# Joint Feature Learning of Image Data with Embedded Metadata to Leverage Snake Species Classification

Benjamin Bracke[1], Mohammadreza Bagherifar[1], Louise Bloch[1,2,3] and Christoph M. Friedrich[1,2]

[1]*Department of Computer Science, University of Applied Sciences and Arts Dortmund (FHDO), Emil-Figge-Straße 42, 44227 Dortmund, Germany*

[2]*Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Essen, Germany*

[3]*Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen, Essen, Germany*

Abstract

Automatic identification of snake species from non-standard photos is an important task to improve medical treatment of snakebites. To address this problem, the SnakeCLEF 2023 competition provides a large data set of photos and metadata information for 1,784 snake species. This paper describes the FHDO Biomedical Computer Science Group's (BCSG) participation in this competition. Through a series of experiments investigating the effects of pre-trained feature extractors, image sizes, metadata integrations, class balance learning and multiple instance pooling methods, a proposed model architecture for joint feature learning of image data and embedded metadata is presented to improve classification of snake species. With this proposal, the best model achieved a macro $F_1$-Score of 81.90 % and challenge-specific metrics of 90.09 % Track 1 and 1, 149 Track 2 on the challenge public test data set.

**Keywords**

Snake species identification, multimodal model architecture, joint feature learning, metadata embedding

## 1. Introduction

This paper presents the participation of the University of Applied Sciences and Arts Dortmund (FHDO) Biomedical Computer Science Group (BCSG) at the Conference of Labs of the Evaluation Forum (CLEF) 2023[1] SnakeCLEF [1] challenge[2] for snake species identification. The code to reproduce the participation is available online on the HuggingFace platform[3].

The SnakeCLEF 2023 Challenge is one of five data-driven challenges of the LifeCLEF 2023 [2, 3, 4] research platform focusing on automated species identification. Specifically, this year's challenge aims to provide data-driven analysis to improve snake species identification, with a focus on accurate identification of venomous and non-venomous snakes based on non-standardised

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

benjamin.bracke002@stud.fh-dortmund.de (B. Bracke); mohammadreza.bagherifar002@stud.fh-dortmund.de (M. Bagherifar); louise.bloch@fh-dortmund.de (L. Bloch); christoph.friedrich@fh-dortmund.de (C. M. Friedrich)

https://www.fh-dortmund.de/friedrich/ (C. M. Friedrich)

0000-0003-4986-7142 (B. Bracke); 0009-0004-1573-6210 (M. Bagherifar); 0000-0001-7540-4980 (L. Bloch); 0000-0001-7906-0038 (C. M. Friedrich)



[1]CLEF 2023: https://clef2023.clef-initiative.eu/, [Last accessed: 2023-06-01]

[2]SnakeCLEF 2023: https://www.imageclef.org/SnakeCLEF2023, [Last accessed: 2023-06-01]

[3]Code Release: https://huggingface.co/BBracke/SnakeCLEF2023 [Last accessed: 2023-06-01]

photographs. As the annual mortality from snakebites is between 81,000 and 138,000 people [5], identifying snake species could help to administer the correct antivenom [6] and thus reduce the number of victims.

Compared to the FHDO-BCSG team's participation in SnakeCLEF 2022 [7] and before[8, 9], the previously proposed workflow for snake species identification based on object detection is abandoned and a new approach is introduced, which focuses more on multimodality to combine the provided image data with the also provided metadata for snake species identification.

The paper is structured as follows: In Section 2 the related work in this research area is described. Section 3 summarises the SnakeCLEF 2023 data set and Section 4 describes the new proposed method for snake species identification. Section 5 shows ablations studies as well as the results obtained using the proposed method. Finally, the results are summarised and concluded in Section 6 and Section 7 gives an outlook on future work.

## 2. Related Work

Automatic snake species classification has a long history, starting with classical ML models. For example, an approach based on manually extracted taxonomic features was implemented in [10] to distinguish between six species. However, manual feature extraction is a tedious task, so field-based approaches have been developed that collect unstructured photographs and extract textural [11] or deep learning features from snake images.

Recently, most of the published studies [12, 13, 14, 15, 16, 17] focus on deep learning-based approaches. Some of these studies were designed as object detection tasks.

For example, different deep learning-based object detection methods were compared to each other in [12]. The data set which was extracted from ImageNet-1k [18] and augmented by a Google Image search included 1,027 images of eleven Australian species. The least frequent class contained 60 images. The best mean Average Precision (mAP) was achieved for a Faster Region-Based Convolutional Neural Network (Faster RCNN) [19] with a ResNet [20] backbone.

A similar approach [13] used Faster RCNN with different detection layers. The data set collected from three data sources contained 250 images of nine species occurring on the Galápagos Islands, Ecuador. To collect the data set, two internet searches were performed on the Google and Flickr platforms, and an image data set provided by the Ecuadorian Institution of Tropical Herping[4] were accessed. Similar to the previously described method, the ResNet backbone achieved the best accuracy of 75 %.

Other studies have performed classification tasks. For example, the performances of three deep learning networks, namely VGG16 [21], DenseNet161 [22], and MobileNetV2 [23] are compared in [14]. The data set contained 3,050 images of 28 species. An accuracy of 72 % was achieved for the test data set and the DenseNet161 architecture.

A deep Siamese network [24] for one-shot learning was developed in [25]. The network was trained on 200 images from the World Health Organization (WHO) venomous snake database[5]. This data set contained three to 16 images per class.

[4]Tropical Herping: https://www.tropicalherping.com/, [Last accessed: 2023-06-01].

[5]WHO Snakebite Data Information Portal: https://snbdatainfo.who.int/?page=Information#tab=tab_3, [Last accessed: 2023-06-01]

Although the previously described methods each examined less than 30 distinguishable species, more than 600 of the world's 3,700 snake species are medically relevant [26].

The SnakeCLEF challenge [27, 26] overcomes this disadvantage by providing a more diverse data set with images of more than 1,000 species. It addresses the problems of high intra-class variance and low inter-class variance as well as the long-tail distribution of snake images. Since snake species strongly vary across countries, the data set also includes location metadata. Several deep learning approaches have been successfully submitted in previous rounds of this challenge.

In SnakeCLEF 2020 [28], the winning approach [29] used a ResNet architecture pre-trained on ImageNet-21k [30] and achieved a macro-averaging $F_1$-Score of $62.54\ \%$. The FHDO-BCSG [8] combined object detection and image classification using a Mask-RCNN [31] instance detection framework and an EfficientNet-B4 [32] classification model. This method reached a macro-averaging $F_1$-Score of $40.35\ \%$. In post competition experiments, the score was optimized to $59.40\ \%$.

The winning approach [33] of SnakeCLEF 2021 combined object detection with an EfficientDet-D1 [34] model, and an EfficientNet-B0 classifier as well as likelihood weighting to fuse image and location information. The best model reached a macro-averaging $F_1$-Score of $90.30\ \%$.

Experiments with several Convolutional Neural Network (CNN) architectures were presented in [35]. The best $F_1$-Score of $83.00\ \%$ was obtained for an ensemble model combining two ResNeSt [36] models with a ResNet [20], and a ResNeXt [37] model. The ensemble was generated by a majority voting of the top 1 predictions of the individual models.

The FHDO-BCSG [9] expanded the SnakeCLEF 2020 workflow by combining object detection with EfficientNets and Vision Transformers (ViTs) [38]. The best model was an ensemble averaging the model predictions of an EfficientNet-B4 model and ViTs. This submission obtained a macro-averaging $F_1$-Score of $78.75\ \%$.

After the challenge, the organizers published an approach [17] that was trained on a subset of the challenge data set and evaluated on the official test set. In the work, ViTs were trained using a two-step approach. First, the model is trained with cross entropy loss on the training data set. Second, the resulting model was fine-tuned with focal loss [39] to improve performance for rare species. The model achieved a macro-averaging $F_1$-Score of $92.20\ \%$.

In the SnakeCLEF 2022 [27] challenge, most teams focused on the combination of image and metadata as well as strategies that solve the problem of long-tail distributions. The approach [40] that produced the best results was an ensemble of different model architectures, namely ConvNeXt [41], VOLO [42], CoLKANet [40], SwinTransformer [43], and ViT. The CoLKANet is a newly developed architecture that combines large kernel attention layers and self attention layers. The model performance was improved by different strategies to enhance the robustness, e.g. the use of TrivialAugment [44], Test-time Augmentation (TTA) [45, 46], pseudo labelling for rare classes, or Exponential Moving Average (EMA). The ensemble achieved an macro averaging $F_1$-Score of $85.40\ \%$ on the private test set.

The second place team [47] trained models based on the ViT architecture. An effective logit adjustment loss (ELAL) [47] which combines the logit adjustment loss [48] with the class-balanced loss [49] was developed to increase the relative margin between logits of rare and common labels. This loss improved the classification, especially for the rare classes. The

final model was an ensemble of two ViT-L models and one ViT-H model and reached a macro $F_1$-Score of $84.57$ % for the private test set.

The third place of the challenge and a macro $F_1$-Score of $82.65$ % for the private test set was reached by [50]. The approach combines supervised and unsupervised training on the training, validation, and test sets using the Simple Framework for Contrastive Learning (SimCLR) [51] with the MetaFormer [52] architecture that combines image data and metadata, TTA, and logit adjustments to reduce the impact of the long-tailed class distribution. The final model is an ensemble combining seven MetaFormer models trained for different epochs and with different hyperparameters.

