


INTERDISCIPLINARY PERSPECTIVE

Challenges and solutions for automated avian recognition in aerial imagery

Zhongqi Miao^{1,2} , Stella X. Yu², Kyle L. Landolt³, Mark D. Koneff⁴, Timothy P. White⁵, Luke J. Fara³, Enrika J. Hlavacek³, Bradley A. Pickens⁶, Travis J. Harrison³ & Wayne M. Getz^{1,7}

¹Department of Environmental Science, Policy, and Management, UC Berkeley, Berkeley, California, USA

²International Computer Science Institute, UC Berkeley, Berkeley, California, USA

³U.S. Geological Survey, Upper Midwest Environmental Sciences Center, La Crosse, Wisconsin, USA

⁴Division of Migratory Bird Management, U.S. Fish & Wildlife Service, Orono, Maine, USA

⁵Environmental Studies Program, Bureau of Ocean Energy Management, Sterling, Virginia, USA

⁶Division of Migratory Bird Management, U.S. Fish & Wildlife Service, Laurel, Maryland, USA

⁷School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

Keywords

Avian conservation, avian image recognition, deep learning, remote sensing

Correspondence

Stella X. Yu, International Computer Science Institute, UC Berkeley, Berkeley, CA, USA.
Tel: +1 734-647-1761.

E-mail: stellayu@berkeley.edu;
stellayu@umich.edu

Mark D. Koneff, Division of Migratory Bird Management, U.S. Fish & Wildlife Service, Orono, ME, USA. Tel: +1 301-980-0125.
E-mail: mark_koneff@fws.gov

Editor: Temuulen Sankey

Associate Editor: Abdulhakim Abdi

Received: 24 June 2022; Revised: 6 October 2022; Accepted: 7 November 2022

Abstract

Remote aerial sensing provides a non-invasive, large geographical-scale technology for avian monitoring, but the manual processing of images limits its development and applications. Artificial Intelligence (AI) methods can be used to mitigate this manual image processing requirement. The implementation of AI methods, however, has several challenges: (1) imbalanced (i.e., long-tailed) data distribution, (2) annotation uncertainty in categorization, and (3) dataset discrepancies across different study sites. Here we use aerial imagery data of waterbirds around Cape Cod and Lake Michigan in the United States to examine how these challenges limit avian recognition performance. We review existing solutions and demonstrate as use cases how methods like Label Distribution Aware Marginal Loss with Deferred Re-Weighting, hierarchical classification, and FixMatch address the three challenges. We also present a new approach to tackle the annotation uncertainty challenge using a Soft-fine Pseudo-Label methodology. Finally, we aim with this paper to increase awareness in the ecological remote sensing community of these challenges and bridge the gap between ecological applications and state-of-the-art computer science, thereby opening new doors to future research.

doi: 10.1002/rse2.318

Introduction

Aerial remote sensing technologies are being increasingly used to monitor and survey wildlife populations (Tuia et al., 2022; Wang et al., 2019). They provide non-invasive tools for detecting, classifying, and assessing the abundance of target species (McEvoy et al., 2016). Traditional wildlife aerial surveys employ human observers to conduct visual counts, often from low-flying aircraft. Although these methods can be efficient in surveying large geographic regions, visual observations from low-flying aircraft are risky (e.g., the risk that observation personnel may encounter a life-threatening scenario increases at low-flying altitudes) (Sasse, 2003) and are subject to various observer

biases such as count bias (Frederick et al., 2003; Redfern et al., 2002). In contrast, aerial remote sensing is a safer alternative that allows flying at a higher altitude, and the method offers the potential for a consistent and reproducible population survey with the addition of an accurate geo-referenced digital format. In addition, aerial imagery surveys conducted at higher altitudes may also reduce animal disturbance (Sasse, 2003). The major disadvantage of using remote sensing for aerial surveys is that covering large areas can generate hundreds of thousands of images, thus hundreds of terabytes of data. Therefore, manually processing remote sensing aerial survey data is time-consuming and prohibitively expensive for many researchers and natural resource agencies (Chabot & Francis, 2016).

Ecologists are increasingly looking to cutting-edge artificial intelligence (AI) methodology, such as deep learning and computer vision technologies, to mitigate the need for labor-intensive processing of digital aerial imagery and to improve monitoring efficiency. For example, deep learning has been applied to aid in digital aerial surveys conducted with various sensor systems, including RGB (i.e., images coded in red green blue colors) (Hong et al., 2019; Liu et al., 2018), thermal (Corcoran et al., 2019), and other sensor systems (Wang et al., 2019). However, several challenges need to be addressed regarding recognition performance for real-world digital aerial imagery applications of AI methods. These include (1) the imbalanced distribution challenge – extremely imbalanced data distributions that generally lead to poor recognition performance; (2) the annotation uncertainty in categorization challenge – uncertainty in annotation caused by various reasons such as varying image resolutions of avian individuals; and (3) the dataset discrepancy challenge – images collected from different study sites (i.e., geographies) that have different characteristics and classes.

To examine these challenges in detail, we use a case study of two real-world digital aerial survey datasets of waterbird species: one collected from the Atlantic Ocean near Cape Cod, Massachusetts, and the other from Lake Michigan near Manitowoc, Wisconsin, USA. We also present solutions, accompanied by brief literature reviews for each challenge, focusing on how the computer science community has previously addressed these types of challenges. We aim to increase awareness of these challenges within the ecological community, clarify the factors affecting AI recognition performance, demonstrate the flexibility of deep learning methods, and promote future research in AI and digital aerial surveys.

Avian Recognition

Aerial images of birds may include a few or many individuals depending on resources being used by those birds and

flocking behavior displayed by species-specific bird groups. Thus implementation of AI methods for identifying the species consists of two distinct tasks (Fig. 1): (1) identifying and cropping out (also referred to as detecting and bounding) each individual in the image and (2) recognizing species and type (e.g., male or female, sub-adult or adult).

Although AI counting of number of individuals in large aggregations is possible (Descamps et al., 2011) [e.g., automated methods for delineating trees in aerial images of forests (Dalponte et al., 2019) and image segmentation methods of various types of geographical objects, such as land cover and land use types, from aerial images (Volpi & Tuia, 2018)], AI cropping methods of birds in aerial images remain to be better developed. Therefore, in this work, we focus on task 2 because task 1 has been addressed by studies like (Hong et al., 2019; Weinstein et al., 2021). In other words, we used only sets of data that consist of images of individuals that have already been cropped out either manually or through the application of AI procedures. The task at hand then is to build an AI model that automatically recognizes (i.e., classifies) avian species from aerial image segments cropped to include only one individual, often at relatively coarse levels of resolution.

Dataset

For our case study, we used an aerial imagery dataset collected from two study sites over bodies of water: the Atlantic Ocean near Cape Cod, Massachusetts, and Lake Michigan near Manitowoc, Wisconsin, USA. After data collection, wildlife experts manually annotated and cropped images of individual birds (i.e., targets; Fig. 1). These images were then passed to a classification algorithm for species classification (Guirado et al., 2019; Liu et al., 2018). The 10 682 individuals identified in the Cape Cod dataset and 236 identified in the Lake Michigan dataset were annotated by experts into the six different classes illustrated in Figure 2A and B:

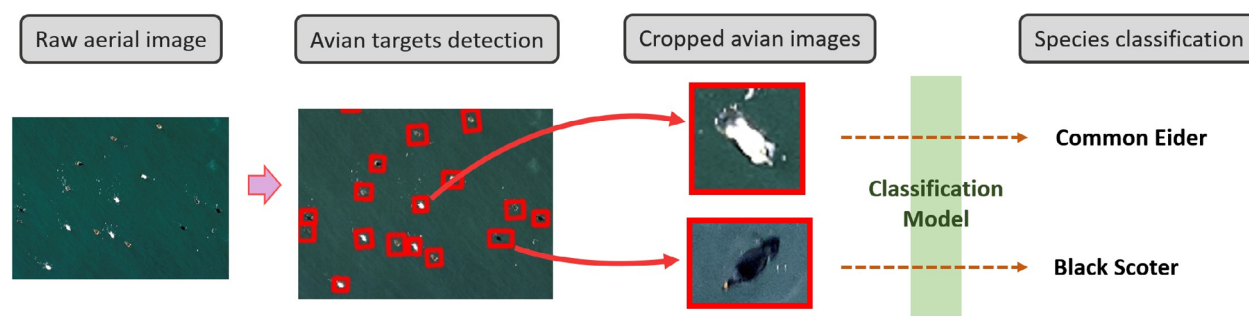


Figure 1. This example shows how a raw aerial image is processed into the final species classification in our case study. Once the raw aerial images are collected, potential objects in the raw images are detected and bounded with boxes either manually or by automatic detection tools (Hong et al., 2019). We used manual bounding boxes from human annotators in our case study. Once the potential objects were cropped around the bounding boxes, our task was to build a deep learning classification model to recognize the actual avian species from these cropped images.

