Modeling Semantic Correlation and Hierarchy for Real-World Wildlife Recognition

Dong-Jin Kim, *Member, IEEE*, Zhongqi Miao, *Member, IEEE*, Yunhui Guo, *Member, IEEE*, and Stella X. Yu, *Member, IEEE*

***Abstract*—We explore the challenges of human-in-the-loop frameworks to label wildlife recognition datasets with a neural network. In wildlife imagery, the main challenges for a model to assist human annotation are two-fold: (1) the training dataset is usually imbalanced, which makes the model's suggestion biased, and (2) there are complex taxonomies in the classes. We establish a simple and efficient baseline, including the debiasing loss function and the hyperbolic network architecture, to address these issues. Moreover, we propose leveraging the semantic correlation to train the model more effectively by adding a co-occurrence layer to our model during training. We demonstrate the efficacy of our method in both a real-world wildlife areal survey recognition dataset and the public image classification dataset, CIFAR100-LT, CIFAR10-LT, and iNaturalist.**

***Index Terms*—Wildlife recognition, active learning, class imbalance.**

I. Introduction and Related Work

Aerial remote sensing technologies [21] are being increasingly used to monitor and survey wildlife populations safely [30], [32], [34]. Previous wildlife aerial surveys require humans to manually count the visual objects on an aircraft which are risky [30], and the monitoring results can be biased for different human observers [6], [27]. In contrast to manual monitoring, aerial remote sensing provides the potential for consistent and reproducible population surveys, with the addition of an accurate geo-referenced digital format [32], [34]. However, as real-world data, such as remote sensing aerial survey data, should cover large areas resulting in thousands of images in the dataset, manually processing such real-world data is time-consuming and expensive for researchers and natural resource agencies [3], [12], [15]. As a result, we explore a human-in-the-loop approach [4], [11], [31] to utilize a deep neural network to collaborate with human annotators to efficiently process real-world wildlife datasets. In particular, our goal is to train a neural network as an image classifier to actively assist human annotators in labeling wildlife recognition datasets [23] by suggesting the class of unlabeled images.

When training a network to suggest image classes for real-world digital aerial imagery applications correctly, we observe two major distinctive points in multi-species datasets. **(1) Imbalanced data distribution.** Like most real-world datasets [33], the wildlife datasets [23] also show extremely imbalanced data distributions in the animal species [26]. Several dominant species are often observed, along with many infrequent species that are sparsely represented in datasets. As illustrated in Fig. 1, in the wildlife dataset from [23], the largest class had 6,246 training images, while the smallest class only had 17 training images. This issue could harm the recognition accuracy as the model prediction can be easily biased towards the abundant class [25]. The fewer training images of a particular species that the model has, the lower the accuracy for that species. **(2) Class hierarchy.** In addition, as illustrated in Fig. 1, the labels in wildlife datasets have a clear hierarchy which is relatively less prominent in standard datasets such as the CIFAR dataset. For example, mutually exclusive `Target` and `Non-target` categories have the highest hierarchy. The `Target` category consists of three second-level categories (super-classes): `Scoter`, `Common Eider`, and `Long-tailed Duck`. Finally, the `Scoter` super-class consists of three fine-grained classes such as `Black Scoter`, `White-winged Scoter`, and `Unknown Scoter`. As the categories in the class list have different levels of hierarchy, each category should be treated differently according to the corresponding hierarchy level. As a baseline to consider the aforementioned characteristics, we first utilize a simple yet effective debiasing loss [22] by leveraging the prior class distribution. In addition, to leverage the hierarchical nature of the class labels, we utilize hyperbolic neural networks [7] to learn the hierarchy in the labels effectively. To further boost the performance, we propose to learn the semantic correlation by modeling the co-occurrence among the classes. In particular, we add a learnable co-occurrence matrix on top of the model's final layer to refine the probability of the class prediction.

Manuscript received 25 November 2022; revised 1 March 2023; accepted 4 March 2023. Date of publication 16 March 2023; date of current version 24 March 2023. The work of Dong-Jin Kim was supported by the Institute of Information and communications Technology Planning and Evaluation funded by the Korea Government, MSIT, through Artificial Intelligence Graduate School Program, Hanyang University, under Grant 2020-0-01373. The work of Stella X. Yu was supported by the USGS and UC Berkeley Cooperative Research. A previous version of this work appeared in NeurIPS 2022 Workshop on Human in the Loop Learning, in close collaboration with Kyle Landolt and Travis Harrison at USGS and Mark Koneff at US Fish and Wildlife Service. The associate editor coordinating the review of this manuscript and approving it for publication was Moacir A. Ponti. *(Corresponding author: Stella X. Yu.)*

Dong-Jin Kim is with the Hanyang University, Seongdong-gu, Seoul 133-791, Korea (e-mail: djnjusa@gmail.com).

