AnimalClue: Recognizing Animals by their Traces
Supplemental Material

A. Implementation Details

Classification. We train VGG-16 [7], ResNet-50 [2], ViT-B [5], and Swin-B [8] with input images resized to 224^2 pixels, while Eff-b1 [12] uses 240^2 pixels. Data augmentation includes random adjustments to brightness, contrast, saturation, hue, and compression rate, as well as vertical/horizontal flipping and rotation. For training, we use a batch size of 128 and the SGD optimizer with a learning rate of 1×10^{-4}. Models are trained for 100 epochs, except for the feather category, which is trained for 50 epochs. This setup ensures that training losses converge.

Detection. We train YOLOv8 [4], YOLOv11 [3], Faster-RCNN [9], DINO [15], and RT-DETR [16] with input images resized to 512^2 pixels. We train the models for 100 epochs for YOLOv8 and YOLOv11, and 50 epochs for DINO, Faster-RCNN, and RT-DETR. For the implementation, we follow YOLOv8 and YOLOv11, and RT-DETR for Ultralytics [13], Faster-RCNN for Detectron2 [14], and DINO for detrex [10].

For Faster R-CNN, we used the Adam optimizer with a learning rate of 1e-4. The ROI head batch size per image was set to 256, and the batch size per iteration was 4. For YOLOv8 and YOLOv11, we set the initial learning rate (lr0) to 0.01, with the final learning rate (lrf) defined as 0.01 × lr0. The batch size was 8. For DINO, we used the DINO-R50-4scale model, which incorporates a ResNet-50 backbone and extracts features from four different resolution levels to enhance multi-scale object detection. The learning rate was set to 1e-4, with a batch size of 16. For RT-DETR, we used the RT-DETR-L model with the AdamW optimizer and a learning rate of 0.001.

Instance Segmentation. We trained YOLOv8 [4], YOLOv11 [3] for 100 epochs, and Mask-RCNN [1], and MaskDINO [6] for 50 epochs with input images resized to 512^2 pixels.

For YOLOv8 and YOLOv11, we set the initial learning rate (lr0) to 0.01, with the final learning rate (lrf) defined as 0.01 × lr0. The batch size was 8. For Mask R-CNN, the batch size per iteration was 4 images, with a learning rate of 1e-4. Additionally, the batch size per image for the Region of Interest (ROI) heads was set to 256. For MaskDINO, we used the ResNet-50 backbone, setting the learning rate to 1e-4 and using a batch size of 16.

B. Removed Images

Here, we describe the types of images that are unsuitable for our AnimalClue and were manually removed.

- Images with overlaid text: Citizen scientists sometimes write labels directly on the images.
- Images containing animals: In many citizen scientists' posts, images include evidence of the animal itself to support the provided labels.
- Distant subjects: Some posts include images where the subject appears too far away.
- Images containing human faces: Although these images are licensed under Creative Commons, we exclude them from our dataset to protect privacy.

Figure 1 shows examples of removed images.

C. Species Distribution

In this study, we collect five types of traces:

- Footprints: 117 species, 46 families, and 20 orders
- Feces: 101 species, 46 families, and 21 orders
- Bones: 269 species, 112 families, and 45 orders
- Eggs: 283 species, 67 families, and 20 orders
- Feathers: 555 species, 89 families, and 30 orders

We visualize the species distribution categorized by order:

- Footprints: Figure 2
- Feces: Figure 3
- Bones: Figure 4, Figure 5
- Eggs: Figure 6, Figure 7
- Feathers: Figure 8, Figure 9, Figure 10

Additional trait details and taxonomy annotations will be made publicly available.

D. Additional Experiments

Other Metrics. For reference, we also report the Top-5 accuracy and balanced accuracy in Table 1. The model used is Swin-B with species-level classification.

CLIP Training. We conducted BioCLIP [11]-style training. We augmented the text prompts during CLIP training to

Figure 1. **Examples of removed images from our dataset.**

	Footprint	Feces	Egg	Bone	Feather
Top5	64.11	68.62	73.63	41.46	82.43
Balanced	28.92	35.41	32.25	15.54	26.55

Table 1. **Species classification results on the footprints, feces, eggs, bones, and feathers datasets (Top-5 accuracy and balanced accuracy)**. The model used is Swin-B.

include both a simple description and a taxonomic context. Specifically, we used the following format: ”A photo of a {species}.” and Ä photo of a {species}, which belongs to {genus}, {family}, {order}, {class}.” This training strategy led to a noticeable performance gain, improving the Top-1 accuracy from 17.6% to 21.7% on footprint species classification.

