Deep Visual-Genetic Biometrics for Taxonomic Classification of Rare Species

Tayfun Karaderi
Dept of Computer Science
University of Bristol
vm19402@bristol.ac.uk

Tilo Burghardt
Dept of Computer Science
University of Bristol
tilo@cs.bris.ac.uk

Raphaël Morard
MARUM
University of Bremen
rmorard@marum.de

Daniela N. Schmidt
School of Earth Sciences
University of Bristol
d.schmidt@bristol.ac.uk

Abstract

Visual as well as genetic biometrics are routinely employed to identify species and individuals in biological applications. However, no attempts have been made in this domain to computationally enhance visual classification of rare classes with little image data via genetics. In this paper, we thus propose aligned visual-genetic learning as a new application domain with the aim to implicitly encode cross-modality associations for improved performance. We demonstrate for the first time that such alignment can be achieved via deep embedding models and that the approach is directly applicable to boosting long-tailed recognition (LTR), particularly for rare species. We experimentally demonstrate the efficacy of the concept via application to microscopic imagery of 30k+ planktic foraminifer shells across 32 species when used together with independent genetic data samples. Most importantly for practitioners, we show that visual-genetic alignment can significantly benefit visual-only recognition of the rarest species. Technically, we pre-train a visual ResNet50 deep learning model using triplet loss formulations to create an initial embedding space. We re-structure this space based on genetic anchors embedded via a Sequence Graph Transform (SGT) and linked to visual data by cross-domain cosine alignment. We show that an LTR approach improves the state-of-the-art across all benchmarks and that adding our visual-genetic alignment improves per-class and particularly rare tail class benchmarks significantly further. Overall, visual-genetic LTR training raises rare per-class accuracy from 37.4% to benchmark-beating 59.7%. We conclude that visual-genetic alignment can be a highly effective tool for complementing visual biological data containing rare classes. The concept proposed may serve as an important future tool for integrating genetics and imageomics towards a more complete scientific representation of taxonomic spaces and life itself. Code, weights, and data splits are published for full reproducibility.

1. Introduction

1.1. Motivation

Visual, Genetic, and Long-Tailed Data in Biology. Both genetic and visual biometrics are extensively utilised to support species and individual identification in biological applications [28, 48, 50]. Yet, modalities are usually learned and processed independently without explicitly considering cross-modal information. In how far information from genetics of a species can assist classification of visuals of the phenotype is of particular practical interest given that visual source or training data for imageomics [3, 46, 48, 53] may be prohibitively limited for some classes (e.g. visual samples for very rare species). In fact, the distribution of most biological species datasets [19, 39, 51] follows a 'long-tailed'

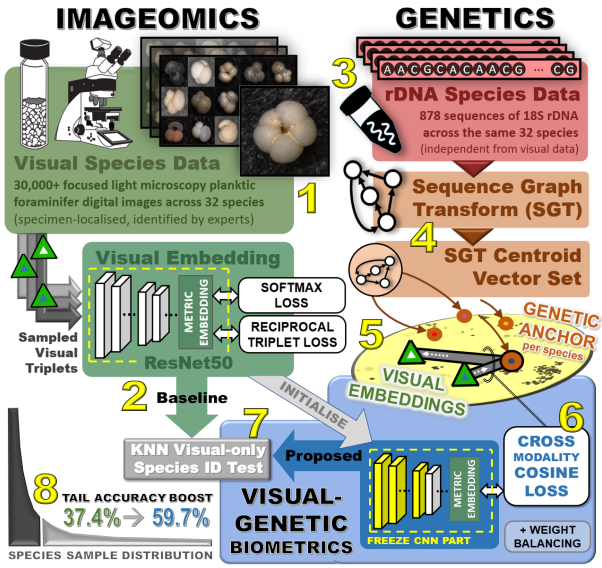

Figure 1. **Visual-Genetic Co-Learning Architecture.** This work combines imageomics with genetics to improve on visual biometric recognition of particularly rare species in biology. ***(1)*** 30k+ images of 32 species from the EndlessForams dataset are used to train ***(2)*** a traditional metric deep learning baseline for species classification. However, we also use ***(3)*** independent rDNA sequence data ***(4)*** transformed into the same embedding space via SGT. Each species can now also be represented via genetic anchor information. ***(5)*** Cross-modality triplets of a genetic anchor and a positive and negative visual embedding are co-used via ***(6)*** the cosine transform to learn a visual-genetic species space adjusting late layers of the visual embedding. We show that ***(7)*** cosine KNN visual-only testing of such a network when weight-balanced can ***(8)*** significantly improve performance, particularly for rare species.

pattern or at least contain many rare classes. Thus, models trained on such data often struggle accurately to encode and consequently recognise less common species.

Cross-Modal Taxonomic Information. In living organisms, the relationship between taxa is traditionally determined by their genetics. However, sister taxa that are closely related commonly share morphological features observable via imaging techniques too. Consequently, visual and genetic feature distances between species as well as their orientation in any overarching, cross-domain feature space should be related to some degree. We therefore hypothesise that enriching imageomic representation spaces via information transfer from genetics may enhance deep visual species representation models particularly when the latter is built under long-tailed training data limitations.

Deep Visual-Genetic Embedding. In this paper and following the above line of argument, we explore enriching deep imageomics for taxonomic species classification with independently sourced genetic information in order to improve visual-only species recognition performance for long-tailed datasets. Fig. 1 provides a schematic overview of the proposed approach and how it combines imageomics with genetics to improve visual classification of rare species. In particular, we propose utilising relative orientation information from rDNA (ribosomal genes DNA) embeddings to optimise visual embeddings. Technically, state-of-the-art (SOTA) triplet loss formulations [23] for learning metric visual classification spaces are expanded across modalities in a second learning stage that uses rDNA anchors and cosine similarity metrics to draw in additional information from the genetic domain. We test the approach on the challenging task of identifying planktic foraminifer species at scale, which is of critical importance for paleoclimatology.

1.2. Paper Contributions

To the best of our knowledge, this work employs rDNA information to guide the orientations of deep visual embeddings for species recognition for the first time. Our key contributions are as follows:

- **Concept.** We propose a new type of modality integration for imageomics combining visual and genetic information in one metric space usable for inference.
- **Implementation.** We provide a deep transfer learning framework that implements the new concept for biological applications and publish the full source code of this visual-genetic co-learning architecture.
- **Experimentation.** We demonstrate that the proposed implementation achieves state-of-the-art accuracy results whilst significantly boosting tail class recognition performance for a visual species recognition task on a large 30k+ example image set covering 32 species of difficult to identify planktic foraminifers. We provide all weights and data splits for full reproducibility.

