Solutions for Fine-grained and Long-tailed Snake Species Recognition in SnakeCLEF 2022

Cheng Zou[1], Furong Xu[1], Meng Wang[1], Wen Li[1] and Yuan Cheng[1]

[1] *Ant Group, Yuan Space 556 Xixi Road, Hangzhou, 310013, China*

Abstract

Automatic snake species recognition is important because it has vast potential to help lower deaths and disabilities caused by snakebites. We introduce our solution in SnakeCLEF 2022 for fine-grained snake species recognition on a heavy long-tailed class distribution. First, a network architecture is designed to extract and fuse features from multiple modalities, i.e. photograph from visual modality and geographic locality information from language modality. Then, logit adjustment based methods are studied to relieve the impact caused by the severe class imbalance. Next, a combination of supervised and self-supervised learning method is proposed to make full use of the dataset, including both labeled training data and unlabeled testing data. Finally, post processing strategies, such as multi-scale and multi-crop test-time-augmentation, location filtering and model ensemble, are employed for better performance. With an ensemble of several different models, a private score 82.65%, ranking the 3rd, is achieved on the final leaderboard.

Keywords

Snake Species Classification, Fine-grained Classification, Long-tailed Class Distribution, Self-supervised Pretraining, SnakeCLEF

1. Introduction

Snakebite is a global health problem, especially in remote geographic areas and developing countries. According to [1], in Asia, up to two million people are envenomed by snakes each year, while in Africa, there are about 435,000 to 580,000 snakebites annually that need treatment, for they can cause permanent disability and disfigurement. Taxonomic knowledge about snakes is crucial in diagnosis and medical response to snakebites, and accurate identification of the snake species is important for the appropriate treatment of snakebite victims since specific antivenoms are effective against specific venomous snakes [2]. Manual identification, e.g. training doctors on each species, is no easy feat, because there are more than 3,500 species of snakes, 600 of which are venomous [3]. So, building an automatic and robust image-based system for snake species identification has the greatest potential to save lives [4].

The difficulty of snake species identification, from both a human and a machine perspective, lies in the high intra-class and low inter-class variance in appearance, which may depend on geographic location, colour morph, sex, or age [4]. Sometimes, having the image alone is not enough, because many species are visually similar to other species, while introducing the

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

wuyou.zc@antgroup.com (C. Zou); booyoungxu.xfr@antgroup.com (F. Xu); darren.wm@antgroup.com (M. Wang); yinian.lw@alibaba-inc.com (W. Li); chengyuan.c@antgroup.com (Y. Cheng)

geographic origin of an observation can help to recognize considerably. The task of SnakeCLEF 2022 challenge [5], as part of the LifeCLEF 2022 [6, 7], aims to recognize a snake species ID given multiple photographs of the same individual and its corresponding geographic locality information.

In this paper, we introduce the solution of team "SAI" in SnakeCLEF 2022 for fine-grained snake species recognition on a severe long-tailed class distribution. First, as discussed above, using the image data alone is not enough, more cues are required for prediction, so a network architecture is designed to extract and fuse features from multiple modalities, i.e. photograph from visual modality and geographic locality information from language modality. Then, because of the heavy long-tailed class distribution, some logit adjustment based methods are studied to relieve the impact caused by the severe class imbalance, which significantly improves the performance. Next, to make full use of the dataset, including both labeled training data and unlabeled testing data, a combination of supervised and self-supervised learning method is utilized for pretraining. Finally, some post processing strategies are employed for better performance, including multi-scale and multi-crop test-time-augmentation, location filtering and model ensemble.

2. Related Work

2.1. Fine-grained Vision Classification

Snake species recognition is basically a task of fine-grained vision classification (FGVC). Modern fine-grained image classification methods can be divided into two parts, use image data only [8, 9, 10, 11], and use image data as well as extra data from other modalities [12, 13, 14, 15]. For the task of snake species recognition, one can solve it by using image data only, but a better choice is to use both image data and geographic locality information. In such studies with extra modalities, [12] is a classic method to introduce geographic information, [13] introduced spatio-temporal information into the network. MetaFormer [15] proposed a unified and flexibly framework to joint the visual appearance and various meta-information.

