Entropy-guided Open-set Fine-grained Fungi Recognition

Huan Ren^{1,†}, Han Jiang^{1,†}, Wang Luo^{1,†}, Meng Meng^{1,†} and Tianzhu Zhang^{1,2,*}

¹University of Science and Technology of China, Hefei, Anhui, China ²Deep Space Exploration Lab, Hefei, Anhui, China

Abstract

FungiCLEF 2023 competition aims at automatically recognizing different species of fungi, which is a challenging open-set, fine-grained task. In this paper, we propose an entropy-guided method for open-set fine-grained fungi recognition. To deal with small inter-class variations in fine-grained recognition, we adopt the MetaFormer as our baseline, which can use meta-information to help differentiate similarly-looking species. Besides, we utilize a Seesaw loss to balance the training process between head classes and tail classes. To handle with the challenge of open-set recognition, we propose an entropy-guided method to discriminate unknown classes, since the entropy can measure the uncertainty of class predictions. By combining these techniques, our method achieves superior F1 Scores, specifically **58.95**% on the public test set and **58.36**% on the private test set, which ranks **1st** place in FungiCLEF 2023 competition. Our code will be available at https://github.com/RenHuan1999/FungiCLEF2023-UstcAIGroup.

Keywords

Fungi Recognition, Open-set Recognition, Fine-grained Recognition, Entropy-guided Method

1. Introduction

FungiCLEF 2023 is a competition held jointly as part of the LifeCLEF 2023 [1] lab¹ of the CLEF 2023 conference, and of the FGVC10 workshop² organized in conjunction with CVPR 2023 conference. The competition aims to automatically identify and classify fungi species, which helps raise awareness about fungi biodiversity and assists ecologists in species identification in the wild [2, 3]. The training data for FungiCLEF 2023 comes mainly from the Danish Fungi 2020 dataset [4] which includes rich metadata such as habit, substrate, time etc. The competition also releases a validation set containing some "Unknown Classes" which are never seen during training. In practice, fungi recognition needs to capture fine-grained visual differences within subordinate categories and identify both the known classes and unknown species. Thus, it is considered as a fine-grained, open-set recognition task.

Fine-grained recognition, as opposed to generic object recognition, is a challenging task due

CEUR Workshop Proceedings (CEUR-WS.org)

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece *Corresponding author.

[†]These authors contributed equally.

[☆] rh_hr_666@mail.ustc.edu.cn (H. Ren); hjiang12@mail.ustc.edu.cn (H. Jiang); lw1998@mail.ustc.edu.cn (W. Luo); meng18@mail.ustc.edu.cn (M. Meng); tzzhang@ustc.edu.cn (T. Zhang)

^{© 2023} Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹https://www.imageclef.org/LifeCLEF2023

²https://sites.google.com/view/fgvc10

to small inter-class variations and large intra-class variations [5]. To handle small inter-class variations, existing methods can be roughly grouped into two categories, including part-based methods [6, 7] and attention-based methods [8, 9], which make the network focus on the most class-discriminative regions. However, these methods ignore additional meta-information that is beneficial for recognition. For example, habitat and substrate are essential sources of information that help differentiate similarly-looking species. Thus, to utilize meta-information effectively for accurate fine-grained recognition, we take MetaFormer [10] as a strong baseline, which is a hybrid network with convolutions to downsample the image and Transformers to fuse visual and meta information. Besides, by analyzing the dataset, we observe a long-tailed category distribution problem. Thus, we adopt a Seesaw loss [11] to balance the training process between head classes and tail classes.

In open-set recognition, the training phase typically involves a set of known classes, and the classifier learns to distinguish and classify instances belonging to these classes [12]. However, during the testing phase, the classifier encounters instances that may belong to either the known classes or unknown classes. The classifier needs to make predictions for the known classes while having the ability to identify instances from unknown classes and assign them a separate "unknown" label. To achieve open-set recognition, existing methods focus on developing open-set classifiers [13, 14], but ignore the design of metrics for identifying known and unknown classes. The entropy [15] can measure the uncertainty of class predictions for the current input sample by a classification model. Based on this uncertainty, unknown classes can be effectively identified since the uncertainty is high for unknown classes. Therefore, we propose an entropy-guided unknown identifier for accurate open-set recognition.

