ISPRS Journal of Photogrammetry and Remote Sensing 197 (2023) 167–180

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

From crowd to herd counting: How to precisely detect and count African mammals using aerial imagery and deep learning?

Alexandre Delplanque [a,*], Samuel Foucher [b], Jérôme Théau [b,c], Elsa Bussière [d], Cédric Vermeulen [a], Philippe Lejeune [a]

[a] TERRA Teaching and Research Centre (Forest Is Life), ULiège, Gembloux Agro-Bio Tech, 2 Passage des Déportés, Gembloux 5030, Belgium
[b] Departement of Applied Geomatics, Université de Sherbrooke, 2500 Boulevard de l'Université, Sherbrooke, QC J1K 2R1, Canada
[c] Quebec Centre for Biodiversity Science (QCBS), Stewart Biology, McGill University, Montréal, QC H3A 1B1, Canada
[d] Wild Sense, PostNet #270, Private Bag X11326, Nelspruit 1200, South Africa

ARTICLE INFO

Keywords:
Deep learning
Livestock
Herd
Convolutional neural networks
Aerial survey
Protected area

ABSTRACT

Rapid growth of human populations in sub-Saharan Africa has led to a simultaneous increase in the number of livestock, often leading to conflicts of use with wildlife in protected areas. To minimize these conflicts, and to meet both communities' and conservation goals, it is therefore essential to monitor livestock density and their land use. This is usually done by conducting aerial surveys during which aerial images are taken for later counting. Although this approach appears to reduce counting bias, the manual processing of images is time-consuming. The use of dense convolutional neural networks (CNNs) has emerged as a very promising avenue for processing such datasets. However, typical CNN architectures have detection limits for dense herds and close-by animals. To tackle this problem, this study introduces a new point-based CNN architecture, HerdNet, inspired by crowd counting. It was optimized on challenging oblique aerial images containing herds of camels (*Camelus dromedarius*), donkeys (*Equus asinus*), sheep (*Ovis aries*) and goats (*Capra hircus*), acquired over heterogeneous arid landscapes of the Ennedi reserve (Chad). This approach was compared to an anchor-based architecture, Faster-RCNN, and a density-based, adapted version of DLA-34 that is typically used in crowd counting. HerdNet achieved a global F1 score of 73.6 % on 24 megapixels images, with a root mean square error of 9.8 animals and at a processing speed of 3.6 s, outperforming the two baselines in terms of localization, counting and speed. It showed better proximity-invariant precision while maintaining equivalent recall to that of Faster-RCNN, thus demonstrating that it is the most suitable approach for detecting and counting large mammals at close range. The only limitation of HerdNet was the slightly weaker identification of species, with an average confusion rate approximately 4 % higher than that of Faster-RCNN. This study provides a new CNN architecture that could be used to develop an automatic livestock counting tool in aerial imagery. The reduced image analysis time could motivate more frequent flights, thus allowing a much finer monitoring of livestock and their land use.

1. Introduction

In sub-Saharan Africa, the rapid growth of the human population over the last decades, combined with very effective sanitary actions on herds, has led to a significant increase in the number of heads of different livestock species (Richard et al., 2019). On the one hand, excessive livestock density can have several adverse effects on the environment, such as soil and vegetation degradation, space and grazing competition with wildlife or spread of diseases (Bengis et al., 2004; Butt & Turner, 2012; De Leeuw et al., 2001; Georgiadis et al., 2007; Vandermeer, 2002). On the other hand, livestock is a major source of income and a livelihood strategy for rural populations (Herrero et al., 2013), and it can enhance agricultural sustainability (Ayantunde et al., 2018) and habitat quality for wildlife if well managed (Fynn et al., 2016). Too-high density of livestock may prompt conflicts over important natural resources within a protected area, such as pastures used for grazing by wild and domestic herbivores (Scholte et al., 2022a; Scholte et al., 2022b; Toutain et al., 2004). Knowledge of livestock density and land use in these areas is therefore necessary to reach both conservation and local communities' goals.

* Corresponding author.
E-mail address: alexandre.delplanque@uliege.be (A. Delplanque).

https://doi.org/10.1016/j.isprsjprs.2023.01.025
Received 12 August 2022; Received in revised form 24 January 2023; Accepted 31 January 2023
Available online 8 February 2023

In large open African areas, livestock and wildlife counting are often carried out by a piloted aircraft, flying at low altitude and following systematic transects while observers count animals in sample strips defined on each side of the aircraft (Caughley, 1977; Grimsdell & Westley, 1981; Norton-Griffiths, 1978). Unfortunately, observers tend to fail to detect and accurately count the true number of animals in the strips, especially when encountering large and dense herds, resulting in biased population estimates (Caughley, 1974; Grimsdell & Westley, 1981; Jachmann, 2002).

For most observers, remote counting from an aircraft becomes inaccurate for groups of 15 or more individuals (Grimsdell & Westley, 1981; Norton-Griffiths, 1978). Photographing large herds has thus become a common practice to improve group size estimates by subsequent counting (Bouché et al., 2012; Craig, 2012; Grimsdell & Westley, 1981; Norton-Griffiths, 1978; Schlossberg et al., 2016). Recently, the use of oblique cameras has been shown to improve wildlife counts, especially for smaller species such as warthog (*Phacochoerus africanus*), Uganda kob (*Kobus kob*), or oribi (*Ourebia ourebi*) (Lamprey et al., 2020a; Lamprey et al., 2020b). Although nadir imagery is increasingly used for aerial survey of wildlife since the growing interest for drones (Linchant et al., 2015), oblique imagery remains a relevant and particularly attractive solution for managers of large protected areas. Oblique imagery has the following advantages over nadir imagery, making it a key research area: the better detection of animals under trees, the better identification of species (side view), the larger sampling area at a same flight height, and the similar viewing configuration with onboard observers (facilitation of detection validation). However, the main drawbacks of this method are: 1) the high volume of imagery generated; and 2) the associated intensive photo-interpretation workload. For instance, Lamprey et al. (2020a) acquired 24,000 images for a survey of a 5037 km^2 reserve in Uganda, and it took 6 weeks for 4 people to interpret the images.

Deep learning architectures, through the use of Convolutional Neural Networks (CNNs), now offer the possibility to semi-automatically detect and identify species in aerial images acquired in heterogeneous landscapes using object detection approaches (Delplanque et al., 2022; Eikelboom et al., 2019; Kellenberger et al., 2017, 2018, 2019a; Naudé and Joubert, 2019; Peng et al., 2020; Torney et al., 2019). These recent approaches allow partially-automated processing of the large volumes of images generated during acquisition campaigns. While these seem to work relatively well for isolated individuals or sparse herds, the case of dense herds remains a complex and challenging task (Delplanque et al., 2022).

In oblique images containing dense herds, factors such as mutual occlusions, close-by bodies, complex background, varying scales, and non-uniform distribution of individuals make common object detection approaches cumbersome if not impossible to accurately locate and count the individuals. Common object detectors are usually anchor-based, meaning that they use anchors during the training process, which are a set of prior box proposals with different scales and aspects centered on potential object locations (Ren et al., 2015). Usually, anchors help the network to converge faster and to obtain better detection performance (Ren et al., 2015; Liu et al., 2016; Lin et al., 2017; Redmon & Farhadi., 2017). However, they are suspected to be the cause of decreased precision in dense herd situations (Delplanque et al., 2022).

The factors mentioned above (i.e., occlusion, complex background, scale variation and non-uniform distribution) are also encountered in crowd detection (Gao et al., 2020), making the task of herd counting very similar to that of crowd counting. While the CNN architectures developed in crowd counting have shown very good results for human counting in densely populated scenes, their transposition to dense terrestrial mammal herd counting in oblique imagery has not yet been explored.

Density-map-based architectures, first proposed by Lempitsky and Zisserman (2010), are popular in crowd counting, due mainly to their improved counting performance compared to detection-based and anchor-based architectures, and for the practicality of dot annotations (Li et al., 2021). Padubidri et al. (2021) have recently shown that density maps can be used to precisely count Steller sea lions (*Eumetopias jubatus*) and African bush elephants (*Loxodonta africana*) in nadir aerial images. Kellenberger et al. (2019a) also proposed density-based approaches that showed great performances using only image-level annotations. However, density-based approaches did not precisely locate individuals in the images, especially in herds; such location capability could be valuable for creating new annotations from unseen images.

This paper presents "HerdNet", a new dense herd CNN-based counting approach, inspired and adapted from crowd counting approaches, which was compared with an anchor-based and density-based baselines.

2. Background

2.1. Pointing, a more natural and efficient way for herd counting

In addition to being a natural way to count objects for humans, pointing is faster than drawing bounding boxes, especially when large numbers of objects are encountered, as in the case of animal herds. Pointing was first proposed by Lempitsky and Zisserman (2010), who presented it as a very attractive and understudied case. Since then, point annotations have been largely used for labeling crowds in images (Li et al., 2021). In recent years, some CNN point-based approaches have also emerged with promising results. While crowd counting CNN architectures generally use points for density map regression, CNN point-based object detectors are often trained to produce a high-resolution map in an encoder-decoder fashion, where points can then be extracted (Ribera et al., 2019; Zhou et al., 2019). An encoder-decoder framework outputs features over the input image's pixel space to obtain precise localization. The encoder block encodes the images into multi-level features' maps of different resolution (i.e., the down-sampling phase), and then the decoder block decodes the encoded features' map while keeping their spatial information (i.e., the up-sampling phase). Other methods also showed that point detection can be achieved on lower-resolution outputs using a simple encoder (i.e., a CNN) but at the expense of a lower position accuracy (Kellenberger et al., 2018; Kellenberger et al., 2019b; Kellenberger et al., 2021).