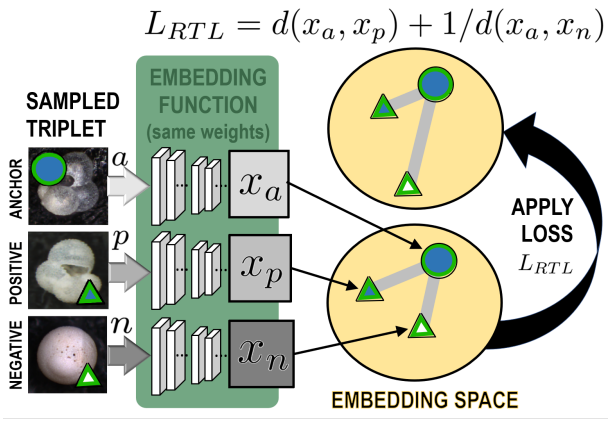

Figure 2. **Metric Learning via Reciprocal Triplet Loss.** Using triplets of an anchor sample (disc), a sample of the same (blue core triangle), and a sample of a different class (white core triangle) yields three vectors x_a, x_p, and x_n. The parameter-free loss function L_{RTL} adjusts Eucledian distances $d(\cdot, \cdot)$ in embedding space by shortening $d(x_a, x_p)$ and lengthening $d(x_a, x_n)$ and can in conjunction with other losses [23] yield state-of-the-art results for learning species spaces directly usable for inference.

- **Analysis.** We structurally analyse the novel visual-genetic spaces created and interpret as well as visualise effects of cross-modal alignment.

2. Related Work

2.1. Microfossil Classification via Metric Learning

Metric Learning. The task of learning a classification-relevant similarity function from data is known as Metric Learning [2, 4, 11, 41, 47], that is creating an embedding function that produces feature vectors in a space where samples of the same class cluster together far away from other data. Back-propagation with contrastive or triplet losses in the mix of cost functions [2, 23] can effectively implement such a system. The resulting distance metric can be used to perform tasks such as classification for both open or closed-set scenarios [2], clustering, and retrieval [12].

Deep Microfossil Classification. Recent taxonomic applications [23] of deep metric learning to visual microfossil identification achieve SOTA performance beyond other CNN approaches [17, 30] when evaluated on the large Endless Forams dataset [17]. Reciprocal Triplet Loss, as illustrated in Fig. 2, together with the SoftMax Loss form the key cost functions used in these SOTA imageomics systems [23]. By combining the two losses both class-relative and class-absolute information can be utilised during learning. However, we note that the principal concept of adjusting distances via an anchor and nearby samples is not bound to a single modality such as vision. Instead, it provides an opportunity to transfer information across modalities by mixing anchor and sample modalities for alignment of different domains within one space (see Section 4.2).

2.2. Long-Tailed Recognition

Natural Data Collections. The distribution of most biological datasets including microfossil [17] datasets and other natural image collections [39, 51] is long-tailed, that is a few classes have a lot more data than many other classes. This uneven distribution causes most machine learning models to perform poorly on the many rare classes.

Specific Long Tail Techniques. Long-tailed Recognition (LTR) techniques are used to improve the performance of models with a focus on rare classes. Different LTR methods have been proposed, such as re-sampling the training data to balance class distributions [18, 36], re-weighting classes and individual training examples [5], transferring feature representations [54] from common classes to rare classes, relating head and tail information [39], decoupling feature learning and classifier learning [21, 56], or using self-supervised or ensemble models [52]. For a more comprehensive overview of LTR, refer to the survey paper [55]. Since our target data are of taxonomic nature and contain rare classes LTR techniques offer a tool to potentially improve performance. In addition, this allows us to separate LTR compensation from the effect achieved by integrating genetic information. Thus, in this work, we explicitly utilise weight balancing [1], one of the latest state-of-the-art regularization approaches to LTR making use of weight decay and Max-Norm constraints (see Section 6.1 for its impact).

2.3. Genetic Data Embedding

Complexity of Genetic Data. In order to integrate genetic information into other spaces sequence data needs to be placed or 'embedded' within them. Genetic sequence embedding is a challenging task due to the structuredness of potentially unaligned sequences of arbitrary length and content. In addition, a good embedding function for sequences has to capture both short- and long-term dependencies between symbols in the sequences.

Embedding of Genetic Sequences. For this task, Ranjan et al. [40] proposed the approach of a Sequence Graph Transform (SGT), a technique that represents sequences via the statistical relationships between symbols and casts this information into a feature vector. We opt to utilise this approach for creating embedding functions since it captures both local and global patterns into fixed-dimensional embedding vectors. In addition, the produced features are indeed interpretable where components represent directional dependencies between symbol pairs. Fig. 3 visualises key features of the SGT technique. Note that, the grouping of pairs of symbols towards a new alphabet ensures the embedding vector is not dimensionality-limited – despite the fact that the square of symbol cardinality is the fixed embedding vector dimension. Section 4.2 details the procedure followed for genetic data in this work.

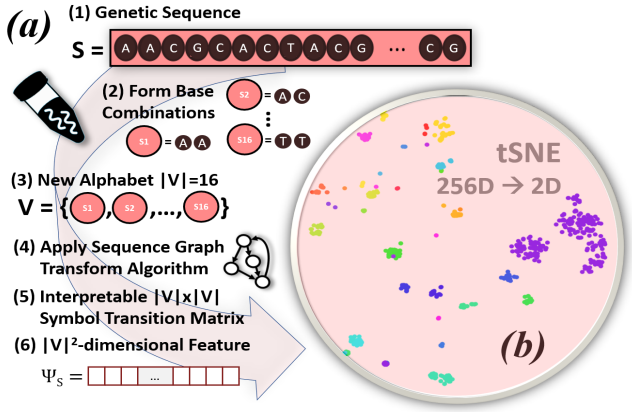

Figure 3. **Genetic Embeddings via Sequence Graph Transform.** ***(a)*** After creating an alphabet V of 16 symbols from pairs of the 4 bases, we use the Sequence Graph Transform to map a given genetic sequence, S into a feature vector Ψ_S of size $|V|\times|V|$. This vector is used as embedding and can also be interpreted as a graph that captures both long and short-term interactions between different symbols of the sequence. ***(b)***: 2D t-SNE visualisation of rDNA sequence embeddings for the 32 microfossil species.

