Say No to the Poisonous Fungi: An Effective Strategy for Reducing 0-1 Cost in FungiCLEF2024

Bao-Feng Tan[1,†], Yang-Yang Li[1,†], Peng Wang[1,†], Lin Zhao[1] and Xiu-Shen Wei[2,*]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[2]School of Computer Science and Engineering, and Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, Southeast University, Nanjing, China

Abstract

The FungiCLEF2024 competition endeavors to precisely identify fungi species leveraging both metadata and image analysis. Pivotal to the success of this competition are two crucial evaluation metrics: minimizing the error rate and the 0-1 cost loss resulting from misclassification. To reduce the identification error rate, we introduce a Dynamic MLP framework, drawing inspiration from [1]. This approach effectively integrates image and metadata embeddings through recursive blocks, utilizing matrix multiplication for deep fusion of information. To further address the issue of 0-1 cost, we devise a novel probability-based screening strategy, which initially consolidates poisonous fungi categories into a single class, then employs marginal expected loss and a threshold parameter α to optimize the recall rate for poisonous species. These approaches significantly reduce the error rate and 0-1 cost associated with misclassification and achieve a score of 0.5548 on the private leaderboard, securing the third-place ranking. The code is available at https://github.com/bftan1949/FungiCLEF2024.

Keywords

Fine-grained image recognition, Open-Set, 0-1 cost loss, Fungi Species Identification

1. Introduction

Fine-grained visual classification, as a core challenge in the field of computer vision and pattern recognition, plays a pivotal role in diverse practical applications [2]. The FungiCLEF2024 [3] challenge, serving as a crucial component of the LifeCLEF2024 [4], aims to promote and incentivize in-depth research on fungi identification algorithms, particularly in complex scenarios that integrate image and metadata inputs. The achievement of this goal not only holds immense value for biodiversity conservation, but also plays a crucial role in maintaining human health.

Prior FungiCLEF challenges have achieved significant progress through deep learning models [5, 6, 7, 8, 9, 10, 11]. To further enhance the practical significance of the competition and effectively address the challenges faced by developers, scientists, users, and the community, this year's organizers have introduced additional constraints. Therefore, the challenges faced by this year's competition can be summarized as follows:

- **Fine-grained Image Recognition:** As a persistent challenge in computer vision, fine-grained image analysis requires participants to conduct deeper research and technological innovations.
- **Feature Fusion:** Effectively fusing metadata features with intuitive image features is crucial, especially when it comes to fine-grained distinction. Relying on subtle cues from metadata to differentiate closely related categories becomes paramount.
- **Open-set Recognition:** Open-set recognition directly affects the robustness and security of artificial intelligence systems. The FungiCLEF2024 challenge specifically includes a large amount of open-set data with unknown categories in the test set.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
†Under the supervision of Xiu-Shen Wei.
tanbf@njust.edu.cn (B. Tan); lyylyyi599@njust.edu.cn (Y. Li); wangpeng@njust.edu.cn (P. Wang); linzhao@njust.edu.cn (L. Zhao); weixs.gm@gmail.com (X. Wei)

Table 1
FungiCLEF2024 dataset statistics for each split.

Subset	Species	Known Species	Unknown Species	All Images	Poisonous Images
Training	1,604	1,604	-	295,938	12,977
Validation	2,713	1,084	1,629	60,832	2,532

- **0-1 Cost Loss:** One of the core issues in FungiCLEF2024, which categorizes fungi into poisonous and non-poisonous, is how to construct a model that minimizes the misclassification of poisonous fungi as non-poisonous to ensure high reliability and safety in identification results.
- **Hardware Constraints:** All algorithms will be executed on the HuggingFace platform, subject to strict limitations of 16GB of GPU memory and a two-hour runtime.

The FungiCLEF2024 dataset is based on data collected through the Atlas of Danish Fungi mobile and Web applications. All fungi specimen observation had to pass the expert validation process, therefore guaranteeing high-quality labels. For the training dataset, it contains 295,938 images - belonging to 1,604 species. The validation dataset contains 60832 images belonging to 2,713 species, 1,084 known from the training set and 1,629 unknown species. The dataset statistics are listed in Table 1.