2.2. Similarities between crowd and herd counting tasks

In crowd counting, there are some challenges that make the task complex, including occlusion, complex background, scale variation, and non-uniform distribution (Gao et al., 2020). These issues are also encountered in herd counting within oblique aerial imagery, which makes the task of herd counting very similar to that of crowd counting (see Fig. 1):

- **Occlusion** (Fig. 1a) - As the herd density increases, the animals will appear to partially occlude each other. This situation is often observed for gregarious and migratory animals which can be grouped around particular places such as watering holes and resource points, and during some practices such as "tightly bunched herding" (Odadi et al., 2018). Such occlusions could limit the performance of traditional object detection architectures.
- **Complex background** (Fig. 1b) - Aerial survey imagery contains mainly background regions that can include many confusing objects (e.g. shadows, rocks). These can lead to a high number of false alarms and bias the counting result.
- **Scale variation** (Fig. 1c) - In oblique aerial images, the size of animals varies both within the same species by the distance from the camera (i.e., intraspecies variation) and between different species (i.e., interspecies variation), increasing the difficulty for accurate detection and identification.

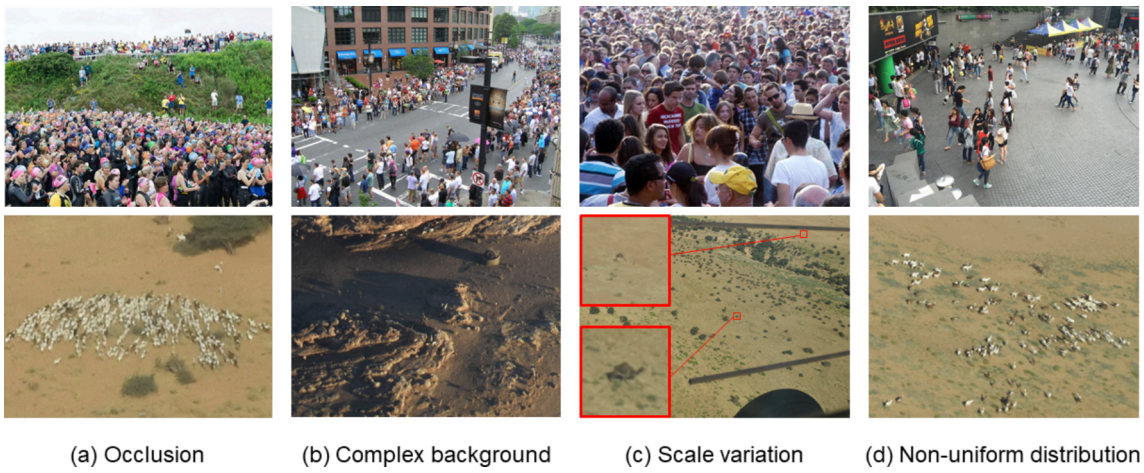

Fig. 1. Examples of challenges faced by crowd counting (top row), extracted from the Shanghaitech dataset (Zhang et al., 2016) and their equivalents in herd counting (bottom row), extracted from the Ennedi dataset.

- **Non-uniform distribution** (Fig. 1d) - Diverse herd distributions and densities may be encountered. The difficulty is further accentuated by the fact that the dataset is dominated by samples containing few individuals, following the patch generation (see Section 3.4.1).

These similarities make crowd counting CNN architectures an interesting approach to tackle the challenges of counting dense herds in oblique aerial imagery. However, unlike crowd counting where the problem is binary (human vs. background), herd counting could be a multi-class problem as several species may be targeted in the same area. The original crowd counting CNN architectures must therefore be adapted accordingly.

2.3. Combining detection and counting tasks

Creating an architecture that can accurately locate individuals in a herd could be valuable. It could be used as a tool for obtaining pre-annotations from new data. However, as mentioned above, while traditional CNN-based object detectors can output object locations, they often fail to detect occluded objects. Density-based architectures could then be used, but at the cost of losing precise location information in dense herd regions. The ideal solution would be one that provides both a relatively accurate count of the herd (as usually given by density map approaches) and the position of the individuals in the herd (as given by detectors). Liang et al. (2021) recently proposed such a solution for crowd counting by using a novel Focal Inverse Distance Transform (FIDT) map which replaced the traditional density map. Their experiments demonstrated that this approach outperforms state-of-the-art localization-based methods and showed competitive counting performances while presenting a strong robustness to background and dense scenes samples. Such robustness is particularly interesting for counting herds in images with a heterogeneous background.

3. Material and methods

This section describes the datasets used, the proposed deep learning architecture (called “HerdNet”) as well as two standard baselines (anchor-based and density-based), and some details on the data processing utilized in this work. The baselines were used to compare the detection and counting capacity of HerdNet, which was optimized on a dataset that contains challenging herds (Ennedi).

3.1. Study area and dataset

The proposed Deep Learning architecture, HerdNet, was developed on a dataset acquired over the Ennedi Natural and Cultural Reserve (ENCR) located in north-eastern Chad during a 2019 aerial survey (Wacher, 2019). The ENCR covers nearly 50,000 km^2 of arid sandstone landscape surrounded by sandy plains. According to the map of Olson et al. (2001), the ENCR encompasses the following biomes: tropical and subtropical grasslands, savannas, and shrublands; deserts and xeric shrublands. It is managed by the African Parks Network (APN), in partnership with the Government of the Republic of Chad. The ENCR is a vital resource for local semi-nomadic groups who need grazing and water for their camels (*Camelus dromedarius*), goats (*Capra hircus*), sheep (*Ovis aries*), donkeys (*Equus asinus*), and rare cattle (*Bos taurus*). The APN’s long-term goal is to get all stakeholders, including local communities that depend on natural resources, to work together to conserve the Sahelo-Saharan heritage of the ENCR, including its archaeological value, while respecting traditions and allowing key species to thrive.

The data were acquired during aerial flights over the ENCR from December 20, 2019 to January 1, 2020. A Cessna 182 equipped with a laser altimeter and external metal strut rod markers calibrated at the observers’ eye level to indicate a 200 m band on each side of the aircraft at survey altitudes of 300 and 350 feet captured the images. Two Nikon D5000 SLR cameras, observer-operated by remote release cable and mounted by suction pads on left and right rear windows, were set up to match each observer’s view of the strut-mounted sample rods and the ground between them. A total of 19 flights were conducted, covering the core of the reserve (i.e., most of the Ennedi massif and the southwestern plains, representing around 23,000 km^2). Flights were conducted at 350 feet (~107 m) along transects spaced 4 km apart over the massif, and at 300 feet (~91 m) along transects spaced 10 km apart in the southwestern plains. Any groups of livestock (camels, donkeys, sheep, and goats) greater than 10 in number were photographed and the images were later used to provide ‘corrected’ counts. The date and time of image acquisitions were used to match the temporal and spatial data (altitude and GNSS coordinates) acquired by the altimeter during the flights. Thus, the images were associated with their respective transect and flight numbers.

‘Corrected’ count and observers’ group identification were used when establishing the ground truth. The annotations were made on Label Studio 1.3 (Tkachenko et al., 2021) by an expert, and consisted of 22,807 body-centered points in a subset of 914 images at 24 megapixels (6,000 × 4,000 pixels), containing major livestock species, i.e., camels, donkeys, and sheep and goats. Sheep and goats have been grouped in a single class (“sheep/goats”) since these two species are not distinguishable and often mixed within herds.

The dataset was split into training, validation and test sets following an allocation of 70, 10 and 20 %, respectively, while considering the

species' distribution, the flight and transect number. Dataset independence is thus ensured, whether the same herd is present in several images, and the species distribution is maintained, which is important in a severely unbalanced class distribution like ours. One transect from each flight was selected to construct the test set, resulting in a set of images containing a wide heterogeneity of landscapes from across the reserve. The images and species distribution for each set are given in Table 1.

3.2. Deep learning architectures

This sub-section provides details about the different deep learning architectures used in this study. These architectures include the following: an anchor-based baseline (Faster-RCNN), a density-based baseline (DLA-34), and the proposed architecture (HerdNet).

3.2.1. Anchor-based Baseline: Faster-RCNN

A naive way to count objects in an image would be to sum the number of detections provided by an object detector. A generic deep learning object detection framework locates and classifies objects within an image through the use of rectangular boxes encompassing the objects, called 'bounding boxes'. Traditional pipelines are anchor-based (Zhao et al., 2019), which means that they rely on anchors, a set of box proposals with different scales and aspects centered on potential object locations. These were first introduced in Faster-RCNN (Ren et al., 2015) and then used by a number of well-known object detectors like SSD (Liu et al., 2016), YOLOv2 (Redmon & Farhadi., 2017) or RetinaNet (Lin et al., 2017) because they improved their detection performance.

While anchor-based object detectors have given good detection performances for large mammals detection in aerial images (Eikelboom et al., 2019; Peng et al., 2020; Torney et al., 2019), Delplanque et al. (2022) recently observed a precision drop in herds and close-by animals resulting in overestimated counts.

In crowd counting, the use of anchor-based or even detection-based frameworks is not recommended because of the expensive labeling cost of bounding boxes and the difficulty of training detectors on heavily occluded objects (Li et al., 2021; Liu et al., 2018). Instead, most crowd counting approaches have relied on point annotations since the study of Lempitsky and Zisserman (2010). Nevertheless, anchor-based detectors are widely used in animal detection on aerial images, and thus remain relevant baselines. As it is one of the most-cited object detectors and the most common baseline, Faster-RCNN was chosen as the anchor-based baseline.

Faster-RCNN (Ren et al., 2015) is a two-stage object detector that:

Generates region proposals using a Region Proposal Network (RPN), which predicts objects' bounds and objectness scores at each position by utilizing anchors; and
Uses the refinement head of Fast R-CNN (Girshick, 2015) for regions of interest (RoIs) classification and bounding box offset regression.

A RPN is a deep fully convolutional network, and Fast R-CNN is composed of a RoI pooling layer and several fully-connected layers. Both share the same CNN features. For architecture comparison consistency, ResNet-34 (He et al., 2016) has been chosen for feature extraction because it has similar numbers of layers and the same convolutional blocks as the proposed architecture encoder (see Section 3.2.3). This choice will minimize any bias that might be caused by the use of a deeper feature extractor.

3.2.2. Density-based Baseline: Adapted DLA-34

Another way to count objects in an image is to estimate a density map whose integral would give the number of objects within that image. This 'density-based' approach was proposed by Lempitsky and Zisserman (2010) and was a real milestone for crowd counting, thanks to its simple framework for object counting and the introduction of point annotation. Since then, numerous Counting CNN (CCNN) architectures have been deployed and have shown excellent crowd counting performances on benchmark datasets (Gao et al., 2020; Li et al., 2021). Density-based CCNNs use CNN as a feature extractor and are trained to regressively learn a mapping between an image and the density map. Ground truth is produced using a density function, typically a normalized 2D Gaussian, convolved over each annotated point (Lempitsky & Zisserman, 2010). When properly trained, density-based CCNNs estimate the object count by integrating the density map they produce, and they provide spatial information about the objects.