2.4. Cross-Domain Transfer Learning

Knowledge Carry-over. Applying knowledge gained from one domain to a different – but related – domain for improved performance is known as Cross-Domain Transfer Learning. Such a methodology is clearly most useful if data is limited in the target domain itself. Since long-tailed taxonomic image collections exhibit exactly this property, we propose to transfer knowledge learned from the genetic domain into the visual domain.

Vision-Language Transfer. Classic transfer learning approaches in the literature often focus on bridging image and text domains. They include DeViSE (Deep Visual-Semantic Embedding) [10] to perform zero-shot image classification via Word2Vec [32] embeddings used to link visual and language-based semantics. Karpathy et al. proposed VSA (Deep Visual-Semantic Alignments) [24] for this task, which uses R–CNN [13] and BRNN [43] to extract image features and text features, respectively. Kiros et al. proposed UVSE (Unifying Visual-Semantic Embedding) [26], which uses a VGG-19 [45] network to extract image features and an LSTM [15] network for text. This was later extended to include hard negative mining in VSE++ [8].

Concept of Compatible Representation. One key feature of most approaches is the mapping of different domains into a common data format where information transfer becomes possible. Our cross-modality learning follows this concept by mapping visuals via ResNet50 and rDNA information via SGT into a common space. We use settings similar to DeViSE for cross-domain training.

Alignment Mechanism. DeViCE establishes a shared

embedding space for visual and semantic features, originally utilizing CNNs for visual descriptors and Word2Vec for semantic descriptors. For alignment, the CNN's late projection layers are trained to align visual descriptors with semantic ones in a shared space, originally capturing image-text relationships. Inspired by DeViCE, we use CNNs for image representations and SGT for genetic ones. Similarly, we train projection layers of the CNN to align visual descriptors with genetic ones, as depicted in Fig. 1 and Fig. 5.

3. Datasets

3.1. Endless Forams Imagery

Taxa with Visual-Genetic Support. We utilise 32 out of 35 species from the public Endless Forams image library [17] for all our experiments. That is we utilise exactly those species for which we were able to gather sufficient genetic rDNA information. Figure 4 shows an associated phylogenetic tree. Endless Forams is one of the largest datasets of its kind and freely available at `endlessforams.org` supporting full reproducibility of our experiments. In its entirety, it contains 34,640 species-labelled and location-centred images of 35 different marine calcareous plankton species (foraminifera) as detailed phylogenetically in Fig. 4. Further, Fig. 6 depicts the covered taxa and Fig. 7 shows a histogram of sample sizes.

3.2. Endless Forams Genetics

Genetic Data Sources. For genetic information, we use a fragment of the 18S rDNA sequences (ribosomal genes DNA) of the 32 species selected in the study to serve as embeddings. We select the sequences from the Planktonic Foraminifera Reference Database [33] publically available at `http://pfr2.sb-roscoff.fr,` and included sequences published afterwards [34, 35, 49]. Overall, we retained 878 sequences covering the entire barcode selected for foraminifera studies [38] and provide the sequence list as supplementary material.

Pylogenetic Tree Inference. As discussed, Figure 4 reconstructs the evolutionary relationships between extant species via phylogenetic inference from sequence representation to a tree structure. We included further species to cover the full phylogenetic spectrum of planktonic foraminifera in this representation. The sequences were automatically aligned using MAFFT v.7 [25], and we inferred the phylogenetic tree with RAxML-NG [27] using the model GTR+I+G that was selected with Modeltest-NG [6] and with 100 non-parametric bootstrap runs.

4. Generating Visual-Genetic Spaces

4.1. Metric Visual-Only Pre-Training.

Constructing Latent Image Embeddings. In order to construct a maximally rich initialisation of a task-aligned data space before visual-genetic integration we first construct a traditional deep mapping from source images to a class-distinctive embedding space [23]. The simplest way of creating such a metric embedding is via the use of a basic contrastive loss L_C [14] using pairs of data points:

$$L_C = \frac{(1-Y)}{2} d(x_1, x_2) + \frac{Y}{2} max(0, \alpha - d(x_1, x_2)), \quad (1)$$

where x_1 and x_2 are the embedded input vectors, Y is a binary label denoting class equivalence/difference for the two inputs, and $d(\cdot, \cdot)$ is the Euclidean distance between two embeddings. However, this formulation cannot put similarities and dissimilarities between different embedding pairs in relation. A triplet loss formulation [42] instead utilises three embeddings x_a, x_p and x_n denoting an anchor, a positive example of the same class, and a negative example of a different class, respectively:

$$L_{TL} = max(0; d(x_a, x_p) - d(x_a, x_n) + \alpha), \quad (2)$$

where α is the margin hyper-parameter. Reciprocal Triplet Loss as visualised in Figure 2 removes the need for this parameter [31] and naturally accounts for offsetting the impact of large margins far away from the anchor:

$$L_{RTL} = d(x_a, x_p) + 1/d(x_a, x_n). \quad (3)$$

As shown by recent work [16,29], including a SoftMax term in this loss can improve performance further. Thus, SoftMax and Reciprocal Triplet Loss can be combined into a standard formulation used here and published in [2] as a mixture with balancing hyper-parameter λ:

$$L = -log(\frac{e^{x_{class}}}{\sum_i e^{x_i}}) + \lambda L_{RTL}. \quad (4)$$

Application-Specific Relevance and Baseline. For the foraminifer classification problem at hand this allows for the use of both relative inter-species information captured by the L_{RTL} component as well as absolute species information captured by the SoftMax term as back-propagation gradient components. Training a latent embedding space as described essentially acts as a single modality baseline (see Fig. 1), replicating a SOTA image-only deep learning solution following [23] to solve the biometric species identification problem.

