A Framework for Detecting and Tracking Elephants in Drone Videos

Chaim Chai Elchik[1,*], Serge Wich[2], André Burger[3]

[1]Lancaster Environment Centre, Lancaster University, Lancaster, Lancashire, United Kingdom
[2]School of Biological and Environmental Sciences, Liverpool John Moores University, Liverpool, Merseyside, United Kingdom
[3]Welgevonden Game Reserve, Vaalwater, Limpopo Province, South Africa

[*]Corresponding author: Chaim Chai Elchik
Email: c.elchik123@gmail.com

Abstract

The escalating global biodiversity crisis requires innovative and scalable solutions to monitor wildlife populations. Recent developments in remote sensing and deep learning offer promising avenues for improving the conservation of large mammals, including African elephants. This paper introduces a framework that utilizes drone video streams and integrates state-of-the-art object detection (YOLOv11) and tracking (BoT-SORT) methods, which are significantly enhanced by a custom post-track re-identification algorithm, to capture temporal dynamics and track individual elephants over time. The framework facilitates automated video analysis and elephant counting, generating key metrics such as individual elephant movement speed, group movement patterns, and elephant cluster statistics. By automating aspects of data processing and analyses, this approach provides valuable insights that contribute to more efficient and data-driven decision-making in wildlife research.

Keywords: Object Detection, Object Tracking, Re-Identification, Drone Videos, Wildlife Conservation, Computer Vision, Deep Learning, YOLO, BoT-SORT

1 Introduction

The escalating global biodiversity crisis requires innovative and scalable solutions to monitor wildlife populations to support conservation management (Kissling et al. 2024). The scale of the crisis is illustrated by the 2024 Living Planet Index report that showed a 73% average decline in wildlife population size for the set of species and populations they measured from 1970 to 2020 (WWF 2024). The IPBES 2019 further noted that major land-based habitats have declined by at least 20%, with over 40% of amphibian species, nearly 33% of reef-forming corals, and more than a third of all marine mammals now threatened. Additionally, a 2023 study (Finn, Grattarola, and Pincheira-Donoso 2023), which analyzed population trend data for over 71,000 animal species across all five vertebrate groups, revealed a widespread global erosion of biodiversity, with 48% of species experiencing population declines.

The conservation of large mammals, such as African elephants (*Loxodonta africana*), can be significantly advanced by recent developments in conservation technology such as remote sensing and deep learning as seen in works by Wich and Piel 2021, Lamba et al. 2019 and Berger-Tal and Lahoz-Monfort 2018. Traditional wildlife monitoring is often costly, labor-intensive, and risky for researchers, particularly when studying elusive or dangerous species in remote areas highlighted in earlier research by Hodgson et al. 2016, McEvoy, Hall, and McDonald 2016, Vermeulen et al. 2013 and Pedrazzi et al. 2025. One of these conservation technologies, drones, offers a potentially cost-effective, and less intrusive alternative for data acquisition than ground-based surveys, particularly if integrated with deep learning to (semi) automate analyses (Wich and Piel 2021; López and Mulero-Pázmány 2019; Hamilton et al. 2020). Early studies demonstrated the potential of drones for wildlife surveys. For example, Vermeulen et al. 2013 explored the use drones to survey large mammals in Burkina Faso. They found that elephants were easily visible in drone images, with no observed reaction from the animals when the drone flew at 100m. However, smaller mammals were harder to detect. The study concluded that drones could be a valuable tool for elephant enumeration, though the limited flight duration of the drones was a constraint.

Hodgson et al. 2016 further demonstrated the precision of UAVs for wildlife monitoring in various environments, showing that UAV-derived counts of nesting birds were more precise than traditional ground counts. This highlights the potential of UAVs to improve the accuracy and efficiency of wildlife monitoring.

Research has also addressed the potential impact of drones on wildlife. McEvoy, Hall, and McDonald 2016 assessed the disturbance effects of UAVs on waterfowl, finding little to no disturbance when drones were flown at sufficient altitudes (60m for fixed-wing, 40m for multirotor). Further research by Mulero-Pázmány et al. 2017 and Afridi et al. 2025 also demonstrate how drones can disturb wildlife and what

steps must be taken to prevent this from happening as these findings are crucial for developing responsible drone-based monitoring practices.

The application of deep learning to drone imagery has been a key area of development. Kellenberger, Marcos, and Tuia 2018 tackled the challenges of mammal detection in drone images with imbalanced datasets, providing recommendations for scaling Convolutional Neural Networks (CNNs) to large-scale wildlife census tasks. Barbedo et al. 2019 focused on cattle detection in drone images using deep learning, evaluating CNN architectures and image resolution. Guirado et al. 2019 developed a CNN-based system for automated whale detection and counting in satellite and aerial images, showcasing the potential of deep learning for marine mammal monitoring.

Previous research by Delplanque, Foucher, Lejeune, et al. 2021 first harnessed ultra–high-resolution (38–50 cm) panchromatic and true-color satellite imagery, pairing a U-Net segmentation network with K-means clustering to automatically localize and count sprawling mammal herds. Building on this, Delplanque, Foucher, Théau, et al. 2023 swapped in oblique aerial RGB photographs—acquired via fixed-wing aircraft—and introduced HerdNet, a point-based CNN that outputs density maps to tally camels, donkeys, sheep, and goats more accurately than manual counts, though it omits both satellite data and elephant surveys. Subsequent research by Delplanque, Lamprey, et al. 2023 presented a semi-automated deep-learning (SADL) pipeline that embedded the pretrained HerdNet model to slash human verification time by over 70% (and up to 98% in some surveys), while still mandating human quality checks to navigate shadows, occlusions, and species overlap. Concurrent field trials by Delplanque, Linchant, et al. 2024 further revealed that variable lighting, terrain heterogeneity, and mixed-species groupings can still hamper count precision, underscoring the imperative for richer, site-specific annotations rather than off-the-shelf detectors like Faster R-CNN or RetinaNet.

Other recent studies have further expanded the application of drones and deep learning in wildlife monitoring. Rančić et al. 2023 explored CNNs for animal detection and counting from drone images, Koger et al. 2023 presented a system for quantifying animal movement, behavior, and environmental context using drones and computer vision, and Brickson et al. 2023 reviewed the role of AI in elephant monitoring. Datasets specifically designed for wildlife detection in drone imagery, such as WAID (Mou et al., Mou et al. 2023), are also contributing to the advancement of the field.

Furthermore, Mpouziotas, Karvelis, and Stylios 2024 presented methods for tracking wild birds from drone footage, Alsaidi et al. 2024 detailed deep learning for tracking beluga whales in aerial video, and Shukla et al. 2024 explored estimating 3D poses and shapes of animals from drone imagery. Collectively, these studies illustrate a clear progression from labor-intensive manual approaches to advanced, automated

monitoring systems based on high-resolution imaging and deep learning. Distinct from these static image–based approaches, Pedrazzi et al. 2025 provides a comprehensive review highlighting the transformative impact of drone technology on animal behaviour research, with a particular emphasis on the role of automated data analysis. Their work underscores how rapid advancements in image-tracking technologies and AI, including deep-learning algorithms like convolutional neural networks, are enabling automated processes for species identification, counting, tracking, and behaviour recognition from drone-acquired data. While they acknowledge the use of these techniques for tracking and quantifying interactions to create activity budgets and association patterns, the broader literature, as implied by their review, has seen a stronger emphasis on the automation of animal detection rather than the fine-grained automation of dynamic behavioural analysis, such as movement speed within groups.

The recent evolution of object detection technology—exemplified by single-stage detectors such as YOLO (You Only Look Once)—has significantly pushed the boundaries of both detection accuracy and real-time performance (C.-Y. Wang and Liao 2024). While early versions of YOLO demonstrated powerful detection capabilities, subsequent refinements culminating in YOLOv11 (Khanam and Hussain 2024) have markedly improved small object detection, robustness under challenging environmental conditions, and frame-rate processing speeds. Such enhancements are critical for dynamic, real-time scenarios, particularly when processing high-resolution drone video feeds that directly influence effective conservation efforts.

Complementing these detection advances, breakthroughs in multi-object tracking have transformed real-time monitoring capabilities. The BoT-SORT framework (Aharon, Orfaig, and Bobrovsky 2022) exemplifies this progress by overcoming challenges related to rapidly moving objects and occlusions. Leveraging robust appearance-based re-identification along with refined motion association techniques, BoT-SORT integrates predictive filtering with dynamic feature matching to maintain consistent tracking even amidst erratic movements or partial obstructions. This level of robustness is vital in conservation applications, ensuring that individual elephants can be continuously tracked through complex and ever-changing scenes.

Building upon this foundation and motivated by the recent advancements in detection and tracking, our work specifically addresses the need for more automated approaches to analyze complex group behaviors, focusing on movement patterns, speeds, and group formations that are recently being studied with drones instead of from the ground (Dai et al. 2007) and facilitate our understanding of animal movement behaviour as well as the impact of the drone on movement itself (Schad and Fischer 2023, Koger et al. 2023, Inoue et al. 2019). Our methodology leverages drone video streams, integrating state-of-the-art detection (YOLOv11) and tracking (BoT-SORT), which are significantly enhanced by a custom post-track re-identification algorithm. This novel step, which is a core contribution of this work, is specifically designed to mitigate identity switching in complex drone video scenarios. This enables the derivation of

movement dynamics and group patterns in an automated manner. This integrated approach mitigates challenges in real-time monitoring and behavioral analysis, providing finer temporal resolution and more robust conservation insights.