Kellenberger et al. (2019a) and Padubidri et al. (2021) have recently shown that density maps can be used for animal counting in nadir aerial images. The former trained an adapted ResNet-18 architecture (He et al., 2016), while the latter trained a U-Net semantic segmentation CNN architecture (Ronneberger et al., 2015) to produce density maps. Unfortunately, precise object location is difficult to obtain from density maps, especially for close-by and occluded objects where 2D Gaussians strongly overlap.

While density-based architectures tend to provide precise object counts in high-density scenes, precise localization is lost. Although the primary goal of aerial surveys is to establish accurate population count, obtaining the precise position of animals in images could be valuable for creating annotations from new data for further model training.

A density-based baseline was therefore established to assess the counting performance of the proposed approach. For comparison consistency, the same feature extractor and decoder as the proposed architecture (i.e., adapted DLA-34 Yu et al. (2018)) was selected. In fact, the architecture is that of HerdNet (Fig. 2), except that the classification head has been removed and the main head generates three density maps (one for each species) instead of one localization map. During the test time, for each species, only the pixels with the maximum value among the three predicted density maps were retained. This process prevents the same individuals from being counted as several species. An adaptive threshold of 0.07 was then applied to the density values to eliminate background noise.

3.2.3. Proposed Architecture: HerdNet

Since the objective was to develop an architecture to accurately locate and count dense herds, the proposed deep learning architecture, HerdNet, is inspired by both point-based object detectors and crowd counting architectures.

Table 1
Details of the Ennedi dataset split. The data was split into training (~70 % of all images), validation (~10 %) and test (~20 %) sets while accounting for data heterogeneity (i.e., species distribution, flight and transect) to maintain independence. The numbers in brackets indicate the relative percentage of data in each set. The last row gives the number of patches containing animals extracted from the 24-megapixel images.

Number of	Training	Validation	Test	Total
Camel	2,608 (69.7 %)	380 (10.2 %)	753 (20.1 %)	3,741
Donkey	861 (70.2 %)	127 (10.3 %)	239 (19.5 %)	1,227
Sheep/Goat	12,486 (70.0 %)	1,774 (9.9 %)	3,579 (20.1 %)	17,839
24 MP images	619 (67.7 %)	122 (13.4 %)	173 (18.9 %)	914
512 × 512 pixel patches	5,826 (75.3 %)	1,039 (13.4 %)	869 (11.3 %)	7,734

MP, megapixel.

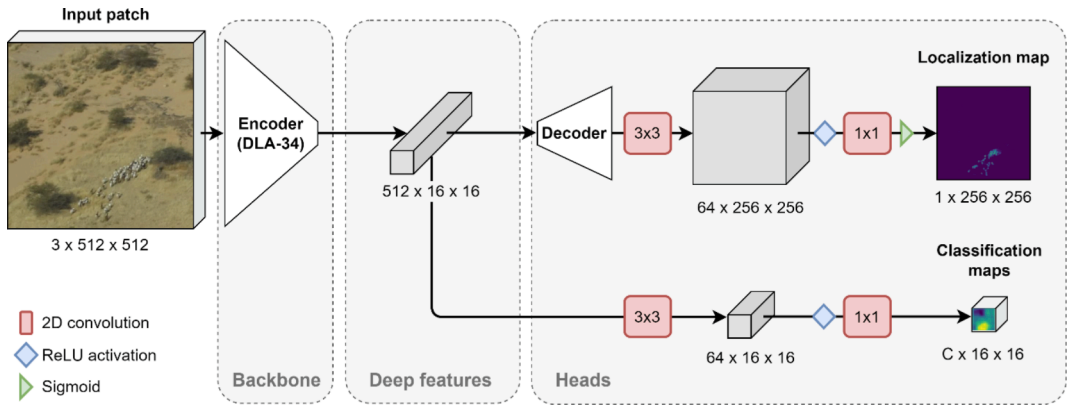

Fig. 2. HerdNet architecture details.

The core of HerdNet is derived from CenterNet, except that only the branch which estimates the objects' center has been retained. This branch corresponds to the localization head. The adapted DLA-34 (Yu et al., 2018) was used as encoder-decoder (Fig. 2) because it gave the best speed vs. accuracy trade-off on the MS COCO (Lin et al., 2014) dataset (Zhou et al., 2019), which is convenient for our application case. As in Zhou et al. (2019), a 3×3 convolutional layer was added on top to obtain specialized features maps for each head, and early experiments showed that 64 channels were adequate to obtain good results. A 1×1 convolution, preceded by a ReLU activation and followed by a sigmoid activation produces the desired localization map. Early experiments showed that a reduction factor of 2 between input and output sizes gives similar results with fewer network parameters than those obtained by keeping the original patch resolution.

For classification, a second low-resolution head was added onto the deep features layer, with one 3×3 convolutional layer with 64 channels on top, as for the location head. Eventually, a 1×1 convolution, preceded by a ReLU activation, produces the C classification maps, C being the number of classes including background (Fig. 2). Ablation studies showed that 16×16 pixel classification maps were sufficient for species identification, and that including the background class in the training objective helped to better learn the landscape heterogeneity (see Appendix S1).

During the testing time, the Local Maxima Detection Strategy (LMDS) proposed by Liang et al. (2021) was used to extract points from the predicted localization map. The LMDS utilizes a 3×3 max-pooling operation to obtain candidate points, which are then filtered using an adaptive threshold, set here at 0.3 times the maximum candidate value. An input patch is considered a negative sample when the maximum candidate value is below 0.1, as in the original paper. Next, the classification maps were used to classify the selected points. The softmax function was used on all classes to obtain classification scores. Then, the most confident class was selected among the foreground classes. With this procedure, a selected point could never be classified as background. Finally, the selected points were used to pin the foreground classes to equivalent locations and the class labels and scores were retrieved.

3.3. Data processing

This sub-section describes all the processes implemented for developing the models. Operations were performed on a Windows-10 workstation using Python 3.8.10. The workstation contained a 64 GB AMD Ryzen 9 5900X central processing unit (CPU) and an 8 GB NVIDIA GeForce RTX 3070 graphics processing unit (GPU). All architectures were implemented in PyTorch 1.11 (Paszke et al., 2019) and experiments were tracked with Weights & Biases 0.10.33 (Biewald, 2020).

3.3.1. Patch generation and stitching

Original 24-megapixel images were cut into patches of 512×512 pixels to maintain initial resolution and because experimenting with original-size images exceeds the memory capacity of current GPUs. To ensure that every animal appears in its entirety during training, a patch overlap was used. After manually measuring the largest individuals in the dataset (i.e., camels close to the lower stream bar), it was concluded that a 160-pixel overlap was a good value, as the widest of these was 156 pixels long.

During the testing, the original-size images were scanned with a sliding window to harvest predictions and then stitch them together. To do so, each patch of 512×512 pixels was evaluated independently and overlapped region predictions were filtered out. Specifically, the common Non-Maximum Suppression (NMS) method was adopted with an Intersection-over-Union (IoU) threshold (Everingham et al., 2015) of 0.5 (as Delplanque et al., 2022; Peng et al., 2020) and a score threshold of 0.4 for Faster-RCNN predictions (Appendix S2). For the adapted DLA-34, predicted density maps were filtered and stitched using Hann windows to reduce the edge-effect, as proposed by Pielawski and Wählby (2020). Next, an adaptive threshold of 0.07 was applied on the stitched image pixel values to eliminate background noise (Appendix S3). Finally, overlapped predicted grid values were averaged before using LMDS for HerdNet.

3.3.2. Model training

3.3.2.1. Hard negative patch mining. Hard negative mining is a training technique used to treat the hard negative samples severely during training (Kellenberger et al., 2018; W. Liu et al., 2016; Shrivastava et al., 2016). In the animal detection domain, hard negative samples correspond to background elements detected as animals with a high confidence score. A Hard Negative Patch (HNP) mining method was adopted here, following the hard negative mining concept, to further reduce the number of false positives produced by the model. After a first training session where the architecture was trained exclusively on animal patches, the model was run on the 24-megapixel training images. HNPs were then mined from the stitched predictions, which are the patches that contain hard negative instances. These HNPs were eventually used to retrain the model a second time to force it to develop more robust features regarding the most confusing background elements. With this method, only the most complex background patches were selected, which makes the task more efficient and less tedious than training on all the patches, as proposed by Kellenberger et al. (2018).

3.3.2.2. Faster-RCNN. For the anchor-based baseline (i.e. Faster-RCNN), bounding boxes were generated from annotated points. For this purpose, a subset of the Ennedi dataset was annotated as bounding boxes. Then, for each species, median height and width were computed

per a 200-pixel horizontal strip in the image, and the maximums of each were selected to create square bounding boxes centered on each annotated point.

Training data was augmented artificially using Albumentations' (Buslaev et al., 2020) random horizontal flip and motion blur data augmentations.

During the first training step, the parameters of the features extractor were initialized using ImageNet (Russakovsky et al., 2015) pretrained parameters. The architecture was then trained and validated on animal-only patches for 100 epochs with a batch size of 4 and a weight decay of 0.005 using the Adam optimizer (Kingma & Ba, 2017). Concerning the learning rate, PyTorch's 'ReduceLROnPlateau' learning rate scheduler was used because it made it possible to automatically decrease the learning rate at the most appropriate time during training. The initial learning rate was set to 10^{-5} after a linear warmup of 100 iterations and could then decrease by a factor of 0.1 until 10^{-6} when no improvement was observed on the validation set over a period of 10 epochs. After a reduction, a delay of 10 epochs was imposed to let the architecture adapt to the new learning rate.

At the end of this first training step, the network's parameters that yielded the best performances on the validation set were selected for initializing the second training step. During the latter, we added the HNPs to the training set and validated on 24-megapixel validation images using the same hyperparameters as the first step, except for the number of epochs and the initial learning rate, which were set at 50 and 10^{-6} respectively. 24-megapixel images were used for validation to focus on both localization and counting within real case scenes during this second training step.

