


2023

Automating sandhill crane counts from nocturnal thermal aerial imagery using deep learning

Emilio Luz-Ricca

RESEARCH ARTICLE

Automating sandhill crane counts from nocturnal thermal aerial imagery using deep learning

Emilio Luz-Ricca¹ , Kyle Landolt², Bradley A. Pickens³ & Mark Koneff⁴¹Institute for Integrative Conservation, William & Mary, 221 North Boundary Street, Williamsburg Virginia, 23185, USA²Upper Midwest Environmental Sciences Center, U.S. Geological Survey, 2630 Fanta Reed Road, La Crosse Wisconsin, 54603, USA³Division of Migratory Bird Management, U.S. Fish and Wildlife Service, 11510 American Holly Drive, Laurel Maryland, 20708, USA⁴Branch of Migratory Bird Surveys I Division of Migratory Bird Management, U.S. Fish and Wildlife Service, 69 Grove Street Extension, Orono Maine, 04469, USA

Keywords

Computer vision, deep learning, sandhill crane, thermal imagery, wildlife monitoring

Correspondence

Emilio Luz-Ricca, Institute for Integrative Conservation, William & Mary, 221 North Boundary Street, Williamsburg, Virginia 23185, USA. Tel: +1 (703) 559-9472. Email: emilio.luz.ricca@gmail.com

Editor: Temuulen Sankey

Associate Editor: Francesco Rovero

Received: 2 February 2022; Revised: 8 July 2022; Accepted: 8 August 2022

doi: 10.1002/rse2.301

Remote Sensing in Ecology and Conservation 2023;9 (2):182–194

Abstract

Population monitoring is essential to management and conservation efforts for migratory birds, but traditional low-altitude aerial surveys with human observers are plagued by individual observer bias and risk to flight crews. Aerial surveys that use remote sensing can reduce bias and risk, but manual counting of wildlife in imagery is laborious and may be cost-prohibitive. Therefore, automated methods for counting are critical to cost-efficient application of remote sensing for wildlife surveys covering large areas. We conducted nocturnal surveys of sandhill cranes (*Antigone canadensis*) during spring migration in the Central Platte River Valley of Nebraska, USA, using midwave thermal infrared sensors. We developed a framework for automated counting of sandhill cranes from thermal imagery using deep learning, assessed and compared the performance of two automated counting models, and quantified the effect of spatial resolution on counting accuracy. Aerial thermal imagery data were collected in March 2018 and 2021; 40 images were analyzed. We applied two deep learning models: an object detection approach, Faster R-CNN and a recently developed pixel-density estimation approach, ASPDNet. Model performance was determined using data independent of the training imagery. The effect of spatial resolution was quantified with a beta regression on relative error. Our results showed model accuracy of 9% mean percent error for ASPDNet and 18% for Faster R-CNN. Most error was related to the undercounting of sandhill cranes. ASPDNet had <50% of the error of Faster R-CNN as measured by mean percent error, root-mean-squared error and mean absolute error. Spatial resolution affected accuracy of both models, with error rate increasing with coarser resolution, particularly with Faster R-CNN. Deep learning models, particularly pixel-density estimators, can accurately automate counting of migratory birds in a dense, aggregate setting such as nocturnal roosting sites.

Introduction

Population monitoring is critical to inform management and conservation of wildlife populations (Nichols & Williams, 2006). Count-based methods using human observers are often applied in wildlife monitoring, but observer bias can be substantial, and adjusting for this bias often requires auxiliary data collection, which can be laborious and costly to collect (Pearse et al., 2008; Williams et al., 2002). Counts from aircraft are often conducted

when the geographic region of interest has a broad spatial extent or is inaccessible by ground. Although aerial surveys using human observers can cover these extents, for small-bodied wildlife such as birds, low-level flight is required, which increases risk for flight crews (Sasse, 2003).

Migratory birds are of international concern and their migratory staging, or stopover, areas are critical to their populations (Kirby et al., 2008; Rakhimberdiev et al., 2018). An estimated 80–85% of the Mid-continent

Population (MCP) of sandhill crane (*Antigone canadensis*) stages in the Central Platte River Valley of Nebraska, USA, during a 3–4 week period in spring (Caven, Buckley, et al., 2020) (Figure 1). Low-level ocular surveys have been conducted along the Central Platte River and North Platte River each spring using consistent methodology since 1982 (Benning & Johnson, 1987) to inform recreational harvest and other population management decisions (Central Flyway Webless Migratory Game Bird Technical Committee, 2018). The current aerial survey is conducted with human observers at a low altitude (60–75 meters above ground level (AGL)) on fixed strip transects; the survey extent was established to include all sandhill crane foraging areas in this important migratory staging region. The survey incorporates the collection of oblique photography by a third crew member in the aircraft rear seat to correct for bias in primary-observer estimates of crane flock sizes (Benning & Johnson, 1987).

Annual indices of sandhill crane abundance generated from this methodology are highly variable (Pearse et al., 2015), leading managers to re-assess the survey design and methodology. Sandhill cranes are dispersed while foraging diurnally when ocular surveys are conducted and there is concern that the current survey coverage is inadequate due to changes in foraging distribution. At night, sandhill cranes aggregate and roost on or near the river channels, which may offer an opportunity for a more efficient and complete survey of the crane population using this area. Nocturnal surveys using thermal

infrared imagery have been identified as a priority need for this population (Association of Fish and Wildlife Agencies' Migratory Shore and Upland Game Bird Support Task Force, 2016). Thermal imagery has the potential to be used for wildlife surveys when sufficient disparity exists in the thermal response of the target and background (Corcoran et al., 2019; Seymour et al., 2017). However, the spatial resolution offered by thermal sensors is coarser as compared to other airborne remote sensing approaches (McKellar et al., 2021), which may hinder detection or differentiation of small targets (Santangeli et al., 2020). The dense aggregation patterns of roosting cranes can exacerbate challenges posed by the lower spatial resolution of thermal sensors due to target merging and occlusion (*i.e.* Figure 2).

Application of remote sensing methods to survey wildlife populations over broad geographic areas can produce large volumes of data. The requirement for manually processing large image datasets may be cost and staff prohibitive for natural resource agencies and may delay the availability of survey results supporting time-sensitive management decisions. Machine learning, particularly deep learning (DL), methods show promise in automating some data processing steps to reduce barriers to implementation of remote sensing survey methods (Corcoran et al., 2021; Weinstein, 2018). DL represents a class of machine learning models based on artificial neural networks, which loosely mimic the structure of a brain by using multiple layers of neurons to make predictions (LeCun et al., 2015). These neural networks learn complicated nonlinear representations of high-dimensional data such as images. Computer vision (CV), defined as DL using images or videos as input, is one of the most active areas of research in DL. The capacity to process and predict from imagery has improved tremendously in the past few years, mainly due to advancements in convolutional neural networks (CNNs) (LeCun et al., 2015).