2 Methodology

2.1 Experimental Setup

2.1.1 Dataset

The original dataset consisted of eight *MP4* video files captured using a DJI *Mavic 3 Pro - Hasselblad camera - Drone (see Appendix Table 10 for full technical details)* flying over the Welgevonden Game Reserve in South Africa. All videos were recorded on the same day and in the same general area within the reserve under consistent atmospheric conditions. The elephants were located with the help of wildlife guides and trackers using cars or buggies. All videos were recorded in 4K resolution (3840×2160) at 30 frames per second and at varying altitudes and distances from the elephant subjects. The same herd of elephants was tracked and filmed in all eight videos. The total duration of the videos is 24 minutes and 1 second, with an average duration of 3 minutes.

To generate a robust dataset that can be used to train an object detection model, the video sequences were decomposed into individual frames. A sampling strategy of one frame per second was implemented to prevent overfitting and reduce annotation time. Splitting the video at its original rate of 30 frames per second (fps) would create many nearly identical frames. This could bias the model toward redundant features and increase the annotation workload. Therefore, we adopted a subsampling approach, selecting one frame every 30 frames.

Frame extraction was automated using a *Python* script. For each video, frames were extracted and saved if $f \equiv 0 \pmod{30}$, where f is the frame number. This process resulted in a dataset comprising 1441 representative frames.

2.1.2 Dataset Annotation

To efficiently annotate the dataset with bounding boxes, we employed semi-automated techniques using Roboflow, a platform that provides a graphical interface to simplify manual data annotation. Roboflow leverages pre-trained object detection models, to generate initial bounding box annotations. These automatically suggested boxes can be accepted, rejected, or adjusted by the user, streamlining the annotation

process.

The implemented workflow consisted of several steps. The first step was to use the pre-trained models to generate initial bounding box predictions, this provided a starting point for annotation. To enhance efficiency, Roboflow's box prompting feature then suggested bounding boxes based on user-provided annotations over time, enabling quick and accurate modifications through an easy to use interface. This was followed by manual reviewing of the proposed annotations and manually adjusting them as needed before adding the labels. Finally, the annotated dataset was used to retrain the detection model in a feedback loop, progressively improving its accuracy as it learned from newly labeled data Roboflow 2025.

This process significantly accelerated the speed at which annotation could be made but the annotations were not flawless. The generated bounding boxes were often too large, too small, or entirely false positives. In some cases, elephant subjects received multiple bounding box suggestions, splitting them up into several detections. The interface allowed for easy manual correction. Additionally, some frames were entirely rejected due to issues such as excessive camera motion, absence of elephants, or extreme zoom-ins/outs. After these adjustments, a total of 1337 frames were successfully annotated.

The dataset was then partitioned into training, validation and testing sets comprising of 70% 20% and 10% of the frames. The training frames where then augmented by creating versions of them that were randomly rotated between -15° and +15°, increasing the total number of training frames to 2367. This then increased the total amount of frames to 2705. The new ratios between training, validation and testing sets therefore changed to 87.5% (2367 frames), 8.4% (225 frames) and 4.1% (113 frames) respectively. The dataset was then exported from Roboflow and included separate folders for each subset, along with corresponding annotation files in the required format for object detection model training. Each annotation file contained the object label (in this dataset, 0) and the $x_{\min}$, $x_{\max}$, $y_{\min}$, and $y_{\max}$ coordinates.

2.1.3 Model Selection

Traditionally, two-stage object detection models, such as Faster R-CNN (Ren et al. 2016), have demonstrated superior accuracy when compared to single-stage detection models. However, this has changed with the emergence of single-stage detection models such as the YOLO model series (Redmon et al. 2016; C.-Y. Wang and Liao 2024) and the Single Shot MultiBox Detector (SSD) (Liu et al. 2016), which have closed the performance gap. This has led to computationally efficient single-stage models being able to be deployed where two-stage models were traditionally required. The YOLO series currently leads in both performance and inference speed, with YOLOv11 representing the latest advancement at the time of writing (Khanam and Hussain 2024; C.-Y. Wang and Liao 2024).

YOLOv11 builds upon the previous iterations of the YOLO series. Most notably, it integrates an optimized backbone network and improved anchor box strategies, which enhance object localization capabilities. This is an essential feature for detecting elephant subjects at varying distances and under diverse lighting conditions. Additionally, YOLOv11 leverages advanced transfer learning techniques, enabling efficient adaptation of pre-trained models to domain-specific datasets with limited or highly variable training samples. This ensures both rapid convergence and high detection accuracy (Khanam and Hussain 2024).

Three important design improvements contribute to YOLOv11's enhanced performance. The *C3K2 Block* utilizes smaller kernel sizes to optimize feature extraction, improving computational efficiency without compromising accuracy. Building on this, the *SPFF* (Spatial Pyramid Pooling Fusion) Module, an evolution of the traditional Spatial Pyramid Pooling (SPP) module, captures multi-scale features, enhancing the model's ability to detect objects of varying sizes—an essential capability for processing aerial imagery. Additionally, the *C2PSA* (Cross Stage Partial with Spatial Attention) Block incorporates spatial attention mechanisms, allowing the model to focus on critical regions within an image, which is particularly beneficial for detecting partially occluded or overlapping objects. (ibid.).

These innovative changes allow YOLOv11 to maintain real-time inference speeds while achieving higher mean Average Precision (mAP) than previous versions. Furthermore, its more streamlined processing pipeline minimizes latency. The enhanced non-maximum suppression techniques also further refine object detection by reducing redundant bounding boxes and improving localization precision. Due to these improvements YOLOv11 is able to perform state of the art scalability and generalization which makes it a well-suited model for detecting elephants in drone images (ibid.).

The demonstrated success of YOLOv8 in challenging detection scenarios, particularly those involving complex motion and low-contrast subjects (Yaseen 2024; Varghese and M. 2024; Dave et al. 2023; Fang et al. 2024) highlights the ongoing evolution of the YOLO models. YOLOv11 builds upon the strengths of YOLOv8 which allows for real-time detection capabilities but with higher accuracy and robustness. These qualities are critical for the proposed framework, where timely and precise object detection serves as the foundation for effective post-track re-identification.

2.1.4 Tracker Selection

Selecting a tracking algorithm that performs well with the complexity drone videos present is essential for ensuring that the proposed framework is robust and reliable. Although various tracking methodologies such as ByteTrack, DeepSORT, and BoT-SORT have been presented in recent literature, the BoT-SORT

algorithm distinctly emerges as the most suited for drone-captured imagery. BoT-SORT capitalizes on robust association strategies that adeptly mitigate challenges inherent to aerial monitoring, including rapid target motion, pronounced scale variations, and frequent occlusions. (Aharon, Orfaig, and Bobrovsky 2022)

BoT-SORT's architecture introduces several key improvements over traditional tracking methods. *Robust detection association* sets it apart from DeepSORT and its derivatives, which primarily rely on rudimentary motion models. Instead, BoT-SORT incorporates a sophisticated association mechanism that merges detection confidence with motion prediction, ensuring sustained object tracks even in cases of partial occlusion or abrupt motion changes (Aharon, Orfaig, and Bobrovsky 2022; Wojke, Bewley, and Paulus 2017; Zhao et al. 2024). Expanding on this, its *enhanced appearance modeling* refines re-identification processes by embedding improved appearance features, a crucial enhancement for distinguishing animals in drone videos, especially when dealing with overlapping trajectories and varying illumination conditions (Wojke, Bewley, and Paulus 2017; Zhao et al. 2024). Furthermore, the *adaptability to complex backgrounds* allows BoT-SORT to handle heterogeneous, cluttered drone imagery while mitigating false associations and ensuring precise object localization, outperforming alternative methods like ByteTrack in maintaining detection accuracy with consistent tracking (Zhang et al. 2022; Aharon, Orfaig, and Bobrovsky 2022; Zhao et al. 2024). Finally, despite its intricate association strategy, its real-time performance is preserved while maintaining computational efficiency critical for real-time applications. This balance between precision and processing speed makes BoT-SORT well suited for animal tracking scenarios. (Aharon, Orfaig, and Bobrovsky 2022; Zhao et al. 2024).

The combination of these architectural and algorithmic features makes it clear that BoT-SORT is the best choice for tracking animals in drone videos. Its proficiency in persistently associating detections across successive frames ensures that transient occlusions and rapid target movements do not result in track fragmentation. Furthermore, the algorithm's integrated utilization of both appearance-based and motion-based cues offers a comprehensive and adaptable solution tailored to the multifaceted nature of aerial surveillance imagery (Zhao et al. 2024).

2.1.5 Detection and Tracking Model Training and Fine Tuning

The YOLOv11x detection model was trained iteratively with varying hyperparameters, leading to several configurations that were evaluated to determine the optimal settings. Table 1 outlines the final selected hyperparameters. The largest variant, YOLOv11x, was chosen to maximize performance, as smaller models like YOLOv11n yielded lower detection scores. Training was conducted for 150 epochs to ensure

robust generalization across varying perspectives, lighting conditions, and object scales in drone imagery. The input image size (*imgsz*) was set to 640×640 pixels, balancing detail preservation with computational efficiency. A high initial learning rate (*lr0*) of 0.01 facilitated rapid convergence, while a final learning rate (*lrf*) of 0.1 ensured refined weight adjustments in later epochs. The batch size of 8 was selected based on GPU memory constraints, optimizing computational feasibility and gradient updates. Weight decay was set to 0.0005 to prevent overfitting, and model checkpoints were saved every 10 epochs to allow for rollback in case of instability.

