

Metaformer Model with ArcFaceLoss and Contrastive Learning for SnakeCLEF2023 Fine-Grained Classification

Notebook for the <LifeCLEF> Lab at CLEF 2023

Zhennan Shi¹, Huazhen Chen², Chang Liu^{1,*} and Jun Qiu¹

¹Beijing Information Science and Technology University, Beijing, 100101, China

²Tianjin University, Tianjin, 300072, China

Abstract

Fine-Grained Visual Classification (FGVC) has always been a significant direction in computer vision. This paper describes our solution for the SnakeCLEF2023 competition. Firstly, we employ the MetaFormer architecture to process both the meta information and image information of the data. Secondly, we utilize ArcFace loss to address the issue of imbalanced data distribution. Next, we leverage the SimCLR contrastive learning method to allow the model to fully utilize the information from the dataset. Lastly, we employ data preprocessing techniques to enhance accuracy. Our approach achieved 88.30% on the private-score-track1 and 1613 on the private-score-track2, securing the third position.

Github: <https://github.com/BAOfanTing/SnakeCLEF2023>

Keywords

Fine-Grained Visual Classification, SnakeCLEF2023, Long Tail Distribution, Multimodal Backbone

1. Introduction

The human eye is an extraordinary organ capable of not only distinguishing broad categories of objects such as bikes, cats, and dogs but also further classifying them into specific subcategories like Garfield cats, Tabby cats, British Shorthair blue cats, and so on. The process of distinguishing different subcategories of cats within the broader category of cats is known as fine-grained visual classification (FGVC). After the rapid development of computer vision, people have attempted to use computer vision instead of human eyes for fine-grained visual classification.


FGVC has applications in our daily lives, industries, and businesses. For example, when taking a photograph of a bird, this technology can be utilized to identify the species of the bird [1]. When capturing an image of a car, this technology can be utilized to determine its brand, model, production year, and other relevant details [2]. FGVC technology is continuously evolving and holds the potential for even more applications in the future.

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

✉ zhennanshi@bistu.edu.cn (Z. Shi); huazhenchen@tju.edu.cn (H. Chen); liu.chang.cn@ieee.org (C. Liu); qiu.jun.cn@ieee.org (J. Qiu)



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

 CEUR Workshop Proceedings (CEUR-WS.org)

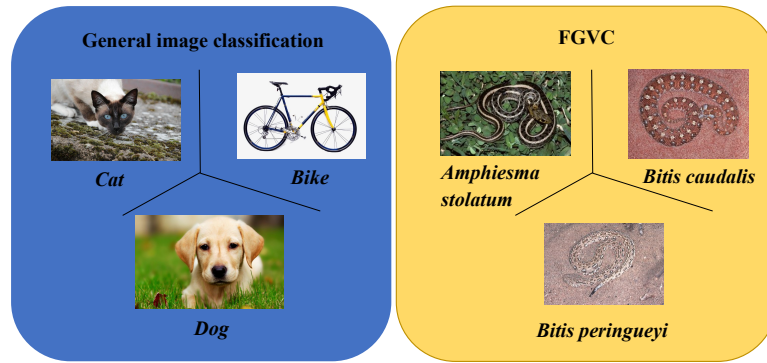


Figure 1: General image classification vs FGVC. We can see that general image classification focuses on distinguishing broad classes of species, while FGVC focuses on subtle differences between classes.

FGVC faces several challenges. Challenge 1: Large intra-class variations. Individuals within the same class can exhibit significant differences in appearance. As shown in Figure 2, the adult and sub-adult plumages of the *Red-crowned Crane* have distinct coloration[3]. Challenge 2: Small inter-class variations. Individuals within the different classes are very similar or closely related in certain aspects. For example, the *Coral Snake* and the *Milk Snake* have striking similarities in their appearance, both having bodies with black, red, and yellow bands. Their colors and patterns are almost identical, except for the arrangement of bands. This competition is the SnakeCLEF2023[4] competition in LifeCLEF2023[5, 6], which focuses on snake species classification, both humans and machines face difficulties due to the large intra-class variations and small inter-class variations in snake appearances. To distinguish them, it is necessary to learn features from the head shape, body shape, appearance, skin texture, and eye structure [7].

