Generalizable Training Techniques for Fine-Grained Long-Tailed Image Recognition: Transferring Methods Optimized for FungiCLEF 2024 to SnakeCLEF 2024

Jack N. Etheredge[1,*]

[1]Twosense, New York, New York, United States

Abstract

Accurate identification of species in fine-grained, long-tailed datasets poses significant challenges due to imbalanced class distributions and the necessity for precise classification while minimizing confusion between dangerous and harmless species. This paper introduces a generalized training and inference methodology designed to tackle these challenges, demonstrated through competitive performance in both the SnakeCLEF and FungiCLEF 2024 challenges. While results for FungiCLEF 2024 are detailed in an accompanying paper, this work primarily explores the application and performance of the same techniques to the SnakeCLEF 2024 challenge. The proposed approach integrates a combination of augmentation techniques, specialized loss functions, and robust model architectures to enhance classification accuracy while jointly minimizing the asymmetric penalty for misclassification of venomous species. For both the public and private leaderboards, my approach achieved second place in all metrics. On the public leaderboard, it scored **81.2** for Track 1, **945** for Track 2, and **33.35** for the F1 score. On the private leaderboard, it scored **79.58** for Track 1, **2557** for Track 2, and **30.29** for the F1 score. These experimental results validate the effectiveness of this methodology, showcasing its robustness across diverse datasets and evaluation metrics. The versatility of this approach indicates its potential applicability to a wide range of similar image recognition tasks. Code and implementation details are available at https://github.com/Jack-Etheredge/snakeclef2024.

Keywords

Fine-grained classification, Long-tailed, Metaformer, CAFormer, SnakeCLEF, FungiCLEF

1. Introduction

Venomous snake bites cause over half a million deaths and disabilities annually, highlighting the need for an effective image-based snake identification system [1]. Such a system could enhance global health efforts, improve ecological and epidemiological data, and optimize antivenom distribution [2]. To this end, the SnakeCLEF 2024 challenge [3] is organized with metrics for both general misclassification rate as well as distinct penalties for the confusion of venomous snakes with other venomous snake species and the confusion of venomous snakes with harmless snakes.

Fine-grained long-tailed image recognition is a challenging task due to the need for high granularity in distinguishing between visually similar classes compounded by significant class imbalance. Competitions like SnakeCLEF and FungiCLEF, both part of the LifeCLEF 2024 [4] lab [1], provide platforms for developing and benchmarking methodologies to tackle these issues. SnakeCLEF 2024 focuses on snake species classification, while FungiCLEF 2024 [5] targets fungi species, including the identification of unknown species and minimizing misclassification between edible and poisonous varieties. Despite differences in datasets and evaluation metrics, both competitions share challenges inherent to fine-grained long-tailed classification, making them ideal for testing the generalizability of my proposed method.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

jack.etheredge@gmail.com (J. N. Etheredge)

https://github.com/Jack-Etheredge (J. N. Etheredge)

0000-0001-5467-3866 (J. N. Etheredge)

[1]https://www.imageclef.org/LifeCLEF2024

2. Related Work

Many different techniques have been explored for the classification of fine-grained images of snakes [6] and fungi [7]. Recent work for both tasks have shown the important role that the inclusion of metadata can play in the final classification performance of different techniques [8, 9, 10, 11, 12]. This year, however, metadata was excluded from the test set for SnakeCLEF. One effect of this is that the geographic regions that the snakes belong to cannot be directly utilized by the models nor can challengers focus efforts on training the classes that belong to the geographic regions present in the test set. Various loss functions and architectures have been successfully applied to SnakeCLEF to deal with the long-tailed fine-grained nature of the data. Seesaw loss [13] and real-world weighted cross-entropy [14] were used by [10]. Focal loss [15] and ArcFace loss [16] were both utilized by [12]. Interestingly, this solution also utilized a training dataset preprocessing step of cropping the images to the region of interest containing the snake. ArcFace and SimCLR [17] were used by [9]. ConvNeXt [18] and Metaformer [19] were top performing model architectures in last year's challenge [6].

3. Methodology

3.1. Dataset

The SnakeCLEF dataset consists of 182,261 images across 1,784 snake species. The training data includes geographical location metadata. FungiCLEF's dataset comprises 295,938 training images of 1,604 species with extensive metadata, including habitat and location. While the metadata was present in the test data for FungiCLEF 2024, it was absent from the test data for SnakeCLEF 2024.

3.2. Competition Objectives and Metrics

Both competitions aim to enhance species recognition accuracy, albeit with differing focuses. SnakeCLEF 2024 evaluates class-balanced metrics, emphasizing the importance of correctly classifying venomous vs. non-venomous species without leveraging metadata at inference time to blind the models to the geographic region. FungiCLEF 2024 includes an open-set component for identifying unknown species and penalizes misclassifications between edible and poisonous fungi. The venomous confusion loss for SnakeCLEF is more complex than the poisonous confusion loss for FungiCLEF, with different costs for misclassification between two venomous classes (2), between two nonvenomous classes (1), venomous $\longrightarrow$ nonvenomous confusion (5), and nonvenomous $\longrightarrow$ venomous confusion (2). Both competitions report the macro-F1 score, but SnakeCLEF additionally incorporates it into the Track 1 score. Track 1 is a weighted average of the accuracies for the four different confusion categories and the macro-averaged F1. Accuracy is also reported for both competitions, but is largely ignored for the results shown in this paper, as it is not reported for the granular results in the overview of either competition last year [6, 7].