Zhongqi Miao is with the University of California, Berkeley, CA 94720 USA (e-mail: zhongqi.miao@berkeley.edu).

Yunhui Guo is with the University of Texas, Dallas, TX 78712 USA (e-mail: yunhui.guo@utdallas.edu).

Stella X. Yu is with the University of Michigan, Ann Arbor, MI 48109 USA, and also with the University of California, Berkeley, CA 94720 USA (e-mail: stellayu@umich.edu).

Digital Object Identifier 10.1109/LSP.2023.3257725

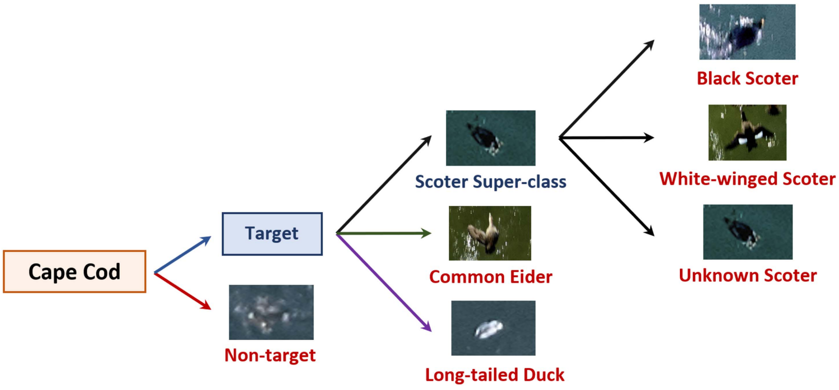

Fig. 1. The six classes in the Cape Cod dataset have a hierarchical relationship described in [23]. `Target` consists of three super-classes, `Scoter Super-class`, `Common Eider`, `Long-tailed Duck`. `Scoter Super-class` could be further divided to `Black Scoter`, `White-winged Scoter`, and `Unknown Scoter`. Moreover, the `Non-target Species` class includes images that cannot be categorized into the other five classes. Note that images between `Target` and `Non-target Species` are mutually exclusive in the dataset.

We use a case study of a real-world digital aerial survey dataset of waterbird species collected from the Atlantic Ocean near Cape Cod, Massachusetts, USA [23]. We show the efficacy of our method with both the wildlife dataset and the public image classification datasets, the CIFAR100-LT, CIFAR10-LT, and iNaturalist datasets [2], [5], [33], where our approach shows favorable performance compared to the baseline methods. In summary, our contributions are three-fold: (1) We explore a new problem to actively label a wildlife recognition dataset, (2) as a model to actively assist the human annotation process, we establish a simple yet effective baseline model to address the class imbalance problem and exploit label hierarchy, and (3) we propose a method to model the semantic correlation in the class labels by adding a co-occurrence layer, which further improves the performance.

II. Proposed Method

Given an input image X and the label Y (among L number of classes), our goal is to learn a classifier $f(\cdot) = \{f_j(\cdot)\}_{j=1}^{L}$ that minimizes the classification error. As mentioned in Section I, wildlife datasets are imbalanced in categories and have a label hierarchy. Therefore, we apply several methods to address these two challenges, (1) debiasing loss, (2) hyperbolic model, and (3) semantic correlation-based learning.

A. Debiasing Loss Function

We leverage one of the popular long-tailed image classification methods named logit adjustment [22] for imbalanced data distribution. Logit adjustment encourages a large relative margin between logits of rare positive versus dominant negative labels to balance learning. In particular, we apply the logit adjustment [22] as a debiasing loss function during training:

$$\mathcal{L}_{LA} = -\log \frac{\exp(f_y(X) + \tau \log \pi_y)}{\sum_{j\in[L]} \exp(f_j(X) + \tau \log \pi_j)}, \tag{1}$$

where $\pi \in \Delta_L$ (for simplex Δ) are estimates of the class priors *e.g.*, the empirical frequencies on the training sample. τ is the hyper-parameter. Other concurrent debiasing loss functions, such as margin-based approaches (LDAM [2]), uniformly increase the margin between a rare positive and *all negatives* and have relatively poor generalizability in most scenarios where the negative classes have heavily biased distribution. In contrast, logit adjustment loss increases the margin between a rare positive and a *dominant* negative in the output probability, which is more suitable for realistic datasets like wildlife datasets [23].

