Received: 22 January 2025 | Accepted: 14 July 2025

DOI: 10.1111/2041-210X.70116

Methods in Ecology and Evolution

RESEARCH ARTICLE

Zero-shot shark tracking and biometrics from aerial imagery

Chinmay K. Lalgudi[1] | **Mark E. Leone**[1] | **Jaden V. Clark**[1] | **Sergio Madrigal-Mora**[2] | **Mario Espinoza**[3]

[1]Stanford University, Stanford, California, USA

[2]Flinders University, Adelaide, South Australia, Australia

[3]Centro de Investigación en Ciencias del Mar y Limnología, Universidad de Costa Rica, San José, Costa Rica

Correspondence
Chinmay K. Lalgudi
Email: clalgudi@stanford.edu

Funding information
Save Our Seas Foundation, Grant/Award Number: 664; American Elasmobranch Society; Mel Lane Student Grants Program from the Stanford Woods Institute for the Environment; Richard D. Green Graduate Fellowship

Handling Editor: Ming Bai

Abstract

1. The recent widespread adoption of drones for studying marine animals provides opportunities for deriving biological information from aerial imagery. The large scale of imagery data acquired from drones is well suited for machine learning (ML) analysis. Development of ML models for analysing marine animal aerial imagery has followed the classical paradigm of training, testing and deploying a new model for each dataset, requiring significant time, human effort and ML expertise.
2. We introduce Frame-Level Alignment and Tracking (FLAIR), which leverages the video understanding of Segment Anything Model 2 (SAM 2) and the vision-language capabilities of Contrastive Language-Image Pre-training (CLIP). FLAIR takes a drone video as input and outputs segmentation masks of the species of interest across the video. Notably, FLAIR leverages a *zero-shot* approach, eliminating the need for labelled data, training a new model or fine-tuning an existing model to generalize to other species.
3. We trained state-of-the-art object detection and instance segmentation models on a new dataset of Pacific nurse sharks. We show that FLAIR massively outperforms these methods and performs competitively against two human-in-the-loop approaches for prompting SAM 2, achieving a Dice score of 0.8. FLAIR readily generalizes to other shark species without additional human effort and can be combined with custom heuristics to automatically extract relevant information including length and tailbeat frequency.
4. FLAIR has significant potential to accelerate aerial imagery analyses, requiring markedly less human effort and expertise than traditional machine learning workflows, while achieving superior accuracy and generalization performance. By reducing the effort required for aerial imagery analysis, FLAIR allows scientists to spend more time interpreting results and deriving insights about marine ecosystems.

KEYWORDS
deep learning, drones in ecology, elasmobranchs, foundation models, marine biometrics, visual tracking

1 | INTRODUCTION

Large marine animals, such as sharks, influence ecosystems through a variety of mechanisms as predators, nutrient transporters and even prey, thus maintaining the balance of marine food webs, regulating species populations and promoting biodiversity (Dedman et al., 2024; Heupel et al., 2014). Overfishing and other anthropogenic threats have greatly reduced shark populations, altering their ecological roles and effects on ecosystems (Ferretti et al., 2010; Stevens et al., 2000). Long-term monitoring of sharks and other large marine animals is key to understanding how animal populations are responding to human impacts and potential environmental shifts caused by climate change (Torres et al., 2022). However, studying sharks is challenging due to their elusive nature, vast ranges and the inherent difficulties in observing their interactions in marine environments (Jorgensen et al., 2022). Thus, there is a need for new technologies to enable further understanding of both their large-scale and precise movement ecology.

Archival tags and acoustic telemetry are commonly used to study shark behaviour across multiple spatial and temporal scales, but tagging requires significant human effort and may be detrimental to the well-being of tagged animals (Kohler & Turner, 2001; Matley et al., 2024). Baited Remote Underwater Video Stations (BRUVS) are another widely adopted method for obtaining data on shark presence and interactions (White et al., 2013). However, BRUVS are limited to capturing localized behaviour and are dependent on sharks maintaining proximity to the deployment site.

The use of unmanned aerial vehicle (UAV) systems is emerging as a promising approach for the non-invasive study of volitional marine animal behaviour and biometrics (Gray, Bierlich, et al., 2019; Hodgson et al., 2013; Ramos et al., 2022; Torres et al., 2022; Torres & Bierlich, 2020). Aerial imagery can be used to compute biometrics such as length, body condition, tailbeat frequency and relative velocities, which can provide key information about animal health (Bierlich et al., 2024), swimming kinematics (DiGiacomo et al., 2023; Porter et al., 2020) and predator–prey interactions (Hansen et al., 2022).

1.1 | Deep learning for marine ecology

Deep learning, the process of training large artificial neural networks to learn complex functions from data, has important applications to the study of aerial wildlife imagery (LeCun et al., 2015). Previous works have used deep learning for the analysis of aerial imagery of marine animals; however, they typically rely on specialized object detection models (Eikelboom et al., 2019; Gray, Bierlich, et al., 2019; Sharma et al., 2018).

Traditionally, object detection models have relied on Convolutional Neural Networks (CNN) architectures and their variants (Alzubaidi et al., 2021; Girshick, 2015; Li et al., 2021; Ren et al., 2015). Perhaps the most popular object detection model with a CNN backbone, You Only Look Once (YOLO) (Jocher et al., 2023), is specialized for inference speed, treating object detection as a regression problem. By predicting classes and bounding boxes in a single pass, YOLO models are suitable for real-time applications. More recently, the Detection Transformer (DETR) (Carion et al., 2020) has achieved state-of-the-art results by integrating transformers into the encoder and decoder of the model. This eliminates the need for many hand-engineered components by doing direct set prediction of object classes and bounding boxes. DETR leverages this global context to achieve impressive results that rival non-transformer architectures. This is particularly beneficial in scenarios discussed in this work, where multiple objects are in close proximity, such as groups of sharks, or partially occluded by factors like glare or high turbidity.