3.3. Training Techniques

To address the long-tailed distribution and fine-grained nature of the datasets, I employed a combination of training techniques and test-time augmentations detailed below.

3.3.1. Data Augmentation

Training was performed with a resize to 768 with bicubic interpolation, square random crop of size 384, TrivialAugment [20], horizontal flip with 50% probability, and random erasing [21] with a probability of 25%, applied in that order. MixUp augmentation [22] and augmentations inspired by it were intentionally excluded due to the fine-grained nature of the dataset, which represents higher intra-class variability and lower inter-class variability than standard classification tasks. However, mixing augmentations in the form of CutMix [23] and RandoMix [24] were previously employed successfully by [10]. Future

work could explore the use of different data augmentations including MixUp and similar techniques during training.

3.3.2. Loss Functions

Multiple loss functions were evaluated for the classification loss. Seesaw loss [13] and a custom venom loss were used to train the models in the final ensemble. Seesaw loss was chosen since it is designed for long-tailed classification. Further, it achieves this without the need for class rebalancing through data sampling by adding additional terms to the standard cross-entropy loss. It employs a mitigation factor to reduce penalties for tail categories based on the ratio of training instances as well as a compensation factor to increase penalties for misclassified instances, thereby reducing the otherwise overwhelming effect of false positives in the tail classes.

A custom venom loss was added to seesaw loss to create the total loss during training. This cost function was formulated by creating a pairwise cost for the confusion for every combination of the target and predicted class. The vector corresponding to the target class was indexed from this cost matrix and the softmax probabilities were multiplied elementwise with the cost vector. The sum of these costs was used as the venom loss. This loss is similar to the real-world weighted cross entropy loss [14], but uses the costs directly instead of utilizing a weighted log loss. Future work could investigate the relative performance of these two loss functions. Since the venom confusion metric is calculated based on the percentage of misclassifications, it is a class-balanced metric. As such, I also experimented with the application of an inverse class weight to the venom loss to account for class imbalance (results shown in Table 5).

Balanced sampling is a simpler alternative to seesaw loss for mitigating the effect of class imbalance. For each epoch, samples were drawn with replacement from the training data with a probability inversely proportional to the number of samples belonging to that class. Focal loss [15] penalizes misclassifications for difficult to classify samples by reducing the loss for well-classified examples (high predicted probability for the correct class) relative to standard cross entropy loss. This is done in an attempt to put more focus on difficult examples dynamically during training. Since the tail classes will likely be more difficult to classify, focal loss should in theory work in conjunction with balanced sampling to improve tail class classification.

Another loss that was evaluated was sub-center ArcFace loss [25]. Sub-center ArcFace loss is a refinement to ArcFace that allows multiple cluster centers per class, which seemed better suited to snake classification than the original ArcFace loss since snakes of the same species can vary widely in their appearance due to age and other factors. Losses that operate directly on the embedding of the model rather than a dense classification are typically used in conjunction with clustering or a distance-based classification relative to ground truth embeddings per class. Instead, I tested the addition of sub-center ArcFace loss to the seesaw and custom venom losses.

LogitNorm [26] was applied to the logits during training before seesaw loss or venom loss were applied. This was done for parity with the models used for the FungiCLEF 2024 challenge [27]. LogitNorm increases class separation in the embedding space of the classifier as well as calibrating the model probabilities. Since the class with the highest predicted probability was selected as the classification in every case, probability calibration was assumed to be of no consequence for individual model classifications. However, since the probabilities are averaged in the model ensembles, it is possible that probability calibration could have an impact on the performance of ensembles.

3.3.3. Optimization and Training Details

The training paradigm used for the SnakeCLEF competition involved several key techniques and methodologies. The dataset was augmented using Trivial Augment and Random Erasing to improve the models' robustness. The AdamW optimizer [28] was used with a weight decay of 0.05. The learning rate was initially set to 1e-3 for the classification output dense layer with the pretrained model frozen for the first 5 epochs, then reduced to 5e-5. Training was conducted with a batch size of 40 for CAFormer-S18, 32

for Metaformer-0, and 24 for CAFormer-S36. CAFormer models were were used with weights pretrained on ImageNet-21K [29] while Metaformer-0 was used with weights pretrained on iNaturalist2021 [30]. A dropout rate of 0.2 was implemented between the dense output classification layer and the penultimate layer to prevent overfitting in all cases unless otherwise stated. Learning rate scheduling was employed, reducing the rate by a factor of 0.1 if the model did not improve the validation loss for 5 consecutive epochs. Early stopping was implemented to prevent overfitting and conserve computational resources. The models were fine-tuned using CAFormer-S18, and a 4x ensemble approach was adopted, utilizing different data splits to improve generalization.

