IET Computer Vision

SPECIAL ISSUE CamTraps, AI and Ecology

ORIGINAL RESEARCH OPEN ACCESS

Stay Local or Go Global: Geo-Referenced Bounding Boxes for Tracking Wildlife in Thermal Drone Videos

Christoph Praschl[1,2] | Vincent Coucke[3] | Anna Maschek[1] | David C. Schedl[1]

[1]University of Applied Sciences Upper Austria, Hagenberg im Mühlkreis, Austria | [2]Johannes Kepler University Linz, Linz, Austria | [3]University of Hasselt, Hasselt, Belgium

Correspondence: Christoph Praschl (christoph.praschl@fh-hagenberg.at)

Received: 28 December 2025 | **Revised:** 25 June 2026 | **Accepted:** 14 July 2026

Keywords: computer vision | global positioning system | object detection | object recognition | object tracking | video cameras

ABSTRACT

Drone-based thermal imaging is increasingly used for wildlife monitoring, but tracking individual animals in aerial footage remains challenging in forested habitats due to low texture, occlusions and dynamic viewpoints. Traditional image-space tracking methods often fail because motion and appearance cues degrade in thermal data. We present a geo-referenced tracking framework that projects detections into world coordinates using drone localisation, camera pose and elevation models. Building on Deep OC-Sort, we introduce (i) a geo-native tracker operating in metric space and (ii) a hybrid tracker combining pixel-space tracking with geo-referenced recovery. Both use physically grounded motion models to improve identity preservation under occlusion. Evaluated on a large airborne thermal dataset (225 videos), our methods achieve the fewest ID switches, reducing identity errors by 17% compared to the best and 92% compared to the worst tested state-of-the-art algorithm—in both cases without custom embeddings, which further improve results. Additionally, geo-referenced outputs enable ecological analyses such as spatial mapping and movement estimation, demonstrating the advantages of global over image-based tracking.

1 | Introduction

Monitoring wildlife populations is a crucial task in ecology and conservation, traditionally accomplished through well-established ground-based methods such as camera traps [1]. Advances in technology have shifted this perspective from the ground to the air: Uncrewed aerial vehicles (UAVs or drones) now offer new possibilities for large-scale, flexible and often nonintrusive wildlife surveys [2–4]. Drones have been applied across a wide range of species and habitats, including (but not limited to) elephant monitoring on African savannahs [5, 6], koala detection in eucalyptus forests [7], deer surveys in temperate zones [8], and multi-species assessments in tropical rainforest canopies [9, 10].

Drones are increasingly deployed in densely vegetated environments such as temperate and tropical forests where traditional ground-based approaches are often limited [7, 9–12]. With the integration of thermal infrared cameras and AI-based image analysis the scope of aerial wildlife monitoring has further expanded [13]. Especially, thermal sensors are valuable for detecting endothermic animals through canopy cover, especially at night or during periods of low ambient temperatures. In addition to these, visual recordings, drones are also being used to collect complementary ecological data, including environmental DNA (eDNA) [12].

Despite these advantages, thermal-based aerial wildlife monitoring presents unique challenges. Thermal imagery lacks the rich textural information found in standard RGB data, typically yielding only coarse animal silhouettes. Moreover, the heat signatures can be ambiguous: unless full and distinct silhouettes or movements are visible, it can be hard to distinguish animals from other warm

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *IET Computer Vision* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

objects in the environment, such as sunlit stones or decomposing vegetation, particularly when using nonradiometric recordings with only relative temperature values instead of radiometric ones with absolute temperature measurements. This makes wildlife monitoring particularly difficult, especially when the forest canopy further partially or completely obscures animals [2].

An additional challenge arises in the context of population estimation: to avoid double-counting, it is essential to not only detect animals but also track and re-identify individuals across frames and flight paths. However, the combination of limited visual features, complex backgrounds and dynamic perspectives during UAV flight complicates reliable re-identification (as illustrated in Figure 1). In addition to the low resolution, minimal texture, and variable viewing angles, this task is particularly difficult due to frequent long-term occlusions caused by topography and vegetation. As a result, traditional visual tracking approaches—such as those relying on pixel-wise motion estimation (e.g., optical flow)—often fail under these conditions (see Figure 2). Although some tracking challenges, such as thermal texture ambiguity and dynamic viewing conditions, affect aerial surveys across all habitat types, the combination of long-term occlusion from canopy cover and the absence of persistent background features makes forests particularly demanding.

To address these challenges, we propose a novel multi-modal tracking algorithm that leverages both visual features and supplemental sensor data—specifically, the global position and global orientation of the drone. By fusing visual detections with geospatial and orientation information, our approach aims to apply tracking based on global coordinates instead of local image coordinates, which improves the accuracy of animal tracking and re-identification (Re-ID) in complex environments such as forests, enabling more robust wildlife population assessments from aerial surveys.

2 | State of the Art

2.1 | General Multi-Object Tracking

Recent advances in deep learning have significantly improved the ability to track multiple objects or points in video data, even in the presence of frequent and long-term occlusions. The dominant approaches follow a tracking-by-detection paradigm, leveraging deep neural networks to extract robust appearance features for associating detections across frames. Methods such as DeepSORT [14], FairMOT [15] and StrongSORT [16] incorporate learnt embeddings to re-identify (Re-ID) targets after occlusion, whereas motion-based strategies—ByteTrack [17], OC-SORT [18], and BoostTrack [19]—improve association by

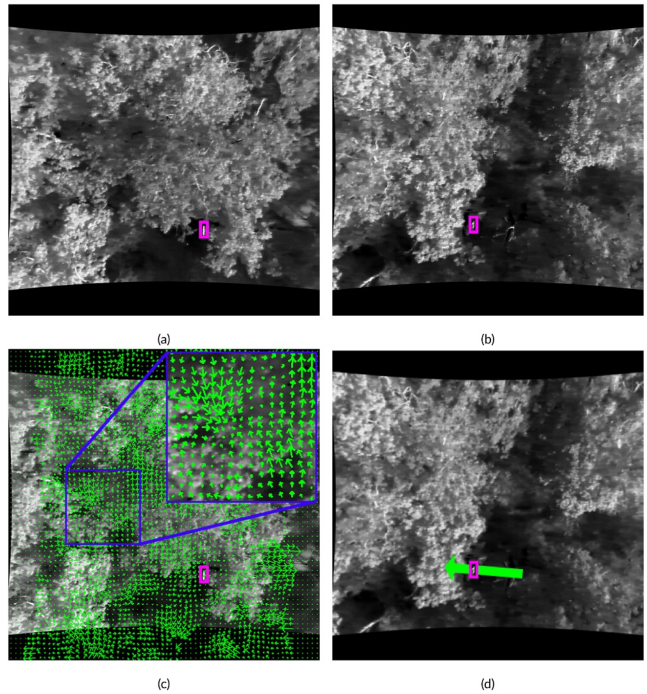


FIGURE 2 | Two consecutive frames of an undistorted thermal video where (a) is the start and (b) shows the view about 2 seconds later. An animal occluded between those frames by a tree is labelled with a bounding box. The fuzzy optical flow between both video frames, shown in (c), highlights the challenges for state-of-the-art tracking algorithms that rely on pixel features—partially zoomed in to improve visibility of flow directions. In our approach, the position of a detection is estimated based on the drone's global movement (d).

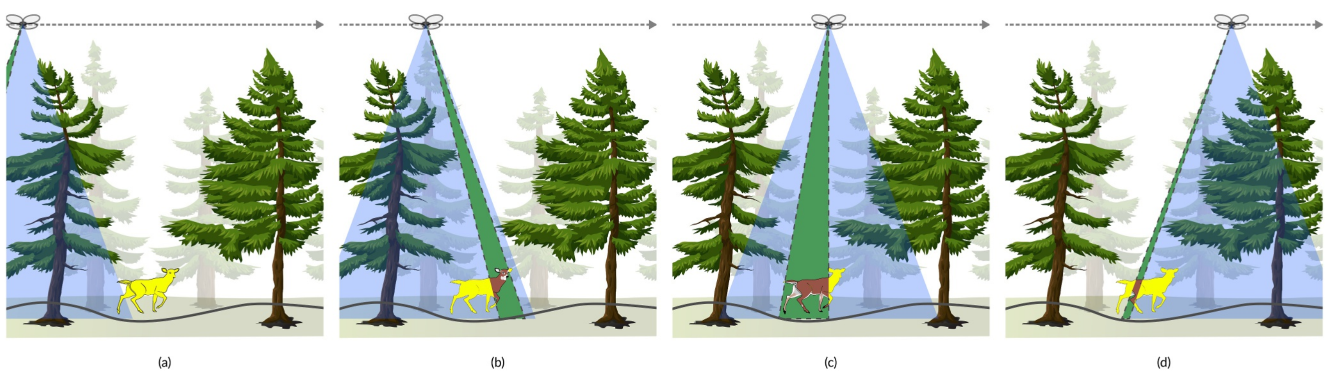


FIGURE 1 | Illustration of the tracking challenge: A UAV conducts an overhead survey of a forested area. Depending on the drone's position and viewing angle, the target animal may be completely occluded, as seen in (a), or only partially visible beneath the forest canopy, with different parts visible as shown in (b)–(d). Such long-term occlusions in combination with variable visibility often cause traditional visual tracking methods to fail in reliably re-identifying the animal across multiple frames or viewpoints. Whenever parts of the deer are not captured by the drone's camera, either because it is out of the view at all or occluded by the canopy, it is illustrated with a yellow overlay. The field of view (FOV) of the drone is shown as a blue frustum. Areas within the FOV, where the tree branches do not cover the animal are highlighted with a green triangle.

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

accounting for both high- and low-confidence detections or by modelling nonlinear motion to recover lost tracks. Algorithms such as Deep OC-SORT [20] and BoT-SORT [21] combine both Re-ID and motion estimation.

Transformer-based trackers (e.g., TrackFormer [22], MOTR [23–25] and MeMOTR [26]) unify detection and tracking in an end-to-end framework, using attention mechanisms and long-term memory to maintain identity through extended occlusions. This allows models to retain and recall distinctive features over many frames, reducing identity switches. Additionally, state space models such as MambaTrack [27] show strong tracking capabilities in occluded scenarios. In addition, point-level and interactive approaches—AllTracker [28], Track Anything [29], PIPs [30] and CoTracker [31]—enable tracking of arbitrary user-specified points, leveraging long-range temporal reasoning and memory modules to reconnect trajectories after occlusion. Recent developments in amodal tracking further improve robustness by predicting object positions during full occlusion [32].

Additionally, tracking-by-segmentation has been explored extensively in prior work, particularly in the context of video object segmentation approaches, such as MaskTrack R-CNN [33], Space-Time Memory Networks (STM; [34]), and more recent transformer-based methods such as AOT [35] and XMem [36], which propagate object masks across frames using learnt temporal representations. Recently, SAM3 and SAM3.1 [37] extend this paradigm by introducing a promptable segmentation framework with integrated temporal memory, enabling robust mask propagation across video frames.

2.2 | Wildlife Detection and Tracking in Aerial Video

Drone-based wildlife surveys have attracted increasing research attention across diverse species and environments [3, 4, 13]. Early work demonstrated that thermal drone imagery combined with deep learning can detect mammals in rainforest canopies [9], eucalyptus forests [7] and open savannah [5, 6]. More recent efforts have addressed deer surveys using enhanced object detectors [8], multi-species detection through sensor fusion of visible and thermal bands [38] and terrestrial surveys combining thermal drones with modern deep learning architectures [10]. For population estimation [39], demonstrated how drone-based detections following a point-based tracking strategy can be projected into world coordinates enabling quantitative behavioural and movement analysis using orthomosaic-based georeferencing. Although approaches for wildlife tracking from UAVs do exist, they predominantly focus on RGB data [6, 40, 41] or environments with few occlusions [42].

2.3 | Tracking in Thermal Imagery

Research on tracking in thermal data is more limited than that in RGB. Existing work demonstrates that adaptations of state-of-the-art MOT methods are feasible when retrained on thermal data [43, 44]—partially also covering wildlife, but not from an aerial view [45]. The BIRDSAI dataset [46] provides a benchmark for aerial thermal detection and tracking of humans and animals in antipoaching contexts, underscoring the relevance and difficulty of this domain. This body of work consistently highlights the challenges posed by limited features in thermal data, especially for visual Re-ID.

3 | Dataset

We adopt the following terminology consistently throughout this section and the rest of the paper. A *(source) video* denotes a raw thermal recording from a single UAV flight. A sequence is a contiguous temporal window extracted from a source video; each source video may yield one or more sequences. A frame is a single image within a video or sequence (29.97 frames per second). An identity refers to a uniquely identifiable individual animal that can be detected and re-identified within a video or sequence. An image crop is a bounding-box-aligned patch extracted around a single animal detection.

Between April 2022 and March 2025, we assembled a comprehensive airborne wildlife dataset, comprised of RGB and, most notably, thermal video data (extension of our preview published in ref. [47]). The dataset encompasses over 400 recorded flights, each capturing a diverse array of mammalian species. The videos were undistorted based on estimated camera intrinsic parameters (for more details see Subsection 4.1). Each video was manually annotated with axis-aligned bounding boxes by wildlife ecologists highlighting species and occlusion status using a key-frame approach.

Within this dataset, red deer (*Cervus elaphus*) and wild boar (*Sus scrofa*) account for approximately 30% of annotated instances each, making them the two most frequently represented species. Fallow deer (*Dama dama*) represent the third most frequently labelled species, whereas roe deer (*Capreolus capreolus*) comprise the fourth largest group. In addition to these dominant classes, the dataset includes a range of other mammals such as chamois (*Rupicapra rupicapra*) or Alpine ibex (*Capra ibex*), which have been recorded in diverse wild animal enclosures as well as natural habitats.

Building upon this foundation, we curated a specialised object-detection subset using 225 source videos, featuring three ecologically significant species in Austria—red deer, wild boar and roe deer Reimoser and Reimoser [48]. Because the primary use case targets density estimation and individual tracking rather than species classification, all animals are annotated under a single "animal" label. This subset serves as the basis for training an object detection model. It consists of 19,252 frames in total, comprising 44,186 bounding-box annotations across 15,730 training frames (~80%), 9823 annotations across 1696 validation frames (~10%) and 4926 annotations across 1826 test frames (~10%). Splits are defined based on individual source videos and flight locations to avoid biases related to environmental, sensor or flight characteristics. As a result, species distributions differ between the subsets. However, because our focus is on general wildlife tracking and we are not considering species classification at this stage, this should not affect downstream results.

In addition, we derived two further tailored versions of the dataset to address specific computer vision tasks, using the same video-level train/test assignments as the detection subset. The first is

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

designed for our MOT use case, supporting the development and benchmarking of tracking algorithms. It consists of 886 sequences in total (714 sequences comprising 239,184 frames and 2434 tracks in train; 172 sequences comprising 61,450 frames and 943 tracks in test), with 676,413 and 223,309 bounding box annotations, respectively; detailed per-split statistics are provided in Table 1. For the MOT dataset, the key-frame bounding boxes were interpolated to create complete, gap-free tracks covering both nonoccluded and occluded animal positions.

The second derived dataset focuses on the task of individual animal Re-ID, enabling more advanced and longitudinal monitoring of wildlife. It comprises 1950 identities in total (1725 identities in train; 225 identities in test), each represented by multiple image crops. The training split contains 32,894 image crops across 1725 identities (19.1 ± 2.2 crops per identity on average). The test split is further divided into a query set and a gallery set: the query set contains 757 image crops across 225 identities (3.4 ± 0.8 crops per identity in average), and the gallery set provides the candidate pool for retrieval with 3235 image crops across the same 225 identities (14.4 ± 2.4 crops per identity in average). All crops are uniformly extracted from the source frames with a size of 128 × 128 px. A summary of all three task-specific datasets is provided in Table 2.

4 | Methodology

Automated UAV missions are executed using predefined flight plans that specify waypoints, flight altitudes, speeds and camera actions, following the setup described by the authors in ref. [49]. These missions are uploaded to the flight controller, allowing the drone to autonomously follow the trajectory while maintaining stable altitude above ground level (AGL) and gimbal orientation. During the flight, a real-time kinematic (RTK; [50]) base station provides real-time differential corrections to the UAV, enabling centimeter-level positioning accuracy. This high-precision localisation is crucial for downstream geo-referencing, as even small positional drifts can accumulate into substantial mapping errors. Following the flight execution, multiple pre-processing steps are applied before the object detection and geo-referenced tracking pipeline is applied. This process is illustrated in Figure 3.

