Remote Sensing in Ecology and Conservation
Open Access
ZSL LET'S WORK FOR WILDLIFE

ORIGINAL RESEARCH

# 21 000 birds in 4.5 h: efficient large-scale seabird detection with machine learning

Benjamin Kellenberger[1,2], Thor Veen[3,4], Eelke Folmer[4] & Devis Tuia[2]

[1]Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, Wageningen, The Netherlands
[2]Environmental Computational Science and Earth Observation Laboratory (ECEO), EPFL, Sion, Switzerland
[3]Quest University, Squamish, Canada
[4]Aeria Solutions Ltd., Squamish, Canada

**Keywords**
coastal birds, convolutional neural network, deep learning, remote sensing, unmanned aerial vehicle, wildlife census

**Correspondence**
Benjamin Kellenberger, Environmental Computational Science and Earth Observation Laboratory (ECEO), Ecole Polytechnique Fédérale de Lausanne (EPFL), Sion, Switzerland. Tel: +41 21 693 80 12; E-mail: benjamin.kellenberger@epfl.ch

Editor: Ned Horning
Associate Editor: Kylie Scales

Received: 17 September 2020; Revised: 15 January 2021; Accepted: 19 February 2021

doi: 10.1002/rse2.200

## Abstract

We address the task of automatically detecting and counting seabirds in unmanned aerial vehicle (UAV) imagery using deep convolutional neural networks (CNNs). Our study area, the coast of West Africa, harbours significant breeding colonies of terns and gulls, which as top predators in the food web function as important bioindicators for the health of the marine ecosystem. Surveys to estimate breeding numbers have hitherto been carried out on foot, which is tedious, imprecise and causes disturbance. By using UAVs and CNNs that allow localizing tens of thousands of birds automatically, we show that all three limitations can be addressed elegantly. As we employ a lightweight CNN architecture and incorporate prior knowledge about the spatial distribution of birds within the colonies, we were able to reduce the number of bird annotations required for CNN training to just 200 examples per class. Our model obtains good accuracy for the most abundant species of royal terns (90% precision at 90% recall), but is less accurate for the rarer Caspian terns and gull species (60% precision at 68% recall, respectively 20% precision at 88% recall), which amounts to around 7% of all individuals present. In sum, our results show that we can detect and classify the majority of 21 000 birds in just 4.5 h, start to finish, as opposed to about 3 weeks of tediously identifying and labelling all birds by hand.

## Introduction

Preservation of biodiversity is of great importance for the maintenance of healthy ecosystems and for human well-being (Cardinale et al., 2012; UN sustainable development goals[1]). The implementation of effective conservation strategies is particularly pressing given the accelerating rates of decline of global biodiversity (Butchart et al., 2010) and is tied to means of measuring biodiversity through indicators, in particular to identify possible threats and for taking effective conservation measures. One such indicator is the abundance and distribution of seabirds, which has the potential to provide a comprehensive measure for ecosystem health, including the lower levels of the food chain (Gregory et al., 2003; Parsons et al., 2008). Many seabird species breed in large colonies that require expansive efforts to count manually, resulting in highly uncertain census estimations. To cope with these issues, recent technological advances are increasingly explored to accelerate and improve the accuracy of monitoring (Andrew & Shephard, 2017; Edney & Wood, 2020; Terletzky & Ramsey, 2016). On the one hand, these advances include unmanned aerial vehicles (UAVs, or 'drones') that acquire high-resolution aerial imagery of breeding colonies from a distant viewpoint. This forgoes the need for entering bird colonies on foot and permits obtaining precise geospatial coordinates of every visible individual. On the other hand, these advances also include automated detection through machine learning, which offers the possibility of detecting and classifying large numbers of birds with reduced manual efforts and

[1]https://sustainabledevelopment.un.org/post2015/transformingourworld/publication

time requirements. In the following two sections, we briefly introduce coastal seabirds as our target and manual census methods (Section 1.1), followed by automated detection with machine learning (Section 1.2).

## Coastal seabirds and manual counting

Coastal ecosystems are at the intersection of terrestrial and aquatic regions and therefore feature high biodiversity (Costanza et al., 1993). The coast of West Africa is productive and biodiverse due to the nutrient-rich oceanic upwelling (Camphuysen & Van der Meer, 2005) and its large estuaries (Baran, 2000). The region is rich in seabirds, which play an important role in the ecosystem as they are top predators in the marine food web. Seabirds are good bioindicators for the health of marine ecosystems but they are challenging to monitor (Einoder, 2009; Parsons et al., 2008; Veen et al., 2018b). Moreover, monitoring seabirds is of importance given the threats of coastal erosion, human disturbance, and increased commercial fisheries (FAO, 2011). The number of breeding pairs is an important basic measurement to track population size over time and identify possible threats. Each nest has two parents who take turns in breeding. While one of the parents breeds, the other is foraging or resting outside of the colony; the number of breeding pairs therefore corresponds directly to the number of identified individuals. In West Africa, a seabird monitoring project has been ongoing for over two decades (Veen et al., 2004, 2018a, 2019) in which a variety of counting methods have been employed in the main breeding areas. Several areas consist of large colonies where birds cannot realistically be individually counted and hence need to be estimated. The most basic method involves scanning the colony with binoculars from a good vantage point. In a first step, the observer estimates the number in a given segment of the colony and subsequently uses this segment as a reference to estimate the numbers in the rest of the colony (the 'block method'[2]). This method is fast and does not require researchers to enter the colony and thus avoids disturbance. The disadvantage is that it is inaccurate depending on experience and hence may lead to large variance among observers, sometimes by almost an order of magnitude (Frederick et al., 2003). If the colony is small, the number of nests can be counted individually using a group of people entering the site. Oftentimes the colonies contain tens of thousands of breeding pairs; manual counts thus take too long and the disturbance can result in depredation and overheating of eggs and chicks (Carney & Sydeman, 1999). In such cases, the number of nests in the colony is estimated by measuring the area of the colony by GPS and by taking estimates of the nest density in quadrants[2]. This second family of methods provides a (much) better estimate of the number of breeding birds, but is more time consuming and requires more people. Furthermore, it involves foot access to the colonies and hence causes increased disturbance.

## Automated seabird mapping

The recent emergence of UAVs provided the opportunity to obtain high resolution images of the breeding colonies, which reduce the disturbance, time and number of people needed in comparison to the traditional survey methods (Ivošević et al., 2015; Linchant et al., 2015). Detecting and counting individuals in aerial images also potentially increases accuracy by orders of magnitude (Hodgson et al., 2016). However, the stage of photointerpretation can be prohibitively time-consuming if done manually (Kellenberger et al., 2018). One possible way to overcome the need for manual annotations, at least in parts, is by employing automated detection methods from computer vision (CV). CV methods, in particular deep learning models like convolutional neural networks (CNNs; Krizhevsky et al., 2012; LeCun et al., 2015), have been successfully applied to a variety of different wildlife detection tasks (Eikelboom et al., 2019; Gray et al., 2019; Hamilton et al., 2020; Kellenberger et al., 2018), including birds as primary targets (Akçay et al., 2020; Borowicz et al., 2018; Hong et al., 2019). However, the focus of these studies has mostly been on very sparsely distributed wildlife, or at best on areas with moderate abundances of individuals. Dense breeding colonies comprised of multiple species against a heterogeneous background provide a number of challenges to CV detectors: different species nest on different substrates and the individual counts vary significantly between species, with areas ranging from containing only single birds to large and dense colonies. These data properties are hypothesized to hamper the usage of off-the-shelf, CNN-based object detection models like the commonly used Faster R-CNN (Ren et al., 2015), due to the high spatial distribution imbalance. Furthermore, the high number of free parameters in conventional CNN architectures requires training on a large set of annotated images, which is antagonistic to our goal of reducing the workload of manual counting, for example, through photointerpretation.