Several studies have proposed using object detection models to track sharks in aerial imagery (Clark et al., 2025; Gorkin III et al., 2020; Sharma et al., 2018). One of the first works utilizing neural networks to analyse aerial imagery of marine life trained a vanilla CNN for sea turtle detection (Gray, Fleishman, et al., 2019) and other works have used transfer learning (fine-tuning a pre-trained CNN) to improve model performance (Desgarnier et al., 2022; Gray, Bierlich, et al., 2019; Sharma et al., 2018). Other studies have trained segmentation models to identify marine animals and extract biometrics from UAV imagery (Bagchi et al., 2025). DeepLabCut, a deep neural network-based framework for markerless pose estimation on animal videos, is an alternate approach for extracting biometrics from UAV imagery (Mathis et al., 2018). DeepLabCut has been used to estimate biometrics of white sharks (including length and tail beat frequency) from UAV videos (DiGiacomo et al., 2023).

Standard object detection and segmentation models require large datasets of high-quality human-annotated data to train models. Furthermore, they often do not generalize well, performing poorly when the inference data distribution differs from the model's training data (Koh et al., 2021). In contrast to conventional approaches, we leverage foundation models (i.e. large deep learning models trained on internet-scale datasets) for marine animal tracking and biometric analysis from aerial imagery (Bommasani et al., 2021). The key advantage of using pre-trained foundation models is that they can be deployed *zero-shot*. That is, they do not require dataset curation or training to adapt to new data and require significantly less human effort and expertise to use. There is a notable lack of adoption of foundation models for the study of marine animals from UAV imagery, but recent works have used Segment Anything Model for surveying biometrics of whales and sharks from UAV imagery (Bagchi et al., 2025; Bierlich et al., 2024; Clark et al., 2025; Kirillov et al., 2023).

In this work, we explore methods for automatically computing segmentation masks and downstream biometrics for sharks using Segment Anything Model 2 (SAM 2), a pre-trained foundation model for promptable image and video segmentation (Ravi et al., 2024) and Contrastive Language-Image-Pretraining (CLIP), an approach for learning shared representations between natural language and pixels (Radford et al., 2021). We present a new method, Frame Level Alignment and Tracking (FLAIR), that uses SAM 2 and CLIP

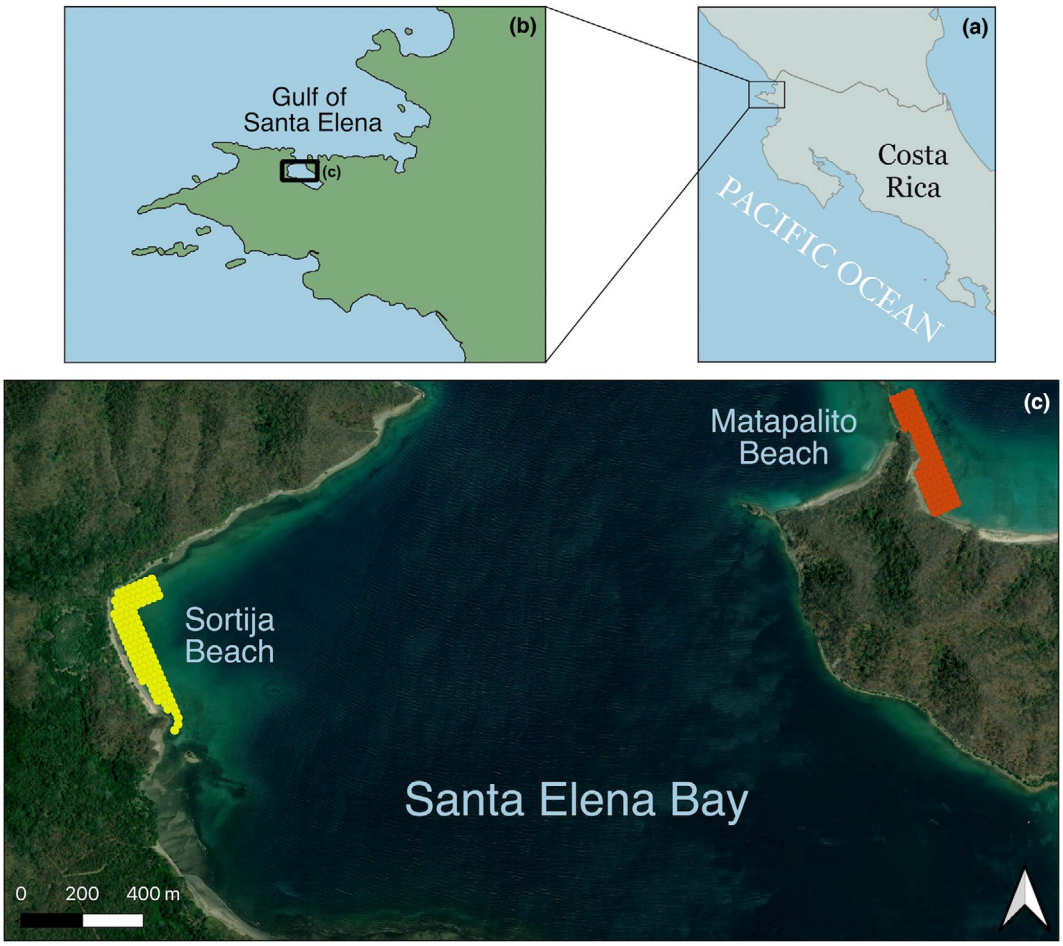

FIGURE 1 A map of the sites where drone video data for the Pacific nurse shark dataset was collected in Santa Elena Bay, Guanacaste, Costa Rica (a). Inset map in (b) shows the Santa Elena Bay study area (c), in which the yellow and orange dots represent the pre-planned flight path locations at Sortija Beach and Matapalito Beach, respectively.

to generate accurate segmentations for several shark species across different environments, leveraging a zero-shot approach that eliminates the need for annotating data, training or fine-tuning. These segmentation masks can be used to compute tailbeat frequency, length, mass, velocities and other downstream biometrics for arbitrary shark species.