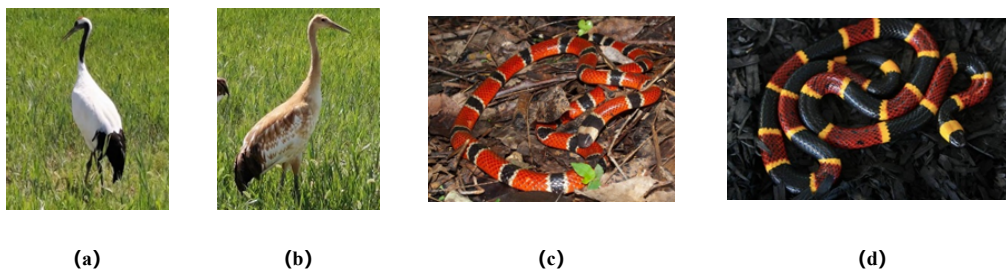


Figure 2: (a) *Adult Red-crowned Crane*, (b) *Sub-adult Red-crowned Crane*, (c) *Coral Snake*, (d) *Milk Snake*. (a) and (b) show the problem of large intra-class variations. (c) and (d) show the problem of small inter-class variations.

2. Related work

FGVC methods can be divided into two categories: methods that only use image information and methods that incorporate meta information, such as geographic location, gender, and shape.

Using only image information. The mentioned network structures proposed various approaches for image recognition. CMAL-Net [8] introduced a cross-layer mutual attention learning network that enabled the model to focus on discriminative regions. Yong Hou et al. [9] proposed a multilayer feature descriptors fusion CNN model that considered both second-order and first-order local feature descriptors from different layers. P-CNN [10] employed a system consisting of three modules: Squeeze-and-Excitation (SE) block, Part Localization Network (PLN), and Part Classification Network (PCN) to enhance fine-grained classification performance. RA-CNN [11] proposed a novel recurrent attention convolutional neural network that recursively learned discriminative region attention and region-based feature representation at multiple scales. MRA-CNN [12] improved RA-CNN by incorporating associations between multiple feature regions. It also introduced a feature scale-dependent (FSD) algorithm to select optimal features as input for the classifier. These network structures proposed different approaches and techniques.

Moving on to snake image recognition, Amiza Amir et al. [13] proposed an image-based method for the automatic identification of snake species, achieving an accuracy of 87%. However, it was limited to identifying only 22 species of snakes. Z. Yang et al. [14] proposed using a detection network to identify the snake's area before classifying its species. Due to limited training data, it could only distinguish 11 snake species. Patel et al. [15] modified and successfully implemented four region-based convolutional neural network (R-CNN) architectures for image classification, achieving an overall accuracy rate of around 75%.

Using meta information. Incorporating meta-information has proven effective. Zhai et al. [16] proposed a joint graph regularized heterogeneous metric learning (JGRHML) algorithm that integrated the structure of different media using joint graph regularization. Geo-Aware [17] systematically investigated various ways of incorporating geolocation information into fine-grained image classification, such as geolocation priors, post-processing, or feature modulation. CVL [18] proposed a two-stream model combining vision and language (CVL) for learning latent semantic representations. The visual stream learned deep features using convolutional neural networks, while the language stream utilized natural language descriptions to indicate distinctive parts or features of each image. The language stream provided a flexible and compact encoding method for salient visual aspects, aiding in the discrimination of subcategories. These models incorporated meta-information in different ways.

Regarding snake image recognition with meta-information, Bloch L et al. [19] utilized YOLOv5 as a detection network and incorporated meta-information such as geographic information for snake classification. However, the issue of imbalanced datasets remained unresolved. I Bolon et al. [20] used Vision Transformer as the backbone and integrated geographic information through binary masking.

In summary, these network structures and methods provide insights for improving image recognition and snake species classification. However, each method has its strengths and limitations. Some methods excel in snake species classification accuracy but are limited by imbalanced data, while others can incorporate rich meta-information but may require more

computational resources. Therefore, it is crucial to consider these pros and cons and choose the appropriate method based on specific requirements.

3. Method

3.1. Dataset

Through analyzing the sample distribution of different categories in the dataset, we discovered a significant class imbalance issue in the dataset. Some classes have as many as 2000 samples, while others have only a few samples. The distribution of images for each category is shown in Figure 3, which exhibits a long-tail distribution pattern. Furthermore, we compared the dataset with the dataset from the previous year and found an increase of over 200 snake species in Table 1. However, the training set has fewer samples this year, with a reduction of 100000 samples. The increased number of classes and the decreased training samples further increase the difficulty of this year's task.