In this paper, we address these problems and attempt to automate the individual detection and classification of seabirds using CNNs. Rather than solely aiming for high detection accuracy, we focus on reducing the manual workload required to train the automated detection

[2]https://www.birdlife.org/sites/default/files/attachments/FIB-Guide-suivi-oiseau-EN.pdf, accessed November 19, 2020

 

model in the first place. We do so by combining a lightweight CNN architecture for detection with prior knowledge about the local distribution patterns of the bird species. We demonstrate our approach on a series of high-resolution, UAV-derived imagery acquired in coastal marine environments in West Africa.

## Methods

### Study area and bird species

The upwelling of nutrient-rich waters makes the Sahelian Upwelling Marine Ecoregion exceptionally productive and the food abundance for seabirds high. However, the number of suitable breeding sites is limited. Along the West African coast between Mauritania and Guinea, only a small number of sites are suitable for breeding but they may hold very large numbers of breeding seabirds (Veen et al., 2004, 2018a, 2019). The main sites from north to south are Parc National du Banc d'Arguin, Parc National de la Langue de Barbarie, Parc National du Delta du Saloum, Bijol Islands (Tanji Bird Reserve), Réserve Ornithologique de Kalissaye, Bijagós Archipélago, Bantambur (near Jeta) and Iles Alcatraz and Naufrage (Fig. 1 left).

In May 2019, we conducted a census along the West African coast to obtain an accurate count of the number of African royal terns (*Thalasseus maximus*), the species in primary focus. We also counted co-occurring Caspian terns (*Hydroprogne caspia*), slender-billed gulls (*Chroicocephalus genei*) and grey-headed gulls (*Chroicocephalus cirrocephalus*). The selection of sites was mainly based on the presence of the African royal tern because all known breeding sites are confined to this region. Furthermore, because royal terns breed rather synchronously in large colonies on a small number of sites, an accurate estimate of the population size can be obtained during a relatively limited amount of time in May (Veen et al., 2003). Because Caspian terns, slender-billed gulls and grey-headed gulls often breed in the same areas, these species were also counted.

### Image acquisition

We used a DJI Phantom 4 Pro to take photographs of the breeding colonies, which has several characteristics that make it very suitable to map seabird colonies. It is quite robust to high wind speeds (up to 10 m/s) and it provides a sufficient flight time of approximately 30 minutes. The camera has a high-resolution (20 Megapixels) 1" CMOS sensor and an 84° field of view lens that provides wide coverage. Its relatively small size makes it easy to transport and it can take off and land in small spaces.

The size of the colony and time available for the survey determines the flight altitude (lower flights cover less area in a given time period). Flight altitude also affects the ground resolution obtained, with higher flights providing

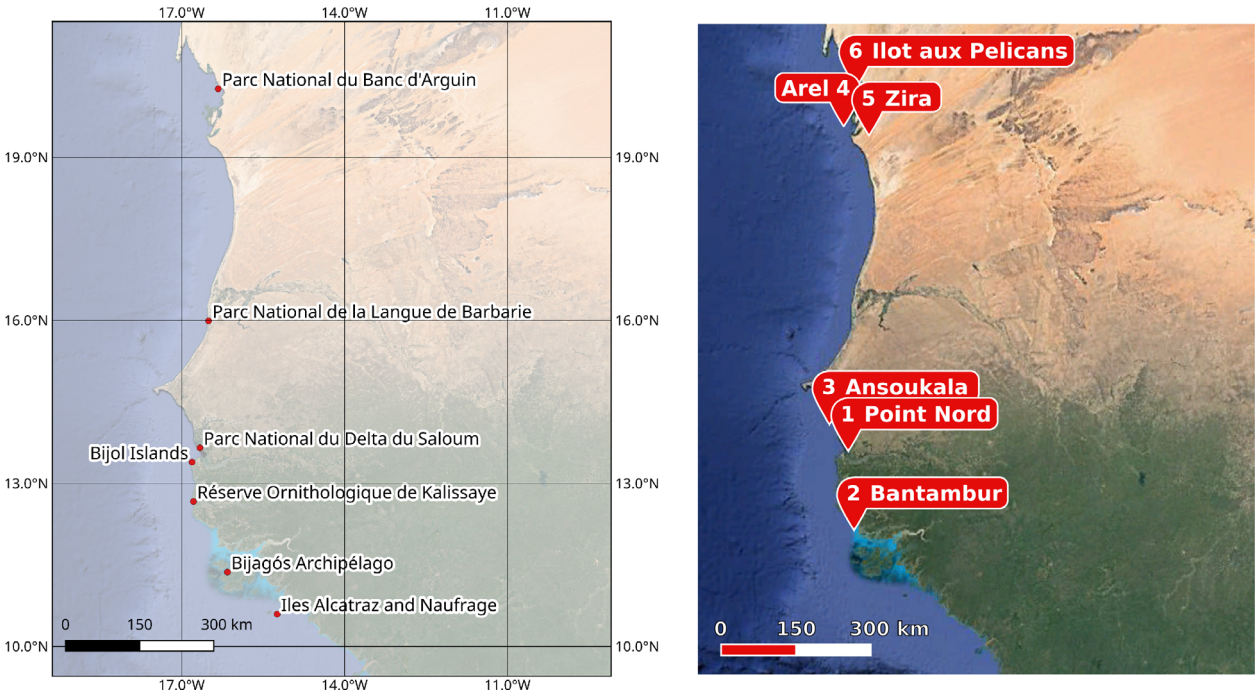


**Figure 1.** Important breeding sites for colonial seabirds along the West African coast (left) and locations of orthomosaics (right).

 

**Table 1.** Details of the six orthomosaics from the UAV survey.

| | | Approx. centre coordinates | | Dimensions | |
|---|---|---|---|---|---|
| Number | Name | Latitude | Longitude | Metres | Pixels |
| 1 | Point Nord | 13.664 | −16.661 | 292.48 × 305.69 | 27 569 × 28 814 |
| 2 | Bantambur | 11.972 | −16.303 | 514.87 × 345.41 | 42 877 × 28 765 |
| 3 | Ansoukala | 13.696 | −16.674 | 223.36 × 215.33 | 21 284 × 20 518 |
| 4 | Arel | 19.901 | −16.504 | 93.50 × 198.25 | 10 194 × 21 615 |
| 5 | Zira | 19.870 | −16.296 | 224.89 × 438.74 | 18 837 × 36 749 |
| 6 | Ilot aux Pelicans | 20.712 | −16.683 | 138.69 × 110.77 | 10 521 × 8403 |

less detailed imagery. We used flight altitudes varying between 20 m and 50 m, which provided high-resolution images allowing for clear identification of the breeding species without causing disturbance. The UAV survey team consisted of a pilot, who controls the UAV, and a visual observer, who surveys the environment for potential problems (e.g. aggressive birds) and notes possible changes in behaviour of the birds caused by the UAV. More specifically, the visual observer monitored the behaviour of the birds prior to take-off and during the flight. The birds do not appear to be disturbed by the UAV and only in rare occasions have we observed a behavioural response to the UAV. The flight altitude was increased in the rare case that behavioural changes were observed.

Colonies were mapped by flying parallel transects at speeds between 3 and 5 m/s at fixed altitude. The camera was set to take a picture every three seconds. The selected speeds, altitude and the distance between transects ensured sufficient overlap between pictures needed to produce high-quality RGB orthomosaics using photogrammetry software (Agisoft Metashape[3]). The aerial survey resulted in six orthomosaics with acquisition locations shown in Figure 1 (right) and numerical details in Table 1.

## Manual photointerpretation

Five orthomosaics were used for model training and one (orthomosaic 1) was retained for testing purposes (i.e. the final, independent evaluation of our results). The reason for this division was due to the circumstance that orthomosaic 1 was the only one that contained all bird species in sufficiently large numbers for testing.