For fine-grained image recognition and feature fusion, this paper employs Dynamic MLP [1] to fully fuse diverse feature information. Additionally, for open-set recognition, an entropy-based approach is utilized, leveraging the model's prediction confidence through entropy to identify open-set images, surpassing previous methods. Furthermore, to minimize the critical 0-1 cost loss caused by misclassifying poisonous fungi as non-poisonous, this paper proposes an easy but quite effective way to mitigate 0-1 cost by utilizing a marginal expected loss function during training, which significantly reduces the cost loss while maintaining accuracy. Details of methods will be discussed in Section 3.

The subsequent sections are organized as follows: In Section 2, we will provide a detailed explanation and interpretation of the dataset and evaluation metrics used in the competition. Section 3 will outline the methodology and core concepts adopted in this paper. Section 4 will focus on presenting our experimental results and actual performance in the competition. Finally, in Section 5, we will provide a comprehensive review and summary of the entire content.

2. Related Work and Evaluation Metrics

2.1. Related Work

Fine-grained image classification: To enhance fine-grained image classification, several approaches have been proposed. For instance, [12, 13, 14, 15, 16] detect the discriminative regions of an image to exploit subtle details. SnapMix [17] utilizes the class activation map (CAM) [18] to mitigate label noise in fine-grained data augmentation. Similarly, Attribute Mix [19] focuses on semantically meaningful attribute features from two images to identify the same super-categories. FixRes [20] investigates data augmentation and resolution strategies to boost classification performance. Other studies focus on extracting more valuable features from multi-channel networks [21] or through contrastive learning [22].

Using additional information: Besides visual information, researchers have incorporated additional information to enhance classification performance. Many existing works [23, 24, 25, 26] combine the image features with additional multi-modal features directly through channel-wise concatenation. Multi-modal features, including images, ages, and dates, were first introduced by Tang et al. [26], who concatenated them from an MLP backbone network to make a joint prediction. Subsequently, Minetto et al. [24] introduced metadata to the geo-spatial land classification task. Further, Salem et al. [25] integrated dense overhead imagery with location and date into a general framework by concatenating

the outputs of the context network.

Open-set recognition: Discriminative models are one of the most important ways for open-set recognition [27]. Traditional methods, such as 1-vs-Set machines based on SVM [28], often suffer from limitations stemming from the weak feature extraction ability of those traditional models. In recent years, deep learning-based methods have garnered increasing attention due to their powerful representation abilities. Bendale et al. [29] first proposed replacing the softmax layer in the network with OpenMax, which calibrates the output probability using the Weibull distribution. A similar work [30] replaced the softmax layer with one-vs-rest units. These methods have pioneered a new direction for the research of open-set recognition.

Previous FungiCLEF work: Most contributions to FungiCLEF2023 were centered on modern Convolutional Neural Network (CNN) or transformer-inspired architectures, such as MetaFormer [31], Swin Transformer [32], and Volo [33]. The winning team [34] achieved 79.28% accuracy using MetaFormer. These results were often enhanced by combining predictions from the same observation and through data augmentations applied during both training and testing. Techniques such as Seesaw loss [35], Focal loss [36], Arcface loss [37], and Sub-Center loss [38] have achieved great success in addressing the unbalanced class distribution. Additionally, metadata was combined with image features to classify fungi categories, further improving the overall performance.

2.2. Evaluation Metrics

FungiCLEF2024 has set a total of 3 evaluation metrics, namely Track1, Track2, and Track3, which will be introduced below.

Track1: The first metric is the standard classification error, which is the average error in predicting class labels. All categories not present in the training set should be correctly classified as the "unknown" category (i.e., labeled as -1). The specific calculation formula is as follows:

$$\text{Track1}(y, q(x))) = \begin{cases} 0 & \text{if } q(x) = y \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

Here, x represents the input image, y represents the true label of the input image, and q represents the trained classification model.