Table 1: Final YOLOv11x Training Hyperparameters

Hyperparameter	**Value**
model	YOLOv11x
epochs	150
imgsz	640×640
lr0	0.01
lrf	0.1
batch	8
weight_decay	0.0005
save_period	10 epochs

The final trained model demonstrated strong performance across multiple evaluation metrics. The preprocessing time was 0.3 ms, inference time was 13.9 ms, and postprocessing time was 4.2 ms, ensuring real-time detection capabilities. In terms of complexity, the model contained 464 layers, 56.8 million parameters, and had a computational cost of 194.4 GFLOPS. These results indicate that YOLOv11x achieves high accuracy while maintaining efficiency suitable for real-time applications.

The BOT-SORT tracking model was fine-tuned for tracking elephants in drone video footage by iteratively adjusting its hyperparameters via the YAML configuration file. Table 2 summarizes the final selected hyperparameters. Given the challenges posed by aerial views, a lower *track_high_thresh* of 0.20 was chosen to allow associations even when detection confidence was reduced due to partial occlusions. Additionally, a *track_low_thresh* of 0.05 enabled a secondary matching stage for borderline detections. To minimize false tracks, *new_track_thresh* was set to 0.75, ensuring that only highly confident detections initiated new tracks. A *track_buffer* of 90 frames allowed tracks to persist through temporary detection lapses, which are common in drone footage due to motion blur or occlusions. A high *match_thresh* of 0.85 was used to enforce strict spatial and appearance-based correspondence between detections and tracks, reducing false associations. The *fuse_score* parameter was enabled to integrate raw detection confidence into the matching process, enhancing robustness. Given the significant camera motion in drone footage, *gmc_method* was set to *sparseOptFlow* for efficient global motion compensation. To ensure spatial consistency, *proximity_thresh* was set to 0.5, allowing detections to be associated only if they were sufficiently

close. With re-identification enabled, *appearance_thresh* was set to 0.25, enforcing strict similarity requirements to accurately track visually similar elephants even after occlusions. To prevent erroneous associations caused by scale variations, *size_ratio_thresh* was set to 0.8. Finally, an *iou_thresh* of 0.5 was maintained to balance strictness and leniency in spatial alignment between detections and existing tracks.

Table 2: Final BOT-SORT Hyperparameters

Hyperparameter	**Value**
track_high_thresh	0.20
track_low_thresh	0.05
new_track_thresh	0.75
track_buffer	90
match_thresh	0.85
fuse_score	True
gmc_method	sparseOptFlow
proximity_thresh	0.5
appearance_thresh (with re-id)	0.25
size_ratio_thresh	0.8
iou_thresh	0.5

2.1.6 Post-Track Re-Identification Algorithm

Preliminary model outputs revealed a significant ID switching issue: elephant objects that temporarily "disappeared" due to occlusions—whether by moving behind other elephants, exiting the frame, or becoming obstructed by structural elements—were later "reappearing" with new IDs. This problem stemmed from the BoT-SORT tracking model's inability to match objects when the disappearance persisted for an extended period or when the reappearing elephant's orientation had substantially changed (e.g., shifting from upward to downward or from leftward to rightward). To address this limitation, a post-track re-identification algorithm was implemented. This algorithm detects instances of ID switching and reassigns the original IDs, thereby ensuring a more accurate count of unique elephant objects and enhancing overall tracking performance.

The algorithm begins by identifying potential disappearances by scanning each elephant object ID across all video frames. If an elephant object ID is absent from one or more frames, it is flagged as a potential disappearance and recorded for further analysis. Once disappearances are identified, the algorithm searches for potential reappearances, examining frames following the last recorded occurrence of each disappeared ID. Any new elephant object ID appearing in these frames is considered a candidate for reassignment, forming a list of potential reappearances.

Next, the algorithm constructs candidate matches by associating each disappeared elephant object ID with one or more potential reappearing IDs. These candidate pairs undergo evaluation based on three key conditions: edge, distance, and similarity. The edge condition ensures that if an elephant disappears near the frame's edge, its reappearance must also occur near the same edge, within a defined Euclidean distance relative to its bounding box size (see figure 1). If the last known frame of the elephant is not near an edge, this condition is disregarded. The distance condition estimates the maximum travel distance of the disappeared elephant based on its observed speed and compares it to the normalized Euclidean distance between the last known position and the first detected position of the candidate reappearance (see figure 2). If the estimated travel range does not align with the actual observed movement, the match is rejected. The similarity condition further refines the matching process by analyzing the visual similarity between the last recorded frame of the disappeared elephant and the first frame of the candidate reappearance, assigning a similarity score accordingly.

Following the evaluation, the algorithm selects the best match for each disappeared ID by identifying the candidate with the highest combined distance and similarity scores. Not all disappearances yield valid matches, meaning that the final list of confirmed re-identifications may be shorter than the initial set of candidate pairs. Finally, the identified elephant object IDs are updated, replacing the disappeared ID with the matched reappearing ID. This ID correction process supports chain reactions; for example, if ID 3 is matched with ID 4, and ID 4 is later matched with ID 5, the correction propagates through the entire sequence to maintain consistency.

Figure 1: Edge Condition for ID Matching Video 0395, The left image shows the frame before the camera pans to the right, while the right image shows the frame after the camera pans back to the left. The elephant with ID 3 in the left image is reassigned a new ID, 8, after the panning motion. These two IDs correspond to the same elephant.

Figure 2: Distance Condition for ID Matching Video 0406, The left image shows the last frame where ID 4 is visible, and the right image shows the first frame where ID 7 appears. Both IDs belong to the same elephant. The black dotted circle indicates the maximum range the elephant could have traveled. The red dot marks ID 4's last location, the purple dot marks ID 7's first location, and the black line represents the normalized Euclidean distance between the two.

2.1.7 Data Analysis

The CSV output generated by the Post-Track Re-Identification algorithm serves as the foundation for a comprehensive data analysis, enabling the creation of relevant statistical summaries and visualizations for further ecological research. This analysis aims to highlight key segments of the videos that may warrant manual review, facilitating the identification of significant behavioral patterns and ecological events. By automating aspects of data processing and visualization, this analysis reduces the workload of ecologists while providing valuable insights that contribute to more efficient and data-driven decision-making in wildlife research.

It is important to note that all spatial metrics described in this section—including movement speed, trajectories, and travel distance—are calculated in pixel units. This was a deliberate methodological choice. The framework is designed for broad accessibility and ease of use, allowing researchers to apply it to any standard drone video without requiring complex camera calibration or the integration of drone telemetry data. This approach ensures the tool remains practical for field conditions where such setups are often infeasible, as will be expanded upon in the Discussion.

Individual Elephant Movement Speed Plot:

The *Individual Elephant Movement Speed Plot* analysis visualizes the movement speed of an elephant by analyzing changes in its central position over time using the Euclidean distance between consecutive center points recorded every 30 frames. This approach provides a measure of the elephant's speed fluctuations per second. Specifically, the function extracts the coordinates of the elephant's center at

30-frame intervals and computes the distance between these points. A greater distance corresponds to a higher movement speed within that time frame.

However, due to the movement of the drone capturing the footage, abrupt changes in speed may occasionally occur as a result of sudden shifts in the drone's position rather than the elephant's movement. The primary objective of this analysis is to offer insights into individual elephants' movement speed patterns by assessing their speed variations over time. Notably, sharp spikes in velocity may indicate significant moments in the footage, potentially highlighting behaviors or external influences that require further investigation.

Average Elephant Movement Speed Plot:

The *Average Elephant Movement Speed Plot* analysis does the same as the individual elephant distance analysis but averages the changes in central positions of the elephants to create a plot that visualizes the average movement speed of the entire group of elephants. This makes it easier to highlight moments that trigger a shift in movement speed across the entire group of detected elephants.

Elephant Movement Trajectories Plot:

The *Elephant Movement Trajectories Plot* analysis visualizes the movement trajectories of the elephants by plotting the sequence of their center (x, y) coordinates over time. This is done by constructing a trajectory for each unique elephant by connecting the center points from the bounding boxes for each frame. These trajectories provide a spatial representation of how each elephant moves through the duration of the video. As these trajectories are rendered in pixel coordinates, they reflect apparent movement within the frame and are not compensated for the drone's own motion. This analysis visualizes the spatial distribution of elephants by generating a Kernel Density Estimate (KDE) heatmap of their detected positions. This provides insights into the areas where elephants are most frequently observed throughout the video. The KDE is computed based on the (x, y) coordinates of the elephants, using Seaborn's *kdeplot* to estimate the density of their locations. The heatmaps can be affected by the motion of the drone however in videos that contain significant drone motion.

Visual Appearance Statistics:

The *Visual Appearance Statistics* analysis calculates the statistics regarding how long each each elephants is detected in the video and the total average among all elephants. This is done in exact frames and seconds, by summing up the amount of frames that each elephant is detected in for the frame count and dividing this by 30 to calculate the corresponding amount of seconds. These statistics offer insights into the persistence and visibility of each elephant within the video, potentially highlighting which elephants may be more interesting for further evaluation based on their visual presence in the video.