We annotated the five training orthomosaics (2–6) with points for each species over patches of bird colonies using the vector editing tool in QGIS[4]. To reduce labelling efforts required, we only annotated a subset of the total number of birds in orthomosaics 2–6 and aimed at obtaining 200 training point annotations or more per species. To do so, we selected a number of clearly distinguishable colonies for each species and annotated all birds present in the areas covered by those colonies, but omitted labelling other areas of the orthomosaics. We also added background polygons indicating areas where no birds were present, graded into 'easy' background (i.e. with homogeneous surfaces), respectively 'hard' backgrounds that also contained clutter and other bird species (grey, respectively white polygons in Fig. 2). An example of the training annotations for orthomosaic 3 is shown in Figure 2.

For the test orthomosaic 1, we created a complete set of annotations for every individual bird we could detect (Fig. 3). To do so, we first divided the orthomosaic into 274 non-overlapping tiles of 800 × 600 pixels, organized on a regular grid. We then created point annotations with class labels in all image tiles using the open source software AIDE[5] (Kellenberger et al., 2020), which allows users to annotate images through a web interface while also recording statistics, such as the date and time an image has been viewed and the time required to provide an annotation. In sum, orthomosaic 1 was used to quantify the manual workload required to annotate the birds by hand, and also to assess the performance of our machine learning-based bird detection model, using the provided annotations as a ground truth.

Unfortunately, initial tests showed that our model is not able to distinguish between grey-headed and slender-billed gulls. The two species are also difficult to discern for humans (see Fig. 4 for an example). We therefore decided to merge the Grey-headed and Slender-billed gulls into a single 'gull' class for all subsequent analyses.

The total number of annotations per species and for each orthomosaic is listed in Table 2.

[5]https://github.com/microsoft/aerial_wildlife_detection

[5]https://github.com/microsoft/aerial_wildlife_detection

 

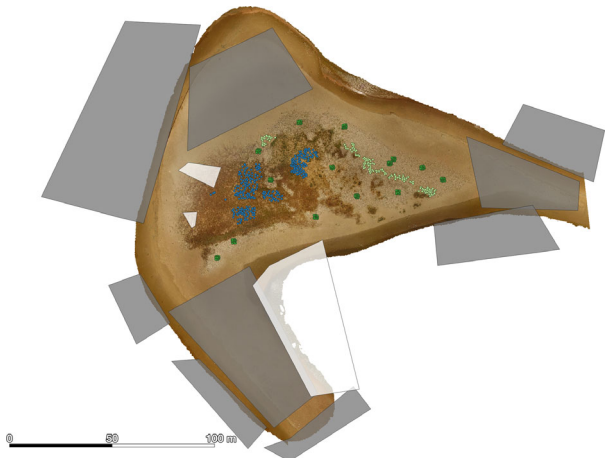


**Figure 2.** Orthomosaic 3 as an example for annotations used to train the detection model. Dark green points: royal terns; light green points: Caspian terns; blue points: grey-headed and slender-billed gulls. Polygons encompass areas without birds, classified into 'easy' (grey) and 'hard' (white) backgrounds.

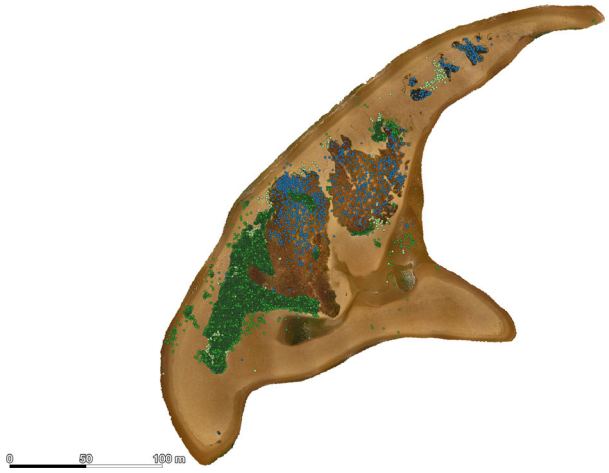


**Figure 3.** Test set orthomosaic 1 with manually annotated individuals per bird species. In total, 21 066 individuals were labelled in this scene.

**Table 2.** Number of manually annotated points per species for each of the six orthomosaics.

| Number | Name/set | Royal Tern | Caspian Tern | gulls |
|---|---|---|---|---|
| 1 | Point Nord | 19 683 | 442 | 941 |
| 2 | Bantambur | 450 | 3 | 499 |
| 3 | Ansoukala | 382 | 169 | 338 |
| 4 | Arel | 295 | 12 | 0 |
| 5 | Zira | 0 | 0 | 628 |
| 6 | Ilot aux Pelicans | 367 | 472 | 0 |
| 2,3,4 | training | 1127 | 184 | 837 |
| 5,6 | validation | 367 | 472 | 628 |
| 1 | test | 19 683 | 442 | 941 |

## Automated detections

### Model details

For automation of the birds' detection, we employed a deep CNN, as shown in Figure 5. CNNs perform sequences of operations, called layers, with each layer processing the output of the previous one. The most common operation in a CNN, the convolution operator, slides filters of predefined size (the first 'Conv' layer in the model, e.g. employs 64 filters, each one of size $7 \times 7$) over the image and calculates the dot product at each location, thereby outputting a spatial map of correlation values between the input and the filters, also called activations. The filter values are free parameters that are learned by the model by means of backpropagation (Rumelhart et al., 1995). To control the size of the outputs, the convolution operator can be applied not at every location in an input, but, for example at every second to produce an output of half the width and height of the input; the parameter that controls the frequency in width and height of operation is denoted as the 'stride'. An alternative operator used to perform spatial downsampling is 'max pooling', which returns the maximum value within the scanning window centred at each location ('Max pool'). In addition to spatial downsampling, max pooling

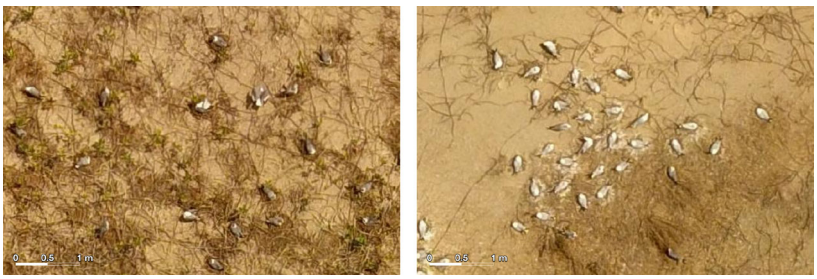


**Figure 4.** Species of the grey-headed gull (left) and slender-billed gull (right) are easily confused, and were therefore merged into a single class.

 

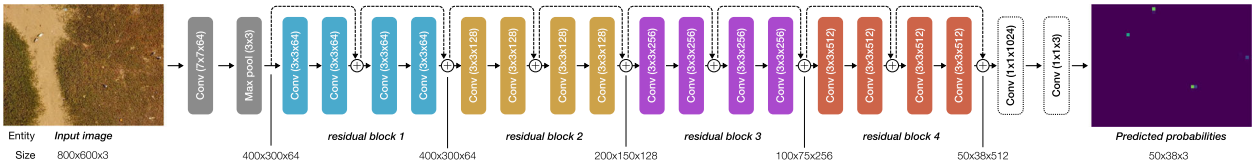


**Figure 5.** Simplified flow chart of the CNN architecture used in this work. The model is based on ResNet-18, which consists of four residual blocks (colour-coded) that contain skip connections (dashed lines). We modify the standard ResNet architecture and replace the last two layers with custom, 1 × 1 convolution operations (dotted rectangles) that map to 1024 channels and then to the three classes respectively. All other layers are initially pre-trained on ImageNet. Normalization and activation functions are omitted for clarity. Sizes of (intermediate) outputs are given in (Width × Height × Channels).

provides invariance of the model to translations (i.e. the precise location of a bird can vary in space to some extent and it still can be recognized by the model). Besides convolutions and pooling, we employed instance normalization (Ulyanov et al., 2016) and rectified linear unit (ReLU) operations after each convolution layer. For an in-depth explanation of the working mode of the different operators, we kindly refer the reader to Goodfellow et al. (2016); for a more remote sensing-based introduction, see Volpi and Tuia (2017).