B. Hyperbolic Models

To effectively learn the hierarchy in the class labels, we add a hyperbolic module [1], [7], [18], [24], [28], [29] in our classifier to create a hyperbolic neural network (HNN) that utilizes the hyperbolic space to embed data with hierarchical structures. In particular, hyperbolic neural networks lift Euclidean features into hyperbolic space for classification. [28] and [29] especially pointed out that hyperbolic space is non-Euclidean space with constant negative curvature, which can be used for embedding tree structures owing to the nature of exponential growth in volume with respect to its radius. Hyperbolic neural networks have shown to be helpful, especially on datasets with known semantic hierarchies (e.g., analyzing single-cell data [18], learning hierarchical word embedding [24], embedding complex networks [1]). However, Hyperbolic neural networks mostly perform worse than Euclidean neural networks on standard benchmarks without clear hierarchies [7]. Therefore, we leverage the clipped hyperbolic classifier [7] method, which is generally effective on standard image classification benchmarks. This method clips the Euclidean feature magnitude in the hybrid architecture connecting Euclidean features to a hyperbolic classifier while training HNNs:

$$\text{CLIP}\left(x^E; r\right) = \min\left\{1, \frac{r}{||x^E||}\right\} \cdot x^E, \tag{2}$$

where x^E is the latent feature in the Euclidean space and r is the clipping value. Hyperbolic features capture the hierarchical structure of the dataset and improve the model's image classification performance.

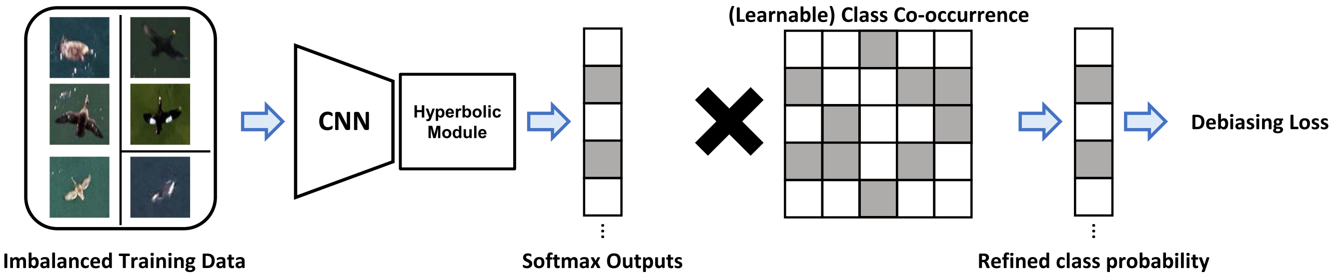

Fig. 2. Our training framework includes a hyperbolic module, debiasing loss function, and our co-occurrence layer to leverage the label correlations. Each box in the matrix refers to the probabilities of the matrix. We color the low and high probability values with white and grey colors, respectively.

C. Learning With Semantic Correlations

Moreover, we propose taking advantage of the natural *correlations* in the bird species. In particular, if a model gives a high likelihood of the black scoter (non-rare class), the model should give a similar likelihood of the white-winged scoter (rare class) as well. In addition, the detection of the Non-target species should preclude other bird classes. In particular, we define a learnable co-occurrence matrix [16], [17] in the form of $L \times L$ followed by a softmax nonlinearity, where L is the number of classes. Each entry of the matrix represents the conditional probability $p(Y = i|Y' = j)$ that the correct class of the given image is $Y = j$ when a primary model prediction Y' is j. By adding the co-occurrence matrix on top of the model's final layer as a *co-occurrence layer*, the probability of the refined class probability is predicted according to the total probability law:

$$P(Y = i|X) = \sum_{j \in \mathcal{Y}} p\left(Y = i|Y' = j\right) * p\left(Y' = j|X\right). \quad (3)$$

This process is illustrated in Fig. 2. Since the co-occurrence layer is learned end-to-end via the back-propagation from the loss function to the model input, it works as prior knowledge [13], [20], [36] that can be automatically learned from the target dataset. In contrast to language-based prior knowledge, which requires external data sources [13], [20], [36], our co-occurrence prior can be automatically learned.

