

# Watch out Venomous Snake Species: A Solution to SnakeCLEF2023

Feiran Hu<sup>1</sup>, Peng Wang<sup>1</sup>, Yangyang Li<sup>1</sup>, Chenlong Duan<sup>1</sup>, Zijian Zhu<sup>1</sup>, Fei Wang<sup>2</sup>, Faen Zhang<sup>2</sup>, Yong Li<sup>1,\*</sup> and Xiu-Shen Wei<sup>1,\*</sup>

<sup>1</sup>*School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China*

<sup>2</sup>*Qingdao AInnovation Technology Group Co., Ltd.*

## Abstract

The SnakeCLEF2023 competition aims to the development of advanced algorithms for snake species identification through the analysis of images and accompanying metadata. This paper presents a method leveraging utilization of both images and metadata. Modern CNN models and strong data augmentation are utilized to learn better representation of images. To relieve the challenge of long-tailed distribution, seesaw loss [1] is utilized in our method. We also design a light model to calculate prior probabilities using metadata features extracted from CLIP [2] in post processing stage. Besides, we attach more importance to venomous species by assigning venomous species labels to some examples that model is uncertain about. Our method achieves 91.31% score of the final metric combined of F1 and other metrics on private leaderboard, which is the 1st place among the participators. The code is available at <https://github.com/xiaoxsparrow/CLEF2023>.

## Keywords

Snake Species Identification, Fine-grained image recognition, Long-tailed, Metadata, SnakeCLEF

## 1. Introduction

Fine-grained visual categorization is a well-established and pivotal challenge within the fields of computer vision and pattern recognition, serving as the cornerstone for a diverse array of real-world applications [3]. The SnakeCLEF2023 competition, co-hosted as an integral part of the LifeCLEF2023 lab within the CLEF2023 conference, and the FGVC10 workshop in conjunction with the esteemed CVPR2023 conference, is geared towards advancing the development of a robust algorithm for snake species identification from images and metadata. This objective holds profound significance in the realm of biodiversity conservation and constitutes a crucial facet of human health preservation.

In this paper, we introduce a method that addresses the recognition of snake species by leveraging both metadata and images. ConvNeXt-v2 [4] and CLIP [2] are used to extract images features and metadata features separately, and the image features and text features are concatenated to be input of MLP classifier, thus getting better representation of examples and

---


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

\*Corresponding author.

✉ hufr@njust.edu.cn (F. Hu); wangpeng@njust.edu.cn (P. Wang); lyylyyi599@njust.edu.cn (Y. Li); duancl@njust.edu.cn (C. Duan); zhuzj@njust.edu.cn (Z. Zhu); wangfei@ainnovation.com (F. Wang); zhangfaen@ainnovation.com (F. Zhang); yong.li@njust.edu.cn (Y. Li); weixs.gm@gmail.com (X.-S. Wei)



© 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)

recognition results. Seesaw loss [1] are utilized in our method, thereby alleviating the long-tailed distribution problem. Notably, our proposed method takes into careful consideration the critical real-world need to distinguish venomous and harmless snake species by using the Real-World Weighted Cross-Entropy (RWWCE) loss [5] and post-processing, resulting in exemplary performance surpassing that of other solutions presented in this year’s competition. Experiments and competition results show that our method is effective in snake species recognition task.

The subsequent sections of this paper provide a comprehensive overview of the key aspects. Section 2 introduces the competition challenges and datasets, accompanied by the examination of the evaluation metric utilized. Section 3 describes our proposed methodologies, offering a comprehensive and detailed introduction to the techniques. Section 4 presents the implementation details, alongside a comprehensive analysis of the principal outcomes achieved. Finally, Section 5 concludes this paper by summarizing the key findings and offering future research directions.