References

[1] Waleed Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*. IEEE Computer Society, 2016. 1

[3] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. 1

[4] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. 1

[5] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[6] Feng Li, Hao Zhang, Huaizhe xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation, 2022. 1

[7] Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 730–734, 2015. 1

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 2017. 1

[10] Tianhe Ren, Shilong Liu, Feng Li, Hao Zhang, Ailing Zeng, Jie Yang, Xingyu Liao, Ding Jia, Hongyang Li, He Cao, Jianan Wang, Zhaoyang Zeng, Xianbiao Qi, Yuhui Yuan, Jianwei Yang, and Lei Zhang. detrex: Benchmarking detection transformers, 2023. 1

[11] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP: A vision foundation model for the tree of life. In *Proceedings of*

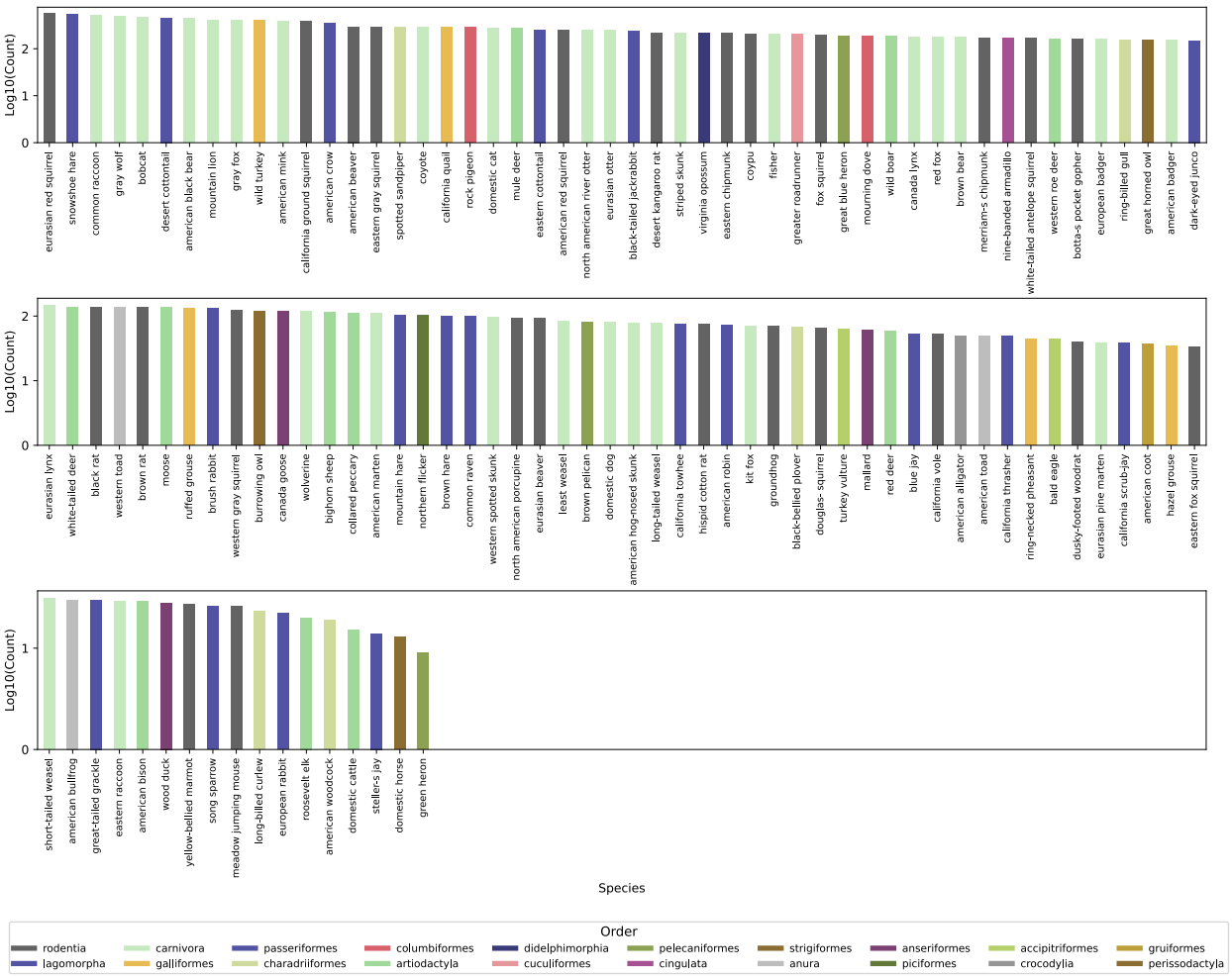

Figure 2. **Species distributions of footprints.**

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 19412–19424, 2024. 1

[12] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 1

[13] Ultralytics. Ultralytics, 2023. 1