The major contributions of this work can be summarized as follows: (1) We utilize MetaFormer as a powerful baseline for fine-grained recognition and employ the Seasaw loss to address the issue of long-tail distribution. (2) We propose a novel entropy-guided open-set recognition method, which is highly effective in addressing open-set recognition challenges. (3) Extensive experimental results demonstrate that the proposed method performs favorably against state-of-the-art methods, and our method ranks **1st** place in FungiCLEF 2023 competition.

2. Related Work

2.1. Fine-grained Vision Classification

Vision-driven method. Current fine-grained vision classification methods can be categorized into two directions: part-based methods [6, 16, 17] and attention-based methods [8, 9]. Part-based methods try to learn discriminative parts to distinguish different species. Zhang et al. [16] propose two deformable part descriptors to pool representations across pose and viewpoint, in turn facilitating fine-grained recognition and attribute prediction. Ge et al. [6], part models are built in a weakly supervised manner to retrieve information hidden in the object proposals. Another line of work attempts to introduce attention mechanisms to boost image recognition accuracy. RA-CNN [8] propose a recurrent attention convolutional neural network, which recursively learns crucial region attention and region-based feature representation at multiple scales. MA-CNN [9] jointly learns part proposals and the feature representations on each part with a multi-attention convolutional neural network. TransFG [18] integrates all

raw attention weights of the Transformer into an attention map to select discriminative image patches. However, these methods ignore some meta-information, making it difficult to recognize vision-similarly species.

Multi-modality-driven method. Besides visual information, some works take advantage of additional meta-information (*e.g., spatio-temporal prior, attribute, and text description*). Tang et al. [19] first introduce additional information and naturally incorporate these features to make final predictions. Dynamic MLP [20] is an efficient structure that exploit the additional information as adaptive perceptron weights to interact with vision features. MetaFormer [10] designs a hybrid structure backbone including convolution and Transformer to fuse vision and meta information to establish joint representations. We choose MetaFormer as our baseline in this competition.

2.2. Open-set Recognition

The objective of open-set recognition is to classify the known classes during the training process and to identify the unknown classes which are absent in the training set while keep the ability for known classification during the test process. Scheirer et al. [21] first present the open-set classification problem and design open set classifiers under the one-vs-set setting to balance the unknown classes. Jain et al. [22] calibrate SVM decision scores to posterior probabilities under the multi-class classifier setting. Another pipeline [23, 24] utilizes uncertainty to measure the feature difference between unknown and known objects. Denouden et al. [23] incorporate the Mahalanobis distance in latent space to better capture these out-of-distribution samples. Liu et al. [24] propose an energy-bounded learning objective to fine-tune the network and design an energy score to better distinguish unknown classes. In this work, we present to use entropy, a simple but effective way to measure uncertainty.

2.3. Fungi Species Classification

The 1st team's paper [25] in FungiCLEF2022 [26] written by Xiong et al. use an ensemble of MetaFormer [10] and ConvNext [27] networks. USTC-IAT-United [28] ensemble several CNN and Transformer architectures, and explore the impact of data augmentation techniques and loss functions. Shen et al. [29] design a novel architecture combined with large kernel convolution and vision Transformer. However, these methods ignore the metrics for identifying known and unknown classes, and we design an entropy-guided unknown identifier to recognize unknown classes more accurately.

3. Our Method

3.1. Model Design

Deep learning model. Figure 1 shows the overall framework of our model. We employ MetaFormer [10] as a strong baseline for fine-grained visual classification. Metaformer is a hybrid framework where the first three stages are convolution blocks and the last two stages



Figure 1: Overview of our framework. We use images and meta data as input and train the MetaFormer model with Seesaw loss. During testing, the proposed entropy-guided unknown identifier is employed for open-set recognition.

Detail settings of MetaFormer-0 and MetaFormer-2. L represents the number of blocks, and C represents the channels of hidden features.

Stages	S0	S1	S2	S3	S4
MetaFormer-0	L=3 C=64	L=2 C=96	L=3 C=192	L=5 C=384	L=2 C=768
MetaFormer-2	L=3 C=128	L=2 C=128	L=3 C=256	L=5 C=512	L=2 C=1024

are Transformer blocks. We combine two series of MetaFormer, namely MetaFormer-0 and MetaFormer-2. The detailed settings are shown in Table 1. In addition, we also try some vision classification models, like InternImage-L [30]. However, due to the memory limits, we do not choose it as our baseline.