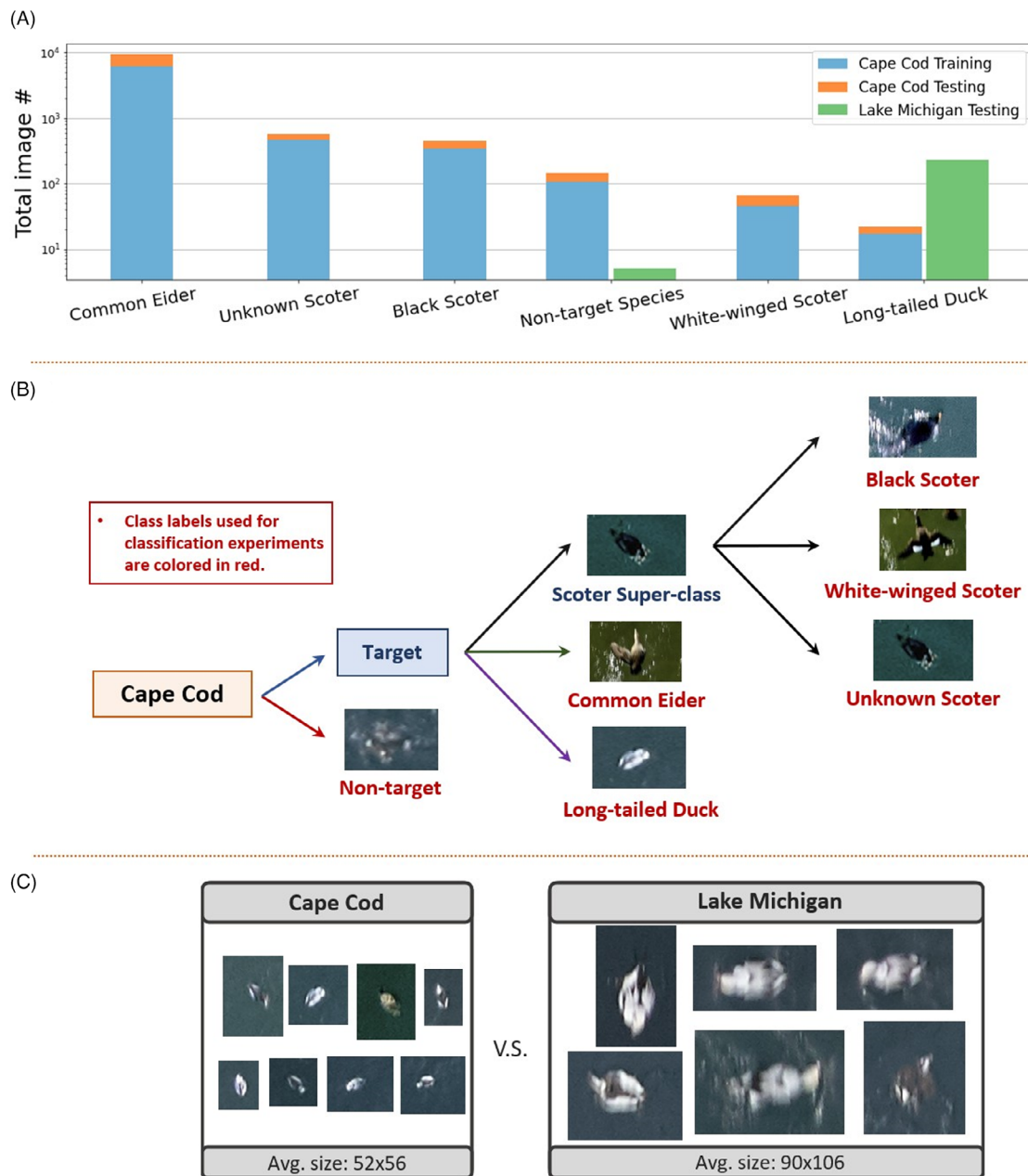


Figure 2. (A) The distribution of our dataset is imbalanced. The classes are sorted by the sample sizes of each class in Cape Cod. The blue, orange, and green colors represent training and testing images from Cape Cod and testing images from Lake Michigan, respectively. In Cape Cod, the largest class, Common Eider, has more than 6246 training images, while the smallest class, Long-tailed Duck, only has 17 training images. In other words, the imbalance ratio of the Cape Cod dataset is 367:1. Lake Michigan dataset only has two classes, Long-tailed Duck and Non-target Species, and it is also imbalanced in terms of class sizes. Long-tailed Duck from Lake Michigan has 231 images, while there are only five images for Non-target Species. y-axis is on log scale. (B) The six classes in the Cape Cod dataset have a hierarchical relationship. Non-target plus three target classes – Scoter Super-class, Common Eider, Long-tailed Duck; and Scoter Super-class could further be categorized/annotated as Black Scoter, White-winged Scoter, and Unknown Scoter. The reason not all Scoter Super-class images could be further categorized to the species level was related to image resolution in our dataset – coarse resolution images posed substantial difficulties for human annotators to make accurate annotations of whether some images were Black Scoter or White-winged Scoter. Specifically, the average image dimension of Unknown Scoter was 56×61 , while the average image dimensions of Black Scoter and White-winged Scoter were 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. On the other hand, we note that the Non-target Species class includes images that do not belong to the other five classes (i.e., mutually exclusive). (C) Long-tailed Duck images from Lake Michigan are 3–4 times larger than those from Cape Cod.

1. Unknown Scoters (scoter individuals that human annotators could not distinguish to the species level)
2. Black Scoter (*Melanitta americana*)
3. White-winged Scoter (*Melanitta deglandi*)
4. Common Eider (*Somateria mollissima*)
5. Long-tailed Duck (*Clangula hyemalis*)
6. Non-target Species (all other avian or non-avian individuals not belonging to the previous classes).

Our model was mainly trained on the Cape Cod dataset. Lake Michigan data were used to evaluate the model's generalization ability (prediction and recognition performance) in different environments and study sites. In other words, we used Lake Michigan data to examine whether the model trained on the Cape Cod dataset could generalize well on the Lake Michigan dataset. Additionally, the Non-target Species class from Lake Michigan does not contain images of target species from Cape Cod. The details of data pre-processing for the experiments are reported in the Appendix S1 (Data section).

Challenges of Avian Recognition in Aerial Imagery

Training with a standard deep learning classification model

We started by applying a standard six-class classification model (i.e., the fundamental classification model without any additional components designed for tasks other than classification) to our Cape Cod dataset because there are six classes (i.e., we treated the Unknown Scoter and the other two scoter classes as three separate classes). The model we used was ResNet-50 (He et al., 2016), a common deep learning convolutional neural network (CNN). The test results from this model are reported in Table 1a. The implementation and hyperparameter tuning details are in the Appendix S1 (Methods section).