[14] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1

[15] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 1

[16] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection, 2023. 1

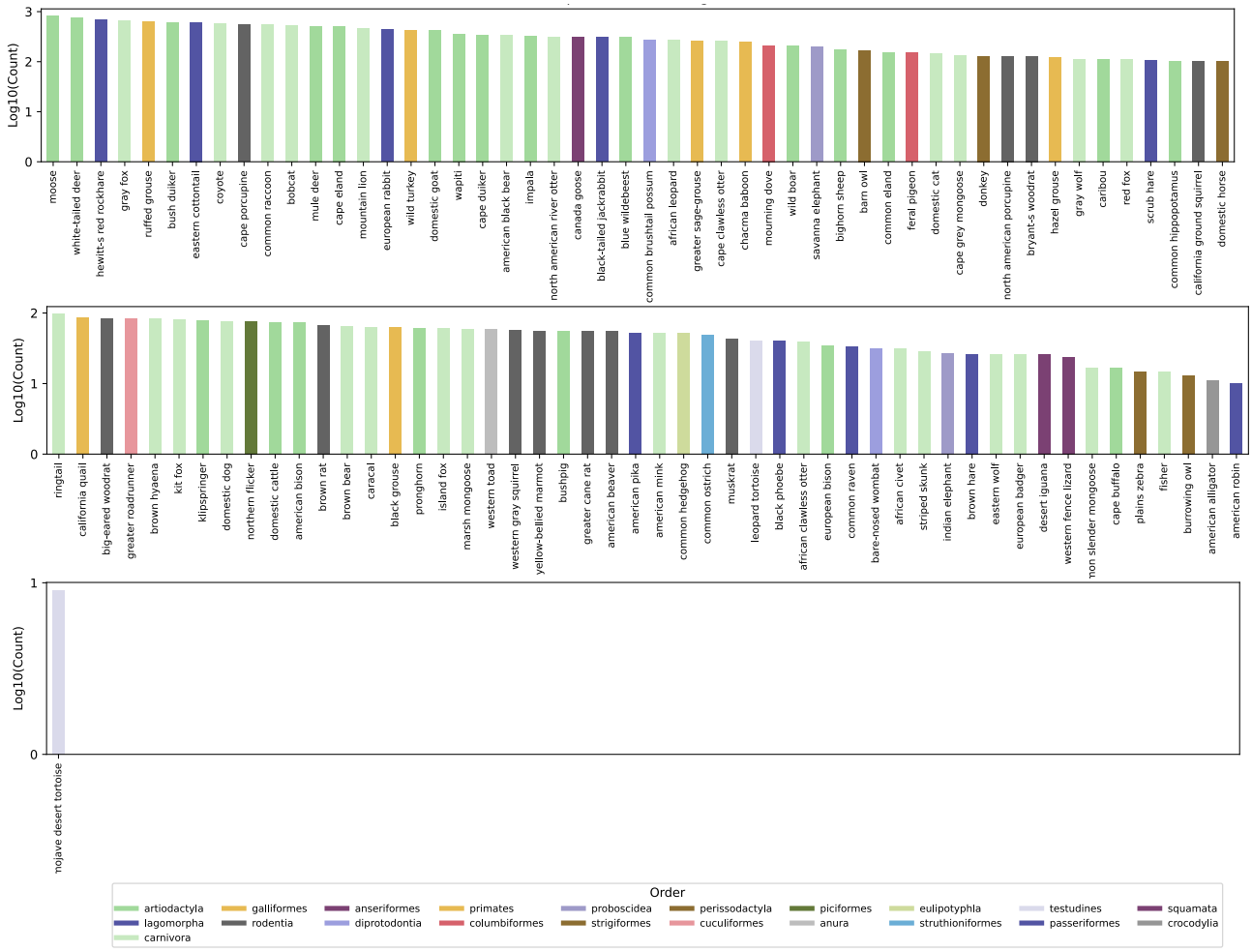

Figure 3. **Species distributions of feces.**