4.2. Genetic to Visual Information Transfer

Multi-Symbol Genetic Embedding. rDNA sequences as resulting from genetic sources as described in Section 3.2 strictly contain four base symbols, that is A, C, G, T. Thus, the SGT algorithm applied to this raw data would produce an embedding vector of very low dimensionality, that is $4 \times 4 = 16$. For rDNA vectors to structurally fit the high-dimensionality required for visual embeddings, we therefore bin every two symbol terms in rDNA sequences to

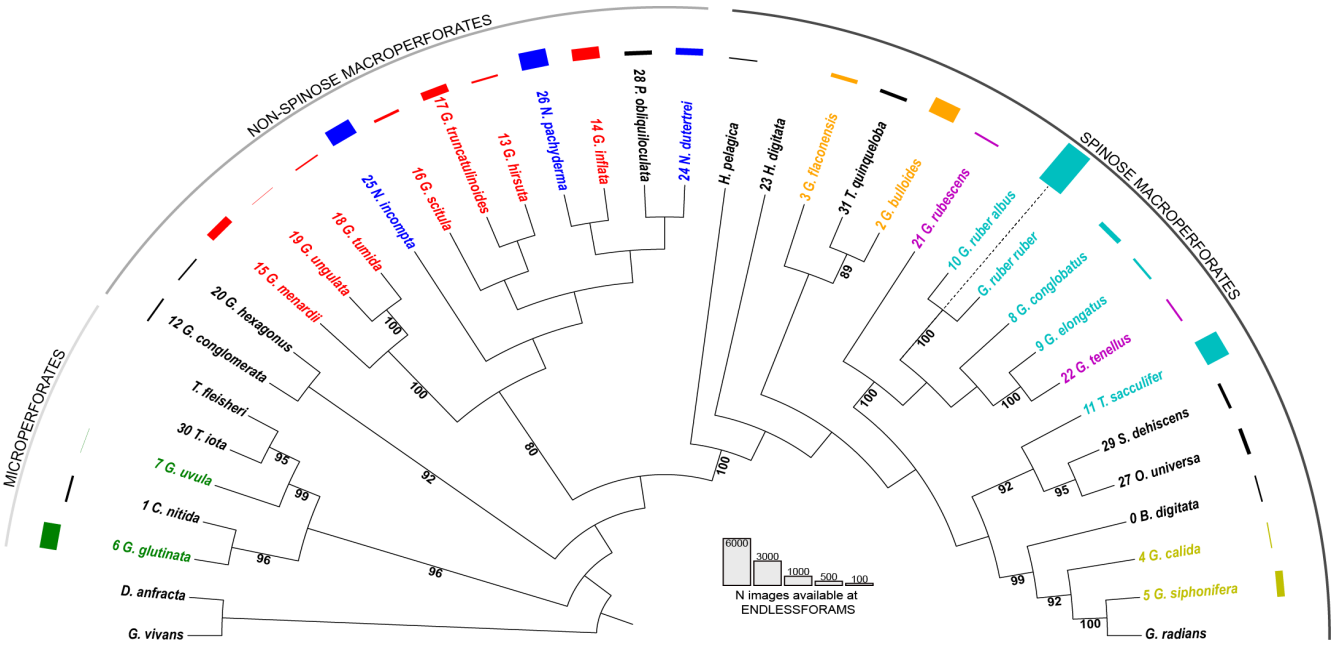

Figure 4. **Phylogenetic Reconstruction of Relevant Foraminifera.** RAxML phylogenetic inference showing the relationships between the extant planktonic foraminifera without branch length. The values next to the branches indicate the bootstrap values and the bars next to the species names the number of images used in the study. Numbers 0 to 31 associated with the taxa used in experiments align with the visual depictions in Fig. 6 and the detailed histogram in Fig. 7. The tree is rooted on the phylogenetically basal *G. vivans* and *D. anfracta*.

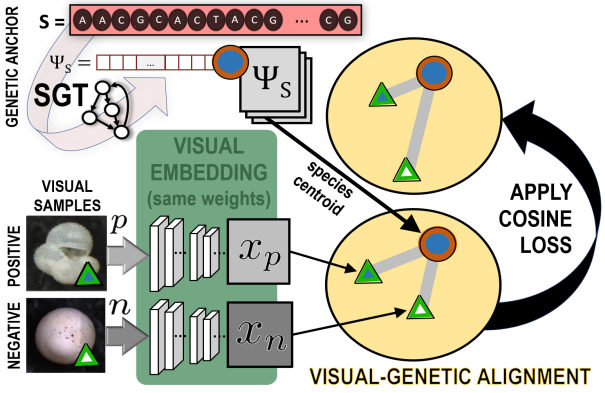

Figure 5. **Visual-Genetic Alignment via Transfer Learning.** The proposed cross-modality alignment of pre-trained ResNet50 visual embeddings towards genetic anchors generated via SGT uses a mixed-modality triplet cosine loss to transfer information. See Fig. 2 for further details regarding symbol semantics.

form a new alphabet of $4 \times 4 = 16$ symbols made of 4 base symbols. Application of SGT to this new symbol set then creates embeddings of size $16 \times 16 = 256$ as required. Figure 3 depicts a 2D t-SNE representation of the resulting genetic embedding space. Note that the embedding dimension may be adjusted to some lower size via compression via a trained fully connected or convolutional layer.

Visual Model Alignment Towards Genetics. The visually pre-trained ResNet50 model (see Section 4.1) is used as initial embedding function. Since the dimensional structure of this space is identical to the rDNA embeddings, the latter can be used to guide visual embedding positions. For each species a single rDNA target embedding is calculated as a genetic anchor defined as the median vector over all available rDNA embeddings of the taxa. Note that the choice of median vector over the mean vector is to eliminate the impact of any outliers that may be present in the datasets. Given this, we freeze the convolutional layers of the ResNet50 model and tune the remaining projection layers to capture cross-domain information. Methodologically, this is achieved by using a triplet formulation that uses Cosine distances. That is, for anchor-positive pairs the loss is defined as $1 - cos(x_1, x_2)$, while for anchor-negative pairs, the loss is defined as $max(0, cos(x_1, x_2)) - m)$ and these two terms are summed to form a loss L_{Cosine}. The parameter $m = 0.5$ is the margin set as recommended in [37]. Repeated application of this loss moves the model towards a higher orientational alignment between visual model and genetic anchors transferring information from the latter towards the visual imageomics classification model.

Choise of Distance Metric. Cosine distances operate invariant to vector scaling, ie. their application is implicitly scale-normalized which is critical for information transfer between non-aligned spaces. Experimentally, our triplet formulation of the cosine embedding loss boosted tail-class performance when compared to using standard Euclidean distances L_{RTL} for alignment.