2.2. Long-tailed Distribution

Snake species recognition in real world is also a task of long-tailed image classification. Recently, long-tailed learning has received plenty of research interests, and among which there are mainly two kinds of solutions, one is post-hoc normalisation of the classification weights [16, 17, 18, 19], and the other is modification of the loss to account for varying class penalties [20, 21, 22]. [23] revisited the classic idea of logit adjustment based on the label frequencies, either applied post-hoc to a trained model, or enforced in the loss during training.

2.3. Snake Species Classification

Before us, there have been a few works tried to build automatic snake species recognition systems. In [1, 24], object detectors were first trained to reduce clutter and drop the unnecessary background for preprocessing, and then the detected snakes were classified by trained deep

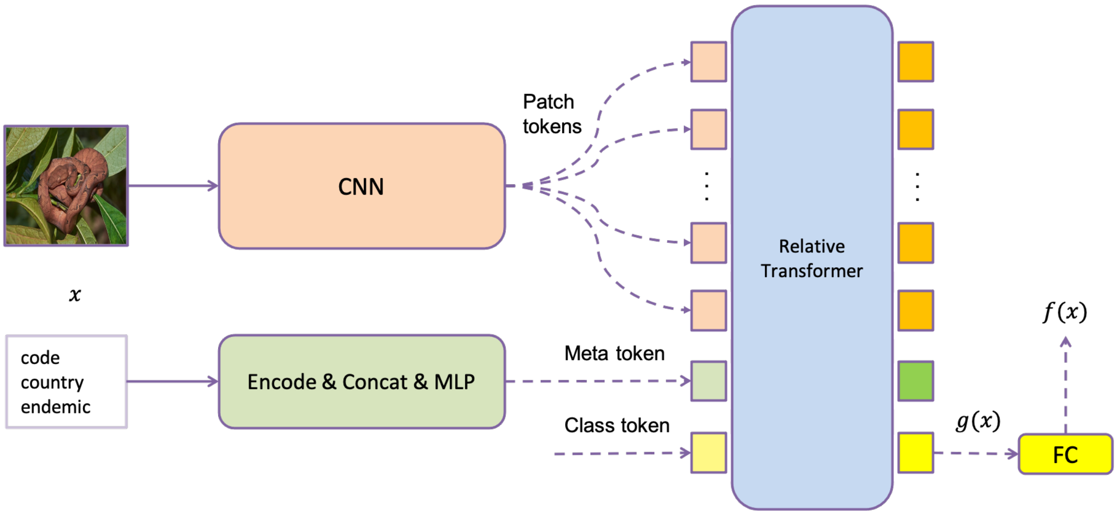

Figure 1: Overall network architecture. It has a hybrid framework, where CNN branch is used to extract vision features, MLP branch is used to encode meta data, and transformer is used to fuse vision features and meta information. $g(x)$ stands for extracted features and $f(x)$ stands for the predicted logits.

models. [25] extracted feature for each image with Inception ResNet V2 and concatenated it with geographic feature, then the concatenated features were classified using a lightweight gradient boost classifier. In [24], EfficientNet and Vision Transformer were trained, and the prior probabilities of location information were multiplied with the model predictions in a subsequent step. Besides, [2, 3, 24, 26] used multiple models to improve the performance.

3. Methodology

3.1. Overview

The proposed solution for fine-grained snake species recognition on a long-tailed class distribution mainly consists of four parts: 1) a network architecture to extract and fuse features from multiple modalities, 2) logit adjustment to relieve the impact caused by the severe class imbalance. 3) a combination of supervised and self-supervised learning method to make full use of both labeled training data and unlabeled testing data, 4) post processing strategies such as multi-scale and multi-crop test-time-augmentation, location filtering and model ensemble.

3.2. Network Architecture

The network design follows MetaFormer [15]. It has a hybrid framework, where on one branch CNN is used to extract vision features, on the other branch MLP is used to encode meta data, and transformer is later used to fuse vision features and meta information. The output feature map of CNN branch is transformed to a series of patch embeddings (denoted as patch tokens), along with the output embedding of MLP branch (denoted as meta token), along with the class token, are fed into the transformer layers for prediction. In the task of snake species recognition, the meta data available consists of location code, country and the tag of endemic.