Due to the substantial imbalance in species instances, class weighting was used in the bounding boxes' classification loss. Satisfactory results were found by setting the class weights' values to 0.1 for background class, and to the unit rounded value of the ratio of the majority class instances to that of the actual class. All other hyperparameters were left at their default values specified in PyTorch. The parameters of the network that yielded the best performances on the full images of the validation set were then selected for testing.

3.3.2.3. Adapted DLA-34. The density-based baseline (i.e., adapted DLA-34) ground truth density maps were generated using a 2D Gaussian function, convolved over each annotated point, for each species class, as in Lempitsky and Zisserman (2010):

$$M_{density,c}(i,j) = \sum_{p \in P} \mathcal{N}\left(i,j;p,\sigma^2\right) \tag{1}$$

where $M_{density,c}(i,j)$ is the density map of a class c, p denotes an equivalent low-resolution annotated point $(x^{'}/2, y^{'}/2)$ within the low-resolution image 2D points set P, and $\mathcal{N}(i,j;p,\sigma^2)$ represents a normalized 2D Gaussian kernel evaluated at pixel (i,j), with the mean centered on p, and an isotropic covariance matrix with spread parameter σ, set at 5 pixels. With this definition, integrating each density map produced gives the total count of each species class N_c:

$$N_c = \sum_{(i,j)} M_{density,c}(i,j) \tag{2}$$

The architecture was then trained using the Structural Similarity Index (SSIM) (Wang et al., 2004) loss between the predicted density maps and the ground truth density maps:

$$\mathcal{L}_{density}(\widehat{Y}, Y) = \frac{1}{C}\sum_{c} w_c(1 - SSIM_c) \tag{3}$$

with:

$$SSIM_c(\widehat{y}_c, y_c) = \frac{\left(2\mu_{\widehat{y}_c}\mu_{y_c} + \lambda_1\right)\left(2\sigma_{\widehat{y}_c y_c} + \lambda_2\right)}{\left(\mu^2_{\widehat{y}_c} + \mu^2_{y_c} + \lambda_1\right)\left(\sigma^2_{\widehat{y}_c} + \sigma^2_{y_c} + \lambda_2\right)} \tag{4}$$

where Y and $\widehat{Y}$ are the ground truth and the predicted density maps, respectively, with y_c and $\widehat{y}_c$ their respective class-specific values, C is the number of species classes, w_c is the class weight, μ and σ are the local mean and variance values, respectively, and λ_1 and λ_2 are set to 10^{-4} and 9×10^{-4}, respectively.

As for Faster-RCNN, the architecture was trained and validated using the same data augmentations, parameter initialization, hyper-parameters, and optimizer. A fixed learning rate of 10^{-5} was used here as a learning rate scheduler gave poorer performances. The HNP mining procedure was discarded here because using it showed an increase in the counting errors.

Class weighting was also applied on the SSIM loss using the same class weights. All other hyperparameters were left at their default values, specified in PyTorch.

3.3.2.4. HerdNet. Low-resolution FIDT maps (Liang et al., 2021) were adopted as ground truth for training the HerdNet's localization branch:

$$M_{loc}(i,j) = \frac{1}{D(i,j)^{(\alpha \times D(i,j)+\beta)} + k} \tag{5}$$

where $M_{loc}(i,j)$ is the FIDT map, $D(i,j)$ represents the euclidean distance between the pixel (i,j) and its nearest equivalent low-resolution animal location $(x^{'}/2, y^{'}/2)$, α and β are FIDT hyper-parameters, set as 0.02 and 0.75 respectively, following Liang et al. (2021), and k is a constant, set to 1 to avoid division by zero. FIDT maps produce local maxima of 1 at each animal's center, with a slow response decay and a background response close to 0.

This branch was trained using the unnormalized penalty-reduced pixel-wise logistic regression with focal loss (Lin et al., 2017), as proposed by Zhou et al. (2019):

$$\mathcal{L}_{loc}(\widehat{Y}_l, Y_l) = -\sum_i \sum_j \begin{cases} \left(1 - \widehat{y}_{l,ij}\right)^{\alpha} \log\left(\widehat{y}_{l,ij}\right), \text{if} y_{l,ij} = 1 \\ \left(1 - y_{l,ij}\right)^{\beta} \left(\widehat{y}_{l,ij}\right)^{\alpha} \log\left(1 - \widehat{y}_{l,ij}\right), \text{otherwise} \end{cases} \tag{6}$$

where Y_l and $\widehat{Y}_l$ are the ground truth and the predicted localization grids, respectively, and $y_{l,ij}$ and $\widehat{y}_{l,ij}$ their values at a specific pixel location (i,j), and α and β are focal loss hyper-parameters, set at 2 and 4, respectively, as indicated in Zhou et al. (2019).

For the classification branch, low-resolution classification maps were produced from equivalent low-resolution animal locations $(x^{'}/32, y^{'}/32)$. Practically, at each equivalent animal location, a 1-pixel border was added and the whole region was defined as the species identifier. This was to ensure a sufficient point coverage area given the low resolution of the classification branch output (16×16 pixel). The common cross-entropy loss was used for training this branch:

$$\mathcal{L}_{class}(\widehat{Y}_c, Y_c) = -\sum_i \sum_j \sum_c w_c y_{ijc} \log\left(\widehat{y}_{ijc}\right) \tag{7}$$

where Y_c and $\widehat{Y}_c$ are the one-hot encoded ground truth and the predicted classification grids, respectively, y_{ijc} and $\widehat{y}_{ijc}$ their values at a specific pixel location (i,j) for a particular class c, and w_c is the class weight.

The overall training objective is then:

$$\mathcal{L} = \mathcal{L}_{loc}(\widehat{Y}_l, Y_l) + \mathcal{L}_{class}(\widehat{Y}_c, Y_c) \tag{8}$$

During the first and second training steps, HerdNet followed the same training procedure and hyperparameters as Faster-RCNN, except that the initial learning was set to 10^{-4}. Again, the best network's parameters were kept based on the performances obtained on the full images of the validation set.

3.3.3. Model evaluation

The trained architectures (or models) were evaluated using both localization and counting metrics. A prediction was defined as a true positive (*TP*) if there was a match with a ground truth and if the animal

identification was correct. In the case where several predictions met the two rules, the best one was selected and the others were considered as false positives (*FP*). Finally, if no matches were found or if the identification was incorrect, the ground truth was considered as false negative (*FN*). To define a match, we used the IoU for Faster-RCNN and set a minimum threshold of 0.3, where the best prediction is the one with the highest IoU. In the HerdNet approach, we used the Euclidean distance between points, with a maximum threshold of 5 pixels, where the best prediction is the one with the minimum Euclidean distance.

Recall, precision and F1 score were then computed for each class (i.e. for each species), as well as for the binary case (animal vs. background):

$$recall = \frac{\sum TP}{\sum TP + \sum FN} \tag{9}$$

$$precision = \frac{\sum TP}{\sum TP + \sum FP} \tag{10}$$

$$F1score = \frac{2 \times recall \times precision}{recall + precision} \tag{11}$$

As it represents the harmonic mean of recall and precision, the F1 score is a good metric with which to assess the compromise between the number of *FPs* and *FNs*. Therefore, the binary F1 score was used as the performance metric during validation. In addition to these metrics, we also compute, for each species, the foreground interclass confusion, which is equal to 0 when all the predictions are correctly classified:

$$confusion(c) = 1 - \frac{n_c}{\sum_{i=1}^{C} n_c} \tag{12}$$

where n_c is the number of predictions identified as class c, and C is the number of foreground classes (i.e., the number of species).

Note that localization metrics could not be applied to the adapted DLA-34 due to the loss of localization information caused by the overlap of the 2D Gaussians in dense herd areas.

The Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) are used as counting metrics, again computed for each class and for the binary case:

$$MAE = \frac{1}{I}\sum_{i=1}^{I} |\hat{n}_i - n_i| \tag{13}$$

$$RMSE = \sqrt{\frac{1}{I}\sum_{i=1}^{I} (\hat{n}_i - n_i)^2} \tag{14}$$

where I is the number of images, and $\hat{n}_i$ and n_i are the predicted and ground truth count of the i-th image, respectively.

Finally, an individual proximity metric was derived by calculating the Minimum Spanning Tree (MST) (Gower & Ross, 1969) on N annotated points in 512 × 512 pixel patches of the test set (Fig. 3). The MST computes a set of $N-1$ straight line segments joining pairs of points with no loops, forming a tree of minimum length. MiSTree Python's package version 1.2.0 was used (Naidoo, 2019). In this package, the MST was initially constructed using a k-nearest neighbor graph, here set at $N-1$, which was then fed to Kruskal's algorithm (Kruskal, 1956). To obtain a metric representative of the proximity of individuals in the patch, we used the median of segment length values instead of the sum. This value was then divided by the threshold value defined above (i.e. 5 pixels) to normalize the metric. Thus, a value close to 1 means a very dense herd where individuals are tightly grouped. Based on this, three proximity classes were defined:

1) High density: patches where the proximity metric varied between 0 and 3;
2) Medium density: patches where the proximity metric varied between 4 and 20; and
3) Low density: patches where the proximity metric was above 20.

4. Results

4.1. Hard negative patch mining

The addition of HNPs to the training set increased the precision of Faster-RCNN and HerdNet by more than 18 % and 30 % respectively, despite a decrease in recall of about 8 % and 7 % respectively (Table 2). This resulted in a better counting performance, with a lower average confusion between species. Consequently, for each of these models, the version trained with HNPs was retained for further analysis on the test set. In contrast, the counting performance of the adapted DLA-34 decreased with the use of this technique (Table 2). Therefore, the version of the adapted DLA-34 using HNP mining was discarded and only the version without this technique was used for the analyses on the test set. This is to compare the best version of each model.

4.2. Model comparison

Overall, HerdNet outperformed the detection and counting performance of the two baselines, Faster-RCNN and the adapted DLA-34, in addition to having a faster processing time (Table 3). However, Faster-RCNN had a lower average confusion level, and the adapted DLA-34 had a lower absolute total counting error.