4.1 | Pre-Processing

The proposed pipeline builds on automated UAV flights conducted with DJI platforms (DJI M3T and M30T). These UAVs record (multi-spectral) videos while simultaneously logging the global position and orientation of both the drone and its gimbal-mounted camera. Global positions are provided in the WGS84 reference system (longitude, latitude, altitude). To enable reliable tracking in geographic space, all longitude–latitude pairs are converted into a local UTM-based coordinate system (for our dataset, UTM zone 33N, EPSG:32633). This transformation replaces angular geographic coordinates with metric planar coordinates, which is essential for computing distances, velocities, and intersection areas between geo-referenced bounding boxes. Using UTM coordinates ensures consistent units (metres) across the tracking pipeline, allowing accurate evaluation of spatial overlap and object motion on the ground—capabilities that are not feasible when working directly with spherical latitude–longitude values.

DJI stores metadata for each video at 29.97 Hz in SubRip Text (.srt) files, including camera position and orientation for every video frame. However, these positions are limited to a 5-digit precision, which corresponds to an approximate spatial accuracy of 1 m as highlighted by the authors in ref. [51]. In parallel, DJI devices generate a proprietary flight log containing higher-precision positioning (13-digit representation), albeit at a reduced frequency of 10 Hz and covering the entire flight, including segments without video recording (e.g., take-off, navigation to the first waypoint, return-to-home and landing). These flight logs can be accessed using third-party tools such as AirData[1].

To obtain high-precision global positioning for each video frame, we synchronize both data sources using their timestamps and an optimisation procedure that minimises the positional

TABLE 1 | Summary statistics of the MOT dataset.

Property	Train	Test
Frames per sequence (mean ± std)	335.0 ± 421.7	357.3 ± 204.6
Frames per sequence (median)	238	325
Animals per frame (mean ± std)	3.04 ± 3.21	3.84 ± 3.63
Sequences with > 1 track (%)	61.8%	75.0%
Occlusion frequency (%)	66.1%	42.9%
Occ. length (frames, mean ± std)	215.5 ± 189.9	172.0 ± 138.7
Species: Red deer	39.3%	28.6%
Species: Wild boar	50.9%	24.6%
Species: Roe deer	9.8%	46.8%

Note: Average values are reported as mean ± standard deviation. Occlusion (Occ.) frequency is defined as the fraction of tracks with at least one frame classified as occluded in the underlying dataset.

TABLE 2 | Overview of all three task-specific datasets derived from the same 225 source videos with shared video-level train/test assignments.

Dataset	Metric	Train	Test
Detection	Frames	15,730	1826
	Labels (bounding boxes)	44,186	4926
MOT	Sequences	714	172
	Frames	239,184	61,450
	Labels (tracks)	2434	943
	With bounding boxes	676,413	223,309
Re-ID	Identities	1725	225
	Image crops	32,894	3992
	Of which: Query	—	757
	Of which: Gallery	—	3235

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

discrepancies between the 5-digit and 13-digit measurements (as illustrated in Figure 4). Subsequently, we use linear interpolation to upsample the high-precision positions to 29.97 Hz. This approach leverages the centimetre-level accuracy offered by RTK-enabled flight operations [52].

Camera intrinsics are estimated using classical calibration procedures [53]. Because standard targets such as checkerboards are unsuitable for our multi-spectral aerial setup, we instead use 2D feature correspondences across RGB and thermal images. Conventional targets fail because thermal sensors capture emitted long-wave infrared radiation rather than reflected light, making printed patterns effectively invisible. Drone-based acquisition introduces further constraints: sensors operate at large varying distances, and any physical target would need to be several metres wide.

To address this, we rely on objects that yield stable, high-contrast signatures in both modalities. Structured objects, such as buildings (as shown in Figure A1), provide sharp contours suitable for cross-modal keypoint selection. Using these manually defined correspondences, we compute homographies and intrinsic parameters, including distortion coefficients, enabling geometric undistortion and consistent spatial alignment between sensors, which is required for accurate projection of points from image to world plane. After undistortion and cropping, the recordings have a resolution of 512 × 512 px (M3T), respectively, 1024 × 1024 px (M30T). Although RGB data are not used for tracking, they can support interpretation of ecological results as shown in Figure A10a.

4.2 | Detection

For our detection stage, we trained a YOLO11x model [54] class-agnostically for general wildlife detection on the described dataset. YOLO was selected for its strong balance of inference speed and detection accuracy and its native support for bounding-box-level outputs compatible with downstream tracking frameworks. The model converged rapidly during training, with all loss components decreasing steeply during the initial epochs and stabilising thereafter, indicating a good fit between the architecture and the thermal imaging domain, as shown in Figure A2 in

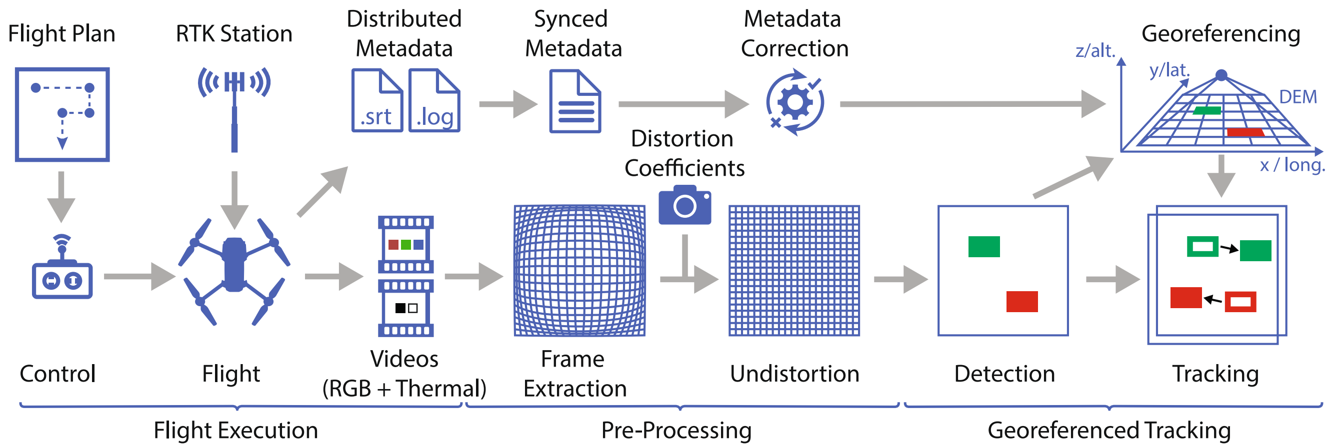


FIGURE 3 | Overview of the geo-referenced drone-based detection and tracking pipeline. The workflow begins with flight execution, including mission planning, control, RTK-based positioning and image metadata acquisition. In the pre-processing stage, video synchronisation, frame extraction, and camera-calibration-based undistortion are performed. The processed frames then enter the geo-referenced tracking stage, consisting of object detection, geo-referencing of bounding boxes and finally multi-frame tracking of detected animals.

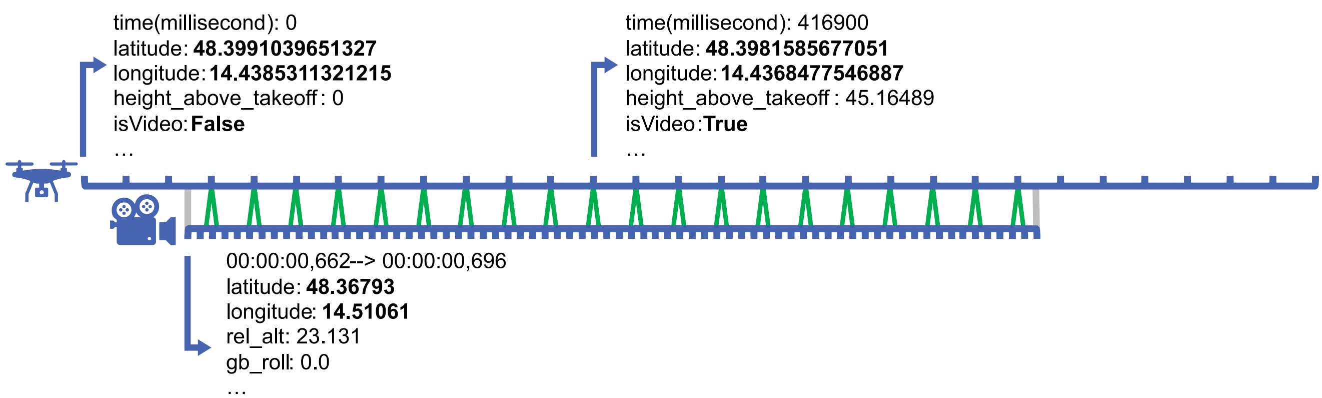


FIGURE 4 | Synchronisation of DJI's high-precision flight log (top) with the lower-precision per-frame SRT metadata (bottom). The flight log provides 13-digit GNSS positions at 10 Hz for the entire mission, including non-recording phases, whereas the SRT file supplies 5-digit GNSS positions and camera metadata at 29.97 Hz only during video capture. By aligning both data sources via timestamps and positional optimisation (green markers), high-precision positions can be interpolated for every video frame.

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the Appendix. Validation performance closely tracked the training curves, showing no signs of overfitting. The detector reached high accuracy across all metrics, achieving a mAP50 of approximately 0.97 and a mAP50–95 near 0.90, demonstrating both robust detection sensitivity and precise spatial localisation despite the diffuse contours characteristic of nonradiometric thermal signatures. Precision increased to above 0.93, whereas recall stabilised around 0.92, indicating that the model produced few false positives while still capturing the vast majority of animal instances. Notably, because of the nonradiometric nature of the data, the model performance may vary significantly across other recordings with lower foreground/background contrast.

4.3 | Re-Identification

As described in the state-of-the-art Section 2, some tracking algorithms incorporate appearance cues extracted through a dedicated Re-ID network. For example, the Deep OC-SORT architecture builds upon the foundational simple online and realtime tracking (SORT) algorithm [55] by incorporating these additional features, among other adaptations. Likewise, Deep OC-SORT substantially enhances robustness in the presence of occlusions, target interactions and complex motion patterns. However, because the original Deep OC-SORT framework, for example, was designed for multi-pedestrian tracking, its integrated Re-ID component is trained exclusively on RGB imagery of persons. This makes it suboptimal for our context, where aerial thermal imagery of wildlife with different visual characteristics forms the primary data source. The same is true for other algorithms.

To address this limitation, we trained a wildlife-specific thermal Re-ID model based on the Omni-Scale Network architecture Zhou et al. [56]. The model was trained on the aforementioned thermal data subset tailored for re-identification tasks, enabling it to capture domain-specific appearance characteristics of wildlife in thermal imagery. In this way, the Re-ID model achieves a Rank-1 accuracy of 94.2%, Rank-5 of 97.0%, Rank-10 of 98.3%, and Rank-20 of 98.7%, demonstrating its strong discriminative capability in challenging field conditions, provided that major parts of the specific animals are visible.

4.4 | Geo-Referenced Tracking

After object detection, each bounding box is projected from image space into geographic coordinates, enabling tracking directly in real-world metric space rather than in the image plane. Object geo-referencing is performed by casting rays from the drone's/camera's global position, directed by its global orientation, through the corners of each 2D detection, and intersecting them with a high-resolution digital elevation model (DEM) represented as a triangular mesh (as illustrated in Figure 5). This yields a set of world-space points describing the footprint of the detected animal on the terrain surface. The final geo-referenced bounding box is derived by taking the axis-aligned minimum and maximum of the projected 3D points. By ignoring the altitude information, we again project these points back to a planar space. Typically, a digital terrain model (DTM) is used as DEM. However, depending on the use case, a digital surface model (DSM) may be required for example, when mapping arboreal animals in trees.

Because residual errors in drone localisation can produce systematic projection bias, we incorporate manually determined correction factors. These corrections may be static for an entire flight or defined for specific frame intervals to compensate for time-varying deviations. The correction values are obtained by projecting the positions of static objects (ideally flat to avoid depth ambiguity) across multiple viewpoints and iteratively adjusting the yaw and altitude offsets to maximise the sharpness of the objects' edges while minimising their apparent spatial displacement (as illustrated in Figure 6). This procedure effectively treats the scene as a geometric reference grid and uses internal image consistency as the optimisation criterion. At the boundaries between adjacent correction intervals, we apply linear interpolation between the two corresponding correction sets to avoid step discontinuities in the projected coordinates; within each interval, the correction is applied as a constant offset. In practice, the magnitude and temporal pattern of these deviations vary considerably between flights. Across all experiments, we were unable to identify a consistent triggering condition or temporal structure that would reliably indicate whether such errors would occur at all, how large they would be or at which point(s) during a flight they might manifest. The only stable trend we observed is that the corrections, when required, are limited to altitude and/or yaw.

The quality and especially the spatial resolution of the DEM directly bounds the projection accuracy, as a coarser DEM introduces larger terrain approximation errors that propagate into the world-coordinate estimates. This relationship can be formalised: when a ray is cast from the camera centre through a detection pixel onto the mesh surface, the DEM cannot represent elevation variation within a single grid cell of size d. On a slope with gradient α, this introduces a worst-case elevation uncertainty of approximately $\Delta h \approx d \cdot \tan(\alpha)$, which in turn produces a horizontal displacement $\Delta r \approx d \cdot \tan(\alpha) \cdot \tan(\theta)$ at the angle θ between the cast ray and the vertical plane. Notably, as the terrain gradient approaches zero ($\alpha \to 0$), the displacement Δr vanishes regardless of grid size d, implying that in flat or gently undulating areas the spatial resolution of the DTM becomes increasingly less critical for geo-referencing accuracy. For geo-referenced detections to be

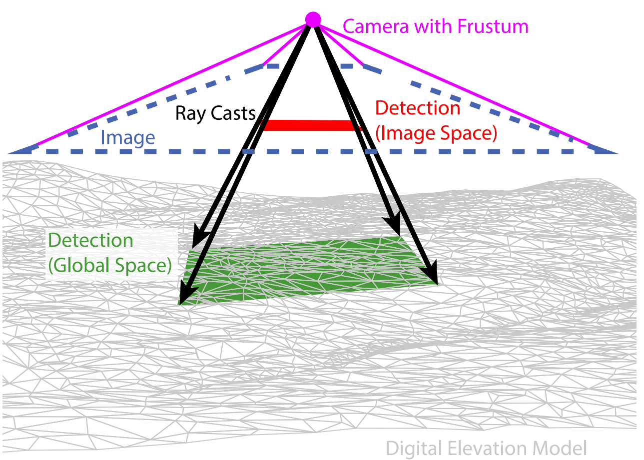


FIGURE 5 | Visualisation of the geo-referencing workflow, where the drone's position and orientation initiate ray casting through the image-space bounding-box corners to generate globally referenced bounding boxes on the digital elevation model.

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ecologically meaningful in more variable topographies, the displacement must remain well below the animal's ground footprint. Red deer (*Cervus elaphus*) occupy approximately 1.6–2.5 m in body length, wild boar (*Sus scrofa*) span roughly 0.9–1.8 m, and roe deer (*Capreolus capreolus*) are the smallest of the three at approximately 1.0–1.3 m. Therefore, sub-metre geo-referencing precision is required to reliably attribute a detection to an individual's true ground position. Considering a moderately steep forested slope of $\alpha \approx 25°$ and an off-nadir angle of $\theta \approx 30°$ near the edge of the sensor's field of view, a 10 m DTM (e.g., Copernicus GLO-30) yields $\Delta r \approx 2.7$ m, exceeding the body dimensions of all three target species and rendering individual-level localisation unreliable. A 1 m DTM as provided by the Austrian Federal Office of Metrology and Surveying (BEV) reduces this error to approximately 0.27 m, which satisfies the sub-animal precision requirement. The DTMs provided by Austrian federal-state open government data portals that we use in this work offer a still finer horizontal grid spacing of 0.5 m, further reducing the theoretical displacement to approximately 0.13 m. The ray-mesh intersection step is the computationally dominant stage of geo-referencing, and its cost scales with the total number of mesh triangles. However, even at 0.5 m resolution the overhead remains modest. For our MOT dataset, geo-referencing a bounding box takes on average only 25.88 ms, while initially loading the DEM requires, on average, 1.77 s (required only once per flight or per area in cases with multiple flights), as shown in Table 3.