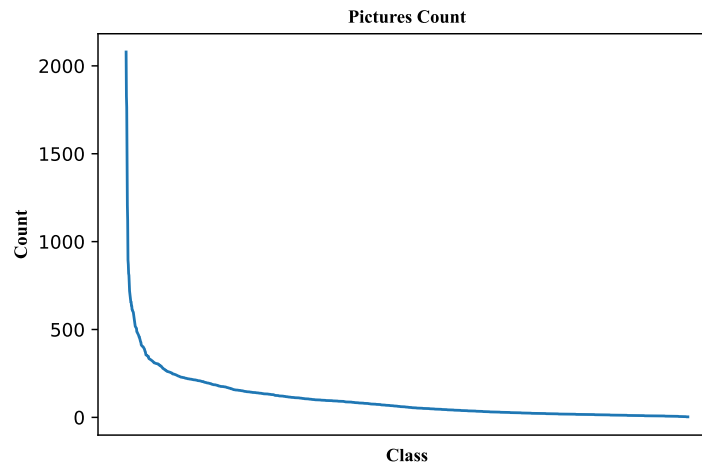


Figure 3: Samples count, with the y-axis representing the number of samples and the x-axis representing the categories in descending order of quantity. It can be seen that the dataset sample numbers are highly unbalanced.

Evaluation Metric: This year, the organizers calculated various metrics. First, they calculated the standard Acc and macro-averaged $F1$. In addition, they calculated the toxicant confusion error, which is the number of samples that confused toxicants as harmless divided by the number of toxicants in the test set.

First consider a function p such $p(s) = 1$ if species s is venomous, otherwise $p(s) = 0$. For

Table 1

Compositions of the 2022 and 2023 datasets were compared. We can find that the number of training samples was much reduced.

	2022	2023
Class	1572	1785
Train samples	Around 270000	Around 180000
Test samples	Around 48000	Around 14000
Meta information	Endemic, Binomial name, Country, Code	Endemic, Binomial name, Code
Image size	240×240, 500×500, Original Image	
Image dimension	RGB	

a correct species y and predicted species \hat{y} , the $lossL(y, \hat{y})$ is given as follows:

$$L(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{if } y \neq \hat{y} \text{ and } p(y) = 0 \text{ and } p(\hat{y}) = 0 \\ 2 & \text{if } y \neq \hat{y} \text{ and } p(y) = 0 \text{ and } p(\hat{y}) = 1 \\ 2 & \text{if } y \neq \hat{y} \text{ and } p(y) = 1 \text{ and } p(\hat{y}) = 1 \\ 5 & \text{if } y \neq \hat{y} \text{ and } p(y) = 1 \text{ and } p(\hat{y}) = 0 \end{cases} \quad (1)$$

The challenge metric private-score-track2 is sum of L over all test observations:

$$private - score - track2 = \sum_i L(y_i, \hat{y}_i) \quad (2)$$

The metric private-score-track1 is a weighted average between the macro $F1$ -score and the weighted accuracies of different types of confusion.

$$private - score - track1 = \frac{(w_1 F_1 + w_2 (100 - P_1) + w_3 (100 - P_2) + w_4 (100 - P_3) + w_5 (100 - P_4))}{\sum_i w_i} \quad (3)$$

where $w_1 = 1.0, w_2 = 1.0, w_3 = 2.0, w_4 = 5.0, w_5 = 2.0$, are the weights of individual terms. F_1 is the macro $F1$ -score, P_1 is the percentage of wrongly classified harmless species as another harmless species, P_2 is the percentage of wrongly classified harmless species as another venomous species, P_3 is the percentage of wrongly classified venomous species as another harmless species, and P_4 is the percentage of wrongly classified venomous species as another venomous species.