HerdNet showed a counting performance that was close to true counts while Faster-RCNN tends to overestimate the true number of

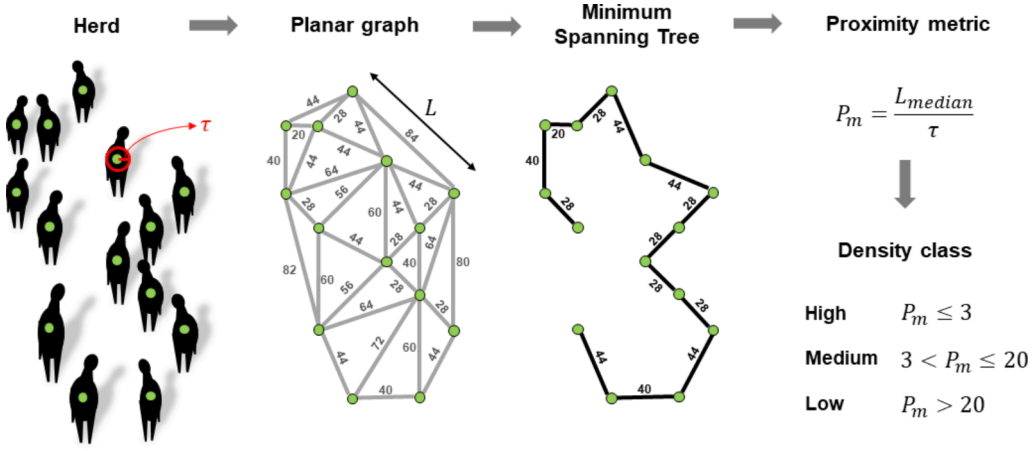

Fig. 3. Conceptual representation of the Minimum Spanning Tree and proximity metric calculation on a schematic herd. τ represents a circular distance threshold (defined here at 5 pixels) and L represents the Euclidean distance between two individuals.

Table 2
Binary (animal vs. background) performances of the three approaches on 24-megapixel images of the validation set, using Hard Negative Patch mining procedure or not. Values in bold indicate the best performance between the two modalities.

Approach	Anchor-based		Density-based		Point-based	
Architecture	Faster-RCNN		DLA-34		HerdNet	
HNP[1]	No	Yes	No	Yes	No	Yes
Recall	**64.1 %**	56.0 %	n/a	n/a	**72.1 %**	64.4 %
Precision	20.4%	**38.5 %**	n/a	n/a	43.5 %	**75.4 %**
F1 score	30.9%	**45.7 %**	n/a	n/a	54.3 %	**69.4 %**
MAE[2]	40.1	**11.3**	**12.3**	12.3	14.3	**6.1**
RMSE[3]	51.7	**16.5**	**19.1**	23.0	19.4	**10.5**
Average confusion	15.0 %	**13.7 %**	n/a	n/a	22.4 %	**17.8 %**
Total counting error	214.4 %	**45.4 %**	**−0.2 %**	−40.8 %	65.5 %	**−14.6 %**

[1] HNP, Hard Negative Patch; [2]MAE, Mean Average Error; [3]RMSE, Root Mean Square Error.

animals (Fig. 4). As for the adapted DLA-34, it tends to underestimate large groups, and overestimate very small groups.

Regarding species identification, HerdNet outperformed Faster-RCNN for all three target species (Table 4). For camels and donkeys, however, Faster-RCNN showed less confusion between species. The adapted DLA-34 showed lower MAE and RMSE values for donkeys than HerdNet, but higher values for camels and sheep/goats.

Taking into consideration both overall and per-species results, HerdNet is the architecture with the best detection and counting performances, especially for sheep/goats, which represent especially challenging herds (Fig. 5).

4.3. Robustness of HerdNet towards animals proximity

Recall and precision were computed for each class of animal proximity defined in section 3.3.3 to assess the robustness of HerdNet towards animal proximity in 512 × 512 pixel patches of test set images. The results indicate that the mean precision of HerdNet was systematically higher than that of Faster-RCNN for each proximity class, while keeping equivalent mean recall values (Fig. 6). This reveals the ability of HerdNet to generate few false positives in both dense (Fig. 7) and sparse herd patterns.

5. Discussion

5.1. Best approach for counting dense herds

Three different approaches were compared to precisely detect and/or count animals in herds within oblique aerial imagery: 1) a CNN-anchor-based object detector (Faster-RCNN); 2) a CNN-density-based detector (an adapted version of DLA-34); and 3) a CNN-point-based object detector (called HerdNet). The first two approaches served as baselines because they have already proved their worth in the field of wildlife detection/counting within aerial imagery.

As previously observed (Delplanque et al., 2022; Peng et al., 2020), the anchor-based architecture showed its limitations in precisely detecting close-by individuals. It produced here a high number of false positives in dense herds, even using score thresholding, resulting in systematic over-counting. This raises questions about the use of such models in entire aerial surveys where it is expected to get images with both remote individuals and dense herds.

The CNN-density-based detector (DLA-34) provided better total counting performance but struggled to correctly count large and dense herds. The counting errors are higher than those obtained by Kellenberger et al. (2019a) and Padubidri et al. (2021) on their nadir datasets. In fact, the DLA-34′s counting errors are low for minority species (i.e., camels and donkeys), but much higher for sheep/goats, which are far more gregarious. This could be explained by the change in scale within the image due to the oblique viewing angle, the higher variance in the number of individuals and the greater heterogeneity of the background. These factors can also limit the performance of crowd counting using density maps (Gao et al., 2020). A solution would be to design a multi-scale architecture such as MCNN, a multi-column architecture that uses different kernel sizes to capture images at different scales (Zhang et al., 2016).

Our CNN-point-based object detector (HerdNet) gave the best detection and counting performances while also being the fastest approach, suggesting that it seems best suited for locating and counting animals in dense herds. Estimating the number of livestock in protected areas is sometimes a politically sensitive issue, as a livestock invasion is detrimental to the biomass of wildlife (Scholte et al., 2022b). Moreover, livestock invasions directly show that the responsible authorities or supporting international non-governmental organizations have failed in their conservation mission. A method that overestimates this figure is therefore undesirable. Hence, HerdNet is the most appropriate and preferred approach for herd counting.

Table 3
Binary (animal vs. background) performances of the three approaches on 24-megapixel images of the test set. Values in bold indicate the best performance among the architectures.

Approach	Anchor-based	Density-based	Point-based
Architecture	Faster-RCNN	DLA-34	HerdNet
Recall	59.5 %	n/a	**70.2 %**
Precision	39.4 %	n/a	**77.5 %**
F1 score	47.4 %	n/a	**73.6 %**
MAE[1]	15.2	15.9	**6.1**
RMSE[2]	26.2	30.4	**9.8**
Average confusion	**11.1 %**	n/a	15.8 %
Total counting error	51.2 %	**7.6 %**	−9.4 %
Processing time (seconds)	5.0	5.5	**3.6**

[1] MAE, Mean Average Error; [2]RMSE, Root Mean Square Error.

5.2. Species identification limits

In terms of species identification, HerdNet was slightly better than Faster-RCNN for sheep/goats but was about 7–8 % worse for minority species, i.e., camels and donkeys. Thus, the class imbalance seems to impact HerdNet more than Faster-RCNN. After manually analyzing the images with the most significant cases of confusion, overall trends were deduced. First, the size and often the low resolution of the individuals were source of confusion for the model, especially for donkeys (Fig. 8). The latter were usually well identified in the higher resolution areas (i.e., near the lower stream bar), but that the identification degraded with the distance to the aircraft. Regarding camels, the lighter ones located in the low-resolution regions of the image (i.e., near the upper stream bar) were often confused with sheep/goats. Furthemore, identification was sometimes incorrect when the animal was positioned to the side (Fig. 8).

This may be explained in part by the fact that the image resolution and high flight height in the massif sometimes did not allow the species to be accurately distinguished during annotation, especially those far from the aircraft. In such cases, identification was solely based on the

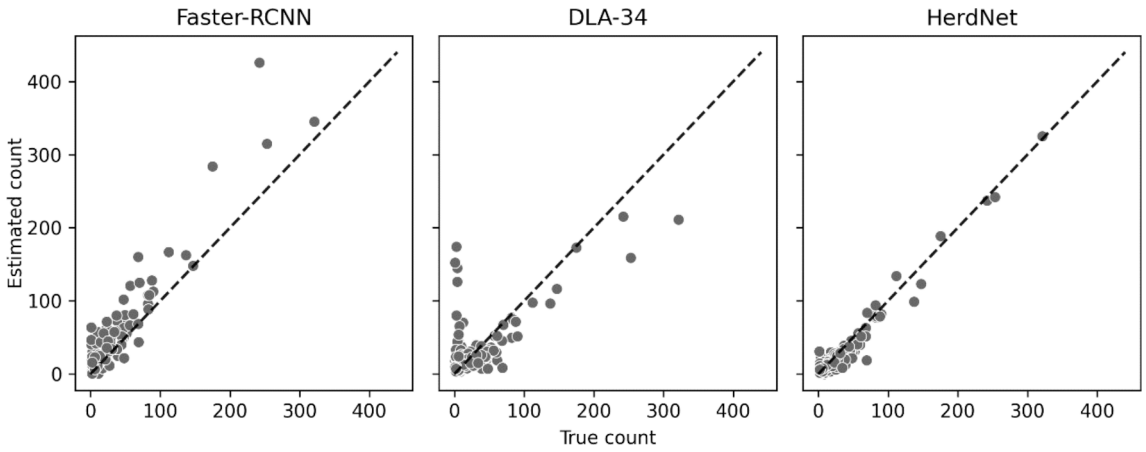

Fig. 4. Estimated counts produced by each architecture versus the true counts in 24-megapixel images of the test set.

Table 4
Performances of the three approaches on 24-megapixel images of the test set according to target species. Values in bold indicate the best performance among the architectures.