We test our approach predominantly on a dataset of the Pacific nurse shark (*Ginglymostoma unami*) from Santa Elena Bay (North Pacific coast of Costa Rica), demonstrating how our method can be used to help better understand the movement ecology of an endangered and understudied species (Madrigal-Mora et al., 2024). We compare segmentation accuracy of FLAIR with multiple approaches, including prompting SAM 2 with a human in the loop, as well as prompting SAM 2 with state-of-the-art object detection models.

Our study suggests that FLAIR is capable of generalizing to other species and we show that FLAIR segmentations can be used to measure biometrics including length and tailbeat frequency. Notably, FLAIR does not require any training or fine-tuning to generalize to other species, highlighting its potential applicability across diverse ecosystems.

2 | MATERIALS AND METHODS

2.1 | Dataset

Drone videos of Pacific nurse sharks were filmed at two field sites (Matapalito Beach and Sortija Beach) in the coastal waters of the Eastern Tropical Pacific Ocean, in Santa Elena Bay, Costa Rica (Figure 1a,b). Data were collected from 2022 to 2024, over a period of 23 months, with varying water visibility (turbidity), illumination and wind/wave conditions (Figure 2a). Images were collected at each site by flying a DJI Mavic 2 drone on a pre-programmed path, recording a continuous video at 30FPS and 3840×2160 resolution. The drone stopped at predetermined waypoints for 3 s each (Figure 1c, yellow and orange dots). Fieldwork was conducted under Costa Rican research permits ACG-PI-021-2017 and ACT-OR-DR-068-18, and all animal procedures were approved by the California State University, Long Beach IACUC (project C1127, protocol 36607066).

More than 6 h of video was recorded in total, during 60 drone surveys, resulting in 648,000 total frames captured. To our knowledge, this is one of the largest open-access datasets of nearshore shark aerial drone imagery. This nurse shark dataset was pruned to

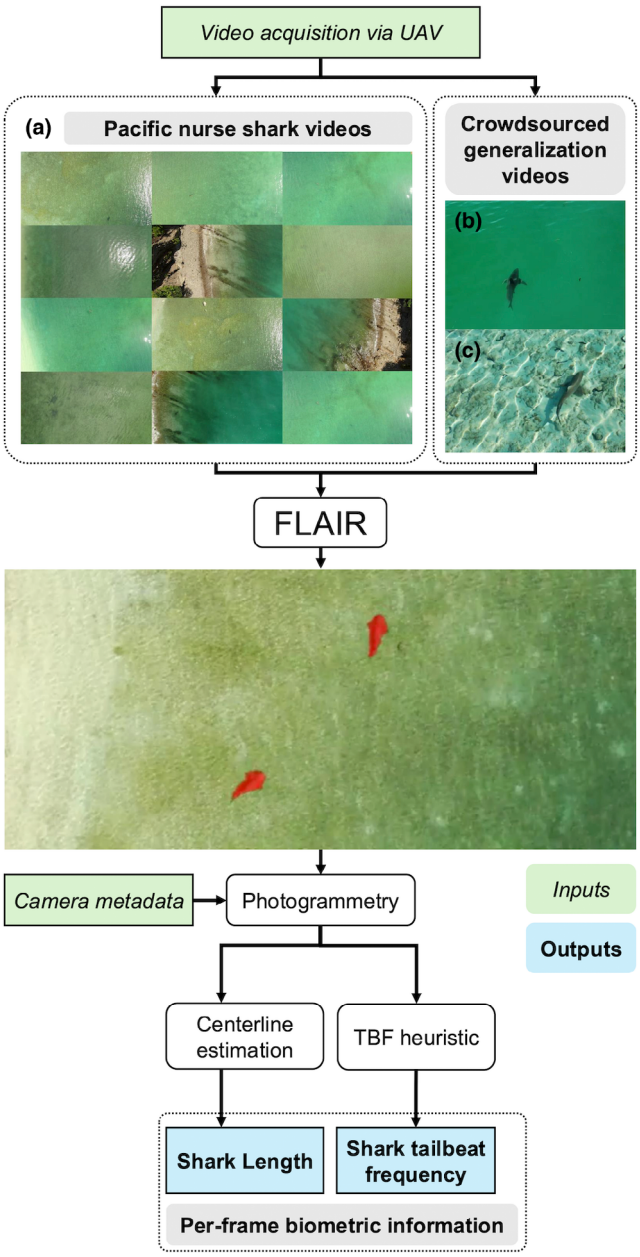

FIGURE 2 Automated biometrics workflow with FLAIR. Inputs are italicized and in green, outputs are bolded and in blue. (a) Representative frames from the Pacific nurse shark UAV video dataset. Tests of generalization of FLAIR beyond our study site and selected species was conducted with crowdsourced videos of a white shark (b) and a blacktip reef shark (c).

include a diverse set of 7 videos from the two sites. Ground-truth bounding boxes were added for each shark with the Computer Vision Annotation Tool (CVAT) (CVATteam, 2020). The images from Videos 1 and 2 (henceforth referred to as the holdout videos) were completely separated from the rest of the dataset, as a test for generalization. The other 5 videos were time-blocked such that each 45 adjacent frames of a video were held together when the data was split into training, validation and test sets. This time-blocking was done to minimize occurrences of consecutive frames being present in the training and test set, which would artificially inflate models' performance metrics. Before object detection and instance segmentation model training, images were rescaled to 1080×1080 pixels and standard data augmentation techniques, including random rotation, brightness and hue adjustments, were applied.

Excluding the images from the holdout videos, our object detection dataset contained 9200 unique positive examples (images containing sharks) and a corresponding 27,000 unique negative examples (images containing no sharks). We selected 9200 negative examples, so that our dataset contained an equal ratio of images with and without sharks. The dataset was then randomly split into training, validation and test sets, in a [80-10-10] ratio, maintaining the images in these 45-frame time blocks. The resulting dataset used for training and testing the object detectors contains 18,400 total images.

For end-to-end instance segmentation, 500 images containing sharks were randomly selected from the 5 training videos, and each instance of a shark was labelled with a ground-truth segmentation mask in CVAT. These 500 images were split into training and validation sets in an [85-15] ratio.