Meta information. In order to classify the similarly-looking species, we make full use of extra meta-information to assist final decision. The meta-information and appearance information are simply fused by relative attention [31] in Transformer layers. Intuitively, we consider geographical, living habit and temporal information including *Observe date (month and day), countryCode, Substrate* and *Habitat*. For temporal information, we map the Observe date [month, day] into $\left[\sin\left(\frac{2\pi month}{12}\right), \cos\left(\frac{2\pi month}{12}\right), \sin\left(\frac{2\pi day}{31}\right), \cos\left(\frac{2\pi day}{31}\right)\right]$. For geographical and living habit information, there are 34, 32 and 31 categories for *countryCode, Substrate* and *Habitat*, respectively. We use one-hot encoding to encode meta-information according to their meta category. Additionally, we employ a non-linear embedding technique to map the different meta-embeddings into a C-dimensional space, which is consistent with the dimension of the image features.

Long-tailed Distribution. As shown in Figure 2, the training set provided by the Fungi-CLEF2023 challenge has a long-tailed class distribution. Among them, the most frequent species has 1913 images while the least frequent species has only 31 images. The imbalanced class



Figure 2: Class distribution of the training set of FungiCLEF2023.

J

distribution will make the training process dominated by head categories, and further lead to misclassification for tail categories. To cope with this challenge, we adopt Seesaw loss [11] instead of cross entropy loss. Seesaw Loss can dynamically decrease the gradients of negative samples for each category by modifying the cross entropy loss as:

$$\mathcal{L}_{Seesaw}(l) = -\sum_{i=1}^{C} y_i \log p_i,$$

$$p_i = \frac{e^{l_i}}{\sum_{i \neq i}^{C} S_{ij} e^{l_j} + e^{l_i}},$$
(1)

where l represents the output logits of the model, y is the one-hot class label and S is a tunable balancing factor across classes. For more details, please refer to [11].

Data augmentation. We use a composed sequence of some common augmentation techniques to improve the results. In the training process, a random cropping is first performed on the image, where the size of the cropped region is randomly chosen between 8% and 100% of the original image size. Afterward, the slice is resized by the bicubic interpolation method and flipped horizontally with a probability of 50 %. Additionally, we adopt RandAugment [32] following the setting of Swin Transformer [31]. This includes both photometric augmentation (*Brightness, Contrast, Color, Sharpness, etc.*) and geometric augmentation (*Shear, Rotate, Translate, etc.*). Finally, a random erasing [33] is used to mask out a random region of images with a probability of 25%. It is worth noting that Mixup [34] and Cutmix [35] are not used in the pre-processing pipeline. During the inference process, we use multi scale & ten crop to perform test time augmentation.

3.2. Entropy-guided Unknown Identifier

In the open-set scenario, the model needs not only to correctly classify between training known categories, but also to identify whether an image is from a previously unseen class, considered an "unknown" category. While existing literature proposes several classical approaches to identify the unknown, such as MSP[13, 36], MLS[37], we introduce a more effective entropy-guided unknown identifier. We next briefly review the previous approaches and then present our method.

Maximum Softmax Probability (MSP). Given a network consisting of a backbone g and a classifier h, the output *logits* of an image x is denoted as l(x) = h(g(x)). The predicted *probabilities* p(x) that sum to one can then be derived by applying the softmax to logits. The model is trained for closed-set classification, specifically, using the cross-entropy loss between the true class labels y and the predicted probabilities p(x). During testing, the maximum softmax probability $p_{max}(x) = \max(p(x))$ is taken for open-set recognition[13, 36]. If $p_{max}(x)$ is above the threshold τ , the image is identified as the corresponding known category, otherwise it is considered to be the unknown.

Maximum Logit Score (MLS). Although most of the literature adopts MSP as the baseline, Vaze et al.[37] suggest using the maximum logit score $l_{max}(x) = \max(l(x))$ instead, and similarly, using threshold τ for $l_{max}(x)$ to determine known/unknown categories. It has been observed that open-set samples tend to have lower feature magnitudes than closedset ones[38, 39]. Since the softmax operation normalizes out most of the feature magnitude information in the logits, they indicate that using the raw logits would lead to better open set detection results.