Track2: The second metric is the cost loss associated with confusing non-toxic and toxic species. Define $\mathrm{d}(\cdot)$ as an indicator function, where if $\mathrm{d}(y) = 1$, it indicates that category y is a toxic category, and if $\mathrm{d}(y) = 0$, it indicates that category y is a non-toxic category. The specific calculation formula for Track2 is as follows:

$$\text{Track2}(y, q(x)) = \begin{cases} 0 & \text{if } \mathrm{d}(q(x)) = \mathrm{d}(y) \\ c_{PSC} & \text{if } \mathrm{d}(q(x)) = 0 \ \ \& \ \ \mathrm{d}(y) = 1 \\ c_{ESC} & \text{if } \mathrm{d}(q(x)) = 1 \ \ \& \ \ \mathrm{d}(y) = 0 \end{cases} \tag{2}$$

In this competition, $c_{PSC} = 100$ and $c_{ESC} = 1$.

Track3: The third metric is the sum of Track1 and Track2. The specific formula is as follows:

$$\text{Track3} = \text{Track1} + \text{Track2} \tag{3}$$

The final ranking of the competition is based on the performance of Track3. Regardless of whether it is Track1, Track2, or Track3, the lower the score, the higher the ranking.

3. Method

In this section, we will introduce our method to handle the recognition problem with open-set and raise an easy way to decrease the Track2 significantly.

3.1. Fine-grained Image Recognition with Feature Fusion

Feature fusion. To enhance the image representation and improve the result of image classification, "Dynamic MLP" proposed in [1] is applied to fully release the potential of the meta information. Marking the input image feature as z_i and the meta feature as z_e respectively, Dynamic MLP is designed to fuse features through a matrix multiplication operation. Specifically, given the input image x_i, we can obtain the image feature through the backbone network[1], and the meta (such as substrate and habitat) feature through a well pre-trained clip text model [39], which can be described as following:

$$z_i = \text{Backbone}(x_i) \tag{4}$$

$$z_e = \text{MLP}(\text{Cat}(\text{Clip}(x_e^{sub}), \text{Clip}(x_e^{hab}))) \tag{5}$$

where Backbone(·) denotes the model before the classification head, and $z_i \in \mathbb{R}^n$, n is the output dimension. x_e^{sub} and x_e^{hab} denote the substrate and habitat data, respectively. Clip denotes a well trained clip text model. Cat(·) denotes the channel-wise concatenation. MLP denotes a residual MLP network, following the descriptions in PriorsNet [40]. After MLP, the meta feature is projected into the same dimension as z_i, i.e., $z_e \in \mathbb{R}^n$. Then, let the original z_i as z_i^0, Dynamic MLP takes z_i^0 and z_e as initial inputs, and the enhanced image representation z_i^N is obtained after N recursive blocks. At last, z_i^N is expanded to align the shape with z_i^0 by a channel-increasing layer for classifying images. The process of Dynamic MLP can be specifically summarized into three steps:

1. Taking into image and meta features and reshaping the meta feature from a 1-d vector to a 2-d matrix, which can be formalized as following:

$$W = \text{Reshape}(f(z_e)) \tag{6}$$

   where Reshape(·) denotes reshape operation, and f denotes a fully connected layer.
2. Obtaining the enhanced image feature z_i^N after N recursions are completed.

$$z_i^{n+1} = \text{ReLU}(\text{LN}(f(W@z_i^n))), n = 0, 1, ..., N \tag{7}$$

   ReLU(·) and LN(·) denote ReLU activation function and layer normalization, respectively. The operator @ denotes the matrix multiplication.
3. Aligning the dimension of z_i^N and z_i^0.

$$\hat{z}_i^N = \text{Layer}(z_i^N) \tag{8}$$

   Layer(·) denotes a channel-increasing layer.

Fine-grained image classification. After the final enhanced image feature $\hat{z}_i^N$ is obtained, we can use it to classify the fine-grained Fungi images:

$$logits = \text{Head}(\hat{z}_i^N) \tag{9}$$

where Head is the last classification head which is used for recognizing images.

[1] In CNN backbones, a image feature are acquired after a pooling layer. In Vit based models, a image feature is the [CLS] token in the last layer.