During the participation in the previous SnakeCLEF challenge, the FHDO-BCSG [7] extended their previous workflow by using object detection with YOLOv5 [53], feature concatenation, and multiplication with prior probabilities. The final ensemble that combines seven models with different architectures (EfficienNet, EfficientNet-v2 [54], and ConvNeXt) reached a macro-$F_1$-Score of $70.80$ % on the private test set.

In this work, the previously developed workflow [7, 8, 9] has been completely revised. The new workflow which is presented in this paper, focuses more on training of multimodal models that combine image data with tabular metadata.

## 3. SnakeCLEF 2023 Data Set

The SnakeCLEF 2023 data set included 196,332 images of 111,215 observations and 1,784 species. The training data set contains 168,144 (85.64 %) images of 95,588 (85.95 %) observations and consists of two data sources. The first one originates from the iNaturalist platform[6] and includes 154,301 (91.77 % of training data set) images of 85,843 (89.81 % of training data set) observations and 1,784 (100.00 %) species. To add images of rare species, additionally, data from Herpmapper is added to the training data set. This data source includes 13,843 (8.23 % of training data set) images of 9,745 (10,19 % of training data set) observations and 889 (49,83 %) species. The validation data set includes 14,117 (7.19 %) images of 7,816 (7.03 %) observations and 1,599 (89.63 %) species. The remaining 14,071 (7.17 %) images of 7,811 (7.02 %) observations were used as a test set.

The distribution of images per snake species is highly imbalanced. The most frequent species was the *Natrix natrix* containing 2,079 images in the training and validation sets. For six species only three images were collected.

In addition to the photographs, metadata that provides information about the country (*code*), and if the species is endemic (*endemic*) is available. The data was collected in 214 countries with Mexico ("MX") being the most frequent country (21,002 images; 10.70 %). For 9,730 images ($4.96$ %) no information about the code was available. 29,198 (14.87 %) of the images show endemic snakes. An additional table is available that contains information if the species is venomous or not. 285 (15.97 %) species in the data set are venomous.

[6] iNaturalist: https://www.inaturalist.org/, [Last accessed: 2023-06-01].

# 4. Proposed Method

The participation of FHDO-BCSG team in SnakeCLEF 2022 [7] and before[8, 9] focused heavily on an object detection based snake species identification workflow where the image data was first cropped to a specific region of interest where the snake is pictured and subsequently classified. The proposed method for this year's participation abandons this workflow and focuses more on a multimodal model that combines the provided image data with the also provided metadata for snake species identification.

## 4.1. ConvNeXt

For feature extraction from image data, the proposed method relies on highly optimised and pre-trained CNNs. Specifically, it uses a ConvNeXt V2 [55] base model with 89M. parameters. The ConvNeXt architecture [41] is a state-of-the-art approach to modernising the most standard CNN architecture (ResNet50) towards the design choices of the popular Vision Transformers models. Therefore, the authors of ConvNeXt conducted several experiments to discover the key components that lead to the performance differences. A key component was changing the multi-stage macro design of the architecture to reduce the stage computation ratio and changing the stem to a simpler "patchify" layer similar to ViT [41]. Other changes included the use of inverted bottleneck blocks with depth-wise convolution, a larger kernel size, and an increased network width to the same number of channels as the Swin-Transformer [41]. ConvNeXt also adopted some features of the micro-scale architecture of transformers, such as replacing the Rectified Linear Units (ReLU) activation function with its smoother Gaussian Error Linear Unit (GELU) [56] variant, using fewer normalization layers, and replacing BatchNorm layers with simple Layer-Normalization [41]. Other performance differences resulted from similar training techniques as for ViT, e.g., the use of the AdamW [57] optimizer, extended training epochs, heavy data augmentation including CutMix [58], RandAugment [59], Random erasing [60], and label smoothing [41]. Further improvements to the ConvNeXt architecture have been made by [55] by adding Global Response Normalisation (GRN) layers to improve inter-channel feature competition, as well as using self-supervised learning techniques such as masked autoencoders. This co-design of architectural improvements and self-supervised learning techniques results in the so-called ConvNeXt V2 model family, which further improves the performance of pure ConvNets. There are different ConvNeXt V2 variants T/S/B/L, which differ only in the number of channels and the number of blocks in each stage.

## 4.2. Leveraging Metadata Information

As mentioned in Section 3, the data set contains additional metadata including region and endemic information for most of the images. The idea is to use the metadata information as additional features to leverage model performance in snake species identification. This requires a multimodal model architecture that can handle both image data and structural metadata.

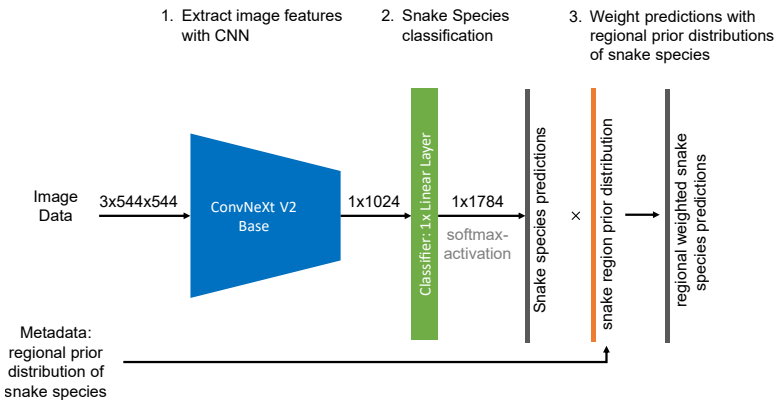



**Figure 1:** Schematic representation of the proposed method, which incorporates regional metadata by multiplying it with regional prior probabilities.

### 4.2.1. Multiplication with Regional Prior Probabilities

A successful approach in recent years of the snake challenge has been to multiply the snake species prediction probabilities of the image classification model by the regional prior probabilities of the snake species, as visualised in Figure 1. This requires estimating the prior probabilities from the relative frequency distribution of observations per snake species and region of the training data set. Two strategies can be used to combine the region and image information. First, the raw regional prior probabilities can be multiplied by the snake species prediction probabilities of the image classification model. The second strategy is to multiply by a binarised version of the prior probabilities so that it acts as a filter mask, filtering out those snake species predictions of the model that do not occur in that region with respect to a given region code in the training data set. For images with missing regional information, as well as for regions not available in the training data set, the prior probability of the ”unknown” class can be used. Unfortunately, this approach only applies to the regional metadata.

### 4.2.2. Joint Feature Learning with Embedded Metadata

A more advanced method of utilising metadata is to use them as additional features alongside the image data and let the model learn how to incorporate them into the identification of snake species, as visualised in Figure 2. This raises the problem of how to represent the discrete, nominal metadata in such a way that the model can utilise them as features, since neural network models usually assume numerical, continuous values as input.

**Embedding layers** , commonly used in natural language processing (NLP) tasks for word embedding, is the proposed solution to embed the metadata as a learnable numerical vector representation. Embedding layers (in particular, the nn.Embedding[7] provided by the PyTorch framework) generally serve as lookup tables to learn a mapping from arbitrary discrete input

[7] https://pytorch.org/docs/stable/generated/torch.nn.Embedding.html, [Last accessed: 2023-06-01]

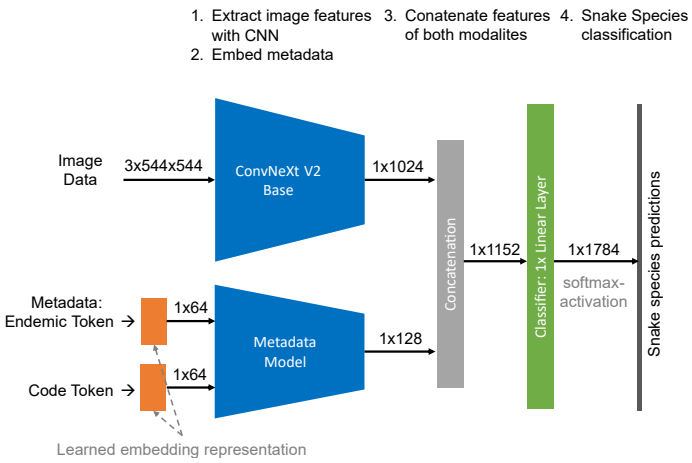



**Figure 2:** Schematic representation of the proposed method for joint feature learning of image data and embedded metadata.

tokens (e.g., words or country codes) to continuous embedding vector representations in a high-dimensional space. To do this, they internally map index values provided by a fixed vocabulary of input tokens to a weight matrix of learnable parameters. During training, the parameters of this weight matrix are updated using backpropagation to minimise the loss function and optimise model performance. With this approach, the representation of the embedding vector of a given input token is ideally optimised to best represent the meaning or context of the input token with respect to the specific training task of the model.

In order to embed the provided metadata of the SnakeCLEF 2023 data set, a fixed vocabulary of input tokens must be defined that maps the unique and alphabetically sorted country codes to integer index values, i.e. {"DE": 0, "US": 1 ..."unknown": 212}. Similarly, a vocabulary of input tokens for the endemic metadata {false: 0, true: 1} must be defined. During training, these predefined tokens are used as input to the embedding layers, which learn a numerical vector representation (embedding) of dimensions 64 for each code and endemic token (manually and heuristically defined so that the combined metadata dimensions are about 1/10 of the image feature dimensions).

**Metadata Model** is used to concatenate the individual embedding representations of code and endemic metadata and to learn a joint representation of both metadata resources. It is a small neural network consisting of two linear layers, both with 128 dimensions, with GELU activation [56] and layer normalisation [61], connected by a drop-out layer.