Entropy. We first try to adopt the MLS as our baseline. To determine the optimal threshold τ , we plot the distribution of the maximum logit values, as shown in Figure 3(a). As can be seen in the third column, there is no clear boundary for the distribution on the test set, so the threshold τ can only be set based on the prior information about the number of the unknown samples[40]. Even with additional training on the validation set, the distribution boundary on the test set is still not obvious, as shown in the fourth column. We therefore conclude that the open-set detection performance is very sensitive to the choice of hyperparameter τ , which is not robust enough. To solve this problem, we propose an entropy-guided unknown identifier. Specifically, we first calculate the entropy of the predicted probabilities, defined as:

$$e(x) = -\sum_{c=1}^{C} p_c(x) \log p_c(x),$$
(2)

which is then used for open-set recognition with threshold τ . The entropy measures the uncertainty of the predictions of the classifier on the input samples. In general, the model is more confident for the known categories, corresponding to a lower entropy. Whereas for the unknown categories the uncertainty is higher and hence the entropy will be higher. Thus, we can effectively distinguish between known/unknown categories based on entropy. As shown in



Figure 3: The distribution of the maximum logit values and the entropies on the validation and test set. The sources of training data are noted in parentheses.

the fourth column of Figure 3(b), after additional training on the validation set, there is a clear boundary in the entropy distribution on the test set, making the choice of threshold τ more intuitive and robust.

3.3. Important Strategies

Training with the validation set additionally. Since the final evaluation is on the test set, it is a common trick to additionally use the data from the validation set for training. For this competition this trick is helpful because the validation set also contains some samples from the unknown category. With these unknown samples, we can estimate the data distribution of unknown categories and achieve better known/unknown separation. Specifically, for the unknown samples in the validation set, we constrain the known classifier to yield a uniform distribution for them, with the target in the corresponding cross-entropy loss formulated as $\hat{y} = \frac{1}{C} \mathbf{1}_C$, where $\mathbf{1}_C$ denote vectors of ones in dimension of C categories.

Enhanced identification of poisonous species. There are some poisonous species of fungi that exist in nature and are dangerous if eaten by mistake. For these poisonous categories, the model should try to identify them, so there should also be an edible/poisonous identification. If a poisonous observation is misclassified as edible, the corresponding cost should increase. To enhance the identification of poisonous species, we design a poisonous classification loss. Specifically, the set of poisonous categories is denoted by C_{poi} , and the set of remaining edible species is denoted by C_{edi} . For those poisonous samples, we index out and then average the top-k of the classification logits corresponding to C_{poi} , as well as the top-k corresponding to C_{edi} . A poisonous/edible classification loss is then used as an additional supervision. The whole process is formulated as:

$$l_{set}(x) = \operatorname{mean}\left(\operatorname{top-}k\left(\{l_c(x)|c \in \mathcal{C}_{set}\}\right)\right), set \in \{poi, edi\},\tag{3}$$

$$\mathcal{L}_{poi}(x) = -\log \frac{e^{l_{poi}(x)}}{e^{l_{poi}(x)} + e^{l_{edi}(x)}}, y(x) \in \mathcal{C}_{poi}.$$
(4)

Model ensemble. Model ensemble is a well proven trick in competitions. However, the organizers of this competition claim that large models are less practical and therefore limit the maximum model size to 1 GB. Within the limitation of the model size, we experiment with ensemble of the MetaFormer-0 (150 MB) and MetaFormer-2 (393 MB) models. The ensemble model (543 MB) can further improve performance while meeting the model size limit.

4. Experiment

4.1. Experimental Settings

Dataset. The training data for the FungiCLEF2023 challenge is primarily derived from the Danish Fungi 2020 dataset[4], which contains 295,938 training images belonging to 1,604 categories. Besides the observed images, a wealth of metadata is also provided, such as habitat, substrate, time, location, country, etc. The validation set contains 60,832 images from 30,131 observations belonging to 2,713 categories, where multiple images may be taken for observations of the same object. The public test set contains 60,225 images from 30,130 observations and has a similar data distribution to the validation set, with an unknown number of categories. To ensure the fairness of the competition and prevent participants from overfitting to the leaderboard, there is another private test set that has not been released by the organizers. The final performance evaluation is conducted on the test set.