Table 2
Comparison of different open-set methods on Fungi dataset. The model and training details of all methods are exactly the same during training, and the results are reported on the validation set.

Method	Top1
OpenMax [29]	61.37
OSRCI [41]	61.58
Entropy(Ours)	**62.19**

3.2. Entropy Based Open-set Identifier

In the Fungi competition, models are not only asked to correctly identify the species in the close-set, but also the pictures in the open-set. As shown in Table 2, existing approaches such as [29, 41] show their superiority on coarse-grained datasets, but are not suitable for large scale fine-grained datasets. So we adopt an easy entropy based method to identify open-set images, which shows better result than [29, 41].

Entropy is defined following to measure the quality of the probability distribution:

$$\text{entropy}(p) = -\sum_{c=1}^{C} p^c \log p^c \tag{10}$$

$$\text{where } p = \text{Softmax}(logits) \tag{11}$$

where Softmax(·) denotes the softmax operation, C denotes the number of categories to be classified in the close-set and p^c denotes the the probability of being identified as the category c. In general, the model is more confident for known categories, corresponding to a lower entropy. Whereas for unknown categories the uncertainty is higher and hence the entropy will be higher. Thus, we can effectively distinguish between known/unknown categories through a entropy threshold τ. Once the threshold τ is determined, we can use the following formula to identify the open-set images:

$$label = \begin{cases} -1 & \text{if entropy}(p) > \tau, \\ \text{Argmax}(p) & \text{if entropy}(p) \leq \tau. \end{cases} \tag{12}$$

where Argmax(·) denotes the argmax function. Since the choice of τ determines the effect of a model, we find the best threshold based on the validation set.

3.3. Probability-guided Poisonous Recognizer

The two tasks of the competition are improving the classification accuracy and reducing the cost of classifying poisonous as non-poisonous, respectively. In fact, the latter task has a greater impact on the final score than the former. In this section, we will introduce an easy but quite effective way to reduce the cost of the latter task.

Put poisonous categories together. The Fungi dataset is a long-tail dataset with uneven distribution, Figure 1 illustrates that the number of categories and the quantity of images for the poisonous are significantly lower compared to edible ones. This imbalance poses a challenge for models to adequately learn robust embeddings for poisonous species, thereby leading to misclassification where poisonous species may be incorrectly labeled as edible. Such errors incur substantial costs. Therefore, we first put all poisonous categories into a single class, effectively reducing the total number of categories from 1604 to 1556 (comprising 1555 edible species and 1 aggregated poisonous class). This approach facilitates the model's focus on the general features of poisonous species, alleviating the need to discern subtle distinctions. Next, two models will be trained separately to classify mixed categories (1555 edible categories and a poisonous category) and only poisonous categories (49 poisonous categories).

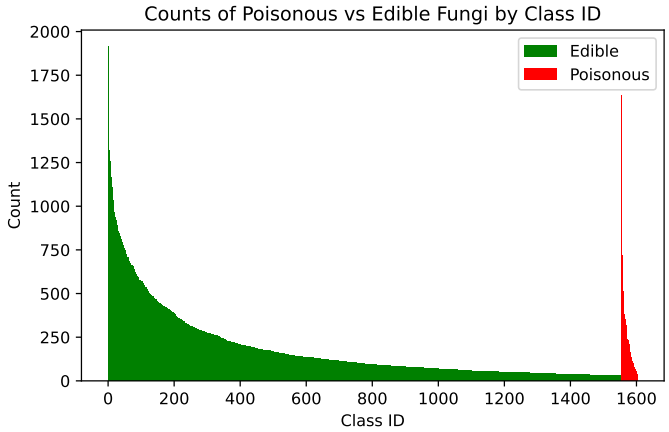

Figure 1: The distribution of the training dataset. The classes are arranged in descending order based on the number of samples in each category.