As a subsequent step, we use the geo-referenced bounding boxes in two complementary ways using adapted versions of the Deep OC-Sort algorithm [20]. First, we change the algorithm to use bounding box definitions within a global space natively instead of the classic pixel-based (local) coordinates, and secondly, we use a hybrid approach, where the global coordinates are used during an additional association stage next to the local coordinates in the initial stages.

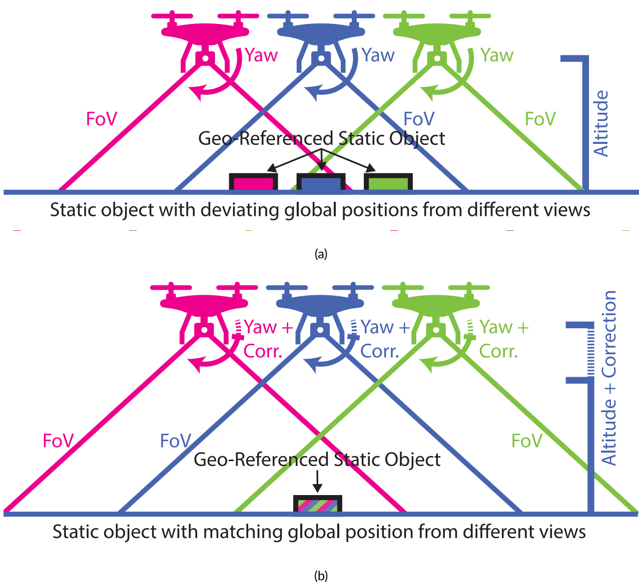


FIGURE 6 | To compensate positional and orientational errors, we apply manual corrections by taking (a) a set of views from different time steps showing the same static object, whose geo-referenced positions are initially inconsistent due to residual localisation errors. A correction factor is then applied to yaw and altitude offsets. Subfigure (b) shows the same static object viewed from the same perspectives after applying the correction, with its geo-referenced footprint now consistently aligned across all viewpoints.

4.4.1 | *Geo-Native Tracking*

The adaptation of the Deep OC-Sort algorithm transforms it from a traditional 2D pixel-plane tracker into a geo-native tracking system. This fundamental shift decouples the object-tracking logic from the camera's perspective, moving the entire inference pipeline into a geo-referenced coordinate system. By redefining the bounding box representation, the tracker no longer operates on screen coordinates but instead processes projected world coordinates such as metres.

The core modification lies within the Kalman filter, which has been re-engineered to track targets in real-world units rather than image pixels. Consequently, the state vector now represents the object's physical centroid, area, and real-world velocity. This transition addresses the perspective distortion issues inherent in standard tracking; for example, when recording with an oblique angle, an object moving at a constant speed towards the camera appears to accelerate in pixel space, confusing linear filters. By operating in a metric space, the constant-velocity assumption becomes physically valid, resulting in significantly more accurate trajectory predictions.

A major architectural change accompanying this shift is the removal of camera motion compensation (CMC) for the tracking state. In the original Deep OC-Sort, camera movement required complex image alignment to prevent stationary objects from drifting in the tracker's view (as shown in Figure 2). However, because the new system tracks coordinates in a fixed geographic frame, the logic becomes inherently invariant to camera ego-motion. Panning or tilting the camera changes where an object appears on the screen but does not alter its geographic coordinates, thereby stabilising tracking without the need to warp the Kalman state.

The data association phase has also been updated to utilise a geometric Intersection over Union (IoU) metric calculated in world space. This "Geo-IoU" standardises the association process because objects now retain consistent physical sizes

TABLE 3 | Summary statistics of digital elevation model (DEM) characteristics and bounding-box (BB) geo-referencing performance across all 886 flights of the MOT dataset.

	Metric	Mean	Std
DEM (n = 886)	Load time [s]	1.773	3.096
	File size [GB]	0.062	0.117
	Area [km^2]	1.086	2.418
	Triangles [10^6]	1.945	4.297
BB (n = 804,854)	Georef. time [ms]	25.879	192.791
	BBs per flight	3593.098	7149.989

Note: Reported values include mean, standard deviation (Std), minimum (Min) and maximum (Max). DEM-related metrics are computed per flight, including loading time as well as physical and spatial extent, whereas BBox geo-referencing time is measured per individual bounding box. This results in a substantially larger sample size for BBox-related measurements, and varying numbers of bounding boxes per flight lead to differing statistical distributions.

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

regardless of their distance from the camera lens, unlike in pixel space, where distant objects appear smaller. Furthermore, the velocity direction consistency checks, which help to filter unlikely matches during association, now operate on real-world velocity vectors. This ensures that the tracker understands the object's true cardinal direction, preventing perspective foreshortening from causing false mismatches. Despite the geo-native core, the system maintains a hybrid input–output interface to support appearance-based re-identification. Although the motion logic and the observation-centric recovery (OCR) pipeline rely strictly on geographic history to recover lost tracks, the system continues to use pixel-space crops to extract visual embeddings, as appearance features exist only in the image plane. The final output is a synthesis of these two worlds, providing robust geographic track IDs that are mapped back to their most recent pixel coordinates for visualisation on the original video feed, respectively, for comparison with classic image-space trackers.

4.4.2 | *Geo-Hybrid Tracking*

Additionally, we propose a hybrid multi-object tracking architecture that extends the Deep OC-Sort paradigm by integrating geo-spatial world coordinates into the association logic. Standard visual trackers often degrade under conditions of severe occlusion, similar visual appearance among targets or erratic camera motion. To mitigate these failure modes, our approach maintains a dual-domain state representation for every tracked object, allowing the system to fall back on absolute world-coordinate positioning when pixel-space metrics become unreliable. The algorithm operates through a rigorous cycle of state prediction, camera motion compensation, and a hierarchical three-stage association cascade.

The fundamental unit of the framework is the track state, which is composed of three distinct vectors: a pixel-space kinematic state, a geo-spatial state and a visual appearance embedding. The pixel-space state uses a Kalman filter with a constant-velocity model to track the bounding box centre, scale and aspect ratio, along with their derivatives. Simultaneously, the geo-spatial state tracks the object's position in projected world coordinates. Unlike the pixel state, the geo-state includes a velocity vector derived from physical units (metres per second), providing a motion model that is invariant to camera ego-motion. Additionally, a visual embedding is maintained using an exponential moving average of ReID features extracted from the detections, ensuring temporal consistency in object appearance. The tracking cycle initiates with a prediction phase. The Kalman filter propagates the pixel-domain state forward in time, whereas the geo-spatial state is linearly extrapolated based on its historical velocity. To account for camera movement, we employ a CMC as in the original Deep OC-Sort. An affine transformation matrix is computed by aligning feature points across consecutive frames, and this matrix is then applied to the predicted means and covariances of the pixel-space Kalman filters. This alignment ensures that the predicted bounding boxes are registered to the current image coordinate system before association begins. Crucially, the geo-spatial predictions do not require this compensation, as they exist in an absolute coordinate frame independent of the camera's perspective.

The association of detections to tracks is performed via a hierarchical cascade that prioritises high-confidence visual matches before resorting to spatial recovery. The first stage implements the standard Deep OC-Sort association logic. A cost matrix is constructed using a weighted combination of IoU, cosine distance between visual embeddings and a velocity-consistency term. This velocity term corrects the Kalman filter's predictions by incorporating the object's estimated speed, thereby handling nonlinear motion more effectively.

The second stage, termed observation-centric recovery (OCR), addresses the "overshooting" problem common in Kalman filtering when objects abruptly stop or change direction. Rather than comparing unmatched detections to the predicted track positions, which may have drifted, the algorithm compares them to the last-known observed pixel bounding boxes of the remaining unmatched tracks. This step is particularly effective for recovering tracks that have been temporarily occluded or stationary, preventing identity switches caused by prediction error.

The final and most novel component of the hierarchy is our geo-referenced association stage. This stage targets difficult cases where pixel-based overlap and visual appearance metrics fail, often due to long-term occlusion or significant displacement caused by camera shifts. Here, the algorithm utilises the projected world coordinates of the remaining unmatched detections and tracks. A cost matrix is computed using the Generalised Intersection over Union (GIoU) of the geo-spatial bounding boxes. GIoU is selected over standard IoU because it provides a nonzero gradient even when boxes do not strictly overlap, allowing the system to associate objects that are spatially proximate but disjoint. For tracks that are significantly displaced, the system applies a secondary fallback metric based on the Euclidean distance between centroids in world coordinates. If the physical distance falls within a strictly defined threshold (e.g., 3 m), the association is recovered. This stage provides a robust safety net, anchoring the tracker to the world.

Upon the conclusion of the association cascade, the system updates the internal states of all matched tracks. The Kalman filter is updated with the new pixel bounding box, the geo-history is appended with the new world coordinates, and the appearance embedding is updated via the moving average. Unmatched detections are initialised as new tentative tracks, while tracks that remain unmatched for a predefined number of frames are considered lost and subsequently terminated. This multi-stage approach ensures that the tracker benefits from the precision of visual data while retaining the robustness of physical spatial logic.

5 | Results

For the evaluation, we have adapted the Deep OC-Sort implementation within the `BoxMOT`[2] framework and utilised the `TrackEval`[3] framework for the evaluation. This allows comparison against state-of-the-art tracking algorithms: BoostTrack [19], BoT-SORT [21], ByteTrack [17], OCSort [57], StrongSORT [16] and the original Deep OC-Sort [20]. Additionally, we

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

integrate the point-based tracking algorithm proposed by the authors in ref. [39]. For this purpose, the centres of the detected bounding boxes are used as tracking foundation.

In addition to these tracking-by-detection approaches, we also integrated SAM3 [37],specifically, its recent update SAM3.1. However, these results must be interpreted with caution. First, SAM3 (and likewise SAM3.1) is primarily trained on RGB data, whereas our dataset consists of thermal imagery, leading to a significant domain gap that can negatively impact performance. Second, SAM3 does not inherently perform identity-aware multi-object tracking. Likewise, each box prompt is treated as an independent track initialisation, without any mechanism to merge overlapping or re-identify trajectories across frames. To enable a meaningful evaluation despite these limitations, we approximated a tracking scenario by leveraging the ground truth annotations: for each annotated trajectory, only the first frame was used to initialise SAM3 via a bounding box prompt. We then evaluated how consistently SAM3 was able to propagate the corresponding mask (and derived bounding box) over subsequent frames.

Furthermore, we integrated MambaTrack [27] into the evaluation. MambaTrack relies on a detector module, a feature extraction module and the Mamba Motion Predictor (MTP) to forecast object trajectories using a state-space model, which has demonstrated strong performance in occluded scenarios. However, the original MPT is trained on the DanceTrack dataset [58], which consists of human dance videos exhibiting movement characteristics substantially different from those of wildlife in thermal aerial footage with additional camera movement. To reduce this domain gap, we retrained the MPT component from scratch on our MOT training split. Additionally, we replace the detection module with our YOLO based detections. The resulting domain-adapted model is then evaluated using TrackEval alongside all other compared approaches, ensuring a consistent and fair evaluation protocol.

All experiments are conducted in a Docker based environment on a system equipped with a NVIDIA RTX4080 GPU (16 GB VRAM) and a Intel Core i9-13900K CPU with 32 GB of RAM with one exception. Due to its requirements of partially 30+ GB of VRAM (depending on the analysed sequence), SAM3.1 has been evaluated on an NVIDIA DGX-A100 system utilising one NVIDIA A100 GPU (40 GB VRAM) and equipped with 640 GB of RAM. The approaches are compared using identity-focused multi-object tracking metrics because all methods (except of the additionally compared SAM3.1) rely on the same YOLO detector (presented in Section 5) and therefore share identical detection quality (except for differences due to bounding box filters). This ensures that the evaluation isolates the performance of the tracking component. Identity consistency is assessed primarily through the ID switch (IDSW; [59]) metric, which counts how often a tracker assigns an incorrect identity to an object during its lifetime. To capture overall identity-preserving performance, we additionally report IDF1 and the association accuracy (AssA) metric [60]. IDF1 measures how well predicted trajectories maintain the correct identities over time, whereas AssA summarises association quality at the sequence level by jointly considering correct and incorrect identity assignments.

Based on the bounding boxes, we also prepared identity embeddings using our trained omni-scale model (presented in Section 6), which again are re-used equally across the different tracking approaches, where applicable. Because not all architectures support such visual cues, result changes are limited to those integrating this feature.

For the evaluation of our geo-referenced wildlife-tracking approach, we relied on the previously described MOT dataset, which is split into training and validation subsets. Because the major part of the tracking component in our current pipeline is purely deterministic and operates solely on provided bounding boxes, no learning-based methodology is involved at this stage (except of the Re-ID model). This is also true for the other compared approaches (except of the MTP module as part of MambaTrack). Consequently, we evaluate both subsets and report the results for each subset independently. In the context of the present work, this separation is not required for model optimisation (especially with BoxMOT's default Re-ID model) but ensures methodological consistency and prepares the dataset structure for future extensions. In particular, it allows seamless integration of end-to-end tracking frameworks that jointly learn detection and re-identification in the future, where the distinction between training and validation data becomes essential. The quantitative results are shown in Tables 4 and 5. A qualitative in-depth comparison for a tracked animal is shown in Figure 7 and with additional examples in the Appendix C with Figures A3–A8.

Our study focuses on tracking performance, with particular emphasis on identity stability in occluded scenarios. Likewise, other metrics provided by `TrackEval`, including detection-oriented metrics, are not thhe focus, but are still shown for comparability in the Appendix in Tables A1, A2.

6 | Discussion

The results demonstrate that incorporating geo-referenced information into the multi-object tracking pipeline improves identity preservation during wildlife monitoring in (thermal) drone videos, especially under conditions of severe occlusion and ambiguous visual appearance. Quantitatively, these advantages translate into a measurable reduction in IDSW. In all tests, the geo-hybrid approach achieves the lowest number of IDSW across the evaluation (test: 34/train: 168 without custom embeddings and 30/162 with such visual cues, respectively) while showing comparable or improved identity-preserving performance, outperforming not only the original Deep OC-SORT (40/202 without custom embeddings and 53/185 with visual cues, respectively) but also all other trackers by 17% (without custom embeddings) and 19% (with custom embeddings) or even higher in the combined results of the train and test splits. The improvement is particularly notable given that the geo-referenced methods do not use additional appearance training (except of the Re-ID model, i.e., however used equally across all methods); gains are achieved entirely by operating in physically meaningful coordinates, which stabilise motion modelling and association. Interestingly, the geo-native approach benefits markedly from custom Re-ID embeddings, with ID switches on the test split reduced from 84 to 33, indicating that there are some ambiguities in global positions, for example, due to close clusters of animals, that can be resolved

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 | Tracking results across the train and test split based on AssA, IDF1, and IDSW with BoxMOT's default embedding-based Re-ID model "osnet_ain_x1_0.pt", where applicable (*).

	Test			Train		
Model	**AssA ↑**	**IDF1 ↑**	**IDSW ↓**	**AssA ↑**	**IDF1 ↑**	**IDSW ↓**
StrongSORT*	55.324	51.171	509	48.565	43.167	1989
ByteTrack	54.992	51.701	289	48.996	43.867	952
BoT-SORT*	56.58	<u>52.283</u>	264	51.738	<u>45.14</u>	718
Koger et al. [39]	50.254	46.42	221	47.67	42.117	628
MambaTrack	51.697	47.125	135	47.581†	42.139†	536†
SAM3.1	43.355	26.911	128	40.467	22.805	412
OCSort	52.435	46.185	108	47.871	41.971	445
BoostTrack*	50.603	43.896	97	46.541	40.933	361
Deep OC-Sort*	56.526	48.806	40	<u>52.434</u>	44.976	202
Geo-native* (ours)	52.703	46.225	84	48.299	42.363	370
Geo-hybrid* (ours)	<u>56.672</u>	48.799	<u>34</u>	52.175	44.878	<u>168</u>

Note: The bias, since the MPT module of MambaTrack is trained on the the train dataset (†). Underlined values indicate the best result for each metric, evaluated column-wise. Bold text identifies our two proposed approaches.