Our particular model, sketched in Figure 5, is based on the ResNet-18 architecture (He et al., 2016). ResNet is widely used in computer vision tasks and has shown high performance in image classification, partially due to the organization of intermediate layers into 'residual blocks' (each residual block is represented by a different colour in Fig. 5) and skip connections (dashed lines). By default, ResNet-18 only predicts a single label per image, which it does by averaging the output of the last residual block in width and height, followed by a fully connected layer, which linearly maps the 512 output channels into the desired number of classes. Contrary to this classical approach, we require spatial predictions rather than single classification scores for the entire patch. Therefore, we removed the average pooling and fully connected layers and replaced them with two 1 × 1 convolution layers, which retain the spatial output size of 50 × 38, but map from the 512 channels of the last residual block to an intermediate 1024 dimensions, and then to 3 for the three bird species. To further adapt the prediction resolution to our needs, we also halved the stride of the very first convolution layer to one, which doubles the predicted size. With that, the full model yields a probability grid with a moderately reduced resolution, where one grid cell approximately corresponds to 16 × 16 cm. We found this to be sufficiently fine for detecting birds in our images. Prior to training the CNN, the weights were initialized with those of a model pre-trained on the ImageNet classification challenge (Russakovsky et al., 2015). This is a common means of warm-starting the model with parameters that are optimized to image recognition tasks. Although ImageNet does not contain overhead images, initializing the model this way has shown to accelerate model training and to lead to improved results, despite the gap between tasks and data sets (Huh et al., 2016). The final output of the model is passed through a sigmoid activation function, which scales every class value per grid cell between zero and one. This way, we can include the 'background' class implicitly by encouraging the model to predict all bird class outputs as zeros when there is no bird; to do so, during prediction we threshold the probability scores and consider all predictions below the threshold as background.

## Model training

For training, we selected a subset of patches from the five orthomosaics in random order and made sure that the total number of training point annotations reached 200 per species – in other words, we selected training patches at random so that each bird species was represented by more or less 200 points. This required users to annotate a total of 600 points for the three species, which can be done in a reasonable amount of time. We set a deliberately low target of 200 points per species to assess the feasibility of training a high-capacity CNN with scarce amounts of training data, but in favour of greatly reduced annotation expenses. We also experimented with fewer points, but found model training to be unstable (see Appendix below). For an assessment of the training stability, we used three different random seeds for the patch order permutation and trained three models accordingly.

Since we split the training images into tiles on a regular grid, these generally do not coincide with how the birds would be labelled. Particularly, in this way, a training patch may cut through a colony of birds. Moreover, this means that an image tile may not be fully annotated, but instead only, for example contains a few labels of a 'cut' bird colony at the border of the tile. Hence, we cannot automatically assign non-annotated pixels as 'background', as we risk assigning unlabelled birds to background. To address this problem, we implemented three modifications as described below.

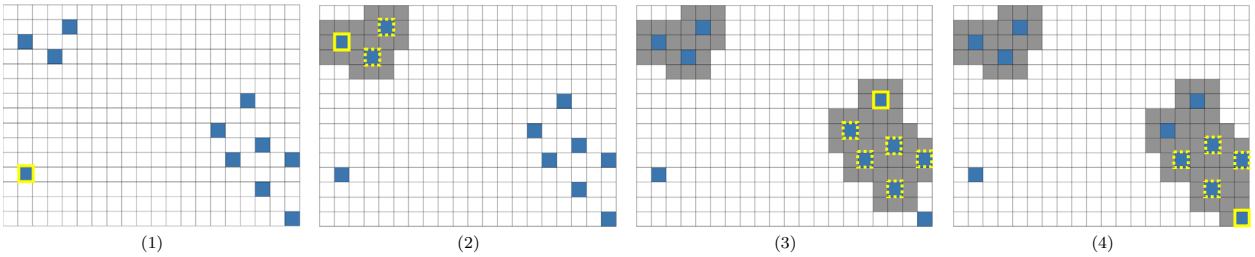


**Figure 6.** Conceptualized example of our convex hull heuristic. Our ground truth points are mapped to a grid the size of the CNN output. We iterate through all points one by one (yellow solid), and locate all other ground truth points (yellow dashed) that fall within a range of eight grid cells distance around the current one. We then draw a convex hull around those selected points and dilate it with a square, 3 × 3 filter. Finally, we assign all empty grid cells within the union of all convex hulls to background (grey). This process is repeated for all species separately.

1 We included background polygons (grey areas in Fig. 2) into the training process to teach the model what background clutter looks like. Background polygons are rasterized and rescaled to the same dimensions as the CNN predictions of 50 × 38 grid cells. We empirically found the model to be highly sensitive to the amount and complexity of background areas; exposing the model to all background polygons from the start generally resulted in a significant drop in bird recall. To mitigate this effect, we resorted to curriculum learning (Bengio et al., 2009), which has been shown to be beneficial for aerial wildlife detection tasks with deep learning (Kellenberger et al., 2018). In detail, we added all images containing the 'easy' background patches to the training set and trained the model for 10 epochs (i.e. 10 passes over all images). Afterwards, we also added the 'hard' background patches and trained the model for a total of 75 epochs. This ensured that the model can learn bird species appearances before potentially getting overwhelmed with the large background polygons and high complexity of them.
2 For our second strategy, we exploited the observation that the *immediate* surroundings of birds often belong to background. To do so, we assigned the eight neighbouring grid cells of a bird to the background class, similar to Kellenberger et al. (2018).
3 Finally, we leveraged the circumstance that our ground truth was annotated on a per-colony basis. Although the image tiles are not labelled completely, the colonies are, which allowed us to assign all grid cells between birds within a colony as background. To do so, we calculated the Euclidean distances between all birds of the same species in an image tile, and created convex hulls between all points whose linear distance amounted to at most eight grid cells. All grid cells within the union of these convex hulls that were unassigned could then be dedicated to background. Figure 6 shows a conceptualized example of this heuristic.

Figure 7 shows a flow chart with examples. The ground truth grid (right) contains grid cells with birds (coloured), background (grey), and un-annotated cells (white). Note that for the example shown, the background class is composed of both a polygon at the bottom of the map (*cf.* point 1 above), the eight neighbours around individual points in the middle and bottom left (*cf.* point 2), and the convex hulls around the remaining birds (point 3).

Finally, we employed a grid cell-wise regression loss as follows:

$$\mathscr{L}(\hat{\mathbf{Y}},\mathbf{Y}) = \sum_i \sum_j \sum_c \left(\hat{y}_{ijc} - y_{ijc}\right)^2, \quad (1)$$

which sums the squared difference between prediction $\hat{y}$ and ground truth label $y$ over all classes $c$, width $i$ and height $j$ positions of the prediction grid $\hat{\mathbf{Y}}$ and ground truth grid $\mathbf{Y}$ respectively. This loss encourages the model to predict a confidence value of 1 for one of the three bird species, if present in the ground truth, or 0 otherwise. Grid cells with neither bird nor background annotations are ignored during training (the loss values are set to zero for these locations). We trained the model for 75 epochs with a batch size of one image using stochastic gradient descent with momentum of 0.9, weight decay of $10^{-4}$, and a learning rate of $10^{-5}$ that gets divided by ten at epoch 50. Due to the large class imbalance, we found the model to often be overwhelmed by the most abundant species (royal tern), or else the background. To reduce this effect, we assigned class weights as follows: 1.0 for royal terns, 10.0 for Caspian terns, 20.0 for gulls and 0.01 for the background class. The model is implemented in PyTorch[6] and was trained on a Linux workstation with an NVIDIA Titan V graphics card.