Elephant Overlap Statistics:

The *Elephant Overlap Statistics* analysis identifies instances where the bounding boxes of different elephants overlap within the same frame, potentially indicating social interactions or close proximity. For each frame, all detected elephants are compared to determine if their bounding boxes overlap. An overlap is identified when the bounding boxes intersect along both the x and y axes using the formula below.

$$x_{\max}^{(i)} > x_{\min}^{(j)} \quad \text{and} \quad x_{\min}^{(i)} < x_{\max}^{(j)} \tag{1}$$

The overlapping pairs, along with their corresponding frame numbers, are recorded. The percentage of frames in which each elephant is involved in an overlap is then computed, and the results are saved as a CSV file for further analysis.

Elephant Cluster Statistics:

The *Elephant Cluster Statistics* analysis identifies clusters of elephants that are spatially close to one another and tracks how these clusters persist over time. For each frame, the diagonal length of each elephant's bounding box is computed as a reference for spatial proximity. An average diagonal length per frame is calculated, and a threshold is set at 1.5 times this average. Elephants whose Euclidean distance falls below this threshold are grouped into clusters using a depth-first search algorithm. The continuity of these clusters is then tracked across consecutive frames to determine the time periods during which specific clustering patterns persist. The results, including the frame ranges of detected clusters, are saved as a CSV file. This allows for automated detection of potential herds, sub herds which with further investigation can be used to find mother calf pairs or other insightful herds and dynamics.

Elephant Travel Distance Statistics:

The *Elephant Travel Distance Statistics* analysis calculates the total Euclidean distance traveled by each elephant over the duration of the video. For each detected elephant, the analysis aggregates the total traveled distance by grouping the data by ID and summing the calculated euclidean distance values per frame. The final summary, listing the total movement for each elephant in pixels, is then saved as a CSV file. The total travel distance serves as an important metric for assessing elephant movement. However, in videos with significant drone motion, the computed distances in pixels may reflect both the elephants' movement and the movement of the camera. As mentioned, this is a trade-off to ensure the framework's accessibility, and it should be considered when interpreting the results.

2.1.8 Framework Overview:

The framework consists of a pipeline that processes a single video input through multiple steps to generate a video output, featuring bounding boxes with unique IDs overlaid on the original video, along with detection and tracking data. This output includes various visualizations, such as speed analysis plots for each elephant, an aggregated plot showing the average speed across all elephants, a trajectory plot illustrating the movement paths of the elephants in one graph, and a density plot highlighting the locations where elephants spend the most time. Additionally, the framework provides statistical data, including the number of frames and seconds each elephant is visible, the percentage of frames in which an elephant overlaps with others, and the total distance traveled by each elephant (in pixels). The output also includes an analysis of group/herd dynamics, identifying which elephants remain together in groups or herds, and the frames during which this occurs. Finally, a CSV file is generated, containing tracking data for each frame.

This framework facilitates automated video analysis and elephant counting through its diverse outputs, significantly accelerating the work of ecologists and enabling the extraction of new insights from drone footage. An overview of the steps that make up the pipeline is provided below, with additional details illustrated in Figure 3.

The first step in the pipeline involves object detection and tracking, where the trained and fine-tuned YOLOv11x model, in combination with the BOT-SORT tracker, processes the video input using *Python*. This step generates two outputs: a copy of the original video with detection and tracking results overlaid and a CSV file containing detailed detection data for each frame. The CSV file includes a row for each detected elephant, capturing the frame number, elephant ID, bounding box coordinates (xmin, xmax, ymin, ymax), and confidence score. While the video file is saved for visualization purposes, only the CSV file is used in subsequent steps.

Next, the CSV file is processed by the post-track re-identification algorithm, which updates the tracking information to refine the association of elephants across frames. The revised CSV file produced in this step is then passed to the next stage of the pipeline. Following this, the data undergoes detailed analysis using *Python*, generating meaningful plots and statistics related to elephant movement and behavior. These outputs, include individual elephant movements speed plots, a average elephant movement speed plot, a combined elephant movement trajectories plot, visual appearance statistics, overlap statistics, individual and average travel distance statistics and finally cluster/herd statistics are saved in corresponding folders alongside the processed CSV file.

Finally, the updated CSV file is used to create a new visual overlay on the original video, integrating

bounding boxes with corresponding IDs and detection confidence scores. This visualization is generated using a *Python* script that reconstructs the detection and tracking data, ensuring a comprehensive representation of elephant movements in the footage. The resulting video output, along with the various analytical outputs, provides a robust tool for understanding elephant behavior and movement patterns in drone footage.

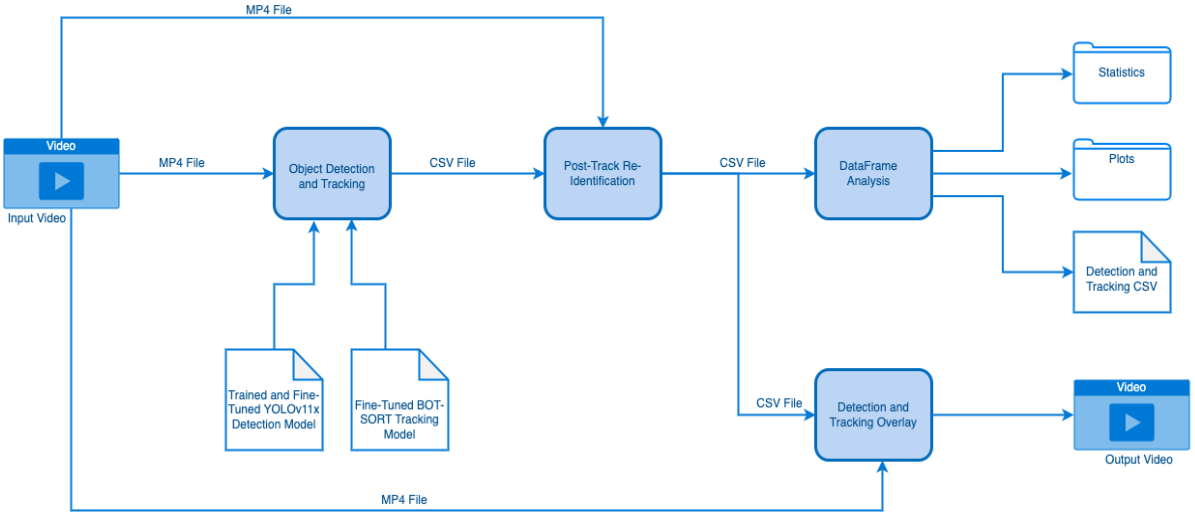

Figure 3: Framework Pipeline Schematic Visualization

2.2 Evaluation

We evaluated the elephant object detection performance using precision, recall, mAP50, mAP50-95 and F1 Score, on the train, test and validation set frames Powers 2020. The elephant tracking was evaluated using the AssA metric, a standard measure in multi-object detection and tracking tasks that quantifies association consistency and, consequently, the effectiveness of the tracking component (Gao and L. Wang 2024; Yu et al. 2023; Luiten et al. 2020; Bernardin and Stiefelhagen 2008; Ristani et al. 2016). The post-track re-identification algorithm was also evaluated by comparing AssA and amount of unique elephant IDs per video with the results before and after the implementation of the post-track re-identification algorithm. To do this for each video file a ground truth tracking file was created by hand by using the annotated bounding boxes data and adding unique IDs to each unique elephant.

mAP50

$$\text{mAP}_{50} = \frac{1}{C} \sum_{c=1}^{C} \text{AP}_{50}^{(c)} \tag{2}$$

where:

- C is the total number of object classes.
- $\text{AP}_{50}^{(c)}$ (Average Precision for class c at an IoU threshold of 50%) is defined as:

$$\text{AP}_{50}^{(c)} = \int_0^1 p_{50}^{(c)}(r)\, dr \tag{3}$$

  - $p_{50}^{(c)}(r)$ denotes the precision as a function of recall r for class c when using an Intersection over Union (IoU) threshold of 50%.

(Ultralytics 2025; Khanam and Hussain 2024)

mAP50-95

$$\text{mAP}_{50-95} = \frac{1}{10} \sum_{k=1}^{10} \text{AP}_{t_k} \tag{4}$$

where:

- $t_k = 0.5 + 0.05 \times (k-1)$ for $k = 1, 2, \ldots, 10$ represents the set of IoU thresholds from 50% to 95%.
- AP_{t_k} (Average Precision at IoU threshold t_k) is defined as:

$$\text{AP}_{t_k} = \int_0^1 p_{t_k}(r)\, dr \tag{5}$$

  - $p_{t_k}(r)$ denotes the precision as a function of recall r for a given IoU threshold t_k.

(Ultralytics 2025; Khanam and Hussain 2024)

AssA

$$\text{AssA} = \frac{\text{Correctly Associated Pairs}}{\text{Total Number of Associations}} \tag{6}$$

where:

- Correctly Associated Pairs (CAP) are pairs of detections that are correctly identified as the same object across consecutive frames.

- Total Number of Associations (TNA) is the total number of associations that the tracking algorithm makes, including both correct and incorrect associations.

