Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Monitoring mammalian herbivores via convolutional neural networks implemented on thermal UAV imagery

Diego Bárbulo Barrios [a,b], João Valente [b,*], Frank van Langevelde [a]

[a] Wildlife Ecology and Conservation Group, Wageningen University & Research, Droevendaalsesteeg 3a, 6708 PB Wageningen, The Netherlands
[b] Information Technology Group, Wageningen University & Research, Hollandseweg 1, 6706 KN Wageningen, The Netherlands

ARTICLE INFO

Keywords:
Animal monitoring
Deep learning
Instance segmentation
Tracking
Livestock

ABSTRACT

Lightweight Unmanned Aerial Vehicles (UAVs) are emerging as a remote sensing survey tool for animal monitoring in several fields, such as precision livestock farming. Together with state-of-the-art computer vision techniques, UAV technology has drastically escalated our ability to acquire and analyse visual data in the field, lowering both costs and complications associated with collection and analysis. This paper addresses monitoring mammalian herbivores using the unexploited field of thermal Multi-Object Tracking and Segmentation (MOTS) in UAV imagery. In our research, a state-of-the-art MOTS algorithm (Track R-CNN) was trained and evaluated in the segmentation, detection and tracking of dairy cattle. Data collection was carried out in two farms with a UAV carrying a thermal camera at various angles and heights, and under different light (overcast/sunny) and thermal (16.5 °C range) conditions. Our findings suggest that dataset diversity and balance, especially regarding the range of conditions under which the data was collected, can significantly enhance tracking efficiency in specific scenarios. For training the algorithm, transfer learning was used as a knowledge migration method. The performance of our best model (68.5 sMOTSA, 79.6 MOTSA, 41 IDS, 100 % counting accuracy, and 87.2 MOTSP), which utilizes 3D convolutions and an association head, demonstrates the applicability and optimal performance of Track R-CNN in detecting, tracking, and counting herbivores in UAV thermal imagery under heterogenous conditions. Our findings demonstrate that 3D convolutions outperform Long-short Term Memory (LSTM) convolutions. However, LSTM convolutions also show optimal performance, offering a viable alternative. Furthermore, our results highlight the inability of Optical Flow to track motionless animals (-15 sMOTSA, −4.1 MOTSA and 2076 IDS) and the proficiency of the association head in differentiating static animals from the background. This research contributes to the growing body of knowledge in automated mammalian herbivore monitoring, with potential applications such as precision livestock farming and wildlife conservation.

1. Introduction

Monitoring animals in the context of extensive livestock farming can provide farmers with knowledge on crucial matters such as health issues, counts, thievery, strayed individuals, incursions by other farms' herds, and the state of the animals' environment (Xu et al., 2020 Apr; Shao et al., 2019; Xu et al., 2020; Barbedo and Koenigkan, 2018; Rivas et al., 2018; Barbedo et al., 2019; Barbedo et al., 2020). Also monitoring wildlife populations is essential in nature conservation, especially in the face of anthropogenic pressures on biodiversity, such as poaching, habitat degradation and agricultural activities (Delplanque et al., 2021; Duporge et al., 2021; Dujon et al., 2021; Linchant et al., 2015; Ceballos et al., 2015; Andrew et al., 2017; de Knegt et al., 2021). Rapid acquisition and analysis of accurate data is crucial to monitor and understand animal distribution (Duporge et al., 2021; Andrew et al., 2017). However, the vast and remote areas that require surveying, together with poor communication infrastructure, difficult ground access, and presence of visual clutter (i.e., vegetation and fog), make effective monitoring particularly challenging (Le et al., 2021). Manned flights of light aircraft, a common survey method, allow for rapid visual inspection of rangelands, but they are costly, dangerous and can disturb the animals (Barbedo and Koenigkan, 2018; Delplanque et al., 2021; Barbedo et al., 2020; Burke et al., 2019; Longmore et al., 2017). Besides, manually carrying out tasks such as locating and counting is prone to observer bias and optical illusions (Rivas et al., 2018; Andrew et al., 2017; Burke et al., 2019; Eikelboom et al., 2019). In this context, remote

[*] Corresponding author.
E-mail address: joao.valente@wur.nl (J. Valente).

https://doi.org/10.1016/j.compag.2024.108713
Received 19 April 2023; Received in revised form 11 January 2024; Accepted 3 February 2024
Available online 9 February 2024

sensing is a potential solution (Xu et al., 2020; Barbedo et al., 2019). Satellites are not well-suited for this task, as cloud cover occludes the animals and images with sufficient resolution are still expensive (Barbedo and Koenigkan, 2018; Duporge et al., 2021). On the other hand, Unmanned Aerial Vehicles (UAVs) combined with deep learning techniques for data processing are emerging as technology capable of revolutionizing animal monitoring (Delplanque et al., 2021; Dujon et al., 2021; Linchant et al., 2015; Andrew et al., 2017). Moreover, UAV technology enables (1) acquisition of high-resolution visual data in areas of difficult access; (2) performance of flights at low altitudes with scant risk of disturbing the monitored animals; and (3) collection of data through a wide array of compatible sensors.

The advantages of remote sensing techniques for animal monitoring are driving a steady adoption of UAVs and computer vision technology in wildlife conservation (Xu et al., 2020; Shao et al., 2019; Barbedo et al., 2019; Duporge et al., 2021) and, to a lower degree, in livestock farming (Barbedo and Koenigkan, 2018; Barbedo et al., 2019; Linchant et al., 2015; Mahmud et al., 2021; García et al., 2020). In both fields, a wide range of deep learning approaches has been explored for the tasks of animal detection, counting, and tracking from aerial imagery (Xu et al., 2020; Shao et al., 2019; Barbedo et al., 2020; Bondi et al., 2018; Rivas et al., 2018; Barbedo et al., 2019; Barbedo et al., 2020; Delplanque et al., 2021). In the specific case of terrestrial mammals, studies involving UAV imagery and deep learning have been conducted on species such as hippopotamus (Lhoest et al., 2015), kangaroos (Lethbridge et al., 2019), wild turkeys (Kassim et al., 2020), rabbits (Burke et al., 2019), and cows (Longmore et al., 2017). Common challenges are differences in animal pose (Barbedo et al., 2020; Van Nuffel et al., 2015), differences in illumination (Rivas et al., 2018; Barbedo et al., 2020), presence of shadows (Rivas et al., 2018), occlusions among animals (Xu et al., 2020; Barbedo et al., 2020; Lhoest et al., 2015), occlusions by vegetation (Barbedo et al., 2020) low image resolution due to long distances between the animals and the sensor (Burke et al., 2019; Bondi et al., 2018; Lhoest et al., 2015; Israel, 2011), and different movement patterns of the UAV and the animals (Rivas et al., 2018; Barbedo et al., 2020; Bondi et al., 2018).

A promising and insufficiently researched approach is the use of camera inclination, different from nadir, for data collection. With few exceptions, such as Barbedo et al. (2020) and Xu et al. (2020), many studies using aerial imagery have chosen vertical angles to attain optimal detection accuracies through a stable ground sample distance. This approach is not viable for monitoring extensive areas, as current UAV flights are not long enough for covering large areas without oblique images (Barbedo et al., 2020).

With few exceptions, animal imagery acquired by UAVs is analysed via Convolutional Neural Networks (CNNs) with one of two approaches: object detection (Shao et al., 2019; Delplanque et al., 2021; Duporge et al., 2021; Dujon et al., 2021) or instance segmentation (Xu et al., 2020; Xu et al., 2020; Le et al., 2021; Kassim et al., 2020), with the latter attaining higher accuracies in animal detection and counting (1). The main reason behind the worse performance of object detection is the overlap of some of the bounding boxes when there is a high number of occlusions among individuals, a common situation when monitoring groups of animals. This issue, stemmed from the high amount of non-target information contained in bounding boxes, leads to lose tracking estimations and ambiguities when detections are compared to ground truths in the evaluation procedure (Xu et al., 2020; Voigtlaender et al., 2019). On the other hand, image segmentation approaches circumvent animal overlapping issues by relying on the pixel-wise delineation of the individuals. Concerning tracking, Multi-Object Tracking and Segmentation (MOTS) architectures (e.g., Track R-CNN and PointTrack) are emerging as a more efficient alternative to classic Multi-Object Tracking (MOT) frameworks (e.g., Faster R-CNN and YOLOv3) due to their higher detection performance through the accurate identification of each individual's pixels (Xu et al., 2020). Nonetheless, to the best of our knowledge, no studies on MOTS applications in aerial animal monitoring have been conducted yet, probably due to this technology's novelty. Regarding the type of imagery used in animal monitoring research, thermal infrared (TIR) imaging has been less explored than Red-Green-Blue (RGB) imaging due to the higher costs of high-resolution TIR sensors in the past (Longmore et al., 2017), and the slower development of TIR technology (Witczuk et al., 2018). Nonetheless, the drop in prices and improvement in the resolution of TIR sensors make it possible for researchers to take advantage of their higher detection rates than traditional colour imagery (Lethbridge et al., 2019; Witczuk et al., 2018) and capacity to detect homeothermic animals at night (Burke et al., 2019).