**Intermediate Fusion** approach [62], is used to concatenate the features of the metadata model with the features of the image feature extractor model. Subsequently, the joint feature representation is passed to a single linear layer of dimension 1784 with softmax activation, which serves as the classifier to predict the snake species. To prevent overfitting, the classifier model is preceded by a drop-out layer.

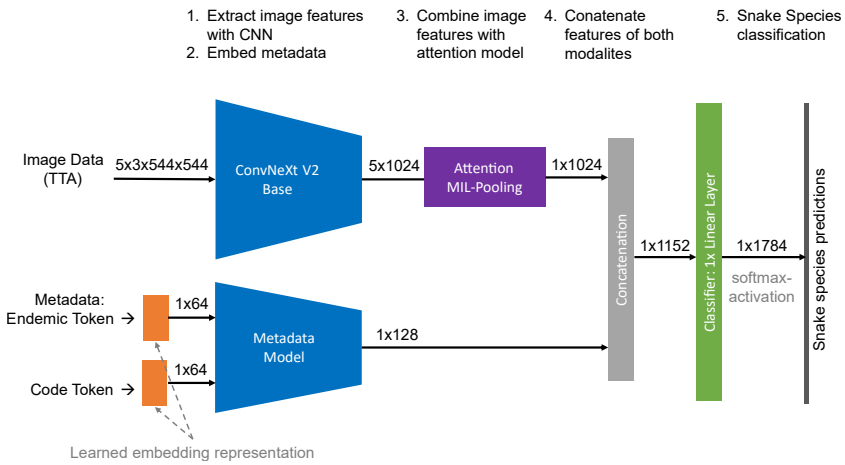



**Figure 3:** Schematic representation of the proposed method with joint feature learning of embedded metadata as well as attention-based bag-level MIL pooled image data of all TTA-, observation instances.

## 4.3. Multi-Instance Learning (MIL) Methods

As described in Section 3, some of the snake observations in the data set may contain more than one image, which makes it necessary to combine the model predictions into one common prediction per observation. This problem is called Multi-Instance Learning (MIL) and there are two different ways to aggregate the results for the instances of the model (MIL pooling): the instance-level approach and the bag-level approach [63]. The main difference between the instance-level approach and the bag-level approach is the operational space of MIL-pooling.

### 4.3.1. Instance-Level MIL Pooling

In the instance-level approach [63], the model architecture requires an instance-level classifier, i.e. the prediction probabilities of the snake species are provided for each instance. The subsequent MIL pooling combines all prediction probabilities of the snake species for each instance into a global prediction.

The proposed method implements mean MIL pooling, which simply averages the prediction probabilities of the classifier output for the snake species across all instances. Another method is weighted average MIL pooling, where the contribution of each instance is weighted differently when averaging the prediction probabilities for the snake species. The normalised values of the highest prediction probability of the model for a snake type are used as the weighting of the instances.

### 4.3.2. Bag-Level MIL Pooling

The bag-level approach [63] works with the feature space (embeddings) and MIL pooling is used to aggregate the embeddings of all instances to obtain a global embedding representation (bag embedding). This bag-level embedding is subsequently classified by a bag-level classifier, and a joint prediction of the snake species is obtained including all instances.

The proposed method implements the bag-level approach with an attention model for MIL pooling, as visualised in Figure 3. Attention MIL pooling [63] is more flexible than, e.g., mean MIL pooling and can adapt the pooling to the task and the provided data and provides better interpretability. Essentially, attention MIL pooling applies a weighted average MIL pooling to the feature space of all instances, with the weights being determined by a small learnable neural network.

### 4.4. Test Time Augmentation

The proposed method uses Test Time Augmentation (TTA) [45, 46] to make model predictions more robust, where multiple augmented versions of an image are presented to the model during inference. The TTA augmentation pipeline consists of resizing the image to 1.25 times the input size of the image feature extraction model and then using the FiveCrops augmentation method to obtain five different cropped representations of the same image, i.e. crops from the four corners and a central crop. As with multi-instance learning, the model predictions for each crop must then be aggregated into a common prediction.

### 4.5. Class Imbalance Learning Methods

As mentioned in Section 3, the distribution of images per snake species in the data set is highly imbalanced. Several approaches have been developed in the past to improve deep learning on imbalanced data sets, i.e using specialised loss functions with class imbalance weighting techniques that are used in the proposed method.

### 4.6. Focal Loss

Focal Loss [39] is such a specialised loss function that applies a dynamic scaling term $(1 - p_t)^\gamma$ to the standard cross-entropy loss function. The scaling factor decays to zero as the confidence of the correctly predicted class $p_t > 0.5$ increases, which automatically down-weights the contribution of easily classifiable examples during training. Vice-versa, the scaling factor increases as the confidence of the correctly predicted class $p_t < 0.5$ decreases, shifting the focus more towards the hard to classify samples during training. With the hyperparameter $\gamma > 0$ the strength of the scaling term can be set exponentially.

In addition, the focal loss can also be combined with a class weighting factor that differently weights the contribution of classes to the loss, allowing to handle the class imbalance problem. Different loss weighting techniques are tried to handle class imbalance, such as inverse class frequency and the so-called effective number of samples technique [49]. The effective number of samples is defined as the volume of samples and can be calculated by a simple formula $(1 - \beta^n)/(1 - \beta)$ (its inverse defines the loss weighting term), with $n$ the number of samples per snake species and $\beta \in [0, 1)$ being a hyperparameter.

### 4.7. ArcFace Loss

ArcFace loss or Additive Angular Margin Loss [64] focuses on learning more discriminative features to enforce higher similarity for intra-class samples and greater diversity for inter-class

**Table 1**
Base configuration of hyperparameters for all experiments.

| Hyperparameter | Base Config |
|---|---|
| batch size | 128 |
| optimizer | AdamW<br>$\beta_1 = 0.9, \beta_2 = 0.999$ |
| CNN feature extractor | |
| lr | 1e-5 |
| layer-wise lr decay | 0.85 |
| stochastic depth | 0.2 |
| weight decay | 1e-8 |
| classifier/ embedding layer/ attention module | |
| lr | 1e-4 |
| weight decay | 0.05 |
| warmup epochs[8] | 5 |

samples, i.e. to generate margin as Support Vector Machines do. It addresses the problem of softmax-based loss functions, such as focal loss, which do not explicitly optimise feature embeddings, resulting in a performance gap for large intra-class variation. ArcFace loss achieves this by learning a projection of features that distributes them on a circular shape, the hypersphere, so that the prediction depends only on the angle between them. Adding a margin to the angle as a penalty increases the distance between classes with similar feature embeddings. The authors introduced the ArcFace loss for deep face recognition and found that it could maximise the decision boundaries between similar looking faces, making it easier to distinguish them.

## 5. Experiments and Results

In order to test the influence of the different proposed methods and come up with a final solution, several ablation studies were conducted, which are described below.

### 5.1. General Experiment Conditions

All experiments are implemented in Python 3.8.13 using several Python packages, but mainly the Python library timm 0.9.2 [65] for pre-trained image classification models and PyTorch 1.13 [66] for training and inference. All experiments use similar base hyperparameters as summarised in Table 1, while experiment-specific hyperparameters will be mentioned in the following subsections. In order to achieve high batch sizes, mixed precision [67] and gradient accumulation are used.

The full-sized version of the SnakeCLEF 2023 data set (training + additional) is used for training. Since some images are missing from the full-sized version of the training data set, a

[8]Training epochs with parameter-frozen feature extractors.

total of 332 images of the smaller "medium" version is added to the training data set. However, a total of 85 images are still missing.

All models are trained with the same image augmentation pipeline, which consists of first resizing the image to $1.25$ times (at smallest image size, keeping the aspect ratio the same) the target model input sizes, then randomly flipping the image vertically and horizontally with probability $0.5$. Subsequently, a random squared crop of the target model input sizes is performed, followed by RandAugment [59] with $n = 2$ and $m = 7$, and normalising the images with ImageNet means $(0.485, 0.456, 0.406)$ and standard deviations $(0.229, 0.224, 0.225)$.

## 5.2. General Validation Conditions

Validation is performed on the full-sized version of the SnakeCLEF 2023 validation data set using an EMA model (exponential moving average of model parameters) that is updated at each training step with a decay rate of $0.9998$.

Also for validation, TTA is used, i.e. each image is first scaled to $1.25$ times (at smallest image size, keeping the aspect ratio the same) the target model input sizes, followed by five square crops of the target model input sizes (4 corner crops, 1 centre crop). Unless otherwise stated, the class predictions of the model for each TTA instance are pooled by mean MIL pooling to obtain a combined snake species prediction of the model. The same applies to combining class predictions of the model for multiple images of the same snake observation. This also means that all metrics for the validation results are gathered on the mean MIL pooled data.

Multiple metrics are taken into account for evaluating the experiment results. First the macro-$F_1$-Score as well as two custom metrics provided by the challenge organizers, namely "Challenge Track 2" Equation 2 and "Challenge Track 1" Equation 3.