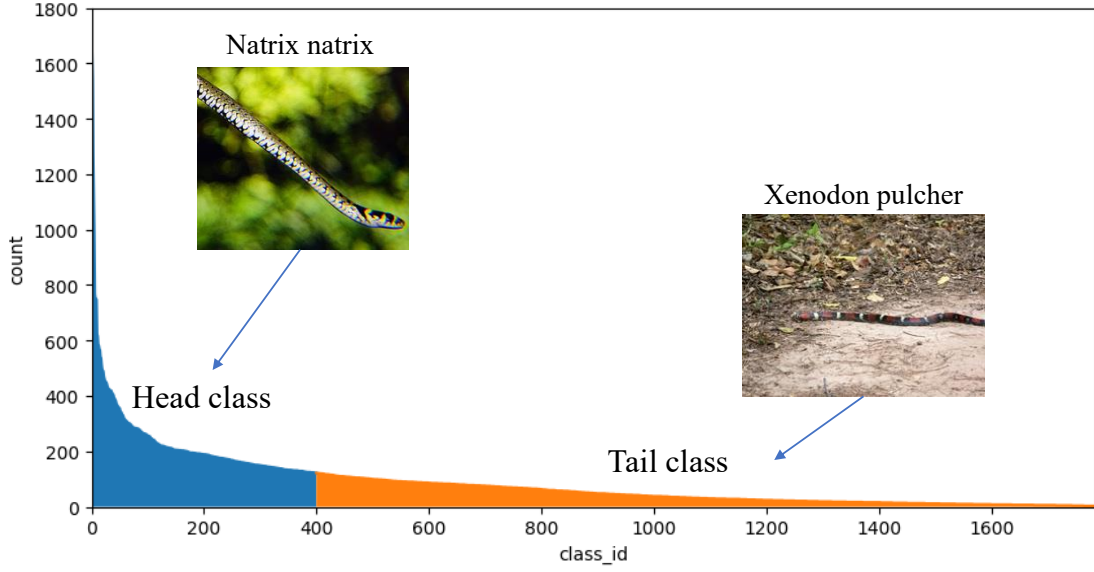
## 2. Competition Description

Understanding datasets and metrics is an essential requirement for engaging in a machine learning competition. Within this section, we aim to introduce our comprehension of the datasets and provide overview of the evaluation metrics employed by the competition organizers.

### 2.1. Challenges of the Competition

Past iterations of this competition have witnessed remarkable accomplishments by machine learning models [6, 7, 8, 9, 10, 11]. To further enhance the competition’s practical relevance and address the exigencies faced by developers, scientists, users, and communities, such as addressing post-snakebite incidents, the organizers have imposed more stringent constraints. The ensuing challenges of this year’s competition can be summarized as follows:

- Fine-grained image recognition: The domain of fine-grained image analysis has long posed a challenging problem within the FGVC workshop, deserving further investigation and study.
- Utilization of metadata: The incorporation of metadata, particularly pertaining to the geographical distribution of snake species, plays a vital role in their classification. Such metadata is commonly employed by individuals to identify snakes in their daily lives. Hence, utilization of location metadata holds significance and needs careful consideration.
- Long-tailed distribution: Long-tailed distributions are common in real-world scenarios, and the distribution of snake species is no exception.
- Identification of venomous and harmless species: The distinction between venomous and harmless snake species is meaningful, as venomous snake bites lead to large number of death each year. Consequently, leveraging deep learning methodologies to address this problem is of paramount urgency.
- Model size limitation: A strict limitation has been imposed on the model size, constraining it to a maximum of 1GB.



**Figure 1:** Long-tailed distribution of the SnakeCLEF2023 training dataset. The blue color means head classes, which means most images in the dataset belong to these classes. The orange color means tail classes, which means most classes in the dataset are tail classes.

## 2.2. Dataset

The organizers provide a dataset, consisting of 103,404 recorded snake observations, supplemented by 182,261 high-resolution images. These observations encompass a diverse range of 1,784 distinct snake species and have been documented across 214 geographically varied regions.

It is worth to note that the provided dataset is in a heavily long-tailed distribution, as shown in Fig. 1. In this distribution, the most frequently encountered species have 1,262 observations consists of 2,079 accompanying images. However, the least frequently encountered species is captured by a mere 3 observations, showing its exceptional rarity within the dataset.