Table 1a shows that in the Cape Cod test set, except for the largest class (i.e., most frequently observed class), Common Eider, which had a 99.0% test accuracy, the remaining classes did not produce accurate recognition performance with the standard classification model. Specifically, the two smallest classes (i.e., least observed classes), Long-tailed Duck and White-winged Scoter, had 0.0 and 9.5% test accuracy. This performance inconsistency is negatively related to the training size of each class. In other words, the fewer training images a class had, the less accurate the model was. We also tested our model on the Lake Michigan data, and the performance was also poor. These results indicate that directly applying a standard classification model on our avian datasets is insufficient to produce good recognition performance. Next, we discuss the causes of this performance inconsistency in

the context of *imbalanced distribution*, *annotation uncertainty in categorization*, and *dataset discrepancies*.

Challenge 1: avian imagery data are naturally imbalanced

Data collected during multi-species surveys tend to have an imbalanced (i.e., long-tailed) species distribution because of the natural composition of animal communities (Pimm et al., 2014). Several dominant species are often observed along with many infrequent species that are sparsely represented in datasets. As illustrated in Figure 2 row (A), in the Cape Cod dataset, the largest class had 6246 training images, while the smallest class only had 17 training images. This training data distribution imbalance leads to a substantial recognition performance inconsistency. The fewer training images of a particular species that the model has, the lower the accuracy for that species. In our experiment, the performance was particularly poor for species with smaller training datasets, such as Long-tailed Duck (17 training images) and White-winged Scoter (45 training images), which had 0.0 and 9.5% test accuracy, respectively (Table 1a). However, Common Eider, the largest class in the dataset (6,246 images), had a 99.0% test accuracy.

Challenge 2: annotation uncertainty in categorization

Sometimes aerial data can be collected from aircraft at various distances from the ground surface, resulting in varying spatial resolutions as measured by ground sampling distances (GSDs; i.e., the ground distance between the centers of neighboring image pixels). Thus, the same species may appear at different image resolutions (i.e., number of pixels) within a dataset (Weinstein et al., 2021). For example, in the Cape Cod dataset, the average image dimension of Unknown Scoter images was 56×61 (number of pixels in image width and height), while the average image dimensions of Black Scoter and White-winged Scoter were 100×107 and 96×103 , respectively. In other words, Unknown Scoter images contain 3–4 times fewer pixels on average. Coarse-resolution images increase human annotators' difficulties in making accurate classifications resulting in a coarse annotation rather than individual species annotation of scoter images. Unknown Scoter is one example of coarsely annotated classes (Fig. 2B).

Directly incorporating this coarsely annotated class as an independent class confused the classification models substantially because the model was forced to distinguish similar-looking avian individuals as different classes. In other words, because Unknown Scoter contains images

Table 1. Experiment results.

(a) The standard classification model trained on the Cape Cod training set performed poorly on the Cape Cod and the Lake Michigan test sets.

Species	Cape Train #	Cod Test #	Lake Michigan Test #	Test accuracy (%)	
				Cape Cod	Lake Michigan
Unknown Scoter	466	114	-	69.3	-
Black Scoter	341	108	-	55.6	-
White-winged Scoter	45	21	-	9.5	-
Common Eider	6246	3172	-	99.0	-
Long-tailed Duck	17	5	231	0.0	0.0
Non-target Species	108	38	5	18.4	20.0
Average accuracy (%)				41.9	10.0

(b) The imbalanced model substantially improved the test performance from the standard model on the Cape Cod test set.

Species	Train #	Test #	Test accuracy (%)	
			Standard	Imbalanced
Unknown Scoter	466	114	69.3	41.2
Black Scoter	341	108	55.6	59.3
White-winged Scoter	45	21	9.5	81.0
Common Eider	6246	3172	99.0	91.9
Long-tailed Duck	17	5	0.0	100.0
Non-target Species	108	38	18.4	81.6
Average accuracy (%)			41.9	75.8

(c) On the Cape Cod test set, our soft-fine pseudo-labeling (SPL) approach improved the performance of White-winged Scoter from the imbalanced model by exploiting Unknown Scoter.

Species	Train #	Test #	Test accuracy (%)			
			Standard	Imbalanced	Imb. + Two-Stage	Imb. + SPL (Ours)
Black Scoter	341	108	96.3	91.7	93.5	89.8
White-winged Scoter	45	21	33.3	85.7	71.4	90.5
Common Eider	6246	3172	99.4	91.2	93.0	91.5
Long-tailed Duck	17	5	0.0	100.0	100.0	100.0
Non-target Species	108	38	36.8	78.9	68.4	81.6
Average accuracy of the two scoter classes (%)			64.8	88.7	82.5	90.1
Average accuracy (%)			53.2	89.5	82.3	90.7

(d) With FixMatch as the adaptation component, our model trained on the Cape Cod dataset performed substantially better than methods without the adaptation component

Species	Test #	Test accuracy (%)		
		Standard	Imbalanced + SPL	Imbalanced + SPL + Adaptation
Long-tailed Duck	231	0.0	50.2	80.9
Non-target Species	5	20.0	80.0	80.0
Average accuracy (%)		10.0	65.1	80.5

that can be either Black Scoter or White-winged Scoter, Unknown Scoter images share similar visual features with Black Scoter and White-winged Scoter. Therefore, for a classification model, it is hard to distinguish instances among these three scoter classes. For example, Figure A.1 (A) shows that although Unknown Scoter and Black

Scoter had relatively sufficient training images (466 and 341 training images, respectively), 40.7% of the Black Scoter images were misclassified as Unknown Scoter, and 23.7% of the Unknown Scoter images were mis-classified as Black Scoter. In addition, 61.9% of the White-winged Scoter were misclassified as Unknown Scoter.

Challenge 3: dataset discrepancies often arise among different study sites

In addition to the imbalanced distribution and annotation uncertainty challenges, in practice, ecological monitoring projects often expand over time (Steenweg et al., 2017). New monitoring and study sites are often added, leading to discrepancies among datasets in lighting conditions, background environment, atmospheric conditions, image capturing distances, and animal species compositions. For example, in our case study, Cape Cod and Lake Michigan datasets have different GSDs, which result in different image resolutions and appearances of avian individuals from the same species (Fig. 2C). The Long-tailed Duck images from Cape Cod have 3–4 times the resolution of Lake Michigan images and thus contain more visual details and features. As a result, a classification model trained on coarse-resolution images (Cape Cod) may perform poorly on images with finer-scale resolution (Lake Michigan). For example, the standard model trained on the Cape Cod dataset only had a 10.0% test performance on the Lake Michigan dataset (Table 1a). We demonstrate in the following “Methods and Results” section that this poor performance did not only come from imbalanced distribution but also from dataset discrepancies.

In addition to image appearance discrepancies in datasets from different study sites, expanding surveys or monitoring programs can also change the composition of animal species recorded (Kays et al., 2020). For example, as data collections continue over time, previously undetected species may be encountered (Prach & Walker, 2011) [e.g., less frequent species (Pimm et al., 2014), recolonizing species (David Mech et al., 2019), reintroduced animals (Taylor et al., 2017), or invasive species that are harmful to the ecosystem (Caravaggi et al., 2016; Clavero & Garcia-Berthou, 2005)]. When novel species are introduced, our standard classification model is no longer effective because conventional AI methods require datasets to have fixed numbers of classes (Arjovsky et al., 2019). Therefore, novel species are typically unrecognizable (i.e., not able to be assigned to a category).

Methods and Results

In this section, we provide brief literature reviews of how the computer science community addresses the challenges mentioned in the previous section and present solutions to each challenge.