3.3. Logit Adjustment for Long-tailed Learning

According to statistics, the dataset has a heavy long-tailed class distribution, where the most frequent species is represented by 6,472 images and the least frequent species by just 5 samples. It is also noteworthy that the evaluation metric, the Mean (Macro) F1-Score, weights recall and precision equally, and a good retrieval algorithm will maximize both precision and recall simultaneously. Thus, moderately good performance on both will be favoured over extremely good performance on one and poor performance on the other. If there is no action to deal with long-tailed distribution problem, the recall for rare classes would be very low, thus the overall performance could not be high.

Logit adjustment [23] is adopted to relieve the long-tailed distribution problem. In post-hoc adjustment, we predict the following instead of the original one,

$$\mathrm{argmax}_{y \in [L]} \exp(f(x)_y)/\pi_y^{\tau} = \mathrm{argmax}_{y \in [L]} f(x)_y - \tau \cdot \log \pi_y \tag{1}$$

In logit adjusted loss, it is combined with the soft target cross entropy loss,

$$L(f(x)_y, y) = -y \cdot \log \mathrm{softmax}(f(x)_y + \tau \cdot \log \pi_y) \tag{2}$$

where $\tau > 0$ is a hyper-parameter, $f(x)_y$ is the output logits of the neural network given input x, and π_y is the estimate of the class prior, e.g., the empirical class frequency on the training data. In practice, one can use either post-hoc logit adjustment or logit adjusted loss.

3.4. Combination of Supervised and Self-supervised Learning

In order to make full use of the dataset, including both labeled training data and unlabeled testing data, a combination of supervised and self-supervised training framework is proposed for pretraining. Specifically, we perform supervised learning on labelled training data and self-supervised learning (SSL) on all the data to obtain a set of task-related parameter initialization. Since the task provides images and meta information, we do pretraining for MetaFormer [15] with meta data, which can jointly take advantage of vision and meta-information.

To obtain a task-relevant initialization instead of imagenet initialization, we combine the self-supervised method SimCLR [27] with MetaFormer. Specifically, for each input pair of image and meta data, we randomly perform two data augmentations (Fig. 2) for the image, but no augmentation on meta data, then a classification loss SoftTargetCE [28, 29] (short for soft target cross entropy) is applied to the labeled data only, and a contrastive loss InfoNCE [30] is applied to all the data. Thus, the loss function for pretraining is,

$$L_{pretrain}(X, Y) = \mathrm{SoftTargetCE}_{Y \neq -1}(f(X), Y) + \alpha \cdot \mathrm{InfoNCE}(g([X; \bar{X}]), Y) \tag{3}$$

where $f(X)$ stands for logits, $g(X)$ stands for extracted features for a given batch X, and Y stands for its corresponding label, $Y = -1$ means unlabeled testing data. $\bar{X}$ is an augmented version of X. α is a hyper-parameter to balance the relative importance between supervised loss and self-supervised loss.

Algorithm 1: Location filtering

```
Input  : logits_adjusted, locations-to-species mapping $L2S$, meta data $Meta$
Output : predicted species name
idx_sort = np.argsort(logits_adjusted)[::-1]
for idx in idx_sort do
    species_name = species_list[idx]
    if species_name in L2S[Meta['code']] then
        break
return species_name
```

3.5. Post Processing

Two kinds of post processing strategies are mainly used, one is multi-scale, multi-crop and multi-model ensemble, the other is location filtering. For model ensemble, average logits is first calculated based on the output logits of different models, then the mean logits is adjusted to deal with long-tailed distribution problem. Specifically, for each single scale/crop input x_i, the $j-\mathrm{th}$ model outputs logits $f_j(x_i)$, then the final logits used for prediction is,

$$\mathrm{logits\_adjusted} = \mathrm{mean}(\sum_j \sum_i f_j(x_i)) - \tau \cdot \log \pi \tag{4}$$

Location Filtering. During training, it can be found that top-5 accuracy is much higher than that of top-1, which implies that if properly chosen from top predictions, the result could be better than the naive argmax one. A locations-to-species mapping is used to heuristically choose the best candidate prediction. For simplicity, we iterate through a sorted logits list until the first certain species appears, whose species name co-occurs with its location code. Algorithm 1 shows the numpy style pseudo code.