## 2.3. Evaluation Metric

In addition to the conventional evaluation metrics of Accuracy (Acc) and Mean F1-Score, this year’s competition incorporates a novel evaluation metric, denoted as “public\_score\_track1” on the leaderboard. This metric combines the F1-Score with an assessment of the confusion errors related to venomous species. It is calculated as a weighted average, incorporating both the macro F1-score and the weighted accuracy of various types of confusions:

$$M = \frac{w_1 F_1 + w_2 (100 - P_1) + w_3 (100 - P_2) + w_4 (100 - P_3) + w_5 (100 - P_4)}{\sum_i^5 w_i}, \quad (1)$$

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. The metric incorporates several percentages, namely  $F_1$  representing the macro F1-score,  $P_1$  denoting the percentage of harmless species misclassified as another harmless species,

$P_2$  indicating the percentage of harmless species misclassified as a venomous species,  $P_3$  reflecting the percentage of venomous species misclassified as another harmless species, and  $P_4$  representing the percentage of venomous species misclassified as another venomous species. This metric is bounded below by 0% and above by 100%. The lower bound is attained when all species are misclassified, including misclassification of harmless species as venomous and vice versa. Conversely, if the F1-score reaches 100%, indicating correct classification of all species, each  $P_i$  value must be zero, leading to an overall score of 100%.

### 3. Method

In this section, we shall introduce the methodologies employed to address the task of snake species classification.

#### 3.1. Data Preprocessing

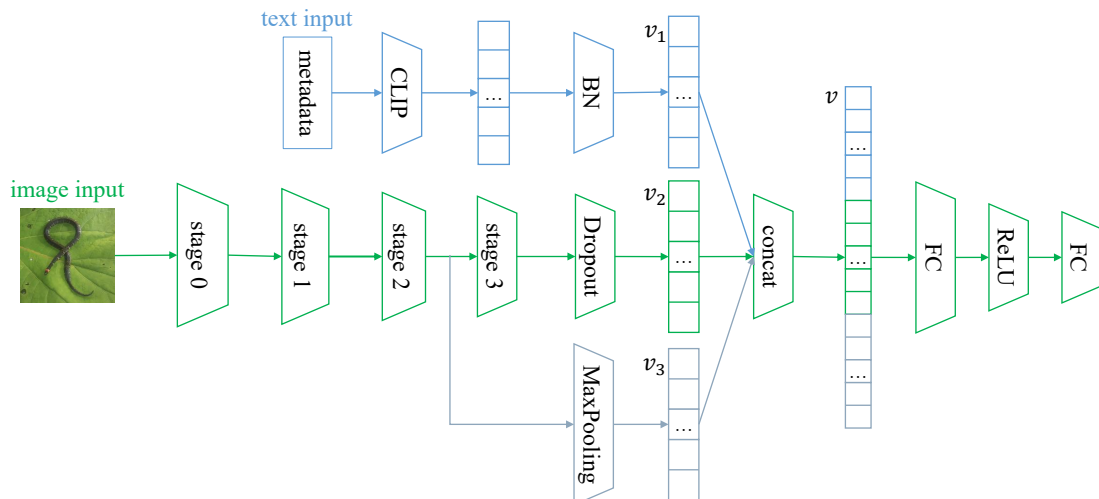
Data preprocessing plays a crucial role in machine learning, as it influences not only the final performance but also the feasibility of problem resolution. Upon obtaining the dataset provided by the competition organizers, several issues emerged. For instance, certain images listed in the metadata CSV file were found to be nonexistent within the corresponding image folders. To address this, we generated a new metadata CSV file by eliminating the affected rows from the original file. Additionally, a subset of images within the dataset was found to be corrupted, potentially due to network transmission or other factors. To mitigate this concern, we utilized OpenCV to read the problematic images and subsequently re-wrote them to the file system, thereby solving the corruption issue.