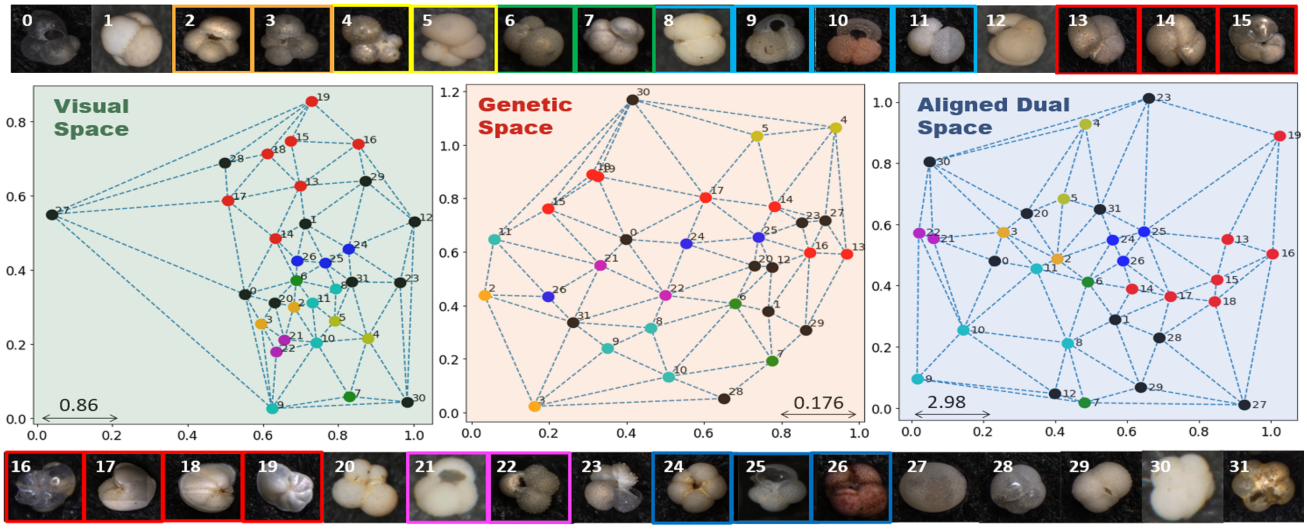

Figure 6. **Structure of Latent Species Spaces.** Visual samples of all 32 numbered taxa together with a 2D visualisation of Delaunay Triangulations of the taxa centroids for three latent spaces constructed: ***(left)*** visual-only baseline space, ***(middle)*** SGT-created genetic space, and ***(right)*** our proposed visual-genetic dual space still permits for visual-only inference and reveals qualitatively improved grouping of genera (shown by colours) together with more equidistant taxa spacing compared to the visual model resulting in superior per-class inference. Note that 256D latent spaces are approximately visualised in 2D by taking cosine distance matrices and minimizing a global energy function via the Kamada Kawai [20] algorithm. Different colors (excluding black) correspond to different genera in line with Fig. 7.

Row	Method	Architecture	Modalities	PC Acc	Tail PC Acc	Head PC Acc	Acc
1	Hsiang et al.* [17]	VGG16	visual-only	69.9	43.3	88.5	87.4
2	Karaderi et al.* [23]	ResNet50 (WD)	visual-only	76.8	43.7	89.1	**89.6**
3	Baseline 1 (Naive)	ResNet50	visual-only	73.6	37.4	88.9	89.1
4	Baseline 2 [23]	ResNet50 (WD)	visual-only	75.9	47.6	87.9	88.3
5	LTR Weight Balancing	ResNet50 (WD+M)	visual-only	77.2	48.2	**89.2**	**89.6**
6	Naive (+A)	Resnet50	visual-genetic	75.3	45.6	88.2	88.0
7	**Ours** (LTR+A)	ResNet50 (WD+M)	visual-genetic	**77.6**	**59.7**	87.4	88.0

LTR genetics

overall tail gains

Table 1. **Quantitative Results with Focus on Per-Class and Tail Performance.** Experimental accuracy (Acc) results for the Endless Forams visual-only test set across the 32 used taxa at 8bit grayscale and 160 × 160 pixels resolution. Starting from a Naive visual vanilla model (row 3) adding LTR weight balancing here for the first time to the domain (row 5) improves all benchmarks and beats the SOTA Baseline 2 model (row 4). Further adding our proposed visual-genetic alignment to this LTR training (row 7) boosts tail performance to 59.7% and can further enhance per-class accuracy at a cost of only 1.6% overall accuracy loss. [Tail Per-Class (PC) accuracy is for classes with less than 100 visual samples. Head PC accuracy is for classes with more than 1,000 samples.] [A: visual-genetic alignment, WD: Weight Decay, M: Maxnorm. *: not directly comparable models with access to 35 visual training classes.]

5. Experimental Setup

5.1. Implementation Details

Basic Training Details. For all experiments, we use a PyTorch-implemented metric learning architecture with a ResNet50 backbone pre-trained on ImageNet [7] using two fully connected projection layers to map the standard ResNet50 feature space of width 2048 to first 1000 and then to our 256-element feature vector, which is the same embedding size as for the genetics. For universal comparability, we utilise a fixed, withheld test set of $6,801$ images for performance stipulation, whilst using the remaining $27,731$ images augmented via rotations, scale, and Gaussian noise transforms for training. Exact sample-accurate data splits are published with this paper for full reproducibility. The network is first tasked to optimize the visual-only loss specified in Eq. 4 combining SoftMax and Reciprocal Triplet Loss components with the mixing parameter $\lambda = 0.01$ [23] as described in Section 4.1 via SGD for 20 epochs. We train both a naive baseline version and one enhanced with LTR weight balancing [1] to separate the effect of LTR learning from genetic alignment. Our published source code [22] provides full details regarding all of the above and result reproducibility.