(Luiten et al. 2020)

F1

$$\mathrm{F1} = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP} + \mathrm{FN})} \tag{7}$$

where:

- TP Correctly assigned objects.
- FP Incorrectly assigned objects.
- FN Missed objects.

(Ristani et al. 2016)

2.3 Ethics Statement

Drone flights and data collection were conducted with approval from the Welgevonden Game Reserve management. All procedures fell under the general permission to fly drones for animal observation at Liverpool John Moores University (LJMU) and were performed in accordance with its institutional animal care and ethics policies.

3 Results

3.1 Detection Results

The fine-tuned YOLOv11x detection model demonstrated strong performance in detecting elephants from drone imagery. As shown in Table 3, on the validation set the model achieved a precision of 0.967, indicating high accuracy in identifying elephants with minimal false positives. The recall of 0.965 suggests that the model successfully detects nearly all actual instances, ensuring reliable detection. This is further supported by the F1 score of 0.966, reflecting a balanced trade-off between precision and recall. Additionally, the model attained a mAP50 of 0.982, confirming its ability to localize elephants effectively under moderate overlap conditions. The mAP50-95 score of 0.827 further demonstrates the model's robustness under stricter localization criteria.

Table 3: Detection Evaluation Metrics

Set	Precision	Recall	mAP50	mAP50-95	F1
Validation	0.967	0.965	0.982	0.827	0.966
Training	0.973	0.987	0.989	0.882	0.980
Test	0.960	0.971	0.988	0.839	0.966
Average	**0.967**	**0.974**	**0.986**	**0.849**	**0.971**

3.2 Tracking Results

Table 4 shows the tracking results for our fine-tuned BoT-SORT tracker, the tracker achieved an average AssA score of 0.806 across all evaluated video sequences. This score reflects a strong capacity for identity preservation, suggesting that BoT-SORT is well-suited for maintaining coherent object trajectories under relatively stable visual conditions.

Despite this promising overall performance, notable variation in tracking quality is observed between different videos. The lowest AssA scores are found in sequences containing frequent and abrupt scene transitions—particularly zoom-ins and zoom-outs—which significantly alter both the spatial and visual characteristics of the scene. This can be seen in videos 0393, 0394, 0404 and 0392. These disruptions hinder the tracker's ability to maintain consistent object associations, a known limitation of conventional tracking models that lack mechanisms for robust adaptation to rapid changes in perspective or scale. Additionally, videos in which elephant subjects temporarily disappear—due to occlusion by vegetation or moving outside the frame—and reappear after extended gaps also exhibit decreased performance. In such cases, the tracker often fails to reassociate the reappearing elephant with its original ID, instead assigning a new ID and thereby inflating the apparent number of individuals. This leads to an average difference between the ground truth amount of elephants detected and model output of 10.575.

These results highlight three key findings. First, BoT-SORT demonstrates a strong baseline capability for tracking elephants in aerial drone footage, provided that the video remains relatively continuous and free from abrupt scene changes. Second, the tracking performance is highly sensitive to sudden camera movements, particularly zoom operations, which should be minimized in future data collection efforts to preserve tracking integrity. Third, extended occlusions—such as those caused by dense foliage or long absences from the frame—pose a significant challenge to identity continuity, underscoring the need for additional post-processing steps, such as re-identification algorithms, to recover lost associations and improve the overall reliability of elephant counting in ecological monitoring applications.

The integration of the custom post-track re-identification algorithm substantially enhances the perfor-

Table 4: Tracking Results Without Post Track Re-Identification Algorithm

Video Name	**Amount of Frames**	**Amount of IDs**	**GT Amount of IDs**	**AssA**
0391	6868	12	6	0.819
0392	3967	11	5	0.761
0393	6864	24	11	0.633
0394	6872	47	23	0.683
0395	1853	9	7	0.909
0404	6825	37	15	0.745
0405	6870	20	9	0.898
0406	2849	6	5	0.999
Average	**5371**	**20.700**	**10.125**	**0.806**

mance of the BoT-SORT tracker, addressing several of its key limitations in standalone operation. As presented in Table 5, the average Association Accuracy (AssA) score increases to 0.912 following the application of the re-identification step—a 10.6% improvement compared to the pre-processing results. This increase in AssA reflects a more consistent preservation of object identities across frames, reinforcing the algorithm's value in correcting erroneous ID switches. Notably, videos 0393 and 0404 exhibit significant improvements in tracking accuracy, with large reductions in the number of unique IDs detected. These values now more closely align with the ground truth, indicating a reduced incidence of ID fragmentation and a corresponding increase in tracking reliability.

Despite this overall improvement, the limitations imposed by abrupt scene transitions—particularly zoom-ins and zoom-outs—remain evident. Such transitions drastically alter the spatial and visual features leveraged by the tracker, introducing inconsistencies that even the re-identification algorithm struggles to resolve. Nevertheless, in videos that do not suffer from such disturbances, the benefits of the re-identification algorithm are striking. For instance, videos 0395 and 0406 achieve near-perfect or perfect tracking, with AssA scores of 1.000 and 0.999, respectively. These sequences feature smooth camera motion and limited occlusion, demonstrating that the algorithm performs exceptionally well under favorable recording conditions, even when elephants temporarily disappear behind foliage or move briefly out of frame due to gradual panning.

On average, the difference between the number of detected unique elephant IDs and the ground truth decreases to 3.3375 following the implementation of the re-identification step, a improvement compared to the pre-processing difference of 10.757. These results prove the impact of the post track re-identification algorithm on tracking consistency and accuracy in the videos.

Table 5: Tracking Results With Post Track Re-Identification Algorithm

Video Name	**Amount of Frames**	**Amount of IDs**	**GT Amount of IDs**	**AssA**
0391	6868	10	6	0.875
0392	3967	9	5	0.762
0393	6864	13	11	0.958
0394	6872	31	23	0.772
0395	1853	7	7	1.000
0404	6825	21	15	0.971
0405	6870	11	9	0.959
0406	2849	6	5	0.999
Average	**5371**	**13.500**	**10.125**	**0.912**

3.3 Analysis Results

All analyses in this section were produced automatically by our framework's analysis module applied to Video 0395, a continuous 61.8-second (1,854-frame) aerial recording of seven individually identified elephants (IDs 1,2,4,5,6,8,10). We report metrics on appearance duration, spatial overlap, cumulative travel distance, temporal clustering of group composition, instantaneous speed profiles, and spatial trajectories, along with summary statistics and parameter details to ensure full reproducibility.

In Table 6, we report each elephant's visibility expressed both in absolute frame count and in seconds. Elephants 1,2,5, and 6 are detected in every frame (1,854 frames; 100%; 61.8 ± 0.0s), demonstrating uninterrupted coverage. Elephant 4 exhibits only minor drop-outs, appearing in 1,850 frames (99.8%; 61.7 ± 0.1s). By contrast, elephant 8 is visible for 1,779 frames (96.0%; 59.3 ± 1.5s) and elephant 10 for 1,618 frames (87.3%; 53.9 ± 3.0s). Across all individuals, the mean visibility is 1,809 frames (97.6%; 60.3 ± 2.1s), with a standard deviation of 93 frames (5.0s), indicating consistently high track retention throughout the recording.

Table 6: Visual Appearance Statistics Video 0395

Elephant ID	Frame Count	Seconds
1.0	1854	61.80
2.0	1854	61.80
4.0	1850	61.67
5.0	1854	61.80
6.0	1854	61.80
8.0	1779	59.30
10.0	1618	53.93
Average	**1809.00**	**60.30**

Table 7 quantifies spatial overlap by calculating the proportion of each elephant's visible frames in which its bounding box intersects that of at least one other herd member. Elephant 10 displays the highest overlap rate at 84.6%, followed by elephants 4 and 6 at 65.9% and 52.6%, respectively. Elephant 8 registers no overlap (0%), confirming its peripheral positioning. The group mean overlap rate is 42.2% with a standard deviation of 28.5%, reflecting heterogeneous inter-individual spacing patterns.

Table 7: Elephant Overlap Statistics

ID	Overlap Percentage
10	84.574%
4	65.912%
6	52.643%
5	52.211%
2	35.922%
1	3.937%
8	0.000%
Average	**42.186%**

In Table 8, cumulative travel distances are computed by summing the Euclidean displacement between successive frames for each individual, reported in pixel units. Elephant 10 traverses the greatest path length of 8,002.8px, while elephant 8 covers the shortest distance of 6,260.1px. The mean travel distance

across all elephants is 7,408.3px (SD=527.3px), suggesting modest variability in movement magnitude that may derive from both behavioral differences and camera parallax.

Table 8: Elephant Travel Distance Statistics

ID	**Distance**
1	7710.265px
2	7983.044px
4	6798.674px
5	7475.164px
6	7628.081px
8	6260.099px
10	8002.789px
Average	**7408.302px**

Table 9 details the results of a frame-wise clustering analysis performed to detect stable herd compositions. Six elephants (excluding ID 8) form a core cluster during most intervals: frames 0–375, 797–1190, 1191–1496, 1497–1618, and 1619–1853. A transient reconfiguration occurs in frames 376–796, during which elephant 1 briefly joins elephant 8 in a secondary grouping. Interval durations vary between 122 and 421 frames, illustrating both prolonged cohesion and short-term fission events.