TABLE 5 | Tracking results across the train and test split based on AssA, IDF1, and IDSW with our custom embedding-based Re-ID model, where applicable (*).

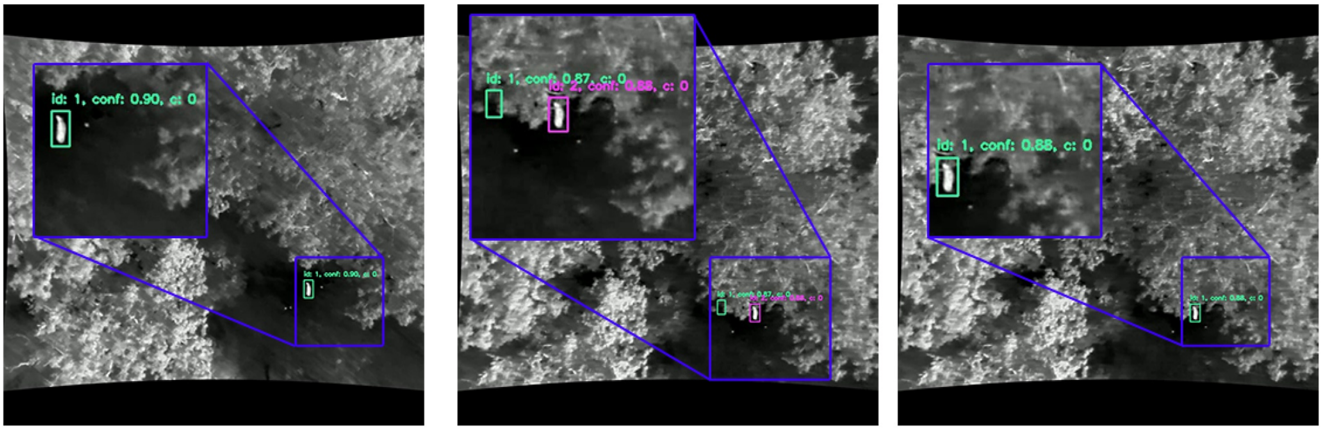
	Test			Train		
Model	**AssA ↑**	**IDF1 ↑**	**IDSW ↓**	**AssA ↑**	**IDF1 ↑**	**IDSW ↓**
StrongSORT*	55.31	51.146	507	48.573	43.174	1986
ByteTrack	54.992	51.701	289	48.996	43.867	952
BoT-SORT*	56.58	<u>52.283</u>	264	51.738	<u>45.14</u>	718
Koger et al. [39]	50.254	46.42	221	47.67	42.117	628
MambaTrack	51.697	47.125	135	47.581†	42.139†	536†
SAM3.1	43.355	26.911	128	40.467	22.805	412
OCSort	52.435	46.185	108	47.871	41.971	445
BoostTrack*	50.92	43.951	94	46.754	40.863	350
Deep OC-Sort*	55.349	48.033	53	<u>52.452</u>	44.994	185
Geo-native* (ours)	56.666	48.694	33	51.876	44.825	167
Geo-hybrid* (ours)	<u>56.714</u>	48.841	<u>30</u>	52.191	44.889	<u>162</u>

Note: The bias, since the MPT module of MambaTrack is trained on the the train dataset (†). Underlined values indicate the best result for each metric, evaluated column-wise. Bold text identifies our two proposed approaches.

using some additional image level comparisons. The geo-hybrid approach also improves, whereas the standard Deep OC-Sort algorithm degrades in the test dataset and BoT-SORT does not change at all. Although especially the original Deep OC-Sort model already achieves strong results on our dataset, the use of global coordinates improves tracking in cases with long-term occlusions (as shown in Figure 7). However, because the dataset contains a large proportion of samples without (long-term) occlusions, the adapted algorithms can only demonstrate measurable improvements in the comparatively few sequences that exhibit long-term occlusions. An additional issue arises during evaluation because, although the detection model was trained in a class-agnostic manner using only deer and wild boars, it partially generalises to other species such as horses (IDs 1, 2, 5, 7, 14, and 15 in Figure 8a,c). These individuals are detected and tracked correctly by the compared approaches, but they are not represented in the ground-truth data, and thus bias the metrics (especially AssA and IDF1).

Consistent with the conservative detection filtering strategy inherited from Deep OC-SORT, the geo-referenced trackers report lower detection recall than high-recall baselines (ByteTrack, BOTSort, StrongSORT), producing approximately 70K (train)/18k (test) fewer false positive track–detection assignments without custom embeddings (as shown in Table A1). This behaviour stems from the deliberate suppression of low-confidence detections to avoid erroneous associations, prioritising identity stability over detection completeness. As a result, DetA is lower (≈ 32 for train/≈ 27 for test) compared to the high-recall baselines (≈ 36 for train/≈ 34 for test), which in turn reduces the HOTA score, which combines DetA and AssA. However, within the same detection regime—that is, among trackers following a similarly conservative filtering strategy (OC-SORT, Deep OC-SORT)—the geo-referenced trackers achieve competitive or superior HOTA, driven by their stronger AssA. The lower HOTA relative to BOTSort or ByteTrack should therefore be interpreted as a consequence of the detection–association trade-

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



(a) (b) (c)

FIGURE 7 | Visual comparison between the original Deep OC-SORT algorithm in image-space (b) and the geo-hybrid implementation (without custom embeddings) (c), demonstrating improved tracking performance when the animal is occluded between frames. Image (a) represents the initial frame, with (b) and (c) captured 1.7 seconds later. With local coordinates, the algorithm fails to maintain consistent identification as highlighted by the different bounding box IDs and colours (green vs. magenta). In contrast, our method successfully tracks the occluded animal throughout the sequence. Zoomed in for better readability. Sequence ID 26_3 in our MOT dataset.

off rather than weaker tracking performance. This interpretation is further supported by the reduced number of IDSW, indicating that stricter detection acceptance effectively suppresses erroneous associations at the cost of recall.

In contrast to the tracking-by-detection approaches, the behaviour of SAM3.1 reveals several fundamental limitations in our setting. Most notably, SAM3.1 is trained exclusively on RGB imagery, which introduces a substantial domain gap when applied to our thermal data. This mismatch becomes particularly evident when using text prompts, which did not yield any meaningful results in our experiments. In contrast, bounding box prompts allow for partial functionality, enabling SAM3.1 to initialise and propagate object masks across frames. Under ideal conditions, the model demonstrates strong short-term temporal consistency and is able to reliably track the prompted region as long as the object remains clearly visible. However, this strength quickly deteriorates in the presence of occlusions as shown in the Appendix in Figure A9. As soon as the target is partially or fully occluded, SAM3.1 fails to maintain a stable representation, often causing the predicted masks and consequently the derived bounding boxes to drift, expand uncontrollably or even collapse. In several cases, bounding boxes grew to cover large portions of the image or were entirely lost, highlighting the lack of robustness in out-of-distribution scenarios (thermal aerial recordings) similar to that of ours. These observations underline that, although SAM3.1 shows promising capabilities for prompt-based mask propagation, its applicability for reliable object tracking in thermal wildlife monitoring remains limited without substantial domain adaptation (e.g., with retraining) and explicit mechanisms for occlusion handling.

MambaTrack, despite being retrained on our domain-specific MOT training data, achieves only a mid-range performance of 135 IDSW on the test split. This places it clearly ahead of methods that suffer heavily from the complexity of our data (such as StrongSORT (509), ByteTrack (289), BoT-SORT (264) and the centroid-based approach proposed by Koger et al. (221) but substantially behind the stronger tracking-by-detection methods such as BoostTrack (97), OC-SORT (108) and Deep OC-SORT (40), as well as both of our geo-referenced variants. The relatively high identity switch count is notable given that MambaTrack's MPT component was explicitly retrained on our training data rather than applied zero-shot. This suggests that the domain gap is not merely a question of motion statistics: the combination of low-texture thermal signatures, highly variable viewing angles and long-term canopy occlusion likely limits the MPT's ability to produce reliable trajectory predictions across the characteristic motion patterns in our dataset, even after retraining. Unlike the tracking-by-detection baselines, MambaTrack's motion predictions are more deeply tied to the learnt representation of trajectory dynamics, which may not transfer well to the erratic and intermittent nature of wildlife movement coupled with drone movements in forested aerial surveys. These findings indicate that state-space model-based tracking approaches require not only domain-specific retraining of the motion model but potentially a broader adaptation of the full pipeline—including the appearance and association components—to handle the specific challenges of thermal aerial wildlife data.

The geo-referenced trackers share a similar conceptual motivation with the approach of Koger et al. [39] in the direction of downstream tasks, particularly in aiming to obtain a global overview of animal positions. However, Koger et al. project drone-based animal detections after tracking from pixel space into world coordinates by generating and geo-referencing an orthomosaic of the entire flight. This process implicitly localises camera and animal positions and enables quantitative analysis of group movement and behaviour. In contrast, our approach does not rely on orthomosaic generation via photogrammetry. Instead, it employs per-frame ray-casting against a pre-existing DEM to estimate world coordinates directly from individual frames and uses this global reference for tracking. This significantly reduces processing overhead since no 3D reconstruction of individual flights is required. This advantage, however, comes at the cost of requiring a pre-existing elevation model.

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

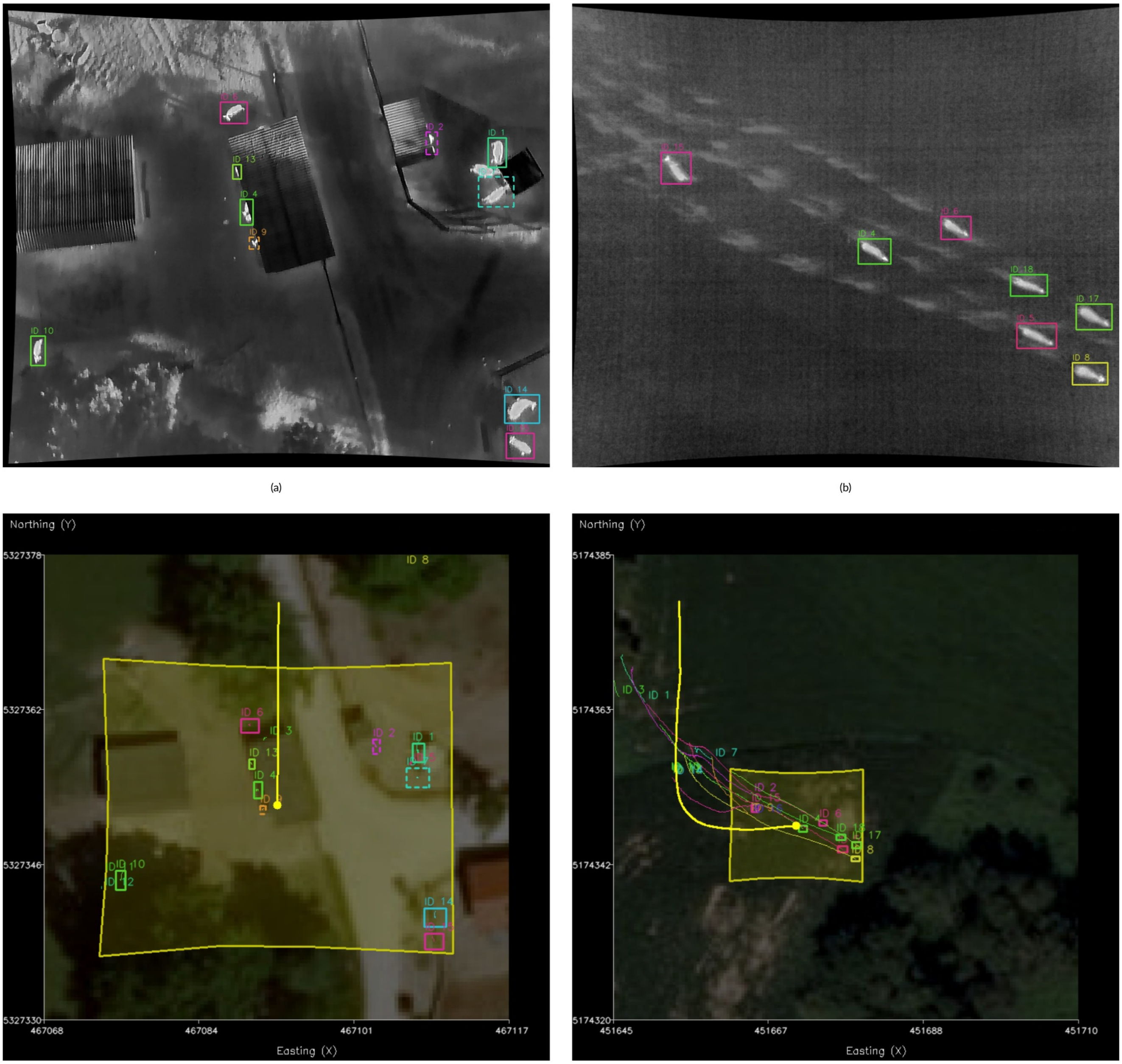


FIGURE 8 | Comparison of tracking results in local image space (top row) and global map space (bottom row), including the geo-referenced bounding boxes, field of view (yellow area), and flight route (yellow line). Left Column (a, c): Detected (continuous box) and interpolated (dashed box) positions of deer (left of street) and horses (right of street)—a result from the generalised detection model. Right Column (b, d): A moving group of deer showing tracked paths allowing examination of movement patterns. This example also contains lost tracks, because animals partially left and re-entered the field of view. Because of that, in combination with the movement, no re-identification can be established. Sequence Ids 14_1 (left column) and 223_1 (right column) in our MOT dataset.

Although this work focuses on thermal imagery, the geo-referenced tracking principles are not inherently modality-specific. The enabling factors—RTK drone localisation, camera pose, and DEM intersection—are available regardless of image modality. In principle, the geo-native and geo-hybrid frameworks also work on RGB video as shown in Figure A10a. The primary challenge for RGB transfer would be a domain-appropriate detection and a Re-ID model, as the current models are trained exclusively on thermal data.

Beyond improving temporal consistency, the use of geo-referenced bounding boxes provides several additional advantages. First, because detections are anchored in a global coordinate system, the animals' positions can be directly documented on a map and can be interpolated when no detection is available for a specific frame (as shown in Figure 8a,c). This makes spatial patterns across monitored areas more comparable over time periods or recurring flights and allows changes in habitat use or population distribution to be visualised more clearly.

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Second, global bounding boxes enable the monitoring of movement patterns (as shown in Figure 8b,d) at a landscape scale as long as the animals remain within the camera's field of view, without the need to generate an orthomosaic using resource-intensive photogrammetry approaches (as, e.g., employed by the authors in ref. [39]) because we only have to geo-reference four points per bounding box instead of requiring overlapping images, feature matching, 3D reconstruction, bundle adjustment, and dense point clouds. This can reduce processing efforts while still providing trajectories and spatial behaviour.

Third, geo-referenced detections allow the extraction of metric measurements directly from the (tight) bounding boxes, enabling quantitative analyses such as estimating animal sizes, distances or spatial densities. Because each bounding box is anchored in a geographic coordinate frame, object dimensions can be estimated in real-world units rather than merely in image space. This not only supports more reliable long-term tracking but also opens the door to ecologically relevant morphometric analyses. Importantly, accurate metric estimation is most reliable on the ground plane, while objects located above ground height inherently introduce greater uncertainty due to depth ambiguity and geometric projection effects. This limitation is particularly pronounced during non-nadir flights, where oblique viewing angles amplify perspective distortions and small elevation differences can lead to disproportionately large errors in reconstructed object dimensions.