Object detection models were evaluated on a test set from the 5 training videos, as well as the 2 holdout videos. The instance segmentation model was tested on the 2 holdout videos, and the training-free approaches—Per-frame Prompting, HiL-Tracking and FLAIR—did not require training and were thus tested on all 7 videos. For computational efficiency, smaller sub-videos containing sharks were used as input into these pipelines, each ranging from 20 to 80 s in length. This trimmed dataset was used for precision and recall metric comparisons across all methods. A random sampling of 100 frames from the 2 holdout videos and 200 frames from the other 5 videos was annotated for segmentation ground-truth masks and biometrics using CVAT, as described in Section 2.5. These ground-truth masks were used to compute the accuracy of segmentation across methods.

To test the generalizability of FLAIR, two UAV videos licensed under Creative Commons Attribution were sourced from YouTube—one of a white shark (*Carcharodon carcharias*) in Southern California, USA (Video W) and one of a blacktip reef shark (*Carcharhinus melanopterus*) in the Great Barrier Reef, Australia (Video B) (Figure 2b,c). Full attribution to the original creators of these videos is provided in the Supporting Information. Segmentation mask ground truths and biometrics were manually annotated on 25 random frames from each of the two videos. Table S1 contains a concise summary of all datasets used in this study.

2.2 | Baselines

2.2.1 | Per-frame prompting

Per-frame prompting requires a human annotator to label bounding boxes for sharks in every frame of a video, which is effective for accurate object detection but extremely laborious (Figure 3a). We hand labelled bounding boxes in CVAT, prompting SAM 2 Image

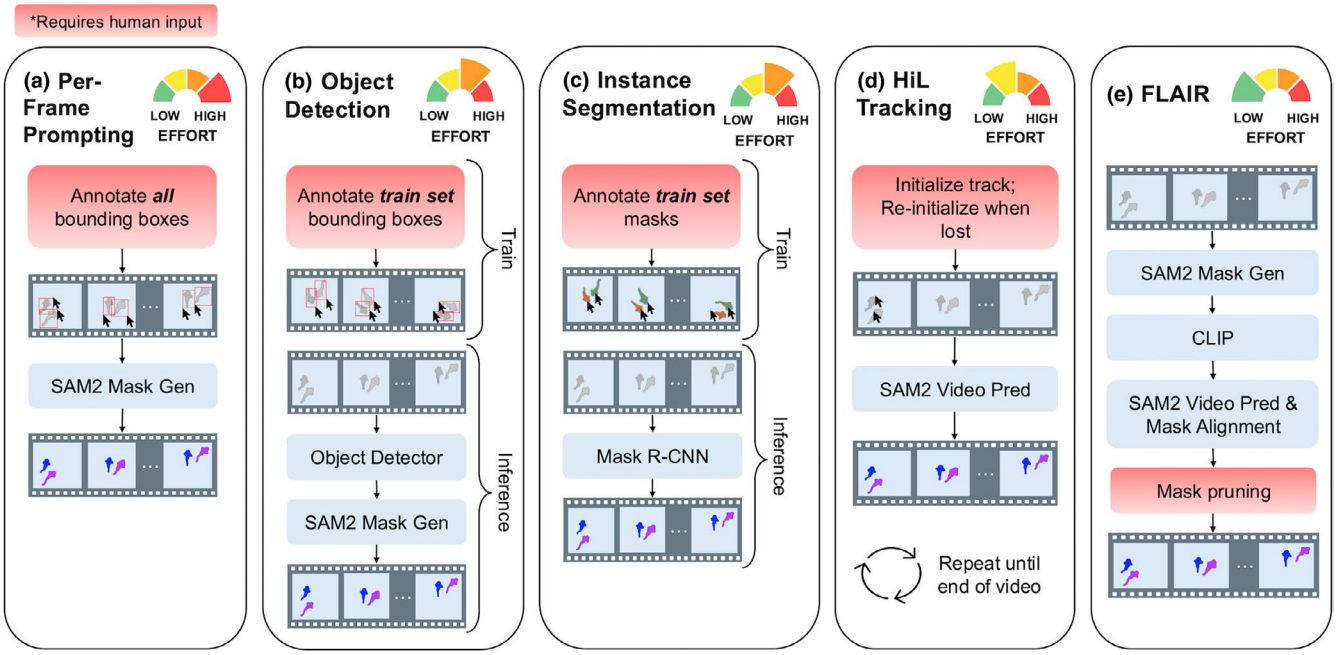

FIGURE 3 Summary of segmentation methods compared in our experiments for Per-frame Prompting with SAM 2 Mask Generation (a), Object Detection models paired with SAM 2 Mask Generation (b), Instance Segmentation with a Mask R-CNN (c), Human-in-the-Loop Tracking with SAM 2 Video Prediction (d) and FLAIR (e). Steps requiring human input/effort are shown in red. Effort is shown in the upper right of each figure.

Predictor with each frame and corresponding bounding boxes to get masks for sharks in every image. Notably, the segmentation aspect of this method does not utilize video understanding, as each frame was segmented as an individual image.

2.2.2 | Object detection models

As object detection models require less manual labelling effort than instance segmentation models, we evaluated the performance of object detectors in tandem with SAM 2 for segmentation tasks. We trained a series of object detection models to prompt SAM 2, predominantly leveraging the Detectron2 model library (Wu et al., 2019). First, we employed several R-CNN model architectures—two-stage detectors that are optimized for accuracy, in contrast to many one-stage detectors that prioritize inference speed (Girshick, 2015). We trained several R-CNNs with various backbones (ResNet and ResNeXt architectures) and Feature Pyramid Networks (FPN), which enhance detection of objects of various scales in images.

We also trained DETR and YOLOv8 models for shark detection (Carion et al., 2020; Jocher et al., 2023). All object detection models were pre-trained on the Imagenet dataset (Deng et al., 2009). The pre-trained medium YOLOv8 model was trained for 40 epochs using default hyperparameters (until loss converged). DETR was trained for 120 epochs, and the rest of Detectron models were trained for 40 epochs each—all with a maximum of 100 objects detected per frame. Each frame, SAM 2 was prompted with bounding boxes output from the object detector to generate segmentation masks (Figure 3b).