4. Experiments

4.1. Experimental Settings

Dataset: The dataset is based on 187,129 snake observations with 318,532 photographs belonging to 1,572 snake species and observed in 208 countries. The data were gathered from the online biodiversity platform iNaturalist. The provided dataset has a heavy long-tailed class distribution, where the most frequent species (Natrix natrix) is represented by 6,472 images and the least frequent species by just 5 samples.
Evaluation Metric: The evaluation metric for this competition is macro F_1-Score. The F_1 score for the i-th species is computed as,

$$F_1^i = 2 \cdot \frac{precision_i \cdot recall_i}{precision_i + recall_i} \tag{5}$$

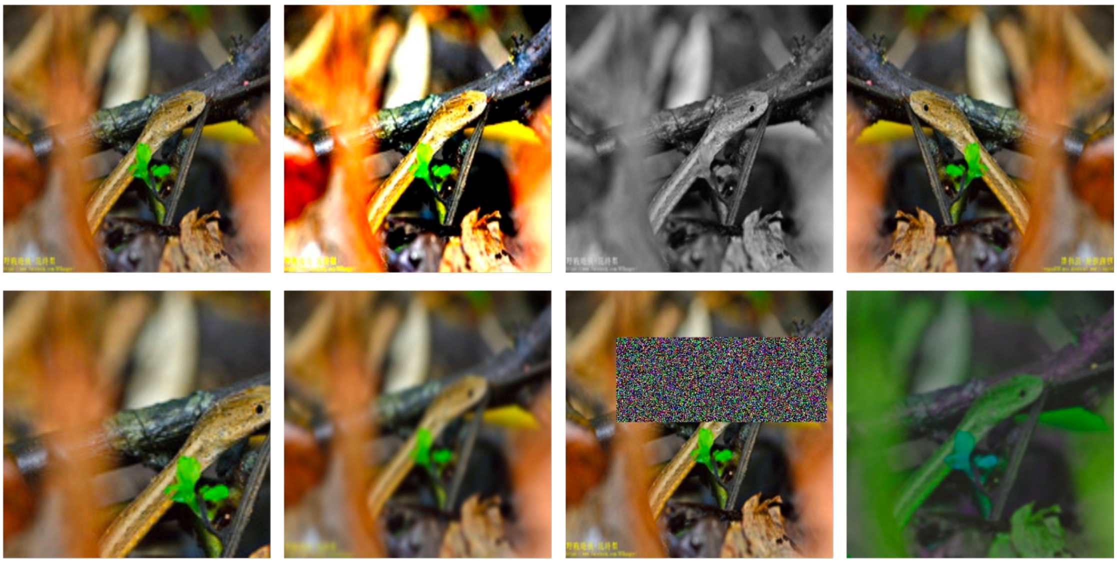

Figure 2: Augmentations for self-supervised learning. From left to right, top to bottom: original image without augmentation, color jittering, gray scale, horizontal flip, random resized crop, Gaussian blur, random erasing, a composed random augmentation of the above.

The macro F_1 is calculated by averaging the F_1 scores over all the species [1],

$$\text{macro } F_1 = \sum_{i=1}^{N} \frac{F_1^i}{N} \tag{6}$$

where N is the number of species. The macro F_1 score is not biased by class frequencies and is more suitable for the long-tailed class distributions. The F_1 metric weights recall and precision equally, and a good retrieval algorithm will maximize both precision and recall simultaneously. Thus, moderately good performance on both will be favoured over extremely good performance on one and poor performance on the other.

Implementation Details: MetaFormer [15] is selected as our base network. More specifically, we use MetaFormer-2 with extra meta information as input for both pretraining and finetuning. The meta data available for snake species consists of location code, country and the tag of endemic. Thus we construct a 2438-d vector to encode all of the above meta data, and then send it to an MLP to generate meta token. Hyper-parameter τ in Eq. 1 is set to 0.55, α in Eq. 3 is set to 0.001. In supervised and self-supervised pretraining, ImageNet-21k pretrained model is loaded as initialization, and the learning rate is initialized as 5×10^{-5}. In later finetuning, the learning rate is initialized as 5×10^{-6}. The models are trained on 8-NVIDIA A100-GPU machines for 300 epochs with a per GPU batch size of 72 for size 384 and 32 for size 512. The augmentations for self-supervised learning is illustrated in Fig. 2, while those for supervised learning and finetuning still follows MetaFormer.