Recently, success with CNNs has motivated more difficult prediction tasks like object detection (*i.e.* Ren et al. (2015), Lin et al. (2017), Redmon and Farhadi (2018)), where objects of interest within an image are identified with a bounding box and classified into a category (*i.e.* bird species). Extracting an object count from a trained object detector can be as simple as counting the number of predicted bounding boxes in an image. Therefore, object detectors could also be considered an approach for the task of object counting within an image (Gao et al., 2020). However, counting *via* object detection requires precise localization of each object within an image and has been shown to perform poorly compared to other methods, such as direct DL-based count regression (Chattopadhyay et al., 2017). Object detection models are also known to struggle with small, densely-packed



Figure 1. Study area near the Platte River in Nebraska, USA, where sandhill cranes stage on their spring migration northward. The study area indicated is the centroid of data collected in 2021, which is roughly indicative of the region surveyed in this study.

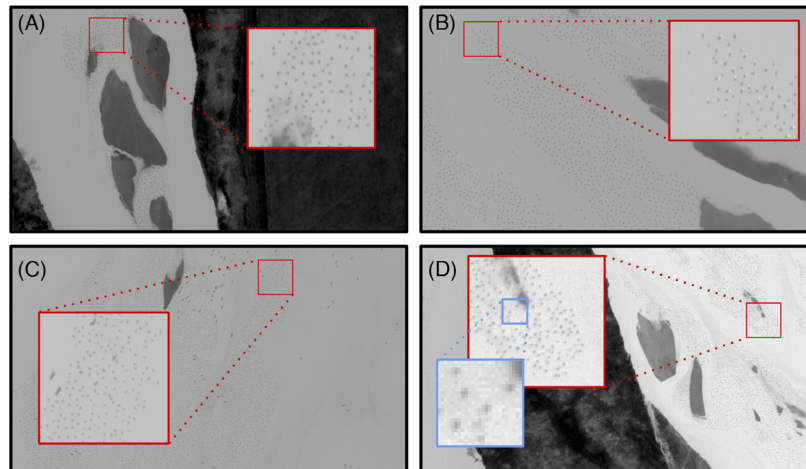


Figure 2. Example images (A–D) from our dataset obtained in March 2018 and March 2021 using a FLIR A8303sc midwave thermal sensor (FLIR Systems, Inc., Nashua, New Hampshire) with a 50-mm lens in 2018 and a 100-mm lens in 2021. Darker areas are characterized as land and vegetation while lighter areas are water or sediments, with zoomed-in areas revealing representative sandhill crane thermal signatures. In particular, B and D contain dense clusters of sandhill crane individuals, and the blue zoomed-in area in D shows issues with resolution, with crane signatures only occupying a few pixels.

objects, especially when image resolution is coarse with respect to object size (Gao et al., 2021; Pang et al., 2019; Pham et al., 2020). As summarized by Gao et al. (2020), density estimation was developed to address the task of counting densely aggregated objects in images. Density estimators indirectly generate an object count for an image by predicting a density map—a count for each pixel or for blocks of pixels in the input image—which avoids the need for localization of each discrete object within an image (Lempitsky & Zisserman, 2010).

A variety of studies have found it advantageous to apply DL models to quickly detect or count birds in very high resolution (*i.e.* <4 cm) diurnal aerial imagery (Akçay et al., 2020; Corcoran et al., 2021; Hong et al., 2019; Kellenberger et al., 2021). Hong et al. (2019) experimented with several different DL object detection models for bird species and found that Faster R-CNN (Ren et al., 2015) consistently performed well. To address situations with very dense aggregations of bird targets, several studies have proposed novel DL-based counting architectures (Arteta et al., 2016; Kellenberger et al., 2021; Kim & Kim, 2020). In particular, Kim and Kim (2020) integrated density estimation into their architecture, with strong counting performance on images of crowded birds from a variety of perspectives. Even with these advancements, factors that affect model performance, such as spatial resolution, are generally unexplored. Additionally, studies predominantly focus on RGB imagery, with little exploration of thermal imagery.

Our study objectives were to: (1) develop and test a DL framework for automated counting of sandhill cranes

from nocturnally acquired thermal imagery, (2) compare the performance of two automated counting approaches and (3) quantify the effect of spatial resolution on counting performance.¹

Materials and Methods

Study area, survey design and imagery collection

Thermal imagery was obtained over the Platte River in central Nebraska, USA on March 20 and 21, 2018 and March 21, 2021 (Fig. 1). We used a FLIR A8303sc mid-wave thermal sensor (FLIR Systems, Inc., Nashua, New Hampshire) with a 50-mm diameter lens to obtain imagery in 2018 and the same sensor was used in 2021 with a 100-mm diameter lens. In the 2018 survey, the thermal sensor was mounted to a Partenavia P-68 Observer 2 aircraft; in the 2021 survey, a Quest Kodiak 100 model aircraft was used. Surveys were conducted at a flying speed of approximately 100 to 130 knots (50 to 65 meters/second) and imagery was obtained at even intervals of one frame per second. The surveys from both years were obtained near the peak levels of crane abundance (Caven, Varner, et al., 2020). The surveys were flown at various altitudes, which resulted in differing spatial resolutions. We measure spatial resolution in terms of ground sample distance (GSD) in centimeters (cm). Flights in 2018 were flown at 610 m and 762 m AGL and flights in 2021 were flown at even intervals between 610 m and 1067 m AGL. The thermal sensors used produced single-band imagery (Fig. 2).

Staging sandhill cranes concentrate each evening on roosting sites within or near the Platte and North Platte River channels. Imagery was obtained at night to exploit a known distribution of cranes and minimize sampling variation associated with crane distributions falling outside the diurnal survey area. In addition, movement of roosting cranes is limited at night, which minimizes double-counting of individuals over multiple images. Imagery collection in 2018 occurred over a large portion of the Central Platte River. Imaging in 2021 targeted several areas that have historically had large aggregations of roosting cranes; at three key areas, ground observers in observation blinds used thermal scopes to identify non-target waterbird species to potentially support species classification, but only three non-target individuals were identified overall (compared to 7500+ estimated sandhill crane individuals).

From the imagery collected in 2018 and 2021, 40 images were selected (10 from 2018 data and 30 from 2021) focusing on frames containing sandhill cranes (*i.e.* Fig. 2). This served to increase the variety of sandhill crane thermal signatures available for training the models and reduced the number of images lacking sandhill cranes, mitigating the overwhelming quantity of background examples during model training (Kellenberger et al., 2018). We also ensured a variety of resolutions were represented in the dataset by selecting imagery from 8.5 cm to 21.2 cm GSD. These 40 images were split into training (24 images), validation (4 images) and testing datasets (12 images). Training data were used only to train the model, validation data were used for model selection and the test data were used to quantify model performance on independent imagery. We will refer to the initial images as 'parent images', which are all 1280×720 pixels. These parent images were split among three annotators, and all sandhill crane signatures were identified in each image *via* bounding box annotation using the software labelImg² 1.8.5. We collected thermal imagery during the time of year when sandhill crane abundance dominates the Platte River ecological landscape (>1 million individuals as per the estimates of Caven, Buckley, et al. (2020)). Additionally, extensive ground observations during our 2021 survey revealed great difficulty in finding non-target species; relatively few were likely to be present. The large body size of sandhill cranes relative to other species commonly present on the Platte River and the characteristic roosting patterns of staging sandhill cranes further ensured the images chosen contained primarily sandhill crane signatures. Therefore, we assumed all thermal signatures in the selected imagery were sandhill cranes. One set of annotations was generated for each image, which was completed entirely by the same annotator. Annotators communicated only to

evaluate rare ambiguous signatures, with annotation decisions made unanimously in these cases.