2.2.3 | Instance segmentation models

To evaluate end-to-end instance segmentation models, we trained a Mask R-CNN with a ResNeXt-101 architecture and an FPN neck using the Detectron2 library. As shown in Figure S2, the model was pre-trained on the COCO dataset (Lin et al., 2014) and trained for 3000 iterations (14 epochs) until loss converged using default hyperparameters.

2.3 | Human-in-the-loop (HiL) tracking

In human-in-the-loop tracking, a human manually annotates a bounding box in the first frame a shark is identified and then SAM 2 is used to track the segmentation through the remainder of the video or until the object is lost (Figure 3c). When the object is no longer detected by SAM 2, the annotator re-initializes the segmentation track with a bounding box. CVAT was used for bounding box initialization.

2.4 | FLAIR

Our proposed method, FLAIR, is an autonomous framework for object tracking—integrating frame-level alignment and video understanding with language prompts, as illustrated in Figure 4. First, individual frames of the video are sampled at a uniform time interval and are passed into the SAM 2 Automatic Mask Generator. In this work, time intervals of 30 frames (1 s) were used. SAM 2 Automatic

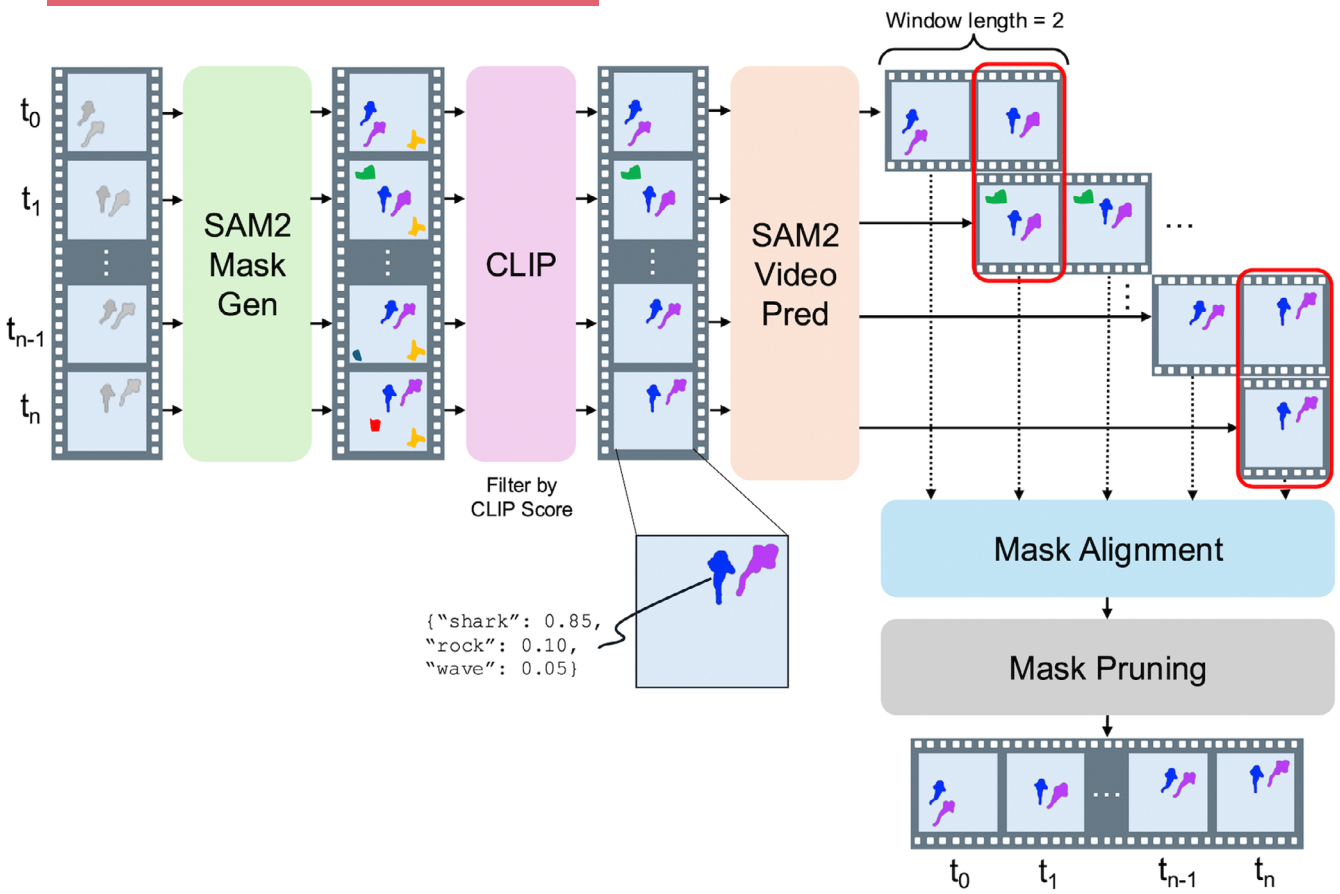

FIGURE 4 Detailed overview of FLAIR architecture. Input video is passed into SAM 2 Mask Generation to segment all objects, which are then filtered by CLIP score given language prompts. Candidate masks are propagated through their window and aligned to eliminate false positives; then a manual mask pruning step further eliminates false positives, resulting in accurate tracking of objects of interest.

Mask Generator generates masks for all possible objects in the image by sampling single-point inputs in a grid, filtering and de-replicating candidate masks and performing further processing for improved quality. Bounding boxes are generated for each mask in the frame and used to crop corresponding image regions. These regions are then passed into CLIP, along with prompts that were tuned for this task. The specific prompts used are included in Supporting Information. The prompts were held constant across all of the nurse shark videos, as well as the white and blacktip reef shark videos, highlighting the generalizability of this method. Each region is assigned a probability associated with every prompt and all regions with a probability assigned to the shark prompt greater than 0.95 are kept as candidate sharks. Frames with their corresponding candidate regions are then initialized in SAM 2 Video Prediction to track the candidate sharks through the video, propagating to both prior and future frames. This video tracking is performed at every time interval. As a mask is tracked through the video, if it overlaps with another candidate mask track initialized in a different time step with an IOU greater than 0.7, the masks are determined to be aligned and thus declared a true positive mask for a shark. For example, we see in Figure 4 that the false positive mask (in green) was propagated through the second time interval track, but was not present in any other tracks. Thus, it was not classified as a true shark mask during alignment.