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

$$\mathbf{L} = \sum_i L\left(y_i, \hat{y}_i\right) \tag{2}$$

$$M = \frac{\left(w_1 F_1 + w_2\left(100 - P_1\right) + w_3\left(100 - P_2\right) + w_4\left(100 - P_3\right) + w_5\left(100 - P_4\right)\right)}{\sum_i^5 w_i} \tag{3}$$

The motivation of "Challenge Track 2" is to penalise misclassifications of venomous species with harmless ones, but not vice versa, as based on this prediction a possible anti-venom might not be injected to the victim. Therefore, different weights for misclassifications are defined in Equation 1, where $p(s) = 1$ if species $s$ is venomous, otherwise $p(s) = 0$, as well as $y$ for ground truth species and $\hat{y}$ for predicted species. The metric "Challenge Track 1" is a weighted average of the overall macro $F_1$-Score and the weighted accuracies of different types of snake species confusion, where $w_1 = 1$, $w_2 = 1$, $w_3 = 2$, $w_4 = 5$, $w_5 = 2$ are the

**Table 2**
Validation results of the first experiment comparing different pre-trained image feature extraction CNNs (ImageNet21k vs. iNaturalist21), which were subsequently fine-tuned with the challenge data set.

| Pre-Train Model | Validation Metrics | | |
|---|---|---|---|
| | **Macro $F_1$-Score** | **Challenge Track 1** | **Challenge Track 2** |
| ImageNet21k | 59.00 % | 87.87 % | 2786 |
| iNaturalist21 | 61.39 % | 88.46 % | 2612 |

weights of individual misclassifications as in Equation 1. $F_1$ is the macro $F_1$-Score, $P_1$ is the percentage of harmless species misclassified as another harmless species, $P_2$ is the percentage of harmless species misclassified as another venomous species, $P_3$ is the percentage of venomous species misclassified as another harmless species, and $P_4$ is the percentage of venomous species misclassified as another venomous species.

## 5.3. Experiment: iNaturalist21 Pre-Training

First, the influence of different pre-trained CNN feature extractor models fine-tuned with the SnakeCLEF 2023 data set is observed. For this purpose, a ConvNeXt V2 base model pre-trained with ImageNet21k ("*convnextv2_base.fcmae_ft_in22k_in1k_384*") from the Python library timm 0.9.2 [65] is compared with a pre-trained iNaturalist21 [68] ConvNeXt V2 base model (provided by ourselves). The pre-training was performed on the iNaturalist21 data set for 10 epochs with an image size of $384 \times 384$ px and normal cross entropy as loss function. Fine-tuning for both models was then performed for 30 epochs with an image size of $384 \times 384$ px and normal cross entropy as the loss function. In order to ensure fairness to other challenge participants, the model weights were published during the course of the challenge[9].

The obtained validation results for this experiment are summarised in Table 2. The validation results show that pre-training on the iNaturalist21 data set leads to an improved downstream snake species classification of about $2.4\ \%$ points macro $F_1$-Score when fine-tuned with the challenge data set.

## 5.4. Experiment: Influence of Image Size

The second experiment focuses on the influence of different image sizes on the snake species classification result. Specifically, the pre-trained iNaturalist21 model fine-tuned in the previous experiment was further tuned for 10 more epochs using image sizes of $464 \times 464$ px, $544 \times 544$ px, and $624 \times 624$ px.

The validation results obtained for this experiment are summarised in Table 3. The validation results show that increased image sizes generally improve the snake species classification result. The biggest gain is seen between image sizes of 384 px and 464 px, of about $1.8\ \%$ points macro $F_1$-Score. However, this comes at the cost of increased training time per epoch.

[9] https://huggingface.co/BBracke/convnextv2_base.inat21_384, [Last accessed: 2023-06-01]

**Table 3**
Validation results of the second experiment comparing different fine-tuning image sizes.

| Model Input Image Size | Training Time[10] per Epoch | Validation Metrics | | |
|---|---|---|---|---|
| | | **Macro $F_1$-Score** | **Challenge Track 1** | **Challenge Track 2** |
| *model of experiment 1 fine-tuned + 10 epochs* | | | | |
| $384 \times 384$ px | 3700 sec. | 61.39 % | 88.46 % | 2612 |
| $464 \times 464$ px | 5300 sec. | 63.18 % | 89.47 % | 2380 |
| $544 \times 544$ px | 7400 sec. | 63.51 % | 89.70 % | 2330 |
| $624 \times 624$ px | 9700 sec. | 64.20 % | 89.75 % | 2307 |
| *new model trained for 40 epochs* | | | | |
| $544 \times 544$ px | 7400 sec. | 65.95 % | 90.17 % | 2197 |

Based on these results, it was investigated whether it would be useful to fine-tune a new model from "the beginning", using the good performing image size of 544 px (taking into account the training time and the improvement in classification results). Therefore, the pre-trained iNaturalist21 model was fine-tuned for 40 epochs (for reasons of comparability, as the effective fine-tuning epochs of the previous results were 30 epochs of the first experiment + 10 epochs of the second experiment) with the selected image size of 544 px and normal cross entropy as the loss function.

The validation results obtained (Table 3) show that this approach could further improve the snake species classification result by about $2.4$ % points macro $F_1$-Score.

## 5.5. Experiment: Leveraging Classification Results with Metadata

The third experiment focuses on the use of the provided metadata, such as endemic and regional code information, in addition to the image data to improve snake species classification. Specifically, the influences of the approaches described in the previous Section 4.2 are compared, such as multiplying model class prediction distributions with regional prior probabilities as well as joint feature learning using embedded endemic and regional code metadata.

As the first approach does not require any re-training, the obtained model from the second experiment is used, but its class prediction outputs are weighted with the previously defined regional prior probabilities of snake species. This approach leverages the snake species classification by about $9.3$ % points macro $F_1$-Score compared to the same model of the second experiment (Table 4) without regional prior probability weighting.

As the model architecture of the second approach differs with embedding layers, metadata model and classifier of the joint feature modalities, a new model needs to be trained for 40 epochs using the selected image size of 544 px from second experiment and normal cross entropy as loss function. This approach also leverages the snake species classification by about $10$ % points macro $F_1$-Score (Table 4) and even marginally outperforms the first mentioned approach of metadata incorporation. Thus, this approach is continued in the following experiments.

[10] for batch size 128 on NVIDIA RTX 6000 Ada GPU

**Table 4**
Validation results of the third experiment comparing different proposed approaches to incorporate metadata into the model architecture.

| Metadata Approach | Validation Metrics | | |
|---|---|---|---|
| | **Macro $F_1$-Score** | **Challenge Track 1** | **Challenge Track 2** |
| No Metadata incorp. | 65.95 % | 90.17 % | 2197 |
| Multiply with regional prior dist. | 75.27 % | 93.25 % | 1511 |
| Joint Feature Learning with Embedded Metadata | 76.04 % | 93.54 % | 1427 |

## 5.6. Experiment: Class Imbalance Learning

The fourth experiment focuses on the strong class imbalance present in the challenge data set and investigates whether the use of specialised loss functions with class balance weighting terms, mentioned in Section 4.5, can improve snake species classification. Specifically, the influences of using focal loss with weak focal value $\gamma = 0.5$ and strong focal value $\gamma = 2.0$ in combination with different class balance weighting terms obtained from the inverse class frequency or "effective number of samples" formula are investigated. A parameter checkpoint from epoch 15 of the joint feature learning model from the previous experiment (trained with normal cross-entropy) is used as weight initialisation. The model is then fine-tuned further for 25 epochs under the influence of the mentioned focal loss function and class balance weighting terms.

The validation results obtained (Table 5) show that using focal loss with different class balance weighting terms has only a marginal effect on snake species classification results. In general, a weak focal value $\gamma = 0.5$ slightly outperforms a strong focal value $\gamma = 2.0$, and using the "effective number of samples" formula as class balance term slightly outperforms the inverse class frequency class balance term. However, the best combination of focal loss with $\gamma = 0.5$ and "effective number of samples" as the class balance term only improves the snake species classification results by about $1.1$ % points macro $F_1$-Score compared to normal cross entropy.

In another experiment, the best combination of focal loss and class weighting term is then combined with the ArcFace loss, which should force the model to learn a better feature representation of the intermediate feature embedding before the classifier. As before, the new model is fine-tuned for 25 epochs under the combined influence of the aforementioned loss functions, using the parameter checkpoint of epoch 15 of the joint feature learning model from the previous experiment as weight initialisation.

Validation results (Table 5) show that this approach further improves the snake species classification result by a substantial amount of about $3.1$ % points macro $F_1$-Score compared to normal cross entropy, making it the best performing model of all experiments, which will be continued in the following experiments.

**Table 5**
Validation results of the fourth experiment comparing different loss functions and class balance terms to handle the strong class imbalance of the challenge data set.

| Loss Function | Class Weighting | Validation Metrics | | |
|---|---|---|---|---|
| | | **Macro $F_1$-Score** | **Challenge Track 1** | **Challenge Track 2** |
| CE | - | 76.04 % | 93.54 % | 1427 |
| Focal Loss $\gamma = 0.5$ | inverse class freq. | 76.58 % | 93.27 % | 1482 |
| Focal Loss $\gamma = 2.0$ | inverse class freq. | 75.93 % | 93.16 % | 1518 |
| Focal Loss $\gamma = 0.5$ | effectiv num. samples | 77.15 % | 93.58 % | 1425 |
| Focal Loss $\gamma = 2.0$ | effectiv num. samples | 76.64 % | 93.34 % | 1466 |
| ArcFace Loss + Focal Loss $\gamma = 0.5$ | effectiv num. samples | 79.10 % | 93.85 % | 1373 |

## 5.7. Experiment: Influence of MIL-Pooling Operators

The fifth experiment focuses on the influence of different MIL pooling methods, either for pooling the model predictions for different TTA instances of the same image, or for multiple image instances of the same snake observation in the data set. As mentioned earlier in this section, in the previous experiments simple mean MIL pooling was used for aggregating TTA instances and image instances of snake observations. This experiment investigates the influence of using different MIL-Pooling methods described in Section 4.3 on the snake species classification results. For the investigated classical MIL pooling methods, no re-training of the model is required, so the best model from the previous experiment is used.