7 | Conclusion and Outlook

This work demonstrates that incorporating geo-referenced information improves the robustness of multi-object tracking in thermal drone videos of wildlife. By projecting detections into a global coordinate system and integrating these coordinates into both geo-native and hybrid tracking pipelines, we mitigate many of the limitations that arise from perspective changes, long-term occlusions, and ambiguous thermal appearance. Our experiments on a large-scale airborne thermal dataset show that both proposed approaches reduce identity switches compared to state-of-the-art image-space trackers, despite using identical visual detections. Beyond improved identity preservation, geo-referenced bounding boxes enable additional downstream benefits, including mapping animal locations, analysing movement patterns at a landscape scale, and supporting metric measurements directly in geographic space. Together, these results highlight the value of moving from local image coordinates towards physically meaningful world coordinates for wildlife monitoring and pave the way for more reliable spatially explicit assessments in drone-based ecological surveys.

The presented tracking approach represents an important intermediate step rather than a final solution for wildlife population estimations. Building on these results, future work will focus on leveraging the generated tracks of bounding boxes within a majority voting–based framework allowing to go from rough "animal" classes to a fine-grained classification including species as well as age and gender identification.

For the present comparison, we have limited our evaluation to learning-free MOT approaches based on upstream detection results. In future work, we plan to extend this analysis not only to end-to-end frameworks such as MOTRv3 [25] but also additional point-level approaches such as AllTracker [28] and assess how they can be incorporated in a global-coordinate-based approach.

Additionally, the projection approach is not limited to bounding boxes and can be generalised to arbitrary geometric data. Building on this, we plan to extend the pipeline to other computer vision tasks such as segmentation. Preliminary results with SAM3 [37] are promising (Figure A10). However, as in the MOT use case, a dedicated evaluation dataset will be required to enable a thorough analysis.

Author Contributions

Christoph Praschl: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, resources, software, validation, visualization, writing – original draft. **Vincent Coucke:** data curation, investigation, methodology, software, validation. **Anna Maschek:** data curation, investigation, methodology, software, validation. **David C. Schedl:** formal analysis, funding acquisition, methodology, project administration, supervision, writing – review and editing.

Acknowledgements

The used dataset was created as part of the research project *BAMBI*, also funded by FFG (program number: 892231), and done together with our partners from the company *Büro für Wildökologie und Forstwirtschaft e.U.*, namely Horst Leitner, Wolfram Jantsch, and Stephanie Wohlfahrt, as well as *Umweltdata GmbH*, namely Günther Bronner, Gudrun Eppich, and Boris Jawecki, and Rudolf Schneeberger (*Viewcopter e.U.*). Additionally, we thank Alexander Wipplinger and Florian Willemsen from the company *SpektakulAIR* for providing samples for the proposed dataset. This paper was prepared with the assistance of the AI tool ChatGPT 5.1 (developed by OpenAI), which was used across all sections exclusively to rephrase and improve the clarity of the text. The content, structure, and scientific contributions of the paper were developed entirely by the authors, without AI-generated input. Open Access funding provided by FH Oberosterreich/KEMÖ.

Funding

This project is funded by the Austrian Research Promotion Agency FFG (project *THUMPER*; programme number: 917796) within the funding programme "Dissertation programme of the University of Applied Sciences Upper Austria 2024", for which the budget is provided by the Federal Republic of Austria.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The implemented tracking algorithms are publicly available on GitHub: https://github.com/bambi-eco/Geo-Referenced-Tracking. Additional scripts related to geo-referencing aerial data are part of two different repositories: https://github.com/bambi-eco/Dataset/ and https://github.com/bambi-eco/alfs. Also, the datasets used in this work are publicly available. The detection dataset can be accessed via Roboflow Universe as well as Zenodo, whereas the MOT and re-identification datasets are hosted exclusively on Zenodo due to Roboflow's current lack of support for these specific use cases. All datasets are available based on their specific ID at the specific platforms: Detection: https://universe.roboflow.com/dseducation and https://doi.org/10.5281/zenodo.15773102.

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Re-ID: https://doi.org/10.5281/zenodo.15532500. MOT: https://doi.org/10.5281/zenodo.15532567. For geo-referencing our drone flights, we relied extensively on digital elevation models with a resolution of 0.5 m provided by various Austrian federal states. Although these DEMs are publicly accessible, redistribution is not permitted. To reproduce our results, you must therefore download the required tiles directly from https://www.data.gv.at/ or from state-specific platforms (e.g., https://www.doris.at/ for Upper Austria) and convert them to meshes with relative EPSG:32633 coordinates. The selection of DEM tiles is defined via JSON description files that also specify the absolute origin (UTM and WGS84) and the spatial extent. These JSON files, together with the manually determined correction files used in our workflow, are published as a replication package on Zenodo (https://doi.org/10.5281/zenodo.17865822). To ensure reproducibility of our tracking approach without the pre-processing steps, we additionally provide our geo-referenced bounding boxes alongside the conventional image-space bounding boxes as part of this replication package. Finally, the trained models are also publicly available via Hugging Face (https://huggingface.co/cpraschl/bambi-models and https://huggingface.co/models?p=1&sort=trending&search=Bambi).

Endnotes

[1] https://airdata.com/ (Accessed 24.06.2026).

[2] https://github.com/mikel-brostrom/boxmot (Accessed on 15.12.2025).

[3] https://github.com/JonathonLuiten/TrackEval (Accessed on 15.12.2025).

References

1. L. Silveira, A. T. Jácomo, and J. A. F. Diniz-Filho, "Camera Trap, Line Transect Census and Track Surveys: A Comparative Evaluation," *Biological Conservation* 114, no. 3 (2003): 351–355, https://doi.org/10.1016/s0006-3207(03)00063-6.

2. L. F. Gonzalez, G. A. Montes, E. Puig, S. Johnson, K. Mengersen, and K. J. Gaston, "Unmanned Aerial Vehicles (Uavs) and Artificial Intelligence Revolutionizing Wildlife Monitoring and Conservation," *Sensors* 16, no. 1 (2016): 97, https://doi.org/10.3390/s16010097.

3. E. Corcoran, M. Winsen, A. Sudholz, and G. Hamilton, "Automated Detection of Wildlife Using Drones: Synthesis, Opportunities and Constraints," *Methods in Ecology and Evolution* 12, no. 6 (2021): 1103–1114, https://doi.org/10.1111/2041-210x.13581.

4. D. Tuia, B. Kellenberger, S. Beery, et al., "Perspectives in Machine Learning for Wildlife Conservation," *Nature Communications* 13, no. 1 (2022): 792, https://doi.org/10.1038/s41467-022-27980-y.

5. W. L. Hartmann, V. Fishlock, and A. Leslie, "First Guidelines and Suggested Best Protocol for Surveying African Elephants (Loxodonta africana) Using a Drone," *Koedoe* 63, no. 1 (2021): 1–9, https://doi.org/10.4102/koedoe.v63i1.1687.

6. C. C. Elchik, S. Wich, and A. Burger, "A Framework for Detecting and Tracking Elephants in Drone Videos," *Drone Systems and Applications* 13 (2025): 1–16, https://doi.org/10.1139/dsa-2025-0032.

7. E. Corcoran, S. Denman, J. Hanger, B. Wilson, and G. Hamilton, "Automated Detection of Koalas Using Low-Level Aerial Surveillance and Machine Learning," *Scientific Reports* 9, no. 1 (2019): 3208, https://doi.org/10.1038/s41598-019-39917-5.

8. H. Lyu, F. Qiu, L. An, D. Stow, R. Lewison, and E. Bohnett, "Deer Survey From Drone Thermal Imagery Using Enhanced Faster r-cnn Based on Resnets and Fpn," *Ecological Informatics* 79 (2024): 102383, https://doi.org/10.1016/j.ecoinf.2023.102383.

9. R. Kays, J. Sheppard, K. Mclean, et al., "Hot Monkey, Cold Reality: Surveying Rainforest Canopy Mammals Using Drone-Mounted Thermal Infrared Sensors," *International Journal of Remote Sensing* 40, no. 2 (2019): 407–419, https://doi.org/10.1080/01431161.2018.1523580.

10. K. Backman, J. Wood, M. Brandimarti, C. T. Beranek, and A. Roff, "Human Inspired Deep Learning to Locate and Classify Terrestrial and Arboreal Animals in Thermal Drone Surveys," *Methods in Ecology and Evolution* 16, no. 6 (2025): 1239–1254, https://doi.org/10.1111/2041-210x.70006.

11. S. Wohlfahrt, C. Praschl, H. Leitner, et al., "Advancing Wildlife Monitoring: Drone-Based Sampling for Roe Deer Density Estimation," in *International Wildlife Congress 2025*, (2025).

12. E. Aucone, S. Kirchgeorg, A. Valentini, L. Pellissier, K. Deiner, and S. Mintchev, "Drone-Assisted Collection of Environmental dna From Tree Branches for Biodiversity Monitoring," *Science Robotics* 8, no. 74 (2023): eadd5762, https://doi.org/10.1126/scirobotics.add5762.

13. Z. Xu, T. Wang, A. K. Skidmore, and R. Lamprey, "A Review of Deep Learning Techniques for Detecting Animals in Aerial and Satellite Images," *International Journal of Applied Earth Observation and Geoinformation* 128 (2024): 103732, https://doi.org/10.1016/j.jag.2024.103732.

14. N. Wojke, A. Bewley, and D. Paulus, "Simple Online and Realtime Tracking With a Deep Association Metric," in *2017 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2017), 3645–3649.

15. Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "Fairmot: On the Fairness of Detection and Re-Identification in Multiple Object Tracking," *International Journal of Computer Vision* 129, no. 11 (2021): 3069–3087, https://doi.org/10.1007/s11263-021-01513-4.

16. Y. Du, Z. Zhao, Y. Song, et al., "Strongsort: Make Deepsort Great Again," *IEEE Transactions on Multimedia* 25 (2023): 8725–8737, https://doi.org/10.1109/tmm.2023.3240881.

17. Y. Zhang, P. Sun, Y. Jiang, et al., "Bytetrack: Multi-Object Tracking by Associating Every Detection Box," in *European Conference on Computer Vision* (Springer, 2022), 1–21.

18. X. Xiao and X. Feng, "Multi-Object Pedestrian Tracking Using Improved yolov8 and oc-sort," *Sensors* 23, no. 20 (2023): 8439, https://doi.org/10.3390/s23208439.

19. V. Stanojević and B. Todorović, "Boosttrack++: Using Tracklet Information to Detect More Objects in Multiple Object Tracking," *arXiv preprint arXiv:2408.13003* (2024).

20. G. Maggiolino, A. Ahmad, J. Cao, and K. Kitani, "Deep oc-Sort: Multi-Pedestrian Tracking by Adaptive Re-Identification," in *2023 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2023).

21. N. Aharon, R. Orfaig, and B. Z. Bobrovsky, "Bot-Sort: Robust Associations Multi-Pedestrian Tracking," *arXiv preprint arXiv:2206.14651* (2022).

22. T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer, "Trackformer: Multi-Object Tracking With Transformers," in *Proceedings of the IEEE/CVF Conference on CV and Pattern Recognition*, (2022).

23. F. Zeng, B. Dong, Y. Zhang, T. Wang, X. Zhang, and Y. Wei, "Motr: End-to-End Multiple-Object Tracking With Transformer," in *European Conference on Computer Vision* (Springer, 2022), 659–675.

24. Y. Zhang, T. Wang, and X. Zhang, "Motrv2: Bootstrapping End-to-End Multi-Object Tracking by Pretrained Object Detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 22056–22065.

25. E. Yu, T. Wang, Z. Li, Y. Zhang, X. Zhang, and W. Tao, "Motrv3: Release-Fetch Supervision for End-to-End Multi-Object Tracking," *arXiv preprint arXiv:2305.14298* (2023).

26. R. Gao and L. M. Wang, "Long-Term Memory-Augmented Transformer for Multi-Object Tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2023), 9901–9910.

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

27. C. Xiao, Q. Cao, Z. Luo, and L. M. Lan, "A Simple Baseline for Multiple Object Tracking With State Space Model," in *Proceedings of the 32nd ACM International Conference on Multimedia*, (2024).

28. A. W. Harley, Y. You, X. Sun, et al., "Alltracker: Efficient Dense Point Tracking at High Resolution," *arXiv preprint arXiv:2506.07310* (2025).

29. J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, and F. Zheng, "Track Anything: Segment Anything Meets Videos," *arXiv preprint arXiv: 2304.11968* (2023).

30. A. W. Harley, Z. Fang, and K. Fragkiadaki, "Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories," in *European Conference on Computer Vision* (Springer, 2022), 59–75.

31. N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht, "Cotracker: It Is Better to Track Together," in *European Conference on Computer Vision* (Springer, 2024), 18–35.

32. C. Y. Hsieh, T. Khurana, A. Dave, and D. Ramanan, "Tracking Any Object Amodally," *CoRR* (2023).

33. L. Yang, Y. Fan, and N. Xu, "Video Instance Segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019).

34. S. W. Oh, J. Y. Lee, N. Xu, and S. J. Kim, "Video Object Segmentation Using Space-Time Memory Networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019).

35. Z. Yang, J. Miao, Y. Wang, Y. Wei, and Y. Yang, "Associating Objects With Transformers for Video Object Segmentation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021).

36. H. K. Cheng, Y. W. Tai, and C. K. X. Tang, "Long-Term Video Object Segmentation With an Atkinson-Shiffrin Memory Model," in *Proceedings of the European Conference on Computer Vision* (ECCV, 2022).

37. N. Carion, L. Gustafson, Y. T. Hu, et al., "Sam 3: Segment Anything With Concepts," *arXiv preprint arXiv:2511.16719* (2025).

38. B. S. Krishnan, L. R. Jones, J. A. Elmore, et al., "Fusion of Visible and Thermal Images Improves Automated Detection and Classification of Animals for Drone Surveys," *Scientific Reports* 13, no. 1 (2023): 10385, https://doi.org/10.1038/s41598-023-37295-7.

39. B. Koger, A. Deshpande, J. T. Kerby, J. M. Graving, B. R. Costelloe, and I. D. Couzin, "Quantifying the Movement, Behaviour and Environmental Context of Group-Living Animals Using Drones and Computer Vision," *Journal of Animal Ecology* 92, no. 7 (2023): 1357–1371, https://doi.org/10.1111/1365-2656.13904.

40. N. N. Dat, T. Richardson, M. Watson, et al., "Wildlive: Near Real-Time Visual Wildlife Tracking Onboard Uavs," *arXiv preprint arXiv: 2504.10165* (2025).

41. A. K. Upadhyay, E. Nepovinnykh, S. M. Rayeed, et al., "Animal Re-Identification via Multiview Spatio-Temporal Track Clustering," in *Computer Vision for Animal Behavior Tracking and Modeling Workshop, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2025).

42. H. Naik, J. Yang, D. Das, M. Crofoot, A. Rathore, and V. H. Sridhar, "Bucktales: A Multi-uav Dataset for Multi-Object Tracking and Re-Identification of Wild Antelopes," *Advances in Neural Information Processing Systems* 37 (2024): 81992–82009, https://doi.org/10.52202/079017-2606.

43. Y. Zhu, Q. Wang, C. Li, J. Tang, C. Gu, and Z. Huang, "Visible–Thermal Multiple Object Tracking: Large-Scale Video Dataset and Progressive Fusion Approach," *Pattern Recognition* 161 (2025): 111330, https://doi.org/10.1016/j.patcog.2024.111330.

44. S. Ma, Y. Yang, and G. Chen, "Aodimp-Tir: Anti-Occlusion Thermal Infrared Targets Tracker Based on Superdimp," *IET Image Processing* 18, no. 7 (2024): 1780–1795, https://doi.org/10.1049/ipr2.13063.

45. Q. Liu, X. Li, Z. He, N. Fan, D. Yuan, and H. Wang, "Learning Deep Multi-Level Similarity for Thermal Infrared Object Tracking," *IEEE Transactions on Multimedia* 23 (2020): 2114–2126, https://doi.org/10.1109/tmm.2020.3008028.

46. E. Bondi, R. Jain, P. Aggrawal, et al., "Birdsai: A Dataset for Detection and Tracking in Aerial Thermal Infrared Videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, (2020), 1747–1756.

47. C. Praschl, L. Böss, K. Probst, and D. C. Schedl, "Towards a Multispectral Airborne Light Field Dataset of Forest Animals," in *Proceedings of the 3rd International Workshop on Camera Traps, AI and Ecology* (University of Jena, 2023).