[6]https://pytorch.org

 

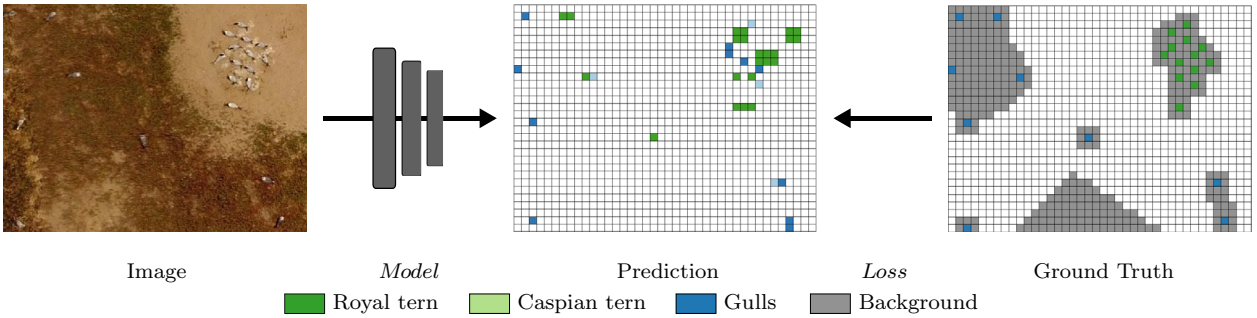


**Figure 7.** Overview of the model training. For each image (left), the CNN predicts a down-sized grid (middle) containing probabilities for each species in each grid cell. For training, images contain a number of bird locations as well as background polygons that get mapped to the prediction grid (legend at the bottom). Grid cells that are undefined (white) are ignored during training.

## Model evaluation

We tested the performance of our model based on precision-recall curves, with precision and recall being calculated as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}; \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

where *TP*, *FP* and *FN* are the number of true positive, false positive and false negative predictions in the entire test orthomosaic. A point predicted by the CNN is treated as a true positive if (and only if) it lies within a distance of 50 cm to the nearest ground truth annotation and has been predicted with the correct class. In practice, we can set a threshold for the minimum predicted probability (e.g. 0.1) for every predicted location; higher thresholds (towards the maximum of 1.0) usually result in higher precision, but lower recall, whereas lower thresholds (towards the minimum of 0.0) have the opposite effect. An optimum threshold depends on the requirements of the application and the desired trade-off between precision and recall. As is common, we calculated per-species precision-recall curves by varying the minimum confidence threshold from 0 to 1 in 50 steps (see, e.g. Fig. 9).

## Model inference

For model evaluation, we split the test orthomosaic into patches of size 800 × 600 pixels on a grid with 50% overlap between the patches. This increases prediction time, but we found it to be worthwhile to eliminate drops in recall towards the borders of the patches. We note that the false positives are primarily found in the neighbourhood of actual birds. To improve precision, we therefore applied non-maximum suppression (NMS), which only retains predictions that are not within distance of other predictions whose maximum bird confidence value is greater. To this end, we exploited the prior knowledge about the expected size and distance of individuals per species: for example, we found that royal terns are generally smaller and breed closer together than Caspian terns, as visible in Figure 2. We therefore estimated the expected distance between individuals per species and used those values to perform class-specific NMS to discard multiple predictions of the same individual. We selected five sample pairs of individuals per species closely together and measured the distances between the birds' centres. Although the expected distance between individuals could also be estimated from the training annotations, we found that the effects on the NMS result of doing so are negligible. We obtained and used 22 cm for royal terns, and 30.5 cm for Caspian terns and gulls. Eventually, this means that whenever the model predicts, for example, two royal terns that are less than 22 cm apart from each other, the prediction with lower confidence value gets discarded.

## Markov random field for post-processing

When looking at the model predictions in more detail (Fig. 8 left), we can notice that the model misclassifies a number of birds in the middle of flocks. Those flocks, however, actually only consist of one species most of the time (royal terns in the figure). We therefore decided to use this prior knowledge and post-processed the predictions with a Markov random field (MRF; Schindler, 2012; Tuia et al., 2018). MRFs post-process the predictions by means of a graph, where the graph's vertices ($\nu$) are the predicted bird points in our case, and its edges ($\epsilon$) are constructed between vertices that, for example are sufficiently close to each other. Both the vertices and edges have attributed a certain notion of cost, respectively energy, which is high if the prediction goes against the prior knowledge (i.e. the points' labels are different from

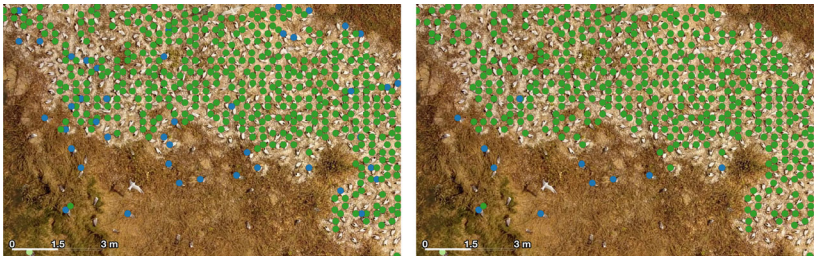

**Figure 8.** The bare, non-maximum suppressed bird predictions of the model (left) often contained spurious misclassifications, such as the gulls (blue) misclassified inside the area of royal terns (dark green). Post-processing the predictions with an MRF corrected most misclassifications (right).

the neighbours), and low otherwise. MRFs then try to minimize the global energy across all graph vertices and edges, which is given by:

$$E = \sum_{p \sim V} (U_p) + \lambda \sum_{pq \in \varepsilon} P_{pq}(l_p, l_q) \quad (3)$$

where

$$U_p = -\log(\phi(X)); P_{pq} = \|l_p - l_q\|_2 \quad (4)$$

Here, $U_p$ is the 'unary' term, which is the negative log of the probabilities per bird species as predicted by the CNN on image $X$ for each point (graph vertex) $p$. In other words, the unary term is the likelihood that a given detection belongs to one of the three classes, as predicted by the CNN ($U_p$ only depends on the CNN). $P_{pq}$ is the pairwise term that applies between predicted graph vertices $p$ and $q$, which in our case is the spatial (Euclidean) distance between the two predicted points' locations $l_p$ and $l_q$. The pairwise term represents our prior knowledge of co-occurrence of species, which in our case is the implication that bird colonies are usually only composed of one species. $\lambda$ is a trade-off constant, which we set to 0.5 to reduce the effect of the pairwise term (i.e. we weight the CNN predictions as more important). Effectively, this formulation compares CNN predictions for each point with all of neighbouring points, to which it is connected by graph vertices. If the model predicts that a point is a Caspian tern, but all connected neighbours have a prediction of 'royal tern', then this point's label will likewise become 'royal tern', unless the model gave it a particularly high confidence in $U_p$. In practice, this is done by calculating energies according to Equation (3) for every label class and selecting the label with the minimum value for each point.

As a neighbourhood around each point, we included all other predicted points within a square of 11 grid cells (around 1.76 m) around the point. We applied the MRF over model predictions in each 800 × 600 patch, prior to NMS. We used the Iterated Conditional Modes (ICM) solver (Besag, 1986), which iterates over all predicted vertices and assigns labels according to the minimum energy, until no labels change anymore. We report results for both the original, NMSed predictions and those with NMS and MRF post-processing in Section 3.2 below.

## Results and Discussion

### Manual annotation statistics

We divided the manual annotation task across five annotators, who labelled all 274 patches concurrently. This resulted in 21 066 point annotations (Fig. 3; Table 2). We monitored the annotation time using the AIDE platform, and obtained a total running time of the manual annotation task of roughly 20 days and 16 h, including breaks. The effective annotation time for all patches, summed over all annotators' contributions, was 3 h and 2 min. On average, annotators required 1.19 s per annotation. After the annotation session, the provided points were visually inspected by one of the authors for correctness.