In this study, we investigate the feasibility of a UAV and TIR imaging system for animal tracking. Our goals were: (1) to assess the performance of a state-of-the-art instance segmentation framework for detecting and tracking mammalian herbivores using aerial thermal imagery; (2) to evaluate the efficiency of the system under different conditions (i.e., temperature, illumination, camera-angle, and height); (3) to compare the performance of 3D and LSTM convolutions for multi-object tracking; and (4) to compare the efficiency of an association head and optical flow warping for linking detections over time. To the best of our knowledge, this is the first study to explore the viability of Track R-CNN and optical flow on thermal multi-object tracking and segmentation.

2. Materials and methods

2.1. Data collection

Thermal aerial imagery of cattle was collected with a UAV in two outdoor cattle farms in the Netherlands (Fig. 1). The farms present a uniform terrain: ground covered by grass and lack of topographic variation and vegetation occlusions. Footage was obtained of a single breed of dairy cattle: Holstein Friesians. Data was collected on three occasions between May and June of 2021. Two collections were carried out in Wageningen, on the 6th of May and 25th of June. The third collection was carried out in Groningen, on the 14th of June. The datasets produced from these surveys depict a rich range of thermic, humidity and light conditions. Both data collections in Wageningen were carried out on overcast days with atmospheric temperatures of 10 °C and 19 °C, while the data acquisition in Groningen was performed on a sunny day under a temperature of 26.5 °C (all at noon). Both the difference in thermal contrast between the animals and the ground and the different absence/presence of sunlight clearly influence the data acquired by our thermal sensor. Fig. 2 depicts three aerial images, each captured on a different day of data collection.

The Parrot ANAFI Thermal was used to acquire thermal videos from the dairy cattle farms. The drone was equipped with thermal (FLIR Lepton 3.5 microbolometer thermal sensor) and RGB (4 k HDR camera) sensors. This makes it possible to switch from thermal to RGB mode or to merge the two in the same recording/image. See Table A1 in the Supplementary Material for the main features of the system.

Parrot ANAFI Thermal offers three colour palettes to cover an array of exploration needs in the infrared electro-magnetic spectrum. In the present study, the Spot palette (see in Fig. 2) was used because it allows the isolation (via colouring) of areas falling within a manually selected temperature range, while leaving grey the areas outside the targeted thermic range. The temperature range adjustments were based on each video's unique thermal contrast between the cattle and the background. To stabilize the thermal camera and avoid errors in the thermogram, data collection began 15 min after turning on the camera.

The UAV was controlled manually via the remote-control system. On each of the data collection trips, different flight heights and recording angles were used to generate a heterogenous database that makes possible testing different detection and tracking algorithms. Flight height ranged between 8 and 28 m and camera angle ranged between 0° (nadir) and 80°. Concerning the flight path, parallel transects were

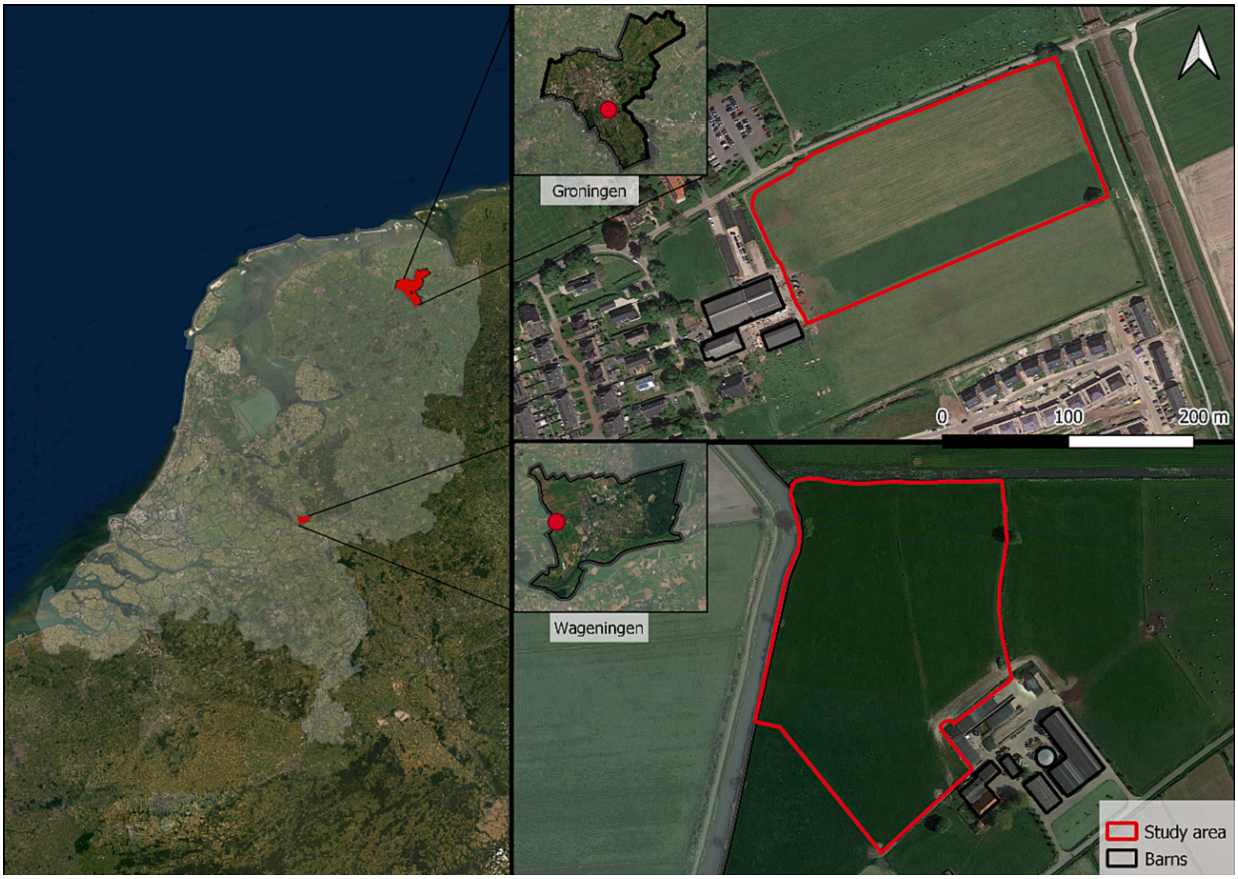

Fig. 1. Study areas: outdoor cattle farms in Groningen (51°58′20 ’’N, 5° 37′ 38′’E) and Wageningen (53° 11′06′’N, 6° 36′ 28′’E). The field in Groningen was 9.786 m^2 (~1 ha) and had 110 milking adult cows, whereas the field in Wageningen was 30.387 m^2 and had around 135 cows.

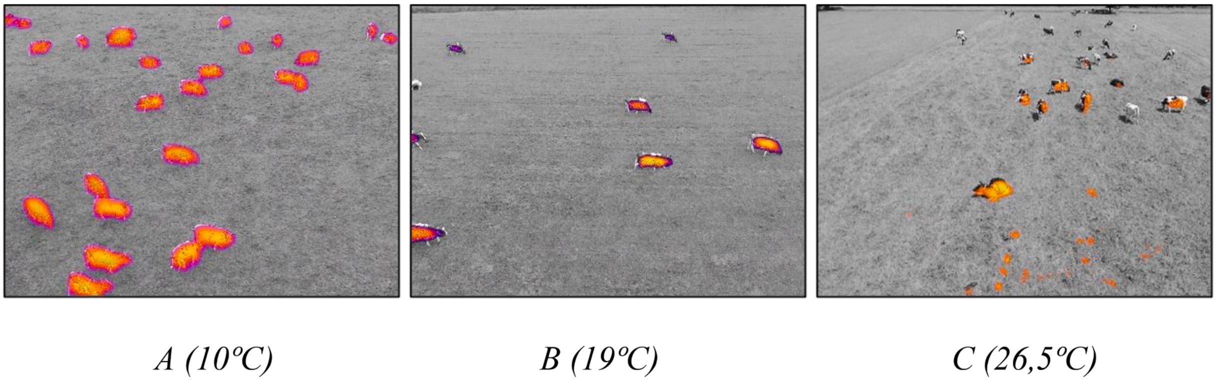

Fig. 2. Instances of thermal data collected under the Spot palette under different atmospheric conditions (temperature and sunlight); from left to right, colder to warmer conditions are portrayed. Concerning light intensity, images ***A*** and ***B*** were captured under overcast conditions, while image ***C*** was captured under sunny conditions.

the most used pattern to cover the maximum possible number of cattle. However, other flight path strategies were also followed: flying from each group of animals to the next (nearest), slow straight flight over static animals, following animals in movement, and slow forward and backward flight with the camera angled between 45° and 80°. The latter was done in order to maximize the number of cows within each video frame.

