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

ORIGINAL RESEARCH

Multispecies detection and identification of African mammals in aerial imagery using convolutional neural networks

Alexandre Delplanque[1], Samuel Foucher[2], Philippe Lejeune[1], Julie Linchant[1] & Jérôme Théau[3,4]

[1]TERRA Teaching and Research Centre (Forest Is Life), ULiège, Gembloux Agro-Bio Tech, 2 Passage des Déportés, Gembloux 5030, Belgium
[2]Computer Research Institute of Montréal, 405 Ogilvy Avenue, Montréal QC, H3N 1M3, Canada
[3]Departement of Applied Geomatics, Université de Sherbrooke, 2500 Boulevard de l'Université, Sherbrooke QC, J1K 2R1, Canada
[4]Quebec Centre for Biodiversity Science (QCBS), Stewart Biology, McGill University, Montréal QC, H3A 1B1, Canada

Keywords
African mammals, CNNs, deep learning, multispecies, UAV, wildlife monitoring

Correspondence
Alexandre Delplanque, ULiège, Gembloux Agro-Bio Tech, TERRA Teaching and Research Centre (Forest Is Life), 2 Passage des Déportés, 5030 Gembloux, Belgium.
Tel: +32 816 226 78;
E-mail: alexandre.delplanque@uliege.be

Editor: Temuulen Sankey
Associate Editor: Anna Carter

Received: 15 April 2021; Revised: 6 July 2021; Accepted: 20 July 2021

doi: 10.1002/rse2.234

Abstract

Survey and monitoring of wildlife populations are among the key elements in nature conservation. The use of unmanned aerial vehicles and light aircrafts as aerial image acquisition systems is growing, as they are cheaper alternatives to traditional census methods. However, the manual localization and identification of species within imagery can be time-consuming and complex. Object detection algorithms, based on convolutional neural networks (CNNs), have shown a good capacity for animal detection. Nevertheless, most of the work has focused on binary detection cases (animal vs. background). The main objective of this study is to compare three recent detection algorithms to detect and identify African mammal species based on high-resolution aerial images. We evaluated the performance of three multi-class CNN algorithms: Faster-RCNN, Libra-RCNN and RetinaNet. Six species were targeted: topis (*Damaliscus lunatus jimela*), buffalos (*Syncerus caffer*), elephants (*Loxodonta africana*), kobs (*Kobus kob*), warthogs (*Phacochoerus africanus*) and waterbucks (*Kobus ellipsiprymnus*). The best model was then applied to a case study using an independent dataset. The best model was the Libra-RCNN, with the best mean average precision (0.80 ± 0.02), the lowest degree of interspecies confusion ($3.5 \pm 1.4\%$) and the lowest false positive per true positive ratio (1.7 ± 0.2) on the test set. This model was able to detect and correctly identify 73% of all individuals (1115), find 43 individuals of species other than those targeted and detect 84 missed individuals on our independent UAV dataset, with an average processing speed of 12 s/image. This model showed better detection performance than previous studies dealing with similar habitats. It was able to differentiate six animal species in nadir aerial images. Although limitations were observed with warthog identification and individual detection in herds, this model can save time and can perform precise surveys in open savanna.

Introduction

Survey and monitoring of animal populations are key management tools in nature conservation and are essential to help fight the pressures they suffer. Anthropogenic pressures, such as poaching, are encountered mainly in developing countries (including most in Africa) where the pressure on biodiversity is very high (Linchant, Lisein, et al., 2015). While large African mammals, such as buffaloes (*Syncerus caffer*) or hippopotamuses (*Hippopotamus amphibius*), play an important role in the dispersion and migration of macro-nutrients within landscapes (Lacher et al., 2019), the average abundance of these populations declined by 59% between 1970 and 2005 (Craigie et al., 2010). Moreover, the latest estimated Living Planet Index indicates a 65% decline in the overall African vertebrate populations between 1970 and 2016 (Almond & Petersen, 2020). Even though humans are dependent on

biodiversity (Almond & Petersen, 2020; Isbell et al., 2017), their impact on the environment is leading us into a period of mass extinction (Ceballos et al., 2015). Moreover, in view of the disruption of future climate conditions (IPCC, 2014), species not able to adapt rapidly could see their populations decline even further (Hetem et al., 2014; Thuiller et al., 2006).

Most of the time, the size of an animal population is estimated through sample counts which consist of estimating the animal density in sample units selected at random or following a systematic scheme. The size of the population corresponds to the product of the mean density inside sample units per surveyed area surface (Norton-Griffiths, 1978). Unfortunately, counting campaigns of this type can rapidly become expensive (Gaidet-Drapier et al., 2006), particularly for large mammal surveys for which the use of a light aircraft is almost indispensable (Jachmann, 1991). Moreover, these aerial campaigns can become dangerous, and their logistics are very complex for operators (Watts et al., 2010; Witmer, 2005).

Although they cannot cover large areas, the use of UAVs (unmanned aircraft vehicles) is presented as a cheaper, more suitable and safer alternative (Chabot & Bird, 2015; Linchant, Lisein, et al., 2015; Vermeulen et al., 2013). In addition, there are sensors that can be embedded and which offer the possibility of acquiring very high-resolution images (Linchant, Lisein, et al., 2015).

Several species of large African mammals have already been studied using UAVs, such as the African elephants (*Loxodonta africana*) (Vermeulen et al., 2013), black (*Diceros bicornis*) and white (*Ceratotherium simum*) rhinos (Mulero-Pázmány et al., 2014), the hippopotamus (Linchant et al., 2018) and many other species (Kellenberger et al., 2018; Rey et al., 2017).

However, counting and identification are not carried out simultaneously and must be deferred when using UAVs. Due to the large amount of data to be analyzed, the size of the study area and the static nature of the animals on the images, counting can become very complex and time-consuming. This problem can be alleviated by utilizing object detection, which finds, locates and classifies objects in images (Zhao et al., 2019). convolutional neural networks (CNNs) have become the basic elements of most computer vision processes and have also proven to be extremely effective in the field of remote sensing (Zhu et al., 2017). These networks have been applied to animal detection in aerial and UAV images and have shown encouraging results (Barbedo et al., 2019; Eikelboom et al., 2019; Kellenberger et al., 2018; Moreni et al., 2021; Naudé & Joubert, 2019a, 2019b; Peng et al., 2020). However, almost all of these studies did not distinguish between species nor were they focused on the case of a single species. It would therefore be interesting to develop a multi-species approach in order to further minimize human resources required for the processing of survey data. To our knowledge, only one study of multispecies animal detection on aerial images using object detection has been conducted to date, Eikelboom et al. (2019), who worked on detecting and identifying three African animal species using aerial oblique images and CNN.

The objective of this study is to compare the performances of three object detection algorithms, based on CNNs, to automatically detect and identify six African mammal species in nadir aerial images: African buffalo, kob (*Kobus kob*), topi (*Damaliscus lunatus jimela*), African warthog (*Phacochoerus africanus*), waterbuck (*Kobus ellipsiprymnus*) and African elephant. The best model is then put into a practical perspective on an independent set of UAV images acquired in a different study area.

Materials and Methods

Dataset

Data collection

We used three different aerial datasets to conduct our study (see details in Table 1). The 'Virunga' and 'Garamba' are two UAV datasets that were taken from a database maintained by the University of Liège, Gembloux Agro-Bio Tech (Belgium). The Aerial Elephant Dataset (AED) is a free dataset provided by Naudé and Joubert (2019a, 2019b).

The Virunga and AED datasets were merged and used as the 'general dataset' to develop the models (training, validation and test), while the Garamba dataset was used as a 'case study' to test the performance of the best model on a complete independent dataset. This was done in order to evaluate the model on a practical use case that did not include all the targeted species and which contained other species.