The SnakeCLEF dataset includes valuable metadata pertaining to the observation locations. Leveraging this location information is of great significance, as certain snake species inhabit geographically confined areas. However, the metadata presents the location in the form of country or region codes, which cannot be directly utilized as inputs for convolutional neural network (CNN) or Vision Transformer (ViT) [12]. To address this challenge, we employ CLIP [2] to extract location features without engaging in fine-tuning. Subsequently, Principal Component Analysis (PCA) [13] is employed to reduce the dimension of the resulting feature vectors.

Data augmentation serves as a key technique in computer vision tasks. Within our methodology, we leverage fundamental image augmentation methods from Albumentations [14]. These methods encompass RandomResizedCrop, Transpose, HorizontalFlip, VerticalFlip, ShiftScaleRotate, RandomBrightnessContrast, PiecewiseAffine, HueSaturationValue, OpticalDistortion, ElasticTransform, Cutout, and GridDistortion. Furthermore, we incorporate data mixing augmentation techniques, such as Mixup [15], CutMix [16], TokenMix [17], and RandomMix [18], during the course of the competition. These data augmentation methods provide strong regularization to models by softening both images and labels, avoiding the model overfitting in training dataset.

#### 3.2. Model

Throughout the competition, we explored various models, including both classical and state-of-the-art architectures, such as Convolutional Neural Networks and Vision Transformers.



**Figure 2:** Architecture of our model. Take ConvNeXt-v2 [4] as the backbone, which is made up of 4 stages, feature vector extracted from metadata ( $v_1$ ), original feature vector ( $v_2$ ) and feature vector from middle stage of the backbone ( $v_3$ ) are concatenated to get the final feature vector  $v$ , a MLP classifier is followed to get the final classification results.

Models employed during the competition include ResNet [19], ViT [20], ConvNeXt [21], BEiT-v2 [22], EVA [23] and ConvNeXt-v2 [4]. The implementation of these models was facilitated using the timm [24] library. In light of the imposed limitations on model parameters and the consideration of the model representation capabilities, we selected ConvNeXt-v2 [4] as the backbone architecture in our final method.

However, relying solely on the visual backbone is insufficient for effectively addressing the task at hand. Given the availability of metadata in the competition and the inherent challenges associated with fine-grained image classification, it becomes necessary to modify the architecture of the vision model to achieve superior performance. The architectural design of the model employed in our final submission is illustrated in Fig. 2.

Following the completion of the third stage of ConvNeXt-v2 [4], the intermediate-level feature map is combined with the high-level image features after the final stage, along with the metadata features. This concatenation process yields a comprehensive representation that captures both the image and metadata information. To mitigate overfitting, we have incorporated MaxPooling [25], BatchNorm [26], and Dropout [27] techniques into our methodology. Once the comprehensive representation is obtained, a classifier comprising two linear layers and ReLU [28] activation functions follows and generates classification results.

### 3.3. Optimization Procedure

Addressing long-tailed recognition is another challenge encountered in the competition. To tackle this issue, we extensively explored various techniques implemented in BagofTricks-LT [29]. In our final submission, we incorporated the seesaw loss [1] as a key component. The seesaw loss formulation can be expressed as follows:

$$L_{\text{seesaw}}(\mathbf{z}) = - \sum_{i=1}^C y_i \log(\hat{\sigma}_i), \quad (2)$$

$$\text{with } \hat{\sigma}_i = \frac{e^{z_i}}{\sum_{j \neq i}^C \mathcal{S}_{ij} e^{z_j} + e^{z_i}},$$

where  $\mathbf{z}$  denotes the output obtained from the fully connected layer,  $C$  represents the total number of classes, and  $y_i$  corresponds to the one-hot label of the image. The hyper-parameters  $\mathcal{S}_{ij}$  are carefully set based on the distribution characteristics inherent in the dataset.

Distinguishing between venomous and non-venomous snake species and the consequential assignment of varying costs to different classification errors are of great importance in this year’s challenge, as demonstrated by Eq. 1. In accordance with these requirements, loss function that effectively models the real-world costs associated with mislabeling [5] is utilized by us. To align with this objective, we incorporate the Real-World Weighted Cross-Entropy (RWWCE) loss function [5] during the final three epochs of training, employing a reduced learning rate.