Approach	Anchor-based			Density-based			Point-based		
Architecture	Faster-RCNN			DLA-34			HerdNet		
Species	Camel	Donkey	Sheep/Goat	Camel	Donkey	Sheep/Goat	Camel	Donkey	Sheep/Goat
n	753	239	3,579	753	239	3,579	753	239	3,579
Recall	57.5 %	18.8 %	60.6 %	n/a	n/a	n/a	**61.8 %**	**37.7 %**	**70.9 %**
Precision	45.8 %	9.3 %	39.6 %	n/a	n/a	n/a	**75.1 %**	**59.6 %**	**75.3 %**
F1 score	51.0 %	12.4 %	47.9 %	n/a	n/a	n/a	**67.8 %**	**46.2 %**	**73.0 %**
MAE[1]	3.3	3.0	13.9	3.6	**1.6**	15.2	**2.6**	2.5	**7.0**
RMSE[2]	5.7	4.3	25.6	6.5	**3.3**	27.8	**4.8**	4.6	**10.7**
Confusion	**0.0 %**	**30.8 %**	2.4 %	n/a	n/a	n/a	7.4 %	39.2 %	**0.8 %**

[1] MAE, Mean Average Error; [2]RMSE, Root Mean Square Error.

observers' survey records. However, as the livestock in this study are typically found in single-species groups, identification of individuals is a bonus and not strictly necessary. Indeed, having a model capable of precisely locating and counting individuals is already a very real help in processing images containing large and dense herds. Identification could be further enhanced by a quick review by the human eye of the surrounding individuals.

5.3. Potential use of HerdNet

The use of HerdNet on other datasets requires prior training on similar data, i.e., with the same viewing angle, the same mammal species, and similar spectral and spatial resolutions. To assess the potential use of HerdNet architecture, it was trained and evaluated on the wildlife nadir aerial images of Delplanque et al. (2022). Results showed that HerdNet produced far fewer false positives than the state-of-the-art model, which was Libra-RCNN (Pang et al., 2019), while maintaining a high recall value, hence showing better counting performances (see Appendix S4 for details). This suggests that this architecture is not limited to its use on oblique imagery only, but has good potential for various types of aerial image, acquired under different acquisition conditions.

However, HerdNet may need to be modified in the case of dense mixed herds. Despite the results of the sensitivity study (see Appendix S1), the low resolution of the classification head could indeed be problematic if different species are within 32 pixels of each other in the input patch. This distance corresponds to one pixel in the 16 × 16-pixel classification maps. This case did not occur in the dataset of this study, as the dense herds were systematically homogeneous in species. Nevertheless, we believe that this should not be an issue for training and using HerdNet on oblique imagery taken with good quality reflex cameras (e.g., 24 MP of resolution), at common camera tilt values (30-45° off nadir) and survey flight heights (300–350 ft). In such case, the ground sampling distance usually does not exceed 3–4 cm/pixel and 6–8 cm/pixel near the lower and upper stream bars, respectively. This means that the low resolution of the classification head could become a concern when two different species would be less than 2–3 m apart in reality, which is rather rare for large wild terrestrial mammals. However, if such a case arises, the architecture can simply be adapted by adding a decoder at the beginning of the classification head, whose depth will depend on the desired output resolution.

Finally, we observed that applying the model on a too-different dataset without re-training led to poor performance, probably due to a data domain gap. This could be solved by using *transductive transfer learning* techniques, which allow the transfer of knowledge learned from a source domain to a different target domain, considering the same learning task (Pan & Yang, 2010). Kellenberger et al. (2019a) have already proposed such a solution for wildlife detection using a Transfer Sampling criterion, allowing their model to be reused for repeated nadir drone image acquisitions. However, the transfer learning from oblique to nadir animal detection does not seem to have been studied and should be explored in future research. At this stage, we can only suggest that future users of HerdNet re-train the architecture on a data domain close to their own to obtain more satisfactory results.

5.4. Model precision practical implications

Precise counts of large mammals within sampling strips are important to obtain minimum-biased population estimates. Since undercounting is one of the major biases of aerial surveys (Caughley, 1974; Grimsdell & Westley, 1981; Jachmann, 2002), the main expectation of automatic approaches is to obtain a model with a high detection rate (i.

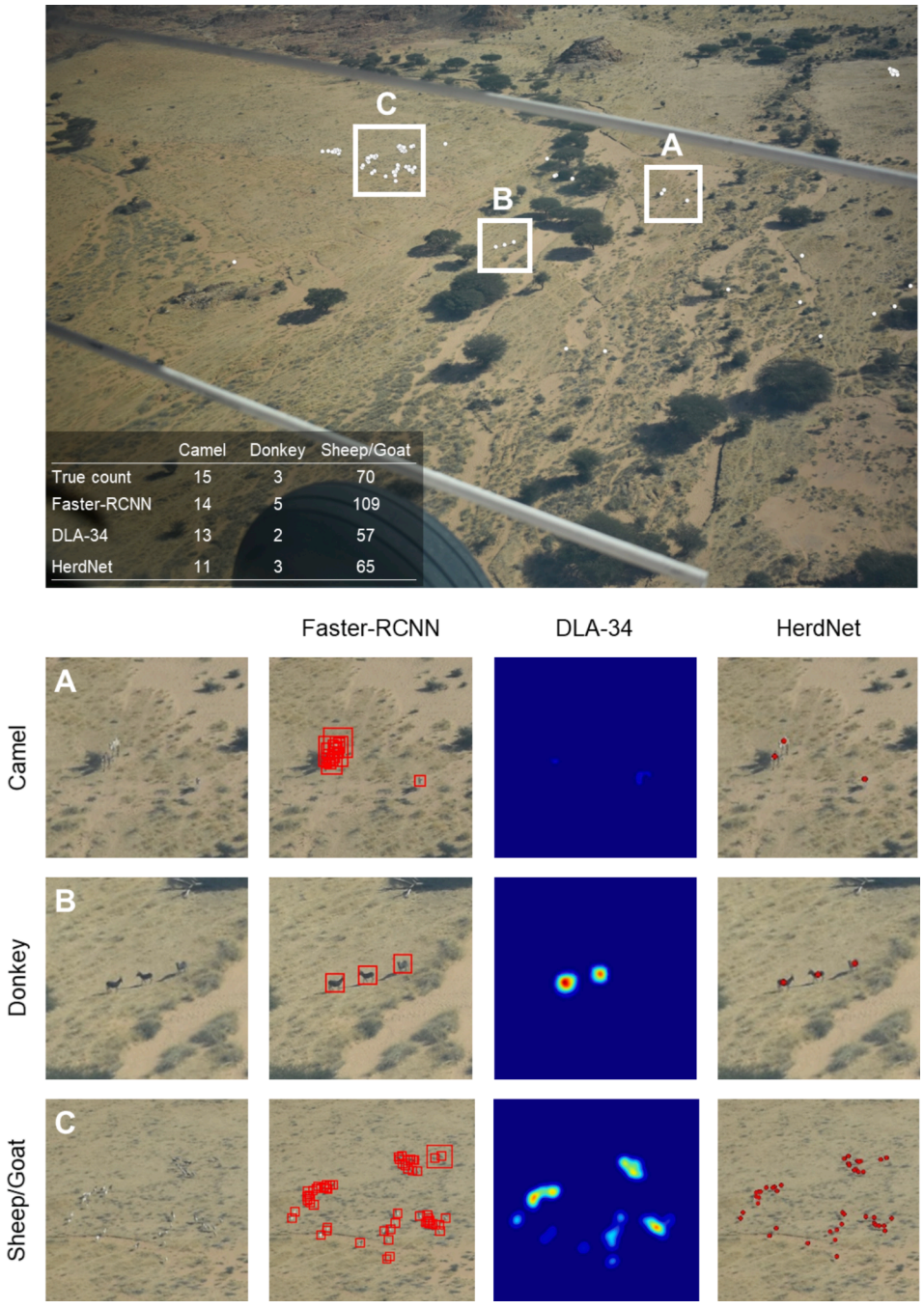

Fig. 5. Predictions of the three trained architectures on a 24-megapixel image containing the three target species (camel, donkey, and sheep/goat). White points correspond to the annotations, red bounding boxes to Faster-RCNN predictions, density maps to the predictions of adapted DLA-34, and red points indicate the HerdNet predictions. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

e. high recall) and few false positives (i.e. high precision). However, recall and precision are often antagonistic: improving the precision of a model usually reduces its recall and vice-versa. When developing tools to assist protected area managers, the recall/precision trade-off depends on the goal. As a semi-automatic model for background image rejection, recall should be preferred, as the detections will be reviewed by humans afterwards. However, protected area managers do not always have dedicated office staff for such specific tasks. For a fully automatic system, the optimal trade-off should be preferred, and a prior estimation of the possible bias is necessary to correct the counts. In this study, we optimized the model on the F1 score to automate the counting of herds and minimize the error in individual images. In view of the results obtained, the model proved to be a good tool for the automatic counting of individuals in individual oblique images from arid environments containing livestock herds.

5.5. *Future work*

Three aspects for future research can be identified. First, the species identification capacity of HerdNet, which could be augmented by confronting it with data sets composed of a large number of species and with some that would be very similar but identifiable by humans from the aircraft (e.g. antelopes). This process would assess the limits of HerdNet regarding the human species' identification ability. Future challenges would involve the adaptation of the model for wildlife species living in herds (elephants, buffaloes, wildebeest, giraffes, etc.), and among those of small size living in small groups (kobs, warthogs, etc.). The use of this approach on complete chains of transect images could then be investigated by automating the management of overlapped images with the aim of obtaining population estimates. Encouraging results will bring us closer to the full automation of aerial surveys. Finally, the