Marginal expected loss. The most direct way to reduce cost is to optimize the cost function itself, but since the calculation of cost is discrete, we use the marginal expected loss function here:

$$MEloss = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\sum_{c=1}^{C} p_i^c \cdot cost^c_{\text{gt}(i)} \tag{13}$$

where $\mathcal{I}$ represents the training images and gt(·) denotes the ground truth label of the image. As defined in Section 2, d(·) indicate poisonous species, where $\text{d}(y) = 1$ if the category y is poisonous, and $\text{d}(y) = 0$ if y is edible. According to the calculation formula of Track2, we can obtain the specific expression of $cost^c_{\text{gt(i)}}$, however, according to Track2, no matter what the ground truth of the picture is, as long as the predicted label meets $\text{d}(\text{gt}(i)) = \text{d}(c)$, the same loss will be obtained, thus disrupting the correct gradient descent direction to the ground truth. Hence, We amend the expression of $cost^c_{\text{gt(i)}}$ to introduce a penalty specifically for instances where $\text{d}(\text{gt}(i)) = \text{d}(c)$ while $\text{gt}(i) \neq c$, to address this issue. The formula is as following:

$$cost^c_{\text{gt}(i)} = \begin{cases} 0 & \text{if gt}(i) = c \\ 5 & \text{if d}(\text{gt}(i)) = \text{d}(c) \ \ \& \ \ \text{gt}(i) \neq c \\ 10 & \text{if d}(\text{gt}(i)) = 0 \ \ \& \ \ \text{d}(c) = 1 \\ 100 & \text{if d}(\text{gt}(i)) = 1 \ \ \& \ \ \text{d}(c) = 0 \end{cases} \tag{14}$$

MEloss will force the model to improve its ability to identify the poisonous and the edible while improving the recognition accuracy.

Increasing the recall rate of the poisonous. The primary aim of reducing Track2 lies in maximizing the recall rate for the poisonous species. To achieve this, we set a probability-guided threshold α, where an image is considered poisonous as soon as the predicted probability of the poisonous category exceeds α. As previously discussed, let define the model classifying mixed categories as h and the model classifying poisonous categories as g, the ultimate decision method is outlined as Algorithm 1.

4. Experiments

In this section, we will introduce the details and main results in detail.

Algorithm 1 Fungi's main inference algorithm

input: models of h and g, threshold τ, α

for image $\{x_i\}_{i \in \mathcal{I}}$ **do**
  $p_i = h(x_i)$
  if p_i > τ **then**
    return -1 #-1 represents the open-set categories
  else if $p_i^{poi} \leq \alpha$ **then**
    return Argmax(p_i)
  else
    $p_i = g(x_i)$
    return Argmax(p_i)
  end if
end for

4.1. Implementation Details

Basic settings. The thresholds of τ and α are 1.5 and 0.01 respectively. The proposed method has been developed utilizing the PyTorch framework [42]. Vit-large [43] and Eva02-large [44], implemented via the timm library [45], serve as the model h and g, respectively. All the models have been pre-training on the ImageNet dataset [46], and are conveniently accessible in HuggingFace. Fine-tuning of these models was performed using 8 Nvidia RTX3090 GPUs. Input size of images is 336. The initial learning rate was set to 2×10^{-5}, and the total number of training epochs was set to 15, with the first epoch dedicated to warm-up by employing a learning rate of 2×10^{-7}. For optimal model training, we employed the AdamW optimizer [47] in conjunction with a cosine learning rate scheduler [48], with the weight decay set to 1×10^{-2}. Since the Fungi dataset with an unbalanced distribution, many studies have suggested solutions [49, 50, 51, 52], considering the practical effect, we finally use seesaw loss [35] and ME loss mentioned above to optimize the model.

Data Augmentations. We employ a composed sequence of common augmentation techniques to enhance results. During training, we first perform random cropping on the image, where the size of the cropped region is randomly chosen between 50% and 100% of the original image size. Subsequently, the slice is resized using the bicubic interpolation method and flipped horizontally and vertically with a probability of 50%. Additionally, we incorporate hue-saturation and brightness-contrast augmentations to randomly adjust the hue, saturation, value, brightness, and contrast of the input image. Finally, standard normalization is applied to all input images. However, due to limitations in both running time and GPU memory on HuggingFace, test-time augmentations are simplified by first resizing images to 336 and then normalizing them using the same mean and std as during training.