III. Experiments

A. Dataset and Details

The main dataset we use is an aerial imagery dataset collected by the U.S. Fish and Wildlife Service at Nantucket Shoals (Cape Cod), Massachusetts, in February 2017. After data collection, wildlife experts manually cropped images of individual birds (*i.e.*, targets) and annotated the images into six different classes, illustrated in Fig. 1: The number of samples for each class is shown in Table I.

The images in the dataset were collected by the U.S. Fish and Wildlife Service at Nantucket Shoals (Cape Cod), Massachusetts, in February 2017. Pixel resolution for the Nantucket Shoals dataset ranged from 0.18 to 1.47 cm. The average image dimension of the dataset is 75 × 79. In particular, the average image dimension of `Unknown Scoter` is 56 × 61, while the average image dimensions of `Black Scoter` and `White-winged Scoter` are 100 × 107 and 96 × 103, respectively. `Unknown Scoter` can be considered a coarse annotation of `Black Scoter` and `White-winged Scoter` because it contains images of either one of the two scoter classes but without species-level annotations.

TABLE I
The Number of Samples in the Wildlife Dataset [23] for Training and Testing Split

Species	Train #	Test #
`Common Eider`	6,246	3,172
`Unknown Scoter`	466	114
`Black Scoter`	341	108
`White-winged Scoter`	45	21
`Long-tailed Duck`	17	5
`Non-target Species`	108	38

The largest class, common eider, has over 6,246 training images, while the smallest class, long-tailed duck, only has 17 training images (imbalance ratio is 367:1). In other words, the imbalance ratio of the dataset is 367:1.

To construct the dataset, the images of individual birds are manually cropped, and every cropped image is annotated by human experts. The Cape Cod dataset consists of a total of 10,682 cropped images. There are six different classes (Table I), one of which is an unknown class only for scoters (`Unknown Scoter`) and a general unknown class (`Non-target Species`) for the instance species that cannot be specified among the classes described in Table I.

We use ResNet50 [8] as our backbone architecture which takes a 224 × 224 image as input. We train the model with a batch size of 64 and a learning rate of 0.001. It takes 90 epochs until convergence. For the hyperbolic module, we used the Poincaré ball model for hyperbolic space, and we set curvature and clipping values as 1 and 15, respectively.

B. Results

Table II shows the quantitative results on the real-world wildlife dataset [23]. The baseline (Vanilla) method shows relatively poor performance, especially for species with limited training samples such as `Long-tailed Duck` (17 training images) and `White-winged Scoter` (45), which have 0.0% and 38.1% test accuracy, respectively. On the other hand,

TABLE II
QUANTITATIVE RESULTS ON THE WILDLIFE DATASET [23]

	Test accuracy (%)					
Species	**Vanilla**	**+LDAM**	**+LA**	**+LA+H**	**+LA+Corr**	**+LA+H+Corr (Ours)**
`Common Eider`	**99.1**	93.8	92.0	86.8	91.5	91.8
`Black Scoter`	95.4	92.6	97.2	87.0	97.2	**98.0**
`White-winged Scoter`	38.1	71.4	85.7	61.9	85.7	**85.9**
`Long-tailed Duck`	0.0	100.0	**100.0**	**100.0**	**100.0**	**100.0**
`Non-target Species`	39.5	68.4	78.9	**84.1**	81.6	81.9
Average accuracy (%)	54.1	85.3	90.8	82.9	91.2	**91.5**

Our final model with logit adjustment loss (+la), hyperbolic module (+h), and our semantic co-occurrence aware learning method (+corr) shows the best performance in most of the classes.

TABLE III
TOP-1 ACCURACY ON CIFAR10-LT AND CIFAR100-LT DATASET [2], [5] WITH THE IMBALANCE RATIO OF 100 ALONG WITH THE INATURALIST DATASET [33]

Methods	CIFAR100-LT	CIFAR10-LT	iNaturalist
Vanilla	38.3	70.4	57.9
LDAM [2]	42.0	77.0	68.0
LA [22]	43.9	77.7	66.4
Ours	**45.3**	**79.0**	**68.7**

Our final model consistently outperforms the baseline methods across different imbalanced datasets.

`Common Eider`, the class with the largest number of training samples (6,246), shows 99.1% test accuracy.

While the classification model with a traditional debiasing loss function (LDAM [2]) already improves the recognition performance over the vanilla model, our model with the logit adjustment loss function (LA [22]) shows even more performance improvement in all the classes (+31.2% average classification accuracy improvement from LDAM and +36.7% from LA). The largest gain comes from the two tail classes, `Long-tailed Duck` and `White-winged Scoter`, from 0.0% to 100.0% and 38.1% to 85.7%, respectively. Despite the improvements in the less abundant classes, the performance of `Common Eider` dropped by 7.1%, which is a common phenomenon of imbalanced methods where the performance of large classes is sacrificed [19], [25], [35].