Automated counting with deep learning

Object detection with faster R-CNN

Faster R-CNN (Ren et al., 2015) is a two-stage object detection architecture that has performed well in conservation monitoring tasks (Corcoran et al., 2019; Duporge et al., 2021; Guirado et al., 2019; Hong et al., 2019). Faster R-CNN splits the object detection process into three main components: initial feature extraction using many convolutional layers, a region proposal network (first stage) that predicts regions that are deemed 'most likely' to contain objects, and a refinement step (second stage) using Fast R-CNN to regress final bounding boxes and generate predicted classes for each bounding box (Akçay et al., 2020; Ren et al., 2015). The input to Faster R-CNN is an image of size $H \times W \times C$, where H is the height of the image in pixels, W the width in pixels and C the number of channels. Faster R-CNN predicts bounding boxes in the form $(x_{min}, y_{min}, x_{max}, y_{max})$, which aims to cover the spatial extent of a particular target, each with a predicted class—in our case, the two classes are sandhill crane and background.

Density estimation with ASPDNet

ASPDNet (Gao et al., 2021) counts objects by training to match a kernel density estimate derived from point annotations for the imagery (Lempitsky & Zisserman, 2010). Because our ground truth values were initially in the form of annotated bounding boxes, we translated bounding boxes to density maps, following Lempitsky and Zisserman (2010). We extracted the centroid of each bounding box to obtain point annotations for each image. We then let $P_I = \{P_1, P_2, \dots, P_{C_I}\}$ denote the set of two-dimensional (2D) coordinates for point annotations in image I , with a total count of C_I . We define the ground truth density map, D_I , as a sum of a normalized 2D Gaussian kernel evaluated at each pixel, so that at any pixel p in image I , the density is defined as:

$$D_I(p) = \sum_{P \in P_I} N(p; P, \sigma^2 \mathbf{1}_{2 \times 2}) \quad (1)$$

where $N(p; P, \sigma^2 \mathbf{1}_{2 \times 2})$ is the evaluation of a Gaussian kernel with a mean of point annotation P .

An appealing aspect of the produced density map is that it preserves the total annotated count for the image in question: we simply integrate over the entire image (or sum, in the finite case) to produce our count. Thus, this density map encourages rough localization while also

providing more ‘continuous’ supervision than simple point annotations. We used a fixed value for σ and found $\sigma = 3$ to work well based on initial experiments.

ASPDNet (Gao et al., 2021) is a recently developed density estimation DL model specifically designed for application to remote sensing imagery with dense objects (i.e. ships, cars, buildings). ASPDNet is an architecture consisting of multiple modules: a truncated version of VGG-16 (Simonyan & Zisserman, 2015) for initial feature extraction, channel- and spatial-attention modules to highlight important features, a scale pyramid module to extract patterns at multiple scales and a deformable convolution module to make the model robust to variations in object orientation. ASPDNet predictions are in the form of a density map of size $\frac{1}{8}$ the spatial dimension of the given input. To ensure that the sizes of the ground truth and predicted densities match, we downsized the ground truth density map using bicubic interpolation, following Gao et al. (2021), to create the final ground truth density map.

Data processing pipeline

To integrate the DL counting component with other key elements, we designed a basic data pipeline (Fig. 3). This pipeline was used for all steps of model development (training, validation and testing).

Model training and testing protocols

Imagery was split into train, validation and test sets at the parent-image level, so that testing and validation tiles were entirely independent of training tiles. Both models

were trained and evaluated on the same images, split into the same train/validation/test subsets. As a whole, the imagery included in the dataset ranged from an estimated spatial resolution of 8.5 cm to 21.2 cm GSD, with an imbalance toward lower GSDs. The test set was balanced to evenly represent all available spatial resolutions. This resulted in two images at each spatial resolution.

For both models, we trained and predicted on smaller 200×200 pixel patches of the parent images, which we will refer to as ‘tiles’ (Fig. 3B). Tiling is an important pre-processing step to avoid excessive computational cost and has been applied previously for training object detectors to identify small bird targets in aerial imagery (Hong et al., 2019); the tile size was chosen through experimentation. Because large portions of each image in the dataset included no sandhill cranes, we limited the number of tiles that lacked sandhill cranes to ensure the networks were not inundated with background examples. During training, each batch consisted of five tiles chosen at random from a single parent image in the training set, with at most one tile where no sandhill cranes were present included per batch. During evaluation and testing, each parent image was zero-padded, meaning black pixels were added to the border of the image to ensure that an integer number of tiles could be extracted. Parent images were then tiled into 28 non-overlapping 200×200 pixel tiles. Training was performed only on portions of each parent image, but testing and validation were performed on the full parent image.

We implemented Faster R-CNN in PyTorch³ 1.8.1 and used a ResNet50 architecture (He et al., 2016) as the initial feature extractor, with weights pre-trained on ImageNet (Russakovsky et al., 2015). The pre-trained

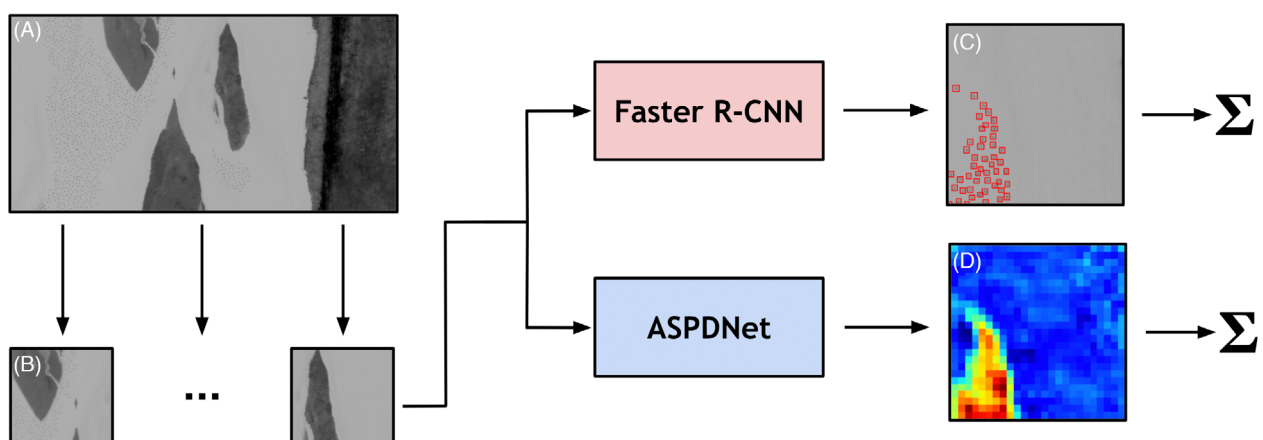


Figure 3. An overview of our process for counting sandhill cranes (*Antigone canadensis*) on the Platte River Valley of Nebraska, USA. We begin in A with a thermal aerial parent image, which is split into non-overlapping tiles in B. These tiles pass through either a Faster R-CNN or ASPDNet convolutional neural network to produce predicted sandhill crane bounding boxes in C or a predicted density map in D, respectively. Finally, detections in C are counted or the density map in D is summed over for all tiles, resulting in a final crane count for the image depicted in A.