The validation results (Table 6) show that applying different combinations of classical MIL pooling operators has only a marginal impact on the snake species classification results, as indicated by the macro $F_1$-Score. However, these results show that the combination of using the weighted average MIL pooling operator for pooling TTA instances and using a simple mean MIL pooling operator for different snake instances of the same observation gives the best results when it comes to more costly errors, as indicated by the "Challenge Track 2" metric.

Furthermore, two different approaches of attention-based MIL pooling are investigated. For both approaches, the best performing model from the previous experiment is further fine-tuned for 10 epochs with an attention module integrated into the model architecture as described in Section 4.3.2. In the first approach, attention is used to pool TTA instances in combination with a classical mean MIL pooling for instances containing the same snake observations. The difference with the second approach is that attention is used to pool both TTA instances and instances of the same snake observations in one step. Due to the VRAM limitations of the GPU, this requires limiting the total number instances, i.e. a maximum of 100 instances are used.

The validation results (Table 6) show that both attention-based MIL pooling approaches can

**Table 6**
Validation results of the fifth experiment comparing different MIL pooling techniques to pool TTA instances as well as instances of the same snake observation. "mean" refers to mean and "w. avg." refers to the weighted average MIL pooling on the instance-level as described in Section 4.3.1. "attention" refers to bag-level MIL pooling described in Section 4.3.2.

| **TTA Instance MIL-Pooling** | **Snake Instance MIL-Pooling** | Validation Metrics | | |
|---|---|---|---|---|
| | | **Macro $F_1$-Score** | **Challenge Track 1** | **Challenge Track 2** |
| *classic MIL pooling methods* | | | | |
| mean | mean | 79.10 % | 93.85 % | 1373 |
| mean | w. avg. | 79.17 % | 93.85 % | 1374 |
| w. avg. | mean | 79.53 % | 94.07 % | 1325 |
| w. avg. | w. avg. | 79.57 % | 94.05 % | 1330 |
| *attention-based MIL pooling methods* | | | | |
| attention | mean | 80.01 % | 94.12 % | 1305 |
| attention | | 79.78 % | 94.16 % | 1297 |

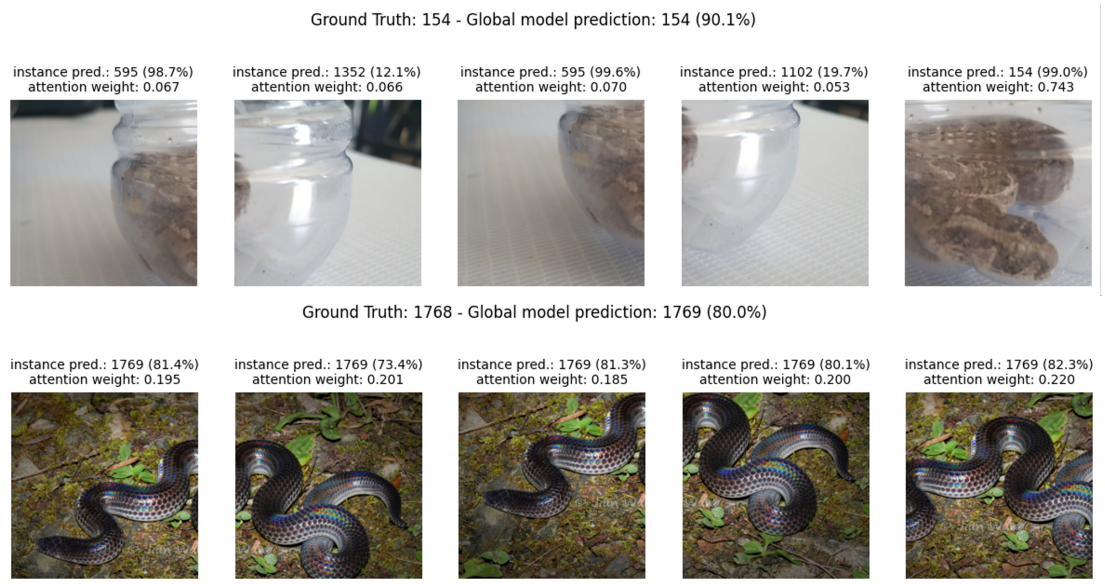



**Figure 4:** Prediction examples of a model with attention-based pooling of TTA instances. "instance pred." refers to the model prediction for each individual image instance. "Global model prediction" refers to the overall model prediction obtained after pooling of TTA instances. "attention weight" refers to the weight that the attention module assigns to the specific instances for TTA instance pooling.
For the example on the top row, attention pooling helped to guide the classifier to more relevant instances where the actual snake is well represented, making the global prediction correct and prevents misclassification. For the example on the bottom row, attention pooling is similar to mean MIL pooling where all TTA instances represent the snake well.

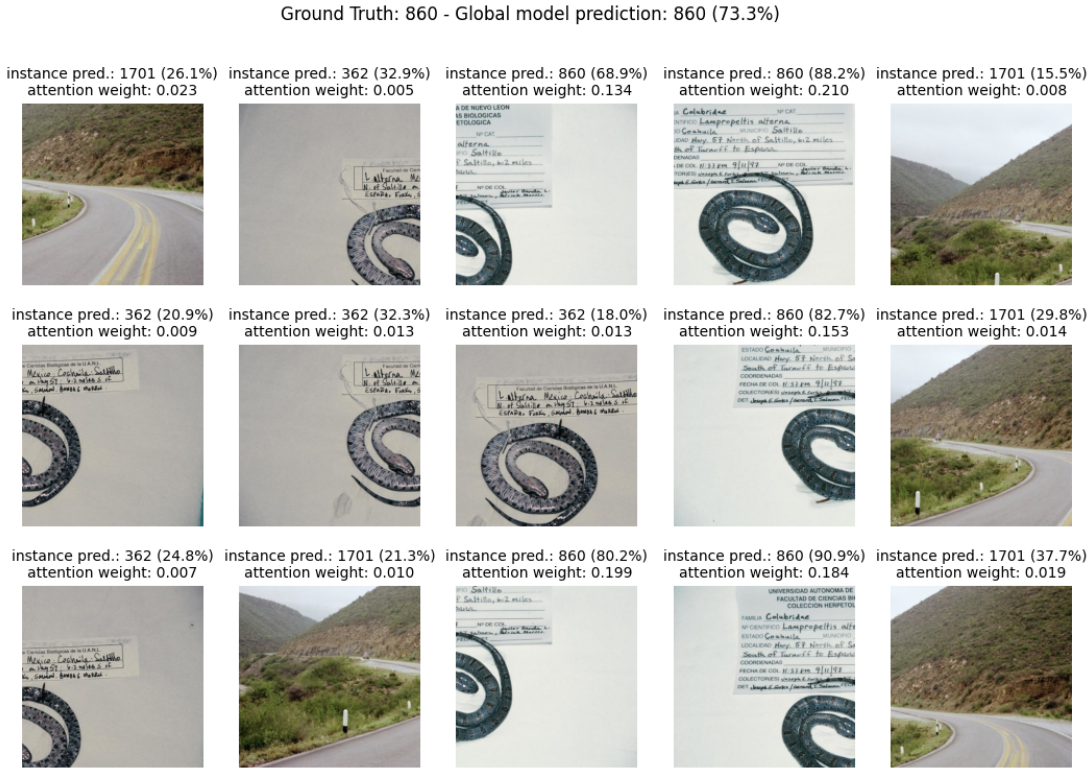



**Figure 5:** Prediction examples of a model with attention-based pooling of TTA instances + multiple instances of the same snake observation. "Instance prediction" refers to the model prediction for each individual image instance. "Global model prediction" refers to the overall model prediction obtained after pooling of TTA instances and instances of the same snake observation. "Attention weight" refers to the weight that the integrated attention module of the model assigns to the specific instances for instance pooling. This value allows to interpret the influence of the individual instances on the overall global prediction of the model.
In this example, the attention module helped to guide the classifier to the more relevant image instances where the actual snake is well represented and reduces the influence of the "junk" instances such as the road images.

reduce the more costly errors, as indicated by the "Challenge Track 2" metric, making these two approach the best performing models so far. Figure 4 shows two examples of the influence of attention-based pooling of TTA instances of the model. For the example in the top row, attention pooling helped to guide the classifier to the more relevant instances where the actual snake is well represented, making the combined prediction from all TTA instances correct and thus preventing misclassification. For the example in the bottom row, attention-based pooling is similar to mean MIL pooling where all TTA instances represent the snake well. Figure 5 shows the influence of attention-based pooling of TTA instances + instances of the same snake observation. In this example, the attention module also helped to guide the classifier to the more relevant image instances where the actual snake is well represented and reduces the influence of the "junk" instances such as the road images.