Solutions for the imbalanced distribution challenge

Imbalanced recognition and long-tailed recognition are areas of machine learning and computer vision research

that address imbalanced classification problems (Cao et al., 2019; Liu et al., 2019; Wang et al., 2020). Common methods include the following:

1. *Training data resampling*: artificially balancing training datasets by either sampling more images from smaller classes (i.e., up-sampling) or sampling fewer images from larger classes (i.e., down-sampling) (He & Garcia, 2009).
2. *Training loss re-weighting*: assigning different weights (i.e., training focus) to the training loss functions based on the number of training images in each class such that the model can have a stronger focus on smaller classes (Cao et al., 2019; Cui et al., 2019; Lin, Goyal, et al., 2017).
3. *Knowledge transfer*: transferring information (such as semantic and visual knowledge) from larger classes to enhance the distinguishability of smaller classes for better classification performance, usually through memory banks and multi-stage training (Kang et al., 2019; Liu et al., 2019; Zhou et al., 2020). For example, Liu et al. (2019) proposed a method that improves recognition performance on smaller classes by exploiting knowledge stored in a memory bank.
4. *Multi-expert models*: combining outputs from multiple experts/sub-models for optimal performance. By assigning data to different experts/sub-models (either through different sampling methods or information complexity metrics), each expert/sub-model can be trained to focus on different parts of the dataset (e.g., abundant or rare classes). As a result, the joint decision of expert models (either through geometric mean or learned fusion mechanisms) can yield more robust performance compared with single-model methods. (Cai et al., 2021; Wang et al., 2020; Zhou et al., 2020).

We used an easy-to-implement yet powerful method called Label Distribution Aware Marginal loss with Deferred Re-Weighting (LDAM-DRW; Cao et al., 2019) to address the imbalanced data distribution in our dataset. Generally speaking, LDAM-DRW is a margin (i.e., sample distance to classifiers) based loss in addition to a scheduled re-weighting technique. LDAM calculates class-specific margins based on the sample size of each class. The fewer training samples a class has, the farther the samples should be from the classifier (i.e., larger margin and thus less confusion), and vice versa. In addition, DRW is a scheduled re-weighting technique that controls when re-weighting based on class sample sizes should be applied to the loss function (Cao et al., 2019; Cui et al., 2019). Compared with traditional re-weighting and re-balancing methods, LDAM-DRW avoids overfitting on rare classes in the early training stage, when the learning rate is relatively larger, and maintains recognition

performance on abundant classes through class-specific margins. In addition, LDAM-DRW does not rely on multiple expert models, which makes the implementation relatively more straightforward. Details of the LDAM-DRW we used are reported in the Appendix S1 (Methods section).

The classification model with an imbalanced component (LDAM-DRW in our experiments) substantially improved the recognition performance on the Cape Cod dataset over the standard classification model (Table 1b). The average class accuracy improved from 41.9 to 75.8%. The largest gain came from the two smallest classes, Long-tailed Duck and White-winged Scoter, from 0.0 to 100.0% and 9.5 to 81.0%, 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 (Liu et al., 2019; Wang et al., 2020).

The confusion matrices (Figure A.1) show that the imbalanced model cleared most of the confusion in Long-tailed Duck and White-winged Scoter because LDAM-DRW assigned larger margins and loss weights to these classes with limited training samples. However, the imbalanced model still struggled to perform well on Unknown Scoter and Black Scoter, with only 41.2 and 59.3% test accuracy, respectively. From Figure A.1(B), it is clear that the confusion within the three scoter classes was still substantial. For example, the imbalanced model misclassified about 44% of Unknown Scoter as either Black Scoter or White-winged Scoter. Meanwhile, about 31% of Black Scoter and 19% of White-winged Scoter were misclassified as Unknown Scoter.

Solutions for the annotation uncertainty in categorization challenge

Since Unknown Scoter introduced substantial confusion to the model and the recognition of Unknown Scoter does not provide species-level information for downstream tasks like population modeling, it is more practical to exclude these coarsely annotated data from model training to eliminate the confusion. When Unknown Scoter is excluded, the task becomes a five-class classification problem. In our experiment, the average class accuracy of fully excluding Unknown Scoter from training and testing improved from 41.9 to 53.2% on the standard model and 75.8 to 89.5% on the imbalanced model compared with the six-class classification results because there was no confusion from Unknown Scoter (Table 1b,c).

However, directly excluding coarsely annotated data is a sub-optimal solution because images with different resolutions can provide complementary information that

ultimately improves the generalization abilities of classification models (Lin, Dollár, et al., 2017). Since Unknown Scoter in our dataset may be the class of relatively coarse resolution images of either Black Scoter or White-winged Scoter, these images may still provide information to improve model performance at the species level, especially when ground-truthed annotations (i.e., human annotations in this context) are limited. For example, although the five-class imbalanced model vastly improved the test accuracy of White-winged Scoter from 9.5 to 85.7% (Table 1c, column Imbalanced), additional performance can be gained by exploiting information contained in Unknown Scoter.

Hierarchical classification is one of the common options addressing uncertain and coarse annotations (Deng et al., 2014). For example, we can split the training process into two stages. In the first stage, we merge Unknown Scoter, Black Scoter, and White-winged Scoter data into one single super-class, Scoter Super-class, and train a classification model on four independent classes (Scoter Super-class, Common Eider, Long-tailed Duck, and Non-target Species). Then, we can train a separate classifier to classify only Black Scoter and White-winged Scoter in the second stage. However, training with multiple stages can quickly become a scaling and model management problem if the dataset has multiple super-classes. Each super-class requires an independent second-stage model and training process. As the number of super-classes increases, the number of models grows as well, such that the overall training time and model management efforts are substantially increased. In addition, the performance error can accumulate because the performance of second-stage models depends on the performance of super-classes in the first stage. For example, on White-winged Scoter, only 17 out of the 21 testing images were classified as Scoter Super-class, and the second stage model was only able to classify from the 17 images and yielded inferior performance compared with single-stage Imbalanced model (Table 1c, column Imb. + Two-Stage).

A novel solution: soft-fine pseudo-labels

An alternative solution is to exploit additional information from coarsely annotated Unknown Scoter images without including it as an independent class while keeping the imbalanced component effective on White-winged Scoter. Therefore, we applied a novel solution called *Soft-fine Pseudo-Labels* (SPL) to address the coarse/uncertain annotation problem that relied only on one stage of training. The method is derived from pseudo-label techniques, a set of techniques in machine learning that use model predictions (i.e., pseudo-labels) to improve the generalization ability of machine learning models (Cascante-Bonilla