48. F. Reimoser and S. Reimoser, "Ungulates and Their Management in Austria," in *European Ungulates and Their Management in the 21st Century*, 2010), 338–356.

49. C. Praschl and D. C. Schedl, "Towards an Automated Biodiversity Modelling Process for Forest Animals Using Uncrewed Aerial Vehicles," in *Proceedings of the 11th International Workshop on Simulation for Energy* (Sustainable Development & Environment, 2023).

50. C. Mekik and M. Arslanoglu, "Investigation on Accuracies of Real Time Kinematic Gps for Gis Applications," *Remote Sensing* 1, no. 1 (2009): 22–35, https://doi.org/10.3390/rs1010022.

51. M. N. Hussein, R. Abdulla, "Localization Using Gps Coordinates in ipv6 Addresses of Wireless Sensor Network Nodes," *Indian Journal of Science and Technology* (2016).

52. C. Praschl, L. Böss, and D. C. Schedl, "Reconstructionless Airborne Radiance Fields," in *ACM SIGGRAPH 2024 Posterspp*, (2024), 1–2.

53. S. Schramm, J. Ebert, J. Rangel, R. Schmoll, and A. Kroll, "Iterative Feature Detection of a Coded Checkerboard Target for the Geometric Calibration of Infrared Cameras," *Journal of Sensors and Sensor Systems* 10, no. 2 (2021): 207–218, https://doi.org/10.5194/jsss-10-207-2021.

54. G. Jocher and J. Qiu, Ultralytics yolo11, (2024), https://github.com/ultralytics/ultralytics.

55. A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple Online and Realtime Tracking," in *2016 IEEE International Conference on Image Processing* (ICIP), 2016), 3464–3468.

56. K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Learning Generalisable Omni-Scale Representations for Person Re-Identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, no. 9 (2021): 5056–5069, https://doi.org/10.1109/tpami.2021.3069237.

57. J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, "Observation-Centric Sort: Rethinking Sort for Robust Multi-Object Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 9686–9696.

58. P. Sun, J. Cao, Y. Jiang, et al., "Dancetrack: Multi-Object Tracking in Uniform Appearance and Diverse Motion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 20993–21002.

59. A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "Mot16: A Benchmark for Multi-Object Tracking," *arXiv preprint arXiv: 1603.00831* (2016).

60. J. Valmadre, A. Bewley, J. Huang, C. Sun, C. Sminchisescu, and C. Schmid, "Local Metrics for Multi-Object Tracking," *arXiv preprint arXiv: 2104.02631* (2021).

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Appendix

Key Point Matching

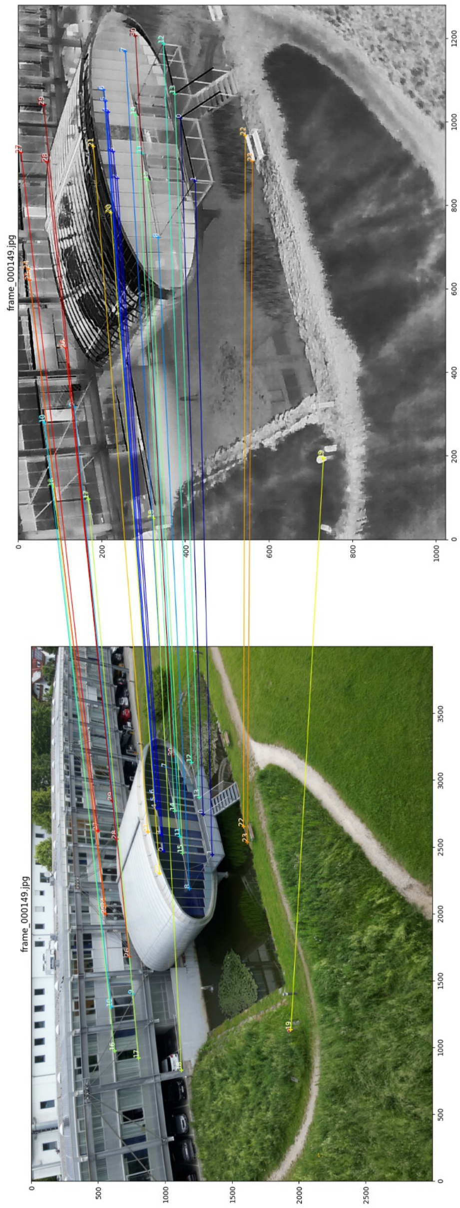


FIGURE A1 | One example image pair with manually annotated keypoints for generating homography and camera intrinsics.

YOLO11x Training Dynamics

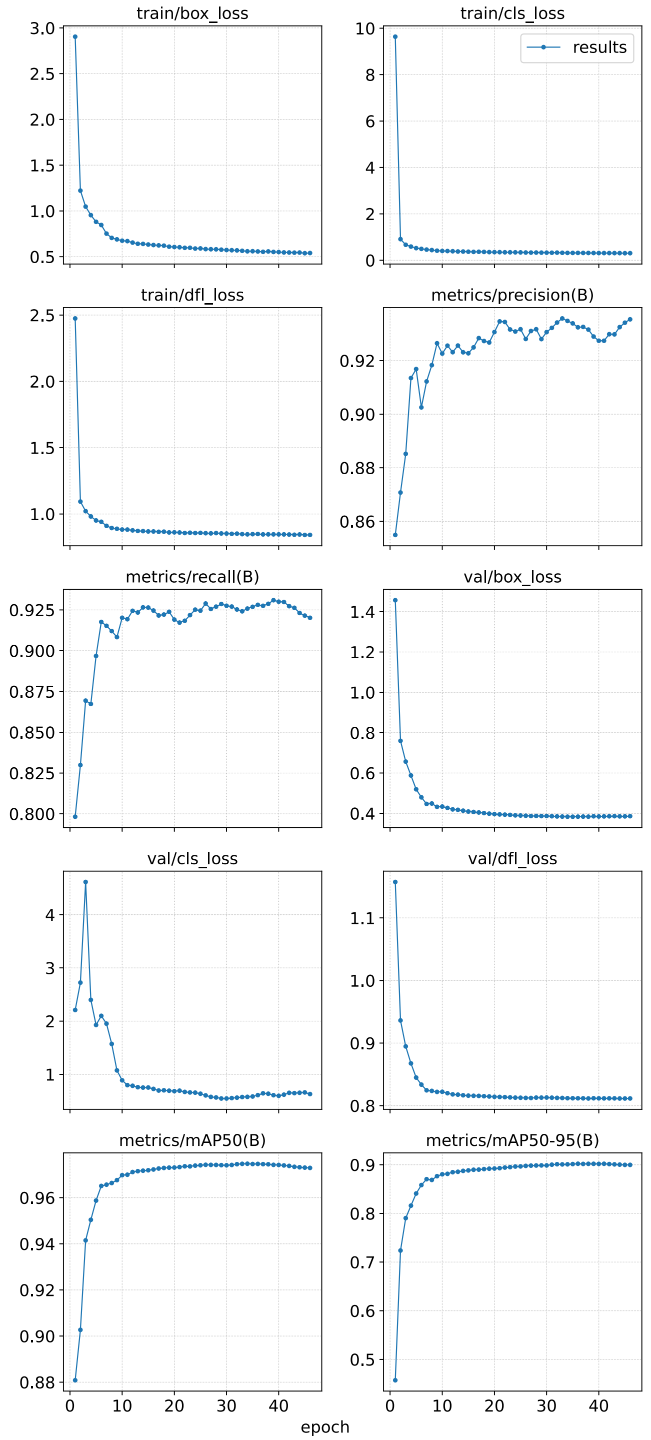


FIGURE A2 | Training dynamics of the YOLO11x class-agnostic wildlife detector on the thermal dataset. The plots illustrate the evolution of training and validation losses (box, classification, and distribution-focal loss) as well as the key performance metrics (precision, recall, mAP50, and mAP50–95) over 50 epochs.

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Long-Term Occlusion Examples

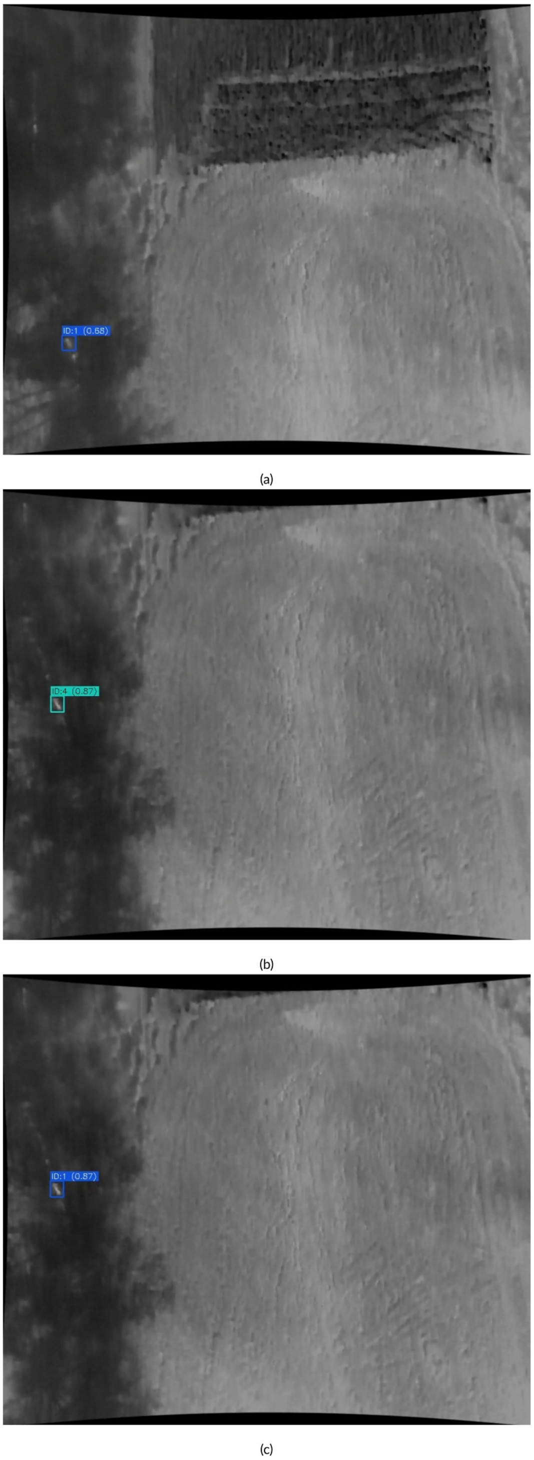


FIGURE A3 | Example of a long-term occlusion: (a) shows the start frame, while (b) and (c) depict another viewpoint after approximately 4 seconds of flight. In (b), Deep OC-SORT fails to maintain the track of the animal with ID 1, whereas in (c), our Geo-Hybrid approach (without custom embeddings) successfully continues the tracking. Sequence ID 294_3 in our MOT dataset.

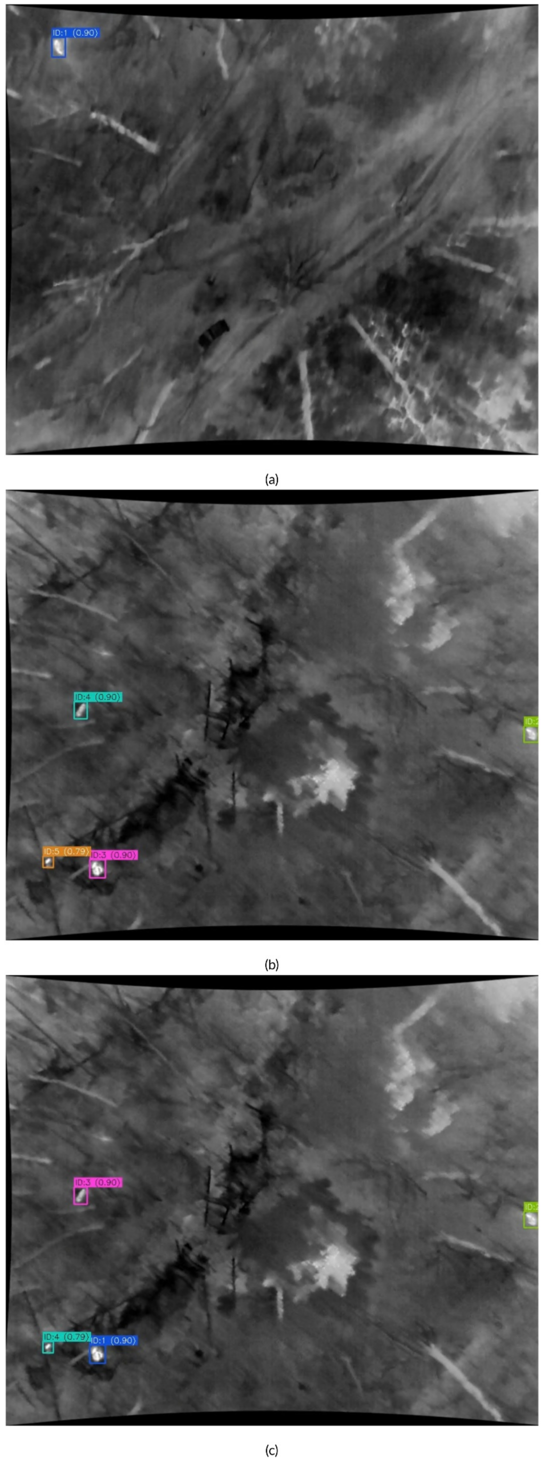


FIGURE A4 | Example of a long-term occlusion: (a) shows the start frame, while (b) and (c) depict another viewpoint after approximately 10 seconds of flight. In (b), Deep OC-SORT fails to maintain the track of the animal with ID 1, whereas in (c) our Geo-Hybrid approach (without custom embeddings) successfully continue the tracking. Sequence ID 243_6 in our MOT dataset.

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

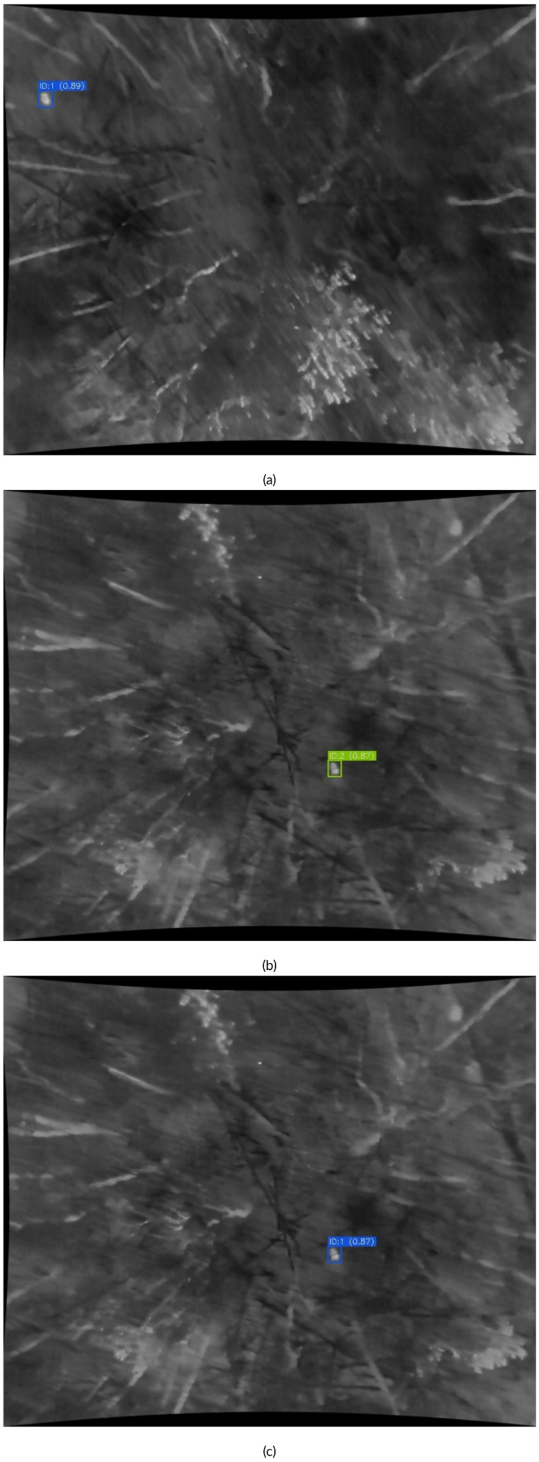


FIGURE A5 | Example of a long-term occlusion: (a) shows the start frame, while (b) and (c) depict another viewpoint after approximately 5 seconds of flight. In (b), Deep OC-SORT fails to maintain the track of the animal with ID 1, whereas in (c) our Geo-Hybrid approach (without custom embeddings) successfully continue the tracking. Sequence ID 241_1 in our MOT dataset.