Implementation Details. We employ the pre-trained MetaFormer[10] as our backbone network, fine-tune it and learn a set of meta tokens and a classification head. Our model is trained using the AdamW[41] optimizer with a cosine decay learning rate scheduler. The learning rate is initialized to $5e^{-5}$ and the weight decay is 0.05. We follow most of the data augmentation and regularization strategies from Swin Transformer[31] in our training. We conduct all the experiments with NVIDIA RTX 3090 (24G) GPU. The threshold τ for the entropy-guided unknown identifier is set to 4, and the k in Equation (3) is 5 in our case.

4.2. Evaluation Metrics

The FungiCLEF2023 challenge uses five evaluation metrics to measure performance from different perspectives, including F₁-Score (higher is better) and four costs (lower is better). Note that during evaluation, all categories not present in the training set are treated as a single "unknown" category, indexed by "-1". Given prediction $q(x^i)$ over observation x^i , ground-truth labels y^i , and a cost function $W_n(y^i, q(x^i))$, the *n*-th cost is calculated by averaging the cost function over all observations:

$$Cost_n = \frac{1}{N} \sum_{i=1}^{N} W_n(y^i, q(x^i)).$$
 (5)

 F_1 -Score. The F_1 score for the *c*-th category is the harmonic mean of the precision and recall, which is formulated as:

$$F_1^c = 2 \cdot \frac{precision^c \cdot recall^c}{precision^c + recall^c} = \frac{2T_p^c}{2T_p^c + F_p^c + F_n^c},\tag{6}$$

where T_p , F_p , F_n represent true-positive, false-positive and false-negative, respectively. The final macro F_1 -Score is calculated by averaging the F_1 scores over all categories:

macro
$$F_1 = \frac{1}{C} \sum_{c=1}^{C} F_1^c.$$
 (7)

Track1: Standard Classification with "unknown" category. The standard classification cost is obtained by averaging the predicted category error over all instances, and the cost function is defined as:

$$W_1(y^i, q(x^i)) = \begin{cases} 0, & y^i = q(x^i) \\ 1, & y^i \neq q(x^i) \end{cases}.$$
(8)

Track2: Cost for confusing edible species for poisonous and vice versa. There is a function *d* that indicates poisonous species as d(y) = 1 if species *y* is poisonous and 0 otherwise. The cost function for poisonous/edible confusion is formulated as:

$$W_2(y^i, q(x^i)) = \begin{cases} 0, & d(y^i) = d(q(x^i)) \\ c_{PSC}, & d(y^i) = 1 \text{ and } d(q(x^i)) = 0 , \\ c_{ESC}, & d(y^i) = 0 \text{ and } d(q(x^i)) = 1 \end{cases}$$
(9)

where $c_{ESC} = 1$ and $c_{PSC} = 100$ because it is more dangerous for a poisonous observation to be misclassified as edible.

Track3: A user-focused loss composes of both the classification error and the poi-sonous/edible confusion. The cost function is given as:

$$W_3(y^i, q(x^i)) = W_1(y^i, q(x^i)) + W_2(y^i, q(x^i)).$$
(10)

Track4: Cost for missing "unknown" species is higher; misclassifying for "unknown" is cheaper than confusing species. The cost function is a variant of the standard classification cost, increasing the weight of the cost of missing an unknown category and decreasing the weight of the cost of wrongly detecting an unknown category, which is written as:

$$W_4(y^i, q(x^i)) = \begin{cases} 0, & y^i = q(x^i) \\ \alpha, & y^i = -1 \text{ and } q(x^i) \neq -1 \\ \beta, & y^i \neq -1 \text{ and } q(x^i) = -1 \\ 1, & y^i \neq q(x^i) \text{ and } y^i \neq -1 \text{ and } q(x^i) \neq -1 \end{cases}$$
(11)

where $\alpha = 10$ and $\beta = 0.1$.

 Table 2

 Public leaderboard of FungiCLEF2023 competition. We rank 1st place.