**Table 7**
Results of the final selected models on the challenge public test data set compared to the submitted models that were further fine-tuned using the training + validation data sets. TTA, MIL pooling methods of the models are given in brackets. "mean" refers to mean MIL pooling and "w. avg." refers to the weighted average MIL pooling of the class probability distribution output by the model as described in Section 4.3.1. "attention" refers to the feature embedding level MIL pooling described in Section 4.3.2.

| **Final Model (TTA + MIL pooling)** | **Fine-Tuned w/ Validation Data** | Test Metrics[11] | | |
|---|---|---|---|---|
| | | **Macro $F_1$-Score** | **Challenge Track 1** | **Challenge Track 2** |
| Model 1. (w.avg + mean) | no | 81.49 % | 94.92 % | 1187 |
| Model 2. (attention + mean) | no | 81.37 % | 94.90 % | 1194 |
| Model 3. (attention) | no | 80.55 % | 94.80 % | 1185 |
| Model 1. (w.avg + mean) | yes | 81.90 % | 95.09 % | 1149 |
| Model 2. (attention + mean) | yes | 81.39 % | 94.95 % | 1187 |
| Model 3. (attention) | yes | 80.97 % | 94.96 % | 1172 |

## 5.8. Final Models + Results on Testdata

As a result of these experiments, a total of three final models were submitted. All models share the same joint feature extractor with embedded metadata trained using the best performing class imbalance learning method of the fourth experiment (Focal loss $\gamma = 0.5$ with "effective num. of samples" class weighting term + ArcFace loss). For submission, this model was further fine-tuned for 20 epochs using the combined training and validation data sets of the challenge. The final models differ only in the way the TTA instances and instances of the same snake observations are pooled, as mentioned in the previous experiment. For the submission models that contain attention-based pooling, the attention modules as well as the classifiers of the models were also retrained for 10 epochs using the combined training and validation data sets of the challenge.

The following Table 7 shows the results of the final models on the public challenge test data set compared to the submitted models further fine-tuned on the challenge training + validation data sets. In general, the test results show that further fine-tuning with the training + validation data sets could only improve the snake species classification results very marginally.

## 5.9. Error Analysis in Detail

This section presents a comprehensive error analysis of a model based on the method proposed in section 4.2.2. It should be noted that the model used for this analysis is none of the submitted models, but is very similar to the Model 1. mentioned previously (w. avg. TTA MIL pooling

[11] on public challenge test data set

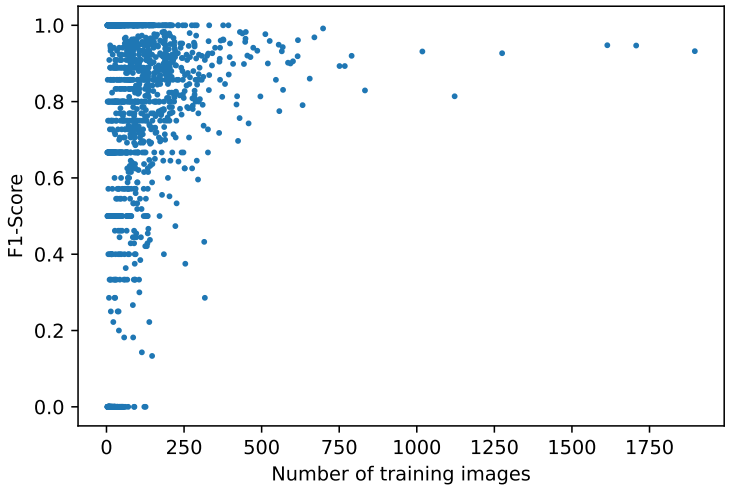
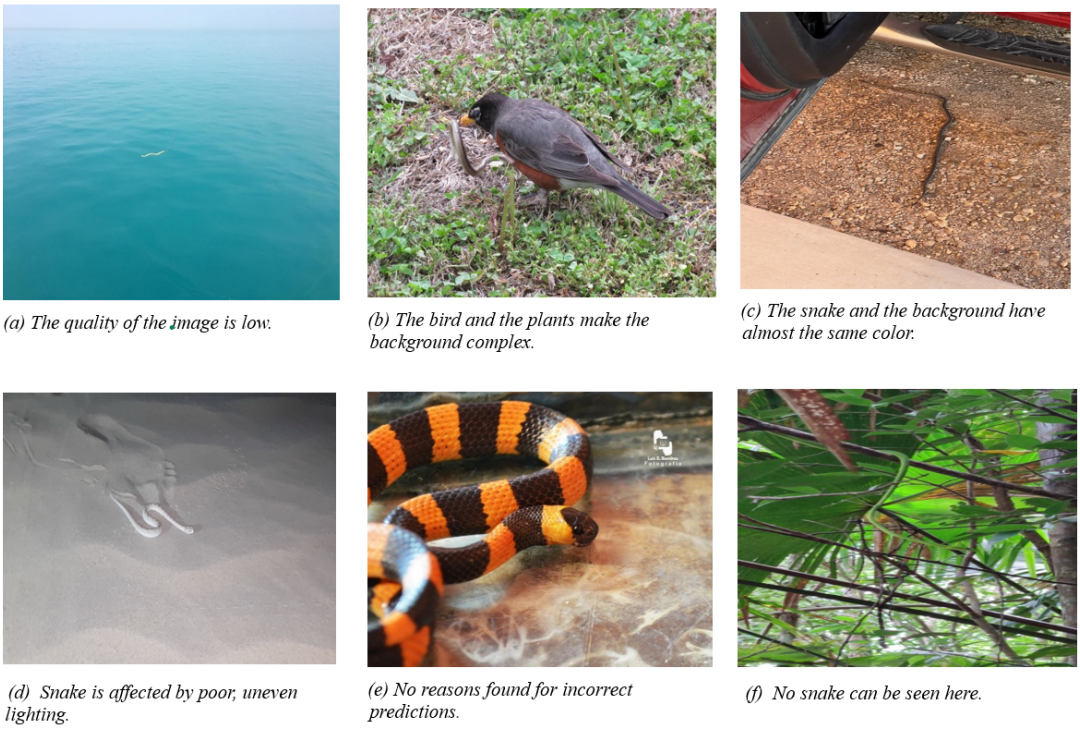



**Figure 6:** Distribution of $F_1$-Score per snake species class in relation to the number of training images per snake species.

+ mean MIL pooling of image instances with the same observation ID). However, due to the similarity of the models, this analysis should still be useful to evaluate and identify shortcomings and limitations of the proposed method.

In general the used model achieved an accuracy score of approximately 83 % by correctly predicting 11,807 images from the validation data set. However, it misclassified 2,305 images, indicating a considerable error rate. Furthermore, the analysis focused on rare snake species classes in the validation data set, defined as classes with fewer than 10 samples, totaling 1,115 classes. Figure 6 represents the relationship between the number of training images and the corresponding $F_1$-Score per species class and it indicates that a considerable number of classes have a relatively small number of training images. The figure suggests that increasing the number of training images for each species has a positive impact on the $F_1$-Score. When the number of training images is higher, the model's performance, as measured by the $F_1$-Score, tends to improve. The fact that most of the plot is in the upper part of the graph and falls within the range of 0.6 to 1 indicates that the model achieves relatively high accuracy and precision in identifying and classifying the species under consideration. Species exhibiting an $F_1$-Score below 0.5 are categorized as rare classes. However, there are certain extremely rare classes that have achieved a remarkable $F_1$-Score of 1.0. This pattern suggests that sufficient training data is crucial for achieving accurate and reliable results. The results revealed that approximately 21 % of the errors were attributed to these rare classes. However, the rare classes also exhibited good performance, accounting for 42 % of the overall correct predictions. This finding suggests that although rare classes may contribute to errors, their presence does not necessarily indicate poor model performance.

To gain deeper insights into the incorrect predictions, a manual examination was conducted on 500 randomly chosen misclassified images. One common reason was attributed to low



*(a) The quality of the image is low.*

*(b) The bird and the plants make the background complex.*

*(c) The snake and the background have almost the same color.*

*(d) Snake is affected by poor, uneven lighting.*

*(e) No reasons found for incorrect predictions.*

*(f) No snake can be seen here.*

**Figure 7:** Difficult snake image examples from the validation dataset that were misclassified by the model assuming the reasons given.

quality images (137 images, about 27 % of the whole dataset), which exhibited issues such as low resolution, compression artifacts, or blurred areas that could have adversely affected the accuracy of the model's predictions (Figure 7a). Another prevalent reason was the presence of a complex background in some images (87 images, about 17 % of the whole dataset), where visually cluttered backgrounds or multiple objects made it challenging for the model to accurately identify the snake object of interest (mostly the trees, branches and leaves) (Figure 7b). Additionally, in certain cases, the snake's colors were almost the same as the background, resulting in the snake blending in with the surroundings, making it difficult for the model to distinguish it accurately (87 images, about 17 % of the whole dataset) (Figure 7c). Furthermore, poor lighting was identified as a factor in some incorrect predictions (62 images, about 12 % of the whole dataset), where images had uneven lighting, shadows, or overexposure, which could have impacted the model's ability to accurately detect the snake (Figure 7d). While these reasons accounted for a sizable portion of the incorrect predictions, there were also cases where no specific reason could be identified for the incorrect prediction (39 images, about 8 % of the whole dataset) (Figure 7e). Another notable observation from the analysis is that approximately 21 % of the incorrect predictions (106 images) were attributed to the snake not being completely captured within the image. In some cases, the head of the snake was not visible in the image, while in other cases, the body was missing. This indicates that the partial presence or absence of the snake within the image played a notable role in contributing to the incorrect predictions made by

the model (Figure 7f). The incomplete presence of the snake in the image can pose challenges for the model in accurately detecting and classifying the snake object, as the model may rely on the complete representation of the snake's features, such as head shape, body pattern, and other visual cues, to make accurate predictions. When crucial parts of the snake are missing in the image, the model's ability to correctly classify the snake can be compromised. In the analysis of the dataset, it was also observed that some images had multiple reasons for incorrect predictions. For example, some images exhibited both low quality and poor lighting issues simultaneously (24 images, about 5 % of the whole dataset), while others had both complex background and color similarity with the background concerns (14 images, about 3 % of the whole dataset). This suggests that certain images may have had multiple factors contributing to the incorrect predictions, making it more challenging for the model to accurately classify the snake object.