3.2. Model

For the competition, we utilized the MetaFormer [21] architecture, which is a useful network architecture for computer vision tasks. MetaFormer has designed a five-stage network structure. The first stage S_0 is a simple three-layer convolutional structure. S_1 and S_2 are MBConv blocks [22] with squeeze-excitation. MBConv blocks are based on an inverted residual mechanism and

a bottleneck design. The inverted residual structure is designed to provide higher nonlinear representation capability while keeping the model lightweight, and the bottleneck uses smaller intermediate layers to reduce computation. S3 and S4 are Transformer blocks with relative position bias. This bias alleviates the problem that the order of the tokens in the input sequence cannot be used in the self-attention operation.

We modify the input meta information of MetaFormer. Specifically, the meta information is modified to include the code, endemic, and binomial names. The workflow of MetaFormer is shown in Figure 4[23].

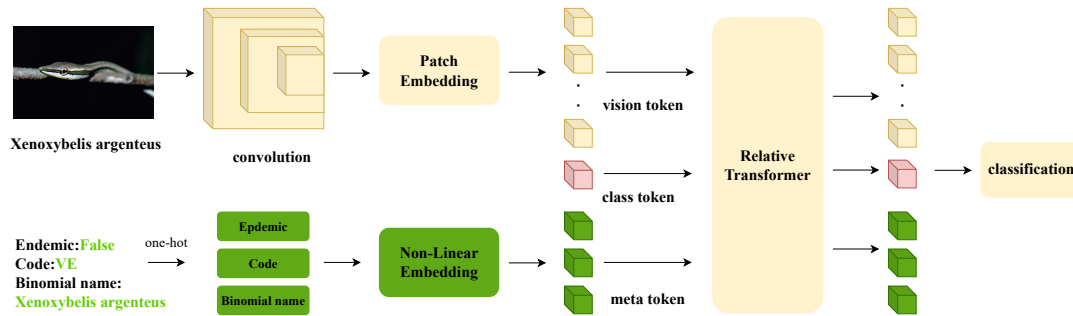


Figure 4: The model employs convolutional layers to extract visual features and then transforms the image features into visual tokens through Patch Embedding. The Code, Endemic, and Binomial name meta information are one-hot encoded and passed through a Non-Linear Embedding layer to obtain meta tokens. The visual tokens, meta tokens, and class tokens are fused using Relative Transformer Layers. The fused tokens are continuously aggregated in the following attention blocks. The final output class token is used for category prediction.

3.3. Long-Tailed loss

In order to address the issue of imbalanced sample distribution in the dataset, we employed a long-tail loss function. There are several commonly used loss functions to cope with the class imbalance problem. ArcFace loss [24] is designed for face recognition tasks. Other loss functions include the Class-Balanced Loss [25], which calculates a small neighborhood associated with each sample for computation. Seesaw Loss [26] mitigates the risk of increased misclassification due to gradient attenuation in negative samples. Equalization Loss v2 [27] discovers a novel gradient-guided reweighting mechanism.

Compared with other loss functions, ArcFace Loss optimizes the measure of feature space by introducing angle cosine values, so that the angles between feature vectors can reflect the similarity between samples. By normalizing the feature vectors and introducing an adjustable parameter, ArcFace Loss enhances the distinguishability of the features and reduces the difference in magnitude of the feature vectors. These features make ArcFace Loss a good performer in fine-grained classification tasks, so we adopt it as our loss function.

ArcFace loss's formula is as follows:

$$L_{\text{arcface}}(\theta) = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(\cos(\theta_{y_i} + m))}}{e^{s(\cos(\theta_{y_i} + m))} + \sum_{j=1, j \neq y_i}^n e^{s \cos \theta_j}} \quad (4)$$

where N represents the number of samples, n represents the number of classes, y_i is the true class of the i -th sample, θ_{y_i} is the angle between the feature vector of the i -th sample and its true class, and θ_j is the angle between the feature vector of the i -th sample and the j -th class. s represents the scale parameter. m represents the margin parameter, which is the inter-class distance. In both the numerator and denominator, a margin is added to each individual term, represented by α and β in Figure 5. By calculating the loss with this margin, the margin increases gradually, resulting in the compression of the region for each class. This effectively enlarges the inter-class distance while reducing the intra-class distance.