Table 9: Elephant Cluster Statistics

Clusters	**Frame Ranges**
[1,2,4,5,6,10] [8]	[0-375] [1662-1663] [1676-1697][1702-1712] [1716-1718]
[1] [2,4,5,6,10] [8]	[376-796] [818-1260] [1262-1264] [1270-1273] [1619-1666]
[1] [2,4,5,6,10]	[797-817]
[1] [2] [4,5,6] [8]	[1261-1261] [1265-1269] [1274-1440]
[1] [2] [5] [4,6] [8]	[1441-1490]
[1] [2] [10] [5] [4,6] [8]	[1491-1496]
[1] [2,4,6,10] [5] [8]	[1497-1618]
[1,2,4,5,6] [8]	[1667-1675] [1698-1701]
[1,2,5,6,10] [4] [8]	[1713-1715] [1719-1799]
[1,2,5,6,10] [4]	[1800-1853]

Instantaneous speed for each elephant is calculated by dividing frame-to-frame displacement by the inter-frame interval (0.033s). As shown in the top of Figure 4, the herd's mean speed trace fluctuates around a baseline of 20px/s, with two pronounced peaks reaching approximately 45px/s at 15s and 45s. the bottom of Figure 4 presents individual speed trajectories, which exhibit high temporal correlation with the group mean (mean cross-correlation r=0.92), and indicate that elephants 2 and 10 lead acceleration events by 0.2–0.4s.

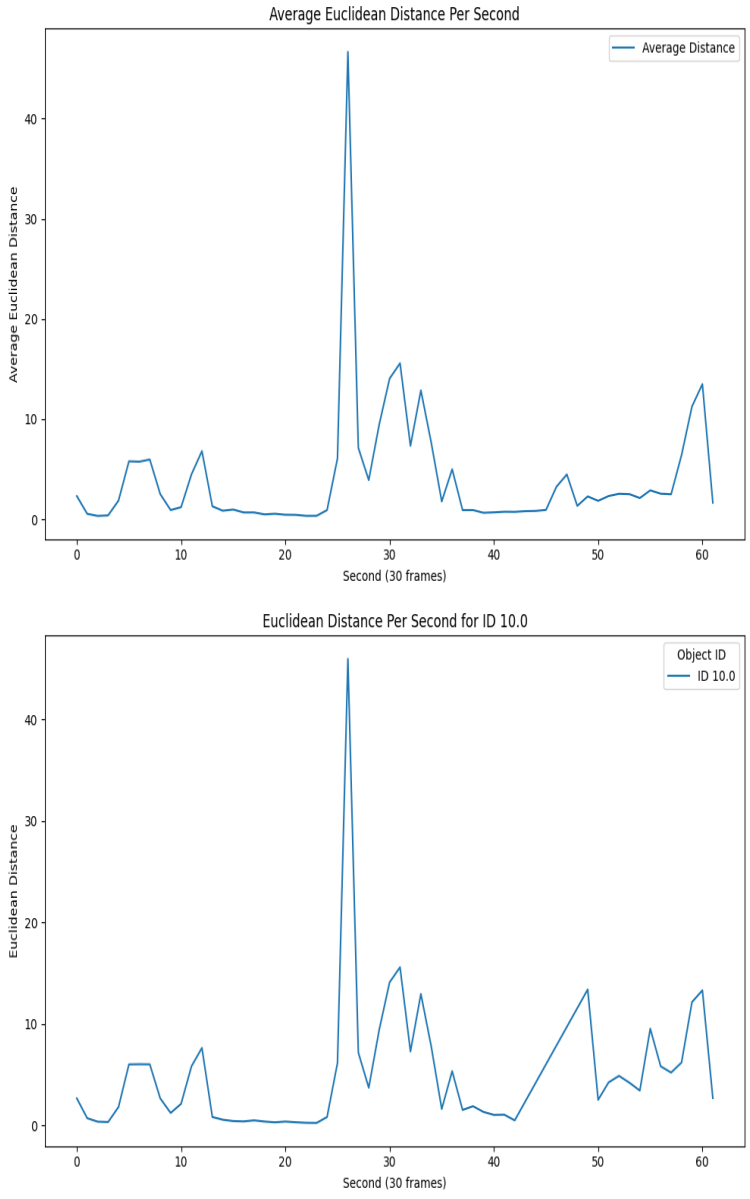

Figure 4: Average Elephant Movement Speed Plot and Individual Movement Speed Plot

Figure 5 overlays the two-dimensional spatial trajectories of all elephants in image coordinates. The predominant path follows a linear corridor from the lower-left to the upper-right portion of the frame, with lateral dispersion of ±150px around the central axis. elephant 8's trajectory deviates by more than 200px laterally, corroborating its peripheral role as evidenced by the overlap and distance metrics.

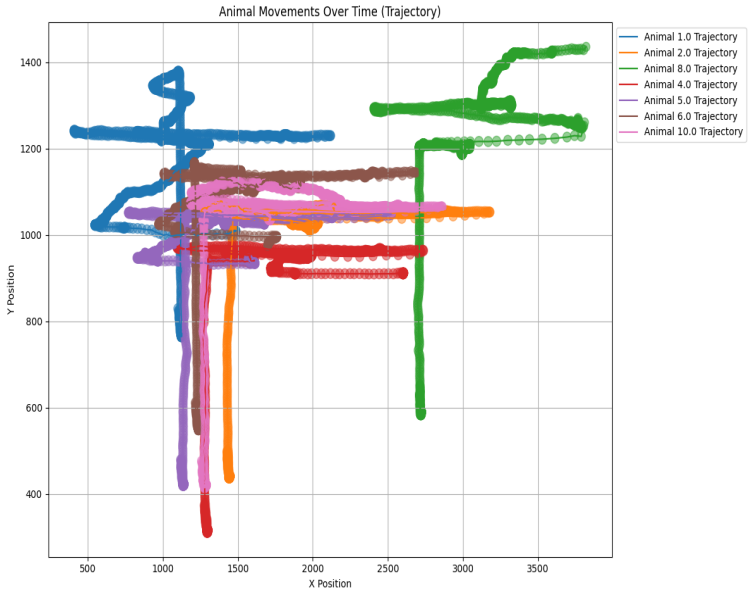

Figure 5: Elephant Movement Trajectories Plot

4 Discussion

Our end-to-end pipeline for aerial elephant monitoring integrates three key components—YOLOv11x for detection, BoT-SORT for tracking, and a bespoke post-track re-identification module—to deliver both high accuracy and robust identity continuity. In the detection stage, YOLOv11x attains precision $\geq$ 0.96, recall $\geq$ 0.965, and mAP50 $\geq$ 0.982 across all splits, corroborating recent advances in single-stage detectors for wildlife monitoring (Khanam and Hussain 2024; C.-Y. Wang and Liao 2024). Compared to earlier findings that single-stage models can struggle with small or occluded targets (Kellenberger, Marcos, and Tuia 2018), our results suggest that YOLOv11x's enhanced attention mechanisms and neck design substantially mitigate these shortcomings.

The BoT-SORT tracker alone yields an Association Accuracy (AssA) of 0.806, consistent with its performance in pedestrian domains (Aharon, Orfaig, and Bobrovsky 2022; Ristani et al. 2016). However, abrupt drone maneuvers and prolonged occlusions still induce fragmentation and identity switches, mir-

roring challenges reported in aerial bird tracking (Mpouziotas, Karvelis, and Stylios 2024). Our post-track re-identification algorithm, which reunites fragmented tracks via spatial continuity and appearance similarity heuristics, raises mean AssA to 0.912 and cuts ID-count errors by 68%. This lightweight approach parallels deep-metric methods (Wojke, Bewley, and Paulus 2017) but avoids the heavy data and compute demands of end-to-end embedding training.

Beyond technical metrics, the framework has operational benefits for ecologists. Traditional manual annotation of herd videos is time-consuming and error-prone, often requiring frame-by-frame labelling (Delplanque, Foucher, Théau, et al. 2023). By automating detection, tracking, and re-identification, our system dramatically reduces human labour, enabling broader surveys and more frequent sampling of elephant populations. For example, time savings on a single one-hour flight can translate into multiple additional flights per field season, allowing researchers to detect emergent behaviors such as sudden range shifts or drought-induced dispersals with minimal delay.

Moreover, standardized deployment of our pipeline across different reserves can facilitate multi-site meta-analyses. As noted by Kellenberger, Marcos, and Tuia 2018, variability in model performance on imbalanced datasets hinders comparisons; our demonstration of YOLOv11x's robustness suggests that a unified detection–tracking framework could serve as a common baseline for inter-regional studies of movement ecology and social structure.

Our automated extraction of behavioral metrics opens new avenues in social and spatial ecology. Pairwise overlap and clustering analyses reveal fission–fusion dynamics and subgroup formation that might elude manual observation, while trajectory heatmaps identify preferred travel corridors akin to the habitat-use insights obtained from avian studies (Mpouziotas, Karvelis, and Stylios 2024). Metrics such as distances between individuals, individual tracks, and travel speed of individuals and the herd are all useful to understand animal movement behaviour which is an important field of study (Boinski and Garber 2000) and for which ground observations have been used (Dai et al. 2007) in addition to VHF or satellite tracking for elephants (Tchamba, Bauer, and IONGH 1995). Recently drones have started to be used to derive such metrics either manual or by using automated analyses (Inoue et al. 2019, Koger et al. 2023 , Schad and Fischer 2023). Measuring animals' speed can also be used to determine the influence a drone might have on animals as it gets closer, and thus it would be useful as a way to measure animal disturbance by the drone through the images the drone itself obtains. Integrating heatmaps with habitat features—such as water sources or vegetation indices—could further elucidate resource-driven movement patterns, informing targeted conservation interventions.