3.4. Inference Techniques

During inference, several techniques were applied to maximize performance. Test-time augmentations were used, including horizontal flips and multi-instance averaging, to increase the robustness of predictions. The resolution of CAFormer-S18 and CAFormer-S36 models was adjusted by resizing from 384x384 to 576x576 for higher resolution inference. Additionally, ensemble averaging was employed, combining predictions from multiple models to improve overall accuracy. Models were ensembled by simple averaging of the prediction probabilities before selecting the class with the highest predicted probability as the prediction. These strategies collectively enhanced the model's performance during the inference phase as shown in Section 4.

Multi-crop refers to the generation and use of three overlapping crops that collectively ensure complete coverage of the entire image. The predicted class probabilities for each of these crops are then averaged to generate the final maximum probability classification. Horizontal flipping (hflip) augmentation involves taking the average predicted class probabilities from both the original and horizontally flipped version of each image in the same manner as multi-crop. Multi-instance refers to averaging the probabilities for each instance in cases where an observation has more than a single instance. In cases where multi-instance is not used, only the first instance for each observation is used to make each prediction. Image size refers to the inference image size that the image was resized to before a square center crop (or multiple square crops in the case of multi-crop) of the same resolution are taken.

Taken collectively, if multi-instance and hflip test time augmentations are both used with an ensemble of CAFormer models, the inference procedure would be as follows: Every image (instance) belonging to each observation would be 1) resized and center cropped to 576x576, and 2) horizontally flipped to keep both the original and mirrored image. Then, probabilities would be generated for each model for both flips of every instance. Finally, the simple average of all of these probabilities would be calculated to determine the class prediction based on the maximum class probability after averaging.

3.5. Model Architectures

An ensemble of CAFormer models [19] were used in the best-performing solution for this competition. These models balance computational efficiency and classification accuracy, making them suitable for both competitions. Notably, the CAFormer models performed consistently well across diverse datasets. A dropout rate of 0.2 was used between the dense output classification layer and the penultimate layer of the network.

3.6. Ensemble of Data Splits

Models were trained on four different training-validation data splits to increase diversity and decrease correlation between the errors of the models comprising the ensemble. The dataset is originally provided as three collections of observations: training, validation, and additional training observations for rare classes. In all cases, the additional training observations are considered part of the original training set. As such, the original dataset can be considered as being provided as a single training and validation split. The A, B, and C data splits were constructed by first combining the original training and validation data for the competition. The A split used the first 90% of the observations per class as the training

Table 1
FixRes fine-tuning. All models were trained with seesaw loss and venom loss and all inference was performed using horizontal flipping, multi-instance averaging, image size 576.

Models (data split)	FixRes	Public Track1↑	Public Track2↓	Public F1↑	Private Track1↑	Private Track2↓	Private F1↑
CAFormer-S18 (A)	-	**79.41**	**1052**	**29.74**	**78.32**	**2729**	**26.18**
CAFormer-S18 (A)	✓	78.08	1140	28.11	76.01	3155	24.24

Table 2
Inclusion of sub-center ArcFace loss. All models were trained with seesaw loss and venom loss and all inference was performed using horizontal flipping, multi-instance averaging, image size 576. * indicates that the model was trained with random erasing. CAFormer-S18 (A, B*, C*, D) was duplicated from the ensemble performance table to simplify comparisons. In cases where multiple data splits are denoted, this refers to an ensemble of multiple models, one per data split (e.g. CAFormer-S18 (A, B*, C*, D) refers to an ensemble of four CAFormer-S18 models, one trained on data split A, another on split B with random erasing, a third on split C with random erasing, and a final with split D).

Models (data split)	ArcFace	Public Track1↑	Public Track2↓	Public F1↑	Private Track1↑	Private Track2↓	Private F1↑
CAFormer-S18 (A, B*, C*, D)	-	**81.2**	**945**	**33.35**	**79.58**	**2557**	**30.29**
CAFormer-S18 (B*, C*, D) + CAFormer-S18 (A w/ ArcFace)	✓	81.09	952	33.22	79.18	2636	29.73

samples and the remaining 10% of the observations per class as the validation samples. In cases where there were fewer than 4 observations per class, all observations were used for training. The B split used the last 90% of samples for training and the C split used the middle 90% of samples for training, both with the same exception regarding tail classes with very few observations. Since the original training observations come before the original validation observations in this combined dataset, the A split is most similar of these 3 splits to the original training and validation split. The D split is the original training and validation split provided by the competition.

3.7. Computational Resources

All experiments were conducted on a single NVIDIA RTX 4090 graphics card, emphasizing the efficiency of our methodology given limited computational resources.