The duration over which each candidate mask is propagated is controlled by a hyperparameter termed the window length. We evaluated multiple window lengths, finding that shorter windows reduced inference time, while longer windows improved the detection of true positives (Figure S1). The optimal window length for this task was determined to be 3 s, maintaining accuracy comparable to full-video propagation. Masks classified as true positives were propagated for the specified window length across the video, or until the mask disappeared. This approach ensures complete tracking of objects without requiring propagation across the entire video, balancing efficiency and tracking accuracy.

Lastly, a brief mask pruning step is conducted to filter potentially erroneous masks, providing an additional layer of quality control through manual inspection. This involves reviewing the mask tracks and selecting those that correspond to objects of interest. FLAIR performance was evaluated with and without this final pruning step.

FLAIR's advantage stems from the relatively small number of false positives that CLIP identifies as potential sharks. Although these false positives can be propagated through the video from a single frame, the same candidate shark mask will likely be absent at future time intervals. Thus, they will not be aligned across tracks—increasing robustness and generalizability.

TABLE 1 Performance metrics (mean Average Precision and Recall at IOU thresholds) of bounding boxes predicted by YOLOv8, DETR, HiL-Tracking, FLAIR without mask pruning and FLAIR on the holdout videos.

	Video 1				Video 2			
	mAP		mAR		mAP		mAR	
Model	[0.5:0.95]	[0.1]	[0.5:0.95]	[0.1]	[0.5:0.95]	[0.1]	[0.5:0.95]	[0.1]
YOLOv8	0.19	0.75	0.28	0.76	0.02	0.03	0.01	0.03
DETR	0.14	0.88	0.24	0.89	0	0	0	0
Mask R-CNN	0.04	0.76	0.11	0.91	0.02	0.16	0.03	0.21
HiL-Tracking	0.05	0.93	0.13	0.95	0.22	1	0.3	1
FLAIR w/o pruning	0.07	0.93	0.17	0.96	0.05	0.16	0.38	1
FLAIR	0.07	0.93	0.17	0.96	0.26	0.9	0.38	1

All analyses were performed using the SAM 2 Hiera Large model on a single NVIDIA L40 GPU. FLAIR requires a minimum of 12 GB of GPU memory and is compatible with a standard NVIDIA T4 GPU (accessible in Google Colab for free). We recommend using NVIDIA L4 or L40 GPUs.

2.5 | Biometric measurements

Tailbeat frequency (TBF) and length were computed from FLAIR-predicted masks and compared against manual calculations for a single individual of each species: a Pacific nurse shark, a blacktip reef shark, and a white shark from open-access videos. Methods for manual measurement for length and TBF are detailed in the Supporting Information. To estimate the length of the shark along its centreline, we first obtain the segmentation mask from FLAIR or HiL. The mask is skeletonized using Zhang's method (Zhang & Suen, 1984), which thins the mask around the boundaries over successive passes, eventually obtaining the skeleton of the mask. The total length of the skeleton is calculated by traversing each point on the backbone, extending the skeleton to include the distal ends of the mask. Lengths are then converted from pixels to meters using Equation S1. Due to the unavailability of drone metadata, biometrics calculations presented in this work should only be used as an example of the capabilities of this workflow and not to draw scientific conclusions.

To calculate TBF, the two furthest points in the mask were identified, with the point closer to the centre of mass (COM) of the mask assigned as the head and the further point deemed the tail. The centre axis of the shark was calculated as the vector from the head point to the COM. The orthogonal distance between the end of the tail and the centre axis is calculated as the magnitude of the orthogonal component of the vector from the COM to the tail relative to the centre axis. This distance is calculated for every mask in each frame of the video, and smoothing is performed by applying a Savitzky–Golay filter across the distances. Then, local extrema above a constant prominence threshold were identified. The first point where the curve intersects the centre axis between each pair of local extrema is classified as a crossing. The tail beat period is calculated as the time between every other crossing of the central axis and the TBF as the reciprocal of the period.

3 | RESULTS

3.1 | Object detection

We evaluated our suite of object detection models on a test dataset to assess the models' ability to learn representations, as well as two holdout videos to assess the generalizability of the models.

Both YOLO and DETR performed better on the test set than the trained Faster R-CNN models with Feature Pyramid Network and Dilated-C5 backbones, with complete results presented in Table S2. It is important to note that although mean Average Precision (mAP) and mean Average Recall (mAR) averaged across Intersection over Union (IOU) thresholds of 0.5 and 0.95 are traditionally used as performance metrics for object detection models, we found that it is not necessary for predicted bounding boxes to have a high IOU with ground-truth boxes to obtain accurate results on downstream segmentation and biometrics tasks.

As seen in Table 1, YOLOv8 and DETR had high precision and recall at lower IOU thresholds on holdout Video 1, but both performed extremely poorly on holdout Video 2, with a mean Average Recall at an IOU threshold of 0.1 (mAR@0.1) being 0.03 and 0, respectively. The Mask R-CNN model demonstrated strong performance on Video 1, but similar to the object detection models, had low accuracy on Video 2, achieving a mAR@0.1 of 0.21. Both FLAIR and HiL achieved high recall in Video 1 and significantly outperformed object detection and instance segmentation models in Video 2, reaching a mAR@0.1 of 1. FLAIR achieved substantially higher precision in Video 2 compared to its variant without the additional mask pruning (0.90 vs. 0.16 at mAP@0.1), highlighting the importance of this step.