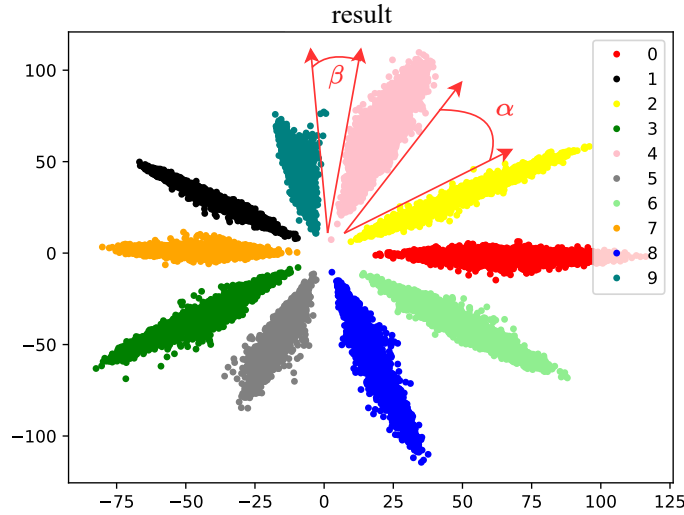


Figure 5: The results of ArcFace loss on the MNIST dataset have shown significant improvements in classification performance. Each point in the figure represents a sample, and each color represents a class.

3.4. SimCLR self-supervised learning

SimCLR [28] is a self-supervised learning framework that has gained significant attention in the field of computer vision. The main goal of SimCLR is to learn meaningful representations of unlabeled data by maximizing agreement between different augmented views of the same image while minimizing agreement between views of different images. The framework's workflow is shown in Figure 6. By maximizing the similarity between the representations of positive image

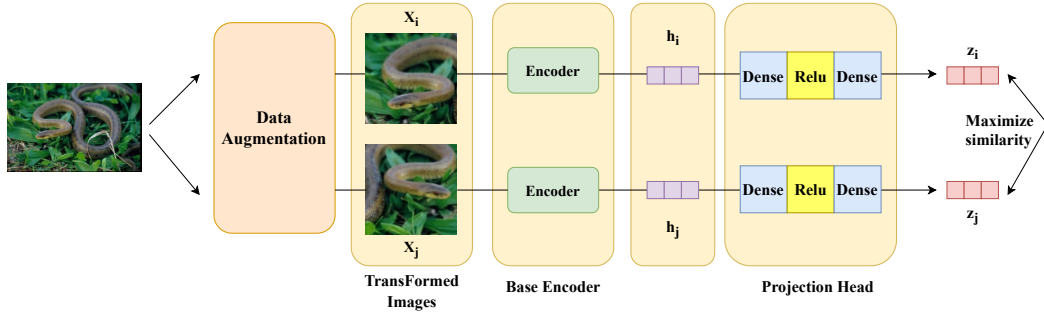


Figure 6: The framework of SimCLR involves the following steps: 1) Take an input image and apply random transformations to generate two augmented images, denoted as x_i and x_j . 2) Pass the augmented images through an encoder to obtain image representations, h_i and h_j respectively. 3) Use a non-linear fully connected layer to map the data to the representation space, z . 4) Maximize the similarity between z_i and z_j , the representations of the positive image pair.

pairs, SimCLR encourages the model to learn meaningful and discriminative features of images.

We also adopted the InfoNCE [29] loss function, commonly used in contrastive learning. It places positive samples in the numerator and negative samples in the denominator, aiming to maximize the similarity of positive samples while minimizing the similarity of negative samples. The formula for InfoNCE is expressed as follows:

$$L_q = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^k \exp(q \cdot k_i / \tau)} \quad (5)$$

where k_+ represents the positive samples for q while the remaining k denotes the negative samples. τ refers to the temperature hyperparameter, which we set to 0.25 in our subsequent experiments.

3.5. Pre-post-process

Data Preprocessing: In order to enhance the generalization ability of the method, we applied image augmentation techniques to the input images, including resizing and center cropping. We also introduced random vertical flipping, random horizontal flipping, and random 45-degree rotation to augment the training dataset.

Data Postprocessing: During the testing phase, we employed the Test Time Augmentation (TTA) strategy. TTA augments the input test data with operations such as expansion, flipping, and rotation to obtain a set of data for an image and take the mean value of the final predictions. For the csv file of the test dataset, each observation-id corresponds to multiple images, and the model predicts one class for each image, so there may be multiple different predicted classes for each observation id. For each observation id, we select the class with the most occurrences as the final prediction result. We also adopt an integrated learning strategy to improve the accuracy by fusing the results of MetaFormer-0, Metaformer-2, and Metaformer-2 with SimCLR

three models. For each observation-id, we select the class with the most occurrences from the csv files generated by the three models as the final prediction result.