### Automated detection results

Figure 9 shows per-species precision-recall curves for the models, with the maximum and minimum values across all three random seeds reported as shaded polygons, and the average precision-recall curves as dashed lines. The left figures show results obtained when only using NMS, the right side shows results with MRF-post-processing and NMS. From these results, we can see a sustained high precision and small value range for royal terns in all models (dark green). Recalls reach up to 88%. For Caspian terns (light green), the variation in precision across random seeds is only marginally larger, but precision is significantly lower. In turn, the gulls (blue) show the lowest consistency across seeds, with variations in precision at highest recall from around 32% to 60% without MRF (left), and 52% to 70% with (right).

The limited precision for the Caspian terns can be traced to two reasons. The first reason is a confusion of

 

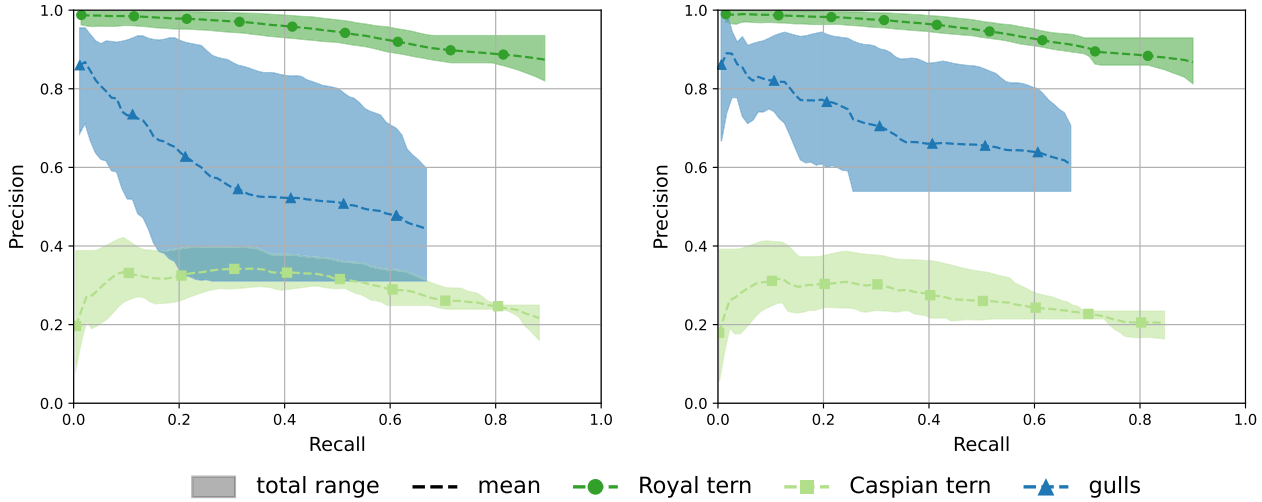


**Figure 9.** Precision-recall curves per species on the test set, showing the total variation of the models trained on three sets of training patches (shaded) and the average performance (dashed). The left figure shows results with just NMS, and the right with NMS and a pairwise MRF on the predicted points.

royal terns and Caspian terns, especially when ever individuals were more loosely distributed in the area. The majority of royal tern flocks were densely packed with barely any distance between birds. Through the receptive field, the CNN likely learned and associated the spatial pattern of high-density flocks with the royal tern class. However, the test orthomosaic nonetheless contains a number of royal terns that are distributed in larger distances compared to dense colonies, such as most of the individuals in Figure 10 (left). However, due to the aforementioned trait, the model ended up mispredicting those individuals as the visually most similar species, which is the Caspian tern.

A second reason for false positives is due to other bird species that had not been labelled, an example of which can be seen in the right panel of Figure 10. Our labelling scheme involved creating dedicated background polygons, and these additional species were included in the 'hard' background polygons. However, due to the reduced background weight, it seems that the learning signal from these extra species was dampened to a too large extent, causing the model to falsely predict those species as well. This may be addressable by an additional class for all other bird species. For the current predictions, the other bird species luckily appear in separate flocks and can quickly be removed in a manual post-processing step.

In the case of the gulls, the limitations are of different nature – the species shows the highest uncertainty across model states, as well as the overall lowest recall. Both effects presumably originate from the large variation in background composition: unlike royal terns and Caspian terns that concentrate on sand, gulls are most frequently found in vegetated patches that cause partial occlusion effects due to grass and other weeds, and further contain debris like rocks. These caused both precision and recall to get reduced. Upon visual inspection, we noticed that the model seemed to particularly struggle over the more heterogeneous parts of the weedy patches, which further

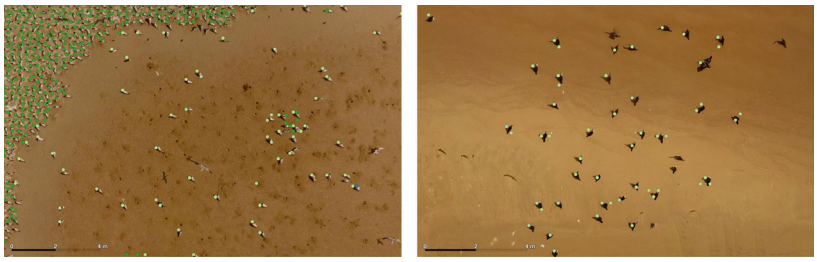

**Figure 10.** Examples of classification errors made by the model. In the left figure, many of the Caspian tern individuals (light green) are predicted twice, possibly due to their slightly larger size compared to, for example royal terns (dark green). The right figure shows individuals of Great cormorant (*Phalacrocorax carbo*), which were not considered in this study and erroneously predicted as Caspian terns.

corroborates this assumption. In sum, the problems with the gulls are not spatially concentrated; instead, the false positives and missed individuals are scattered across the entire orthomosaic. This unfortunately means that manual post-processing is more tedious than in the case of, for example, falsely assigned background species discussed above.

When looking at the results after NMS and the MRF (Fig. 9, right), the effect of the MRF seems to be strongest in the case of the gulls, raising the averaged precision at maximum recall from around 48% to 60%. Similar to the confusion between Caspian terns and royal terns, the model seemed to occasionally predict royal terns as gulls (*cf.* Fig. 8). This seems rather surprising, given the differences between royal terns and gulls in terms of their appearances and background composition. The MRF primarily resolves confusions in these dense flocks of royal terns by exploiting the proximity patterns and the high homogeneity of species within colonies. For areas where birds are spaced apart further, the neighbouring effect diminishes. Since the falsely predicted gulls get reverted to royal terns, the MRF also has a positive effect on the latter, producing more homogeneous flocks of royal terns. However, this effect is not visible in Figure 9, due to the sheer number of royal tern individuals contained in the test orthomosaic.

Figure 11 shows all predictions by the model with confidence of 0.1 or greater, NMS and MRF post-processing, over the test orthomosaic. Compared to the ground truth (Fig. 3), it can be seen that the CNN located the large patches of royal terns (green) and the general area of the gulls (blue) very well indeed. The total number of predicted birds (between 17 485 and 23 288, depending on the random seed) also matches the number of individuals in the ground truth (21 066) reasonably well and further corresponds to estimations we obtained during on-site censuses with binocular scans from two observers (18 000 and 24 500). Hence, prediction results of our CNN are on-par with manual surveys. However, the model clearly struggled identifying the more rare species, such as the gulls, and occasionally misclassified background clutter as birds. In part, this can be attributed to the exceptionally low amount of training data – 200 samples per species is unusually little for deep learning models having millions of free parameters, and we empirically found it to be the lower limit to obtain a reasonably stable prediction (see also experiments with fewer points in Appendix). However, the focus species of the survey and most abundant class, the royal tern, is recognized and mapped with an exceptionally high accuracy. The other two species occur in greatly reduced numbers and any mispredictions can be rectified with low amounts of human intervention. Furthermore, since the royal tern is the key species indicating major breeding sites along the West African coast and primary target of our study, additional prediction of the other species is primarily a surplus and implications of lower accuracy are not as severe. With this in mind, we argue that a model like the proposed one may at least serve as a useful pre-detector, alleviating researchers from the tedium of annotating tens of thousands of birds.