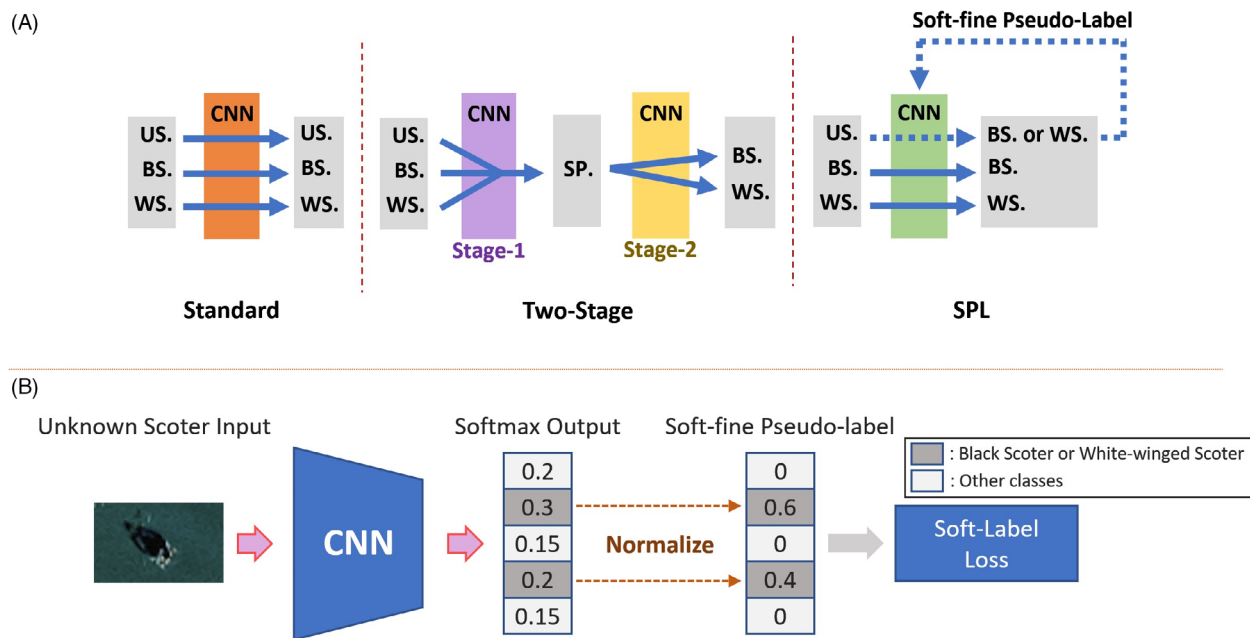


Figure 3. (A) We graphically depict how the three scoter classes can be classified using different approaches. Under the standard approach, Unknown Scoter is treated as an independent class. Under the two-stage setting, the three scoter classes are firstly grouped into one super-class, Scoter Super-class, and then a separate model is trained solely for Black and White-winged Scoter classification. With our SPL approach, coarsely annotated Unknown Scoter images are converted to finer-grained pseudo-labels and used to improve model generalization. US. is Unknown Scoter. BS. is Black Scoter. WS. is White-winged Scoter. SP. is Scoter Super-class. (B) A diagram of our SPL approach to solving the annotation uncertainty challenge using a novel soft-fine pseudo-labeling method. The soft-fine labels are generated by normalizing the Softmax outputs of Unknown Scoter images. To generate the soft-fine pseudo-labels for coarsely annotated Unknown Scoter images, first, we normalized the two values representing Black Scoter and White-winged Scoter from the Softmax outputs to 1 and set the other values to zero. Then, the new vectors were used as the pseudo-labels with soft supervision (i.e., the supervision values are less than 1) on either Black Scoter or White-winged Scoter. With this approach, the supervision from Unknown Scoter is not as strong as independent classes but still relevant to force the model to recognize the images as scoters with higher probabilities than the other classes. Further, this framework does not rely on multiple stages of training and class merging, such that the imbalanced model can still be effective on White-winged Scoter. SPL, Soft-fine pseudo-labels.

et al., 2021; Lee, 2013; Sohn et al., 2020). Pseudo-label techniques were originally developed to address semi-supervised learning, where the predictions of unlabeled/unannotated data are used as *pseudo-labels* (i.e., labels that are generated and not ground-truthed) to tune the model iteratively (Lee, 2013). However, the creation of pseudo-labels for improved model generalization is not limited to the semi-supervised setting.

Figure 3A illustrates the differences in treating the three scoter classes using our SPL approach compared with standard and two-stage models. Unlike conventional pseudo-label approaches in semi-supervised learning that generate pseudo-labels for all possible classes, in our approach, we only use Unknown Scoter images to generate finer-grained Black and White-winged Scoter pseudo-labels. Specifically, we first normalized the outputs of the classification model (five dimension vectors) with a Softmax function (Hinton et al., 2015). Then we normalized the values that represent Black Scoter and White-winged Scoter to 1 and set the other three values to 0 (Fig. 3B).

We used these normalized Softmax values as our soft-fine labels on Unknown Scoter images with an Averaged Binary Cross-entropy (ABCE) loss, which is a loss function traditionally used for samples with multiple co-occurring labels (Tsoumakas & Katakis, 2007).

Our SPL approach forces the model to distinguish between scoter versus non-scoter images because the generated soft-fine pseudo-labels (i.e., training signals) have zeros on the dimensions that represent non-scoter classes. In addition, since the generated soft-fine pseudo-labels had co-occurring labels for both Black Scoter and White-winged Scoter that were treated independently by ABCE loss, the confusion between these two scoter classes was also suppressed. The implementation details can be found in the Appendix S1 (Methods section), and the results of our case study are reported in Table 1c.

As a result, incorporating Unknown Scoter for complementary information with our SPL approach further improved the test accuracy of White-winged Scoter from 85.7 to 90.5% compared with the imbalanced model,

which did not use Unknown Scoter data during training (column Imb. + SPL in Table 1c). In addition, the performance of the non-scooter classes was also improved, especially Non-target Species, which increased from 78.9 to 81.6% compared with the imbalanced model. These improvements indicate that the use of SPL with coarsely annotated data not only relieved the confusion among the scooter classes (Black Scoter and White-winged Scoter) but also among other classes (Figure A.2). However, in hierarchical classification, the performance of other classes was almost the same, if not worse (except for Black Scoter), compared with the six-class classification results (Table 1b). Further, the only additional components to the imbalanced model are the SPL normalization and ABCE loss, making this approach scalable without requiring multiple stages of training. These advantages make SPL more versatile and universal compared with simple hierarchical classification approaches.

However, despite the scalability and the exclusion of coarse annotation confusion, the proposed SPL can sacrifice some performance in the two scooter classes (Black Scoter and White-winged Scoter) compared with hierarchical classification. For example, the effective test accuracy of Black Scoter was 93.5% using two-stage training, whereas it was only 89.8% using our soft-fine label approach (Table 1c). The uncertainty during SPL training likely was the leading cause of this performance drop because the model was trained to identify all species at once with pseudo-labels (i.e., labels that are not ground-truthed).

Unknown Scoter evaluation

In our five-class classification experiments, we only focused on the test performance of samples with finer-grained annotations. With our SPL approach, although Unknown Scoter data were exploited during training, they were excluded from testing because of the lack of finer-grained annotations. Directly applying classification models to provide fine-grained predictions can be particularly challenging when images are too blurry for species to be recognized, as we noted in some of the Unknown Scoter images (Figure A.3). How to efficiently address these blurry images during test time is one of the future research directions.

Solutions for the dataset discrepancy challenge

Visual discrepancies

We tested the performance of our SPL model trained from the Cape Cod dataset on the Lake Michigan dataset. Because of the visual discrepancies between the Cape Cod

and Lake Michigan datasets, the average accuracy of the two classes in Lake Michigan dropped substantially from 90.7 to 65.1% (Table 1c,d). Only 50.2% of the Long-tailed Ducks in Lake Michigan data were correctly classified when the test accuracy on the Cape Code dataset was 100.0% (Table 1c,d, Imbalanced + SPL). These results also show that after the imbalanced distribution challenge was addressed, the model still did not perform well on Long-tailed Duck from Lake Michigan.

One of the most common approaches to address the challenge of incorporating data from new study sites is to fine-tune existing models with new annotations (i.e., transfer learning) (Yosinski et al., 2014). In our example, we need to provide sufficient annotated Lake Michigan data to fine-tune our Cape Cod model such that the model can recognize targets from both study sites. Although the total number of images in the Lake Michigan dataset is relatively small (236 images in our case study) and thus easy to annotate, the effort and resources needed for human annotation on larger datasets are not trivial.

Two machine learning techniques that can help an AI classification model adapt to new sets of data that look different without human annotations are the following:

1. *Domain adaptation*: This technique adapts models trained from one domain (study sites in this context) to other domains either with or without annotations (Peng et al., 2019; Saenko et al., 2010; Venkateswara et al., 2017). Although they can be recognized as the same classes by humans, images from different domains tend to have different distributions in terms of color, texture, and visual appearance. These differences result in distribution discrepancies of the learned feature/latent vectors at the end of CNN models. Thus, *distribution confusion* is the most commonly adopted domain adaptation technique. The technique *confuses* the feature vector distributions of each domain (usually without class annotations) such that the models cannot distinguish which domain the feature vectors come from and learn to use more fundamental information (e.g., structural similarities) to make recognition (Hoffman et al., 2018; Liu, Miao, et al., 2020; Tzeng et al., 2017). Although domain adaptation approaches with distribution confusion may perform better than most other methods, they usually require complicated distribution matching and confusion techniques. For example, one of the state-of-the-art methods, Open Compound Domain Adaptation (OCDA; Liu, Miao, et al., 2020), requires four stages of training and tuning and largely relies on considerable training data, which can be too complicated for small datasets with a limited number of classes and training data.