The species selection was based on the availability of at least 100 individuals in the general dataset to ensure minimal model configurations. In addition, to optimize the speed of model development, images that did not contain animals were not used.

General dataset data splitting

The distribution of individuals of each species according to training, validation and test sets is given in Table 2. The approximate targets of distribution were 70% of the individuals in the training, 10% in the validation and 20% in the test datasets.

The distribution of the number of individuals by species and by flight was considered in performing the split.

Table 1. Dataset specifications and details.

Dataset	Virunga	AED (Naudé & Joubert, 2019a, 2019b)	Garamba
Location	DRC (Virunga national park)	Parks, games and reserves in Botswana, Namibia and South Africa	DRC (Garamba National Park)
Land cover (Mayaux et al., 2004)	Savanna	Deciduous woodland, open deciduous shrubland, closed grasslands	Savanna
Dates	April–June 2016	2014–2018	May 2015
Time of day	Early morning	Full day	Early morning
System	Falcon (UAV)	SkyReach BushCar (A/C)	Falcon (UAV)
Camera(s)	Sony-A6000, Sony-Nex7	Canon 6D	Sony-Nex7
Flight altitude	100 m	220–2270 m	90 m
Number of flights	9	8	6
Image dimension	6000 × 4000 pixels	Various (5472 × 3648 pixels, 5496 × 3670 pixels, 5521 × 3687 pixels, 5525 × 3690 pixels)	6000 × 4000 pixels
GSD	2.4 cm	2.4–13.0 cm	2.0 cm
Species	Hippopotamus, buffalo, kob, topi, warthog, waterbuck	Elephant	Hartebeest (*Alcelaphus buselaphus*), hippopotamus, buffalo, kob, warthog, waterbuck, giraffe (*Giraffa camelopardalis*)
Images selected	897	400	All (7034)

GSD, ground sampling distance; AED, aerial elephant dataset; DRC, Democratic Republic of Congo; UAV, unmanned aerial vehicle; A/C, aircraft.

This step was required in order to avoid the splitting of some consecutive images containing the same individuals and to thereby maintain the independence of the three sets.

Ground truth

For the Virunga and Garamba datasets, the annotations (points and labels) were provided with the images. The individuals were previously located and identified manually by two operators on the UAV images using the software WIMUAS (Linchant, Lhoest, et al., 2015; http://www.gembloux.ulg.ac.be/gf/outilslogiciels/VolDrone2016.7z). The AED dataset also provided annotations (points) with the images (Naudé & Joubert, 2019a, 2019b). We assumed that all pre-identifications were correct. Bounding boxes were manually defined by a co-author of this study using the Colabeler AI annotation tool (http://www.colabeler.com/).

Methodology

Detection algorithms and implementation on the general dataset

Three object detection algorithms were tested: Faster-RCNN (Ren et al., 2017), Libra-RCNN (Pang et al., 2019) and RetinaNet (Lin, Goyal, et al., 2017). These algorithms were selected based on their performance on the benchmark datasets and on the availability of the code at the time of the study.

Table 2. Number of individuals according to species, training, validation and test sets.

Species	Training	Validation	Test	Total
Buffalo	1058 (70%)	102 (7%)	349 (23%)	1509
Elephant	2012 (68%)	264 (9%)	688 (23%)	2964
Kob	1732 (73%)	161 (7%)	477 (20%)	2370
Topi	1678 (62%)	369 (13%)	675 (25%)	2722
Warthog	316 (73%)	43 (10%)	74 (17%)	433
Waterbuck	166 (69%)	39 (16%)	36 (15%)	241
Total	6962 (68%)	978 (10%)	2299 (22%)	10 239

The different rows show the distribution of individuals in each set and the relative percentage (in parentheses).

Faster-RCNN

This object detection algorithm (Ren et al., 2017) takes images as input and constructs feature maps using a CNN (also called backbone). Based on these features maps, a region proposal network generates region proposals and assigns a probability of containing an object to each region. The predicted region proposals are then reshaped and eventually, classification and bounding box regression is performed to predict the presence and

location of objects in the input images. These types of networks are commonly called 'two-stage detectors' due to their two-step process (Soviany & Ionescu, 2018). Faster-RCNN was chosen because it is used in many studies as a baseline.

Libra-RCNN

This algorithm, developed by Pang et al. (2019), is also a two-stage detector that does basically the same thing as Faster-RCNN. Its particularity is that it balances the training process at three levels which initially limit the detection performance:

1 the sample level, by balancing the distribution of training samples close to that of challenging samples (called hard negatives). This addresses the problem of the random sampling scheme that often results in selected samples dominated by the easy ones (Pang et al., 2019);
2 the feature level, by balancing the low-level and high-level features of each layer in the backbone, which are complementary for object detection;
3 the objective level, by balancing the tasks of localization and classification, thus avoiding one of the two tasks being overwhelmed by the other.

Thanks to its multi-level balanced approach to training, Libra-RCNN allows for greater precision and recall than Faster-RCNN, which is why it has been selected for comparison.

RetinaNet

This third algorithm is a 'single-stage detector', unlike the first two algorithms presented above. Algorithms of this type treat object detection as a simple regression problem by taking an input image and learning the class probabilities and bounding box coordinates directly (Soviany & Ionescu, 2018). Its architecture is composed of a backbone that takes input images, builds feature maps at different scales and generates region proposals for each scale in the form of anchors (Lin, Goyal, et al., 2017). These anchors are then used as inputs for two sub-networks, the first one classifies the object and the second one simultaneously performs the regression of the bounding boxes. RetinaNet was used by Eikelboom et al. (2019) for the detection and identification of three African mammal species based on oblique aerial images. This algorithm was therefore chosen to evaluate its performance on nadir UAV images.

For all three algorithms, the backbone consists of a ResNet-101 (He et al., 2016) connected to a feature pyramid network (Lin, Dollár, et al., 2017). These algorithms were used through their implementation in the adapted mmdetection toolbox version 1.0.0 (Chen et al., 2019) with PyTorch 1.4.0, TorchVision 0.5.0, OpenCV 4.4.0, MMCV 0.6.0, CuDNN 7.6.3 and Magma 2.5.1 libraries. All the codes and libraries were implemented and transcribed into Jupyter notebooks to run on Google Colaboratory. Training and detection runs were then performed with an NVIDIA Tesla P100-PCIE 16GB GPU running on an Ubuntu 18.04 LTS Colab Linux platform, with CUDA 10.1.243. These were followed by statistical tests conducted using Python's SciPy 1.5.4 library.

Image subdivision and stitching algorithm on the general dataset

All the images were cut into sub-frames of 2000 × 2000 pixels, the maximum size that can be supported by the GPU memory. During the subdivision process, some individuals were cut into several parts and some of them no longer appeared in their entirety. Only individuals whose partial bounding box represented more than 25% of the original surface area were kept. This limit was chosen because below this threshold, individuals are difficult to identify manually.

Only sub-frames containing animals were kept for training (Fig. 1). For the validation and the test sets, the cutting was done with an overlap of 50% on each edge of the sub-frames, and all sub-frames were kept. These steps were taken in order to avoid missing any individuals and to ensure that each individual would appear in its entirety in at least one sub-frame. Moreover, this approach allowed the predictions to be stitched into the initial image frame.