ResNet50 weights were frozen for the duration of model training, whereas the rest of the network was randomly initialized and fully trainable. We tuned several key hyperparameters through experimentation: the maximum number of allowable bounding box detections per image was set to 500 (greater than the maximum number of birds observed in tiles from our dataset), the bounding box non-maximum suppression threshold was set to 0.3 (partially motivated by tests carried out by Hong et al. (2019)), and the bounding box score threshold was set to 0.4. Adjusting the score threshold reduced false positives (*i.e.* detecting a crane where there were none), and adjusting the maximum number of box detections reduced false negatives (*i.e.* failing to detect a crane that was present). We also applied random brightness and contrast shifts using the Python data augmentation library Albumentations⁴ 1.0.0 to simulate future imagery obtained using alternate thermal sensors or under varying weather conditions, to support model generalization. We trained for 40 epochs using stochastic gradient descent (SGD) with an initial learning rate of 1e-3, a momentum value of 0.9 and weight decay (L_2 regularization term) of 5e-3 (Ren et al., 2015). We halved the learning rate whenever the validation metrics plateaued for five epochs. To determine whether each bounding box predicted by Faster R-CNN was a true or false positive, the intersection-over-union (IoU) was used; only bounding box predictions with IoUs above a selected threshold (between 0 and 1) were deemed true positives (see Hong et al. (2019)).

We implemented ASPDNet in PyTorch 1.8.1, slightly modified from Gao et al. (2021). Because it is impossible for tiles to have a negative number of targets, we thresholded all final density values in the output density map to zero by appending a rectified linear unit (ReLU) to the model. We froze the weights in the initial VGG-16 feature extractor, which were obtained through pre-training on ImageNet (Russakovsky et al., 2015), because this improved results during experimentation. Following Gao et al. (2021), we trained using SGD with an initial learning rate of 1e-7, a momentum value of 0.95, a weight decay of 5e-4 and applied random horizontal/vertical flips using Albumentations 1.0.0 in addition to the augmentations for Faster R-CNN. Every 30 epochs we reduced the learning rate by a factor of 10. We trained for 200 epochs in total, saving the best model based on performance on the validation set.

Evaluating automated crane counts

To evaluate the performance of the two models on individual parent images, we used four metrics. For sandhill crane detection (*i.e.* Faster R-CNN), average precision (AP) was used. AP varies between 0 and 100, with 100

implying high detection ability at a selected IoU threshold. We used implementations provided by Padilla et al. (2021) to calculate AP and chose a relatively low IoU threshold of 0.3 to avoid mislabeling correct detections as false positives, a potential issue when the target size is very small relative to the image size (see Hong et al. (2019)). AP is useful in determining the performance of an object detector in localizing a target instance and correctly regressing a bounding box around said instance; however, it does not inform us of counting accuracy.

To determine the counting accuracy of each trained model, we use three counting metrics: mean absolute error (MAE), root mean squared error (RMSE) and mean percent error (MPE). Assume we have an evaluation dataset with N images, with \hat{C}_i predicted sandhill cranes and C_i annotated sandhill cranes in image i . Then, these three metrics are defined as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{C}_i - C_i| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{C}_i - C_i)^2} \quad (3)$$

$$MPE = \frac{1}{N} \sum_{i=1}^N 100 \times \left| \frac{\hat{C}_i - C_i}{C_i} \right| \quad (4)$$

MAE is a measure of absolute, or raw, counting error, and is commonly used in object counting contexts (Gao et al., 2020). RMSE provides additional insight into the counting accuracy of the models, with higher absolute errors penalized more heavily than in MAE. MPE describes relative error, enabling comparison of counting performance across models and studies. MPE explicitly accounts for the magnitude of the annotated count C_i . For Faster R-CNN, performance as measured by these three metrics depends strongly on choices for key hyperparameters, such as the IoU threshold and score threshold used during inference.

To determine if counting errors in test data were associated with spatial resolution when applying both the ASPDNet and Faster R-CNN models, we evaluated model performance on the test dataset. Relative error per observation (*i.e.* percent error divided by 100) for each predicted count was used as the response variable, and spatial resolution, measured in GSD (cm), was the single covariate. We used a beta regression with a logit link because our dependent variable was a continuous proportion (Douma & Weedon, 2019; Ferrari & Cribari-Neto, 2004). Beta regression is similar to a Generalized Linear Model, but the beta regression is suited for data bounded by 0 and 1 and is flexible in terms of the shape of data distribution (Douma & Weedon, 2019). Because

beta regression cannot contain zeros as a proportion, we adjusted one value that had 0.0000 relative error to a near-equivalent value of 0.0001 relative error. Wald tests were used to determine statistical differences ($\alpha = 0.05$), and beta regressions were conducted in R (R Core Team, 2020) using the 'betareg' package (Cribari-Neto & Zeileis, 2010).

Results

Final image dataset and annotations

In total, 63,539 sandhill cranes were identified across the 40 images (Table 1 and Fig. 4). Annotated bounding boxes were between 2 and 176 pixels in area, with the average sandhill crane occupying 29 pixels. Variation in bounding box size was largely a function of spatial resolution, with the largest bounding boxes being rare instances of sandhill cranes in flight. On average, each parent image had 1,589 sandhill cranes, ranging between 116 and 4,292 (Fig. 5). Annotation time for each parent image varied based on annotator and number of sandhill cranes in the image, but was generally between 1 and 3 hours per parent image. Ambiguous signatures requiring inter-annotator communication were identified in only two images.

Model inference and evaluation

Overall, the two DL models showed reasonable accuracy in predicted counts (Table 1), with MPE 9–18% (Table 2). Most errors were due to models undercounting the actual number of sandhill cranes present (Fig. 6). When accounting for the various spatial resolutions in the test set, ASPDNet had <50% of the error of Faster R-CNN, as measured by RMSE, MAE and MPE (Table 2).

Relative error differed significantly across spatial resolutions for both ASPDNet (beta regression, 3 degrees of freedom, $z = 2.33$, $P = 0.02$, $\beta = 0.13 \pm 0.06$ standard error, pseudo r-squared = 0.23) and Faster R-CNN (beta

regression, 3 degrees of freedom, $z = 3.52$, $P < 0.001$, $\beta = 0.21 \pm 0.06$ standard error, pseudo r-squared = 0.27). In both cases, error in counts increased when spatial resolution was more coarse, and the association was stronger for Faster R-CNN than for ASPDNet (Figs. 6 and 7).

Discussion

Overview

We faced several challenges in applying DL methods to automate counting of roosting sandhill cranes: target objects were relatively small with respect to image GSD, some occupying only a few pixels and targets were very dense, which are known challenges for DL-based object counting approaches (Gao et al., 2021; Pang et al., 2019; Pham et al., 2020). Despite these challenges, both DL models demonstrated reasonably high counting accuracy. For context, Johnson et al. (2010) compared sandhill crane aerial observer counts with manual counts from aerial photos; they found aerial observers underestimated abundance by 21.3% on average, with a range in observer flock counts differing from photography by –56.9% to +25%.⁵ The framework we developed could be applied to other broad-scale surveys of medium to small-bodied birds where image resolution and bird aggregation density are of concern. We recognize, however, that our specific results are conditioned on the characteristics and capabilities of the midwave thermal sensors we used, the environmental conditions experienced during image acquisition and the thermal response of sandhill crane targets in relation to complexity and emissivity of the background.