Upon the logit adjustment loss, adding both the hyperbolic module or our co-occurrence-based learning method gives noticeable performance improvements in the `Non-target Species` class by better learning the correlation among the categories. However, the hyperbolic module shows a performance drop in most of the other classes. We conjecture that it is because the number of classes in the wildlife dataset is too small (which might have a relatively weak hierarchy among the labels), and the hyperbolic module requires a large number of parameters (which causes an overfitting problem). Note that combining both the hyperbolic module and our co-occurrence-based learning method leads to the best performance by compensating the weakness of each other.

We also evaluate our method on standard image classification datasets with class imbalance, CIFAR100-LT and CIFAR10-LT [2], [5] with the imbalance ratio of 100, in order to validate the efficacy of our method to alleviate imbalance problems. We also evaluate our method on a realistic imbalanced classification dataset, iNaturalist [33]. The result is shown in Table III. As shown in the table, our method shows the best performance compared to the recent powerful debiasing methods, LDAM [2] or logit adjustment [22]. This signifies that our method is also effective in alleviating class imbalance on popular benchmarks, even on a realistic setup.

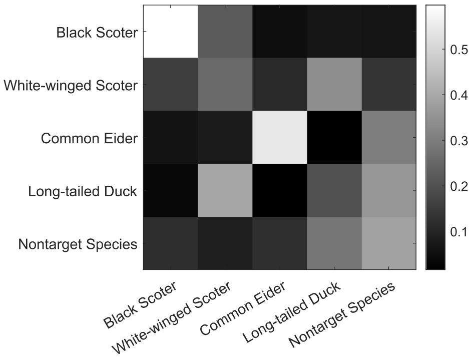

Fig. 3. The visualization of the learned co-occurrence matrix. Along the Y-axis is the given action, and the X-axis enumerates conditional actions.

Finally, we visualize the learned co-occurrence matrix from our method in Fig. 3. The correlation between `Black Scoter` and `White-winged Scoter` and the correlation between `Long-tailed Duck` and `Non-target Species` can be found in the co-occurrence matrix. An interesting observation is that there is a correlation between `White-winged Scoter` and the `Long-tailed Duck`, which is not intuitive.

IV. CONCLUSION

We explore how to utilize a network to collaborate with human annotators to process real-world wildlife datasets. We present solutions in terms of our debiasing loss function, a hyperbolic model, and our semantic correlation learning. We hope our challenges and solutions inspire future researchers in wildlife surveys. One of the possible future research directions would be to apply our method to more practical active labeling frameworks [4], [11]. Other possible future works include higher-level visual tasks, such as caption generation [12], [15] with wildlife datasets. Furthermore, our method can be extended to other domains that are vulnerable to data issues, such as human-object interaction detection [16], [17] and sign language recognition [9], [10].

REFERENCES

[1] G. Alanis-Lobato, P. Mier, and M. A. Andrade-Navarro, "Efficient embedding of complex networks to hyperbolic space via their Laplacian," *Sci. Rep.*, vol. 6, no. 1, pp. 1–10, 2016.

[2] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, “Learning imbalanced datasets with label-distribution-aware margin loss,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1567–1578.

[3] D. Chabot and C. M. Francis, “Computer-automated bird detection and counts in high-resolution aerial images: A review,” *J. Field Ornithol.*, vol. 87, no. 4, pp. 343–359, 2016.

[4] J. W. Cho, D.-J. Kim, Y. Jung, and I.S. Kweon, “MCDAL: Maximum classifier discrepancy for active learning,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 22, 2022, doi: 10.1109/TNNLS.2022.3152786.

[5] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, “Class-balanced loss based on effective number of samples,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9268–9277.

[6] P.C. Frederick, B. Hylton, J. A. Heath, and M. Ruane, “Accuracy and variation in estimates of large numbers of birds by individual observers using an aerial survey simulator,” *J. Field Ornithol.*, vol. 74, no. 3, pp. 281–287, 2003.

[7] Y. Guo, X. Wang, Y. Chen, and S. X. Yu, “Clipped hyperbolic classifiers are super-hyperbolic classifiers,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11–20.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[9] Y. Jang, Y. Oh, J. W. Cho, D.-J. Kim, J. S. Chung, and I. S. Kweon, “Signing outside the studio: Benchmarking background robustness for continuous sign language recognition,” in *Proc. Brit. Mach. Vis. Conf.*, 2022.