4. Results

My methodology demonstrated competitive performance in both SnakeCLEF and FungiCLEF 2024. For SnakeCLEF, my model achieved second place in all competition metrics on the public and private leaderboards, successfully differentiating between venomous and non-venomous species without the ability to overfit to geographic regions utilizing the metadata. For FungiCLEF, my approach excelled in recognizing unknown species while minimizing edible-poisonous misclassifications. My models achieved 1st place for Track1 classification score, macro F1, and Accuracy, while achieving competitive performance in the other two metrics [27].

FixRes involves not only inference at a higher resolution relative to training, but also fine-tuning the final layers of the model at the desired inference resolution without training augmentations. FixRes fine-tuning did not improve performance on any metric when inference was performed on a resolution of 576, as can be seen in Table 1.

Multiple loss functions were evaluated in addition to seesaw loss and the custom venom loss. One of the losses that was evaluated in addition to seesaw loss was sub-center ArcFace loss. Table 2 shows the addition of sub-center ArcFace loss to the training of one of the models in the ensemble. CAFormer-S18

Table 3
Balanced focal loss with higher dropout rate. In all cases, the models are CAFormer-S18 trained with venom loss using data split D. Inference was performed at a resolution of 768. Private leaderboard results omitted where unavailable. Track1, Track2, Public, and Private are abbreviated Trk1, Trk2, Pub, and Priv respectively.

Models (data split)	loss	dropout	Pub Trk1↑	Pub Trk2↓	Pub F1↑	Priv Trk1↑	Priv Trk2↓	Priv F1↑
CAFormer-S18 (D)	seesaw	0.2	**78.24**	**1125**	**27.50**	**76.95**	**2977**	**25.68**
CAFormer-S18 (D)	balanced focal	0.2	74.45	1361	20.91	-	-	-
CAFormer-S18 (D)	balanced focal	0.4	74.66	1353	21.92	73.4	3486	18.87

Table 4
Metaformer-0 vs CAFormer-S18. Both models were trained with seesaw loss and venom loss on data split D. Both models used an inference image resolution of 384. No test time augmentations were used by either model. CAFormer-S18 outperforms Metaformer-0 in every metric except private leaderboard Track 2.

Models (data split)	Public Track1↑	Public Track2↓	Public F1↑	Private Track1↑	Private Track2↓	Private F1↑
CAFormer-S18 (D)	**76.16**	**1251**	**23.29**	**73.73**	**3558**	**20.26**
Metaformer-0 (D)	74.53	1358	21.38	73.15	3554	18.51

Table 5
Addition of class weight to venom loss. All models were trained with seesaw loss and venom loss and used multi-instance averaging at inference. Two different combinations of data split, horizontal flipping ("hflip"), and image size are shown with different row colors denoting each. The best results for each combination of data split, image size, and hflip are shown in bold. In both the case of image size 768 without horizontal flipping and image size 576 with hflip, the addition of class weight to venom loss is harmful to all metrics. Track1, Track2, Public, and Private are abbreviated Trk1, Trk2, Pub, and Priv respectively.

Models (data split)	weighted venom loss	hflip	image size	Pub Trk1↑	Pub Trk2↓	Pub F1↑	Priv Trk1↑	Priv Trk2↓	Priv F1↑
CAFormer-S18 (A)	-	✓	576	**79.41**	**1052**	**29.74**	**78.32**	**2729**	**26.18**
CAFormer-S18 (A)	✓	✓	576	77.91	1149	27.51	75.67	3191	23.08
CAFormer-S18 (D)	-	-	768	**78.24**	**1125**	**27.50**	**76.95**	**2977**	**25.68**
CAFormer-S18 (D)	✓	-	768	76.28	1248	24.21	74.89	3273	20.60

with data split A was trained both with and without the sub-center ArcFace loss. In both cases, the classifications from the dense layer were used for predictions rather than utilizing the embeddings directly. The ensemble that contained a model with sub-center ArcFace loss had poorer performance across all metrics. This suggests that the addition of sub-center ArcFace loss to the seesaw and custom venom losses did not further mitigate the impact of the tail classes with very few observations despite the loss optimizing the separation of classes in the pre-classification model embedding.

Another loss that was evaluated for the multiclass classification was focal loss, which was paired with balanced sampling to directly address class imbalance. Focal loss with balanced sampling did not perform as well as seesaw loss, as can be seen in Table 3. Increasing the dropout rate for the penultimate layer may be slightly beneficial, with the greatest percent improvement in metrics being the F1 score (which increased 0.99), but this difference is trivial compared to the difference across all metrics for seesaw loss vs focal loss with balanced sampling. F1 increases nearly 7 points when focal loss with balanced sampling is replaced with seesaw loss.

Initial experiments with Metaformer-0 [19] showed that CAFormer-S18 gave better performance across all metrics. While Metaformer shows remarkable performance on fine-grained datasets, partic-

Table 6
Ensemble performance. CAFormer-S36 is also included in the comparison as a strong single model baseline. * indicates that the model was trained with random erasing. All models were trained with seesaw loss and venom loss and used horizontal flipping, multi-instance averaging, and an image size of 576 at inference. In cases where multiple data splits are denoted, this refers to an ensemble of multiple models, one per data split (e.g. CAFormer-S18 (A*, C*) refers to an ensemble of two CAFormer-S18 models, one trained on data split A and one trained on data split C, both with random erasing).