4. Experiments

4.1. Implementation Details

We conducted all the experiments with one NVIDIA GeForce RTX 3090. We used the AdamW[30] optimizer with a weight decay of 0.05 and a base learning rate of 5e-5. The batch size is determined by the maximum number that the GPU can handle, usually an integer multiple of 2. We use the batch size of 22. And the initial learning rate is modified according to batch size (That is, the learning rate multiplied by the batch size multiplied by the number of GPUs divided by 512.). Also, set the number of training epochs to 100. During training, we used data augmentation and rotation. We also use the CosineLRScheduler of the timm library to modify the learning rate. At first the learning rate increases from the warmup learning rate 5e-8 to the base learning rate, and then enters the cosine annealing phase, where the learning rate is adjusted by the cosine function, decreasing with increasing epochs until it decreases to min learning rate 5e-7.

4.2. Ablation Studies

Firstly, we attempted the approach based on EfficientNet [31], the loss function used was CrossEntropy Loss. However, we found the model couldn't incorporate meta information, resulting in lower accuracy than expected. Therefore, we switched to the MetaFormer model. Comparison of above two models is shown in Table 2. The results demonstrated the superiority of the MetaFormer backbone. Thus, we use MetaFormer-2 as the backbone of our method.

Table 2

Results of different models. Accuracy is the number of samples correctly predicted in the validation dataset divided by the total number of samples in the validation dataset.

Model	Accuracy	Size of image	Parameters
EfficientNet-B7	0.675	384×384	66M
MetaFormer-0	0.726	384×384	28M
MetaFormer-2	0.764	384×384	81M

To further enhance accuracy, we employed the SimCLR contrastive learning method to train the model and fine-tuned it with an input of size 512x512. As shown in Table 3, our highest accuracy was achieved using the SimCLR contrastive learning method. The results show the effectiveness of the contrastive learning method SimCLR.

5. Conclusion

In this paper, we presented our solution for the snakeCLEF2023 competition. We adopted the MetaFormer architecture to incorporate effective meta information, utilized the ArcFace loss

Table 3

Results of SimCLR and input of size 512x512. We found that using 512x512 large size fine tuning does not work very well.

Model	Accuracy	Epochs
SimCLR+MetaFormer-2	0.838	100
MetaFormer-2	0.734	100

function to address the issue of long-tail data distribution, employed the SimCLR contrastive learning method with pre-trained models to improve accuracy, and applied data augmentation techniques to enhance the model's generalization ability.

Our solution achieved 88.30% on the private-score-track1 and 1613 on the private-score-track2, securing the third position. Due to time and resource constraints, we were unable to further explore the long-tail loss function and new contrastive learning methods. However, the results demonstrate the effectiveness of our approach and highlight the potential for further improvements. Future work could involve the following aspects: 1) Exploring or designing new long-tail loss functions. 2) Investigating other contrastive learning methods, such as the MAE [32] pre-training method, to further enhance the performance of pre-training.

References

- [1] A. Miyaguchi, J. Yu, B. Cheungvivatpant, D. Dudley, A. Swain, Motif mining and unsupervised representation learning for birdclef 2022, arXiv preprint arXiv:2206.04805 (2022).
- [2] B. Cynthia Sherin, K. Jayavel, Effective vehicle classification and re-identification on stanford cars dataset using convolutional neural networks, in: Proceedings of 3rd International Conference on Artificial Intelligence: Advances and Applications: ICAIAA 2022, Springer, 2023, pp. 177–190.
- [3] X. Wang, X. Sun, M. Cao, Y. Zhang, M. Zhu, A multi-scale approach to investigating the wintering habitat selection of red-crowned cranes in the yancheng nature reserve, china., Pakistan Journal of Zoology 48 (2016).
- [4] L. Picek, M. Šulc, Chamidullin, A. M. Durso, Overview of snakeclef 2023: Snake identification in medically important scenarios, in: CLEF 2023-Conference and Labs of the Evaluation Forum, 2023.
- [5] A. Joly, H. Goëau, S. Kahl, L. Picek, C. Botella, D. Marcos, M. Šulc, M. Hruz, T. Lorieul, S. S. Moussi, M. Servajean, B. Kellenberger, E. Cole, A. Durso, H. Glotin, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, I. Eggel, P. Bonnet, H. Müller, Lifeclef 2023 teaser: Species identification and prediction challenges, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 568–576.
- [6] A. Joly, C. Botella, L. Picek, S. Kahl, H. Goëau, B. Deneu, D. Marcos, J. Estopinan, R. Chamidullin, M. Šulc, M. Hruz, M. Servajean, B. Kellenberger, E. Cole, H. Glotin, et al., Overview of lifeclef 2023: evaluation of ai models for the identification and prediction of