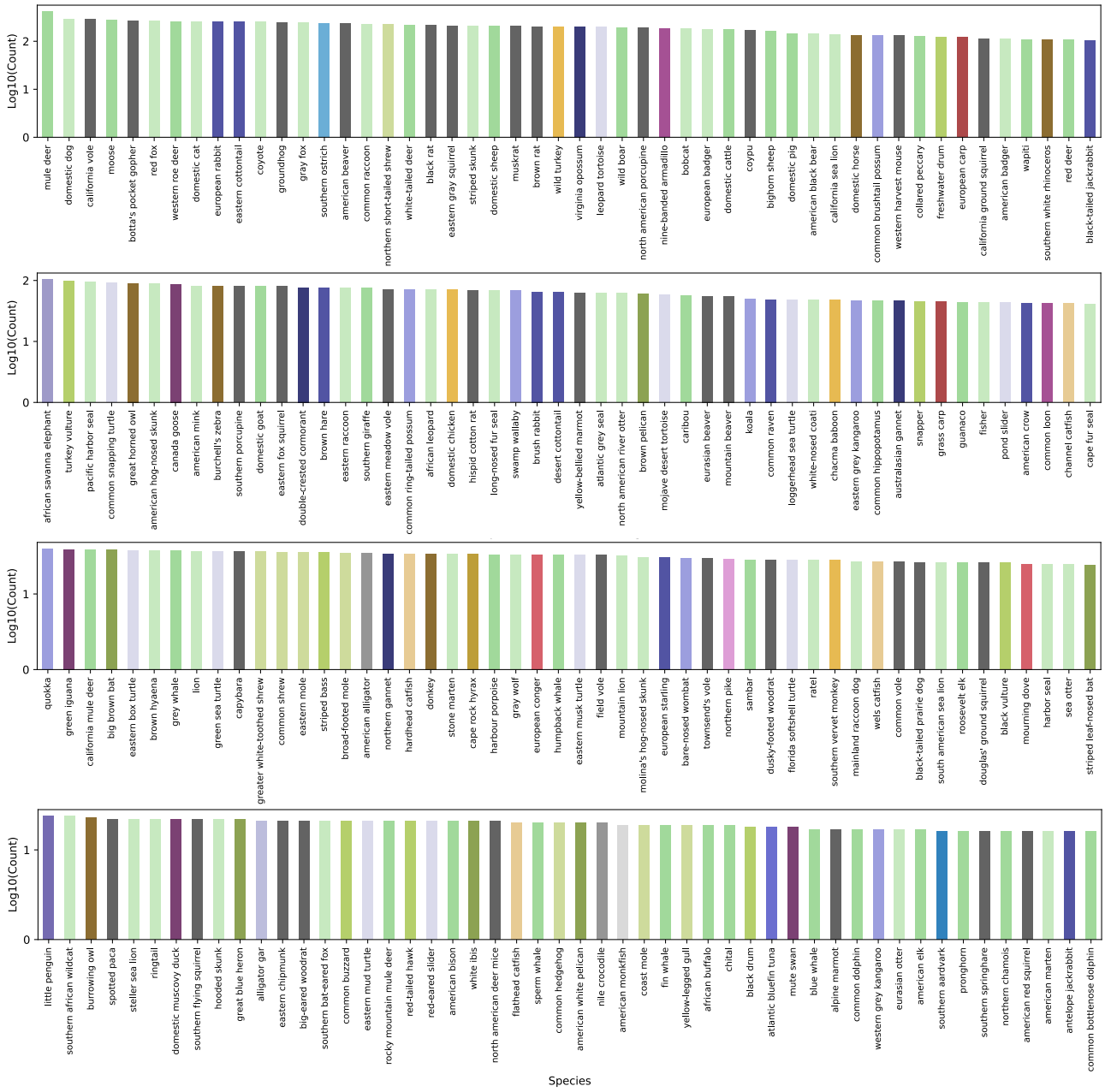

Figure 4. **Species distributions of bones (A).**

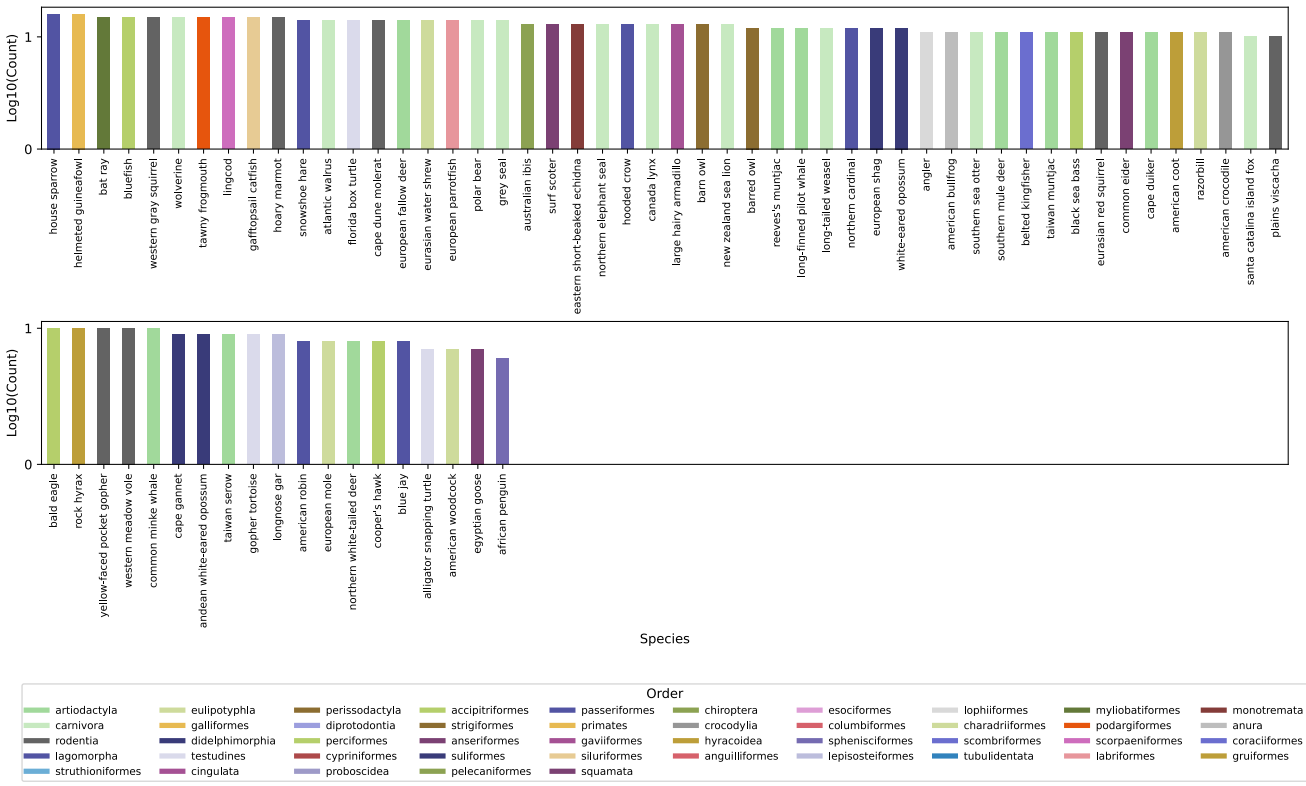

Figure 5. **Species distributions of bones (B).**

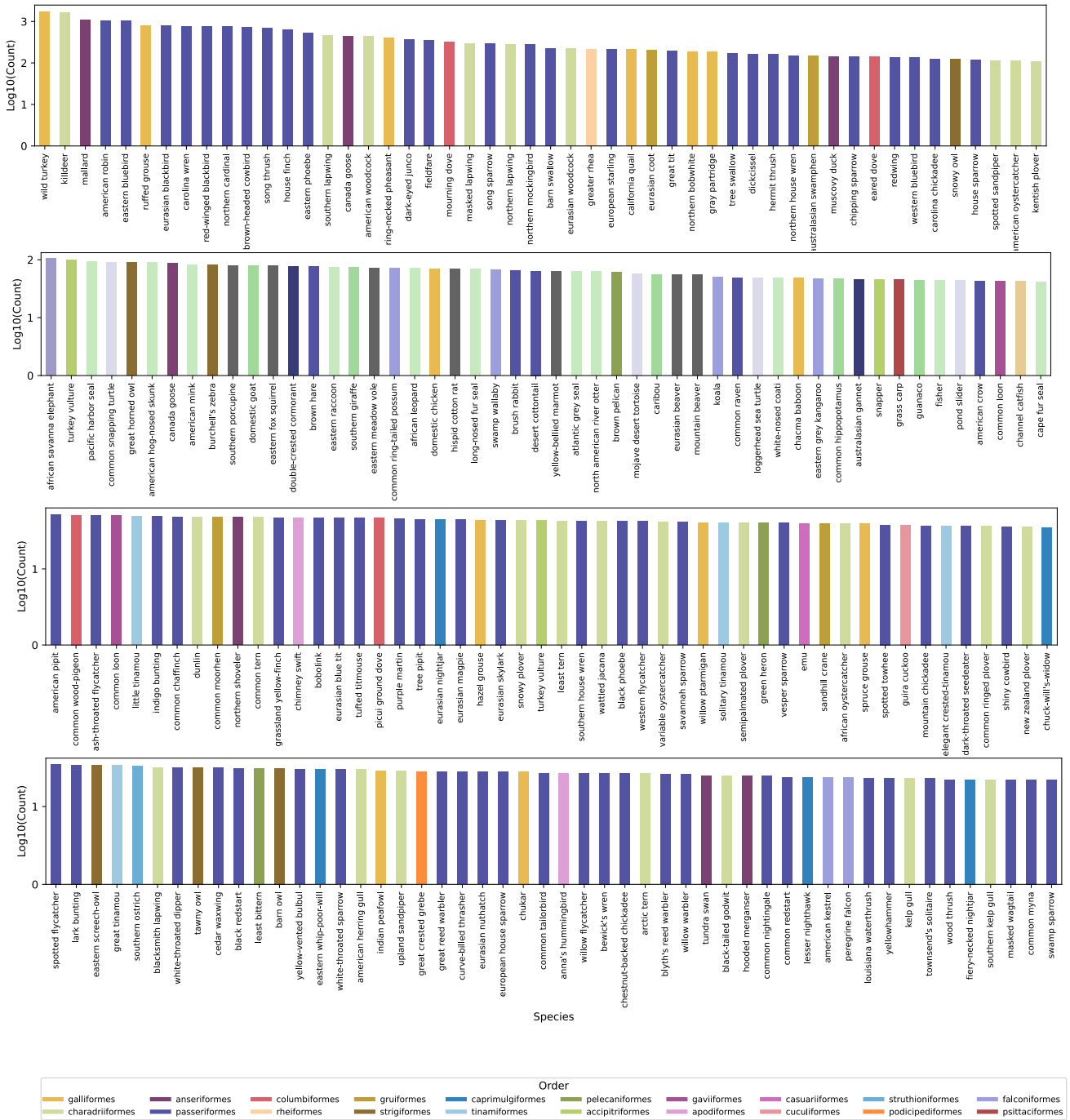

Figure 6. **Species distributions of eggs (A).**

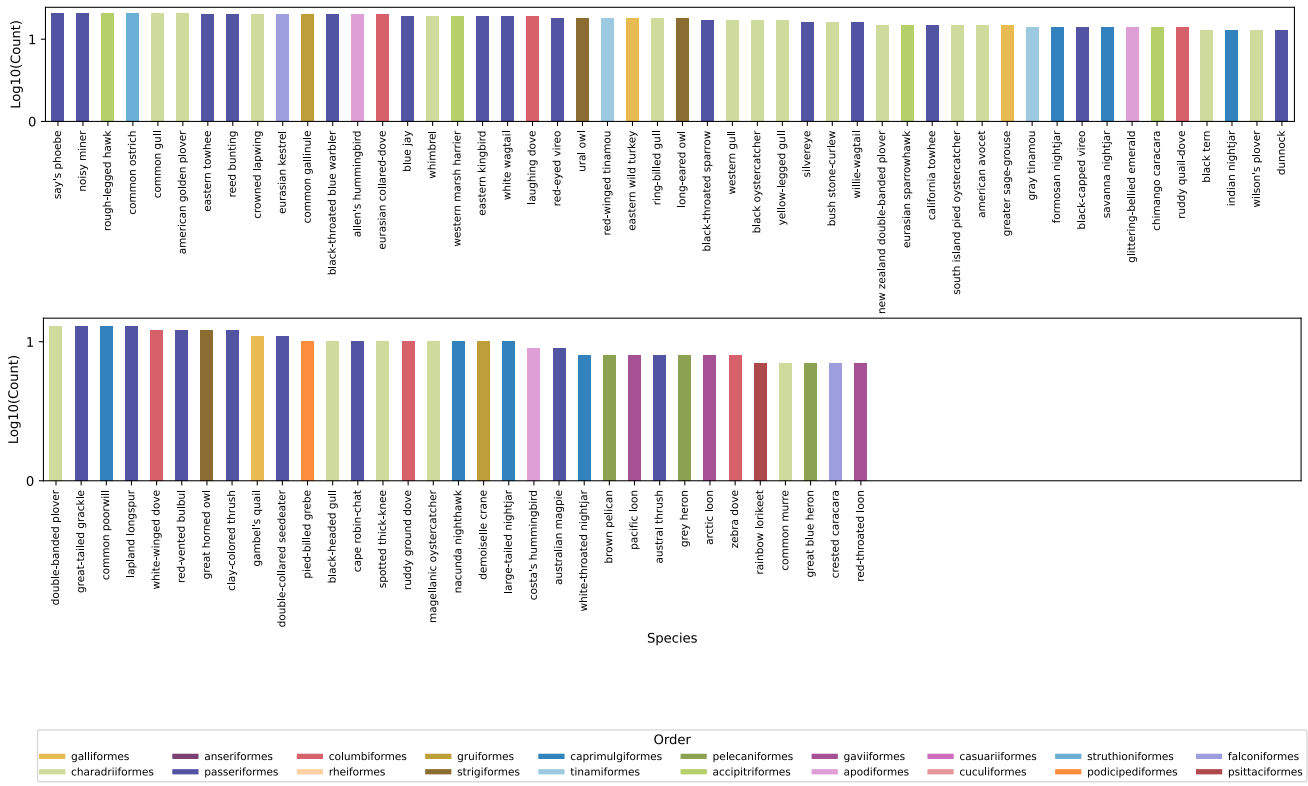