2. *Semi-supervised learning*: This technique is an alternative option and is usually more straightforward in terms of implementation (Lee, 2013; Sohn et al., 2020). Semi-supervised learning uses unlabeled/unannotated data to improve the generalization ability of AI models, usually through the generation of pseudo-labels (Zhu & Goldberg, 2009). Intrinsically, similar to the mechanisms of advanced domain adaptation approaches with distribution confusion, semi-supervised learning also expands the feature vector distribution by learning from unannotated data (Zhu & Goldberg, 2009). In practice, when data are collected from new study sites, they are treated as unannotated data, and pseudo-labels are then generated for fine-tuning existing models.

Here, we explored how a relatively easy-to-implement semi-supervised learning method, FixMatch (Sohn et al., 2020), adapted our Cape Cod model to Lake Michigan images. FixMatch is a pseudo-label method combined with a technique called *consistency regularization*. With consistency regularization, models are trained to produce consistent outputs of the same inputs (images in this context) that vary by different perturbations such as data augmentation (French et al., 2017). In other words, model outputs of the same inputs are expected to be the same regardless of the perturbation so that the models can focus more on the invariant (or consistently distinguishing) features of inputs, thus improving generalization ability (Xie et al., 2019, 2020). FixMatch perturbs the same inputs (unannotated Lake Michigan data in our experiments) with two augmentation procedures: weak and strong augmentations. Specifically, weakly augmented data are similar to the raw inputs and thus easier for the model to recognize. On the contrary, strongly augmented data are largely distorted from raw inputs and thus hard to recognize. FixMatch uses predictions of weakly augmented inputs as pseudo-labels to train strongly augmented counterparts, which intrinsically regularizes the consistency of the same inputs from two different perturbations (Sohn et al., 2020). In our experiments, every Lake Michigan image was augmented weakly and strongly, and the model regularized the outputs from both augmentations of the same image. The implementation of FixMatch is straightforward because the only extra component required by FixMatch is a two-branch training data augmentation procedure. It can be plugged into our SPL model and other existing AI models without complicated components. The details of this method are provided in Appendix S1 (Methods section).

In Table 1d, we report the results of applying FixMatch as the adaptation component to fine-tuning the Cape Cod model on the Lake Michigan data. Although FixMatch was not initially designed for domain adaptation

(i.e., only for semi-supervised learning tasks), it still substantially improved the classification accuracy on the Lake Michigan Lake dataset without any annotations. Compared with our SPL approach without the adaptation component, the class averaged accuracy improved from 65.1 to 80.5%. Most of the improvements came from Long-tailed Duck, which increased its accuracy from 50.2 to 80.9%.

Novel species

When novel species are introduced, domain adaptation and semi-supervised learning methods are no longer effective because conventional AI recognition methods require datasets to have fixed numbers of classes (Arjovsky et al., 2019). Therefore, novel species are typically unrecognizable. Similar to adapting models to new domains, model fine-tuning through transfer learning with annotated data is also one of the most widely adopted methods to expand the models' recognition capacity (Yosinski et al., 2014). However, since it is uncertain which individuals in the newly collected datasets are of novel species, a complete annotation (i.e., a considerable amount of human effort) is necessary for model fine-tuning.

In such circumstances, improving the efficiency of human annotation becomes a challenge. Ideally, it is possible to automatically identify all the images of novel species, and human effort can focus solely on these images rather than all the newly collected data. Out-of-distribution detection (OOD; DeVries & Taylor, 2018; Scheirer et al., 2013) is one of the related research areas in machine learning that attempts to discover novel samples during test time.

Modern OOD approaches for deep learning usually apply prediction confidence calibration to separate known and novel samples (Liang et al., 2018; Liu, Wang, et al., 2020). In other words, since traditional Softmax-based deep learning models are often overly confident (even on novel samples) (Guo et al., 2017), calibrating the confidence of sample predictions can be effective at separating known and novel samples. Common approaches include the following:

1. *Confidence enhancement*: using additional functions such as smoothed Softmax or energy function to reduce the overconfidence of model predictions such that it is easier to find an effective prediction confidence threshold that separates known *versus* novel samples (Grathwohl et al., 2019; Guo et al., 2017; Hsu et al., 2020; Liang et al., 2018; Liu, Wang, et al., 2020; Szegedy et al., 2016).
2. *Distance-based OOD*: using geometrical distances between samples (e.g., Euclidean distance,

Mahalanobis distance, and cosine similarity) in learned feature/latent embedding spaces as a novelty metric. The distance can be calculated between samples and class centroids (geometric mean) or nearest neighbors (Chen et al., 2020; Liu et al., 2019; Miao et al., 2019; Ren et al., 2021; Techapanurak et al., 2020).

3. *Novel sample generation*: generating artificial novel samples (through data augmentation and generative models) to train AI models to produce lower confidence predictions on novel samples during testing (Goodfellow et al., 2020; Hein et al., 2019).

A more straightforward approach can be applied when non-target species are in the dataset. In most real-world datasets, especially aerial imagery of small-bodied animals with uncertain human annotations, there are often instances of non-target animal species. When we treat these non-target instances as a single class, we can train AI models to classify target versus non-target animal species. Then all the images that are classified as non-target during test time can be sent to human experts for verification. Intrinsically, target versus non-target classification is an OOD technique. For example, Figure 2B shows that target versus non-target species are usually mutually exclusive, and a classifier can be learned between these two sets of classes. Thus, during test time, AI models are very likely to classify images of novel species (unknown to the Cape Cod model in our experiment) as non-target species.

In comparing the methods listed in Table 1d, we set an independent class in both Cape Cod and Lake Michigan datasets for non-target species. Non-Target Species in Lake Michigan does not contain target species in Cape Cod. In the Lake Michigan data (Table 1d), our model successfully identified most of the non-target species (with 80.0% test accuracy), although non-target species in Lake Michigan did not necessarily overlap with those in the Cape Cod dataset. In addition, since the model is trained from Cape Cod, it can classify all four target species from Cape Cod (such as Long-tailed Duck). However, when there are insufficient training data for non-target species, it can be difficult for classification models to generalize well on novel species, and thus, more advanced OOD methods may be necessary.

Conclusion

We tackled three challenges of automated avian recognition in aerial imagery datasets and how various methods can be applied to address these challenges. We evaluated how well existing and our novel SPL approach performed with respect to these three challenges using data from Cape Cod and Lake Michigan.

First, we demonstrated that the classification performance of a standard model is severely curtailed by an imbalance of the number of images of particular species. We showed that this imbalanced distribution challenge can be substantially mitigated by applying a simple imbalanced recognition method (LDAM-DRW), especially on classes with limited training samples like Long-tailed Duck and White-winged Scoter.

Second, we demonstrated that the classification performance of both standard one-stage and hierarchical classification methods was poor on data that included uncertainty in human annotations because of coarse resolution issues. This annotation uncertainty in categorization challenge results in some images being assigned to a coarse annotation (Unknown Scoter). We then demonstrated that classification performance could be much improved using our novel SPL approach that provides a link between coarse and fine-grained annotations. In particular, our approach generated soft-fine pseudo-labels from coarse Unknown Scoter annotations to improve the model's generalization/recognition ability on Black Scoter and White-winged Scoter classes. With our approach, we were able to exploit coarsely annotated data for better model generalization and keep the imbalanced component effective on White-winged Scoter.

Third, we demonstrated that the test performance could be substantially improved using FixMatch when adapting models from data at one site to classifying data at another site. The dataset discrepancies challenge may often cause inconsistent classification performance. In our experiments, we attached FixMatch onto our SPL approach to address resolution discrepancies between datasets from Cape Cod and Lake Michigan and achieved better performance than baselines on the Lake Michigan data without additional annotations. We also experimented with the possibility of using a non-target class, Non-target Species, to detect novel species during testing. Our results show that the model could identify most of the Non-target Species images from the Lake Michigan dataset.