## 6. Conclusion

This work presents a multimodal deep learning based model for snake species identification using image data in combination with additional metadata, including regional and endemic information. The presented model architecture allows joint feature learning of both modalities, obtained from a ConvNeXt V2 base image feature extractor CNN and a custom model that embeds the provided metadata, in an intermediate feature concatenation approach.

Subsequent ablation studies have investigated the influence of selected hyperparameters as well as deep learning techniques to further improve the proposed method. The results of the ablation studies showed, that pre-training on large fine-grained data sets, such as iNaturalist21, as well as using large image sizes could improve the downstream fine-tuning of the image feature extractor CNN. Furthermore, the results showed that the identification of snake species can be leveraged when additional metadata is considered. The proposed joint feature model proved to be a good approach that could even outperform other approaches considered in previous SnakeCLEF challenges, like weighting model outputs with regional prior distributions of snake species. The problem of the highly class imbalanced challenge data set was addressed by using the focal loss function with class balance re-weighting terms. The obtained results showed that this approach only marginally affected the snake species identification performance. The considered ArcFace loss that directly optimise the joint modality feature embedding to enforce higher similarity for intra-class samples and greater diversity for inter-class samples, proved to be a much better approach to improve snake species identification performance. Further refinements in terms of less costly snake species identification errors were achieved by integrating learnable attention-based MIL pooling over classical non-trainable operators into the model architecture to pool both TTA instances as well as instances of the same snake observation. Increasing the size of the training data set by including the validation data set only improved snake species identification very slightly, resulting in the best performing model achieving a macro $F_1$-Score of 81.9 % and challenge-specific metrics of 95.09 % Track 1 and 1149 Track 2.

## 7. Further Research

The proposed method offers several possibilities for further research. These generally include experiments of different learning techniques with better optimised hyperparameters or adjustments to the proposed model architecture (i.e. metadata model size or embedding layer sizes) that were not explicitly considered in the conducted ablation studies.

A more specific approach for future work would be a two-stage snake species classification process that first predicts whether the provided snakes in the image instances belong to a venomous or non-venomous species, as this information is also provided by the challenge data set. This 'venomous' information could then also be embedded as additional metadata using the proposed model architecture to predict the actual snake species. This approach may be able to further reduce the challenge specific metrics as they explicitly account for the confusion between venomous and non-venomous species.

Since the considered ArcFace loss performed well by enforcing higher embedding similarity for intra-class samples and higher diversity for inter-class samples, it may be worthwhile for future work to test deep learning methods that follow a similar direction. These would include unsupervised pre-training methods using contrastive loss techniques such as in SimCLR [51], or simpler approaches such as Bootstrap Your Own Latent (BYOL) [69].

## Acknowledgments

The work of Louise Bloch was partially funded by a PhD grant from University of Applied Sciences and Arts Dortmund, Dortmund, Germany.

## References

[1] L. Picek, M. Šulc, R. 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.

[2] 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: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022): 2022-09-05 – 2022-09-08, Bologna, Italy, Springer, 2022, pp. 257 – 285.

[3] A. Joly, H. Goëau, S. Kahl, L. Picek, C. Botella, D. Marcos, M. Šulc, M. Hrúz, 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.

[4] A. Joly, C. Botella, L. Picek, S. Kahl, H. Goëau, B. Deneu, D. Marcos, J. Estopinan, R. Chamidullin, M. Šulc, M. Hrúz, 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.

[5] J. M. Gutiérrez, J. J. Calvete, A. G. Habib, R. A. Harrison, D. J. Williams, D. A. Warrell, Snakebite envenoming, Nature Reviews Disease Primers 3 (2017). doi:10.1038/nrdp.2017.63.

[6] H. F. Williams, H. J. Layfield, T. Vallance, K. Patel, A. B. Bicknell, S. A. Trim, S. Vaiyapuri, The urgent need to develop novel strategies for the diagnosis and treatment of snakebites, Toxins 11 (2019). doi:10.3390/toxins11060363.

[7] L. Bloch, J. Böckmann, B. Bracke, C. M. Friedrich, Combination of object detection, geospatial data, and feature concatenation for snake species identification, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 1982–2013. URL: https://ceur-ws.org/Vol-3180/paper-158.pdf.

[8] 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, in: Working Notes of the 11th Conference and Labs of the Evaluation Forum (CLEF 2020): 2020-09-22 – 2020-09-25, Thessaloniki, Greece, 2020, p. 201. URL: http://ceur-ws.org/Vol-2696/paper_201.pdf.

[9] L. Bloch, C. M. Friedrich, EfficientNets and Vision Transformers for snake species identification using image and location information, in: Working Notes of the 12th Conference and Labs of the Evaluation Forum (CLEF 2020): 2021-09-21 – 2021-09-24, Bucharest, Romania, 2021, pp. 1477–1498. URL: http://ceur-ws.org/Vol-2936/paper-126.pdf.

[10] A. James, D. Kumar, B. Mathews, S. Sugathan, Discriminative histogram taxonomy features for snake species identification, Human-Centric Computing and Information Sciences 4 (2014). doi:10.1186/s13673-014-0003-0.

[11] A. Amir, N. A. H. Zahri, N. Yaakob, R. B. Ahmad, Image classification for snake species using machine learning techniques, in: S. Phon-Amnuaisuk, T.-W. Au, S. Omar (Eds.), Proceedings of the Computational Intelligence in Information Systems Conference (CIIS 2016): 2016-11-18 – 2016-11-20, Brunei, Brunei Darussalam, Springer International Publishing, Cham, 2017, pp. 52–59. doi:10.1007/978-3-319-48517-1_5.

[12] Z. Yang, R. Sinnott, Snake detection and classification using deep learning, in: Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS 2021): 2021-01-05 – 2021-01-08, Maui, Hawaii, US, 2021, pp. 1212–1221. doi:10.24251/hicss.2021.148.

[13] A. Patel, L. Cheung, N. Khatod, I. Matijosaitiene, A. Arteaga, J. W. Gilkey, Revealing the unknown: Real-time recognition of galápagos snake species using deep learning, Animals 10 (2020) 806. doi:10.3390/ani10050806.

[14] M. Vasmatkar, I. Zare, P. Kumbla, S. Pimpalkar, A. Sharma, Snake species identification and recognition, in: Proceedings of the IEEE Bombay Section Signature Conference (IBSSC

2020): 2020-12-04 – 2020-12-06, Mumbai, India, 2020, pp. 1–5. doi:10.1109/IBSSC51096.2020.9332218.

[15] C. Abeysinghe, A. Welivita, I. Perera, Snake image classification using Siamese networks, in: Proceedings of the 3rd International Conference on Graphics and Signal Processing (ICGSP 2019): 2019-06-01 – 2019-06-03, Hong Kong, Hong Kong, Association for Computing Machinery, New York, NY, USA, 2019, p. 8–12. doi:10.1145/3338472.3338476.

[16] I. S. Abdurrazaq, S. Suyanto, D. Q. Utama, Image-based classification of snake species using convolutional neural network, in: Proceedings of the International Seminar on Research of Information Technology and Intelligent Systems (ISRITI 2019): 2019-12-05 – 2019-12-06, Yogyakarta, Indonesia, Institute of Electrical and Electronics Engineers (IEEE), 2019, pp. 97–102. doi:10.1109/isriti48646.2019.9034633.

[17] 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) 1–19. doi:10.1371/journal.pntd.0010647.

[18] J. Deng, W. Dong, R. Socher, L. 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 (CVPR 2009): 2009-06-20 – 2009-06-25, Miami Beach, Florida, US, Institute of Electrical and Electronics Engineers (IEEE), 2009, pp. 248–255. doi:10.1109/cvpr.2009.5206848.

[19] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 1137–1149. doi:10.1109/tpami.2016.2577031.

[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceesings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016): 2016-07-27 – 2016-07-30, Las Vegas, Nevada, US, Institute of Electrical and Electronics Engineers (IEEE), 2016, pp. 770–778. doi:10.1109/cvpr.2016.90.

[21] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Y. Bengio, Y. LeCun (Eds.), Conference Track Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015): 2015-05-07 – 2015-05-09, San Diego, California, US, 2015.

[22] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017): 2017-07-21 – 2017-07-26, Honolulu, Hawaii, 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243.

[23] 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 (CVPR 2018): 2018-06-18–2018-06-22, Salt Lake City, Utah, US, 2018, pp. 4510–4520. doi:10.1109/CVPR.2018.00474.

[24] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah, Signature verification using a Siamese time delay neural network, in: J. Cowan, G. Tesauro, J. Alspector (Eds.), Advances in Neural Information Processing Systems (NIPS 1993): Denver, Colorado, US, volume 6, Morgan-Kaufmann, 1994, pp. 737–744. URL: https://proceedings.neurips.cc/paper/1993/

file/288cc0ff022877bd3df94bc9360b9c5d-Paper.pdf.