Models (data split)	Public Track1↑	Public Track2↓	Public F1↑	Private Track1↑	Private Track2↓	Private F1↑
CAFormer-S18 (A*, C*)	79.17	1073	30.19	77.55	2901	26.74
CAFormer-S18 (B*, C*)	80.2	1000	30.62	78.11	2798	27.64
CAFormer-S18 (A*, B*, C*)	80.78	965	31.76	78.42	2743	27.87
CAFormer-S18 (A, B*, C*, D)	**81.2**	**945**	**33.35**	79.58	2557	**30.29**
CAFormer-S18 (B*, C*, D) + CAFormer-S36 (D)	81.07	954	33.28	**79.96**	**2481**	30.2
CAFormer-S36 (D)	79.95	1013	29.69	79.18	2607	28.23

Table 7
Averaging test-time augmentations. All models were trained with seesaw loss and venom loss. An image resolution of 576 is used for inference in all cases. Row color is used to differentiate data split and random erasing combinations. Best performance for each metric is in bold per data split and random erasing combination. * indicates that the model was trained with random erasing. Track1, Track2, Public, and Private are abbreviated Trk1, Trk2, Pub, and Priv respectively.

Models (data split)	hflip	multi-crop	multi-instance	Pub Trk1↑	Pub Trk2↓	Pub F1↑	Priv Trk1↑	Priv Trk2↓	Priv F1↑
CAFormer-S18 (D)	-	-	-	78.39	1109	26.75	77.43	2857	25.02
CAFormer-S18 (D)	-	✓	✓	**79.94**	1024	**31.23**	**78.05**	**2782**	**27.08**
CAFormer-S18 (D)	-	-	✓	79.87	1025	30.57	77.71	2825	26.26
CAFormer-S18 (D)	✓	-	✓	79.92	**1023**	30.89	77.88	2798	26.60
CAFormer-S18 (A)	-	-	✓	79.19	1065	29.22	77.90	2807	26.07
CAFormer-S18 (A)	✓	-	✓	**79.41**	**1052**	**29.74**	**78.32**	**2729**	**26.18**
CAFormer-S18 (C*)	-	-	✓	78.06	1133	26.80	76.26	3088	24.34
CAFormer-S18 (C*)	✓	-	✓	**78.28**	**1118**	**27.09**	**76.19**	**3101**	**24.32**

Table 8
Inference resolution. The same model is used with different inference image resolutions (image size). The model was trained with seesaw loss and venom loss. No additional test time augmentations were performed (horizontal flip averaging, multiple crop averaging, multi-instance averaging). An image resolution of 576 dramatically outperforms 384, whereas it only slightly outperforms 768 on Track 1 and Track 2 metrics. An image resolution of 768 achieves the best F1 score of the three resolutions.

Models (data split)	image size	Public Track1↑	Public Track2↓	Public F1↑	Private Track1↑	Private Track2↓	Private F1↑
CAFormer-S18 (D)	384	76.16	1251	23.29	73.73	3558	20.26
CAFormer-S18 (D)	576	**78.39**	**1109**	26.75	**77.43**	**2857**	25.02
CAFormer-S18 (D)	768	78.24	1125	**27.50**	76.95	2977	**25.68**

ularly when metadata is available, it appears that CAFormer models may be more performant when metadata is unavailable.

Class weighted venom loss was evaluated as an alternative to the venom loss, since the custom venom loss did not account for class imbalance. In the case of two different data splits, the addition of this weight term to the venom loss negatively impacted all metrics. Results are shown in Table 5. Weighted venom loss did not improve the generalizability of the Track 2 score.

Table 9
Random erasing. All models were trained with seesaw loss and venom loss and utilized multi-instance averaging and image size 576 at inference. Despite slight performance improvements on a local validation set, random erasing appears to harm performance on the public and private leaderboards. * indicates that the model was trained with random erasing (RE).

Models (data split)	*RE	Public Track1↑	Public Track2↓	Public F1↑	Private Track1↑	Private Track2↓	Private F1↑
CAFormer-S18 (A)	-	**79.19**	**1065**	**29.22**	**77.9**	**2807**	**26.07**
CAFormer-S18 (A*)	✓	78.32	1121	27.99	75.59	3239	24.31

Several different model ensembles were evaluated based on different splits of the dataset, the inclusion of random erasing, and the CAFormer-S18 vs CAFormer-S36 architecture. Results are shown in Table 6. The best performing ensemble of models included on the private leaderboard comprised a CAFormer-S36 model trained on split D without random erasing, and three CAFormer-S18 models trained on splits B and C with random erasing and on split D without random erasing. The private leaderboard Track1 score of 79.96, Track 2 score of 2481, and F1 score of 30.2 achieved second place for all three metrics. An ensemble of all CAFormer-S18 models slightly outperformed this ensemble for private leaderboard F1 score (30.29 vs 30.2). Interestingly, this ensemble comprisingly solely CAFormer-S18 models performed best across all public leaderboard metrics. The all CAFormer-S18 ensemble has the same composition as the ensemble mentioned above with the exception of replacing the CAFormer-S36 model with a CAFormer-S18 model trained on data split A without random erasing. In all cases, there was a large disparity between the public and private leaderboard performance, particularly with respect to the Track 2 venomous → harmless confusion loss. In all cases, the private Track 2 loss was over two fold higher than the public Track 2 loss.