2.2. Data preparation and pre-processing

The recorded videos were captured at 9 frames per second, had a resolution of 1440x1080 pixels, and were saved in MOV format. They were manually trimmed into shorter sections to prevent multiple detection of the same individuals (i.e., double counting) and to discard sections in which the thermal sensor got mis-calibrated. Moreover, the following factors were prioritized in the trimming selection: average abundance of animals recorded, presence of animals in movement, and changes of camera inclination. Out of the resulting trimmed videos, 5 were chosen for training and 2 were chosen for testing the Convolutional Neural Networks (CNNs) (Table A2 in in the Supplementary Material). All videos were cropped into PNG images (1440x1080). The resolution was optimal for downscaling and upscaling the CNN algorithms, as they were divisible by 2 at least 6 times (Xu et al., 2020).

Ground truth for the seven datasets was labelled manually with the Computer Vision Annotation Tool (CVAT) (openvinotoolkit/cvat, 2022). The CVAT platform was selected because of the time-saving benefits of its interpolation feature. Every individual cow in every frame was manually annotated with a polygon. CVAT was run in a docker container (Documentation, 2022).

2.3. Overview of framework

A visual overview of the steps followed to conduct our study is provided in Fig. 3. Two detection and tracking algorithms were tested: Track R-CNN (Voigtlaender et al., 2019) and PWC-Net (Sun et al., 2020). These algorithms were selected based on their performance on the 2020 CVPR MOTSChallenge (Challenge and Results, 2022), a benchmark for performance evaluation of multi-target tracking and segmentation, and on the public availability of their code at the time of our study.

Two datasets, COW MOTS and the well-known KITTI MOTS (Voigtlaender et al., 2019), were used in this study. All models were trained and tested on COW MOTS and fine-tuned on KITTI MOTS. See Table A3 in the Supplementary Material for a description of these two datasets.

Extended from the detection and segmentation framework Mask R-CNN (He et al., 2017), Track R-CNN addresses tracking via an association head and two 3D convolutional layers that enable it to associate object identities over time (Voigtlaender et al., 2019). Time is the additional third dimension in the 3D convolutions, which are integrated on top of a ResNet-101 backbone; augmenting its features with temporal context. The augmented features are then inputted to the Region Proposal Network (RPN).

The association head is a fully connected layer used to link detections over time. It receives region proposals as inputs and predicts an association vector for each proposal. Each association vector represents the identity of a cow. The distance between vectors is larger for those belonging to different instances than for those belonging to the same. In this manner, new detections are linked to existing tracks based on their association vector similarity. The association head is trained using an adaptation to video sequences of the batch hard triplet loss schemed by (Hermans et al., 2017). A Hungarian algorithm is used for matching.

A powerful architecture for mask propagation, PWC-Net (Sun et al., 2020) tracks pixels across frames through optical flow estimation (Voigtlaender et al., 2019) by building a feature pyramid from each pair of adjacent images. A detailed description of the algorithm's functioning is given in (Sun et al., 2020). In the present study, we experiment with optical flow warping (mask propagation scores) as an alternative to the association vector similarities that Track R-CNN uses for tracking cow's detections across frames.

2.4. Model implementation

All our models used Track R-CNN for segmentation and detection and differed in their temporal component (3D convolutions vs. LSTM layers) and their tracking mechanism (association head vs. optical flow). We experimented with three types of structures: two models used 3D convolutional layers + association head; one used LSTM layers + association head; and one used 3D convolutional layers + optical flow. A comparison of the alternative parameters was carried out by evaluating the four Track R-CNN models on our two testing datasets (Table A2 in the Supplementary Material), which represent cold (10 °C) and warm (26.5 °C) temperatures, respectively. A tracking mechanism relying on the Euclidean distances between bounding box centres was tested and discarded after being outperformed by the mask-based association head in a preliminary assessment. Table 1 displays the configuration parameters used for each model; the best performance for all models (point of convergence) was found experimentally to be at 23 epochs.

Since the size of the annotated dataset is relatively small, training the models from scratch could have led to rapid overfitting (Xu et al., 2020; Courtney and Sreenivas, 2020). To solve this issue, fine-tuning and pre-training were used on all the models. These transfer learning techniques allowed the migration of knowledge from other datasets to ours. A Resnet101 pre-trained model on the COCO (Lin et al., 2014) and Mapillary (ICCV, 2017) datasets was used to initialize the Mask R-CNN section of the Track R-CNN framework, and all the models we used were fine-tuned on the KITTI MOTS dataset by initializing the association head with weights obtained by training on KITTI MOTS.

The models were trained using the Adam optimizer with a learning rate of 5x10^{-7} and the hyper-parameters (number of epochs and batch size) were tuned for each model following an empirical approach. The number of epochs for all models was set to 44. As Track R-CNN does not output validation loss data, versions of each model were saved at different points of the training procedure to later find the point of convergence experimentally by running the evaluation procedure on each version. Due to the limitations in data availability, a dedicated validation set was not utilized in this study. Hence, the test datasets were employed for selecting the best number of training epochs. We recognize that this approach may lead to potential data leakage, which could result in overestimating the modelś performance on unseen data. However, this method was necessary given the studýs constraints and still provides valuable insights into the modelś behaviour in the studýs context. Batch sizes of 4 and 8 were used. After training, to increase the performance of our models, the detection and tracking parameters were optimized using random search with 1000 iterations per experiment. During each iteration, the model was trained and evaluated on the training data. The best combination of hyperparameters on the training data were evaluated one final time on the test dataset(s) to obtain their metrics on previously unseen data. The tuned detection and tracking hyper-parameters were: Detection Confidence Thresholds, Re-identification (ReID) Weights, Mask and Bounding Box IOU Weights, Bounding Box Center Weights, Association Thresholds, Keep Alive Parameters, New ReID Thresholds and Flag, and Box Offset and Scale.

As evaluation measures, we followed the MOTS metrics devised by (Voigtlaender et al., 2019) as an adaptation of the well-established CLEAR MOT metrics for multi-object tracking. The used MOTS evaluation measures were: 1) the soft Multi Object Tracking and Segmentation Accuracy (sMOTSA), 2) the Multi Object Tracking and Segmentation Accuracy (MOTSA), and 3) the Multi Object Tracking and Segmentation Precision (MOTSP). Their formulas are as follows:

$$sMOTSA = \frac{\widetilde{\mathrm{TP}} - |FP| - |IDS|}{|M|}, \tag{1}$$

$$MOTSA = \frac{|TP| - |FP| - |IDS|}{|M|}, \tag{2}$$

$$MOTSP = \frac{\widetilde{\mathrm{TP}}}{|FP|}, \tag{3}$$

TP stands for true positives and refers to the masks hypothesized by the algorithm that are mapped to ground truth masks and have an IoU higher than 0.5. *TP* stands for soft true positives and refers to all hypothesized masks mapped to ground truth masks. For example, if two cows with identities “Cow1” and “Cow2” are detected by the model with an IoU, respectively, of 80 % and 44 %; the *TP* would be 50 % (average of 100 % for Cow1 and 0 % for Cow2); while the $\widetilde{TP}$ would be 62 % (average of 80 % for Cow1 and 44 % for Cow2). *FP* stands for false positives and refers to hypothesized masks that are not mapped to any ground truth mask; this metric is expected to be higher when the thermal contrast between the animals and the ground is low, as thermal signatures from hot spots in the ground can be mistaken by those of cows. *IDS* refers to the ID switches over the same identity, which are expected to increase with occlusions and fast movements of the UAV and/or the camera. *M* is the number of ground truth masks.