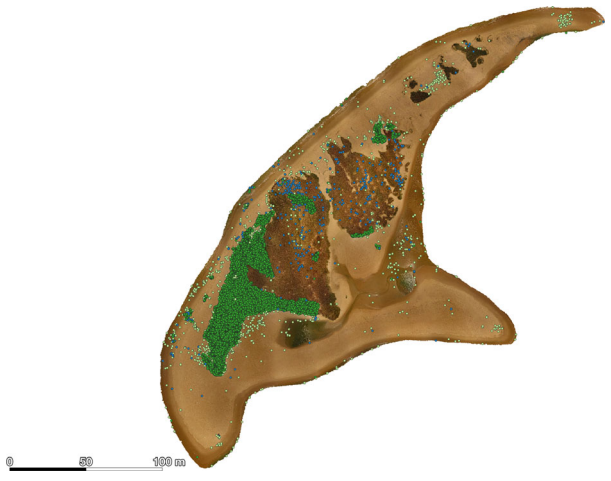


**Figure 11.** Predictions by the CNN on the test orthomosaic with class confidence 0.1 or greater, post-processed with NMS and the MRF.

Furthermore, a model as proposed helps reducing total analysis time by a large margin. Obtaining a result as shown in Figure 11 required labelling 600 points and a few polygons (about 30 min for creating all training annotations ), training the model for 75 epochs (3:58 h with our unoptimized code), and predicting individuals (3:14 min; 3:34 with MRF), resulting in about 4.5 h of total time required. This is in stark contrast to the multiple days that were required by the manual photointerpretation survey (Section 2.3), especially considering that the amount of manual intervention can be reduced to half an hour. Crucially, model running time is low, and the model can be employed to also predict birds in the remaining orthmosaics (about 20 min of additional time required). In total, all six orthomosaics are estimated to contain about 70 000 birds; as a result, the time gain of employing our CNN as opposed to manual counting scales exponentially with the number of images involved. The training time may further be reduced in practice through early stopping (we deliberately trained the model for more epochs than needed to ensure convergence). Hence, we argue that using machine learning is a viable way to obtain spatial, high-density bird census estimations in a faster and significantly less tedious manner compared to traditional surveys and manual photo-interpretation efforts.

 

## Limitations and Future Perspectives

Although our results indicate that CNNs can be used to obtain comparably reliable detections with a promising decrease in manual labour effort required, the proposed pipeline has a number of limiting factors. First, the appearance similarity of the two gull species and low abundance of one of them required merging the species together; our trials to predict them separately resulted in one of the two being almost completely ignored. It is questionable whether more training points could help the model overcome this issue, due to the visual similarity of the species.

Second, despite a satisfactory recall, our model struggled in predicting Caspian terns and gulls with satisfying precision. In other words, this means that most Caspian terns in the test image were correctly detected (high recall), but that many false positives were also predicted for this class (low precision). Visual inspection of the predicted maps, as in Figure 11, confirms these findings: the model identified various different background objects as Caspian terns, including weeds, rocks and sand formations, debris, and other bird species (*cf.* Fig. 10). Although the model still yielded a moderate to very high recall (68% for gulls, 88% for Caspian terns), these false positives have the effect of reducing precision down to 60% for gulls and 20% for Caspian terns, as shown in Figure 9. Our annotated 'hard' background polygons (see Section 2.4.2) include such more heterogeneous background formations and hence partially reduced the number of false positives, but were not sufficient to do so to a satisfactory degree. Increasing the number of labelled birds (up to 600 per class; see Appendix) did not solve this problem. Future works may address this by adding more background polygons and introducing dedicated classes for other bird species and types of clutter.

Furthermore, polygons have the potential to facilitate the labelling process of birds in the first place. For annotators, it is faster and more intuitive to simply draw freehand polygons around colonies of birds, instead of labelling every individual with points. If the effort is lowered on the users' side, there is also less information at hand for the model, which can complicate training. Studying the effectiveness of this kind of weakly supervised learning logic (Kellenberger et al., 2019) in this high point density scenario is a future direction of our work.

Furthermore, we noticed during training that the model requires particularly careful tuning of hyperparameters, such as class weights, in order to predict the species in the right amounts. While this effect is common to all supervised machine learning models, it gets amplified in our case due to the large imbalance and low abundance of annotations for training. With periodic checks during training, a stable model state can be achieved, but further work may attempt to improve model stability by, for example adding regularizers, or incorporating more advanced weighting schemes (Madhyastha & Jain, 2019).

Finally, predictions could be improved by incorporating more advanced priors. In our work we post-processed the predictions by means of an MRF that respects positive spatial autocorrelation, that is neighbouring similarities. A next possible option could be to instead estimate the average bird density per species and reason on a flock level about the distribution, and the species class that fits best. This could potentially also be incorporated by means of random fields, but we leave this idea to future studies.

## Conclusion

We developed a CNN that can detect terns and gulls in UAV-derived imagery over dense colonies against a heterogeneous background. This task is by itself challenging due to the high and variable colony density, and we set a further requirement to minimize the amount of manual work required to get the CNN to work properly. To this end, we used our recently proposed annotation platform AIDE, which is open source and readily deployable. As a result, where normally a large effort is needed to annotate training data sets for complex imagery, our CNN performed well for the most abundant class, the royal tern, with annotations that required as little as 30 min to create. Our model struggled with less abundant classes (Caspian terns and gulls), predicting large numbers of false positives. However, given those classes only constitute less than 7% of the total 21 066 individuals, we believe our approach can still greatly facilitate the monitoring of West African colonies by reducing the analysis time and improving the accuracy of the estimates. It furthermore allows for the extraction of additional data beyond ground-based field surveys, such as the location of the detected birds, which can provide insights of small changes in breeding location. Further work may focus on two aspects: first, the improvement of the precision of the model by introducing dedicated prediction classes for other bird species present; and second the investigation of the transferability to other ecosystems and scenarios. The latter may require more training images from more diverse ecosystems beyond the six islands and/or from acquisitions in different seasons to make the model more robust to data variations, but could pave the way for the widespread adoption of (semi-) automated bird identification with CNNs across censuses.

## Author Contributions

The authors Benjamin Kellenberger, Thor Veen, Eelke Folmer and Devis Tuia declare that they have all seen and

 

approved the version of the manuscript submitted to *Remote Sensing in Ecology and Conservation*. They also confirm that the manuscript has not been submitted, or published, elsewhere, in any way, and that it also is not in press or under consideration for publication in another journal.

## Acknowledgements

We thank Hanneke Dallmeijer, Jan Veen and Wim Mullié for their invaluable advice regarding bird monitoring in West Africa and BirdLife Dakar and the MAVA foundation for organizing and supporting the census work. Special thanks go to Dakotah Daily and Bryce Andreasen, for their help in annotating the test orthomosaic. We also gratefully acknowledge the support of the NVIDIA Corporation with the donation of a Titan V GPU used for this research, as well as the Microsoft AI for Earth program and team.

## Funding Information

We also gratefully acknowledge the support of the NVIDIA Corporation with the donation of a Titan V GPU used for this research, as well as the Microsoft AI for Earth program and team.

### References

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

Andrew, M.E. & Shephard, J.M. (2017) Semi-automated detection of eagle nests: an application of very high-resolution image data and advanced image analyses to wildlife surveys. *Remote Sensing in Ecology and Conservation*, **3**(2), 66–80.

Baran, E. (2000) Biodiversity of estuarine fish faunas in West Africa, Naga. *The International Center for Living Aquatic Resources Management Quarterly*, **23**(4), 4–9.