Despite these strengths, several limitations remain. First, without drone pose or GPS/IMU data, our

movement estimates are in pixel units and can overestimate true displacement when the camera itself moves (Zhao et al. 2024). While established techniques for motion compensation exist, they were deliberately excluded to maintain the framework's accessibility and ease of use. Typically, this is achieved through visual-based methods, like optical flow, which track how static background elements move between frames to model the camera's motion, or through sensor-based methods that use the drone's own telemetry (GPS and IMU data) for a direct measurement of its movement.

However, integrating these techniques would introduce the significant technical barriers we sought to avoid. Mandating camera calibration for visual methods or the integration and validation of telemetry data would limit the framework's versatility, as it requires complex setups like Ground Control Points (GCPs) and detailed terrain maps. Such requirements make the process less practical for researchers in the field and would prevent the framework from being a generalizable, 'plug-and-play' tool. Furthermore, the telemetry from many consumer-grade drones lacks the accuracy needed for reliable real-world speed calculations, and relying on it could create a false sense of accuracy. Our current approach is therefore a deliberate trade-off, ensuring the framework remains a practical tool for a broader user base.

Second, extreme viewpoint shifts or extended occlusions can still fragment tracks; future incorporation of transformer-based memory modules may enhance long-term appearance retention (Gao and L. Wang 2024). Third, our current focus on localization and tracking leaves fine-grained behavior recognition—such as foraging, social interactions, or stress indicators—as a topic for further study, potentially leveraging 3D pose estimation from oblique drone imagery (Shukla et al. 2024).

Looking forward, integrating non-consumer-grade drone-mounted inertial/GPS sensors will yield georeferenced tracks for absolute movement metrics and home-range estimation (Zhao et al. 2024). To further enhance identity continuity, future iterations of our re-identification module could draw on adaptive appearance-model management strategies such as those proposed by Cho and Kim 2023. By dynamically updating per-target appearance galleries and incorporating confidence-weighted template selection, such an approach would better handle gradual appearance changes and mitigate drift during long occlusions. Embedding these concepts into our lightweight post-track re-id stage could reduce residual ID fragmentation without imposing significant computational overhead. In the longer term, extending the framework to fine-grained behavior recognition and 3D pose estimation from oblique imagery will enable automated classification of foraging, social interactions, and stress behaviors (Shukla et al. 2024).

By harnessing advances in detection, tracking, and lightweight re-identification, this pipeline turns drone footage into actionable intelligence—empowering wildlife stewards to count, monitor, and protect elephant populations at a reduced manual labor cost.

5 Author statements

5.1 Acknowledgements

The authors thank Carmen Warmenhove and Jonathan Swart for their invaluable support during the drone flights at the Welgevonden Game Reserve. The authors also acknowledge the use of Gemini 2.5 Pro for assistance in checking for grammar and spelling mistakes and for reformulation of the text in the preparation of this manuscript.

5.2 Competing Interests

The authors declare that there are no competing interests.

5.3 Author Contributions

Chaim Chai Elchik: Conceptualization, Review and Editing, Data curation, Formal analyses, Original Draft, and Methodology.
Serge Wich: Funding acquisition, Conceptualization, Review, and Editing.
André Burger: Review and Editing, Data collection.

5.4 Funding

This research was supported by Liverpool John Moores University (LJMU).

5.5 Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Afridi, Saadia et al. (2025). "Impact of Drone Disturbances on Wildlife: A Review". In: *Drones* 9.4, p. 311. DOI: 10.3390/drones9040311. URL: https://www.mdpi.com/2504-446X/9/4/311.

Aharon, Nir, Roy Orfaig, and Ben-Zion Bobrovsky (2022). *BoT-SORT: Robust Associations Multi-Pedestrian Tracking*. arXiv: 2206.14651 [cs.CV]. URL: https://arxiv.org/abs/2206.14651.

Alsaidi, Muhanad et al. (2024). "Localization and tracking of beluga whales in aerial video using deep learning". In: *Frontiers in Marine Science* 11, p. 1445698. ISSN: 2296-7745. DOI: 10.3389/fmars.2024.1445698.

Barbedo, Jayme Garcia Arnal et al. (2019). "A Study on the Detection of Cattle in UAV Images Using Deep Learning". In: *Sensors* 19.24, p. 5436. DOI: 10.3390/s19245436.

Berger-Tal, Oded and José J. Lahoz-Monfort (2018). "Conservation technology: The next generation". In: *Conservation Letters* 11.6, e12458. DOI: 10.1111/conl.12458. URL: https://conbio.onlinelibrary.wiley.com/doi/abs/10.1111/conl.12458.

Bernardin, Keni and Rainer Stiefelhagen (2008). "Evaluating multiple object tracking performance: the clear MOT metrics". In: *EURASIP Journal on Image and Video Processing* 2008, pp. 1–10.

Boinski, Sue and Paul A Garber (2000). *On the move: how and why animals travel in groups*. University of Chicago Press.

Brickson, Leandra et al. (2023). "Elephants and algorithms: A review of the current and future role of AI in elephant monitoring". In: *J. R. Soc. Interface* 20.197, p. 20230367. ISSN: 1742-5689. DOI: 10.1098/rsif.2023.0367.

Cho, Yeong-Jun and Dohyung Kim (Jan. 2023). "Rethinking Multi-Object Tracking Based on Re-Identification and Appearance Model Management". In: *IEEE Access*. Volume PP, Pages 1-1 often indicate early access or in press. Check for updated publication details. DOI: 10.1109/ACCESS.2023.3274662.

Dai, Xiaohua et al. (Feb. 2007). "Short-Duration Daytime Movements of a Cow Herd of African Elephants". In: *Journal of Mammalogy* 88.1, pp. 151–157. ISSN: 0022-2372. DOI: 10.1644/06-MAMM-A-035R1.1. eprint: https://academic.oup.com/jmammal/article-pdf/88/1/151/2736489/88-1-151.pdf. URL: https://doi.org/10.1644/06-MAMM-A-035R1.1.

Dave, Brahm et al. (2023). "Wild Animal Detection using YOLOv8". In: *Procedia Computer Science* 230. 3rd International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN 2023), pp. 100–111. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2023.12.065. URL: https://www.sciencedirect.com/science/article/pii/S1877050923020707.

Delplanque, Alexandre, Samuel Foucher, Philippe Lejeune, et al. (Aug. 2021). "Multispecies detection and identification of African mammals in aerial imagery using convolutional neural networks". In: *Remote Sensing in Ecology and Conservation* 8. DOI: 10.1002/rse2.234.

Delplanque, Alexandre, Samuel Foucher, Jérôme Théau, et al. (Feb. 2023). "From crowd to herd counting: How to precisely detect and count African mammals using aerial imagery and deep learning?" In: *ISPRS Journal of Photogrammetry and Remote Sensing* 197, pp. 167–180. DOI: 10.1016/j.isprsjprs.2023.01.025.

Delplanque, Alexandre, Richard Lamprey, et al. (Nov. 2023). "Surveying wildlife and livestock in Uganda with aerial cameras: Deep Learning reduces the workload of human interpretation by over 70%". In: *Frontiers in Ecology and Evolution* 11. DOI: 10.3389/fevo.2023.1270857.

Delplanque, Alexandre, Julie Linchant, et al. (June 2024). "Will artificial intelligence revolutionize aerial surveys? A first large-scale semi-automated survey of African wildlife using oblique imagery and deep learning". In: *Ecological Informatics* 82, p. 102679. DOI: 10.1016/j.ecoinf.2024.102679.

Fang, Chengwu et al. (2024). "Enhancing Livestock Detection: An Efficient Model Based on YOLOv8". In: *Applied Sciences* 14.11. ISSN: 2076-3417. DOI: 10.3390/app14114809. URL: https://www.mdpi.com/2076-3417/14/11/4809.

Finn, Catherine, Florencia Grattarola, and Daniel Pincheira-Donoso (May 2023). "More losers than winners: investigating Anthropocene defaunation through the diversity of population trends". In: *Biological reviews of the Cambridge Philosophical Society* 98. DOI: 10.1111/brv.12974.

Gao, Ruopeng and Limin Wang (2024). *MeMOTR: Long-Term Memory-Augmented Transformer for Multi-Object Tracking*. arXiv: 2307.15700 [cs.CV]. URL: https://arxiv.org/abs/2307.15700.

Guirado, Emilio et al. (2019). "Whale counting in satellite and aerial images with deep learning". In: *Scientific reports* 9.1, p. 14271. DOI: 10.1038/s41598-019-50795-9.

Hamilton, Grant et al. (2020). "When You Can't See the Koalas for the Trees: Using Drones and Machine Learning in Complex Environments". In: *Biological Conservation* 247, p. 108598. DOI: 10.1016/j.biocon.2020.108598. URL: https://www.sciencedirect.com/science/article/abs/pii/S000632071931537X.

Hodgson, Jarrod C. et al. (2016). "Precision wildlife monitoring using unmanned aerial vehicles". In: *Scientific reports* 6, p. 22574. DOI: 10.1038/srep22574.