[25] G. Koch, R. Zemel, R. Salakhutdinov, Siamese neural networks for one-shot image recognition, in: Proceedings of the Deep Learning workshop of the International Conference on Machine Learning (ICML 2015): 2015-06-06 – 2015-06-11, Lille, France, volume 2, 2015. URL: https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf.

[26] A. M. Durso, G. K. Moorthy, S. P. Mohanty, I. Bolon, M. Salathé, R. Ruiz de Castañeda, Supervised learning computer vision benchmark for snake species identification from photographs: Implications for herpetology and global health, Frontiers in Artificial Intelligence 4 (2021) 17. doi:10.3389/frai.2021.582110.

[27] 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 the 13th Conference and Labs of the Evaluation Forum (CLEF 2022): 2022-09-05 – 2022-09-08, Bologna, Italy, 2022, pp. 1957 – 1969.

[28] L. Picek, R. Ruiz De Castañeda, A. M. Durso, P. M. Sharada, Overview of the SnakeCLEF 2020: Automatic snake species identification challenge, in: Proceedings of the 11th Conference and Labs of the Evaluation Forum (CLEF 2020): 2020-09-22 – 2020-09-25, Thessaloniki, Greece, 2020, p. 258. URL: http://ceur-ws.org/Vol-2696/paper_258.pdf.

[29] M. G. Krishnan, Impact of pretrained networks for snake species classification, in: Proceedings of the 11th Conference and Labs of the Evaluation Forum (CLEF 2020): 2020-09-22 – 2020-09-25, Thessaloniki, Greece, 2020, p. 194. URL: http://ceur-ws.org/Vol-2696/paper_194.pdf.

[30] T. Ridnik, E. Ben-Baruch, A. Noy, L. Zelnik-Manor, ImageNet-21k pretraining for the masses, in: Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Datasets and Benchmarks Track (Round 1): 2021-12-06 – 2021-12-14, Online-only Conference, 2021, pp. 1–12. URL: https://openreview.net/forum?id=Zkj_VcZ6ol.

[31] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017): 2017-10-22 – 2017-10-29, Venice, Italy, Institute of Electrical and Electronics Engineers (IEEE), 2017, pp. 2980–2988. doi:10.1109/iccv.2017.322.

[32] M. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning (ICML 2019): 2019-06-10 – 2019-06-15, Long Beach, California, US, volume 97, 2019, pp. 6105–6114. URL: http://proceedings.mlr.press/v97/tan19a.html.

[33] R. Borsodi, D. Papp, Incorporation of object detection models and location data into snake species classification, in: Working Notes of the 12th Conference and Labs of the Evaluation Forum (CLEF 2021): 2021-09-21 – 2021-09-24, Bucharest, Romania, 2021, pp. 1499–1511. URL: http://ceur-ws.org/Vol-2936/paper-127.pdf.

[34] M. Tan, R. Pang, Q. V. Le, EfficientDet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2020): 2020-06-14 – 2020-06-19, Online-only conference, 2020, pp. 10781–10790. URL: https://openaccess.thecvf.com/content_CVPR_2020/papers/Tan_EfficientDet_Scalable_and_Efficient_Object_Detection_CVPR_2020_paper.pdf.

[35] R. Chamidullin, M. Šulc, J. Matas, L. Picek, A deep learning method for visual recognition

of snake species, in: Working Notes of the 12th Conference and Labs of the Evaluation Forum (CLEF 2021): 2021-09-21 – 2021-09-24, Bucharest, Romania, 2021, pp. 1512–1525. URL: http://ceur-ws.org/Vol-2936/paper-128.pdf.

[36] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, A. Smola, ResneSt: Split-attention networks, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022, pp. 2736 – 2746. doi:10.1109/cvprw56347.2022.00309.

[37] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017): 2022-07-21 – 2022-07-26, Honolulu, Hawaii, 2017, pp. 5987–5995. doi:10.1109/CVPR.2017.634.

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: Proceedings of the 9th International Conference on Learning Representations (ICLR 2021): 2021-05-03 – 2021-05-07, Online-only conference, 2021, pp. 1–21. URL: https://openreview.net/forum?id=YicbFdNTTy.

[39] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999–3007. doi:10.1109/ICCV.2017.324.

[40] Y. Shen, X. Sun, Z. Zhu, When large kernel meets vision transformer: A solution for SnakeCLEF & FungiCLEF, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 2199 – 2211. URL: https://ceur-ws.org/Vol-3180/paper-175.pdf.

[41] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11966 – 11976. doi:10.1109/cvpr52688.2022.01167.

[42] 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) 1 – 13. doi:10.1109/tpami.2022.3206108.

[43] Z. Liu, Y. Lin, Y. Cao, J. Qiu, Y. Wei, Z. G. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012 – 10022. doi:10.1109/iccv48922.2021.00986.

[44] S. Müller, F. Hutter, TrivialAugment: Tuning-free yet state-of-the-art data augmentation, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 754 – 762. doi:10.1109/iccv48922.2021.00081.

[45] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in: F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), Advances in Neural Information Processing Systems (NIPS 2012): 2012-12-03 – 2012-12-06, Lake Tahoe, Nevada, US, volume 25, Curran Associates, Inc., 2012, pp. 1097–1105. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), 2015-06-07 – 2015-06-12, Boston, Massachusetts, US, 2015, pp. 1–9. doi:10.1109/CVPR.2015.7298594.

[47] L. Yang, X. Li, R. Song, K. Zhu, G. Li, Solution for SnakeCLEF 2022 by tackling long-tailed categorization, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 2253–2261. URL: https://ceur-ws.org/Vol-3180/paper-180.pdf.

[48] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, S. Kumar, Long-tail learning via logit adjustment, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=37nvvqkCo5.

[49] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9268 – 9277. doi:10.1109/cvpr.2019.00949.

[50] C. Zou, F. Xu, M. Wang, W. Li, Y. Cheng, Solutions for fine-grained and long-tailed snake species recognition in SnakeCLEF 2022, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 2291–2300. URL: https://ceur-ws.org/Vol-3180/paper-183.pdf.

[51] T. Chen, S. Kornblith, M. Norouzi, G. E. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning (ICML), volume 1, 2020, pp. 1597 – 1607.

[52] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, S. Yan, MetaFormer is actually what you need for vision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022): 2022-06-19 – 2022-06-24, New Orleans, Louisiana, US, 2022, pp. 10819–10829.

[53] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Only Look Once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016): 2016-06-26 – 2016-07-01, Las Vegas, Nevada, US, 2016, pp. 779–788. doi:10.1109/CVPR.2016.91.

[54] M. Tan, Q. Le, EfficientNetV2: smaller models and faster training, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning (ICML 2021): 2021-07-18 – 2021-07-24, Online-only conference, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 10096–10106. URL: http://proceedings.mlr.press/v139/tan21a/tan21a.pdf.

[55] 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, 2023. arXiv:2301.00808v1.

[56] D. Hendrycks, K. Gimpel, Bridging nonlinearities and stochastic regularizers with Gaussian error linear units, CoRR abs/1606.08415 (2016). URL: http://arxiv.org/abs/1606.08415.

[57] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: Proceedings of the International Conference on Learning Representations (ICLR 2019) 2019-05-06 – 2019-05-

09, New Orleans, Louisiana, US, 2019, pp. 1–18. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[58] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, J. Choe, CutMix: Regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019): 2019-10-27 -– 2019-11-02, Seoul, Korea, 2019, pp. 6022–6031. doi:10.1109/ICCV.2019.00612.

[59] E. D. Cubuk, B. Zoph, J. Shlens, Q. V. Le, RandAugment: Practical automated data augmentation with a reduced search space, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR 2020): 2020-06-14 – 2020-06-19, Online-only conference, 2020, pp. 3008–3017. doi:10.1109/CVPRW50498.2020.00359.

[60] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, Proceedings of the 34 Conference on Artificial Intelligence (AAAI 2020): 2020-02-07 – 2020-02-12, New York, New York, US 34 (2020) 13001–13008. doi:10.1609/aaai.v34i07.7000.

[61] J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, 2016. arXiv:1607.06450.

[62] S. Y. Boulahia, A. Amamra, M. R. Madi, S. Daikh, Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition, Machine Vision and Applications 32 (2021). URL: https://doi.org/10.1007%2Fs00138-021-01249-8. doi:10.1007/s00138-021-01249-8.

[63] M. Ilse, J. M. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: International Conference on Machine Learning, volume 80, 2018, pp. 2127 – 2136.

[64] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, S. Zafeiriou, ArcFace: Additive angular margin loss for deep face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (2022) 5962 – 5979. doi:10.1109/tpami.2021.3087709.

[65] R. Wightman, PyTorch image models, 2019. doi:10.5281/zenodo.4414861.

[66] 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: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems (Neurips 2019): 2019-12-08 – 2019-12-14, Vancouver, Canada, Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[67] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, H. Wu, Mixed precision training, in: Proceedings of the International Conference on Learning Representations (ICLR 2018): 2018-04-30 – 2018-05-03, Vancouver, Canada, 2018, pp. 1–12. URL: https://openreview.net/forum?id=r1gs9JgRZ.

[68] m. Grant Van Horn, iNat challenge 2021 - FGVC8, 2021. URL: https://kaggle.com/competitions/inaturalist-2021.

[69] J.-B. Grill, F. Strub, F. Altchè, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. Guo, B. Azar, M. G. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap your own latent - a new approach to self-supervised learning, in: Conference on Neural Information Processing Systems (NeurIPS), 2020.