To both eliminate unnecessary partial bounding boxes and to reassemble the sub-frames, a stitching algorithm was constructed. Each image first undergoes a subdivision into overlapped sub-frames of 2000 × 2000 pixels. These sub-frames are then passed through the trained algorithm (i.e. model) to obtain predictions that contain bounding boxes, species names and confidence scores. The coordinates of the predicted bounding boxes of each sub-frame are then modified to be placed in the initial image plan. Next, the NMS (non-maximum suppression) algorithm is applied to a filter the predicted bounding boxes based on the IoU criteria (Everingham et al., 2010):

$$\mathrm{IoU} = \frac{\mathrm{area}(\mathrm{box}\, A \cap \mathrm{box}\, B)}{\mathrm{area}(\mathrm{box}\, A \cup \mathrm{box}\, B)}. \quad (1)$$

Here, a threshold of 0.5 was chosen. This high threshold was deliberately chosen in order to avoid missing some individuals in herds or some juveniles that are very close to their mothers (Appendix S2).

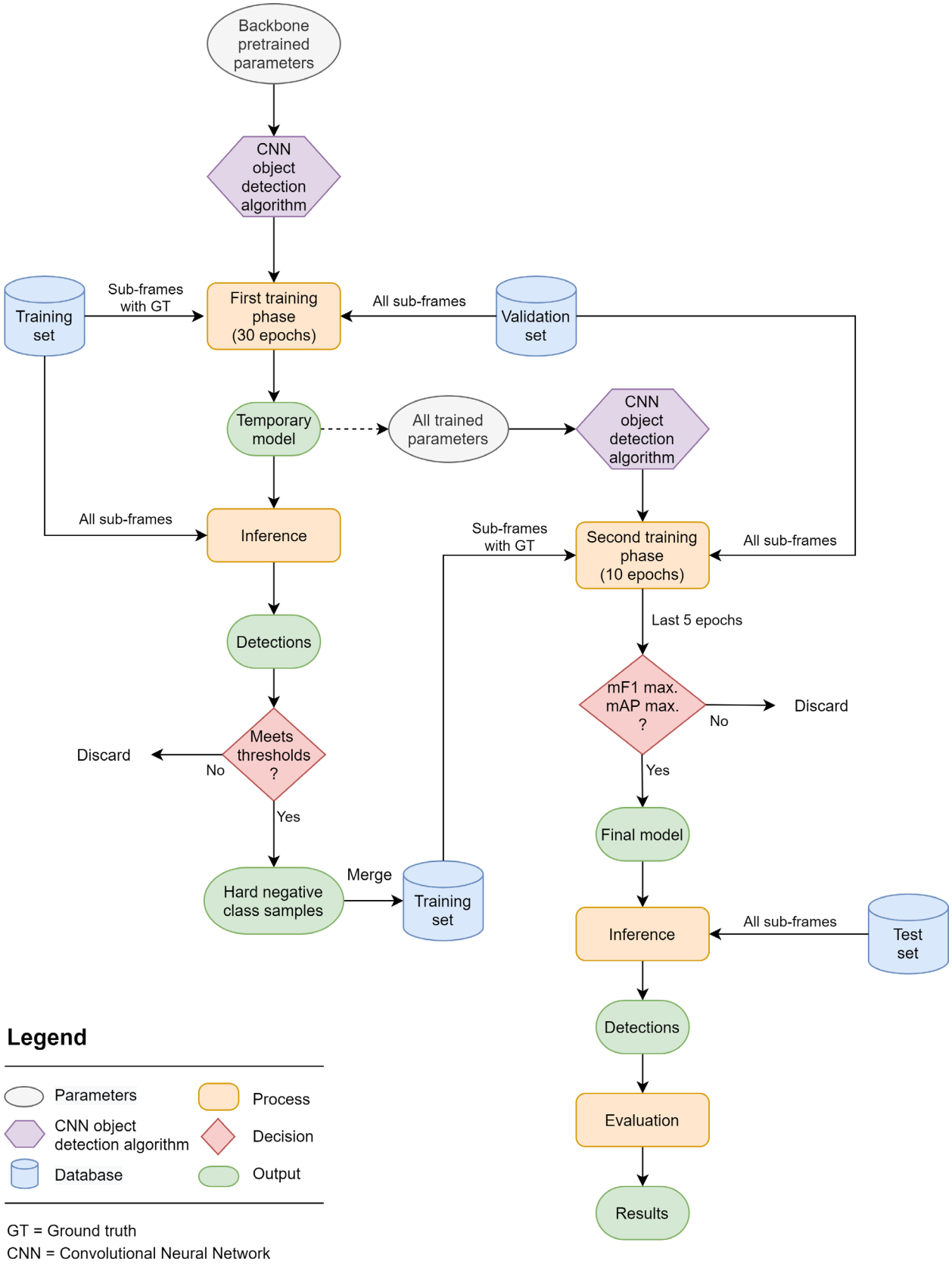

Figure 1. Flowchart of the methodology used to train, validate and test each of the three object detection algorithms, using the general dataset. The results after evaluation were then used for comparison.

Training on the general dataset

Because of the imbalance of the different classes, it is possible that during training, the majority species in terms of numbers dominate the others, leading to a decrease in the performance for minority species. Moreover, since the size of the training dataset is relatively small, training the different algorithms from scratch could lead to serious overfitting problems. To overcome these problems, four techniques were used: fine-tuning, data augmentation, class weighting and the hard negative class.

Fine-tuning

Each algorithm's backbone was initialized by pre-trained parameters (Fig. 1) on the ImageNet training dataset (Russakovsky et al., 2015). Next, all the parameters, except the first layer of the backbone, were trained on our dataset, with an adjustment of the number of classes at the head of the network.

Data augmentation

In addition to common strategies (i.e. rotation, mirroring and flipping, horizontal and vertical views), we used other strategies to detect animals in the various situations that can be encountered in aerial images: random blur, random contrast and random brightness.

Class weighting

Used by Kellenberger et al. (2018), this technique led to an improvement in animal detection performance. In our study, satisfactory results were obtained by weighting the species-related terms in the class loss function according to

$$w_i = \frac{\min(\{n_1, \ldots, n_i, \ldots, n_k\})}{n_i}, \quad (2)$$

where n_i is the number of annotations within a class i in the sub-frames training set, and k is the number of classes.

Hard negative class

The hard negative class (Peng et al., 2020) was used to limit the number of false positives (FP) (Fig. 1). This method treats hard negatives (high-scoring FPs) as foreground objects to make the model more sensitive to them. The score threshold was chosen to have a class size of between 2000 and 2200 to avoid a too-high class imbalance. Note that for the validation and test sets, the hard negative class-predicted bounding boxes were discarded and only species classes were maintained. Preliminary analysis showed that the inclusion of the hard negative class increased the models' performance (Appendix S3).

The training for the first training phase was done during 30 epochs with the Stochastic Gradient Descent as an optimizer (a momentum of 0.9 and a weight decay of 10^{-3}), and with a learning rate decreasing from 10^{-3} to 10^{-5} by steps of 10 epochs. The hard negative class was included from the 31st epoch (second training phase, Fig. 1, Appendix S4) and training continued for 10 more epochs with a learning rate of 10^{-4} for the first five epochs and 10^{-5} for the last five epochs.

Evaluation of CNN models

For each complete image, a comparison between ground truth and predictions was performed in order to determine the true positives (TP), FP and false negatives (FN).

A detection was then considered as a TP if the labels between the predicted bounding box and the ground-truth bounding box correspond and if the IoU between these boxes exceed 0.30. When several detections overlapped the same ground truth, the one with the maximum IoU was selected and the others were then considered as FP. Finally, if the labels did not match or if the ground truth was not detected by the model, the ground truth was considered as FN.

Precision/recall curves for each species were constructed to evaluate the performance of each model. These curves were calculated by varying the confidence score threshold associated with each predicted bounding box, between 0 and 1:

$$p(k) = \frac{n_{\mathrm{TP}}(k)}{n_{\mathrm{TP}}(k) + n_{\mathrm{FP}}(k)}, \quad (3)$$

$$r(k) = \frac{n_{\mathrm{TP}}(k)}{n_{\mathrm{TP}}(k) + n_{\mathrm{FN}}(k)}, \quad (4)$$

where p is the precision and r the recall, k is the confidence score threshold, and n_{TP}, n_{FP} and n_{FN} are the numbers of the TP, FP and FN, respectively.