4.2. Fungi Dataset Experiments

The key experimental results are presented in Table 4. As evident from the table, Dynamic MLP exhibits superior feature fusion ability, effectively reducing the error rate for Vit and Eva. Concurrently, the strategy of consolidating all poisonous categories mitigates the model's need to discern subtle differences between poisonous classes, enabling it to concentrate on macro differences instead. Notably, as observed in the last two lines of the table, Track2 and Track1 occupy opposing ends of the seesaw, which is a natural consequence of the poisonous recognition strategy outlined in this paper. However, from the standpoint of Track3, the advantages of optimizing Track2 outweigh those of Track1, thereby reaffirming the arguments put forth in section 3.

Table 3 demonstrates the impact of different α on Track2. It is evident from the table that as the threshold value decreases, Track2 steadily drops. This is because α directly affects the model's recall rate for poisonous classes; the smaller α is, the higher the recall rate of the model, thus reducing the

Table 3
The table shows how different values of α impact Track2. The backbones are Vit-large and Eva02-large, and all the methods mentioned in Section 3 are adopted. The results are reported on the validation set.

α	Track2
0.2	0.2577
0.15	0.2366
0.1	0.2037
0.05	0.1841
0.01	**0.1226**

Table 4
The Table shows results of different methods and settings, all of which are reported on the validation set. Track1 represents the error rate of classification (both on close-set and open-set). Track2 represents the cost caused by misclassification and Track3 is equivalent to Track1 plus Track2. All these metrics are as small as possible. DM represents Dynamic MLP, PPT represents for putting the poisonous categories together, cat represents the concatenate operation in channel-wise with meta and image features.

Backbone	Feature fusion	Open-set	0-1 cost	Track1	Track2	Track3
Vit-large	cat	-	-	0.4623	0.5095	0.9718
Eva-large	cat	-	-	0.4547	0.5587	1.0134
Vit-large	DM	-	-	0.4581	0.4854	0.9435
Eva-large	DM	-	-	0.4438	0.4709	0.9147
Vit-large & Eva-large	DM	-	PPT	0.4313	0.3234	0.7647
Vit-large & Eva-large	DM	Entropy	PPT	**0.3651**	0.4834	0.8485
Vit-large & Eva-large	DM	Entropy	ME-loss & PPT	0.3809	0.2868	0.6677
Vit-large & Eva-large	DM	Entropy	ME-loss & PPT & threshold α	0.3951	**0.1226**	**0.5177**

Table 5
The table shows the final scores of different teams on the private leaderboard, Track1 represents the error rate of image recognition (including closed and open sets), Track2 represents the cost loss caused by wrong recognition, Track3 is equal to Track1 plus Track2, and the ranking is based on Track3, the smaller the better.

Rank	Team	Track1	Track2	Track3
1	IES	0.3107	0.0904	0.3621
2	jack-etheredge	**0.2436**	0.1629	0.4075
3	**upupup(Our)**	0.3898	**0.0718**	0.513
4	chirmy	0.2693	0.4149	0.6667
5	TingTing1999	0.2749	0.4378	0.6934
6	glhr	0.4996	0.6511	1.1526
7	DS@GT	0.3907	1.604	2.0443

probability of identifying poisonous classes as non-poisonous.

Table 5 illustrates the performance of our team in comparison to other competitors on the private leaderboard, where we secured the 3rd position, surpassing the 4th place in Track2 and achieving a performance even better than the 1st place. These outcomes collectively demonstrate the efficacy of the method presented in Section 3.3. Nonetheless, due to the simplicity of the approach we adopted for open-set recognition, we fell short in Track1 compared to other teams, highlighting an area that requires enhancement in our future endeavors.

5. Conclusion

The core challenges of FungiCLEF2024 are identifying fine-grained fungi images in an open-set environment and minimizing the 0-1 cost of misclassification. To address the first challenge, we use Dynamic MLP, a recursive structure utilizing matrix multiplication, for feature fusion to improve accuracy. To mitigate the 0-1 cost, we propose an easy yet effective approach that first places poisonous fungi categories into a single class and then employs ME loss and α to optimize the recall rate for poisonous species. However, open-set fine-grained fungi recognition remains a significant challenge. Our current approach relies solely on entropy for classifying open-set species, which has proven to be overly simplistic and inefficient. Consequently, the open-set problem stands as an enduring challenge that necessitates further investigation and innovation.