The Track R-CNN framework was implemented on a high-

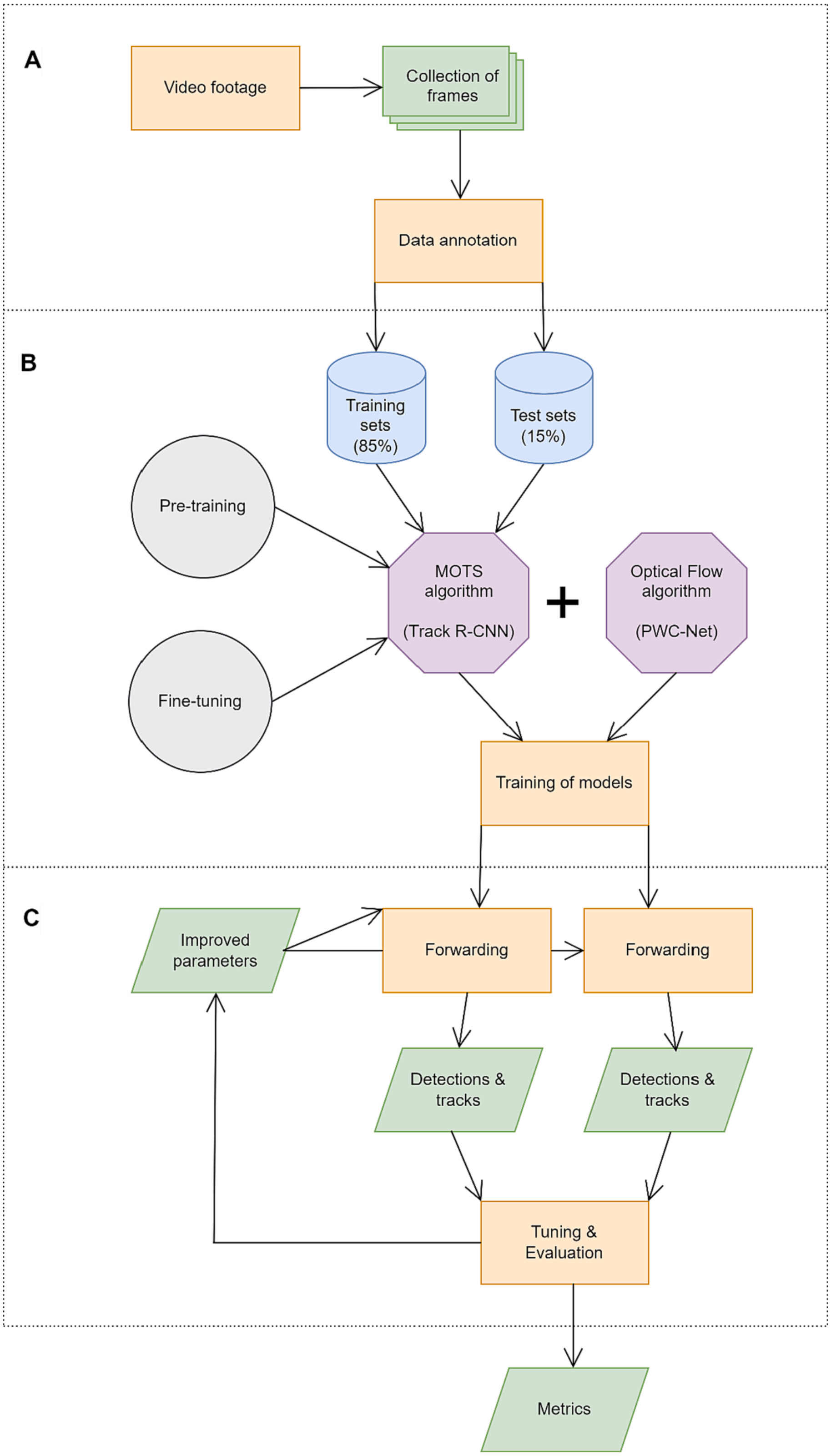

Fig. 3. Flowchart of the method used to train and test the algorithms. After evaluation, the results of the different models were compared. The three main sections of the method are: A) data pre-processing; B) data processing via the two algorithms; and C) forwarding of the results, tuning of the parameters and evaluation of the model performance.

Table 1
Models configuration.

	Temporal component	Tracking mechanism	Batch size	Epoch number
3Dconv4_23	2x3Dconvolutions	Association head	4	23
3Dconv8_23	2x3Dconvolutions	Association head	8	23
LSTM8_23	2xLSTMconvolutions	Association head	8	23
4_23_optical	2x3Dconvolutions	Optical flow	4	23

performance computer equipped with a NVIDIA Titan RTX video card with 24 GB GDDR6 with a Linux OS (Ubuntu 20.04.1 LTS). Calculations were supported by 64 GB of RAM and an Intel® Core^{TM} i9-10940X CPU @ 3.30GHZ x 28.

3. Results

3.1. Model performance on the testing dataset

The detection metric MOTSP remained rather constant for all models, varying within a range of 0.3 %. 3Dconv4_23 was the best performing model, showing that a batch size of 4 yields slightly better tracking results than a batch size of 8: the tracking metrics sMOTSA and MOTSA improved 2.3 and 3.0 %, respectively, while the number of ID switches of both models were the same. As for the counting accuracy (Table 2 & Fig. 4), a batch size of 4 yields a perfect score, while the model using a batch size of 8 underestimates the count by 8.6 %. Concerning the temporal component, the tracking results of 3Dconv8_23 and LSTM8_23 show that, for the same batch size, 3D convolutions performed marginally better than LSTM convolutions in SMOTSA and MOTSA, while a considerable relative improvement of 21.2 % in the number of ID switches is observed. The counting accuracy is not affected by the temporal component, with both models (3Dconv8_23 and LSTM8_23) underestimating the count by 8.6 %. The model 4_23_optical shows that optical flow performed poorly as an association method, degrading to negative values the SMOTSA and MOTSA obtained by the same model (3Dconv4_23) with an association head, and producing a 4963 % increase (from 41 to 2076) in the number of ID switches.

3.2. Model performance under different temperatures

The detection metric MOTSP remained stable, with a 0.4 variation between the models (Table 3 and Fig. 4). The models 3Dconv4_23 and 3Dconv8_23 show that batch size affected tracking and counting performance. sMOTSA and MOTSA have respectively 3.1 and 3.7 % improvement with the smaller batch size of 4 (3Dconv4_23) under cold conditions; and 3.6 and 2 % gain with the bigger batch size of 8 (3Dconv8_23) under warm conditions. ID switches followed the same trend, with both models 3Dconv4_23 and 3Dconv8_23 attaining a drop of 5 % under their best performing thermal conditions. Counting accuracy improves with the smaller batch size of 4 under cold conditions, with 3Dconv4_23 outperforming the other models by 7.1 %, and with a bigger batch size of 8 under warm conditions, with 3Dconv8_23 and LSTM8_23 attaining a perfect score. Table 4 shows a comparison of the best performing parameters (those of the model 3Dconv4_23) when used for training on the whole dataset, only on warm data, and only on cold data. The results illustrate the effect of each testing dataset's characteristics on detection and tracking. An improvement in all metrics can be observed for the model (3dconv4_23_warm) trained without the coldest (10 °C) datasets, with respect to the same model (3Dconv4_23) trained on the whole COW MOTS dataset (i.e., on the three sampled temperatures: 10 °C, 19 °C, and 26.5 °C). The improvement in mask detection accuracy is pronounced, with an increase of 10.7 % in MOTSP. Concerning tracking results, the improvement is even greater: 68.5 % in SMOTSA, 64.3 % in MOTSA, and an attainment of 0 ID switches by 3Dconv4_23_warm. Detections and tracks of the models 3Dconv4_23_cold and 3Dconv4_23_warm can be observed in Fig. 5.

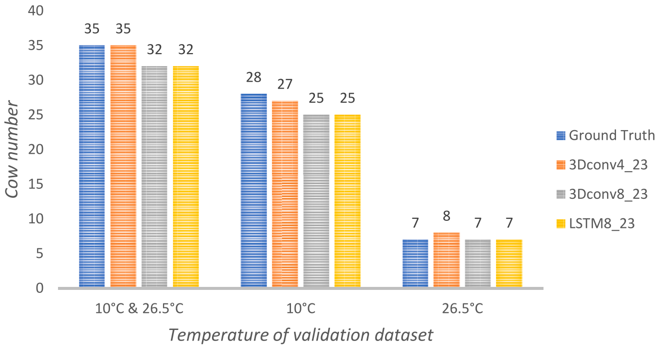

Fig. 4. Comparison of counting results of three different models evaluated on the whole testing dataset (including data collected under 10 °C and 26.5 °C), only on the cold testing dataset (10 °C), and only on the warm testing dataset (26.5 °C). The total ground truth is included as a baseline to assess the models accuracy.

4. Discussion

Our results demonstrate the applicability and optimal performance of a MOTS algorithm (Track R-CNN) to detect, track and count mammalian herbivores using UAV thermal imagery taken from different heights and angles, under different temperatures and light conditions.

In this study, while the methodological approach of using test data for selecting training epochs presents a risk of data leakage, we believe that the fundamental objectives of the study remain sound. These objectives include assessing the performance of a state-of-the-art instance segmentation architecture in tracking mammalian herbivores using aerial thermal imagery, evaluating the systeḿs efficiency under various conditions, comparing the performance of 3D and LSTM convolutions for multi-object tracking, and analysing the efficiency of two different tracking mechanisms. The broad and thorough scope of the research analysis supports the validity and importance of our findings despite the mentioned methodological constraints.