Cross-Modality Alignment. After these 20 epochs of visual-only training, we engage our genetic anchors ob-

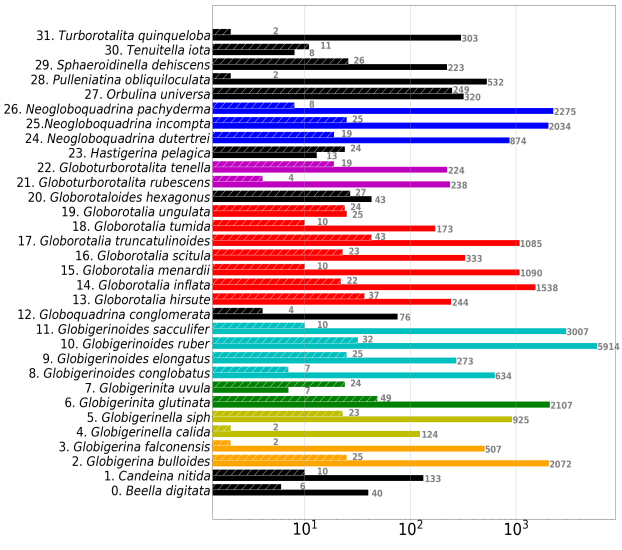

Figure 7. **Histogram of Available Samples Across Taxa.** A visualisation of the highly variable sample number associated to the 32 taxa used in the study (Dashed Upper Bars: rDNA, Filled Lower Bars: Visual). Groupings of taxa known as genera are indicated by colouration in line with Fig. 6.

tained through SGT and align the space across modalities using a Cosine embedding loss with triplet formulation as explained in Sec. 4.2 for another 5 epochs via SGD. We note that the choice of this loss function is critically important when integrating visual-genetic spaces. Neither using SGT embeddings as anchors with the standard Euclidean-based L_{RTL} loss nor utilizing extra trainable integration layers yielded any integrative success and cross-modality data transfer. Consequently, the use of the proposed orientational cosine alignment between visual embeddings and genetic anchors (see Fig. 5) proved a critical design choice and effective cost function for information transfer. Full implementation details are available via the source code repository published with this paper.

6. Results and Discussion

6.1. Quantitative Performance Benchmarks

Baselines. Table 1 quantifies performance for the system and across various baselines. We show performance of the original Endless Forams paper in *row 1* and the current SOTA work in *row 2* recalculating performance for our 32 covered taxa from the exact models built from all 35 taxa. For fair comparability we then retrain the SOTA model in *row 4* on only our 32 taxa covered (ie. where genetic data was available). For full ablation of effects we also stipulate performance for the exact model used, but without genetic alignment and without LTR techniques given as 'Naive' model in *row 3*. All baselines show low performance on rare tail classes compared to per-class or general accuracy.

Improving SOTA Performance via LTR. This work is the first to apply a SOTA LTR technique to the problem at hand. When adding weight balancing [1] in (row 5) to the 'Naive' Baseline 1, we were able to outperform all benchmarks and also beat the directly comparable SOTA Baseline 2 (row 4) across all measures without degradation of overall top accuracy. However, without using genetic information tail performance gains of LTR techniques are still limited.

Visual-Genetic Alignment. When adding visual-genetic alignment techniques to the 'Naive' Baseline 1 without LTR techniques (see *row 6*) an overall increase in per-class accuracy of 1.7% can be seen and for rare classes (N$<$100) per-class accuracy improved significantly by 8.2%. However, in order to improve performance for rare classes further we tested combining visual-genetic alignment with LTR-improved latent spaces.

Improving Per-Class and Rare Class Benchmarks. Combining the proposed visual-genetic alignment with LTR training (row 7) boosts tail performance significantly by another 11.5% to 59.7% and can further enhance per-class accuracy at only 1.6% overall accuracy trade-off. Given the practical importance of determining the presence of rare classes in paleoclimatology [9] this cost at the distribution's head is negligible in many important tasks. Moreover, the effect of genetically-driven improvements is particularly pronounced for the rarest classes where least image data is available. Fig. 9 plots per-class accuracy against cardinality of available class samples. This analysis confirms that a visual-genetic LTR model consistently yields superior performance to component ablations and that relative gains grow with sample rarity.

Visual-Genetic Alignment Control. In order to show that genetic information indeed has a domain-specific effect, we also test visual-genetic alignment with a ResNet50 model trained only on the ImageNet, that is one which has never seen images of foraminifers. We aligned this model as before to genetic anchors. As shown in Fig. 8*(a)*, such alignment still improves both overall accuracy (by 16.9%) and the per-class accuracy (by 15.5%) from the generic off-the-shelf ImageNet pre-training model. This confirms that visual-genetic alignment indeed transfers domain-specific information.

6.2. Qualitative and Class-Specific Discussion

Structure of Species Spaces. Based on these results, Fig. 6 explains and depicts a structural visualisation of the underlying latent spaces. It can be seen that superior per-class inference observed in the proposed visual-genetic dual space on the *(right)* is underpinned by more equidistant taxa spacing compared to the visual-only Baseline 1 model on the *(left)*. In addition, an improved clustering of genera (see colours) shows that aligned information from phenotype and genotype can indeed efficiently encode relationships between taxa in the metric structure of a latent space.

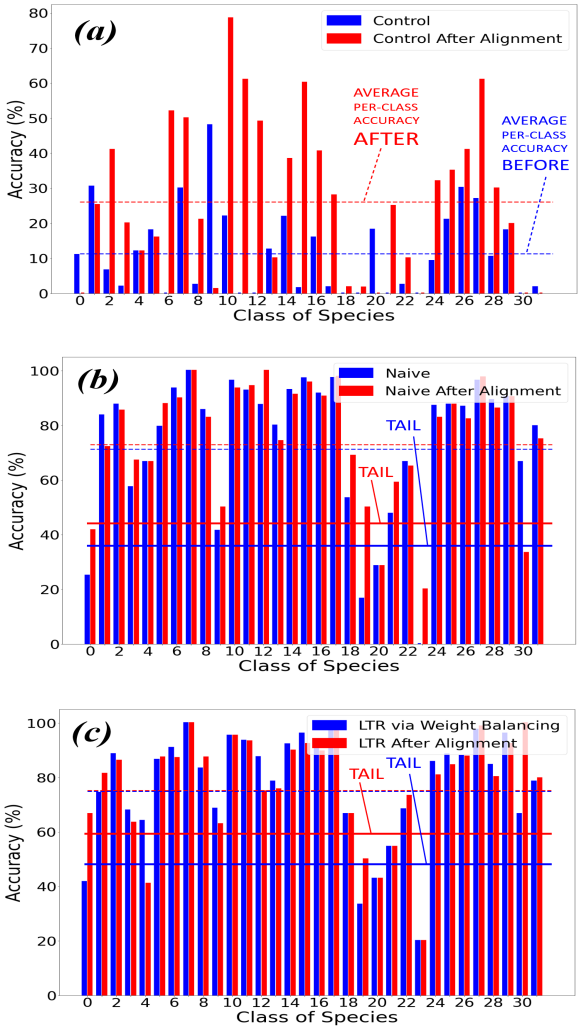

Figure 8. **Alignment Effect on Per-Class Accuracy.** Visual-genetic alignment consistently improves visual-only per-class test accuracy particularly for tail classes. ***(a)*** Alignment improvements for a out-of-domain control network initialised on ImageNet highlight effective information transfer from the genetic domain; ***(b)*** Alignment of the Baseline 1 'Naive' visual-only model shows significantly raised tail performance; ***(c)*** Alignment of an LTR visual-only model shows further per-class accuracy improvements with high gains on tail classes well beyond LTR-only performance.