- birds, plants, snakes and fungi, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 14th International Conference of the CLEF Association, CLEF 2023*, Thessaloniki, Greece, September 18–23, 2023, Proceedings, Springer, 2023.
- [7] N. I. Progga, N. Rezoana, M. S. Hossain, R. U. Islam, K. Andersson, A cnn based model for venomous and non-venomous snake classification, in: *Applied Intelligence and Informatics: First International Conference, AII 2021*, Nottingham, UK, July 30–31, 2021, Proceedings 1, Springer, 2021, pp. 216–231.
 - [8] D. Liu, L. Zhao, Y. Wang, J. Kato, Learn from each other to classify better: Cross-layer mutual attention learning for fine-grained visual classification, *Pattern Recognition* 140 (2023) 109550.
 - [9] Y. Hou, H. Luo, W. Zhao, X. Zhang, J. Wang, J. Peng, Multilayer feature descriptors fusion cnn models for fine-grained visual recognition, *Computer Animation and Virtual Worlds* 30 (2019) e1897.
 - [10] J. Han, X. Yao, G. Cheng, X. Feng, D. Xu, P-cnn: Part-based convolutional neural networks for fine-grained visual categorization, *IEEE transactions on pattern analysis and machine intelligence* 44 (2019) 579–590.
 - [11] 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.
 - [12] S. Fayou, H. Ngo, Y. Sek, Combining multi-feature regions for fine-grained image recognition, *Int. J. Image Graph. Signal Process* 14 (2022) 15–25.
 - [13] A. Amir, N. A. H. Zahri, N. Yaakob, R. B. Ahmad, Image classification for snake species using machine learning techniques, in: *Computational Intelligence in Information Systems: Proceedings of the Computational Intelligence in Information Systems Conference (CIIS 2016)*, Springer, 2017, pp. 52–59.
 - [14] Z. Yang, R. Sinnott, Snake detection and classification using deep learning (2021).
 - [15] A. Patel, L. Cheung, N. Khatod, I. Matijosaitiene, A. Arteaga, J. W. Gilkey Jr, Revealing the unknown: real-time recognition of galápagos snake species using deep learning, *Animals* 10 (2020) 806.
 - [16] X. Zhai, Y. Peng, J. Xiao, Heterogeneous metric learning with joint graph regularization for cross-media retrieval, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, 2013, pp. 1198–1204.
 - [17] 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.
 - [18] 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.
 - [19] L. Bloch, J.-F. Böckmann, B. Bracke, C. M. Friedrich, Combination of object detection, geospatial data, and feature concatenation for snake species identification (2022).
 - [20] I. Bolon, L. Picek, A. M. Durso, G. Alcoba, F. Chappuis, R. Ruiz de Castañeda, An artificial intelligence model to identify snakes from across the world: Opportunities and challenges for global health and herpetology, *PLoS neglected tropical diseases* 16 (2022) e0010647.
 - [21] 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).
- [22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.
 - [23] Y. Huang, A. Huang, W. Zhu, Y. Fang, J. Feng, Explored an effective methodology for fine-grained snake recognition, arXiv preprint arXiv:2207.11637 (2022).
 - [24] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4690–4699.
 - [25] 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.
 - [26] 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.
 - [27] J. Tan, X. Lu, G. Zhang, C. Yin, Q. Li, Equalization loss v2: A new gradient balance approach for long-tailed object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 1685–1694.
 - [28] Y. Liu, W. Tu, S. Zhou, X. Liu, L. Song, X. Yang, E. Zhu, Deep graph clustering via dual correlation reduction, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 7603–7611.
 - [29] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).
 - [30] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).
 - [31] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.
 - [32] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.