References

[1] L. Yang, X. Li, R. Song, B. Zhao, J. Tao, S. Zhou, J. Liang, J. Yang, Dynamic mlp for fine-grained image classification by leveraging geographical and temporal information, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10945–10954.

[2] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, S. Belongie, Fine-grained image analysis with deep learning: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (2021) 8927–8948.

[3] L. Picek, M. Sulc, J. Matas, Overview of FungiCLEF 2024: Revisiting fungi species recognition beyond 0-1 cost, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

[4] A. Joly, L. Picek, S. Kahl, H. Goëau, V. Espitalier, C. Botella, B. Deneu, D. Marcos, C. Leblanc, T. Larcher, M. Šulc, M. Hrúz, M. Servajean, et al., Overview of LifeCLEF 2024: Challenges on species distribution prediction and identification, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024.

[5] L. Picek, M. Šulc, J. Matas, J. Heilmann-Clausen, Overview of fungiclef 2022: Fungi recognition as an open set classification problem, Working Notes of CLEF (2022).

[6] J. Yu, H. Chang, K. Lu, G. Xie, L. Zhang, Z. Cai, S. Du, Z. Wei, Z. Liu, F. Gao, et al., Bag of tricks and a strong baseline for FGVC, Working Notes of CLEF (2022).

[7] G. Fan, C. Zining, W. Weiqiu, S. Yinan, S. Fei, Z. Zhicheng, C. Hong, Does closed-set training generalize to open-set recognition?, Working Notes of CLEF (2022).

[8] K. Desingu, A. Bhaskar, M. Palaniappan, E. A. Chodisetty, H. Bharathi, Classification of fungi species: A deep learning based image feature extraction and gradient boosting ensemble approach, Working Notes of CLEF (2022).

[9] S. Wolf, J. Beyerer, Transformer-based fine-grained fungi classification in an open-set scenario, Working Notes of CLEF (2022).

[10] F. Hu, P. Wang, Y. Li, C. Duan, Z. Zhu, Y. Li, X.-S. Wei, A deep learning based solution to fungiclef2023, Aliannejadi et al.[1] (2023) 2051–2059.

[11] F. Hu, P. Wang, Y. Li, C. Duan, Z. Zhu, F. Wang, F. Zhang, Y. Li, X.-S. Wei, Watch out venomous snake species: A solution to snakeclef2023, arXiv preprint arXiv:2307.09748 (2023).

[12] A. Behera, Z. Wharton, P. R. Hewage, A. Bera, Context-aware attentional pooling (cap) for fine-grained visual classification, in: Proceedings of the AAAI conference on artificial intelligence, volume 35, 2021, pp. 929–937.

[13] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, Z. Zhang, The application of two-level attention models in deep convolutional neural network for fine-grained image classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 842–850.

[14] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, L. Wang, Learning to navigate for fine-grained classification, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 420–435.

[15] N. Zhang, J. Donahue, R. Girshick, T. Darrell, Part-based r-cnns for fine-grained category detection,