In addition to the choice of loss functions, the selection of an optimizer and an appropriate learning rate decay strategy are important in the training of our models. For optimization, we adopt the AdamW optimizer [30]. To enhance convergence speed and overall performance, we implement cosine learning rate decay [31] coupled with warmup techniques during the training process. These strategies collectively facilitate more effective and efficient model convergence.

### 3.4. Post-processing

In this year’s challenge, the task requires the solution to accurately identify the venomous nature of snakes, particularly focusing on distinguishing the venomous species, with the limited model capacity. It is challenging but fortunately, the organizers provided a metadata repository, with a particular focus on geographical information. In practical contexts, where reliance solely on visual cues may prove insufficient performance on fine-grained classification, the supplementation of geographical details assumes a crucial role in assisting human experts in making judgment. Thus, the integration of geographical information within the metadata exhibits the potential to enhance the decision-making prowess of classification models.

Inspired by [32], assuming the above-mentioned trained model as  $f$ , we developed a simple prior model denoted as  $g$ . This prior model is simple but efficiently, composed of three fully connected layers with non-linear activation function and employed dropout regularization. In the training process of this light model, we adopt the AdamW [30] optimizer and performed balanced sampling on the training data, to mitigate the impact of the long-tail distribution in the dataset. The objective of this training process was to minimize the following loss function:

$$\mathcal{L}_{loc}(\mathbf{x}, \mathbf{r}, \mathbf{O}, y) = \lambda \log(s(g(\mathbf{x})\mathbf{O}_{:,y})) + \sum_{\substack{i=1 \\ i \neq y}}^C \log(1 - s(g(\mathbf{x})\mathbf{O}_{:,i})) + \sum_{i=1}^C \log(1 - s(g(\mathbf{r})\mathbf{O}_{:,i})), \quad (3)$$

where the metadata features extracted from CLIP is denoted as  $\mathbf{x}$ .  $\mathbf{O}$  is the category embedding matrix, where each column is the prototype of different category, pre-computed by our trained model  $f$ , e.g., ConvNeXt-v2 [4]. Furthermore,  $\mathbf{r}$  signifies a uniformly random location data point, and  $\lambda$  serves as a hyper-parameter for weighting positive observations. It is important to note that if a category  $y$  has been observed at the spatial location  $\mathbf{x}$  within the training set, the value of  $s(g(\mathbf{x})\mathbf{O}_{:,y})$  should approximate 1. Conversely, if the category has not been observed, the value should approximate 0.

During the inference stage, our prior model efficiently calculates the prior class embeddings denoted as  $\mathbf{P}$ . Utilizing the following equation:

$$\mathbf{S}' = \text{Softmax}(\mathbf{P}) \odot \mathbf{S}, \quad (4)$$

where  $\mathbf{S}$  is the prediction score computed by  $f$ . We derive the final class scores  $\mathbf{S}'$  by computing the joint probability of predictions from the two models  $f$  and  $g$ . In real-world scenarios, misclassifying a non-venomous snake as venomous carries significant consequences and is deemed unacceptable. To address this concern, we implement a robust post-processing approach. When the predicted confidence of an image  $\mathbf{x}$  is relatively low, we analyze its top-5 predictions. If any of these predictions include a venomous class, we classify the image as venomous. This post-processing technique represents a well-considered compromise between precision and recall. Notably, this approach enable us to get the 1st place in the private leaderboard. We firmly believe that this strategy possesses considerable advantages for practical applications.

## 4. Experiments

In this section, we will introduce our implementation details and main results.

### 4.1. Experiment Settings

The proposed methodology is developed utilizing the PyTorch framework [33]. All models employed in our approach have been pre-trained on the ImageNet dataset [34], readily available within the timm library [24]. Fine-tuning of these models was conducted across 4 Nvidia RTX3090 GPUs. The initial learning rate was set to  $2 \times 10^{-5}$ , and the total number of training epochs was set to 15, with the first epoch dedicated to warm-up, employing a learning rate of  $2 \times 10^{-7}$ . To optimize model training, we utilized the AdamW optimizer [30] in conjunction with a cosine learning rate scheduler [31], setting the weight decay to  $2 \times 10^{-5}$ . During inference on the test dataset, we adopted test time augmentation. Furthermore, considering that an observation may consist of multiple images, we adopted a simple averaging approach to obtain a singular prediction for each observation.