Developing a framework for automating sandhill crane counts

Methodologies for application of thermal imagery for large mammal monitoring are relatively well established, but application of this technology to bird monitoring over large areas has been considered problematic because of the coarse resolution of thermal sensors and the small body size of birds (Chabot & Francis, 2016). More recent studies with state-of-the-art sensors have utilized thermal imagery to count penguins (Bird et al., 2020) and identify individual nesting birds (McKellar et al., 2021; Santangeli et al., 2020) over relatively limited areas (*i.e.* tens to hundreds of meters). Sandhill cranes have been surveyed using thermal aerial sensors in the past, but the low spatial resolution of the thermal sensors used resulted in difficulty discriminating crane signatures when they were located on less emissive overland backgrounds and in difficulty distinguishing individuals within a group from

Table 1. The number of images, number of annotated sandhill crane signatures and total counts produced by both convolutional neural network models (ASPDNet and Faster R-CNN) for the train, validation and test splits as well as the full dataset.

Dataset split	Number of images	Number of annotations	ASPDNet total count	Faster R-CNN total count
Train	24	36,835	37,766	28,988
Validation	4	6,682	6,425	5,905
Test	12	20,022	19,473	15,713
Full Dataset	40	63,539	63,664	50,606

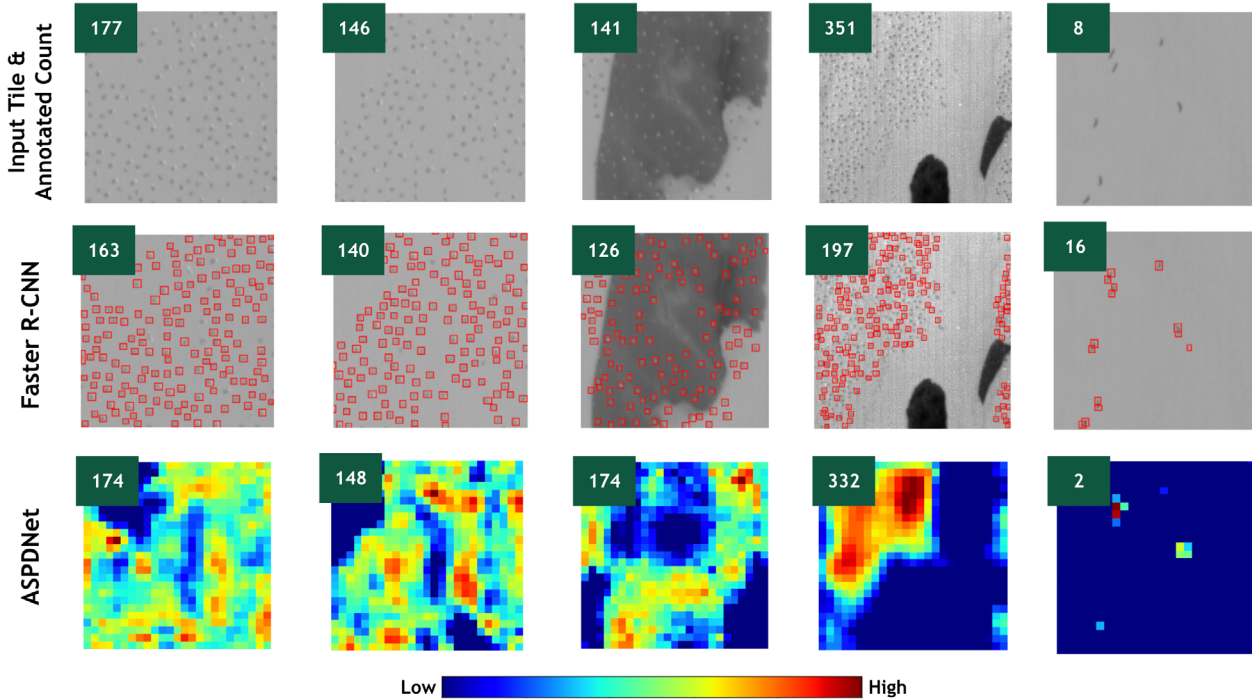


Figure 4. A sampling of 200×200 pixel input tiles derived from our test imagery with each column also containing the annotated count and predictions for both deep learning models. Counts are included in the top left corner for each cell of the grid. Depicted are both high-quality predictions (predicted count is very close to annotated count) and tiles with high prediction error (large difference between predicted and annotated). In predicted density maps, densities range from low (dark blue) to high (dark red), but the scale differs between tiles.

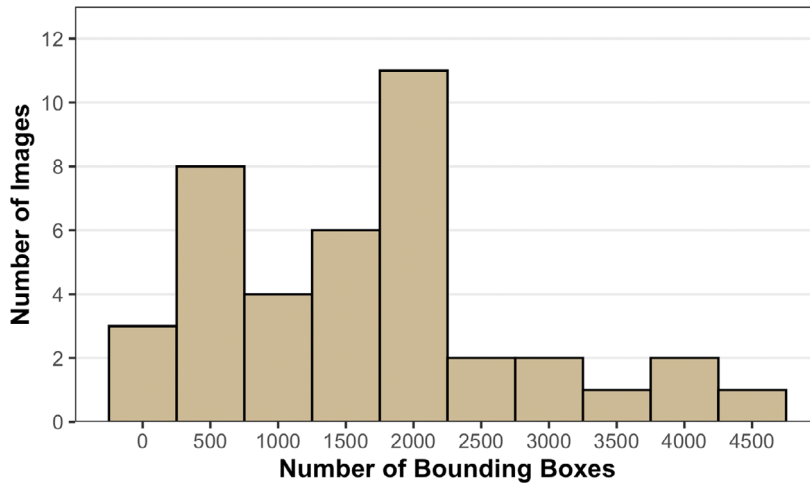


Figure 5. The frequency distribution of sandhill crane bounding boxes annotated per image for all 40 parent images in the dataset.

higher acquisition altitudes (Kinzel et al., 2006). The mid-wave sensors and lenses we used enabled discrimination of individual cranes in dense roosting aggregations from higher acquisition altitudes and over a wider range of background emissivity.

Commonly, a spectral threshold approach is taken to distinguish and count individual animals in imagery (Bird

et al., 2020; Chabot & Francis, 2016). However, this technique requires substantial manual calibration to account for diverse thermal signatures and spectral variation, potentially limiting applicability of the resulting algorithm. In our study, DL using CNNs has the advantage of automatically learning complex relationships among the thermal response, shape and size of features without the

Table 2. The performance of two convolutional neural network models, ASPDNet and Faster R-CNN, for predicting counts of sandhill cranes in thermal imagery. RMSE is root mean squared error, MPE is mean percent error, MAE is mean absolute error and AP is average precision. AP was calculated at an intersection-over-union threshold of 0.3. To contextualize MAE: the test imagery had on average 1,665 sandhill cranes per image (standard deviation = 1,242).