$F1$ scores are usually used to define the combination of precision and recall that produces an optimal compromise between the number of FP and FN. An $F1$ score essentially represents the harmonic mean of precision and recall:

$$F1\,\mathrm{score} = \frac{2 \times p \times r}{p + r}, \quad (5)$$

where p and r are the precision and recall, respectively.

In this study, we used a mean $F1$ score (m$F1$): $F1$ scores were calculated for each species and then the whole was averaged. This metric was used because it gives an overall idea of the compromise between the FP and the FN.

The average precision (AP) (Everingham et al., 2010), representing the area under the precision/recall curve, was then calculated for each animal species in order to evaluate the performance of the detection algorithm in detecting a particular species. Finally, the mean average precision (mAP) was calculated to quantify the overall performance of each detection algorithm and thus allow their comparison. The mAP represents the average AP of all the species.

Each algorithm was trained for five runs with different fixed seeds. This step allowed us to control the stochastic aspect related to the training of an object detection algorithm. From these five runs, paired sample t-Student tests and confidence intervals were computed to compare the models and determine if the differences in performance were significantly different.

After each epoch, each trained algorithm (i.e. each model) was saved and tested on the stitched validation image set to verify that it was not falling into overfitting. In addition, for the last five epochs of each three algorithms of interest, the model with the best performance on the validation set was selected for testing (Fig. 1). To determine the best performance, the epoch with the maximum m$F1$ score was first selected. Next, the mAP corresponding to the epochs that presented an equivalent m$F1$ value (i.e. with two significant digits retained) was analyzed. From among these, the epoch with the highest mAP was finally selected for testing. This method enabled the selection of a globally efficient model (i.e. a high mAP) with a good compromise between FP and FN (i.e. a high m$F1$ value).

Processing of the case study dataset

To choose the model to apply to the case study (the Garamba dataset), we first selected the algorithm that showed the best performances on the test set. Then, we selected the best model based on the five tested runs, using the same selection method as in the validation set. The Garamba dataset's images were previously cut into subframes according to the same methodology as the validation and test sets (see Section 2.2.2). Detections were then stitched together according to an inference approach using the same stitching algorithm and evaluated using the same evaluation methodology as for the general dataset (see Section 2.2.3). Note that due to the high similarity between the two species (see Appendix S1) and the impossibility to distinguish them on UAV images, hartebeests and topis were merged into the same class during the inference step.

Results

Species detection

Topi, buffalo and kob were very well detected by all the trained algorithms (i.e. the three models) studied, with only slightly poorer results for elephants (Fig. 2). Given the results, warthogs and waterbucks appear to be more difficult to detect. Nevertheless, waterbucks were very well detected by Libra-RCNN (AP = 0.89) but very poorly detected by RetinaNet (AP = 0.01). RetinaNet was the model that had the most difficulty in detecting minority species (warthogs and waterbucks).

False positives were particularly high for elephants and warthogs, for all the models, as indicated by the poor precision at the highest recall value of these species (Fig. 2).

Libra-RCNN was the model that presented the highest AP for each of the species, except for elephants, where it equalled the AP of the Faster-RCNN model.

Model comparison

The results of the independent t-Student tests showed a significant difference in performance on the test set between the three models for mAP, m$F1$ and mean interspecies confusion, but not for recall. There was a significant difference in the FP/TP ratio for Faster-RCNN and Libra-RCNN with RetinaNet but not between Faster-RCNN and Libra-RCNN (see Appendix S5 for details).

The Libra-RCNN model produced the best mAP, the best m$F1$ score and the lowest average level of interspecies confusion in the test set (Fig. 3). In contrast, the RetinaNet model had the lowest mAP and m$F1$ values, along with the highest average interspecies confusion score. Finally, Faster-RCNN's performance ranks it between the other two.

Regarding the percentage of animals detected, all three models detected on average the same percentage of animals (true detection rate), with 94.5% ($\pm$0.5%) for Faster-RCNN, 94.3% ($\pm$0.5%) for Libra-RCNN and 94.6% ($\pm$0.3%) for RetinaNet, where the confidence intervals represent the 95% t-Student confidence interval (4 d.f.), computed from the results of the five seeds.

Finally, in terms of the FP/TP ratio (binary case), Libra-RCNN presented the lowest value (1.7 ± 0.2), closely followed by Faster-RCNN (1.8 ± 0.1). RetinaNet had the highest ratio with an average of nearly nine FN per TP (9.0 ± 0.7), a very high number of false alarms.

These results suggest that the Libra-RCNN model is more suitable for multi-species animal detection than the other two, with superior detection performance compared to Faster-RCNN and RetinaNet. Therefore, Libra-RCNN was selected and applied to the case study dataset.

