machine learning, beyond fine-grained visual categorization. The datasets may serve as a benchmark for 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.

Acknowledgements

LP was supported by the Ministry of Education, Youth and Sports of the Czech Republic project No. LO1506, and by the grant of the UWB project No. SGS-2019-027. MŠ and JM were supported by Toyota Motor Europe and by CTU grant SGS20/171/OHK3/3T/13.
References


[27] Jonathan Krause, Michael Stark, Jia Deng, and Fei Fei Li. 3d object representations for fine-grained categorization. 11 2013.