Several test-time augmentations were evaluated including horizontal flipping (hflip), averaging multiple crops (multi-crop), multi-instance averaging, and inference at a higher resolution than the training resolution. The results for all of these augmentations are summarized in Table 7 with the exception of increasing the inference image resolution, the results of which are shown in Table 8. Each of the averaging-based test-time augmentations improve performance, in isolation or in combination. Multi-instance is the most computationally demanding, but also provides the greatest lift in performance of hflip, multi-crop, and multi-instance. Hflip provides a similar lift to multi-crop, but involves doubling rather than tripling the number of images that must pass through the models. The most impactful augmentation for is inference at 576 image resolution instead of inference at the training resolution of 384, as shown in Table 8.

Random erasing was included in many of the models in an effort to increase the generalizability of the models. It appears that too much of the class-specific information was obscured by the erasure leading to a slight degradation in performance. As shown in Table 9, the public Track 1 score is worse by 0.87 while the private Track 1 score is worse by 2.31 when random erasing is included in the training augmentations.

Since the learning rate reduction and early stopping was decided based on validation loss, it was necessary to perform all training with a training-validation split. In order to utilize all the available data and to increase the diversity of the models in the final ensemble, different training-validation splits of the data were used to train otherwise identical models. Table 10 shows the impact of these different splits on the performance of the models. The difference between the best and worst performing splits is greater than the differences between the inclusion of LogitNorm (Table 12), random erasing (Table 9), horizontal flipping (Table 7), or multiple crops (Table 7). The difference was also greater than the difference between a larger ensemble and averaging multiple image resolutions (Table 11). This suggests that different splits of the data can have a significant impact on final performance of the models, particularly if individual models are used instead of being combined into an ensemble.

Averaging the predicted probabilities from multiple image (multi-res) resolutions was investigated as a test-time augmentation. However, since this requires performing inference through the same model

Table 10
Different data splits. All models were trained with seesaw loss and venom loss and utilized multi-instance averaging and image size 576. All the models were trained with random erasing.

Models (data split)	Public Track1↑	Public Track2↓	Public F1↑	Private Track1↑	Private Track2↓	Private F1↑
CAFormer-S18 (A*)	78.32	1121	27.99	75.59	3239	24.31
CAFormer-S18 (B*)	**79.47**	**1049**	**29.86**	**77.25**	**2928**	**26.43**
CAFormer-S18 (C*)	78.06	1133	26.80	76.26	3088	24.34

Table 11
Multi-res vs larger ensemble. The 4-model ensemble is duplicated from the ensemble performance table to facilitate a simpler comparison. All models were trained with seesaw loss and venom loss. Horizontal flipping, multi-instance averaging test-time augmentations were applied. At a similar compute budget, a larger ensemble outperforms multi-res. Track1, Track2, Public, and Private are abbreviated Trk1, Trk2, Pub, and Priv respectively. Multiple data splits are indicated per experiment. Each case refers to an ensemble of models. For example, "CAFormer-S18 (C*, D)" denotes an ensemble comprising two CAFormer-S18 models: one trained on data split C with random erasing and another trained on data split D without random erasing. * indicates that the model was trained with random erasing.

Models (data split)	image size	Pub Trk1↑	Pub Trk2↓	Pub F1↑	Priv Trk1↑	Priv Trk2↓	Priv F1↑
CAFormer-S18 (A, B*, C*, D)	576	**81.2**	**945**	**33.35**	**79.58**	**2557**	**30.29**
CAFormer-S18 (C*, D)	576, 652	80.41	998	32.65	78.56	2712	28.06

Table 12
LogitNorm ablation. The addition of LogitNorm does not appear to improve the performance on any metric. Both models have image resolution 576 but no other test-time augmentations. Both models were trained with random erasing. Private leaderboard results unavailable.

Models (data split)	LogitNorm	Public Track1↑	Public Track2↓	Public F1↑
CAFormer-S18 (A*)	✓	77.14	1190	25.14
CAFormer-S18 (A*)	-	**77.67**	**1155**	**25.71**

Table 13
Venom loss ablation. The addition of venom loss significantly improves the performance of the models across all metrics. Both models have image resolution 576 but no other test-time augmentations. Both models were trained with random erasing. The same baseline model results are shown in Table 12 for ease of comparison. Private leaderboard results unavailable.