Model	RMSE	MAE	MPE	AP
ASPDNet	212.14	138.18	9%	-
Faster R-CNN	562.49	367.92	18%	76.03

need for manual and explicit understanding of feature characteristics as needed for a threshold approach. In addition, CNNs can be completely automated after model training. Although we do not address it here, operational implementation of our methods would require that imagery be mosaicked before tiling (see Fig. 3) to avoid double-counting individuals in the overlapping regions of contiguous imagery. This can be readily integrated using existing mosaicking implementations; see, for example, Borowicz et al. (2018) or Kellenberger et al. (2021).

Comparison of two deep learning approaches: Faster R-CNN and ASPDNet

The accuracy of the two DL models appeared to differ according to counting metrics with the density estimator,

ASPDNet, having <50% of the error estimated for the object detector, Faster R-CNN, as measured with RMSE, MAE and MPE. The MPE of 9% for ASPDNet is similar to previous bird monitoring studies using DL with very high resolution visible spectrum imagery. Hong et al. (2019) tested a variety of models and reported MPE on counts ranging from 3.5–10.5%. The undercounting of sandhill cranes was the primary source of inaccuracy for both our models, which can be partially explained by image spatial resolution (see below). Our study is the first to incorporate ASPDNet in a wildlife application, and our results indicate ASPDNet can provide more accurate counts (compared to Faster R-CNN) in images with high densities of roosting sandhill cranes. In addition, the performance of ASPDNet was less affected by image spatial resolution over the range of GSDs we examined. Given the low MPE for both models, future application of these methods as a replacement for manual counting seems promising, as long as future imagery remains consistent with the context, or domain, of the training imagery (see Kellenberger et al. (2019)). Additionally, careful inspection of model performance when applied to a full imagery dataset, composed of thousands of images, is important to ensure models do not generate excessive false positives, as has been observed in previous monitoring studies (Kellenberger et al., 2018). Future modeling efforts may benefit from experimentation with recent object detection algorithms designed for small, dense targets in remote

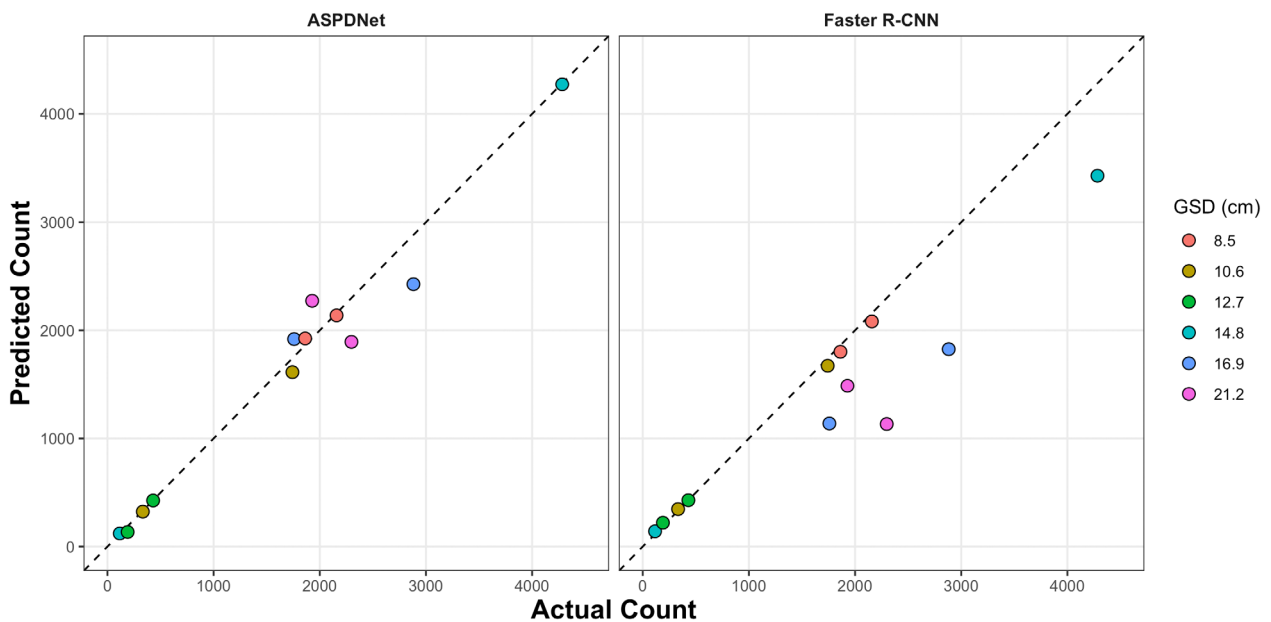


Figure 6. Sandhill crane actual versus predicted counts based on thermal imagery and convolutional neural network (CNN) models. Counting errors are depicted as distance from the 1:1 line with the ASPDNet architecture (left) and Faster R-CNN architecture (right) for each parent image in the test set. The dashed line shows where predicted counts match actual counts.

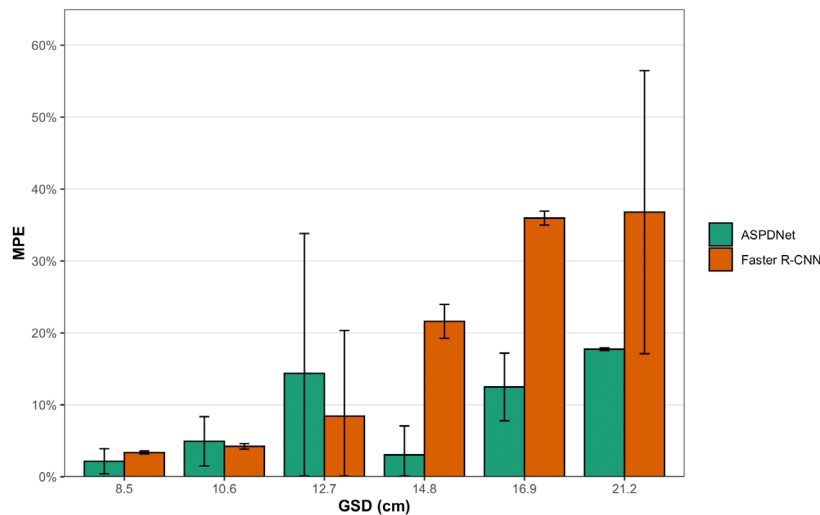


Figure 7. Mean percent error (MPE) of sandhill crane counts in thermal imagery surveys. Bars represent predictions from two deep learning models, ASPDNet (green) and Faster R-CNN (orange). Two samples were available at each available ground sample distance (GSD), and error bars show ± 1 standard deviation.

sensing imagery (*i.e.* Pham et al. (2020) or Pang et al. (2019)) alongside further exploration of density estimation models (*i.e.* Liu et al. (2020) or Wen et al. (2021)).

The effect of spatial resolution on counting error

We found that image GSD was positively, and linearly, associated with counting error in imagery ranging from 8.5 to 21.2 cm GSD. To our knowledge, only one other study has examined the effects of image spatial resolution on the performance of DL models applied to wildlife survey: Santangeli et al. (2020) found that false negatives in detections of bird nests from thermal imagery collected using an unoccupied aerial vehicle increased significantly from a 25 m to 50 m flight altitude. Spatial resolution in aerial sensors is determined by a combination of hardware and the distance AGL when imagery is obtained. Although imagery could be obtained at a finer resolution closer to ground level, this is balanced against a narrower field-of-view, the safety of the aircraft, and potential disturbances to the wildlife targets. For example, Kinzel et al. (2006) obtained infrared thermal imagery of sandhill cranes at 1,200 m during a first pass to ensure a broad field-of-view, but their methodology required a second imaging pass at a 300 m altitude to produce image resolutions suitable for identification of individuals. Our study indicates for thermal imaging of sandhill cranes at night, an image GSD of ≤ 14.8 cm is necessary to obtain accurate counts and imagery of ≤ 10.6 cm is ideal.