Figure 7. **Species distributions of eggs (B).**

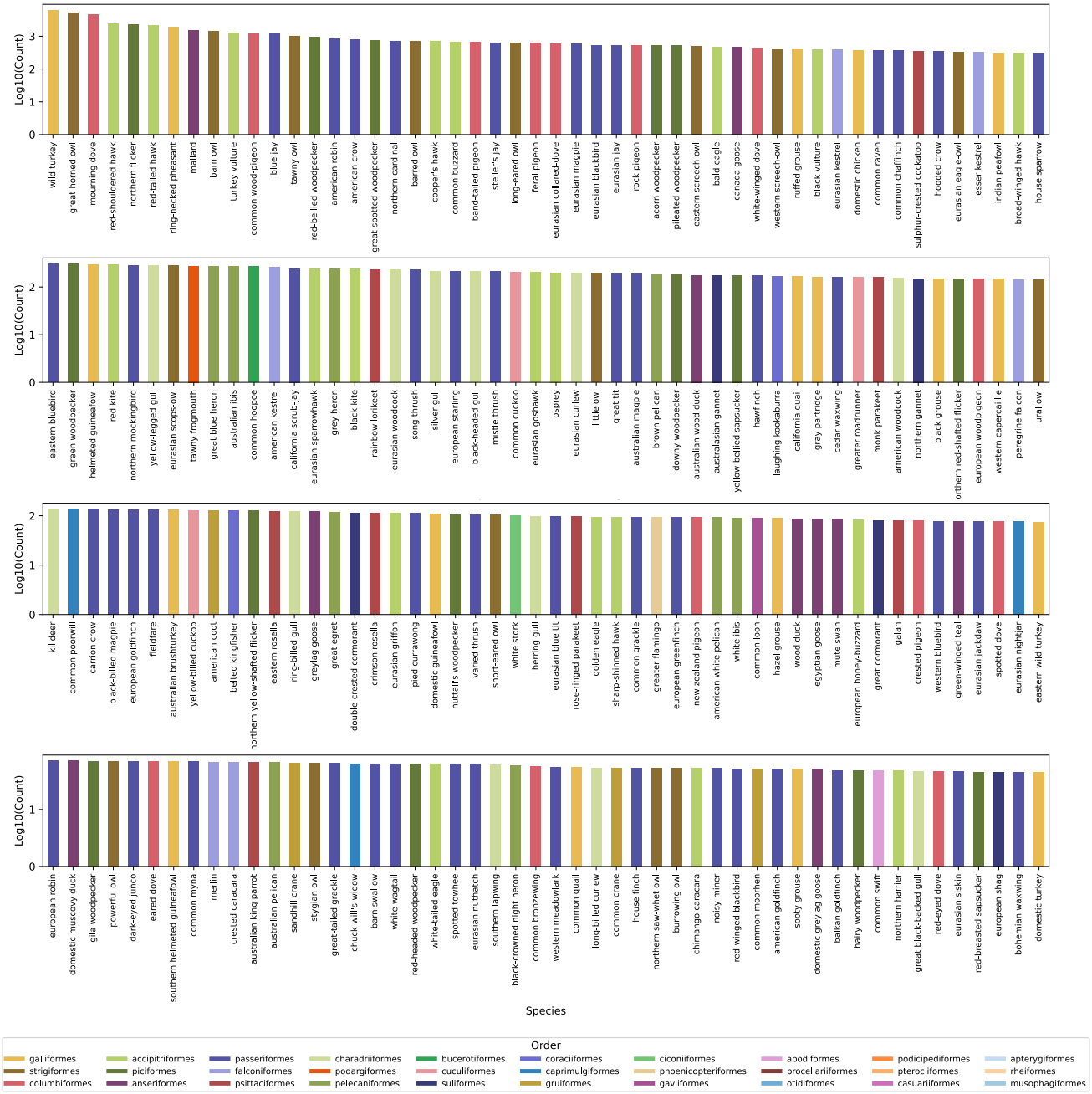

Figure 8. **Species distributions of feathers (A).**

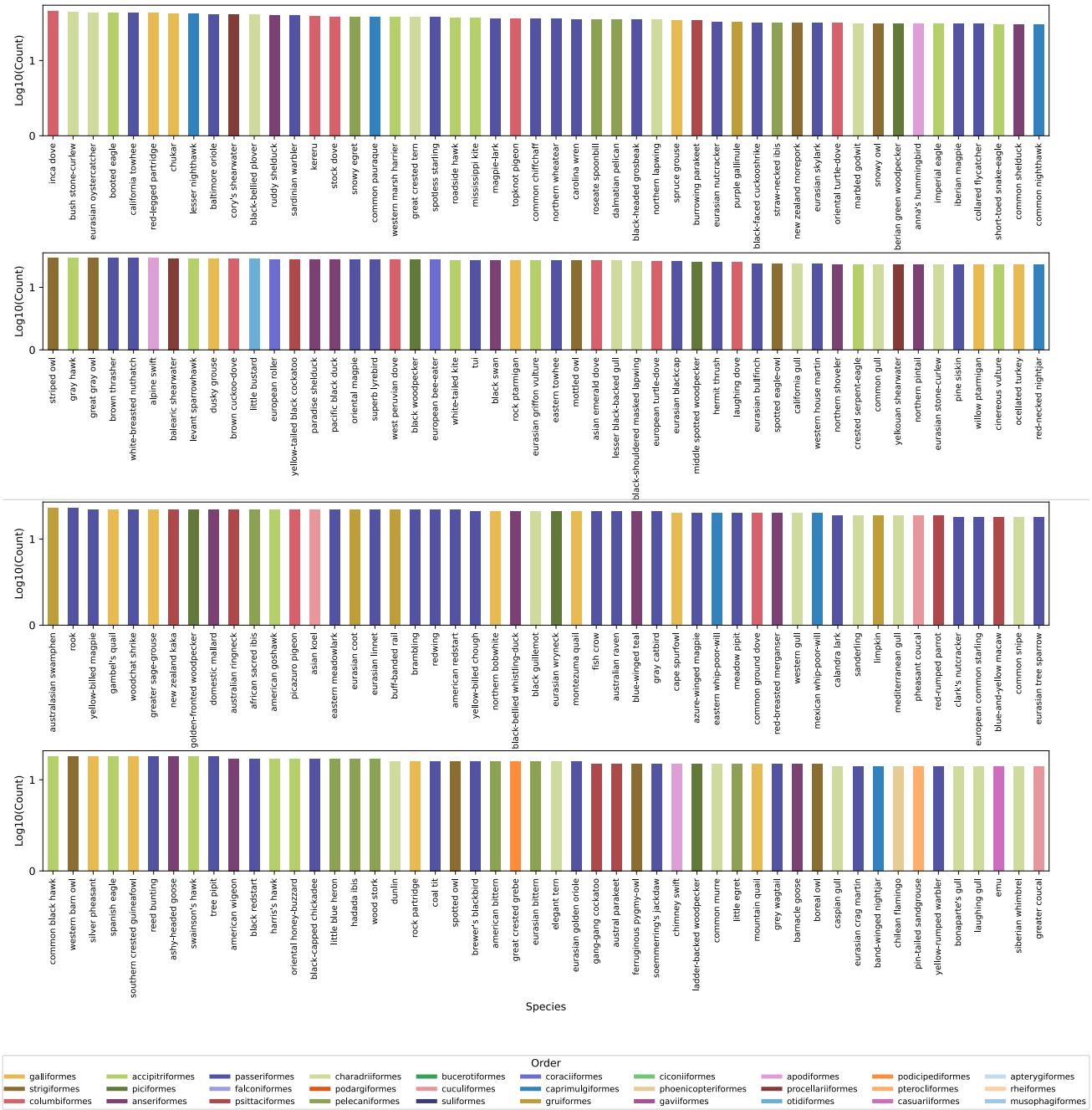

Figure 9. **Species distributions of feathers (B).**

Figure 10. **Species distributions of feathers (C).**