Models (data split)	Venom loss	Public Track1↑	Public Track2↓	Public F1↑
CAFormer-S18 (A*)	✓	**77.14**	**1190**	**25.14**
CAFormer-S18 (A*)	-	74.46	1375	23.17

for n resolutions, the increased performance must be weighted against this increase in compute cost. Since multi-res requires inference through the same model n resolutions number of times, the compute cost should be comparable between an ensemble that is twice as large vs averaging two resolutions. Better performance is achieved across all metrics using a larger ensemble, as shown in Table 11.

All final models were trained with LogitNorm. To determine whether its inclusion was beneficial, an identical model was trained without LogitNorm and evaluated using the same settings. The inclusion of LogitNorm may slightly degrade performance of the models as shown in Table 12. Since it significantly improves performance on FungiCLEF [27], it may be of greater benefit to open-set classification, and

Table 14
Public leaderboard performance for teams with selected models. My models (bold) achieve 2nd place in all metrics.

Rank	Team Name	Track1↑	Track2↓	F1↑
1	upupup	85.63	687	43.66
2	**jack-etheredge**	**81.2**	**945**	**33.35**
3	ZCU-KKY	69.92	1660	15.44
4	Autohome AI	59.11	2431	11.59

Table 15
Private leaderboard performance for teams with selected models. My models (bold) achieve 2nd place in all metrics.

Rank	Team Name	Track1↑	Track2↓	F1↑
1	upupup	83.57	1840	34.58
2	**jack-etheredge**	**79.58**	**2557**	**30.29**
3	ZCU-KKY	67	4611	13.29
4	Autohome AI	54.15	7063	9.22

thus if the task will never be open-set, it seems that LogitNorm can be safely excluded from the training.

All final models were trained with venom loss. To determine whether its inclusion was beneficial, an identical model was trained without venom loss and evaluated using the same settings. As shown in Table 13, the custom venom loss improves performance on the Track2 metric as expected. Since the Track1 metric is also influenced by the venomous ⟶ harmless confusion, it is unsurprising that Track1 would improve with the inclusion of venom loss. What is more surprising is that the F1 score was improved by the venom loss. This shows that a real-world cost matrix for pairwise class confusion can be utilized without sacrificing overall classification performance. Future work could investigate how broadly applicable this is beyond this specific dataset.

4.1. Final model ensemble and leaderboard performance

The best performing ensembles both utilized horizontal flipping, multi-instance averaging, and a higher resolution image of 576x576 relative to the training resolution of 384x384. An ensemble of CAFormer-S18 models trained on data splits A and D without random erasing and data splits B and C with random erasing performed best for all public leaderboard metrics as well as F1 on the private leaderboard. However, this ensemble was outperformed for Track 1 and Track 2 on the private leaderboard by swapping the CAFormer-S18 model trained on data split A for a CAFormer-S36 model trained on data split D as shown in Table 6 and described previously in Section 4.

Table 14 shows the public leaderboard performance of each team and Table 15 shows the private leaderboard performance. In both cases, my method achieves 2nd place across all metrics. Notably, there is a larger gap between the performance of my models and 3rd place than the difference in performance of my models relative to 1st place for all metrics. Interestingly, there is a large disparity in the performance of Track2 between the public leaderboard and the private leaderboard for all participants. This suggests that either the public and private leaderboard have different data distributions or all competitors overfit their solutions to the public leaderboard. Since the other metrics do not show such a large disparity, this suggests that the ratio of difficult to classify venomous species may be greater in the private leaderboard test set.

5. Conclusions

I presented in this work a robust training and inference methodology that generalizes well across different fine-grained long-tailed image recognition tasks. The similarities between SnakeCLEF and FungiCLEF, such as asymmetric penalties for misclassification, highlight the effectiveness and generalizability of my approach. Differences, such as the lack of metadata in SnakeCLEF and the presence of unknowns in FungiCLEF, necessitated specific adjustments. Future work could explore few-shot learning techniques to further enhance performance for classes with few examples. Additional future work could investigate the potential for geographic metadata to increase model bias against the successful identification of invasive snake species in comparison to models not using that metadata. My approach's competitive performance on both SnakeCLEF and FungiCLEF 2024 suggests its potential applicability to other similar challenges.

Acknowledgments

The author would like to thank Jillian Etheredge for constructive criticism of the manuscript.

References

[1] I. Bolon, A. M. Durso, S. B. Mesa, N. Ray, G. Alcoba, F. Chappuis, R. R. d. Castañeda, Identifying the snake: First scoping review on practices of communities and healthcare providers confronted with snakebite across the world, PLOS ONE 15 (2020) e0229989. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0229989. doi:10.1371/journal.pone.0229989, publisher: Public Library of Science.

[2] R. R. de Castañeda, A. M. Durso, N. Ray, J. L. Fernández, D. J. Williams, G. Alcoba, F. Chappuis, M. Salathé, I. Bolon, Snakebite and snake identification: empowering neglected communities and health-care providers with AI, The Lancet. Digital Health 1 (2019) e202–e203. doi:10.1016/S2589-7500(19)30086-X.

[3] L. Picek, M. Hruz, A. M. Durso, Overview of SnakeCLEF 2024: Revisiting snake species identification in medically important scenarios, 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, J. Estopinan, 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. 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.