Table 1

Representative experimental results about the importance of: 1) meta information, 2) different pretrained models, 3) solving long-tailed class distribution problem, 4) location filtering, 5) different input image size, 6) model ensemble. It's noteworthy that different parts in the table have different baselines, which is mainly caused by limited number of submissions.

Experimental Settings	macro F_1	Comments
The importance of meta information		
Vision Only	66.64 %	ImageNet-21k pretraining + Location filtering
Vision + Meta-208d	71.11 %	ImageNet-21k pretraining + Location filtering
Vision + Meta-2438d	74.19 %	ImageNet-21k pretraining + Location filtering
Different pretrained models		
ImageNet-1k [15]	63.76 %	Without location filtering or logit adjustment
ImageNet-21k [15]	66.64 %	Without location filtering or logit adjustment
iNat21 [15]	64.32 %	Without location filtering or logit adjustment
Supervised+SSL	68.83 %	Without location filtering or logit adjustment
Solving long-tailed class distribution		
None	74.19 %	ImageNet-21k pretraining + Location filtering
Seesaw [31]	76.49 %	ImageNet-21k pretraining + Location filtering
Logit adjusted loss [23]	78.01 %	ImageNet-21k pretraining + Location filtering
Post-hoc logit adjustment [23]	78.57 %	ImageNet-21k pretraining + Location filtering
Location filtering		
Without	64.32 %	iNat21 + No meta data + No logit adjustment
With	69.09 %	iNat21 + No meta data + No logit adjustment
Different input image size		
384	81.18 %	Meta+Logit adjustment+SSL+Location filtering
512	82.03 %	Meta+Logit adjustment+SSL+Location filtering
Model ensemble		
Model1	81.18 %	input image size 384
Model2	82.03 %	input image size 512
Ensemble	82.72 %	

4.2. Experimental Results

In this part, we report some representative experimental results, including: 1) the importance of meta information, 2) the importance of pretrained models, 3) the importance of solving long-tailed class distribution, 4) the importance of location filtering, 5) the importance of larger input size, 6) the importance of model ensemble.

The experimental results are summarized in Tab. 1. Academically, these comparisons here are not strict ablation studies, because they are obtained by few limited submissions during the competition. However, these results provided a meaningful and effective path to optimize the solution, which did improve the online judge performance.

As shown in Tab. 1, training the model with additional meta information significantly improves the performance from 66.64% to 74.19%, which indicates the importance of data from multiple modalities. For pretrained models, the proposed task-related supervised+SSL pretraining for Snake performs better than those commonly used ones, which indicates the importance of using unlabeled testing data. In long-tailed learning, logit adjustment is proved to be more effective, which improves the score from 74.19% to 78.57%. Location filtering, a task-specific post processing, using the statistics prior from the whole dataset to remove illegal predictions, improves the score from 64.32% to 69.09%. Also, training with larger input size 512 improves the performance from 81.18% to 82.03%.

With an ensemble of seven different models, we got private score 82.65% on the final leaderboard. The improvement is marginal compared to a single model, because the differences among these models are small, i.e., different epochs, different hyper-parameters. Interestingly, in late submission, we find it inferior to 82.72%, the ensemble of only two models.

5. Conclusion and Future Work

In this paper, we introduce our solution in SnakeCLEF 2022 for fine-grained snake species recognition on a severe long-tailed class distribution. Attentions have been mainly focused on four parts: 1) fusion of vision features and meta information, 2) solving long-tailed class distribution problem, 3) making full use of both the labeled training data and unlabeled testing data via supervised and self-supervised pretraining, 4) post processing strategies such as location filtering and model ensemble.

Though great improvements have been made, there still exist some actions of great potential for future work: 1) hard example mining for fine-grained and long-tailed dataset, 2) treating it as a retrieval task not a classification task, 3) using a snake detection model to get more precise bounding box for data pre-processing. We have tried some of the above but none of them contributed to the final performance during the competition, but they have great potential if further studied.

References

[1] R. Borsodi, D. Papp, Incorporation of object detection models and location data into snake species classification., in: CLEF (Working Notes), 2021, pp. 1499–1511.

[2] R. Chamidullin, M. Šulc, J. Matas, L. Picek, A deep learning method for visual recognition of snake species (2021).