However, a larger sample size would be helpful to refine recommendations and to identify any specific thresholds that affect the accuracy of DL models. Given the ubiquitous importance of the effect of spatial resolution on DL model performance in wildlife survey applications, further study on an array of species and contexts would be beneficial. In addition, land and water temperature, cloud cover, time of survey, substrate type and other factors likely affect model counting accuracy (Kinzel et al., 2006; Santangeli et al., 2020), and additional research to support more general conclusions would be helpful.

Conclusions and Future Directions

We developed an approach to automate counts of roosting sandhill cranes at night at a key staging area during spring migration. Through application of two DL models to thermal imagery, we found that the density estimator, ASPDNet, exhibited substantially lower counting error compared to the object detector, Faster R-CNN. Both models performed particularly well when data were collected at a relatively fine spatial resolution (≤ 14.8 cm GSD). Our methods have the potential for broader application to improve the efficiency of surveys of densely aggregated birds. This could include cranes (family: *Gruidae*), which spend considerable time aggregated at migratory staging areas around the world (Archibald & Meine, 1996), or waterbird aggregations, such as shorebirds, swans, or waterfowl, at their staging areas. However, bird species identity may be difficult to discern from coarse resolution thermal imagery. Coincident ground

observations may be necessary to positively identify bird species in thermal imagery and to evaluate image characteristics necessary for species determination, and is an important area for further research.

Acknowledgments

We thank Brian Lubinski (U.S. Fish and Wildlife Service, retired), Garrett Wilkerson (U.S. Fish and Wildlife Service), Larry Robinson (U.S. Geological Survey, retired), Benjamin Finley, David Brandt, and Andrew Strassman (U.S. Geological Survey), and members of the Crane Trust for their efforts in collecting, curating and processing the data. We thank Anna Bragger (University of Wisconsin-La Crosse) for helping with image annotation. We thank Aaron Pearse (U.S. Geological Survey) for giving feedback on an early draft of the manuscript and thank the three anonymous peer reviewers as well as the U.S. Geological Survey internal reviewer for helping to improve the manuscript during the revision process. The first author acknowledges funding and logistical support from the Institute for Integrative Conservation at William & Mary, and thanks Rob Rose and Erica Garrouette for mentorship throughout.

The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the U.S. Fish and Wildlife Service but do represent the views of the U.S. Geological Survey. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Endnotes

1 All data and metadata used in this study is publicly available at <https://doi.org/10.5066/P9DZKFKQ3> (Lubinski et al., 2022). All code used to produce models and analyses is available in the following public GitHub repository: <https://github.com/emiliolr/counting-cranes>.

2 <https://github.com/tzutalin/labelImg>.

3 <https://pytorch.org/>.

4 <https://alumentations.ai/>.

5 Observer error in Johnson et al. (2010) is expressed as a ratio of the flock count from imagery (equivalent to the annotated image count in our study) to the observer's estimated flock count. Formally, this is $R = C_i / \hat{C}_i$, which we convert to a percent error as $PE = 100 \cdot (R^{-1} - 1)$ to facilitate comparison of results.

References

Akçay, H. G., Kabasakal, B., Aksu, D., Demir, N., Oz, M., and Erdogan, A. (2020). Automated bird counting with deep learning for regional bird distribution mapping. *Animals*, **10** (7):1207.

- Archibald, G.W. & Meine, C.D. (1996) *The cranes: status survey and conservation action plan*. Gland, Switzerland: IUCN, p. 281.
- Arteta, C., Lempitsky, V. & Zisserman, A. (2016) Counting in the wild. In: *Computer Vision – ECCV 2016*. Cham, Switzerland: Springer International Publishing, pp. 483–498.
- Association of Fish and Wildlife Agencies' Migratory Shore and Upland Game Bird Support Task Force. (2016) *Priority information needs for sandhill cranes ii: A funding strategy*. Technical report. Washington, DC: Association of Fish and Wildlife Agencies.
- Benning, D. & Johnson, D. (1987) Recent improvements to sandhill crane surveys in Nebraska's Central Platte Valley. In: *Proceedings of the 1985 Crane Workshop*. Lake Wales, FL: National Audubon Society, pp. 10–16.
- Bird, C.N., Dawn, A.H., Dale, J. & Johnston, D.W. (2020) A semi-automated method for estimating Adélie penguin colony abundance from a fusion of multispectral and thermal imagery collected with unoccupied aircraft systems. *Remote Sensing*, **12**(22), 3692.
- Borowicz, A., McDowall, P., Youngflesh, C., Sayre-McCord, T., Clucas, G., Herman, R. et al. (2018) Multi-modal survey of Adélie penguin mega-colonies reveals the danger islands as a seabird hotspot. *Scientific Reports*, **8**(1), 3926.
- Caven, A.J., Buckley, E.M.B., King, K.C., Wiese, J.D., Baasch, D.M., Wright, G.D. et al. (2020) Temporospatial shifts in sandhill crane staging in the Central Platte River valley in response to climatic variation and habitat change. *Monographs of the Western North American Naturalist*, **11** (1), 33–76.
- Caven, A., Varner, D. & Drahotka, J. (2020) Sandhill crane abundance in Nebraska during spring migration: making sense of multiple data points. *Transactions of the Nebraska Academy of Sciences and Affiliated Societies*, **40**, 6–18.
- Central Flyway Webless Migratory Game Bird Technical Committee. (2018) *Management guidelines for the midcontinent population of sandhill cranes*. Technical report. Lakewood, CO: U.S. Fish and Wildlife Service, Division of Migratory Bird Management, https://pacificflyway.gov/Documents/Msc_plan.pdf.
- 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.
- Chattopadhyay, P., Vedantam, R., Selvaraju, R.R., Batra, D. & Parikh, D. (2017) Counting everyday objects in everyday scenes. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA: IEEE Computer Society, pp. 4428–4437.
- 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), 3208.
- Corcoran, E., Winsen, M., Sudholz, A. & Hamilton, G. (2021) Automated detection of wildlife using drones: synthesis,