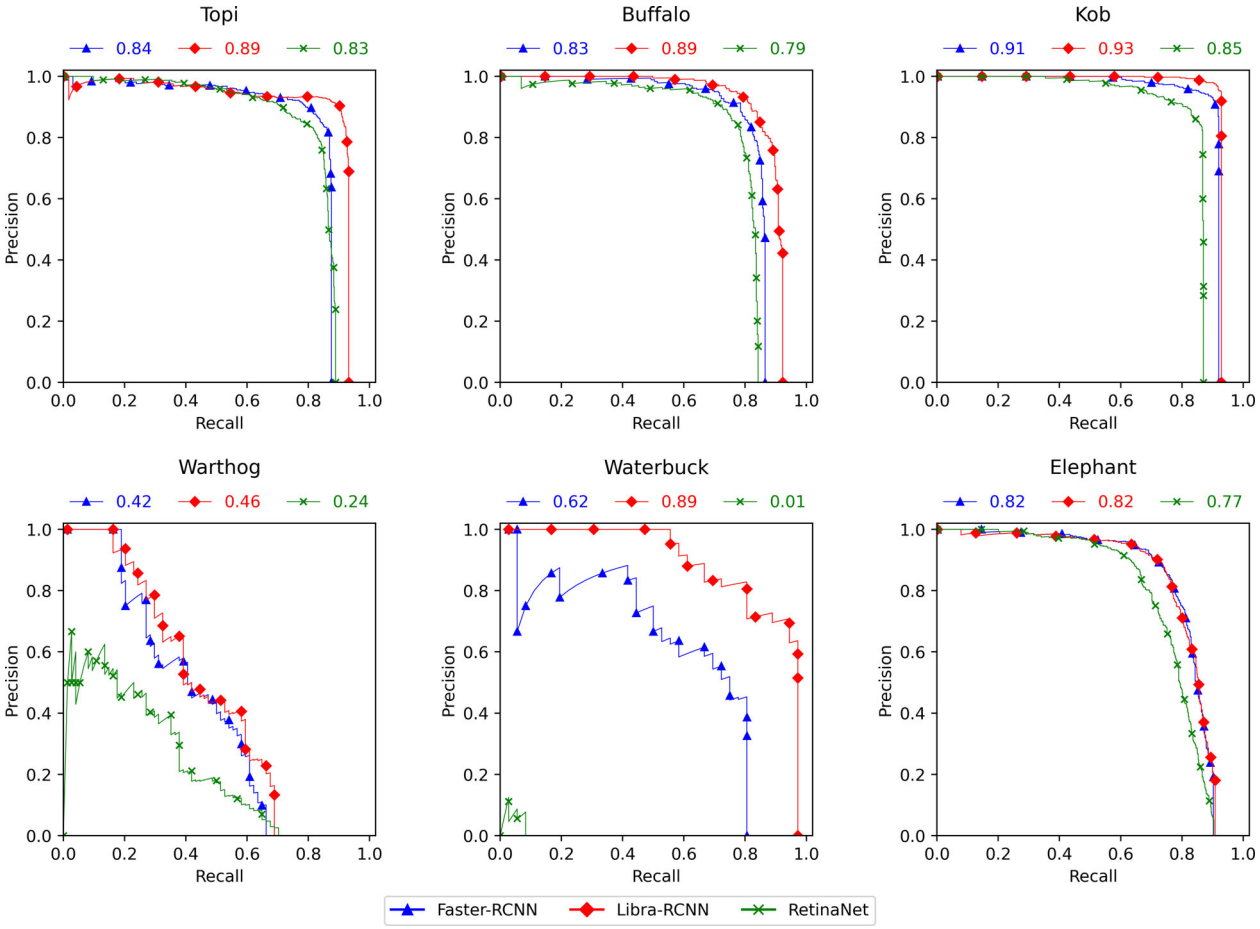

Figure 2. Precision/Recall curves of the three detection algorithms for the six targeted species on the test set. Axis legend represents the average precision (AP) of the corresponding curve. These curves were calculated for each of the algorithms using the model with the best mean average precision (mAP) among the five seeds.

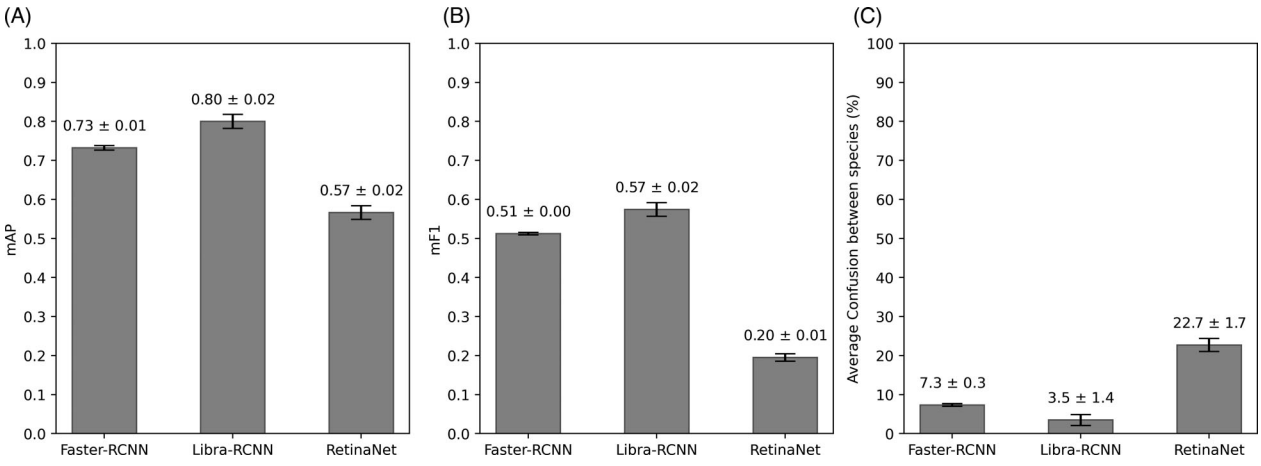

Figure 3. Bar plots of mAP (A), m*F*1 (B) and average interspecies confusion (C) calculated from the detection results of the test set. The error bars represent the 95% *t*-Student confidence interval (4 d.f.), computed from the results of the five seeds. mAP, mean average precision.

Case study (Garamba dataset)

To evaluate the performance of the best-developed model, the Libra-RCNN model was applied to the Garamba dataset. The total processing time (with a single GPU) was 23h26 for all the flight images, with an average of 12 s/image. Detections were present in 9% of the images ($\frac{607}{7034} \cong 0.09$), and among the 180 images containing ground truths, 9% were missed by the model ($\frac{16}{180} \cong 0.09$). However, no or almost no images were missed for some species (Table 3). For all six targeted species, 73% were correctly identified, with a relatively wide variation between the species. Furthermore, 64 individuals were correctly detected but misidentified. The same trend in species detection as observed on the test set results can be observed here as well: the majority species are better detected and identified than the minority species (Table 3).

Among the 305 individuals of other species initially identified during the annotation step, 43 were found by the model: 29 hippopotamuses out of 196, 13 giraffes out of 43 and 1 undetermined species out of 7. In addition, all FP with an IoU of 0 with the ground truths were reviewed; among these 945 FP detections, 133 were in fact individuals of our six targeted species that were missed during the annotation phase. Of these, 55 were correctly identified by the model (Table 3).

Discussion

The Libra-RCNN model showed better detection performance on the test set than other published models dealing with the detection of mammals in similar habitats and landscapes (Eikelboom et al., 2019; Kellenberger et al., 2018; Rey et al., 2017). Moreover, the models presented here were able to differentiate six animal species on nadir aerial images, which to the best of our knowledge has never been tried before in the literature. The performance of our best model (Libra-RCNN) on the test set surpasses that of the latest multi-species model published (Eikelboom et al., 2019) in terms of global recall, global FP/TP ratios, mAP and F1 scores. Finally, it showed good performance on a complete independent raw dataset from another park (i.e. Garamba) and was able to detect additional individuals, some belonging to other species.

Species detection

Our best model, the Libra RCNN, showed very good detection, identification and generalization results for the majority species (topi, buffalo and kob) and was even surprisingly good at detecting one minority species, the waterbuck. For topi, buffalo and elephant detection, we observed that all three models were less precise for herds. The lower precision was mainly due to the overlap of the bounding boxes within the herds (Fig. 4). Indeed, this box overlap probably made it more difficult for the algorithms to converge during training. In images containing herds, a large number of boxes were therefore created during the inference step, but despite the application of the NMS, some detections persisted and were therefore qualified as FP, as several boxes defined the same individual. After revision, herding represented about 40% of the FPs for the Libra-RCNN model on the test set, and about 41% for that model on the Garamba dataset.

Table 3. Results of the Libra-RCNN model applied on the case study dataset (Garamba) for the six targeted species: hartebeest (considered as topi due to high similarity), buffalo, kob, warthog, waterbuck and elephant.

	Number of images			Individuals					Number of false positives		
Species	With GT	With detections[1]	Missed	GT	Recall	Precision	F1 score	Misclassified	Total	Human missed[2]	Other species[3]
Hartebeest	29	102	0	151	0.59	0.34	0.43	45	174	4	1
Buffalo	55	148	5	547	0.87	0.44	0.58	7	620	19	10
Kob	62	95	1	321	0.67	0.62	0.64	5	133	26	6
Warthog	24	158	9	82	0.40	0.09	0.14	0	349	6	18
Waterbuck	10	122	1	14	0.14	0.01	0.03	7	144	0	6
Elephant	0	54	0	0	n/a	n/a	n/a	0	171	0	2
All	180	607	16	1115	0.73	0.34	0.46	64	1591	55	43

The last row corresponds to the results of the whole set of six species. Note that the number of images with detections considering all six species (last row) is not equal to the sum of the images with detections by species. This difference is due to the fact that the model sometimes detected several species within the same image, and so these images appear in the detected images of multiple species. GT, ground truth; n/a, not applicable (since this species is absent from the dataset).

[1]Number of images with detections (i.e. that contain predictions).

[2]Number of false positives that were in fact animals missed by humans during annotation, but correctly detected and identified by the model.

[3]Number of other species not belonging to the set of six targeted species (i.e. hippopotamus, giraffe and unknown), but detected by the model.