3.2 | Shark segmentation

Beyond object detection, segmentation accuracy is an especially relevant metric for evaluating the performance of models on downstream scientific tasks, such as biometrics. Dice score was used to evaluate segmentation performance, which measures the spatial overlap between the predicted and ground-truth masks, with a

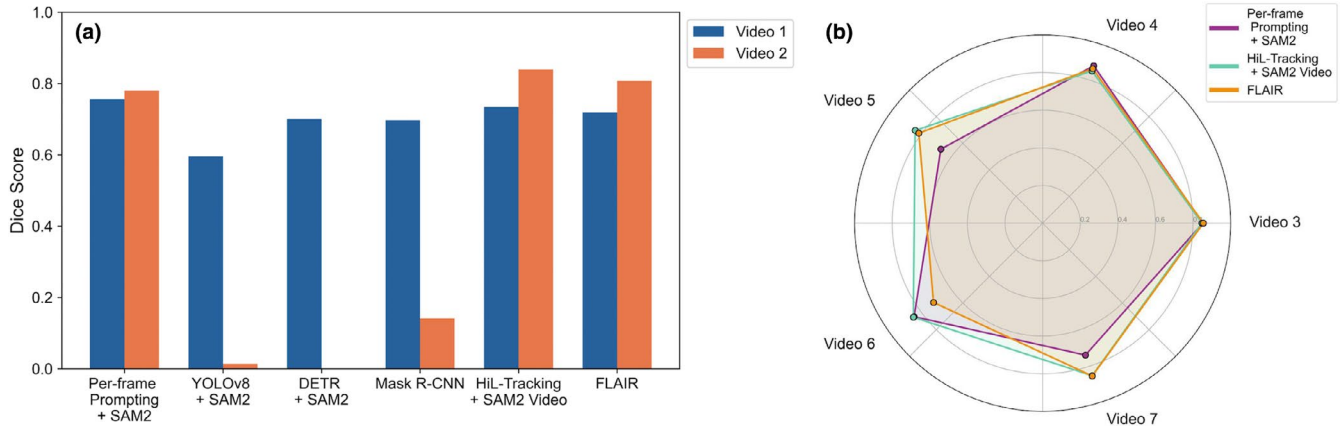

FIGURE 5 (a) Dice score comparison of Per-frame Prompting + SAM 2, YOLOv8 + SAM 2, DETR + SAM 2, Mask R-CNN, HiL-Tracking + SAM 2 Video and FLAIR on 2 holdout videos of nurse sharks. Object detector methods have near-zero segmentation accuracy on the second video. (b) Dice score comparison of Per-frame Prompting, HiL-Tracking and FLAIR on 5 videos containing nurse sharks. FLAIR has competitive performance with both methods that require a human in the loop.

value of 1 indicating perfect agreement and 0 indicating no overlap. All models had relatively high Dice scores for Video 1, as it was taken on the same day and location as two of the training videos (Figure 5a). However, both object detectors (YOLOv8 and DETR) were unable to identify bounding boxes in Video 2, which was taken on a different day, resulting in substantially lower segmentation accuracy. The Mask R-CNN outperformed object detectors on Video 2, but had considerably lower segmentation accuracy than HiL-Tracking and FLAIR. On Video 2, Per-frame Prompting, HiL-Tracking, FLAIR had high Dice scores of 0.780, 0.839 and 0.807, respectively (Table S3).

HiL-Tracking and FLAIR both outperform Per-frame Prompting on videos 3–7, with mean Dice scores of 0.847, 0.821 and 0.796, respectively. FLAIR even surpasses HiL-Tracking and Per-frame Prompting on videos 3 and 7, thus highlighting the advantage of using video understanding in instance segmentation. Both FLAIR and HiL-Tracking generated accurate segmentations for the white and blacktip shark videos as well, with Dice scores of 0.919 and 0.881, respectively (Table S3).

3.3 | Biometrics case study

Body length and tailbeat frequency (TBF) were calculated from FLAIR and HiL-Tracking masks for sampled frames from three individual videos of a white shark, a Pacific nurse shark and a blacktip reef shark. These measurements were compared to manual measurements of body length and TBF. The FLAIR-predicted body lengths align closely with the manually measured lengths, as these values are tightly concentrated around the line $y = x$ in Figure 6a. For each species, the distributions of FLAIR and HiL-Tracking-predicted body length are nearly identical, and they closely follow manually measured body length (Figure 6e–g). The body lengths derived from FLAIR-predicted masks (reported as mean ± standard deviation) for the white shark, Pacific nurse shark and blacktip reef shark, were 5.3 ± 0.8 m, 1.5 ± 0.1 m and 1.0 ± 0.3 m, respectively. Similarly, the body lengths derived from manual annotation for the white shark, Pacific nurse shark and blacktip reef shark, were 5.0 ± 0.8 m, 1.4 ± 0.1 m and 1.0 ± 0.3 m, respectively. In addition, visual inspection of masks and predicted centrelines shows accurate segmentation and centre line estimation of the sharks (Figure 6b–d). The large variance in white shark and blacktip reef shark length measurements is due to the unavailability of drone metadata for these open-access videos. The drone metadata (camera zoom and angle, drone altitude, etc.) varied throughout these videos, but were assumed to be constant. Recording drone metadata for each video and maintaining constant altitude and nadir camera angle is recommended for more systematic biometrics analyses.

Tail beat frequency was calculated by analysing the raw tail displacement signal relative to the centre line, as shown in Figure 7a–c. Tail displacement signals and TBF measurements for FLAIR and HiL were nearly identical. TBF predicted from FLAIR very closely followed manually calculated TBF, with a mean error of 2.1% across all three species. All FLAIR-predicted TBF measurements were within 7% error from the manual measurements. TBF was relatively constant for nurse and white sharks across time, but increased in the blacktip shark across the video, peaking at 0.86 tail beats per second, corresponding to 1.16 s per tail beat. (Figure 7d). Additional details on biometrics measurements are available in the Supporting Information section.