Taxa-Specific Discussion. In order to quantify performance fine-grained on taxa level, Fig. 8 depicts a breakdown of the effect of transferring genetic information towards the visual domain for control, 'Naive' visual, and LTR enhanced models. For the rare tail classes 0,12,19,20,22,29,31 in particular accuracy increases after alignment for classes 0,19,22,29,31. For the remaining two rare classes 12 and 20 the genetic placement on the phytologentic tree of foraminifers (see Fig. 4) is in fact unclear [44] according to domain experts, thus casting doubt

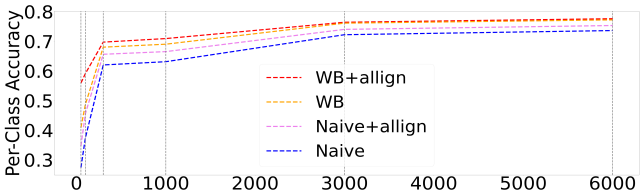

Figure 9. **Tail-Class Performance vs Sample Abundance.** Average per-class accuracy for classes with samples less than the abscissa value. The proposed combination of LTR weight balancing and visual-genetic alignment outperforms other ablations noting that gains improve with sample rarity.

over whether genetic information can at all be reliable for improving visual classification in these taxa. Overall, per-class accuracy *consistently* improves after visual-genetic alignment with highest impact on tail classes quantified in Fig. 9. Thus, the novel concept of proposed deep visual-genetic biometrics is demonstrably effective in the tested domain of taxonomic species classification.

7. Conclusion and Future Work

Visual-Genetic Biometrics for Rare Taxa. We presented visual-genetic biometrics, a novel domain for improving visual taxonomic classification in datasets with rare species via information transfer from the genetic domain. We provided a deep proof-of-concept implementation that leverages rDNA data to align imageomic and genetic information to create a multi-domain embedding space. Using 30k+ visuals across 32 taxa from the Endless Forams dataset we first demonstrated that traditional CNN application can be enhanced by LTR techniques to outperform the state-of-the-art on all benchmarks. We then showed that visual-genetic alignment can further improve per-class performance, particularly for rare classes. This establishes a new benchmark and confirms the effectiveness of visual-genetic biometrics by proof-of-concept. We note that the latent species space built with LTR techniques is particularly receptive to genetic information transfer.

Future Integration of Imageomics and Genetics. We believe that the implementation of artificial intelligence systems that organise life based on both phenotype and genotype will be important and significant with respect to reconciling genetic and imageomic spaces. This stems from a widely unexplored potential for a data-driven formal integration of the various approaches to the classification of life and for establishing tractable interfaces between the forms and levels life exhibits.

Acknowledgements

TK was supported by the UKRI CDT in Interactive Artificial Intelligence under the grant EP/S022937/1. DNS was supported by NERC grant NE/P019439/1. Thanks also to Allison Y. Hsiang and the Endless Forams team.

References

[1] Shaden Alshammari, Yuxiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. *CVPR*, 2022. 3, 6, 7

[2] William Andrew, Jing Gao, Siobhan Mullan, Neill Campbell, Andrew W. Dowsey, and Tilo Burghardt. Visual identification of individual Holstein-Friesian cattle via deep metric learning. *Computers and Electronics in Agriculture*, 185:106133, 2021. 2, 4

[3] T. Berger-Wolf and et al. Imageomics Institute, Ohio State University, 2022. Link: https://imageomics.osu.edu/about, last accessed: 12/04/2023. 1

[4] Otto Brookes, Majid Mirmehdi, Hjalmar Kühl, and Tilo Burghardt. Triple-stream deep metric learning of great ape behavioural actions. *18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, 5, 2023. 2

[5] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Journal of Artificial Intelligence Research*, 2019. 3

[6] Diego Darriba, David Posada, Alexey M Kozlov, Alexandros Stamatakis, Benoit Morel, and Tomas Flouri. Modeltest-ng: a new and scalable tool for the selection of dna and protein evolutionary models. *Molecular biology and evolution*, 37(1):291–294, 2020. 4

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[8] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *BMVC*, 2018. 3

[9] S Fenton, N Pearson, J Dunkley, and A. Purvis. Environmental predictors of diversity in recent planktonic foraminifera as recorded in marine sediments. *PLoS One*, 2016. 7

[10] A Frome, G S Corrado, J Shlens, and et al. Devise: a deep visual-semantic embedding model. *Advances in Neural Information Processing Systems*, 2013. 3

[11] Jing Gao, Tilo Burghardt, and Neill Campbell. Label a herd in minutes: Individual holstein-friesian cattle identification. *21st International Conference on Image Analysis and Processing Workshop (ICIAPW) on Learning in Precision Livestock Farming (LPLF)*, 2021. 2

[12] W Ge, W Huang, D Dong, and M.R. Scott. Deep metric learning with hierarchical triplet loss. *ECCV*, 2018. 2

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 3

[14] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2:1735–1742, 2006. 4

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[16] T. Hodan, P. Haluza, S. Obdrzalek, J. Matas, M. Lourakis, and X. Zabulis. T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. *Winter Conference on Applications of Computer Vision (WACV), IEEE*, 0:880–888, 2017. 4

[17] Allison Y. Hsiang et al. Endless Forams: > 34,000 modern planktonic foraminiferal images for taxonomic training and automated species recognition using convolutional neural networks. *Paleoceanography and Paleoclimatology*, 34:1157–1177, 2019. 2, 3, 4