In addition, for elephants, the images were taken at any time of the day, unlike the other datasets. This led to greater variability of shadows, colours and brightness within the images, and thus to poorer detection results, as observed in Rey et al. (2017). Moreover, this dataset (AED) comes from parks and reserves with varying landscapes and terrain features that differ from those of Virunga, such as denser tree cover in some images. However, training the models on these field variations normally made them more robust to heterogeneous terrain features (Kellenberger et al., 2018).

This difference in the landscape also explains the lower percentage of animals detected by Libra-RCNN on the Garamba dataset. In addition, we observed that these differences in terrain features were also the cause of many FPs within the Garamba images. For example, the model detected a large number of termite mounds as animals. This terrain characteristic was indeed much less present in the training set.

Despite the class weighting during training, the models struggled to correctly identify warthogs, most probably due to a lack of training samples. Furthermore, this animal was the smallest mammal in this study. Its small size generated a large number of FP due to insufficient pixel resolution and because some acquisition drawbacks (blur, contrast) did not allow the model to distinguish some of this small mammal's attributes. It could therefore easily be mistaken for small rocks, common in the African landscapes where this species is found.

The surprisingly high number of the Garamba dataset's FPs (133) that were in fact real animals can be explained by the overlap of the images and the initial methodology of annotating the individuals. Indeed, only 84 individuals were actually real human-missed animals. The other animals had already been tagged during the manual annotation phase in the previous frame or would be in the next frame within a succession of overlapping images. Annotating everything was not required, although recommended for the purpose of that specific survey, as double counting was not desired at the time of the census, despite the possibilities to differentiate between first observation and double counting in the software.

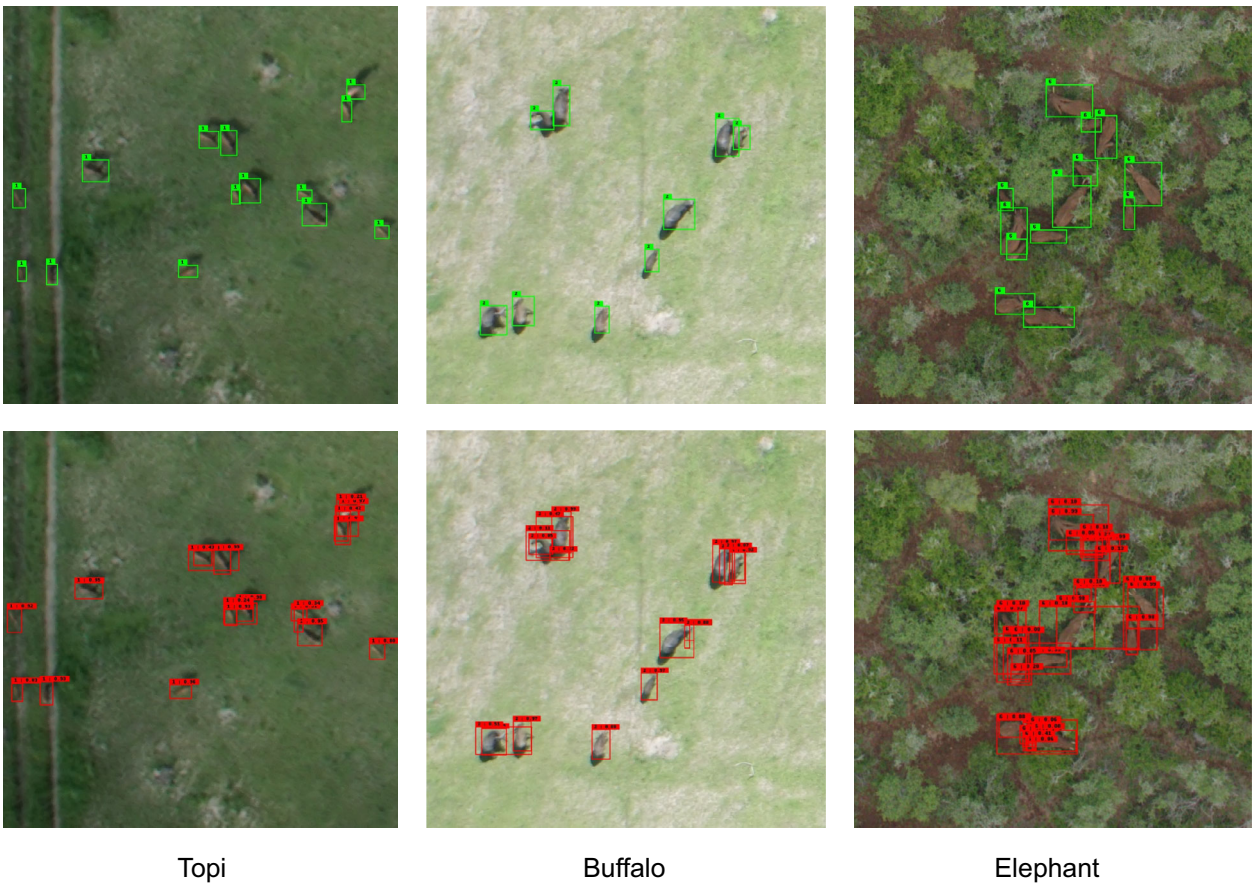

Figure 4. Detections examples of the Libra-RCNN model, on partial test images showing the major cause of the high number of false positives. Note that ground truths are in green (first row) and detections are in red (second row).

Consequently, attention was focused on the individuals that had not yet been tagged, and so sometimes the individuals present in the periphery were not tagged again. From these 84 new individuals, 55 were correctly identified by the model.

Model comparison

Two-stage detection models (Faster-RCNN and Libra-RCNN) seemed to detect animal species more precisely than a single-stage model (RetinaNet). This difference in performance was as expected (Soviany & Ionescu, 2018).

The Faster-RCNN and Libra-RCNN models were very similar in terms of their detection performance. The differences that we observed between these two models on the test set (Fig. 3) were probably due to the Libra-RCNN L1-balanced loss and its rebalancing at the training sample distribution level. These components caused the algorithm to focus on difficult cases during training, which leads to better detection and classification performances (Pang et al., 2019).

Operational implications

The Libra-RCNN model presents interesting perspectives as a good semi-automatic detection and identification tool for African mammal species. It could be used in practice to save human time, create new training data and establish initial, rapid population counts, with human verification of detected individuals as post-processing. However, our experience in reviewing the FPs shows that this screening must necessarily be performed with the animals' surrounding context, which is crucial for decision making by the human eye.

The model developed here can mainly be applied in open savanna or sparsely wooded areas and for the detection of our six studied species. Indeed, our results show that in order to develop a model that can be used in various ecosystems, it would be necessary to have a training