[10] Y. Jang,, “Self-sufficient framework for continuous sign language recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023.

[11] D.-J. Kim, J. W. Cho, J. Choi, Y. Jung, and I. S. Kweon, “Single-modal entropy based active learning for visual question answering,” in *Proc. Brit. Mach. Vis. Conf.*, 2021.

[12] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, “Dense relational captioning: Triple-stream networks for relationship-based captioning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6271–6280.

[13] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, “Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 2012–2023.

[14] D.-J. Kim et al., “Modeling semantic correlation and hierarchy for real-world wildlife recognition,” in *Proc. NeurIPS Workshop Hum. Loop Learn.*, 2022.

[15] D.-J. Kim, T.-H. Oh, J. Choi, and I. S. Kweon, “Dense relational image captioning via multi-task triple-stream networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7348–7362, Nov. 2022.

[16] D.-J. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon, “Detecting human-object interactions with action co-occurrence priors,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 718–736.

[17] D.-J. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon, “ACP++: Action co-occurrence priors for human-object interaction detection,” *IEEE Trans. Image Process.*, vol. 30, pp. 9150–9163, 2021.

[18] A. Klimovskaia, D. Lopez-Paz, L. Bottou, and M. Nickel, “Poincaré maps for analyzing complex hierarchies in single-cell data,” *Nature Commun.*, vol. 11, no. 1, 2020, Art. no. 2966.

[19] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, “Large-scale long-tailed recognition in an open world,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2537–2546.

[20] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, “Visual relationship detection with language priors,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 852–869.

[21] J. F. McEvoy, G. P. Hall, and P. G. McDonald, “Evaluation of unmanned aerial vehicle shape, flight path and camera type for waterfowl surveys: Disturbance effects and species recognition,” *PeerJ*, vol. 4, 2016, Art. no. e1831.

[22] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, “Long-tail learning via logit adjustment,” in *Proc. Int. Conf. Learn. Representations*, 2021.

[23] Z. Miao et al., “Challenges and solutions for automated avian recognition in aerial imagery,” *Remote Sens. Ecol. Conservation*, 2022.

[24] M. Nickel and D. Kiela, “Poincaré embeddings for learning hierarchical representations,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6341–6350.

[25] Y. Oh, D.-J. Kim, and I. S. Kweon, “DASO: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9786–9796.

[26] L. Stuart et al., “The biodiversity of species and their rates of extinction, distribution, and protection,” *Science*, vol. 344, 2014, Art. no. 6187.

[27] J.V. Redfern, P.C. Viljoen, J.M. Kruger, and W.M. Getz, “Biases in estimating population size from an aerial census: A case study in the kruger national park, South Africa: Starfield festschrift,” *South Afr. J. Sci.*, vol. 98, no. 9, pp. 455–461, 2002.

[28] F. Sala, C. De Sa, A. Gu, and C. Ré, “Representation tradeoffs for hyperbolic embeddings,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4460–4469.

[29] R. Sarkar, “Low distortion delaunay embedding of trees in hyperbolic plane,” in *Proc. Graph Drawing: 19th Int. Symp.*, 2011 pp. 355–366.

[30] D. B. Sasse, “Job-related mortality of wildlife workers in the United States, 1937-2000,” *Wildlife Soc. Bull.*, vol. 31, pp. 1015–1020, 2003.

[31] I. Shin, D.-J. Kim, J. W. Cho, S. Woo, K. Y. Park, and I. S. Kweon, “Labor: Labeling only if required for domain adaptive semantic segmentation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 8588–8598.

[32] D. Tuia et al., “Perspectives in machine learning for wildlife conservation,” *Nature Commun.*, vol. 13, no. 1, 2022, Art. no. 792.

[33] G.V. Horn et al., “The inaturalist species classification and detection dataset,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8769–8778.

[34] D. Wang, Q. Shao, and H. Yue, “Surveying wild animals from satellites, manned aircraft and unmanned aerial systems (UASs): A review,” *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1308.

[35] X. Wang, L. Lian, Z. Miao, Z. Liu, and S. X. Yu, “Long-tailed recognition by routing diverse distribution-aware experts,” in *Proc. Int. Conf. Learn. Representations*, 2021.

[36] R. Yu, A. Li, V. I. Morariu, and L. S. Davis, “Visual relationship detection with internal and external linguistic knowledge distillation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1974–1982.