4.1. Analysis of models trained on the whole dataset

Under cold conditions (10 °C), the high MOTSP values show that our algorithm's detection performance overcomes many variations in animal size caused by the range of heights and angles used to monitor the animals. The values of sMOTSA, MOTSA and IDS demonstrate our algorithm's capacity to deal with the challenging features of the colder

Table 2
Mask tracking, detection and counting results of our models on the COW MOTS testing datasets.

	sMOTSA	MOTSA	IDS	FP	FN	Counting accuracy	MOTSP
3Dconv4_23	**68.5**	**79.6**	**41**	**149**	**313**	**100 %**	87.2
3Dconv8_23	66.2	76.6	41	134	402	91.4 %	87.5
LSTM8_23	65.0	75.5	52	126	427	91.4 %	87.3
4_23_optical	−15	−4.1	2076	–	–	–	87.3

Table 3
Mask tracking, detection and counting results of our best models on cold and warm data of COW MOTS dataset.

	Cold							Warm						
	sMOTSA	MOTSA	IDS	FP	FN	Counting accuracy	MOTSP	sMOTSA	MOTSA	IDS	FP	FN	Counting accuracy	MOTSP
3Dconv4_23	**76.0**	**86.0**	**29**	**38**	**237**	**96.4 %**	88.7	14.3	33.7	12	111	76	85.7 %	74.0
3Dconv8_23	72.9	82.3	34	40	310	89.3 %	89.0	**17.9**	**35.7**	**7**	**94**	**92**	**100 %**	73.8
LSTM8_23	72.1	81.8	41	30	325	89.3 %	88.6	13.6	30.3	11	96	102	100 %	74.7

Table 4
Mask tracking results, on the warm and cold testing datasets, of our best performing model (3Dconv4_23) trained on different temperatures[2].

	Cold						Warm					
	sMOTSA	MOTSA	IDS	FP	FN	MOTSP	sMOTSA	MOTSA	IDS	FP	FN	MOTSP
3Dconv4_23	**76.0**	**86.0**	**29**	**38**	**237**	**88.7**	14.3	33.7	12	111	76	74.0
3Dconv4_23_warm	–	–	–	–	–	–	**82.8**	**98.0**	**0**	**4**	**2**	**84.7**
3Dconv4_23_cold	72.2	81.0	46	42	303	88.6	–	–	–	–	–	–

[2] The first row shows the mask tracking results, on the warm and cold testing datasets, of our best performing model (3Dconv4_23). The second and third rows show mask tracking results of the same model trained on and warm data. 3Dconv4_23_warm has been trained on temperatures of 19 °C and 26.5 °C; 3Dconv4_23_cold has been trained on temperatures of 10 °C and 19 °C.

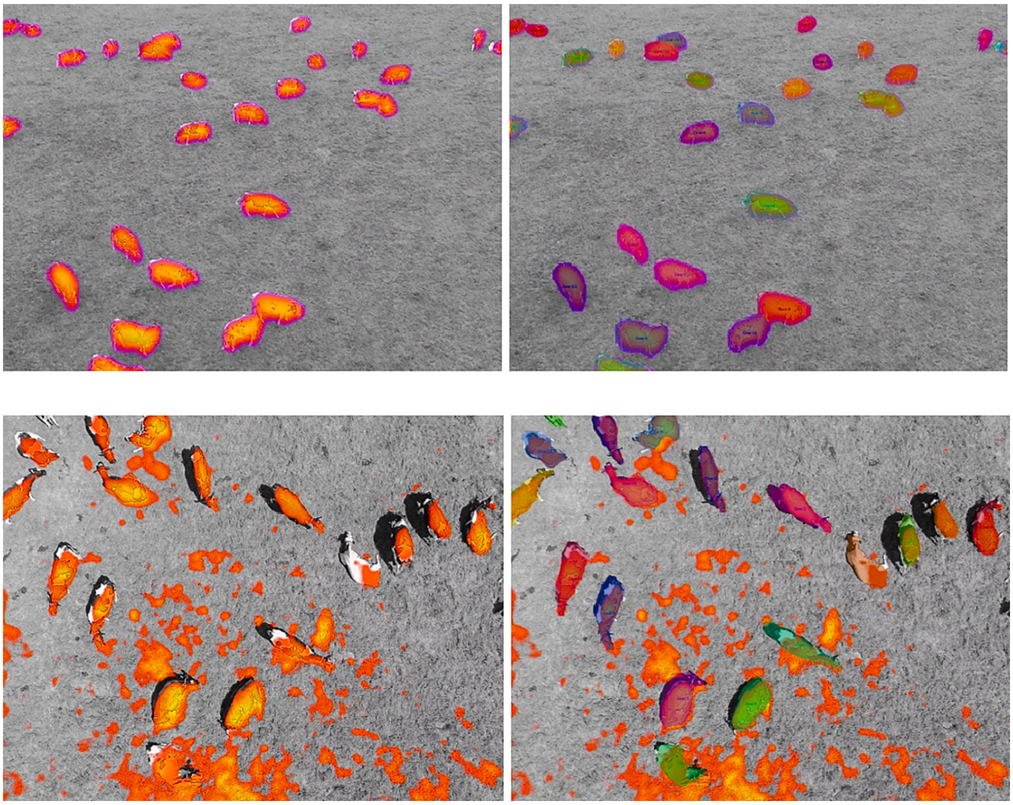

Fig. 5. Original images (top-left and bottom-left) from the two testing datasets and their tracks (top-right and bottom-right). The upper left image belongs to the cold dataset (10 °C). The bottom left image belongs to the warm dataset (26.5 °C).

testing dataset (0003), such as the recurrent occlusions caused by the high density of animals, the use of oblique images and the frequent animal movements. Visual inspection of the tracks shows that most ID switches take place on the animals farthest from the UAV, probably because their features are harder to discern. Therefore, we believe that an increase in the dataset's size and in the videos resolution would further reduce misidentifications. As for counting accuracy (Fig. 4 & Table 3), results show that all our models display good performance under the challenging conditions of the colder scenario.

Under warm conditions (26.5 °C), the MOTSP values (Table 3) of all our models show a lower detection efficiency than under cold conditions. Even though all cows are detected, visual inspection reveals that the algorithm tends to make more than one identification in each cow. We argue that these errors are caused by the higher presence of cattle under cold than under warm conditions in the training dataset. The notable visual differences between the animals' thermal signatures at different temperatures biased the neural networks to recognize the features of the well-defined thermal signatures at 10 °C more accurately than those of the fragmented thermal signatures at 26.5 °C. Having multiple detections on individuals is behind the poor MOTSA, sMOTSA and IDS metrics observed in Table 3. On the other hand, the counting performance (Fig. 4 and Table 3) of our models is higher under the warm scenario than under the cold one, with the most likely cause being the lack of occlusions and animal size variations in the warm dataset. The challenges posed by oblique recording angles (occlusions and animal size variations) are illustrated in Fig. 6, where erroneous detections and identity assignments are more prevalent in animals further from the sensor and in animals overlapping each other. These results exhibit that counting and detection accuracy are influenced differently by the conditions under which the records were taken and by the composition of the training dataset.

A notable challenge in our study was the unstable detection and tracking of overlapping cattle at some oblique angles. Track R-CNN proved uncapable of reliably handling these dynamic overlaps under such conditions, resulting in inconsistent identity assignments across frames, as illustrated in Fig. 7. However, it is important to note that on some instances Track R-CNN successfully detected and tracked overlapping cattle, as shown in Fig. 8. To overcome the observed challenges, we propose exploring advanced post-processing techniques to maintain accurate identity assignments despite partial overlaps. Future work in this area is paramount for improving the reliability of UAV-based animal monitoring in scenarios with frequent animal interactions.

Our two temporal components, 3D convolutions and LSTM convolutions, encode the relationship between spatial and temporal information to make predictions in sequential data (Shi et al; Nabavi, 2018; Mahadevan et al., 2020; Tran et al., 2018). Both types of convolutions have been successfully applied to semantic segmentation tasks (Nabavi, 2018; Mahadevan et al., 2020) and have given similar results in instance segmentation in the KITTI MOTS dataset (Voigtlaender et al., 2019). Our results show that 3D convolutions slightly outperform LSTM convolutions on all tracking metrics when applied to the COW MOTS dataset (Table 2). The performance gap between the two models is more pronounced when applied to the warm da- taset than when applied to the cold dataset or the entire COW MOTS dataset (Table 3). We attribute this to the warm datasetś unique characteristics: it is a small-sized dataset depicting complex and fragmented thermal signatures on cattle, and similar thermal signatures on the ground. In such scenarios, the spatial processing capabilities of 3D convolutions in short temporal ranges become particularly advantageous (Zhu et al., 2017). These models are adept at discerning complex spatial patterns in the data when the temporal aspect is minimal, a task that is challenging for LSTMs (Zhu et al., 2017; Zhang et al). Since in the warm dataset cattle are mostly motionless and observed with a nadir view, temporal variability is reduced compared to the cold dataset. In this context, the long-term temporal sensitivity of LSTMs does not confer a significant advantage. Consequently, 3D convolutions demonstrate superior performance in this spatially complex but temporally uniform environment.Another reason could be the greater susceptibility of LSTM convolutions to overfit (Courtney and Sreenivas, 2020). However, the evolution of the training loss of 3Dconv8_24 and LSTM8_24 follows a similar pattern over the 23 epochs, making it unlikely that LSTM8_24 did overfit. Further research is needed to understand the better performance of 3Dconvolutions in COW MOTS.