Although each solution we have discussed has its intrinsic limitations, these methods are often flexible and can be combined to accommodate specific requirements. For example, the imbalanced model with LDAM-DRW was combined with SPL and FixMatch to address imbalanced distribution, coarse annotations, and domain discrepancies. We have also demonstrated that existing methods can be easily adjusted for specific tasks. For example, our SPL approach is derived from pseudo-label approaches from semi-supervised learning and multi-label classification.

In addition, the solutions can be easily replaced by more advanced methods in the future if necessary. For

example, when the number of training classes gets bigger, the imbalanced ratio among classes gets larger, and the data distribution gets more long-tailed (i.e., a larger proportion of classes have limited training samples), LDAM-DRW can be replaced by state-of-the-art long-tailed recognition methods like Routing Diverse Distribution-Aware Experts (RIDE; Wang et al., 2020) to produce optimal results. When the domain discrepancies among datasets get more complicated, such as multiple types of backgrounds, FixMatch can be replaced by domain adaptation methods like OCDA for unlimited possibilities of target domains.

On the other hand, some of the challenges we have listed are not specific to aerial avian recognition. For example, imbalanced and long-tailed distribution also exists in ecological datasets derived from other sensor systems such as camera traps (Miao et al., 2021) and bioacoustic monitors (Chronister et al., 2021) because natural animal communities are imbalanced (Pimm et al., 2014). In addition, real-world challenges are not limited to the three examples we displayed in this paper. For example, when combined with automated detection processes, more complicated challenges are intertwined. Through the examples presented here and the literature cited, we have discussed the ways to challenge decomposition and hope to demonstrate the flexibility of deep learning methods, open doors to the ecological community, and promote further research.

Acknowledgments

We thank Brian Lubinski and Dave Fronczak for their assistance with the data collection and annotation of data used in the study. The findings and conclusions in this article are those of the author(s) and do not necessarily represent the views of the U.S. Fish and Wildlife Service.

Funding Information

This project is funded by U.S. Geological Survey, grant number G19AC00203.

Disclaimer

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Data Availability Statement

Data will be released upon publication through: <https://www.sciencebase.gov/catalog/item/63bf21cdd34e92aad3cdac5a> (Miao et al., 2023).

References

- Arjovsky, M., Bottou, L., Gulrajani, I. & Lopez-Paz, D. (2019) Invariant risk minimization. *arXiv* [preprint] arXiv:1907.02893.
- Cai, J., Wang, Y. & Hwang, J.-N. (2021) Ace: ally complementary experts for solving long-tailed recognition in one-shot. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 112–121.
- Cao, K., Wei, C., Gaidon, A., Arechiga, N. & Ma, T. (2019) Learning imbalanced datasets with label-distribution-aware margin loss. *arXiv* [preprint] arXiv:1906.07413.
- Caravaggi, A., Zaccaroni, M., Riga, F., Schai-Braun, S.C., Dick, J.T.A., Montgomery, W.I. et al. (2016) An invasive-native mammalian species replacement process captured by camera trap survey random encounter models. *Remote Sensing in Ecology and Conservation*, 2(1), 45–58.
- Cascante-Bonilla, P., Tan, F., Qi, Y. & Ordonez, V. (2021) Curriculum labeling: revisiting pseudo-labeling for semi-supervised learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8), 6912–6920.
- Chabot, D. & Francis, C.M. (2016) Computer-automated bird detection and counts in high-resolution aerial images: a review. *Journal of Field Ornithology*, 87(4), 343–359.
- Chen, X., Lan, X., Sun, F. & Zheng, N. (2020) A boundary based out-of-distribution classifier for generalized zero-shot learning. In: Vedaldi, A., Bischof, H., Brox, T. & Frahm, J.M. (Eds.) *European conference on computer vision*. Cham: Springer, pp. 572–588.
- Chronister, L.M., Rhinehart, T.A., Place, A. & Kitzes, J. (2021) An annotated set of audio recordings of eastern north American birds containing frequency, time, and species information. *Ecology*, 102, e03329.
- Clavero, M. & Garcia-Berthou, E. (2005) Invasive species are a leading cause of animal extinctions. *Trends in Ecology & Evolution*, 20(3), 110.
- Corcoran, E., Denman, S., Hanger, J., Wilson, B. & Hamilton, G. (2019) Automated detection of koalas using low-level aerial surveillance and machine learning. *Scientific Reports*, 9(1), 1–9.
- Cui, Y., Jia, M., Lin, T.-Y., Song, Y. & Belongie, S. (2019) Class-balanced loss based on effective number of samples. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA. pp. 9260–9269. <https://doi.org/10.1109/CVPR.2019.00949>
- Dalponte, M., Frizzera, L. & Gianelle, D. (2019) Individual tree crown delineation and tree species classification with hyperspectral and LiDAR data. *PeerJ*, 6, e6227.
- David Mech, L., Isbell, F., Krueger, J. & Hart, J. (2019) Gray Wolf (*Canis lupus*) recolonization failure: a Minnesota case study. *The Canadian Field-Naturalist*, 133(1), 60–65.
- Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengio, S. et al. (2014) Large-scale object classification using label relation graphs. In: Fleet, D., Pajdla, T., Schiele, B. &

- Tuytelaars, T. (Eds.) *Computer vision – ECCV 2014. ECCV 2014. Lecture notes in computer science*. Cham: Springer, pp. 48–64.
- Descamps, S., Béchet, A., Descombes, X., Arnaud, A. & Zerubia, J. (2011) An automatic counter for aerial images of aggregations of large birds. *Bird Study*, **58**(3), 302–308.
- DeVries, T. & Taylor, G.W. (2018) Learning confidence for out-of-distribution detection in neural networks. *arXiv [preprint]* arXiv:1802.04865.
- Frederick, P.C., Hylton, B., Heath, J.A. & Ruane, M. (2003) Accuracy and variation in estimates of large numbers of birds by individual observers using an aerial survey simulator. *Journal of Field Ornithology*, **74**(3), 281–287.
- French, G., Mackiewicz, M. & Fisher, M. (2017) Self-ensembling for visual domain adaptation. *arXiv [preprint]* arXiv:1706.05208.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S. et al. (2020) Generative adversarial nets. *Communications of the ACM*, **63**(11), 139–144.
- Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., Norouzi, M. & Swersky, K. (2019) Your classifier is secretly an energy based model and you should treat it like one. *arXiv [preprint]* arXiv:1912.03263.
- Guirado, E., Tabik, S., Rivas, M.L., Alcaraz-Segura, D. & Herrera, F. (2019) Whale counting in satellite and aerial images with deep learning. *Scientific Reports*, **9**(1), 1–12.
- Guo, C., Pleiss, G., Sun, Y. & Weinberger, K.Q. (2017) On calibration of modern neural networks. *International Conference on Machine Learning, PMLR*, **70**, 1321–1330.
- He, H. & Garcia, E.A. (2009) Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, **21**(9), 1263–1284.
- He, K., Zhang, X., Ren, S. & Sun, J. (2016) Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV. IEEE, pp. 770–778.
- Hein, M., Andriushchenko, M. & Bitterwolf, J. (2019) Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 41–50.
- Hinton, G., Vinyals, O. & Dean, J. (2015) Distilling the knowledge in a neural network. *arXiv [preprint]* arXiv:1503.02531.
- Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K. et al. (2018) Cycada: Cycle-consistent adversarial domain adaptation. *Proceedings of the 35th International Conference on Machine Learning, PMLR*, **80**, 1989–1998.
- Hong, S.-J., Han, Y., Kim, S.-Y., Lee, A.-Y. & Kim, G. (2019) Application of deep-learning methods to bird detection using unmanned aerial vehicle imagery. *Sensors*, **19**(7), 1651.
- Hsu, Y.-C., Shen, Y., Jin, H. & Kira, Z. (2020) Generalized Odin: detecting out-of-distribution image without learning from out-of-distribution data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10951–10960.
- Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J. et al. (2019) Decoupling representation and classifier for long-tailed recognition. *arXiv [preprint]* arXiv:1910.09217.
- Kays, R., Arbogast, B.S., Baker-Whitton, M., Beirne, C., Boone, H.M., Bowler, M. et al. (2020) An empirical evaluation of camera trap study design: how many, how long and when? *Methods in Ecology and Evolution*, **11**, 700–713.
- Lee, D.-H. (2013) Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. *Workshop on Challenges in Representation Learning, ICML*, **3**(2), 896.
- Liang, S., Li, Y. & Srikant, R. (2018) Enhancing the reliability of out-of-distribution image detection in neural networks. *ICLR*.
- Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. & Belongie, S. (2017) Feature pyramid networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2117–2125.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollar, P. (2017) Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*. pp. 2980–2988.
- Liu, W., Wang, X., Owens, J.D. & Li, Y. (2020) Energy-based out-of-distribution detection. *arXiv [preprint]* arXiv:2010.03759.
- Liu, Y., Sun, P., Highsmith, M.R., Wergeles, N.M., Sartwell, J., Raedeke, A. et al. (2018) Performance comparison of deep learning techniques for recognizing birds in aerial images. *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*. Guangzhou, China. IEEE, pp. 317–324.
- Liu, Z., Miao, Z., Pan, X., Zhan, X., Lin, D., Yu, S.X. et al. (2020) Open compound domain adaptation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 12406–12415.
- Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B. & Stella, X.Y. (2019) Large-scale long-tailed recognition in an open world. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2537–2546.
- McEvoy, J.F., Hall, G.P. & McDonald, P.G. (2016) Evaluation of unmanned aerial vehicle shape, flight path and camera type for waterfowl surveys: disturbance effects and species recognition. *PeerJ*, **4**, e1831.
- Miao, Z., Gaynor, K.M., Wang, J., Liu, Z., Muellerklein, O., Norouzzadeh, M.S. et al. (2019) Insights and approaches using deep learning to classify wildlife. *Scientific Reports*, **9**(1), 1–9.
- Miao, Z., Liu, Z., Gaynor, K.M., Palmer, M.S., Yu, S.X. & Getz, W.M. (2021) Iterative human and automated identification of wildlife images. *Nature Machine Intelligence*, **3**(10), 885–895.
- Miao, Z., Yu, S.X., Landolt, K.L., Koneff, M.D., White, T.P., Fara, L.J. et al. (2023) Images and annotations to automate