in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, Springer, 2014, pp. 834–849.
[16] X.-S. Wei, C.-W. Xie, J. Wu, C. Shen, Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization, Pattern Recognition 76 (2018) 704–714.
[17] S. Huang, X. Wang, D. Tao, Snapmix: Semantically proportional mixing for augmenting fine-grained data, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 1628–1636.
[18] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.
[19] H. Li, X. Zhang, Q. Tian, H. Xiong, Attribute mix: Semantic data augmentation for fine grained recognition, in: 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), IEEE, 2020, pp. 243–246.
[20] H. Touvron, A. Vedaldi, M. Douze, H. Jégou, Fixing the train-test resolution discrepancy, Advances in neural information processing systems 32 (2019).
[21] D. Chang, Y. Ding, J. Xie, A. K. Bhunia, X. Li, Z. Ma, M. Wu, J. Guo, Y.-Z. Song, The devil is in the channels: Mutual-channel loss for fine-grained image classification, IEEE Transactions on Image Processing 29 (2020) 4683–4695.
[22] Y. Gao, X. Han, X. Wang, W. Huang, M. Scott, Channel interaction networks for fine-grained image categorization, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 10818–10825.
[23] G. Mai, K. Janowicz, B. Yan, R. Zhu, L. Cai, N. Lao, Multi-scale representation learning for spatial feature distributions using grid cells, arXiv preprint arXiv:2003.00824 (2020).
[24] R. Minetto, M. P. Segundo, S. Sarkar, Hydra: An ensemble of convolutional neural networks for geospatial land classification, IEEE Transactions on Geoscience and Remote Sensing 57 (2019) 6530–6541.
[25] T. Salem, S. Workman, N. Jacobs, Learning a dynamic map of visual appearance, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12435–12444.
[26] K. Tang, M. Paluri, L. Fei-Fei, R. Fergus, L. Bourdev, Improving image classification with location context, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1008–1016.
[27] C. Geng, S.-j. Huang, S. Chen, Recent advances in open set recognition: A survey, IEEE transactions on pattern analysis and machine intelligence 43 (2020) 3614–3631.
[28] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, T. E. Boult, Toward open set recognition, IEEE transactions on pattern analysis and machine intelligence 35 (2012) 1757–1772.
[29] A. Bendale, T. E. Boult, Towards open set deep networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1563–1572.
[30] L. Shu, H. Xu, B. Liu, Doc: Deep open classification of text documents, arXiv preprint arXiv:1709.08716 (2017).
[31] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, S. Yan, Metaformer is actually what you need for vision, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10819–10829.
[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
[33] L. Yuan, Q. Hou, Z. Jiang, J. Feng, S. Yan, Volo: Vision outlooker for visual recognition, IEEE transactions on pattern analysis and machine intelligence 45 (2022) 6575–6586.
[34] H. Ren, H. Jiang, W. Luo, M. Meng, T. Zhang, Entropy-guided open-set fine-grained fungi recognition., in: CLEF (Working Notes), 2023, pp. 2122–2136.
[35] J. Wang, W. Zhang, Y. Zang, Y. Cao, J. Pang, T. Gong, K. Chen, Z. Liu, C. C. Loy, D. Lin, Seesaw loss for long-tailed instance segmentation, 2021, pp. 9695–9704.
[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings

of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
[37] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4690–4699.
[38] J. Deng, J. Guo, T. Liu, M. Gong, S. Zafeiriou, Sub-center arcface: Boosting face recognition by large-scale noisy web faces, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, Springer, 2020, pp. 741–757.
[39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.
[40] O. Mac Aodha, E. Cole, P. Perona, Presence-only geographical priors for fine-grained image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9596–9606.
[41] L. Neal, M. Olson, X. Fern, W.-K. Wong, F. Li, Open set learning with counterfactual images, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 613–628.
[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.
[43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[44] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, Y. Cao, Eva-02: A visual representation for neon genesis, arXiv preprint arXiv:2303.11331 (2023).
[45] R. Wightman, Pytorch image models, https://github.com/rwightman/pytorch-image-models, 2019.
[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, 2009, pp. 248–255.
[47] I. Loshchilov, F. Hutter, Fixing weight decay regularization in adam (2017).
[48] I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts, 2017.
[49] X.-S. Wei, S.-L. Xu, H. Chen, L. Xiao, Y. Peng, Prototype-based classifier learning for long-tailed visual recognition, Science China Information Sciences 65 (2022) 160105.
[50] Y.-Y. He, J. Wu, X.-S. Wei, Distilling virtual examples for long-tailed recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 235–244.
[51] Y. Zhang, X. Wei, B. Zhou, J. Wu, Bag of tricks for long-tailed visual recognition with deep convolutional neural networks, 2021, pp. 3447–3455.
[52] B. Zhou, Q. Cui, X.-S. Wei, Z.-M. Chen, Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9716–9725. doi:10.1109/CVPR42600.2020.00974.