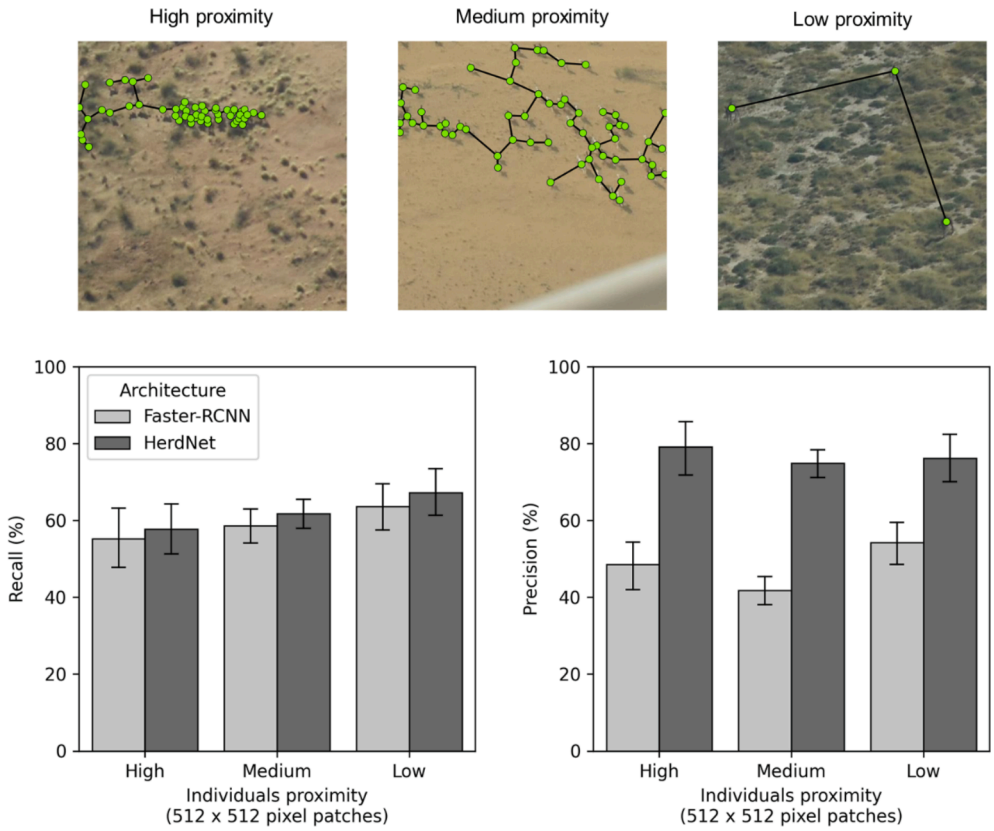

Fig. 6. Recall and precision mean values of Faster-RCNN and HerdNet, computed on 512 × 512 pixel patches for each class of animal proximity metric based on a minimum spanning tree. The error bars correspond to the 95 % confidence interval.

Fig. 7. Examples of HerdNet predictions for challenging dense sheep/goat herds. The first row contains sample patches selected from 24-megapixel images, while the second row shows the respective predicted points in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

generalizability of HerdNet should be further developed by studying its response to background, viewing angle, and species variability, and possible generalization solutions (e.g. using domain adaptation techniques). A general model or an adaptation approach, that should be simple and require limited technical and human resources, would allow its practical use by protected area managers. That sort of approach would enable them to easily adapt the model for use in the savannah, for example, during both the dry and rainy seasons.

6. Conclusion

In large protected areas in Africa, large mammals are usually surveyed by human observers using aircraft. Unfortunately, the difficulty of observers to precisely count large groups has led to the use of aerial

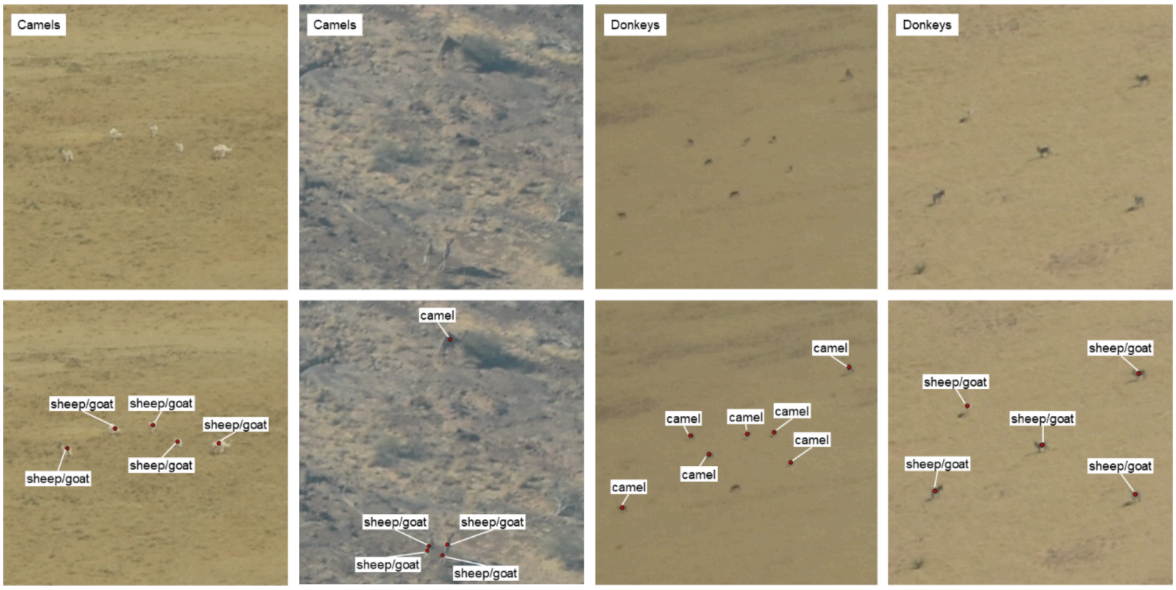

Fig. 8. Examples of HerdNet identification confusion for camels and donkeys. The first row contains sample patches with homogeneous species groups, selected from 24-megapixel images, while the second row shows the respective predicted points in red and the associated species name. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

imagery. In such images, the manual counting of individuals is time consuming and the latest Deep Learning approaches have shown their limitations in detecting dense herds. Inspired by crowd counting, the point-based Deep Learning architecture proposed in this study, HerdNet, addresses this problem by precisely detecting and counting animals regardless of individual proximity. Outperforming both anchor-based and density-based baselines, the proposed model has proven to be the fastest and the most suitable approach for detecting and counting closed-by large mammals. It could therefore be used as an automatic livestock counting tool on oblique aerial images acquired in arid areas, and it could be extended to other areas and wildlife species after prior retraining.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are very grateful to the European Union and the Dutch Postcode Lottery for providing the financial means to conduct the 2019 aerial survey of the Ennedi Massif and south-western plains. We thank the African Parks Network (APN), especially the ENCR management team for the planning, preparation, and implementation of the survey, as well as the aircrew, composed of Tim Wacher, Alexis Peltier, Moussa Sougui and Issa Hamid. We also thank the Zoological Society of London for analyzing the data and producing the final report of the survey. Finally, we thank both East and West Ennedi Provinces for authorizing low-altitude flights for the purpose of this study.

Funding information

This work was supported by the Fund for Research Training in Industry and Agriculture (FRIA, F.R.S.-FNRS). The 2019 aerial survey of the Ennedi Massif and south-western plains was financially supported by the European Union and the Dutch Postcode Lottery.

Code accessibility

All code for training and testing HerdNet, as well as pre-trained models, is available at https://github.com/Alexandre-Delplanque/HerdNet. Any updates will be published on this GitHub repository.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.isprsjprs.2023.01.025.

References

Ayantunde, A.A., Duncan, A.J., van Wijk, M.T., Thorne, P., 2018. Review: Role of herbivores in sustainable agriculture in Sub-Saharan Africa. Animal 12, s199–s209. https://doi.org/10.1017/S175173111800174X.

Bengis, R.G., Leighton, F.A., Fischer, J.R., Artois, M., Morner, T., Tate, C.M., 2004. The role of wildlife in emerging and re-emerging zoonoses. Revue scientifique et technique-office international des epizooties 23 (2), 497–512.

Biewald, L. (2020). *Experiment Tracking with Weights and Biases*. https://www.wandb.com/.

Bouché, P., Lejeune, P., & Vermeulen, C. (2012). How to count elephants in West African savannahs? Synthesis and comparison of main gamecount methods. *Biotechnologie, Agronomie, Société et Environnement*, *16*(1), 77-91.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: Fast and Flexible Image Augmentations. *Information*, *11*(2), 125. https://doi.org/10.3390/info11020125.

Butt, B., Turner, M.D., 2012. Clarifying competition: The case of wildlife and pastoral livestock in East Africa. Pastoralism: Res., Policy Practice 2 (1), 9. https://doi.org/10.1186/2041-7136-2-9.

Caughley, G., 1974. Bias in Aerial Survey. J. Wildl. Manag. 38 (4), 921–933. https://doi.org/10.2307/3800067.

Caughley, G., 1977. Sampling in Aerial Survey. J. Wildl. Manag. 41 (4), 605–615. https://doi.org/10.2307/3799980.

Craig, G. C. (2012). *Aerial Survey standards for the MIKE Programme. Version 2.0*. CITES MIKE programme.

De Leeuw, J., Waweru, M. N., Okello, O. O., Maloba, M., Nguru, P., Said, M. Y., Aligula, H. M., Heitkönig, I. M. A., & Reid, R. S. (2001). Distribution and diversity of wildlife in northern Kenya in relation to livestock and permanent water points. *Biological Conservation*, *100*(3), 297–306. https://doi.org/10.1016/S0006-3207(01)00034-9.

Delplanque, A., Foucher, S., Lejeune, P., Linchant, J., Théau, J., 2022. Multispecies detection and identification of African mammals in aerial imagery using convolutional neural networks. Remote Sens. Ecol. Conserv. 8 (2), 166–179. https://doi.org/10.1002/rse2.234.

Eikelboom, J.A.J., Wind, J., van de Ven, E., Kenana, L.M., Schroder, B., de Knegt, H.J., van Langevelde, F., Prins, H.H.T., 2019. Improving the precision and accuracy of animal population estimates with aerial image object detection. Methods Ecol. Evol. 10 (11), 1875–1887. https://doi.org/10.1111/2041-210X.13277.

Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2015. The Pascal Visual Object Classes Challenge: A Retrospective. Int. J. Comput. Vis. 111 (1), 98–136. https://doi.org/10.1007/s11263-014-0733-5.

Fynn, R. W. S., Augustine, D. J., Peel, M. J. S., & de Garine-Wichatitsky, M. (2016). Strategic management of livestock to improve biodiversity conservation in African savannahs: A conceptual basis for wildlife–livestock coexistence. *Journal of Applied Ecology*, *53*(2), 388–397. .

Gao, G., Gao, J., Liu, Q., Wang, Q., & Wang, Y. (2020). CNN-based Density Estimation and Crowd Counting: A Survey. *ArXiv:2003.12783*.