Concerning the tracking mechanism, the association head drastically outperforms optical flow warping (Table 2), illustrating PWC-Net's inability to track in COW MOTS scenarios. Visual inspection of PWC-Net's detection results shows that the poor tracking metrics of 4_23_optical are caused by the optical flow framework's inability to detect motionless animals, as can be observed in Fig. 9. This failure presents a meaningful issue in COW MOTS, as the warm (26.5 °C) datasets are mostly composed by resting cattle (as a behavioural consequence of the higher temperature) and, even though the cold datasets present high levels of movement by most individuals, their motion is not always constant throughout the videos. To test the extent to which the warm testing dataset was affecting the metrics and the level of animal movement required by PWC-Net to perform optimally, a second evaluation was conducted solely on the cold (10 °C) testing dataset. An insufficient improvement in the tracking metrics MOTSA and sMOTSA (from −15 to −10.2 in sMOTSA, and from −4.1 to −0.5 in MOTSA) of this second evaluation proves that irregular levels of motion by all individuals throughout the videos are still insufficient for adequate tracking via optical flow. Therefore, it can be concluded that while the association head's use of association vectors to link detections over time can successfully deal with erratic or lack of movement by the animals, optical flow's reliance on pixel motion makes it unsuitable for the task.

Regarding the batch size, there is no consensus in the literature on whether generalization to test data is better with small or large batches. Smaller batch sizes are touted as a better way to achieve convergence in fewer epochs, while larger batches are considered to offer better

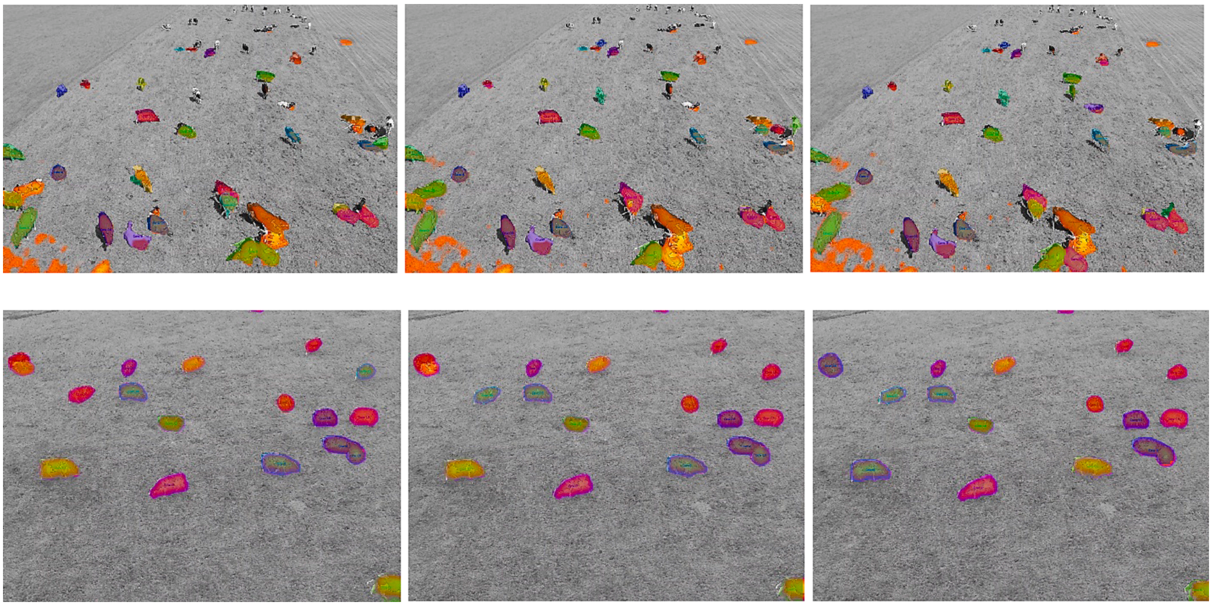

Fig. 6. Examples of detection and tracking problems under warm (26.5 °C) and cold (10 °C) scenarios, due to animal overlapping and size variations, for data collected under oblique angles. The upper three images are consecutive frames of a dataset not used for training or testing. The lower three images are consecutive frames of the cold test dataset (0003).

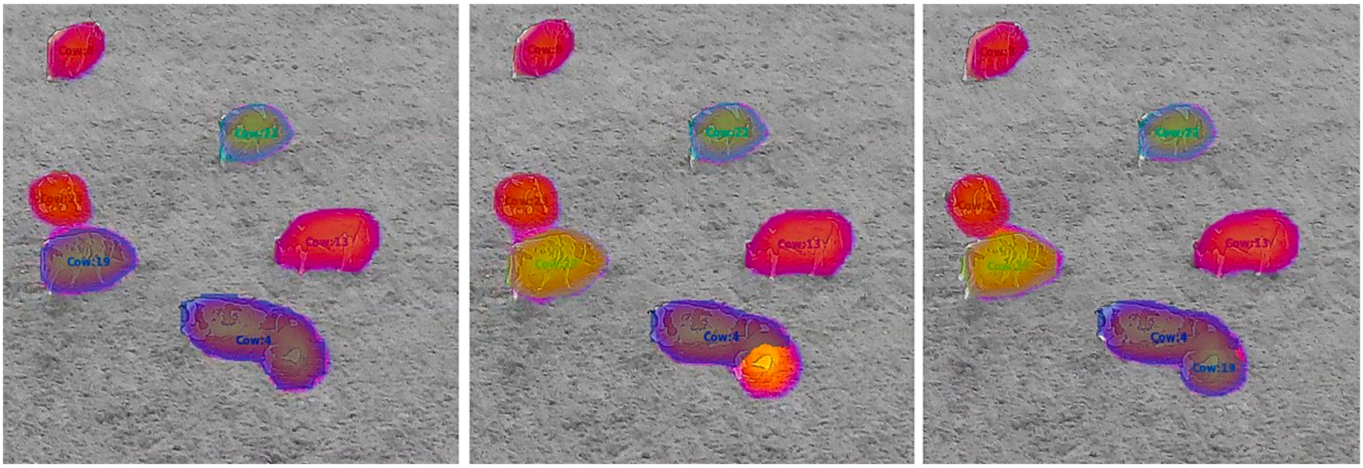

Fig. 7. Examples of detection and identity tracking challenges by 3Dconv4_23 in overlapping cattle (bottom of the image) at oblique UAV-recorded angles. Three consecutive frames of the test cold dataset (10 °C) are depicted from left to right.

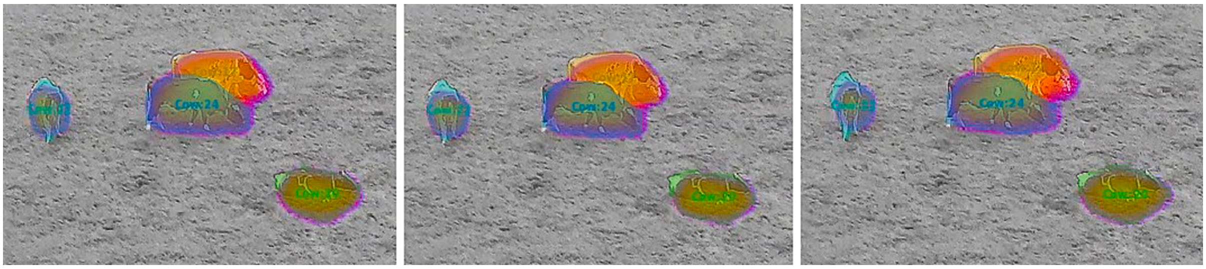

Fig. 8. Successful detection and identity tracking of overlapping cattle (top of the image) by 3Dconv4_23 in oblique UAV imagery. Three consecutive frames of the test cold dataset (10 °C) are depicted from left to right.