Rank	Team Name	$F_1 (\uparrow)$	Track1 (↓)	Track2 (↓)	Track3 (↓)	Track4 (↓)
1	meng18	58.95	0.2072	0.1742	0.3814	1.4762
2	stefanwolf	56.27	0.3528	0.2133	0.5662	2.9296
3	word2vector	55.46	0.3519	0.2561	0.6080	2.8167
4	SSSAMMMM	52.76	0.4124	0.3270	0.7395	3.3302

Private leaderboard of FungiCLEF2023 competition. We rank 1st place.

Rank	Team Name	$F_1 (\uparrow)$	Track1 (\downarrow)	Track2 (\downarrow)	Track3 (↓)	Track4 (↓)
1	meng18	58.36	0.2409	0.1269	0.3702	1.771
2	stefanwolf	55.31	0.3473	0.1904	0.556	1.9045
3	word2vector	54.34	0.3601	0.2324	0.6034	2.9269
4	SSSAMMMM	51.67	0.4408	0.3264	0.7673	3.6493

4.3. Comparison on the Leaderboard

Table 2 and Table 3 show the performance of our team compared to other teams on the leaderboard of FungiCLEF 2023 competition. The leaderboards are divided into public and private test sets, in both of which we win **1st** place, outperforming the 2nd place by a significant margin. For example, in the public leaderboard in Table 2, our F_1 -Score exceeds that of the 2nd place by more than 2%, indicating that our method can better solve the open-set fine-grained recognition problem. On both the standard classification cost represented by Track1 and the unknown classification cost represented by Track4, our cost is reduced by nearly half compared to the 2nd place, demonstrating that our method is better at identifying unknown species. Our method also achieves a lower cost on the poisonous classification cost denoted by Track2, indicating that our method discriminates the poisonous/edible category better. Similar performance gains can be observed in the private leaderboard in Table 3.

4.4. Ablation Studies

To analyze the impact of each design, we conduct a series of ablation studies on the public test set, as detailed below.

Model architecture. Table 4 shows the impact of the model architecture on performance. The comparison between **1** and **2** reveals that the accumulated steps in the training process does not have a significant impact on performance. Certain performance gains are achieved by performing model ensemble for **1** and **2**. InternImage[30], a recent model that uses pure images to classify, performs worse compared to MetaFormer, indicating the importance of meta-information for fine-grained classification. By performing model ensemble on MetaFormer and InternImage, performance is higher than **3**, indicating that ensemble is more effective with different model architectures. However, in our final submission, we do not employ InternImage due to the model size limit.

Ablation studies about different model architectures. Accu4 means that the accumulated steps is set to 4 during training. We use the maximum logit scores to identify the unknown and do not use additional validation set for training.

Model Architecture	$F_1 (\uparrow)$	Track1 (↓)	Track2 (\downarrow)	Track3 (↓)	Track4 (\downarrow)
MetaFormer-2 (accu4)	55.94	0.3530	0.2160	0.5690	2.9591
MetaFormer-2 (accu8)	55.97	0.3550	0.2398	0.5948	2.9412
❸ Ensemble (❶ + ❷)	56.68	0.3499	0.2263	0.5762	2.9283
InternImage	54.32	0.3598	0.2265	0.5863	2.9543
⑤ Ensemble (① + ④)	56.50	0.3486	0.1962	0.5448	2.9159

Table 5

Ablation studies about different sources of training data. We use the maximum logit scores to identify the unknown with MetaFormer-2.

Training Data	$F_1 (\uparrow)$	Track1 (↓)	Track2 (\downarrow)	Track3 (↓)	Track4 (↓)
train	55.94	0.3530	0.2160	0.5690	2.9591
train+val	54.98	0.3381	0.1896	0.5277	2.8853

Table 6

Ablation studies about different unknown identifier. MLS indicates the Maximum Logit Score. We use additional validation set for training MetaFormer-2.

Unknown Identifier	$F_1 (\uparrow)$	Track1 (\downarrow)	Track2 (\downarrow)	Track3 (\downarrow)	Track4 (\downarrow)
MLS	54.98	0.3381	0.1896	0.5277	2.8853
entropy	57.84	0.2101	0.2544	0.4645	1.4815

Training data. Table 5 shows the impact of different training data sources on performance. As can be seen from the table, the additional use of the validation set for training can lead to performance gains on the Track1-Track4 metrics. It can also be seen in Figure 3 that adding the validation set for training can improve the discrepancy between known/unknown categories on the validation set and further generalize to the test set for better open-set recognition.