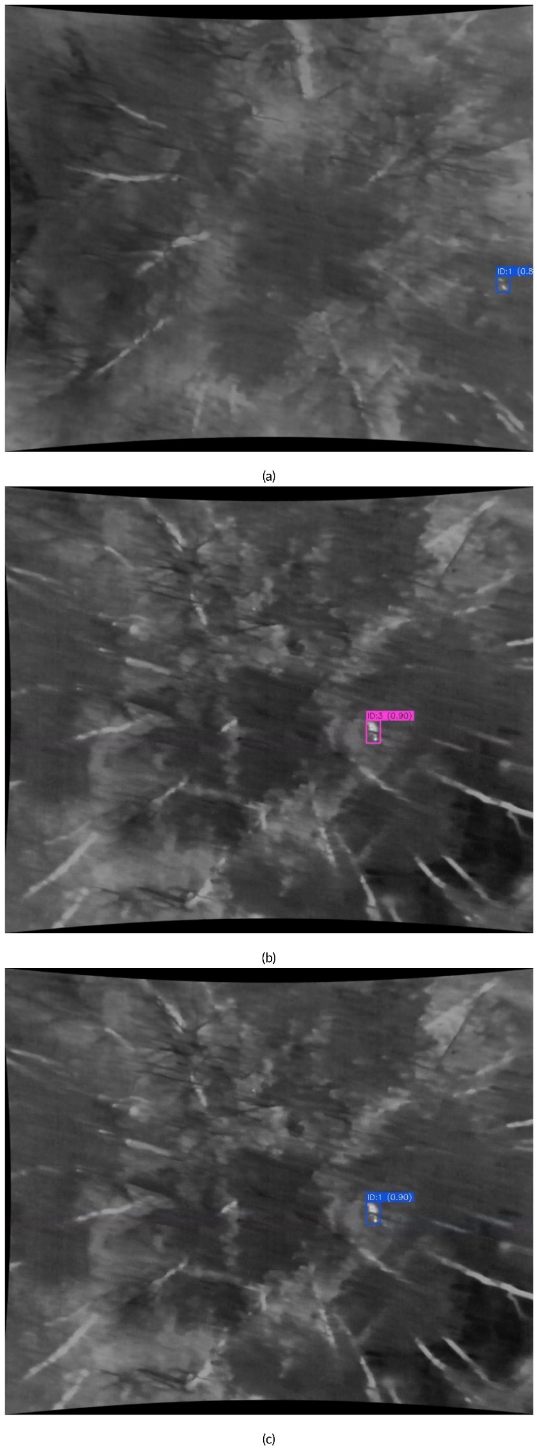


FIGURE A6 | Example of a long-term occlusion: (a) shows the start frame, while (b) and (c) depict another viewpoint after approximately 4 seconds of flight. In (b), Deep OC-SORT fails to maintain the track of the animal with ID 1, whereas in (c) our Geo-Hybrid approach (without custom embeddings) successfully continue the tracking. Sequence ID 238_2 in our MOT dataset.

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

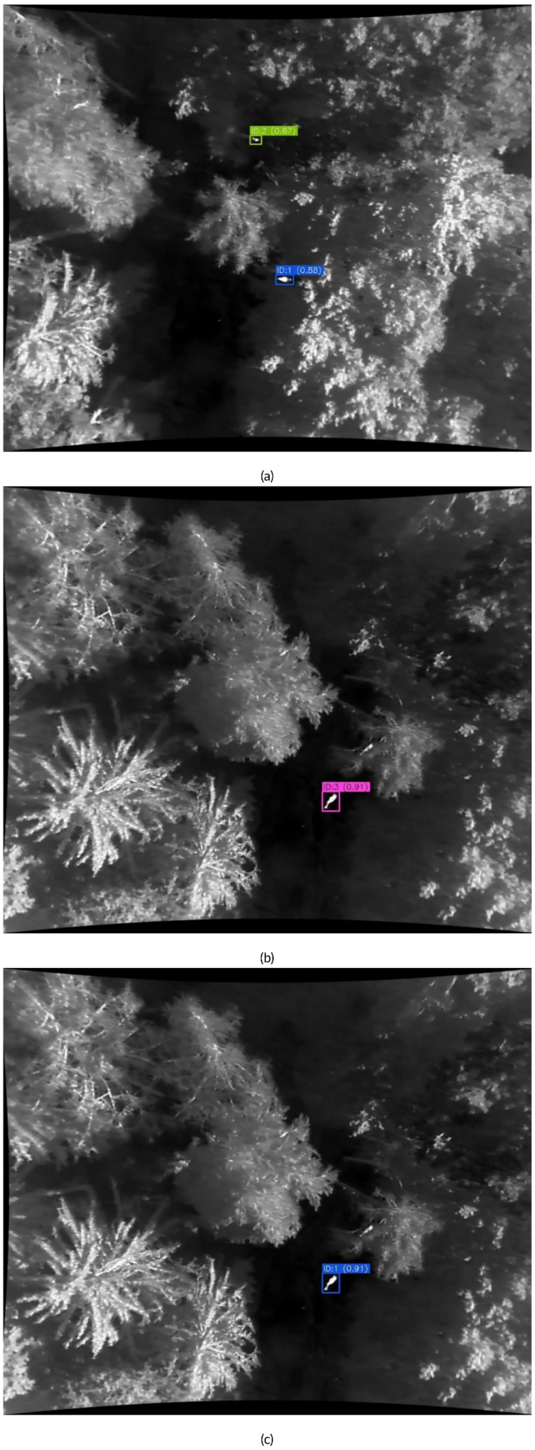


FIGURE A7 | Example of a long-term occlusion: (a) shows the start frame, while (b) and (c) depict another viewpoint after approximately 2 seconds of flight. In (b), Deep OC-SORT fails to maintain the track of the animal with ID 1, whereas in (c) our Geo-Hybrid approach (without custom embeddings) successfully continue the tracking. Sequence ID 211_7 in our MOT dataset.

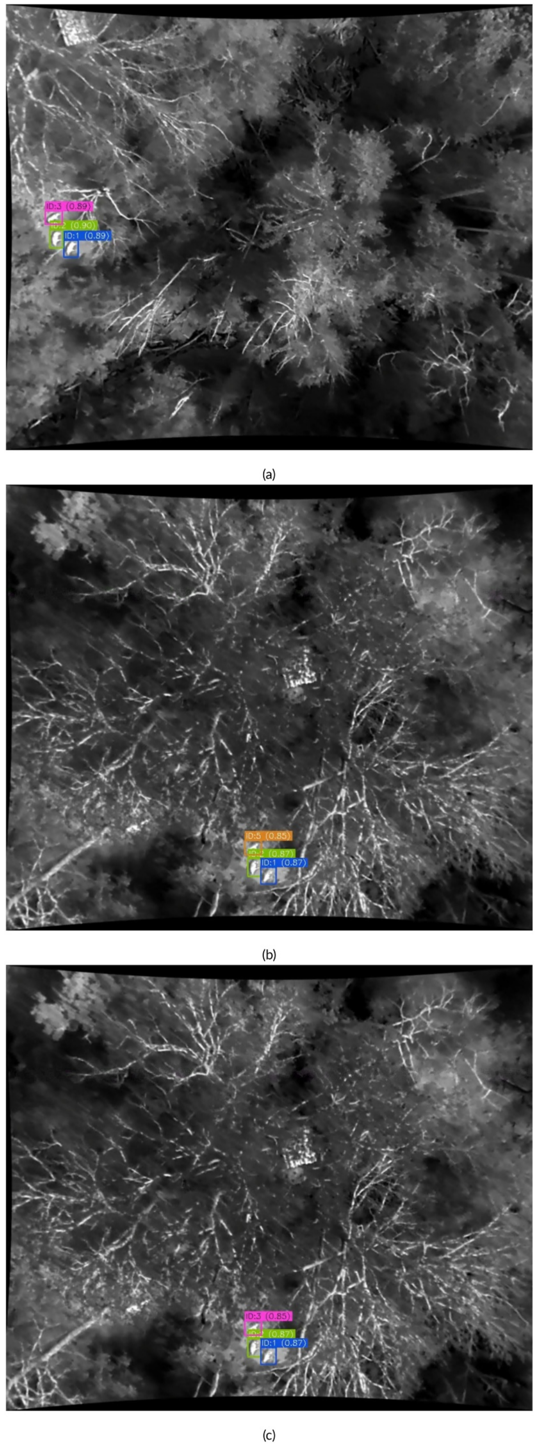


FIGURE A8 | Example of a long-term occlusion: (a) shows the start frame, while (b) and (c) depict another viewpoint after approximately 8 seconds of flight. In (b), Deep OC-SORT fails to maintain the track of the animal with ID 3, whereas in (c) our Geo-Hybrid approach (without custom embeddings) successfully continues the tracking. Sequence ID 198_2 in our MOT dataset.

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

SAM3.1 Tracking Results

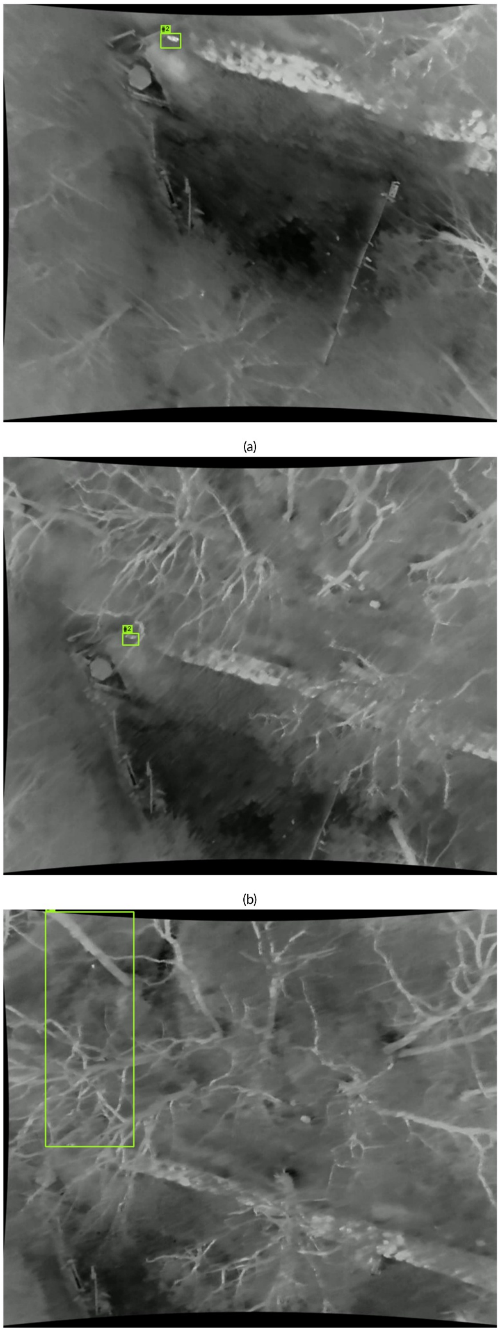


FIGURE A9 | Qualitative tracking example with SAM 3.1. (a) Shows accurate localisation from an initial box prompt. In (b), partial occlusions occur, but tracking remains consistent. In (c), tracking fails: the mask drifts and the bounding box expands excessively. This highlights sensitivity to occlusions and limited robustness in out-of-distribution settings like thermal imagery. Sequence ID 100_2 in our MOT dataset.

Extended Results

TABLE A1 | Extended tracking results across the train and test split with BoxMOT's default embedding-based Re-ID model "osnet_ain_x1_0.pt", where applicable (*).

Model	Split	AssA ↑	IDF1 ↑	IDSW ↓	IDP ↑	IDR ↑	IDTP ↑	IDFN ↓	IDFP ↓	Frag ↓	MT ↑	PT	ML ↓	HOTA ↑	DetA ↑	DetRe ↑	DetPr ↑	AssRe ↑	AssPr ↑	LocA ↑
StrongSort*	Test	55.324	51.171	509	61.732	43.695	97575	125733	60487	3034	287	256	400	42.679	33.864	39.323	55.555	59.195	74.505	72.534
	Train	48.565	43.167	1989	44.922	41.545	280997	395375	344526	11098	589	871	974	40.276	35.602	43.904	47.473	53.972	67.687	68.620
ByteTrack	Test	54.992	51.701	289	61.729	44.476	99318	123990	61576	2779	291	260	392	42.806	34.137	39.851	55.310	59.336	72.925	72.578
	Train	48.996	43.867	952	45.271	42.548	287783	388589	347902	9319	584	908	942	40.739	35.743	44.373	47.214	54.416	66.540	68.672
BoT-SORT*	Test	56.580	52.283	264	62.613	44.878	100217	123091	59841	2792	298	248	397	43.348	34.132	39.744	55.449	60.820	73.818	72.549
	Train	51.738	45.140	718	46.642	43.732	295791	380581	338386	9528	604	875	955	41.846	35.822	44.339	47.289	57.352	67.194	68.582
Koger et al. [39]	Test	50.254	46.420	221	64.389	36.292	81042	142266	44821	2092	162	314	467	37.496	28.776	32.244	57.208	54.352	71.860	72.845
	Train	47.670	42.117	628	47.541	37.803	255690	420682	282141	9031	360	901	1173	38.010	32.096	38.305	48.173	53.833	65.147	68.638
MambaTrack	Test	51.697	47.125	135	65.836	36.696	81946	141362	42524	2095	164	306	473	37.973	28.641	32.021	57.447	54.907	75.092	72.910
	Train	47.581	42.139	536	48.223	37.418	253085	423287	271740	8650	360	886	1188	37.704	31.670	37.627	48.492	52.090	68.704	68.777
Sam3.1	Test	43.355	26.911	128	28.920	25.162	56190	167119	138106	5531	124	315	504	30.924	22.830	30.055	34.543	50.483	51.066	64.681
	Train	40.467	22.805	412	23.850	21.848	116890	418132	373214	10322	260	808	1167	28.617	21.052	28.898	31.546	48.463	48.538	64.662
OCSort	Test	52.435	46.185	108	66.933	35.256	78729	144579	38895	1484	158	295	490	37.396	27.338	30.362	57.642	55.720	75.337	72.944
	Train	47.871	41.971	445	48.925	36.748	248550	427822	259469	7739	374	882	1178	37.570	31.261	36.811	49.010	52.378	69.154	68.849

(Continues)

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE A1 | (Continued)

Model	Split	AssA ↑	IDF1 ↑	IDSW ↓	IDP ↑	IDR ↑	IDTP ↑	IDFN ↓	IDFP ↓	Frag ↓	MT ↑	PT	ML ↓	HOTA ↑	DetA ↑	DetRe ↑	DetPr ↑	AssRe ↑	AssPr ↑	LocA ↑
BoostTrack*	Test	50.603	43.896	97	67.099	32.617	72837	150471	35714	1388	124	309	510	35.525	25.554	28.167	57.944	53.600	75.820	73.093
	Train	46.541	40.933	361	51.718	33.870	229086	447286	213864	6223	328	853	1253	35.244	28.144	32.580	49.749	50.867	69.367	69.339
Deep OC-Sort*	Test	56.526	48.806	40	70.121	37.429	83582	139726	35615	1456	175	279	489	39.082	27.741	30.835	57.767	60.013	75.715	72.981
	Train	52.434	44.976	202	51.986	39.632	268057	408315	247572	7842	413	849	1172	39.684	31.716	37.415	49.079	57.583	68.858	68.870
Geo-native* (ours)	Test	52.703	46.225	84	67.790	35.679	79673	143635	37856	1473	158	294	491	37.710	27.314	30.328	57.623	56.590	75.761	72.913
	Train	48.299	42.363	370	49.775	37.245	251916	424456	254189	7716	375	876	1183	37.904	31.171	36.699	49.045	53.759	68.640	68.887
Geo-hybrid* (ours)	Test	56.672	48.799	34	69.860	37.495	83729	139579	36124	1476	177	277	489	39.211	27.855	30.976	57.714	60.336	75.189	72.943
	Train	52.175	44.878	168	51.784	39.598	267828	408544	249371	7885	412	856	1166	39.623	31.783	37.513	49.058	58.011	67.623	68.862