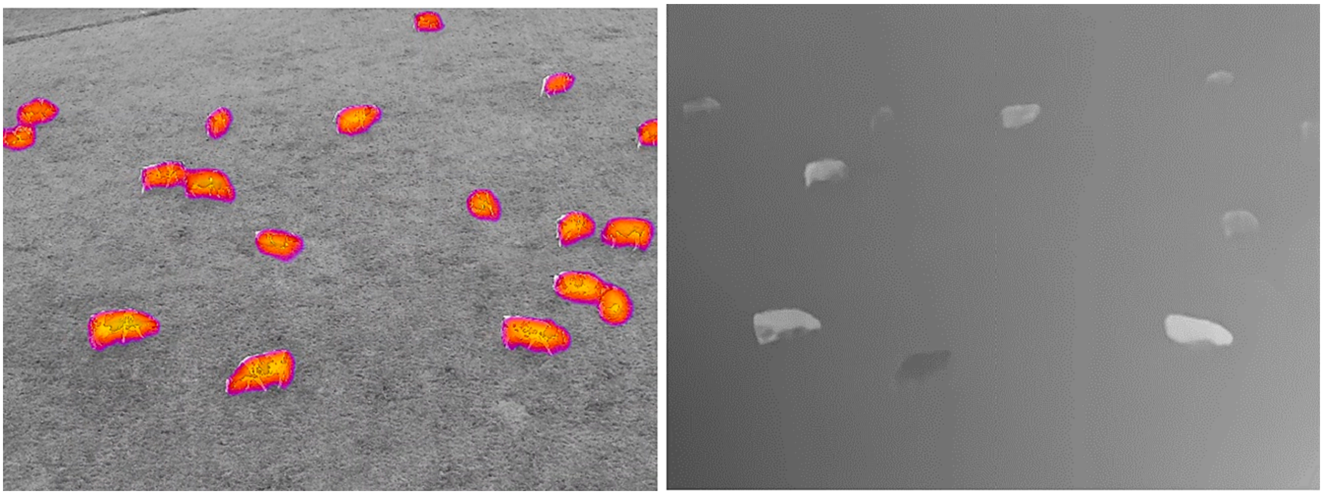

Fig. 9. U channel of the optical flow image (right) computed with PWC-Net and the flowiz (GitHub - georgegach/flowiz, 2021) package, and the original image (left), belonging to the colder testing dataset (0003).

computational efficiency (Devarakonda and Naumov, 2017). As for test accuracy, both (Devarakonda and Naumov, 2017) and (Smith et al., 2018) state that smaller batches achieve higher test accuracy, while (Radiuk and Radiuk, 2017) and (Tran et al., 2018) argue the opposite. Even though our results agree with those of (Devarakonda and Naumov, 2017) and (Smith et al., 2018), showing a slightly better overall performance with a smaller batch size of 4 (Table 2), it must be noted that the difference between both sizes is smaller than those used in the cited literature. We found no differences in the time needed to reach convergence. Future research in animal monitoring should experiment with bigger differences in batch size and with increasing the batch size over training; a promising method for solving the trade-off between smaller and larger batches which provided optimal results to (Devarakonda and Naumov, 2017).

4.2. Analysis of models trained on different temperatures

Surprisingly, the model 3Dconv4_23_cold produced worse tracking metrics under cold conditions (10 °C) than its homologous 3Dconv4_23, despite the first having been exclusively trained on data for the cold conditions. We hypothesize that there are several possible factors contributing to this behaviour: 1) the diversification of the dataset with warm data can prevent the model from overfitting to the characteristics of the data collected under cold conditions and learn more robust features; 2) while there are significant differences between the data collected under warm and cold conditions, the inclusion of the warm data may improve the model's understanding of the universal features that define cattle; 3) the constant speed at which warm data was recorded may be compensating for the pitfalls of the several rapid turns of the camera in training datasets collected at 10 °C (0000 and 0001), with each one blurring most cows over a few frames. To test this conjecture, we ran the evaluation procedure of 3Dconv4_23_cold with the tracking parameters of 3Dconv4_23, which further degraded the tracking metrics. This illustrates that the problem lays in the features learned by the network during training and not on the tracking parameters. We conclude that diversifying the dataset and increasing its size helped the model to generalize better in the context of tracking, and that Track R-CNN is highly sensible to abrupt changes in recording speed and recommend future studies to avoid them. The detection metric, on the other hand, is similar in both models. The reason is probably the high thermal contrast between the animals and the background at 10 °C, which makes the thermal signature of the cows easily singled out against a grey background; regardless of whether warm datasets were also used for training or not.

On the other hand, the model 3Dconv4_23_warm shows a considerable gain in detection and, especially, tracking accuracy compared to 3Dconv4_23, which can be attributed to the substantial increase in the proportion of warm training data. The slightly lower detection results obtained with 3Dconv4_23_warm than with 3Dconv4_23_cold can be attributed to the higher difficulty of detecting cows with thermal imagery when background temperatures are similar to those of the animals. In this scenario, thermal signatures appear in the grass (Witczuk et al., 2018) and the ones on the cows tend to be small or fragmented. Moreover, shadows make for potential false positives, as there are black cows in the herd. However, the drastically low number of false positives and false negatives attests to the good performance of Track R-CNN. We hypothesize that the tracking metrics mostly benefit from the lack of occlusions, prevented by the nadir view and the low activity of the animals under high temperatures.

Even though we have obtained optimal detection and tracking results, we cannot ignore the benefits of collecting data in the homogenous grasslands of the farms. The greater heterogeneity encountered in the wild means that future research on animal monitoring should assess MOTS performance under more challenging scenarios with more than one species. With these results, our study contributes to the research on automated monitoring of mammalian herbivores, which can be applied to various contexts such as wildlife ecology and conservation, and precision livestock farming.

5. Conclusions

This study demonstrates the effective application of Track R-CNN for detecting, tracking, and counting mammalian herbivores on UAV thermal imagery collected under a spectrum of outdoor farm conditions. Our results show that, among the temporal components tested, 3D convolutions outperform LSTM convolutions, especially in scenarios with high spatial complexity and low temporal variation. However, LSTM convolutions also show optimal performance, offering viable alternatives in similar applications. Regarding the tracking mechanisms, our study favors the association head over optical flow. The latter proved to be unable to differentiate between static animals and the background, highlighting a significant limitation in tracking scenarios without constant target movement, which is a common occurrence in animal monitoring. Additionally, our findings suggest that dataset diversity and balance, concerning the range of conditions under which the data was collected, significantly enhance tracking efficiency in specific conditions. This improvement in generalization capacity is notable even when considering factors such as animal size variability and occlusions. Moreover, this study sheds light on the limitations of Track R-CNN's tracking capabilities in scenarios involving rapid UAV camera movements, suggesting a need for further refinement in dynamic monitoring contexts.

While we have obtained optimal detection and tracking results, we cannot ignore the advantages of collecting data in the relatively homogenous grasslands of outdoor farm settings. Future research should explore these MOTS techniques in more heterogenous and challenging settings, including multiple species, to fully assess their applicability in precision livestock farming, wildlife ecology, and conservation. By pushing the boundaries of existing knowledge in thermal multi-object tracking and segmentation, our study makes a significant contribution to the novel field of automated mammalian herbivore monitoring, providing a valuable reference for future advancements.

6. Data accessibility

We make publicly available our annotated thermal livestock dataset at https://doi.org/10.5281/zenodo.6370315 and provide a repository illustrating the implementation of the deep learning frameworks: https://github.com/Diego-Barbulo/Instance-Segmentation-and-Tracking.

CRediT authorship contribution statement

Diego Barbulo Barrios: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **João Valente:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing. **Frank van Langevelde:** Investigation, Project administration, Supervision, Writing – original draft, Writing – review & editing.

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.

Data availability

The data is available in a public repository

Acknowledgements

We would like to thank Wouter Schimmel and Henk-Jan Hoving for offering their farms for the data collection, and Hilmy Baja for his valuable advice on implementing Optical Flow within Track R-CNN.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2024.108713.

References

Andrew, M.E., Shephard, J.M., Margaret Andrew, C.E., Buchanan, G., 2017. Semi-automated detection of eagle nests: an application of very high-resolution image data and advanced image analyses to wildlife surveys. Remote Sens. Ecol. Conserv. 3 (2), 66–80.

Barbedo, J.G.A., Koenigkan, L.V. Perspectives on the use of unmanned aerial systems to monitor cattle: https://doi.org/101177/0030727018781876. 2018 Jun 24.

Barbedo, J.G.A., Koenigkan, L.V., Santos, T.T., Santos, P.M. A study on the detection of cattle in UAV images using deep learning. Sensors 2019, Vol 19, Page 5432019 Dec 10;19(24):5436.

Barbedo, J.G.A., Koenigkan, L.V., Santos, P.M., Ribeiro, A.R.B. Counting cattle in UAV Images—dealing with clustered animals and animal/background contrast changes. Sensors 2020, Vol 20, Page 2126. 2020 Apr 10;20(7):2126.