[18] Wen-Yuan Wang Hui Han and Bing-Huan Mao. a new over-sampling method in imbalanced data sets learning. *International Conference on Intelligent Computing*, 2005. 3

[19] P. M. Hull. and A. Y. Hsiang. Endless Forams Most Beautiful, 2020. Link: http://endlessforams.org, last accessed: 25/01/2023. 1

[20] Tomihisa Kamada, Satoru Kawai, et al. An algorithm for drawing general undirected graphs. *Information processing letters*, 31(1):7–15, 1989. 6

[21] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *ICLR*, 2020. 3

[22] T. Karaderi. Deep visual genetic biometrics for taxonomic classification of rare species. https://github.com/TayfunKaraderi/WACV-2024---Deep-Visual-Genetic-Biometrics-for-Taxonomic-Classification-of-Rare-Species, year = 2024, . 6

[23] T Karaderi, T Burghardt, AY Hsiang, J Ramaer, and DN Schmid. Visual microfossil identification via deep metric learning. *3rd International Conference on Pattern Recognition and Artificial Intelligence (ICPRAI), Lecture Notes in Computer Science (LNCS)*, 13363, June 2022. 2, 4, 6

[24] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 3

[25] Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013. 4

[26] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *NIPS Deep Learning Workshop*, 2014. 3

[27] Alexey M Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, and Alexandros Stamatakis. Raxml-ng: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455, 2019. 4

[28] Hjalmar S Kühl and Tilo Burghardt. Animal biometrics: quantifying and detecting phenotypic appearance. *TREE*, 28(7):432–441, 2013. 1

[29] M. Lagunes-Fortiz, D. Damen, and W. Mayol-Cuevas. Learning discriminative embeddings for object recognition on-the-y. *2019 International Conference on Robotics and Automation (ICRA), IEEE*, 0:2932–2938, 2019. 4

[30] R. Marchant, M. Tetard, A. Pratiwi, M. Adebayo, and T. de Garidel-Thoron. Automated analysis of foraminifera fossil records by image classification using a convolutional neural network. *Journal of Micropalaeontology*, 39(2):183–202, 2020. 2

[31] A. Masullo, T. Burghardt, D. Damen, T. Perrett, and M. Mirmehdi. Who goes there? Exploiting silhouettes and wearable signals for subject identification in multi-person environments. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1599–1607, 2019. 4

[32] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR*, 2013. 3

[33] Raphaël Morard, Kate F Darling, Frédéric Mahé, Stéphane Audic, Yurika Ujiié, Agnes KM Weiner, Aurore André, Heidi A Seears, Christopher M Wade, Frédéric Quillévéré, et al. Pfr2: a curated database of planktonic foraminifera 18s ribosomal dna as a resource for studies of plankton ecology, biogeography and evolution. *Molecular Ecology Resources*, 15(6):1472–1485, 2015. 4

[34] Raphaël Morard, Angelina Füllberg, Geert-Jan A Brummer, Mattia Greco, Lukas Jonkers, André Wizemann, Agnes KM Weiner, Kate Darling, Michael Siccha, Ronan Ledevin, et al. Genetic and morphological divergence in the warm-water planktonic foraminifera genus globigerinoides. *PloS one*, 14(12):e0225246, 2019. 4

[35] Raphaël Morard, Nele M Vollmar, Mattia Greco, and Michal Kucera. Unassigned diversity of planktonic foraminifera from environmental sequencing revealed as known but neglected species. *PLoS One*, 14(3):e0213936, 2019. 4

[36] Lawrence O Hall Nitesh V Chawla, Kevin W Bowyer and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16, 2002. 3

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. https://pytorch.org/docs/stable/generated/torch.nn.CosineEmbeddingLoss.html, 2019. Accessed: 2023-04-17. 5

[38] Jan Pawlowski and Maria Holzmann. A plea for dna barcoding of foraminifera. *The Journal of Foraminiferal Research*, 44(1):62–67, 2014. 4

[39] Toby Perrett, Saptarshi Sinha, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Use your head: Improving long-tail video recognition, 2023. 1, 3

[40] Chitta Ranjan, Samaneh Ebrahimi, and Kamran Paynabar. Sequence graph transform (sgt): A feature embedding function for sequence data mining. *Data Mining and Knowledge Discovery*, 2016. 3

[41] Stefan Schneider, Graham W. Taylor, Stefan S. Linquist, and Stefan C. Kremer. Similarity learning networks for animal individual re-identification - beyond the capabilities of a human observer. *WACV*, abs/1902.09324, 2019. 2

[42] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A united embedding for face recognition and clustering. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2:815–823, 2015. 4

[43] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. 3

[44] Schwager. Globoquadrina conglomerata. http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm. Accessed: 2023-02-21. 8

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 3

[46] Maria Stennett, Daniel I Rubenstein, and Tilo Burghardt. Towards individual grevy's zebra identification via deep 3d fitting and metric learning. *IEEE/IAPR International Conference on Pattern Recognition (ICPR) Workshop on Visual Observation and Analysis of Vertebrate And Insect Behavior (VAIB)*, 2022. 1

[47] Maria Stennett, Daniel I Rubenstein, and Tilo Burghardt. Towards individual grevy's zebra identification via deep 3d fitting and metric learning. *26th IEEE/IAPR International Conference on Pattern Recognition (ICPR) Workshop on Visual Observation and Analysis of Vertebrate And Insect Behavior (VAIB)*, 2022. 2

[48] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):1–15, 2022. 1

[49] Yurika Ujiié and Yoshiyuki Ishitani. Evolution of a planktonic foraminifer during environmental changes in the tropical oceans. *PLoS One*, 11(2):e0148847, 2016. 4

[50] Alice Valentini, François Pompanon, and Pierre Taberlet. Dna barcoding for ecologists. *Trends in ecology & evolution*, 24(2):110–117, 2009. 1

[51] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 1, 3

[52] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *ICLR*, 2020. 3

[53] Xinyu Yang, Tilo Burghardt, and Majid Mirmehdi. Dynamic curriculum learning for great ape detection in the wild. *International Journal of Computer Vision*, pages 1–19, 2023. 1

[54] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. feature transfer learning for face recognition with under-represented data. *CVPR*, 2019. 3

[55] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 2021. 3

[56] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn. Bilateral-branch network with cumulative learning for long-tailed visual recognition. *CVPR*, 2020. 3