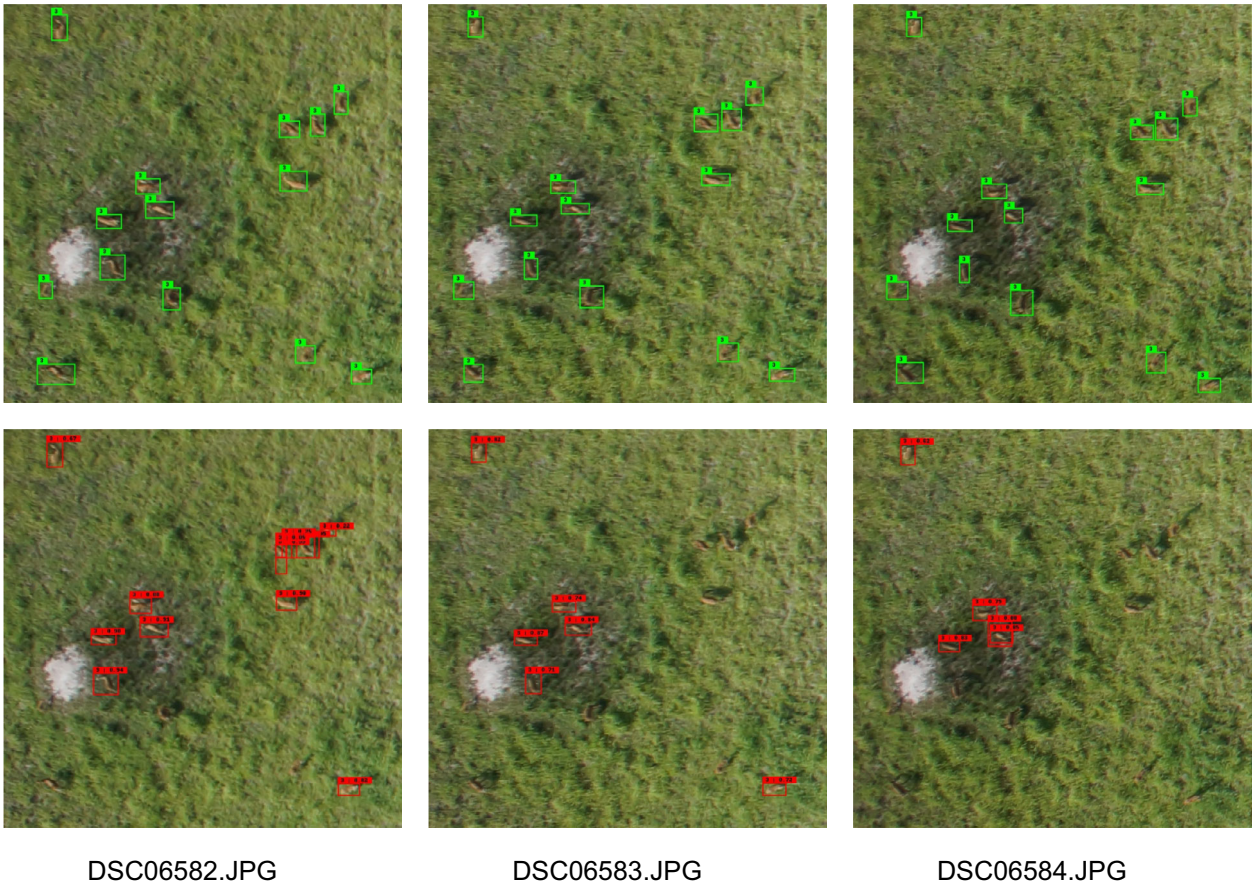

Figure 5. Kob detections of the Libra-RCNN model on consecutive Garamba's partial images, showing that the images overlap made it possible to detect a maximum of individuals thanks to the slight viewing angle changes. Note that ground truths are in green (first row) and detections are in red (second row).

set with a large variability of landscapes and terrain features.

Generally, detection performance improves when more training data are used. Unfortunately, the acquisition and pre-processing of aerial animal training data are costly. Developing a semi-automatic animal detection tool, such as those presented here, requires significant upstream work. From the manual identification and location of animals to the dataset training, the workload is quite large and requires significant human resources with highly technical skills. Moreover, as with any deep-learning application, training an algorithm requires a large computing capacity and a huge amount of data. Luckily, more and more open-source data (images and annotations) are being made available (Eikelboom et al., 2019; Kellenberger et al., 2018; Naudé & Joubert, 2019a, 2019b).

Finally, in an attempt to automate the counting of individuals, the thorny problem of images overlap remains an obstacle. Our results from Garamba were presented here without accounting for multiple detections. We observed that this overlap is crucial to detect all possible individuals. Indeed, in Garamba, some individuals were only detected in a few images thanks to a slight change in the viewing angle (Fig. 5). This need for overlap leads the model to slightly overestimate the number of real FN.

Research perspectives

In surveying animal species, the problem of class imbalance will always be present due to the natural distribution of species within ecosystems. Nevertheless, more and more studies are looking into this recurrent problem in multi-class object detection (Oksuz et al., 2020). There is also the challenging problem of the large number of FP. Newer methods, such as synthetic data generation, could help to address this problem by generating images with heterogenous backgrounds (Beery et al., 2020). In addition, it could be beneficial to consider switching from boxes to points (Ribera et al., 2019) or masks (Xu et al., 2020) to avoid the problems of overlapping boxes in herds and in an attempt to automate the counting. These solutions should be investigated in future works.

Acknowledgements

The work of Alexandre DELPLANQUE was supported by the Fund for Research Training in Industry and Agriculture (FRIA, F.R.S.—FNRS). This project was financed in part by the Ministère de l'Économie et de l'Innovation (MEI) of the province of Québec. We would like to thank the teams involved in the collection of the aerial images in the different areas and the observers who produced the ground truth associated with these images. Special thanks go to all those involved in the acquisition of data in the DRC under the Forest and Climate Change in Congo project (FCCC/2014-2016) funded by the European Union (EU, Grant Number DCI-ENV/2012/309-143) and granted by the Center for International Forestry Research (CIFOR). Data acquisition in the field was greatly facilitated by the support of R&D office, the ICCN (Institut Congolais pour la Conservation de la Nature), African Parks Network (Garamba NP) and the Virunga Foundation. We would also like to thank Simon LHOEST and Cédric VERMEULEN for reviewing the manuscript and bringing their expertise to the discussion of the results.

Funding Information

The work of Alexandre DELPLANQUE was supported by the Fund for Research Training in Industry and Agriculture (FRIA, F.R.S.—FNRS). This project was financed in part by the Ministère de l'Économie et de l'Innovation (MEI) of the province of Québec. Special thanks go to all those involved in the acquisition of data in the DRC under the Forest and Climate Change in Congo project (FCCC/2014-2016) funded by the European Union (EU, Grant Number DCI-ENV/2012/309-143) and granted by the Center for International Forestry Research (CIFOR).

Data Availability Statement

Our image data, annotations, results files and Jupyter Notebooks are available on a ULiège, Gembloux Agro-Bio Tech server, at the repository: http://gxgfservplateforme.gxabt.ulg.ac.be/garamba/.

References

Almond, G.M. & Petersen, T. (2020) *Living planet report 2020 - bending the curve of biodiversity loss*. Gland, Switzerland: WWF.

Barbedo, J.G.A., Koenigkan, L.V., Santos, T.T. & Santos, P.M. (2019) A study on the detection of cattle in UAV images using deep learning. *Sensors*, **19**(24), 5436. https://doi.org/10.3390/s19245436

Beery, S., Liu, Y., Morris, D., Piavis, J., Kapoor, A., Joshi, N. et al. (2020) Synthetic examples improve generalization for rare classes. In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. Snowmass, CO: Institute of Electrical and Electronics Engineers (IEEE), pp. 852–862. https://doi.org/10.1109/WACV45572.2020.9093570

Ceballos, G., Ehrlich, P.R., Barnosky, A.D., Garcïa, A., Pringle, R.M. & Palmer, T.M. (2015) Accelerated modern human-induced species losses: entering the sixth mass extinction. *Science Advances*, **1**(5), e1400253. https://doi.org/10.1126/sciadv.1400253

Chabot, D. & Bird, D.M. (2015) Wildlife research and management methods in the 21st century: where do unmanned aircraftt in? *Journal of Unmanned Vehicle Systems*, **3**(4), 137–155. https://doi.org/10.1139/juvs-2015-0021

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X. et al. (2019) MMDetection: open MMLab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155.