Model	Split	OWTA ↑	HOTA (0) ↑	LocA (0) ↑	HOTALocA (0) ↑	MOTA ↑	MOTP ↑	MODA ↑	CLR Re ↑	CLR Pr ↑	MTR ↑	PTR ↑	MLR ↓	CLR TP ↑	CLR FN ↓	CLR FP ↓	sMOTA ↑
StrongSort*	Test	46.320	67.251	60.798	40.888	24.240	68.959	24.468	47.625	67.284	30.435	27.147	42.418	106350	116958	51712	9.456
	Train	45.312	71.769	53.935	38.709	2.850	64.707	3.145	47.813	51.700	24.199	35.785	40.016	323396	352976	302127	−14.024
ByteTrack	Test	46.552	68.225	60.720	41.426	23.972	69.138	24.101	48.076	66.725	30.859	27.572	41.569	107357	115951	53537	9.135
	Train	45.925	73.997	53.769	39.787	2.265	64.881	2.406	48.195	51.280	23.993	37.305	38.702	325979	350393	309706	−14.660
BoT-SORT*	Test	47.107	68.569	60.767	41.668	24.418	68.989	24.537	48.106	67.116	31.601	26.299	42.100	107425	115883	52633	9.500
	Train	47.121	75.799	53.770	40.757	2.438	64.699	2.544	48.153	51.357	24.815	35.949	39.236	325693	350679	308484	−14.560
Koger et al. [39]	Test	39.939	57.909	61.441	35.580	22.236	69.231	22.335	39.349	69.813	17.179	33.298	49.523	87869	135439	37994	10.128
	Train	41.983	68.202	54.021	36.843	3.984	64.653	4.077	41.797	52.563	14.790	37.017	48.192	282702	393670	255129	−10.790
MambaTrack	Test	40.390	58.461	61.710	36.077	22.580	69.237	22.640	39.190	70.309	17.391	32.450	50.159	87514	135794	36956	10.524
	Train	41.548	66.725	54.426	36.316	5.180	64.716	5.260	41.427	53.389	14.790	36.401	48.809	280200	396172	244625	−9.437
Sam3.1	Test	35.775	71.788	42.561	30.554	−36.025	60.908	−35.968	25.520	29.331	13.150	33.404	53.446	56988	166321	137308	−46.002
	Train	33.841	69.685	40.926	28.520	−47.334	62.577	−47.257	22.174	24.206	11.633	36.152	52.215	118635	416387	371469	−55.632
OCSort	Test	39.625	57.493	61.910	35.594	21.789	69.187	21.837	37.255	70.729	16.755	31.283	51.962	83194	140114	34430	10.309
	Train	41.210	66.031	54.713	36.127	6.174	64.720	6.240	40.675	54.154	15.366	36.237	48.398	275111	401261	232908	−8.176
BoostTrack*	Test	37.485	54.318	62.062	33.711	20.328	69.500	20.371	34.491	70.954	13.150	32.768	54.083	77021	146287	31530	9.808
	Train	38.271	60.579	55.535	33.643	7.846	65.297	7.900	36.694	56.031	13.476	35.045	51.479	248191	428181	194759	−4.888
Deep OC-Sort*	Test	41.440	59.847	61.984	37.096	22.347	69.231	22.365	37.871	70.950	18.558	29.586	51.856	84570	138738	34627	10.695
	Train	43.560	69.874	54.760	38.263	6.539	64.745	6.569	41.402	54.308	16.968	34.881	48.151	280029	396343	235600	−8.057
Geo-native* (ours)	Test	39.956	57.996	61.870	35.882	21.794	69.152	21.831	37.231	70.740	16.755	31.177	52.068	83140	140168	34389	10.309
	Train	41.561	66.658	54.797	36.526	6.503	64.730	6.558	40.692	54.382	15.407	35.990	48.603	275229	401143	230876	−7.849
Geo-hybrid* (ours)	Test	41.589	60.150	61.910	37.239	22.254	69.211	22.269	37.970	70.746	18.770	29.374	51.856	84791	138517	35062	10.563
	Train	43.505	69.880	54.733	38.247	6.458	64.744	6.483	41.475	54.239	16.927	35.168	47.905	280524	395848	236675	−8.164

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE A2 | Extended tracking results across the train and test split with our custom embedding-based Re-ID model, where applicable (*).

Model	Split	AssA ↑	IDF1 ↑	IDSW ↓	IDP ↑	IDR ↑	IDTP ↑	IDFN ↓	IDFP ↓	Frag ↓	MT ↑	PT	ML ↓	HOTA ↑	DetA ↑	DetRe ↑	DetPr ↑	AssRe ↑	AssPr ↑	LocA ↑
StrongSort*	Test	55.310	51.146	507	61.701	43.675	97529	125779	60537	3034	287	256	400	42.673	33.863	39.323	55.553	59.181	74.504	72.534
	Train	48.573	43.174	1986	44.929	41.552	281043	395329	344487	11096	589	871	974	40.280	35.603	43.905	47.473	53.982	67.674	68.620
ByteTrack	Test	54.992	51.701	289	61.729	44.476	99318	123990	61576	2779	291	260	392	42.806	34.137	39.851	55.310	59.336	72.925	72.578
	Train	48.996	43.867	952	45.271	42.548	287783	388589	347902	9319	584	908	942	40.739	35.743	44.373	47.214	54.416	66.540	68.672
BoT-SORT*	Test	56.580	52.283	264	62.613	44.878	100217	123091	59841	2792	298	248	397	43.348	34.132	39.744	55.449	60.820	73.818	72.549
	Train	51.738	45.140	718	46.642	43.732	295791	380581	338386	9528	604	875	955	41.846	35.822	44.339	47.289	57.352	67.194	68.582
Koger et al. [39]	Test	50.254	46.420	221	64.389	36.292	81042	142266	44821	2092	162	314	467	37.496	28.776	32.244	57.208	54.352	71.860	72.845
	Train	47.670	42.117	628	47.541	37.803	255690	420682	282141	9031	360	901	1173	38.010	32.096	38.305	48.173	53.833	65.147	68.638
Mamba Track	Test	51.697	47.125	135	65.836	36.696	81946	141362	42524	2095	164	306	473	37.973	28.641	32.021	57.447	54.907	75.092	72.910
	Train	47.581	42.139	536	48.223	37.418	253085	423287	271740	8650	360	886	1188	37.704	31.670	37.627	48.492	52.090	68.704	68.777
Sam3.1	Test	43.355	26.911	128	28.920	25.162	56190	167119	138106	5531	124	315	504	30.924	22.830	30.055	34.543	50.483	51.066	64.681
	Train	40.467	22.805	412	23.850	21.848	116890	418132	373214	10322	260	808	1167	28.617	21.052	28.898	31.546	48.463	48.538	64.662
OCSort	Test	52.435	46.185	108	66.933	35.256	78729	144579	38895	1484	158	295	490	37.396	27.338	30.362	57.642	55.720	75.337	72.944
	Train	47.871	41.971	445	48.925	36.748	248550	427822	259469	7739	374	882	1178	37.570	31.261	36.811	49.010	52.378	69.154	68.849
BoostTrack*	Test	50.920	43.951	94	67.249	32.642	72893	150415	35500	1362	124	309	510	35.599	25.507	28.113	57.917	53.718	76.566	73.079
	Train	46.754	40.863	350	51.714	33.776	228451	447921	213305	6227	313	869	1252	35.243	28.042	32.457	49.695	50.799	70.395	69.323
Deep OC-Sort*	Test	55.349	48.033	53	68.601	36.954	82521	140787	37770	1489	178	281	484	38.811	27.918	31.065	57.670	60.286	72.569	72.939
	Train	52.452	44.994	185	51.998	39.653	268205	408167	247596	7847	413	849	1172	39.707	31.737	37.439	49.093	57.709	68.677	68.875
Geo-native* (ours)	Test	56.666	48.694	33	69.889	37.548	83847	139461	36125	1487	177	279	487	39.264	27.871	30.998	57.697	60.443	75.179	72.935
	Train	51.876	44.825	167	51.774	39.587	267754	408618	249407	7885	411	858	1165	39.617	31.780	37.509	49.057	57.999	67.618	68.862
Geo-hybrid* (ours)	Test	56.714	48.841	30	69.975	37.537	83824	139484	35967	1472	177	279	487	39.277	27.836	30.954	57.704	60.381	75.633	72.937
	Train	52.191	44.889	162	51.791	39.602	267859	408513	249329	7890	412	856	1166	39.628	31.788	37.517	49.064	58.020	67.627	68.863

Model	Split	OWTA ↑	HOTA (0) ↑	LocA (0) ↑	HOTALocA (0) ↑	MOTA ↑	MOTP ↑	MODA ↑	CLR Re ↑	CLR Pr ↑	MTR ↑	PTR ↑	MLR ↓	CLR TP ↑	CLR FN ↓	CLR FP ↓	sMOTA ↑
StrongSort*	Test	46.314	67.241	60.798	40.881	24.239	68.958	24.466	47.625	67.282	30.435	27.147	42.418	106350	116958	51716	9.455
	Train	45.316	71.781	53.936	38.716	2.850	64.707	3.144	47.813	51.700	24.199	35.785	40.016	323397	352975	302133	−14.025
ByteTrack	Test	46.552	68.225	60.720	41.426	23.972	69.138	24.101	48.076	66.725	30.859	27.572	41.569	107357	115951	53537	9.135
	Train	45.925	73.997	53.769	39.787	2.265	64.881	2.406	48.195	51.280	23.993	37.305	38.702	325979	350393	309706	−14.660
BoT-SORT*	Test	47.107	68.569	60.767	41.668	24.418	68.989	24.537	48.106	67.116	31.601	26.299	42.100	107425	115883	52633	9.500
	Train	47.121	75.799	53.770	40.757	2.438	64.700	2.544	48.153	51.357	24.815	35.949	39.236	325693	350679	308484	−14.560
Koger et al. [39]	Test	39.939	57.909	61.441	35.580	22.236	69.231	22.335	39.349	69.813	17.179	33.298	49.523	87869	135439	37994	10.128
	Train	41.983	68.202	54.021	36.843	3.984	64.653	4.077	41.797	52.563	14.790	37.017	48.192	282702	393670	255129	−10.790

(Continues)

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE A2 | (Continued)

Model	Split	OWTA ↑	HOTA (0) ↑	LocA (0) ↑	HOTALocA (0) ↑	MOTA ↑	MOTP ↑	MODA ↑	CLR Re ↑	CLR Pr ↑	MTR ↑	PTR ↑	MLR ↓	CLR TP ↑	CLR FN ↓	CLR FP ↓	sMOTA ↑
MambaTrack	Test	40.390	58.461	61.710	36.077	22.580	69.237	22.640	39.190	70.309	17.391	32.450	50.159	87514	135794	36956	10.524
	Train	41.548	66.725	54.426	36.316	5.180	64.716	5.260	41.427	53.389	14.790	36.401	48.809	280200	396172	244625	−9.437
Sam3.1	Test	35.775	71.788	42.561	30.554	−36.025	60.908	−35.968	25.520	29.331	13.150	33.404	53.446	56988	166321	137308	−46.002
	Train	33.841	69.685	40.926	28.520	−47.334	62.577	−47.257	22.174	24.206	11.633	36.152	52.215	118635	416387	371469	−55.632
OCSort	Test	39.625	57.493	61.910	35.594	21.789	69.187	21.837	37.255	70.729	16.755	31.283	51.962	83194	140114	34430	10.309
	Train	41.210	66.031	54.713	36.127	6.174	64.720	6.240	40.675	54.154	15.366	36.237	48.398	275111	401261	232908	−8.176
BoostTrack*	Test	37.563	54.419	62.038	33.761	20.260	69.485	20.302	34.421	70.913	13.150	32.768	54.083	76865	146443	31528	9.757
	Train	38.269	60.570	55.499	33.615	7.705	65.278	7.757	36.535	55.938	12.859	35.703	51.438	247110	429262	194646	−4.981
Deep OC-Sort*	Test	41.174	59.621	61.902	36.907	22.266	69.205	22.290	38.079	70.689	18.876	29.799	51.326	85033	138275	35258	10.540
	Train	43.584	69.918	54.775	38.298	6.556	64.748	6.583	41.422	54.316	16.968	34.881	48.151	280165	396207	235636	−8.046
Geo-native* (ours)	Test	41.648	60.299	61.893	37.321	22.244	69.202	22.259	37.992	70.716	18.770	29.586	51.644	84839	138469	35133	10.543
	Train	43.500	69.870	54.732	38.241	6.458	64.744	6.483	41.472	54.239	16.886	35.251	47.864	280504	395868	236657	−8.164
Geo-hybrid* (ours)	Test	41.654	60.375	61.904	37.375	22.233	69.205	22.247	37.945	70.736	18.770	29.586	51.644	84735	138573	35056	10.548
	Train	43.510	69.881	54.739	38.252	6.455	64.745	6.479	41.472	54.237	16.927	35.168	47.905	280506	395866	236682	−8.166

Geo-Referenced Segmentations

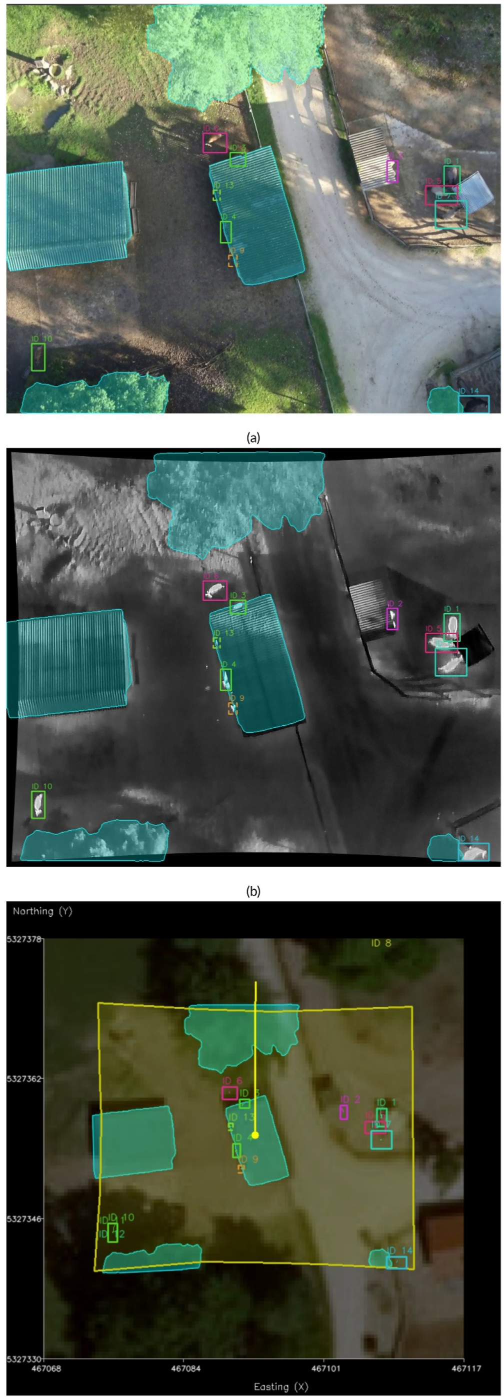


FIGURE A10 | Geo-referenced multi-modal visualisation of detections and segmentations. (a) RGB image with SAM3-based semantic segmentations of trees and buildings. (b) Corresponding thermal image, where the segmentation masks are projected using the calibrated RGB–thermal camera model. (c) Global map view showing all geo-referenced bounding boxes, semantic masks, and the thermal camera's field of view (yellow area). Sequence ID 14_1 in our MOT dataset.

17519640, 2026, 1, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/cvi2.70077 by Universiteit Hasselt, Wiley Online Library on [19/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License