- the classification of avian species. *U.S. Geological Survey data release*. <https://doi.org/10.5066/P9YL80R6>
- Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K. & Wang, B. (2019) Moment matching for multi-source domain adaptation. *ICCV*.
- Pimm, S.L., Jenkins, C.N., Abell, R., Brooks, T.M., Gittleman, J.L., Joppa, L.N. et al. (2014) The biodiversity of species and their rates of extinction, distribution, and protection. *Science*, **344**, 6187.
- Prach, K. & Walker, L.R. (2011) Four opportunities for studies of ecological succession. *Trends in Ecology & Evolution*, **26** (3), 119–123.
- Redfern, J.V., Viljoen, P.C., Kruger, J.M. & Getz, W.M. (2002) Biases in estimating population size from an aerial census: a case study in the Kruger National Park, South Africa: Starfield festschrift. *South African Journal of Science*, **98**(9), 455–461.
- Ren, J., Fort, S., Liu, J., Roy, A.G., Padhy, S. & Lakshminarayanan, B. (2021) A simple fix to mahalanobis distance for improving near-OOD detection. *arXiv [preprint]* arXiv:2106.09022.
- Saenko, K., Kulis, B., Fritz, M. & Darrell, T. (2010) Adapting visual category models to new domains. In: Daniilidis, K., Maragos, P. & Paragios, N. (Eds.) *Computer vision – ECCV 2010. ECCV 2010. Lecture notes in computer science*. Berlin, Heidelberg: Springer.
- Sasse, D.B. (2003) Job-related mortality of wildlife workers in the United States, 1937–2000. *Wildlife Society Bulletin*, **31**(4), 1015–1020.
- Scheirer, W.J., Rocha, A., Sapkota, A. & Boulton, T.E. (2013) Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 1757–1772.
- Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E.D. et al. (2020) Fixmatch: simplifying semi-supervised learning with consistency and confidence. *arXiv [preprint]* arXiv:2001.07685.
- Steenweg, R., Hebblewhite, M., Kays, R., Ahumada, J., Fisher, J.T., Burton, C. et al. (2017) Scaling-up camera traps: monitoring the planet's biodiversity with networks of remote sensors. *Frontiers in Ecology and the Environment*, **15**(1), 26–34.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. (2016) Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas. pp. 2818–2826.
- Taylor, G., Canessa, S., Clarke, R.H., Ingwersen, D., Armstrong, D.P., Seddon, P.J. et al. (2017) Is reintroduction biology an effective applied science? *Trends in Ecology & Evolution*, **32**(11), 873–880.
- Techapanurak, E., Suganuma, M. & Okatani, T. (2020) Hyperparameter-free out-of-distribution detection using cosine similarity. *Proceedings of the Asian Conference on Computer Vision*.
- Tsoumakas, G. & Katakis, I. (2007) Multi-label classification: an overview. *International Journal of Data Warehousing and Mining (IJDDWM)*, **3**(3), 1–13.
- Tuia, D., Kellenberger, B., Beery, S., Costelloe, B.R., Zuffi, S., Risse, B. et al. (2022) Perspectives in machine learning for wildlife conservation. *Nature Communications*, **13**(1), 792. Available from: <https://doi.org/10.1038/s41467-022-27980-y>
- Tzeng, E., Hoffman, J., Saenko, K. & Darrell, T. (2017) Adversarial discriminative domain adaptation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI.
- Venkateswara, H., Eusebio, J., Chakraborty, S. & Panchanathan, S. (2017) Deep hashing network for unsupervised domain adaptation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, pp. 5385–5394.
- Volpi, M. & Tuia, D. (2018) Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, **144**, 48–60.
- Wang, D., Shao, Q. & Yue, H. (2019) Surveying wild animals from satellites, manned aircraft and unmanned aerial systems (UASs): a review. *Remote Sensing*, **11**(11), 1308.
- Wang, X., Lian, L., Miao, Z., Liu, Z. & Yu, S.X. (2020) Long-tailed recognition by routing diverse distribution-aware experts. *arXiv [preprint]* arXiv:2010.01809.
- Weinstein, B.G., Garner, L., Saccomanno, V.R., Steinkraus, A., Ortega, A., Brush, K. et al. (2021) A general deep learning model for bird detection in high resolution airborne imagery. *Ecological Applications*, **32**, e2694.
- Xie, Q., Dai, Z., Hovy, E., Luong, M.-T. & Le, Q.V. (2019) Unsupervised data augmentation for consistency training. *arXiv [preprint]* arXiv:1904.12848.
- Xie, Q., Luong, M.-T., Hovy, E. & Le, Q.V. (2020) Self-training with noisy student improves imagenet classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10687–10698.
- Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. (2014) How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, **2**, 3320–3328.
- Zhou, B., Cui, Q., Wei, X.-S. & Chen, Z.-M. (2020) BBN: bilateral-branch network with cumulative learning for long-tailed visual recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9719–9728.
- Zhu, X. & Goldberg, A.B. (2009) Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, **3**, 1–130.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table A.1. Details of Cape Cod training-testing split, and Lake Michigan testing set.

Table A.2. List of hyperparameters used in the baseline experiments.

Table A.3. Augmentation pool for FixMatch fine-tuning.

Figure A.1. On the Cape Cod test set, the model generally performed poorly because of the imbalanced data distribution, and substantial recognition confusion exists among the three scoter classes.

Figure A.2. Our SPL approach further reduced the confusion within Black Scoter and White-winged Scoter from the imbalanced model within the Cape Cod test set.

Figure A.3. It is hard to verify the predictions of Unknown Scoter images.