Craigie, I., Baillie, J., Balmford, A., Carbone, C., Collen, B., Green, R. et al. (2010) Large mammal population declines in Africa's protected areas. *Biological Conservation*, **143**, 2221–2228. https://doi.org/10.1016/j.biocon.2010.06.007

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. https://doi.org/10.1111/2041-210X.13277.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J. & Zisserman, A. (2010) The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, **88**(2), 303–338. https://doi.org/10.1007/s11263-014-0733-5

Gaidet-Drapier, N., Fritz, H., Bourgarel, M., Renaud, P.-C., Poilecot, P., Chardonnet, P. et al. (2006) Cost and efficiency of large mammal census techniques: comparison of methods for a participatory approach in a communal area, Zimbabwe. *Biodiversity and Conservation*, **15**(2), 735–754. https://doi.org/10.1007/s10531-004-1063-7

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*. pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

Hetem, R.S., Fuller, A., Maloney, S.K. & Mitchell, D. (2014) Responses of large mammals to climate change. *Temperature*, **1**(2), 115–127. https://doi.org/10.4161/temp.29651

IPCC. (2014) Climate change 2014: synthesis report. In: Pachauri, R. & Meyer, L. (Eds.) *Contribution of working groups I, II and III to the fifth assessment report of the intergovernmental panel on climate change*. Geneva, Switzerland: IPCC, p. 151.

Isbell, F., Gonzalez, A., Loreau, M., Cowles, J., Dıaz, S., Hector, A. et al. (2017) Linking the influence and dependence of people on biodiversity across scales. *Nature*, **546**(7656), 65–72. https://doi.org/10.1038/nature22899

Jachmann, H. (1991) Evaluation of four survey methods for estimating elephant densities. *African Journal of Ecology*, **29**, 188–195. https://doi.org/10.1111/j.1365-2028.1991.tb01001.x

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. https://doi.org/10.1016/j.rse.2018.06.028

Lacher, T.E., Davidson, A.D., Fleming, T.H., Gómez-Ruiz, E.P., McCracken, G.F., Owen-Smith, N. et al. (2019) The functional roles of mammals in ecosystems. *Journal of Mammalogy*, **100**(3), 942–964. https://doi.org/10.1093/jmammal/gyy183

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. & Belongie, S. (2017) Feature pyramid networks for object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 936–944. https://doi.org/10.1109/CVPR.2017.106

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*. pp. 2999–3007. https://doi.org/10.1109/ICCV.2017.324

Linchant, J., Lhoest, S., Quevauvillers, S., Lejeune, P., Vermeulen, C., Ngabinzeke, J.S. et al. (2018) UAS imagery reveals new survey opportunities for counting hippos. *PLoS One*, **13**(11), 1–17. https://doi.org/10.1371/journal.pone.0206413

Linchant, J., Lhoest, S., Quevauvillers, S., Semeki, J., Lejeune, P. & Vermeulen, C. (2015) WIMUAS: developing a tool to review wildlife data from various UAS flight plans. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, **40**, 379–384. https://doi.org/10.5194/isprsarchives-XL-3-W3-379-2015

Linchant, J., Lisein, J., Ngabinzeke, J., Lejeune, P. & Vermeulen, C. (2015) Are unmanned aircraft systems (UAS) the future of wildlife monitoring? A review of accomplishments and challenges. *Mammal Review*, **45**, 239–252. https://doi.org/10.1111/mam.12046

Mayaux, P., Bartholomé, E., Fritz, S. & Belward, A. (2004) A new land-cover map of Africa for the year 2000. *Journal of Biogeography*, **31**(6), 861–877. https://doi.org/10.1111/j.1365-2699.2004.01073.x

Moreni, M., Theau, J. & Foucher, S. (2021) Train fast while reducing false positives: improving animal classification performance using convolutional neural networks. *Geomatics*, **1**(1), 34–49. https://doi.org/10.3390/geomatics1010004

Mulero-Pázmány, M., Stolper, R., Van Essen, L., Negro, J.J. & Sassen, T. (2014) Remotely piloted aircraft systems as a rhinoceros antipoaching tool in Africa. *PLoS One*, **9**(1), 1–10. https://doi.org/10.1371/journal.pone.0083873

Naudé, J. & Joubert, D. (2019a) The aerial elephant dataset: a new public benchmark for aerial object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 48–55.

Naudé, J. & Joubert, D. (2019b) Data from: the aerial elephant dataset: a new public benchmark for aerial object detection. Zenodo. https://doi.org/10.5281/zenodo.3234780

Norton-Griffiths, M. (1978) *Counting animals: revised second edition. Handbook no. 1. serengeti ecological monitoring programme*. Nairobi, Kenya: African Wildlife Leadership Foundation.

Oksuz, K., Cam, B.C., Kalkan, S. & Akbas, E. (2020) Imbalance problems in object detection: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1. https://doi.org/10.1109/TPAMI.2020.2981890

Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W. & Lin, D. (2019) Libra R-CNN: towards balanced learning for object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 821–830. https://doi.org/10.1109/CVPR.2019.00091

Peng, J., Wang, D., Liao, X., Shao, Q., Sun, Z., Yue, H. et al. (2020) Wild animal survey using UAS imagery and deep learning: modified Faster R-CNN for kiang detection in Tibetan Plateau. *ISPRS Journal of Photogrammetry and Remote Sensing*, **169**, 364–376. https://doi.org/10.1016/j.isprsjprs.2020.08.026

Ren, S., He, K., Girshick, R. & Sun, J. (2017) Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Rey, N., Volpi, M., Joost, S. & Tuia, D. (2017) Detecting animals in African Savanna with UAVs and the crowds. *Remote Sensing of Environment*, **200**, 341–351. https://doi.org/10.1016/j.rse.2017.08.026

Ribera, J., Guera, D., Chen, Y. & Delp, E.J. (2019) Locating objects without bounding boxes. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6479–6489.

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. https://doi.org/10.1007/s11263-015-0816-y

Soviany, P. & Ionescu, R.T. (2018) Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction. In *Proceedings of the 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*. pp. 209–214. https://doi.org/10.1109/SYNASC.2018.00041

Thuiller, W., Broennimann, O., Hughes, G., Alkemade, J.R.M., Midgley, G.F. & Corsi, F. (2006) Vulnerability of African mammals to anthropogenic climate change under conservative land transformation assumptions. *Global Change Biology*, **12**(3), 424–440. https://doi.org/10.1111/j.1365-2486.2006.01115.x

Vermeulen, C., Lejeune, P., Lisein, J., Sawadogo, P. & Bouché, P. (2013) Unmanned aerial survey of elephants. *PLoS One*, **8**(2), 1–7. https://doi.org/10.1371/journal.pone.0054700

Watts, A.C., Perry, J.H., Smith, S.E., Burgess, M.A., Wilkinson, B.E., Szantoi, Z. et al. (2010) Small unmanned aircraft systems for low-altitude aerial surveys. *The Journal of Wildlife Management*, **74**(7), 1614–1619. https://doi.org/10.1111/j.1937-2817.2010.tb01292.x

Witmer, G.W. (2005) Wildlife population monitoring: some practical considerations. *Wildlife Research*, **32**(3), 259–263. https://doi.org/10.1071/WR04003

Xu, B., Wang, W., Falzon, G., Kwan, P., Guo, L., Chen, G. et al. (2020) Automated cattle counting using Mask R-CNN in quadcopter vision system. *Computers and Electronics in Agriculture*, **171**, 105300. https://doi.org/10.1016/j.compag.2020.105300

Zhao, Z.-Q., Zheng, P., Xu, S.-T. & Wu, X. (2019) Object detection with deep learning: a review. *IEEE Transactions on Neural Networks and Learning Systems*, **30**(11), 3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F. et al. (2017) Deep learning in remote sensing: a comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, **5**(4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307

Supporting Information

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

Appendix S1. Image samples of the six targeted species.
Appendix S2. Preliminary tests for NMS (non-maximum suppression) threshold.
Appendix S3. Comparison of models' performances with and without the inclusion of the hard negative class.
Appendix S4. Detection algorithms' training details.
Appendix S5. Independent t-Student test results on the test set (general dataset).