Unknown identifier. Table 6 demonstrates the impact of different unknown identification methods on performance. As can be seen from the table, the proposed entropy-guided unknown identifier can lead to a significant performance improvement, especially for the identification of unknown categories, as indicated by Track1 and Track4. Furthermore, as shown in Figure 3, entropy is more discriminative for known/unknown categories than MLS, indicating that entropy is more suitable for open-set recognition. However, it is worth noting that when utilizing entropy, Track2 (poisonous identification) exhibits a higher value compared to MLS. This discrepancy arises due to the fact that the "poisonous identification" task only focuses on "known categories", while entropy encourages the model to identify more samples as unknown classes. As a result, the model's performance in the "poisonous identification" task for known categories may be compromised. To address this issue and enhance the identification of poisonous species, we have designed a specific poisonous/edible classification loss, which is elaborated upon in detail in Section 3.3.

Ablation studies about the enhanced identification of poisonous species. We employ the entropy-guided unknown identifier and use additional validation set for training MetaFormer-0.

\mathcal{L}_{poi}	F1 (†)	Track1 (↓)	Track2 (\downarrow)	Track3 (↓)	Track4 (\downarrow)
X	57.80	0.2088	0.2865	0.4953	1.4256
1	58.11	0.2069	0.2067	0.4136	1.3936

Table 8

Performance comparison of the final submitted models on the public test set.

Submitted Model	$F_1 (\uparrow)$	Track1 (↓)	Track2 (\downarrow)	Track3 (↓)	Track4 (\downarrow)
MetaFormer-0 (150M)	58.11	0.2069	0.2067	0.4136	1.3936
MetaFormer-2 (393M)	57.63	0.2123	0.1943	0.4066	1.4984
Ensemble (543M)	58.95	0.2072	0.1742	0.3814	1.4762

Poisonous/edible identification. Table 7 demonstrates the impact of the poisonous/edible classification loss \mathcal{L}_{poi} on performance. As can be seen from the table, Track2 is significantly reduced with the incorporation of explicit poisonous/edible classification supervision, indicating a stronger identification of poisonous categories.

Submitted models. Table 8 shows a comparison among the performance of the three submitted models on the public test set. Within the limitation of the maximum model size of 1GB, we employ the ensemble of MetaFormer-0 and MetaFormer-2, yielding further performance gains.

5. Conclusion

In this paper, we propose an entropy-guided method for open-set fine-grained fungi classification. The proposed method is based on a modern model MetaFormer, which can utilize meta-information to differentiate similarly-looking species for accurate fine-grained recognition. We also explore the Seesaw loss for long-tail recognition. More importantly, we propose an entropy-guided unknown identifier to discriminate unknown classes with the support of open-set scenarios. Extensive results demonstrate the effectiveness of the proposed method, and our method ranks **1st** place in Fungi2023 competition.

References

- A. Joly, H. Goëau, S. Kahl, L. Picek, C. Botella, D. Marcos, M. Šulc, M. Hrúz, T. Lorieul, S. S. Moussi, et al., Lifeclef 2023 teaser: Species identification and prediction challenges, in: European Conference on Information Retrieval, Springer, 2023, pp. 568–576.
- [2] M. Sulc, L. Picek, J. Matas, T. Jeppesen, J. Heilmann-Clausen, Fungi recognition: A practical use case, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2316–2324.
- [3] L. Picek, M. Šulc, J. Matas, J. Heilmann-Clausen, T. S. Jeppesen, E. Lind, Automatic fungi recognition: Deep learning meets mycology, Sensors 22 (2022) 633.