Barbedo, J.G.A., Koenigkan, L.V., Santos, P.M. Cattle detection using oblique UAV images. Drones 2020, Vol 4, Page 75. 2020 Dec 8;4(4):75.

Bondi, E., Fang, F., Hamilton, M., Kar, D., Dmello, D., Choi, J, et al. SPOT poachers in action: augmenting conservation drones with automatic detection in near real time. Thirty-Second AAAI Conference on Artificial Intelligence. 2018 Apr 27.

Burke, C., Rashman, M., Wich, S., Symons, A., Theron, C., Longmore, S. Optimizing observing strategies for monitoring animals using drone-mounted thermal infrared cameras. https://doi.org/101080/0143116120181558372. 2019 Jan 17;40(2): 439–67.

Ceballos, G., Ehrlich, P.R., Barnosky, A.D., García, A., Pringle, R.M., Palmer, T.M., 2015. Accelerated modern human-induced species losses: entering the sixth mass extinction. Sci. Adv.

MOT Challenge - Results [Internet]. [cited 2022 Feb 2]. Available from: https://motchallenge.net/results/CVPR_2020_MOTS_Challenge/.

Courtney, L., Sreenivas, R. Comparison of Spatiotemporal Networks for Learning Video Related Tasks. 2020 Sep 15.

de Knegt, H.J., Eikelboom, J.A.J., van Langevelde, F., Spruyt, W.F., Prins, H.H.T. Timely poacher detection and localization using sentinel animal movement. Scientific Reports 2021 11:1. 2021 Feb 25;11(1):1–11.

Delplanque, A., Foucher, S., Lejeune, P., Linchant, J., Théau, J., 2021. Multispecies detection and identification of African mammals in aerial imagery using convolutional neural networks. Remote Sens. Ecol. Conserv.

Devarakonda, A., Naumov, M., 2017 Dec. Garland M. Adaptive Batch Sizes for Training Deep Neural Networks, AdaBatch, p. 6.

Docker Documentation | Docker Documentation [Internet]. [cited 2022 Feb 2]. Available from: https://docs.docker.com/.

Dujon, A.M., Ierodiaconou, D., Geeson, J.J., Arnould, J.P.Y., Allan, B.M., Katselidis, K.A., et al., 2021. Machine learning to detect marine animals in UAV imagery: effect of morphology, spacing, behaviour and habitat. Remote Sens. Ecol. Conserv. 7 (3), 341–354.

Duporge, I., Isupova, O., Reece, S., Macdonald, D.W., Wang, T., 2021. Using very-high-resolution satellite imagery and deep learning to detect and count African elephants in heterogeneous landscapes. Remote Sens. Ecol. Conserv. 7 (3), 369–381.

Eikelboom, J.A.J., 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 Ecol. Evol. 10 (11), 1875–1887.

García, R., Aguilar, J., Toro, M., Pinto, A., Rodríguez, P., 2020. A systematic literature review on the use of machine learning in precision livestock farming. Comput. Electron. Agric. 179 (September), 105826.

GitHub - georgegach/flowiz: Converts Optical Flow files to images and optionally compiles them to a video. Flow viewer GUI is also available. Check out mockup right from Github Pages: [Internet]. [cited 2021 Dec 31]. Available from: https://github.com/georgegach/flowiz.

He, K., Gkioxari, G., Dollar, P., Girshick, R. Mask R-CNN. 2017. p. 2961–9.

Hermans, A., Beyer, L., Leibe, B. In Defense of the Triplet Loss for Person Re-Identification. 2017 Mar 22.

ICCV 2017 Open Access Repository [Internet]. [cited 2022 Feb 20]. Available from: https://openaccess.thecvf.com/content_iccv_2017/html/Neuhold_The_Mapillary_Vistas_ICCV_2017_paper.html.

Israel, M. A UAV-based roe deer fawn detection system. 2011 Nov 1.

Kassim, Y.M., Byrne, M.E., Burch, C., Mote, K., Hardin, J., Larsen, D.R., et al. Small object bird detection in infrared drone videos using mask R-CNN deep learning. IS and T International Symposium on Electronic Imaging Science and Technology. 2020 Jan 26;2020(8).

Le, H., Samaras, D., Lynch, H.J., 2021. A convolutional neural network architecture designed for the automated survey of seabird colonies. Remote Sens. Ecol. Conserv.

Lethbridge, M., Stead, M., Wells, C., Lethbridge, M., Stead, M., Wells, C., 2019. Estimating kangaroo density by aerial survey: a comparison of thermal cameras with human observers. Wildl. Res. 46 (8), 639–648.

Lhoest, S., Linchant, J., Quevauvillers, S., Vermeulen, C., Lejeune, P., 2015. HOW MANY HIPPOS (HOMHIP): Algorithm for automatic counts of animals with infra-red thermal imagery from UAV. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. Microsoft COCO: Common Objects in Context. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2014;8693 LNCS(PART 5):740–55.

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

Longmore, S.N., Collins, R.P., Pfeifer, S., Fox, S.E., Mulero-Pázmány, M., Bezombes, F., et al. Adapting astronomical source detection software to help detect animals in thermal images obtained by unmanned aerial systems. https://doi.org/101080/0143116120171280639. 2017 May 19;38(8–10):2623–38.

Mahadevan, S., Athar, A., Sa, A., Sep, O., Hennen, S., Leal-Taixé, L., et al. Making a Case for 3D Convolutions for Object Segmentation in Videos. 2020 Aug 26.

Mahmud, M.S., Zahid, A., Das, A.K., Muzammil, M., Khan, M.U., 2021. A systematic literature review on deep learning applications for precision cattle farming. Comput. Electron. Agric. 187 (July), 106313.

Nabavi, S.S., Rochan, M., Wang, Y. Future Semantic Segmentation with Convolutional LSTM. British Machine Vision Conference 2018, BMVC 2018. 2018 Jul 20.

openvinotoolkit/cvat: Powerful and efficient Computer Vision Annotation Tool (CVAT) [Internet]. [cited 2022 Feb 2]. Available from: https://github.com/openvinotoolkit/cvat.

Radiuk, P., Radiuk, P.M., 2017. Impact of Training Set Batch Size on the Performance of Convolutional Neural Networks for Diverse Datasets. 20, 20–24.

Rivas, A., Chamoso, P., González-Briones, A., Corchado J.M. Detection of cattle using drones and convolutional neural networks. Sensors 2018, Vol 18, Page 2048. 2018 Jun 27;18(7):2048.

Shao, W., Kawakami, R., Yoshihashi, R., You, S., Kawase, H., Naemura, T. Cattle detection and counting in UAV images based on convolutional neural networks. https://doi.org/101080/0143116120191624858. 2019 Jan 2;41(1):31–52.

Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C., et al. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting.

Smith, S.L., Kindermans P.J., Ying, C., Le, Q.V. Don't Decay the Learning Rate, Increase the Batch Size. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings. 2017 Nov 1.

Sun, D., Yang, X., Liu, M.Y., Kautz, J., 2020. Models matter, so does training: an empirical study of cnns for optical flow estimation. IEEE Trans. Pattern Anal. Mach. Intell. 42 (6), 1408–1423.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.A. Closer Look at Spatiotemporal Convolutions for Action Recognition. 2018. p. 6450–9.

Van Nuffel, A., Zwertvaegher, I., Van Weyenberg, S., Pastell, M., Thorup, V.M., Bahr, C., et al. Lameness detection in dairy cows: Part 2. use of sensors to automatically register changes in locomotion or behavior. Animals 2015, Vol 5, Pages 861-885. 2015 Aug 28;5(3):861–85.

Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Seka,r B.B.G., Geiger, A., et al. MOTS: Multi-object tracking and segmentation. 2019. p. 7942–51.

Witczuk, J., Pagacz, S., Zmarz, A., Cypel, M., 2018. Exploring the feasibility of unmanned aerial vehicles and thermal imaging for ungulate surveys in forests - preliminary results. Int. J. Remote Sens. 39 (15–16), 5504–5521.

Xu, B., Wang, W., Falzon, G., Kwan, P., Guo, L., Sun, Z, et al. Livestock classification and counting in quadcopter aerial images using Mask R-CNN. https://doi.org/101080/0143116120201734245. 2020 Nov 1;41(21):8121–42.

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. Comput. Electron. Agric. 1 (171), 105300.

Zhang, L., Zhu, G., Shen, P., Song, J., Afaq Shah, S., Bennamoun, M. Learning Spatiotemporal Features using 3DCNN and Convolutional LSTM for Gesture Recognition.

Zhu, G., Zhang, L., Shen, P., Song, J., 2017. Multimodal gesture recognition using 3-D convolution and convolutional LSTM. IEEE Access 5, 4517–4524.