### 4.2. Main Results

In this section, we present our primary findings attained throughout the challenge, as illustrated in Tab. 1. The ‘‘Metric’’ column within the table corresponds to the public track1 metric featured on the leaderboard.

**Table 1**  
Main results of SnakeCLEF.

Backbone	Resolution	Metric (%)	Comments
ResNet50 [19]	224 × 224	72.22	baseline
BEiT-v2-L [22]	224 × 224	82.59	stronger backbone
BEiT-L [35]	384 × 384	88.74	cutmix
EVA-L [23]	336 × 336	86.82	cutmix
Swin-v2-L [36]	384 × 384	88.19	cutmix
VOLO [20]	448 × 448	88.50	cutmix
ConvNeXt-v2-L [4]	384 × 384	88.98	seesawloss + randommix
ConvNeXt-v2-L [4]	384 × 384	89.47	seesawloss + cutmix
ConvNeXt-v2-L [4]	512 × 512	90.86	seesawloss + cutmix + metadata
ConvNeXt-v2-L [4]	512 × 512	91.98	seesawloss + cutmix + metadata + middle-level feature
ConvNeXt-v2-L [4]	512 × 512	93.65	seesawloss + cutmix + metadata + middle-level feature + post-processing

As indicated by Tab. 1, the model parameters and image resolution hold crucial significance in image recognition tasks, aligning with conventional expectations. An increase in model parameters and image resolution corresponds to improvement in the public leaderboard score. Furthermore, data augmentation plays as a key factor in enhancing the generalization capacity of models. Notably, CutMix [16] outperforms alternative data mixing augmentation techniques, such as RandomMix [18], based on our experimental observations.

Metadata plays a pivotal role in the recognition of snake species, enabling models to acquire enhanced representations of observations and thereby achieve superior classification results. In our experiments, the utilization of metadata facilitated the acquisition of enriched contextual information, leading to improved model performance. Additionally, the incorporation of the Seesaw loss [1] demonstrated notable efficacy in mitigating the challenges posed by long-tailed distributions, surpassing the conventional CrossEntropy loss. Moreover, the integration of middle-level features proved effective in alleviating the complexities associated with fine-grained image recognition, enabling more precise discrimination between similar snake species.

Given that the final evaluation metric takes into account the demands of real-world applications and imposes greater penalties for misclassifying a venomous snake species as harmless compared to misclassifying a harmless species as venomous, we place significant emphasis on post-processing techniques. Specifically, when the model exhibits uncertainty in its predictions for a particular observation, we adopt a cautious approach and classify it as a venomous snake species based on the top-5 predictions. This post-processing strategy has proven highly advantageous, leading to substantial improvements in both the public leaderboard and the private test data performance, as evidenced by Tab. 1.



## 5. Conclusion

Fine-grained visual analysis holds great practical significance, particularly in accurately discerning the toxicity of snakes within the domain of snake sub-classification. This paper focuses on addressing the snake classification problem by harnessing the valuable metadata present in the dataset for posterior filtering. Additionally, a robust post-processing technique is employed to facilitate toxicity identification. These approaches have culminated in our noteworthy achievement of securing the first-place position in the challenge, attaining an impressive overall evaluation score of 91.31% on the private leaderboard.