Inoue, Sota et al. (2019). "Spatial positioning of individuals in a group of feral horses: A case study using drone technology". In: *Mammal Research* 64.2, pp. 249–259.

IPBES (2019). *Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. Version 1. DOI: 10.5281/zenodo.6417333. URL: https://doi.org/10.5281/zenodo.6417333.

Kellenberger, Benjamin, Diego Marcos, and Devis Tuia (2018). "Detecting Mammals in UAV Images: Best Practices to address a substantially Imbalanced Dataset with Deep Learning". In: arXiv: 1806.11368 [cs.CV].

Khanam, Rahima and Muhammad Hussain (2024). *YOLOv11: An Overview of the Key Architectural Enhancements*. arXiv: 2410.17725 [cs.CV]. URL: https://arxiv.org/abs/2410.17725.

Kissling, W. Daniel et al. (2024). "Development of a cost-efficient automated wildlife camera network in a European Natura 2000 site". In: *Basic and Applied Ecology* 79, pp. 141–152. DOI: 10.1016/j.baae.2024.06.006. URL: https://doi.org/10.1016/j.baae.2024.06.006.

Koger, Benjamin et al. (2023). "Quantifying the movement, behaviour and environmental context of group-living animals using drones and computer vision". In: *Journal of Animal Ecology* 92.7, pp. 1357–1371. ISSN: 0021-8790. DOI: 10.1111/1365-2656.13904.

Lamba, Aakash et al. (2019). "Deep learning for environmental conservation". In: *Current Biology* 29.19, R977–R982. DOI: 10.1016/j.cub.2019.08.016. URL: https://www.sciencedirect.com/science/article/pii/S0960982219310322.

Liu, Wei et al. (2016). "SSD: Single Shot MultiBox Detector". In: *Computer Vision – ECCV 2016*. Springer International Publishing, pp. 21–37. ISBN: 9783319464480. DOI: 10.1007/978-3-319-46448-0_2. URL: http://dx.doi.org/10.1007/978-3-319-46448-0_2.

López, Jesús Jiménez and Margarita Mulero-Pázmány (2019). "Drones for Conservation in Protected Areas: Present and Future". In: *Drones* 3.1, p. 10. DOI: 10.3390/drones3010010. URL: https://www.mdpi.com/2504-446X/3/1/10.

Luiten, Jonathon et al. (Oct. 2020). "HOTA: A Higher Order Metric for Evaluating Multi-object Tracking". In: *International Journal of Computer Vision* 129.2, pp. 548–578. ISSN: 1573-1405. DOI: 10.1007/s11263-020-01375-2. URL: http://dx.doi.org/10.1007/s11263-020-01375-2.

McEvoy, John F., Graham P. Hall, and Paul G. McDonald (2016). "Evaluation of unmanned aerial vehicle shape, flight path and camera type for waterfowl surveys: disturbance effects and species recognition". In: *PeerJ* 4, e1831. DOI: 10.7717/peerj.1831.

Mou, Chao et al. (2023). "WAID: A Large-Scale Dataset for Wildlife Detection with Drones". In: *Appl. Sci.* 13.18, p. 10397. ISSN: 2076-3417. DOI: 10.3390/app131810397.

Mpouziotas, Dimitris, Petros Karvelis, and Chrysostomos Stylios (2024). "Advanced Computer Vision Methods for Tracking Wild Birds from Drone Footage". In: *Drones* 8.6, p. 259. ISSN: 2504-446X. DOI: 10.3390/drones8060259.

Mulero-Pázmány, Margarita et al. (2017). "Unmanned aircraft systems as a new source of disturbance for wildlife: A systematic review". In: *PLOS ONE* 12.6, e0178448. DOI: 10.1371/journal.pone.0178448. URL: https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0178448.

Pedrazzi, Lucia et al. (2025). "Advancing animal behaviour research using drone technology". In: *Animal Behaviour* 222, p. 123147. ISSN: 0003-3472. DOI: https://doi.org/10.1016/j.anbehav.2025.123147. URL: https://www.sciencedirect.com/science/article/pii/S0003347225000740.

Powers, David MW (2020). "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation". In: *arXiv preprint arXiv:2010.16061*. arXiv: 2010.16061 [stat.ML].

Rančić, Kristina et al. (2023). "Animal Detection and Counting from UAV Images Using Convolutional Neural Networks". In: *Drones* 7.3, p. 179. ISSN: 2504-446X. DOI: 10.3390/drones7030179.

Redmon, Joseph et al. (2016). *You Only Look Once: Unified, Real-Time Object Detection*. arXiv: 1506.02640 [cs.CV]. URL: https://arxiv.org/abs/1506.02640.

Ren, Shaoqing et al. (2016). *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. arXiv: 1506.01497 [cs.CV]. URL: https://arxiv.org/abs/1506.01497.

Ristani, Ergys et al. (2016). *Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking*. arXiv: 1609.01775 [cs.CV].

Roboflow (2025). *Roboflow Documentation: Build Vision Models with Roboflow*. Accessed: 2025-03-06. URL: https://docs.roboflow.com/.

Schad, Lukas and Julia Fischer (2023). "Opportunities and risks in the use of drones for studying animal behaviour". In: *Methods in Ecology and Evolution* 14.8, pp. 1864–1872.

Shukla, Vandita et al. (2024). "Towards Estimation of 3D Poses and Shapes of Animals from Oblique Drone Imagery". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-2-2024*, pp. 379–386. DOI: 10.5194/isprs-archives-XLVIII-2-2024-379-2024.

Tchamba, MN, H Bauer, and HH DE IONGH (1995). "Application of VHF-radio and satellite telemetry techniques on elephants in northern Cameroon". In: *African Journal of Ecology* 33.4, pp. 335–346.

Ultralytics (2025). *Ultralytics Documentation: YOLO Performance Metrics*. Accessed: 2025-03-06. URL: https://docs.ultralytics.com/guides/yolo-performance-metrics/.

Varghese, Rejin and Sambath M. (2024). "YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness". In: *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pp. 1–6. DOI: 10.1109/ADICS58448.2024.10533619.

Vermeulen, Cédric et al. (2013). "Unmanned aerial survey of elephants". In: *PloS one* 8.1, e54700. DOI: 10.1371/journal.pone.0054700.

Wang, Chien-Yao and Hong-Yuan Mark Liao (2024). *YOLOv1 to YOLOv10: The fastest and most accurate real-time object detection systems*. arXiv: 2408.09332 [cs.CV]. URL: https://arxiv.org/abs/2408.09332.

Wich, Serge A. and Alex K. Piel, eds. (2021). *Conservation Technology*. Oxford: Oxford University Press. ISBN: 9780198850243. URL: https://global.oup.com/academic/product/conservation-technology-9780198850243.

Wojke, Nicolai, Alex Bewley, and Dietrich Paulus (2017). *Simple Online and Realtime Tracking with a Deep Association Metric*. arXiv: 1703.07402 [cs.CV]. URL: https://arxiv.org/abs/1703.07402.

WWF (2024). *Living Planet Report 2024 – A System in Peril*. Gland, Switzerland: WWF.

Yaseen, Muhammad (2024). *What is YOLOv8: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector*. arXiv: 2408.15857 [cs.CV]. URL: https://arxiv.org/abs/2408.15857.

Yu, En et al. (2023). "MOTRv3: Release-Fetch Supervision for End-to-End Multi-Object Tracking". In: *arXiv preprint arXiv:2305.14298*. arXiv: 2305.14298 [cs.CV].

Zhang, Yifu et al. (2022). *ByteTrack: Multi-Object Tracking by Associating Every Detection Box*. arXiv: 2110.06864 [cs.CV]. URL: https://arxiv.org/abs/2110.06864.

Zhao, Xun et al. (2024). "A Vision-Based End-to-End Reinforcement Learning Framework for Drone Target Tracking". In: *Drones* 8.11. ISSN: 2504-446X. DOI: 10.3390/drones8110628. URL: https://www.mdpi.com/2504-446X/8/11/628.

6 Appendix

Table 10: Key Technical Specifications of the DJI Mavic 3 Pro

Characteristic	**Specification**
Hasselblad Camera	
Sensor	4/3 CMOS, 20 MP
Lens FOV	84°
Equivalent Focal Length	24 mm
Aperture	f/2.8 to f/11 (adjustable)
ISO Range (Video)	100-12800
Shutter Speed	8 s – 1/8000 s
Video & Imaging	
Max Video Resolution	5.1K: 5120×2700@50fps
	DCI 4K: 4096×2160@120fps
	4K: 3840×2160@120fps
Video Formats	MP4/MOV (MPEG-4 AVC/H.264, HEVC/H.265)
	Apple ProRes 422 HQ, 422, 422 LT (Cine Model)
Color Profiles	Normal, HLG, 10-bit D-Log M
Max Video Bitrate	H.264/H.265: 200 Mbps
Digital Zoom	Hasselblad Camera: 1-3×
	Medium Tele Camera: 3-7×
	Tele Camera: 7-28×
Gimbal	
Stabilization	3-axis mechanical (tilt, roll, pan)
Mechanical Range	Tilt: $-135°$ to $100°$
	Roll: $-45°$ to $45°$
	Pan: $-27°$ to $27°$
Controllable Range	Tilt: $-90°$ to $35°$
Max Control Speed (Tilt)	$100°/s$