[6] L. Picek, R. Chamidullin, M. Hrúz, A. M. Durso, Overview of SnakeCLEF 2023: Snake Identification in Medically Important Scenarios, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.

[7] L. Picek, M. Šulc, R. Chamidullin, J. Matas, Overview of FungiCLEF 2023: Fungi Recognition Beyond 1/0 Cost, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.

[8] 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, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022.

[9] Z. Shi, H. Chen, C. Liu, J. Qiu, Metaformer Model with ArcFaceLoss and Contrastive Learning for SnakeCLEF2023 Fine-Grained Classification, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.

[10] 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, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, arXiv, 2023. URL: http://arxiv.org/abs/2307.09748, arXiv:2307.09748 [cs].

[11] H. Ren, H. Jiang, W. Luo, M. Meng, T. Zhang, Entropy-guided open-set fine-grained fungi recognition, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023. URL: https://api.semanticscholar.org/CorpusID:264441405.

[12] L. Bloch, A. Boketta, C. Keibel, E. Mense, A. Michailutschenko, O. Pelka, J. Rückert, L. Willemeit, C. Friedrich, Combination of Image and Location Information for Snake Species Identification using Object Detection and EfficientNets, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023. URL: https://api.semanticscholar.org/CorpusID:225071467.

[13] 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. URL: http://arxiv.org/abs/2008.10032. doi:10.48550/arXiv.2008.10032, arXiv:2008.10032 [cs].

[14] Y. Ho, S. Wookey, The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling, IEEE Access 8 (2020) 4806–4813. URL: https://ieeexplore.ieee.org/document/8943952/. doi:10.1109/ACCESS.2019.2962617.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal Loss for Dense Object Detection, 2018. URL: http://arxiv.org/abs/1708.02002. doi:10.48550/arXiv.1708.02002, arXiv:1708.02002 [cs] version: 2.

[16] 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 (CVPR), 2019.

[17] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A Simple Framework for Contrastive Learning of Visual Representations, in: Proceedings of the 37th International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607. URL: https://proceedings.mlr.press/v119/chen20j.html, iSSN: 2640-3498.

[18] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022).

[19] Q. Diao, Y. Jiang, B. Wen, J. Sun, Z. Yuan, MetaFormer: A Unified Meta Framework for Fine-Grained Recognition, 2022. URL: http://arxiv.org/abs/2203.02751. doi:10.48550/arXiv.2203.02751, arXiv:2203.02751 [cs].

[20] S. G. Müller, F. Hutter, TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation, 2021. URL: http://arxiv.org/abs/2103.10158. doi:10.48550/arXiv.2103.10158, arXiv:2103.10158 [cs].

[21] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random Erasing Data Augmentation, 2017. URL: http://arxiv.org/abs/1708.04896. doi:10.48550/arXiv.1708.04896, arXiv:1708.04896 [cs].

[22] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, 2018. URL: https://arxiv.org/abs/1710.09412. arXiv:1710.09412.

[23] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, J. Choe, CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features, 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 6022–6031. URL: https://ieeexplore.ieee.org/document/9008296/. doi:10.1109/ICCV.2019.00612, conference Name: 2019 IEEE/CVF International Conference on Computer Vision (ICCV) ISBN: 9781728148038 Place: Seoul, Korea (South) Publisher: IEEE.

[24] X. Liu, F. Shen, J. Zhao, C. Nie, RandoMix: a mixed sample data augmentation method with multiple mixed modes, Multimedia Tools and Applications (2024). URL: https://link.springer.com/10.1007/s11042-024-18868-8. doi:10.1007/s11042-024-18868-8.

[25] J. Deng, J. Guo, T. Liu, M. Gong, S. Zafeiriou, Sub-center ArcFace: Boosting Face Recognition by Large-Scale Noisy Web Faces, volume 12356, Springer International Publishing, Cham, 2020, pp. 741–757. URL: https://link.springer.com/10.1007/978-3-030-58621-8_43. doi:10.1007/978-3-030-58621-8_43, book Title: Computer Vision – ECCV 2020 Series Title: Lecture Notes in Computer Science.

[26] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, Y. Li, Mitigating Neural Network Overconfidence with Logit Normalization, 2022. URL: http://arxiv.org/abs/2205.09310. doi:10.48550/arXiv.2205.

09310, arXiv:2205.09310 [cs].
[27] J. N. Etheredge, OpenWGAN-GP for Fine-Grained Open-Set Fungi Classification, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.
[28] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, 2019. URL: http://arxiv.org/abs/1711.05101. doi:10.48550/arXiv.1711.05101, arXiv:1711.05101 [cs, math].
[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255. URL: https://ieeexplore.ieee.org/document/5206848. doi:10.1109/CVPR.2009.5206848, iSSN: 1063-6919.
[30] G. Van Horn, O. Mac Aodha, iNat Challenge 2021 - FGVC8. Kaggle. (2021). URL: https://kaggle.com/competitions/inaturalist-2021.