## References

- [1] 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 Conference on Computer Vision and Pattern Recognition, 2021, pp. 9695–9704.
- [2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
- [3] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, S. Belongie, Fine-grained image analysis with deep learning: A survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2021) 8927–8948.
- [4] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, S. Xie, ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders, *arXiv preprint arXiv:2301.00808* (2023).
- [5] Y. Ho, S. Wookey, The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling, *IEEE Access* 8 (2019) 4806–4813.
- [6] L. Pícek, I. Bolon, A. M. Durso, R. R. de Castañeda, Overview of the snakeclef 2020: Automatic snake species identification challenge, *Working Notes of CLEF* (2020).
- [7] L. Pícek, A. M. Durso, I. Bolon, R. R. de Castañeda, Overview of snakeclef 2021: Automatic snake species identification with country-level focus, *Working Notes of CLEF* (2021).
- [8] L. Pícek, M. Hružík, A. M. Durso, I. Bolon, Overview of snakeclef 2022: Automated snake species identification on a global scale, *Working Notes of CLEF* (2022).
- [9] L. Bloch, A. Boketta, C. Keibel, E. Mense, A. Michailutschenko, O. Pelka, J. Rückert, L. Willemeit, C. M. Friedrich, Combination of image and location information for snake species identification using object detection and efficientnets, *Working Notes of CLEF* (2020).
- [10] R. Chamidullin, M. Šulc, J. Matas, L. Pícek, A deep learning method for visual recognition of snake species, *Working Notes of CLEF* (2021).
- [11] C. Zou, F. Xu, M. Wang, W. Li, Y. Cheng, Solutions for fine-grained and long-tailed snake species recognition in snakeclef 2022, *arXiv preprint arXiv:2207.01216* (2022).
- [12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Trans-

formers for image recognition at scale, Proceedings of the International Conference on Learning Representations (2021).

- [13] A. Maćkiewicz, W. Ratajczak, Principal components analysis (PCA), *Computers & Geosciences* 19 (1993) 303–342.
- [14] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, A. A. Kalinin, Albumentations: Fast and flexible image augmentations, *Information* 11 (2020) 125.
- [15] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, *arXiv preprint arXiv:1710.09412* (2017).
- [16] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6023–6032.
- [17] J. Liu, B. Liu, H. Zhou, H. Li, Y. Liu, Tokenmix: Rethinking image mixing for data augmentation in vision transformers, in: European Conference on Computer Vision, Springer, 2022, pp. 455–471.
- [18] X. Liu, F. Shen, J. Zhao, C. Nie, Randommix: A mixed sample data augmentation method with multiple mixed modes, *arXiv preprint arXiv:2205.08728* (2022).
- [19] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
- [20] L. Yuan, Q. Hou, Z. Jiang, J. Feng, S. Yan, Volo: Vision outlooker for visual recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
- [21] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.
- [22] Z. Peng, L. Dong, H. Bao, Q. Ye, F. Wei, Beit v2: Masked image modeling with vector-quantized visual tokenizers, *arXiv preprint arXiv:2208.06366* (2022).
- [23] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, Y. Cao, Eva: Exploring the limits of masked visual representation learning at scale, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 19358–19369.
- [24] R. Wightman, Pytorch image models, <https://github.com/rwightman/pytorch-image-models>, 2019.
- [25] Y.-L. Boureau, F. Bach, Y. LeCun, J. Ponce, Learning mid-level features for recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2559–2566.
- [26] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, PMLR, 2015, pp. 448–456.
- [27] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, *arXiv preprint arXiv:1207.0580* (2012).
- [28] V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in: International Conference on Machine Learning, PMLR, 2010, pp. 807–814.
- [29] Y. Zhang, X. Wei, B. Zhou, J. Wu, Bag of tricks for long-tailed visual recognition with deep convolutional neural networks, in: Proceedings of AAAI Conference on Artificial

- Intelligence, 2021, pp. 3447–3455.
- [30] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).
  - [31] I. Loshchilov, F. Hutter, Sgdr: Stochastic gradient descent with warm restarts, arXiv preprint arXiv:1608.03983 (2016).
  - [32] 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.
  - [33] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, 2019, pp. 8024–8035.
  - [34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
  - [35] H. Bao, L. Dong, S. Piao, F. Wei, Beit: Bert pre-training of image transformers, arXiv preprint arXiv:2106.08254 (2021).
  - [36] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al., Swin transformer v2: Scaling up capacity and resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 12009–12019.