Bengio, Y., Louradour, J., Collobert, R. & Weston, J. (2009) Curriculum learning. In: Bottou, L. and Littman, M. (Eds.) *International conference on machine learning*. Madison, WI: Omnipress, pp. 41–48.

Besag, J. (1986) On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B (Methodological)*, **48**(3), 259–279.

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), 1–9.

Butchart, S.H., Walpole, M., Collen, B., Van Strien, A., Scharlemann, J.P., Almond, R.E. et al. (2010) Global biodiversity: indicators of recent declines. *Science*, **328**(5982), 1164–1168.

Camphuysen, C. & Van der Meer, J. (2005) Wintering seabirds in West Africa: foraging hotspots off Western Sahara and Mauritania driven by upwelling and fisheries. *African Journal of Marine Science*, **27**(2), 427–437.

Cardinale, B.J., Duffy, J.E., Gonzalez, A., Hooper, D.U., Perrings, C., Venail, P. et al. (2012) Biodiversity loss and its impact on humanity. *Nature*, **486**(7401), 59–67.

Carney, K.M. & Sydeman, W.J. (1999) A review of human disturbance effects on nesting colonial waterbirds. *Waterbirds*, **22**(1), 68–79.

Costanza, R., Kemp, W.M. & Boynton, W.R. (1993) Predictability, scale, and biodiversity in coastal and estuarine ecosystems: implications for management. *Ambio*, **22**(2–3), 88–96.

Edney, A.J. & Wood, M.J. (2020) Applications of digital imaging and analysis in seabird monitoring and research. *Ibis*, **163**(2), 317–337.

Eikelboom, J.A., Wind, J., van de Ven, E., Kenana, L.M., Schroder, B., de Knegt, H.J. et al. (2019) Improving the precision and accuracy of animal population estimates with aerial image object detection. *Methods in Ecology and Evolution*, **10**(11), 1875–1887.

Einoder, L.D. (2009) A review of the use of seabirds as indicators in fisheries and ecosystem management. *Fisheries Research*, **95**(1), 6–13.

FAO (2011) Review of the state of world marine fishery resources. FAO. Available from http://www.fao.org/3/i2389e/i2389e.pdf

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.

Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. (2016) *Deep learning*. Vol. **1**. Cambridge, MA: MIT press Cambridge.

Gray, P.C., Fleishman, A.B., Klein, D.J., McKown, M.W., Bézy, V.S., Lohmann, K.J. et al. (2019) A convolutional neural network for detecting sea turtles in drone imagery. *Methods in Ecology and Evolution*, **10**(3), 345–355.

Gregory, R.D., Noble, D., Field, R., Marchant, J., Raven, M. & Gibbons, D. (2003) Using birds as indicators of biodiversity. *Ornis Hungarica*, **12**(13), 11–24.

Hamilton, G., Corcoran, E., Denman, S., Hennekam, M.E. & Koh, L.P. (2020) When you can't see the koalas for the trees: using drones and machine learning in complex environments. *Biological Conservation*, **247**, 108598.

He, K., Zhang, X., Ren, S. & Sun, J. (2016) Deep residual learning for image recognition. In: Bajcsy, R., Li, F.-F. and Tuytelaars, T., (Eds.) *Computer vision and pattern recognition*. New York, NY: IEEE, pp. 770–778.

 

Hodgson, J.C., Baylis, S.M., Mott, R., Herrod, A. & Clarke, R.H. (2016) Precision wildlife monitoring using unmanned aerial vehicles. *Scientific reports*, **6**(1), 1–7.

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.

Huh, M., Agrawal, P. & Efros, A.A. (2016) *What makes ImageNet good for transfer learning?* arXiv preprint arXiv:1608.08614.

Ivošević, B., Han, Y.-G., Cho, Y. & Kwon, O. (2015) The use of conservation drones in ecology and wildlife research. *Journal of Ecology and Environment*, **38**(1), 113–118.

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., Marcos, D. & Tuia, D. (2019) When a few clicks make all the difference: improving weakly-supervised wildlife detection in UAV images. In: Davis, L., Torr, P. and Zhu, S.-C., (Eds.) *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. New York, NY: IEEE.

Kellenberger, B., Morris, D. & Tuia, D. (2020) AIDE: accelerating image-based ecological surveys with interactive machine learning. *Methods in Ecology and Evolution*, **11**(12), 1716–1727.

Krizhevsky, A., Sutskever, I. & Hinton, G.E. (2012) ImageNet classification with deep convolutional neural networks. In: Bartlett, P. and Pereira, F., *Advances in neural information processing systems*. Red Hook, NY: Curran Associates, Inc., pp. 1097–1105.

LeCun, Y., Bengio, Y. & Hinton, G. (2015) Deep learning. *Nature*, **521**(7553), 436–444.

Linchant, J., Lisein, J., Semeki, J., Lejeune, P. & Vermeulen, C. (2015) Are unmanned aircraft systems (UASs) the future of wildlife monitoring? A review of accomplishments and challenges. *Mammal Review*, **45**(4), 239–252.

Madhyastha, P. & Jain, R. (2019) *On model stability as a function of random seed*. arXiv preprint arXiv:1909.10447.

Parsons, M., Mitchell, I., Butler, A., Ratcliffe, N., Frederiksen, M., Foster, S. et al. (2008) Seabirds as indicators of the marine environment. *ICES Journal of Marine Science*, **65**(8), 1520–1526.

Ren, S., He, K., Girshick, R. & Sun, J. (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: Cortes, C. and Lawrence, ND, (Eds.) *Advances in neural information processing systems*. Red Hook, NY: Curran Associates, Inc., pp. 91–99.

Rumelhart, D.E., Durbin, R., Golden, R. & Chauvin, Y. (1995) Backpropagation: the basic theory. In: Chauvin, Y. & Rumelhart, D.E. (Eds.), *Backpropagation: theory, architectures and applications*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, Ch. 1, pp. 1–34.

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.

Schindler, K. (2012) An overview and comparison of smooth labeling methods for land-cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, **50**(11), 4534–4545.

Terletzky, P.A. & Ramsey, R.D. (2016) Comparison of three techniques to identify and count individual animals in aerial imagery. *Journal of Signal and Information Processing*, **7**(3), 123–135.

Tuia, D., Volpi, M. & Moser, G. (2018) Decision fusion with multiple spatial supports by conditional random fields. *IEEE Transactions on Geoscience and Remote Sensing*, **56**(6), 3277–3289.

Ulyanov, D., Vedaldi, A. & Lempitsky, V. (2016) *Instance normalization: the missing ingredient for fast stylization*. arXiv preprint arXiv:1607.08022.

Veen, J., Dallmeijer, H., Schlaich, A.E., Veen, T. & Mullié, W.C. (2019) Diet and foraging range of Slender-billed gulls Chroicocephalus genei breeding in the Saloum Delta, Senegal. *Ardea*, **107**(1), 33–46.

Veen, J., Dallmeijer, H., Van Damme, C.J., Leopold, M.F. & Veen, T. (2018a) Analyzing pellets and feces of African royal terns (*Thalasseus maximus albididorsalis*) results in different estimates of diet composition. *Waterbirds*, **41**(3), 295–304.

Veen, J., Dallmeijer, H. & Veen, T. (2018b) Selecting piscivorous bird species for monitoring environmental change in the Banc d'Arguin, Mauritania. *Ardea*, **106**(1), 5–18.

Veen, J., Peeters, J., Leopold, M., Van Damme, C. & Veen, T. (2003) *Les oiseaux piscivores comme indicateurs de la qualité de l'environnement marin: suivi des effets de la plche littorale en afrique du nord-ouest*. Alterra: Tech. rep.

Veen, J., Peeters, J., Mullié, W.C. & Diagana, C. (2004) *Manual for monitoring seabird colonies in West Africa*. Dakar: Wetlands International.

Volpi, M. & Tuia, D. (2017) Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, **55**(2), 881–893.

## Supporting Information

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

**Appendix 1.** Additional results obtained with more and fewer training points.

 