Georgiadis, N.J., Ihwagi, F., Olwero, J.G.N., Romañach, S.S., 2007. Savanna herbivore dynamics in a livestock-dominated landscape. II: Ecological, conservation, and management implications of predator restoration. Biol. Conserv. 137 (3), 473–483. https://doi.org/10.1016/j.biocon.2007.03.006.

Girshick, R. (2015, December). Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 1440-1448). https://doi.org/10.1109/ICCV.2015.169.

Gower, J.C., Ross, G.J.S., 1969. Minimum Spanning Trees and Single Linkage Cluster Analysis. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) 18 (1), 54–64. https://doi.org/10.2307/2346439.

Grimsdell, J.J.R., Westley, S., 1981. Low-level aerial survey techniques. International Livestock Centre for Africa.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, 770–778. https://doi.org/10.1109/CVPR.2016.90.

Herrero, M., Grace, D., Njuki, J., Johnson, N., Enahoro, D., Silvestri, S., & Rufino, M. C. (2013). The roles of livestock in developing countries. *Animal*, *7*, 3–18. https://doi.org/10.1017/S1751731112001954.

Jachmann, H. (2002). Comparison of aerial counts with ground counts for large African herbivores. *Journal of Applied Ecology*, *39*(5), 841–852. https://doi.org/10.1046/j.1365-2664.2002.00752.x.

Kellenberger, B., Volpi, M., Tuia, D., 2017. Fast animal detection in UAV images using convolutional neural networks. IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 2017, 866–869. https://doi.org/10.1109/IGARSS.2017.8127090.

Kellenberger, B., Marcos, D., Tuia, D., 2018. Detecting mammals in UAV images: Best practices to address a substantially imbalanced dataset with deep learning. Remote Sens. Environ. 216, 139–153. https://doi.org/10.1016/j.rse.2018.06.028.

Kellenberger, B., Marcos, D., Lobry, S., Tuia, D., 2019a. Half a Percent of Labels is Enough: Efficient Animal Detection in UAV Imagery Using Deep CNNs and Active Learning. IEEE Trans. Geosci. Remote Sens. 57 (12), 9524–9533. https://doi.org/10.1109/tgrs.2019.2927393.

Kellenberger, B., Marcos, D., & Tuia, D. (2019b). When a Few Clicks Make All the Difference: Improving Weakly-Supervised Wildlife Detection in UAV Images. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1414–1422. https://doi.org/10.1109/CVPRW.2019.00182.

Kellenberger, B., Veen, T., Folmer, E., Tuia, D., 2021. 21 000 birds in 4.5 h: Efficient large-scale seabird detection with machine learning. Remote Sens. Ecol. Conserv. 7 (3), 445–460. https://doi.org/10.1002/rse2.200.

Kingma, D.P., Ba, J., 2017. Adam: A Method for Stochastic Optimization. ArXiv: 1412.6980.

Kruskal, J.B., 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. Am. Math. Soc. 7 (1), 48–50. https://doi.org/10.1090/S0002-9939-1956-0078686-7.

Lamprey, R., Ochanda, D., Brett, R., Tumwesigye, C., Douglas-Hamilton, I., 2020a. Cameras replace human observers in multi-species aerial counts in Murchison Falls, Uganda. Remote Sens. Ecol. Conserv. 6 (4), 529–545. https://doi.org/10.1002/rse2.154.

Lamprey, R., Pope, F., Ngene, S., Norton-Griffiths, M., Frederick, H., Okita-Ouma, B., Douglas-Hamilton, I., 2020b. Comparing an automated high-definition oblique camera system to rear-seat-observers in a wildlife survey in Tsavo, Kenya: Taking multi-species aerial counts to the next level. Biol. Conserv. 241, 108243 https://doi.org/10.1016/j.biocon.2019.108243.

Lempitsky, V., & Zisserman, A. (2010). Learning To Count Objects in Images. *Advances in Neural Information Processing Systems*, *23*.

Li, B., Huang, H., Zhang, A., Liu, P., Liu, C., 2021. Approaches on crowd counting and density estimation: A review. Pattern Anal. Appl. 24 (3), 853–874. https://doi.org/10.1007/s10044-021-00959-z.

Liang, D., Xu, W., Zhu, Y., Zhou, Y., 2021. Focal Inverse Distance Transform Maps for Crowd Localization and Counting in Dense Crowd. ArXiv:2102.07925.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (pp. 740–755). Springer International Publishing. https://doi.org/10.1007/978-3-319-10602-1_48.

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017, October). Focal Loss for Dense Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 2999-3007). https://doi.org/10.1109/ICCV.2017.324.

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 Rev. 45, 239–252. https://doi.org/10.1111/mam.12046.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 21–37). Springer International Publishing. https://doi.org/10.1007/978-3-319-46448-0_2.

Liu, J., Gao, C., Meng, D., & Hauptmann, A. G. (2018). DecideNet: Counting Varying Density Crowds Through Attention Guided Detection and Density Estimation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5197–5206. https://doi.org/10.1109/CVPR.2018.00545.

Naidoo, K. (2019). MiSTree: A Python package for constructing and analysing Minimum Spanning Trees. *Journal of Open Source Software*, *4*(42), 1721. https://doi.org/10.21105/joss.01721.

Naudé, J., & Joubert, D. (2019). The Aerial Elephant Dataset: A New Public Benchmark for Aerial Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 48-55).

Norton-Griffiths, M., 1978. Counting animals. Serengeti Ecological Monitoring Programme, African Wildlife Leadership.

Odadi, W.O., Riginos, C., Rubenstein, D.I., 2018. Tightly bunched herding improves cattle performance in African savanna rangeland. Rangel. Ecol. Manage. 71 (4), 481–491. https://doi.org/10.1016/j.rama.2018.03.008.

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V., Underwood, E.C., Kassem, K.R., 2001. Terrestrial Ecoregions of the World: A New Map of Life on EarthA new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. Bioscience 51 (11), 933–938. https://doi.org/10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2.

Padubidri, C., Kamilaris, A., Karatsiolis, S., Kamminga, J., 2021. Counting sea lions and elephants from aerial photography using deep learning with density maps. Anim. Biotelem. 9 (1), 27. https://doi.org/10.1186/s40317-021-00247-x.

Pan, S.J., Yang, Q., 2010. A Survey on Transfer Learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191.

Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., & Lin, D. (2019). Libra R-CNN: Towards Balanced Learning for Object Detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 821–830. https://doi.org/10.1109/CVPR.2019.00091.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, *32*.

Peng, J., Wang, D., Liao, X., Shao, Q., Sun, Z., Yue, H., Ye, H., 2020. Wild animal survey using UAS imagery and deep learning: Modified Faster R-CNN for kiang detection in Tibetan Plateau. ISPRS J. Photogramm. Remote Sens. 169, 364–376. https://doi.org/10.1016/j.isprsjprs.2020.08.026.

Pielawski, N., & Wählby, C. (2020). Introducing Hann windows for reducing edge-effects in patch-based image segmentation. *PLOS ONE*, *15*(3), e0229839. https://doi.org/10.1371/journal.pone.0229839.

Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7263-7271.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems*, *28*.

Ribera, J., Güera, D., Chen, Y., & Delp, E. J. (2019). Locating Objects Without Bounding Boxes. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6472-6482). https://doi.org/10.1109/CVPR.2019.00664.

Richard, D., Alary, V., Corniaux, C., Duteurtre, G., Lhoste, P., 2019. Dynamique des élevages pastoraux et agropastoraux en Afrique intertropicale. Éditions Quae, p. (p. 268)..

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (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.

Schlossberg, S., Chase, M. J., & Griffin, C. R. (2016). Testing the Accuracy of Aerial Surveys for Large Mammals: An Experiment with African Savanna Elephants (Loxodonta africana). *PLoS ONE*, *11*(10), e0164904. https://doi.org/10.1371/journal.pone.0164904.

Scholte, P., Kari, S., & Moritz, M. (2022a). Thousands of pastoralists seek refuge in Waza National Park, Cameroon. *Oryx*, *56*(3), 330-330. https://doi:10.1017/S0030605322000217.

Scholte, P., Pays, O., Adam, S., Chardonnet, B., Fritz, H., Mamang, J.-B., Prins, H.H.T., Renaud, P.-C., Tadjo, P., Moritz, M., 2022b. Conservation overstretch and long-term decline of wildlife and tourism in the Central African savannas. Conserv. Biol. 36, e13860.

Shrivastava, A., Gupta, A., & Girshick, R. (2016). Training Region-Based Object Detectors with Online Hard Example Mining. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 761–769. .

Tkachenko, M., Malyuk, M., Shevchenko, N., Holmanyuk, A., & Liubimov, N. (2021). *Label Studio: Data labeling software*. https://github.com/heartexlabs/label-studio.

Torney, C.J., Lloyd-Jones, D.J., Chevallier, M., Moyer, D.C., Maliti, H.T., Mwita, M., Kohi, E.M., Hopcraft, G.C., 2019. A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images. Methods Ecol. Evol. 10 (6), 779–787. https://doi.org/10.1111/2041-210X.13165.

Toutain, B., Visscher, M.-N.-D., Dulieu, D., 2004. Pastoralism and Protected Areas: Lessons Learned from Western Africa. Hum. Dimens. Wildl. 9 (4), 287–295. https://doi.org/10.1080/10807120490505963.

Vandermeer, J.H., 2002. Tropical Agroecosystems. CRC Press.

Wacher, T., 2019. Aerial survey of the Ennedi Massif. Zoological Society of London.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Trans. Image Process. 13 (4), 600–612. https://doi.org/10.1109/TIP.2003.819861.

Yu, F., Wang, D., Shelhamer, E., & Darrell, T. (2018). Deep Layer Aggregation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2403–2412. https://doi.org/10.1109/CVPR.2018.00255.

Zhang, Y., Zhou, D., Chen, S., Gao, S., & Ma, Y. (2016). Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 589–597. https://doi.org/10.1109/CVPR.2016.70.

Zhao, Z.-Q., Zheng, P., Xu, S.-T., Wu, X., 2019. Object Detection With Deep Learning: A Review. IEEE Trans. Neural Networks Learn. Syst. 30 (11), 3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865.

Zhou, X., Wang, D., Krähenbühl, P., 2019. Objects as Points. ArXiv:1904.07850.