- opportunities and constraints. *Methods in Ecology and Evolution*, **12**(6), 1103–1114.
- Cribari-Neto, F. & Zeileis, A. (2010) Beta regression in R. *Journal of Statistical Software*, **34**(2), 1–24.
- Douma, J.C. & Weedon, J.T. (2019) Analysing continuous proportions in ecology and evolution: a practical introduction to beta and Dirichlet regression. *Methods in Ecology and Evolution*, **10**(9), 1412–1430.
- Duporge, I., Isupova, O., Reece, S., Macdonald, D.W. & Wang, T. (2021) Using very-high-resolution satellite imagery and deep learning to detect and count African elephants in heterogeneous landscapes. *Remote Sensing in Ecology and Conservation*, **7**(3), 369–381.
- Ferrari, S. & Cribari-Neto, F. (2004) Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**(7), 799–815.
- Gao, G., Gao, J., Liu, Q., Wang, Q. & Wang, Y. (2020) CNN-based density estimation and crowd counting: a survey. *arXiv preprint arXiv:2003.12783*.
- Gao, G., Liu, Q. & Wang, Y. (2021) Counting from sky: a large-scale dataset for remote sensing object counting and a benchmark method. *IEEE Transactions on Geoscience and Remote Sensing*, **59**(5), 3642–3655.
- 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), 14259.
- He, K., Zhang, X., Ren, S. & Sun, J. (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Las Vegas, NV: IEEE, pp. 770–778.
- Hong, S., Han, Y., Kim, S., Lee, A. & Kim, G. (2019) Application of deep-learning methods to bird detection using unmanned aerial vehicle imagery. *Sensors*, **19**(7), 1651.
- Johnson, D., Solberg, J. & Amundson, C. (2010) Countability of sandhill cranes in aerial surveys. In: *Proceedings of the Eleventh North American crane workshop*. Baraboo, WI: North American Crane Working Group, pp. 89–97.
- Kellenberger, B., Marcos, D., Lobry, S. & Tuia, D. (2019) Half a percent of labels is enough: efficient animal detection in UAV imagery using deep CNNs and active learning. *IEEE Transactions on Geoscience and Remote Sensing*, **57**(12), 9524–9533.
- Kellenberger, B., Marcos, D. & Tuia, D. (2018) Detecting mammals in UAV images: best practices to address a substantially imbalanced dataset with deep learning. *Remote Sensing of Environment*, **216**, 139–153.
- Kellenberger, B., Veen, T., Folmer, E. & Tuia, D. (2021) 21 000 birds in 4.5 h: efficient large-scale seabird detection with machine learning. *Remote Sensing in Ecology and Conservation*, **7**(3), 445–460.
- Kim, S. & Kim, M. (2020) Learning of counting crowded birds of various scales via novel density activation maps. *IEEE Access*, **8**, 155296–155305.
- Kinzel, P.J., Nelson, J.M., Parker, R.S. & Davis, L.R. (2006) Spring census of mid-continent sandhill cranes using aerial infrared videography. *Journal of Wildlife Management*, **70**(1), 70–77.
- Kirby, J.S., Stattersfield, A.J., Butchart, S.H.M., Evans, M.I., Grimmett, R.F.A., Jones, V.R. et al. (2008) Key conservation issues for migratory land- and waterbird species on the world's major flyways. *Bird Conservation International*, **18** (S1), S49–S73.
- LeCun, Y., Bengio, Y. & Hinton, G. (2015) Deep learning. *Nature*, **521**(7553), 436–444.
- Lempitsky, V. & Zisserman, A. (2010) *Learning to count objects in images, NIPS 2010*. Red Hook, NY: Curran Associates Inc., pp. 1324–1332.
- Lin, T.Y., Goyal, P., Girshick, R., He, K. & Dollár, P. (2017) Focal loss for dense object detection. In: *Proceedings of the IEEE international conference on computer vision*. Venice, Italy: IEEE, pp. 2980–2988.
- Liu, L., Jiang, J., Jia, W., Amirgholipour, S., Wang, Y., Zeibots, M. et al. (2020) Denet: a universal network for counting crowd with varying densities and scales. *IEEE Transactions on Multimedia*, **23**, 1060–1068.
- Lubinski, B., Robinson, L.R., Finley, B.C., Wilkerson, G., Strassman, A.C., Baker, A., et al. (2022) *Aerial thermal imagery of the Central Platte River Valley and bounding box annotations of sandhill cranes*. U.S Geological Survey data release. <https://doi.org/10.5066/P9DZKFQ3>
- McKellar, A.E., Shephard, N.G. & Chabot, D. (2021) Dual visible-thermal camera approach facilitates drone surveys of colonial marshbirds. *Remote Sensing in Ecology and Conservation*, **7**(2), 214–226.
- Nichols, J.D. & Williams, B.K. (2006) Monitoring for conservation. *Trends in Ecology and Evolution*, **21**(12), 668–673.
- Padilla, R., Passos, W., Dias, T., Netto, S. & da Silva, E. (2021) A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics*, **10**(3), 279.
- Pang, J., Li, C., Shi, J., Xu, Z. & Feng, H. (2019) R²-cnn: fast tiny object detection in large-scale remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, **57**(8), 5512–5524.
- Pearse, A., Gerard, P., Dinsmore, S., Kaminski, R. & Reinecke, K. (2008) Estimation and correction of visibility bias in aerial surveys of wintering ducks. *Journal of Wildlife Management*, **72**(3), 6.
- Pearse, A.T., Krapu, G.L., Brandt, D.A. & Sargeant, G.A. (2015) Timing of spring surveys for midcontinent sandhill cranes. *Wildlife Society Bulletin*, **39**(1), 8793.
- Pham, M., Courtrai, L., Friguet, C., Lefevre, S. & Baussard, A. (2020) Yolo-fine: one-stage detector of small objects under various backgrounds in remote sensing images. *Remote Sensing*, **12**(15), 2501.
- R Core Team. (2020) *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

- Rakhimberdiev, E., Duijns, S., Karagicheva, J., Camphuysen, C.J., Dekinga, A., Dekker, R. et al. (2018) Fuelling conditions at staging sites can mitigate Arctic warming effects in a migratory bird. *Nature Communications*, **9**(1), 4263.
- Redmon, J. & Farhadi, A. (2018) YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.
- Ren, S., He, K., Girshick, R. & Sun, J. (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS 2015*. Cambridge, MA: MIT Press, pp. 91–99.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S. et al. (2015) ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, **115**(3), 211–252.
- Santangeli, A., Chen, Y., Klun, E., Chirumamilla, R., Tiainen, J. & Loehr, J. (2020) Integrating drone-borne thermal imaging with artificial intelligence to locate bird nests on agricultural land. *Scientific Reports*, **10**(1), 10993.
- Sasse, D.B. (2003) Job-related mortality of wildlife workers in the United States, 1937–2000. *Wildlife Society Bulletin (1973–2006)*, **31**(4), 1015–1020.
- Seymour, A.C., Dale, J., Hammill, M., Halpin, P.N. & Johnston, D.W. (2017) Automated detection and enumeration of marine wildlife using unmanned aircraft systems (UAS) and thermal imagery. *Scientific Reports*, **7**(1), 45127.
- Simonyan, K. & Zisserman, A. (2015) Very deep convolutional networks for large-scale image recognition. In: *International Conference on Learning Representations*. San Diego, CA: ICLR.
- Weinstein, B.G. (2018) A computer vision for animal ecology. *Journal of Animal Ecology*, **87**(3), 533–545.
- Wen, L., Du, D., Zhu, P., Hu, Q., Wang, Q., Bo, L. et al. (2021) Detection, tracking, and counting meets drones in crowds: A benchmark. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Manhattan, NY: IEEE, pp. 7812–7821.
- Williams, B.K., Nichols, J.D. & Conroy, M.J. (2002) *Analysis and management of animal populations*. San Diego, CA: Academic Press, p. 817.