- [4] L. Picek, M. Šulc, J. Matas, T. S. Jeppesen, J. Heilmann-Clausen, T. Læssøe, T. Frøslev, Danish fungi 2020-not just another image recognition dataset, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1525–1535.
- [5] Y. Ding, Y. Zhou, Y. Zhu, Q. Ye, J. Jiao, Selective sparse sampling for fine-grained image recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6599–6608.
- [6] W. Ge, X. Lin, Y. Yu, Weakly supervised complementary parts models for fine-grained image classification from the bottom up, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3034–3043.
- [7] C. Liu, H. Xie, Z.-J. Zha, L. Ma, L. Yu, Y. Zhang, Filtration and distillation: Enhancing region attention for fine-grained visual categorization, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 11555–11562.
- [8] J. Fu, H. Zheng, T. Mei, Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4438–4446.
- [9] H. Zheng, J. Fu, T. Mei, J. Luo, Learning multi-attention convolutional neural network for fine-grained image recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5209–5217.
- [10] Q. Diao, Y. Jiang, B. Wen, J. Sun, Z. Yuan, Metaformer: A unified meta framework for fine-grained recognition, arXiv preprint arXiv:2203.02751 (2022).
- [11] 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, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9695–9704.
- [12] K. Joseph, S. Khan, F. S. Khan, V. N. Balasubramanian, Towards open world object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5830–5840.
- [13] 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.
- [14] Z. Ge, S. Demyanov, Z. Chen, R. Garnavi, Generative openmax for multi-class open set classification, arXiv preprint arXiv:1707.07418 (2017).
- [15] P. J. Coles, M. Berta, M. Tomamichel, S. Wehner, Entropic uncertainty relations and their applications, Reviews of Modern Physics 89 (2017) 015002.
- [16] N. Zhang, R. Farrell, F. Iandola, T. Darrell, Deformable part descriptors for fine-grained recognition and attribute prediction, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 729–736.
- [17] 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.
- [18] J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, C. Wang, Transfg: A transformer architecture for fine-grained recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 852–860.
- [19] 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.

- [20] L. Yang, X. Li, R. Song, B. Zhao, J. Tao, S. Zhou, J. Liang, J. Yang, Dynamic mlp for finegrained 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.
- [21] 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.
- [22] L. P. Jain, W. J. Scheirer, T. E. Boult, Multi-class open set recognition using probability of inclusion, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13, Springer, 2014, pp. 393–409.
- [23] T. Denouden, R. Salay, K. Czarnecki, V. Abdelzad, B. Phan, S. Vernekar, Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance, arXiv preprint arXiv:1812.02765 (2018).
- [24] W. Liu, X. Wang, J. Owens, Y. Li, Energy-based out-of-distribution detection, Advances in neural information processing systems 33 (2020) 21464–21475.
- [25] Z. Xiong, Y. Ruan, Y. Hu, Y. Zhang, Y. Zhu, S. Guo, W. Zhu, B. Han, An empirical study for fine-grained fungi recognition with transformer and convnet, Working Notes of CLEF (2022).
- [26] L.Picek, M.Sulc, J. S. Matas, J.Heilmann-Clausen, Overview of fungiclef 2022: Fungi recognition as an open set classification problem, in: Conference and Labs of the Evaluation Forum, 2022.
- [27] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.
- [28] 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).
- [29] Y. Shen, X. Sun, Z. Zhu, When large kernel meets vision transformer: A solution for snakeclef & fungiclef, Working Notes of CLEF (2022).
- [30] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, et al., Internimage: Exploring large-scale vision foundation models with deformable convolutions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14408–14419.
- [31] 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.
- [32] E. D. Cubuk, B. Zoph, J. Shlens, Q. V. Le, Randaugment: Practical automated data augmentation with a reduced search space, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 702–703.
- [33] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 13001– 13008.
- [34] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, arXiv preprint arXiv:1710.09412 (2017).
- [35] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE/CVF international

conference on computer vision, 2019, pp. 6023-6032.

- [36] 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.
- [37] S. Vaze, K. Han, A. Vedaldi, A. Zisserman, Open-set recognition: A good closed-set classifier is all you need, arXiv preprint arXiv:2110.06207 (2021).
- [38] A. R. Dhamija, M. Günther, T. Boult, Reducing network agnostophobia, Advances in Neural Information Processing Systems 31 (2018).
- [39] G. Chen, P. Peng, X. Wang, Y. Tian, Adversarial reciprocal points learning for open set recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (2021) 8065–8081.
- [40] Z. Xiong, Y. Ruan, Y. Hu, Y. Zhang, Y. Zhu, S. Guo, B. Han, 1st place solution for fungiclef 2022 competition: Fine-grained open-set fungi recognition, CEUR Workshop Proceedings 3180 (2022) 2227 – 2237.
- [41] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).