[3] L. G. Coca, A. T. Popa, R. C. Croitoru, L. P. Bejan, A. Iftene, Uaic-ai at snakeclef 2021: Impact of convolutions in snake species recognition., in: CLEF (Working Notes), 2021, pp. 1540–1546.

[4] L. Picek, A. M. Durso, I. Bolon, R. R. de Castañeda, Overview of snakeclef 2021: Automatic snake species identification with country-level focus (2021).

[5] L. Picek, A. M. Durso, M. Hrúz, I. Bolon, Overview of SnakeCLEF 2022: Automated snake species identification on a global scale, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022.

[6] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Durso, H. Glotin, R. Planqué, W.-P. Vellinga, A. Navine, H. Klinck, T. Denton, I. Eggel, P. Bonnet, M. Šulc, M. Hruz, Overview of lifeclef 2022: an evaluation of machine-learning based species identification and species distribution prediction, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2022.

[7] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Durso, I. Bolon, et al., Lifeclef 2022 teaser: An evaluation of machine-learning based species identification and species distribution prediction, in: European Conference on Information Retrieval, Springer, 2022, pp. 390–399.

[8] T.-Y. Lin, A. RoyChowdhury, S. Maji, Bilinear cnn models for fine-grained visual recognition, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1449–1457.

[9] H. Zheng, J. Fu, Z.-J. Zha, J. Luo, Learning deep bilinear transformation for fine-grained image representation, Advances in Neural Information Processing Systems 32 (2019).

[10] J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, C. Wang, A. Yuille, Transfg: A transformer architecture for fine-grained recognition, arXiv preprint arXiv:2103.07976 (2021).

[11] 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.

[12] G. Chu, B. Potetz, W. Wang, A. Howard, Y. Song, F. Brucher, T. Leung, H. Adam, Geo-aware networks for fine-grained recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[13] 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.

[14] X. He, Y. Peng, Fine-grained image classification via combining vision and language, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5994–6002.

[15] 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).

[16] J. Zhang, L. Liu, P. Wang, C. Shen, To balance or not to balance: A simple-yet-effective approach for learning with long-tailed distributions, arXiv preprint arXiv:1912.04486 (2019).

[17] B. Kim, J. Kim, Adjusting decision boundary for class imbalanced learning, IEEE Access 8 (2020) 81674–81685.

[18] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, Y. Kalantidis, Decoupling representation and classifier for long-tailed recognition, arXiv preprint arXiv:1910.09217 (2019).

[19] H.-J. Ye, H.-Y. Chen, D.-C. Zhan, W.-L. Chao, Identifying and compensating for feature deviation in imbalanced deep learning, arXiv preprint arXiv:2001.01385 (2020).

[20] J. Tan, C. Wang, B. Li, Q. Li, W. Ouyang, C. Yin, J. Yan, Equalization loss for long-tailed object recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11662–11671.

[21] K. Cao, C. Wei, A. Gaidon, N. Arechiga, T. Ma, Learning imbalanced datasets with label-distribution-aware margin loss, Advances in neural information processing systems 32 (2019).

[22] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: Proceedings of the IEEE/CVF conference on computer vision and pattern

recognition, 2019, pp. 9268–9277.

[23] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, S. Kumar, Long-tail learning via logit adjustment, arXiv preprint arXiv:2007.07314 (2020).

[24] L. Bloch, C. M. Friedrich, Efficientnets and vision transformers for snake species identification using image and location information., in: CLEF (Working Notes), 2021, pp. 1477–1498.

[25] K. Desingu, M. Palaniappan, J. Kumar, Snake Species Classification Using Transfer Learning Technique, Technical Report, EasyChair, 2021.

[26] L. Kalinathan, P. Balasundaram, P. Ganesh, S. S. Bathala, R. K. Mukesh, Automatic snake classification using deep learning algorithm., in: CLEF (Working Notes), 2021, pp. 1587–1596.

[27] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: ICML, 2020.

[28] R. Müller, S. Kornblith, G. E. Hinton, When does label smoothing help?, Advances in neural information processing systems 32 (2019).

[29] C.-B. Zhang, P.-T. Jiang, Q. Hou, Y. Wei, Q. Han, Z. Li, M.-M. Cheng, Delving deep into label smoothing, IEEE Transactions on Image Processing 30 (2021) 5984–5996.

[30] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).

[31] 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.

A. Online Resources

The code and models are available at Google Drive.