3.4 | Efficiency comparisons

We assessed the manual annotation time required across methods. For frames with at least one shark present, labelling bounding boxes took an average of 10.5 s per frame, while labelling segmentation masks took an average of 45 s per frame. At training time, the object detectors used a dataset of 9200 images with sharks, which took approximately 27 h to label with bounding boxes. The

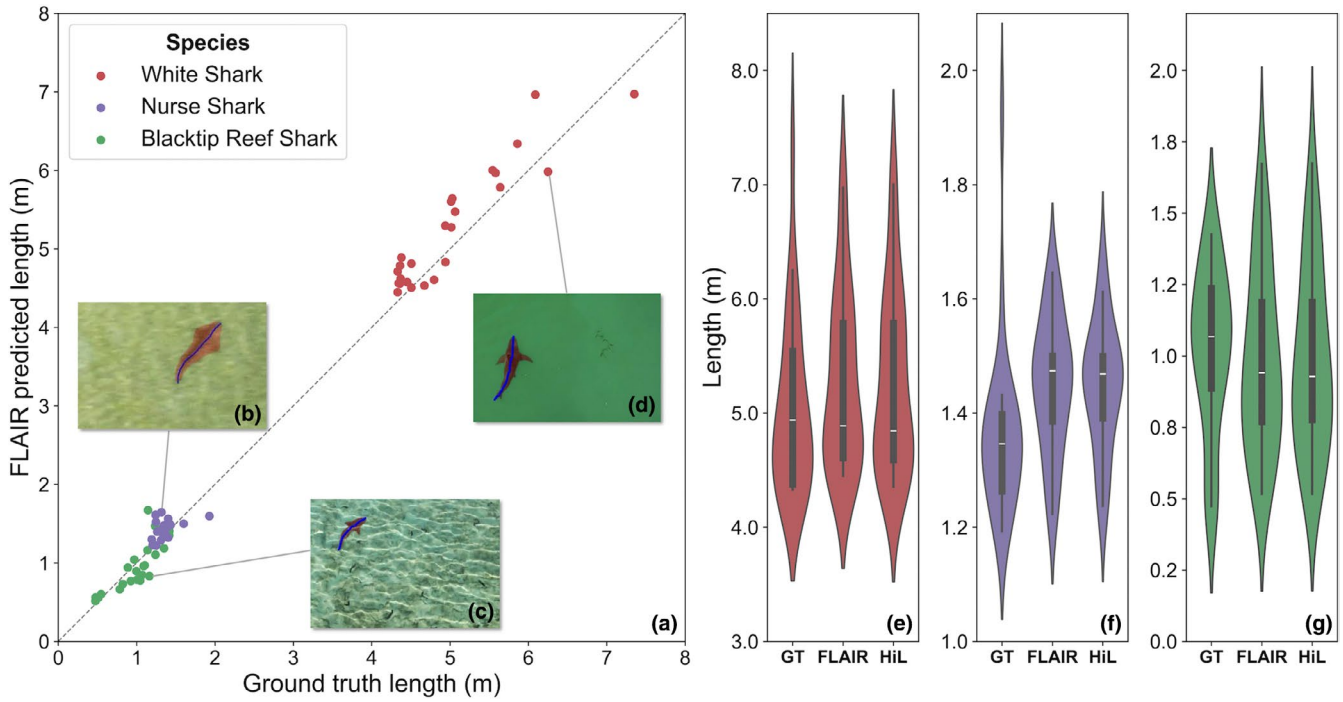

FIGURE 6 (a) Comparison between manual length measurements and automated length calculation from FLAIR masks across three shark videos for (b) Pacific nurse shark, (c) blacktip reef shark and (d) white shark. Violin plots show the distribution of length measurements measured manually and calculated from FLAIR and HiL masks from sampled frames across white shark (e), Pacific nurse shark (f) and blacktip reef shark (g) videos. Internal box plots show the median (white horizontal line), interquartile range (25th–75th percentile, grey rectangles) and the grey vertical lines extend to the minimum and maximum. This analysis is limited to one shark per species. The observed variation in lengths for each shark may be attributable to the absence of relevant drone metadata.

Mask R-CNN was trained with 500 images containing sharks, which took approximately 6 h to label with segmentation masks. In addition, significant engineering effort (on the order of days) was required for training and validating the object detection and instance segmentation models. Per-frame prompting and HiL-Tracking require no manual effort or computational resources at training time.

FLAIR requires some degree of CLIP prompt engineering and hyperparameter tuning. The extent of manual effort varies with the complexity and diversity of the visual environment of the dataset. However, tuning can be conducted on a small (and representative) subset of the data and does not require further manual effort when applying FLAIR to additional videos. In practice, applying FLAIR to new aerial datasets required less than 5 min of tuning to achieve excellent performance.

Computational inference times across methods were measured for a 5-min-long, 30 FPS video (9000 total frames) in which four sharks were present at various points. Since the video had sharks present in 1710 frames, manual labelling of segmentation masks would take approximately 21 h. Object detection methods required 8 min to generate masks for the entire video, while instance segmentation using Mask R-CNN took around 18 min. In contrast, per-frame prompting was considerably slower, estimated at approximately 5 h. Approaches using video understanding (HiL and FLAIR) were more efficient, completing in 19 and 54 min, respectively. FLAIR optionally supports manual mask pruning, enabling users to eliminate all false positives in under a minute. Additional details on inference time calculations are available in the Supporting Information.

The total runtime for FLAIR is primarily determined by two stages: SAM 2 Automatic Mask Generation and SAM 2 Video Propagation, while CLIP filtering contributes negligibly to overall processing time. The duration of the Automatic Mask Generation phase scales linearly with video length, while the time for Video Propagation depends on the number of object candidates retained after CLIP-based filtering. This dependency reflects both the number of actual objects present in the video and the effectiveness of the CLIP text prompts used. As shown in Figure S1, the majority of the runtime is attributable to the Automatic Mask Generation step, which remains constant across all window lengths. As segmentation methods continue to advance—particularly with the introduction of batch processing and other optimizations—this stage will become increasingly faster, ultimately enabling frame alignment techniques like FLAIR to operate in real time.

4 | DISCUSSION

In this study, we present FLAIR, an automated text-prompted instance segmentation method that markedly improves the efficiency of studying marine animals in aerial drone videos. We found FLAIR performed better than several state-of-the-art object detection and instance segmentation models, achieving higher detection rates

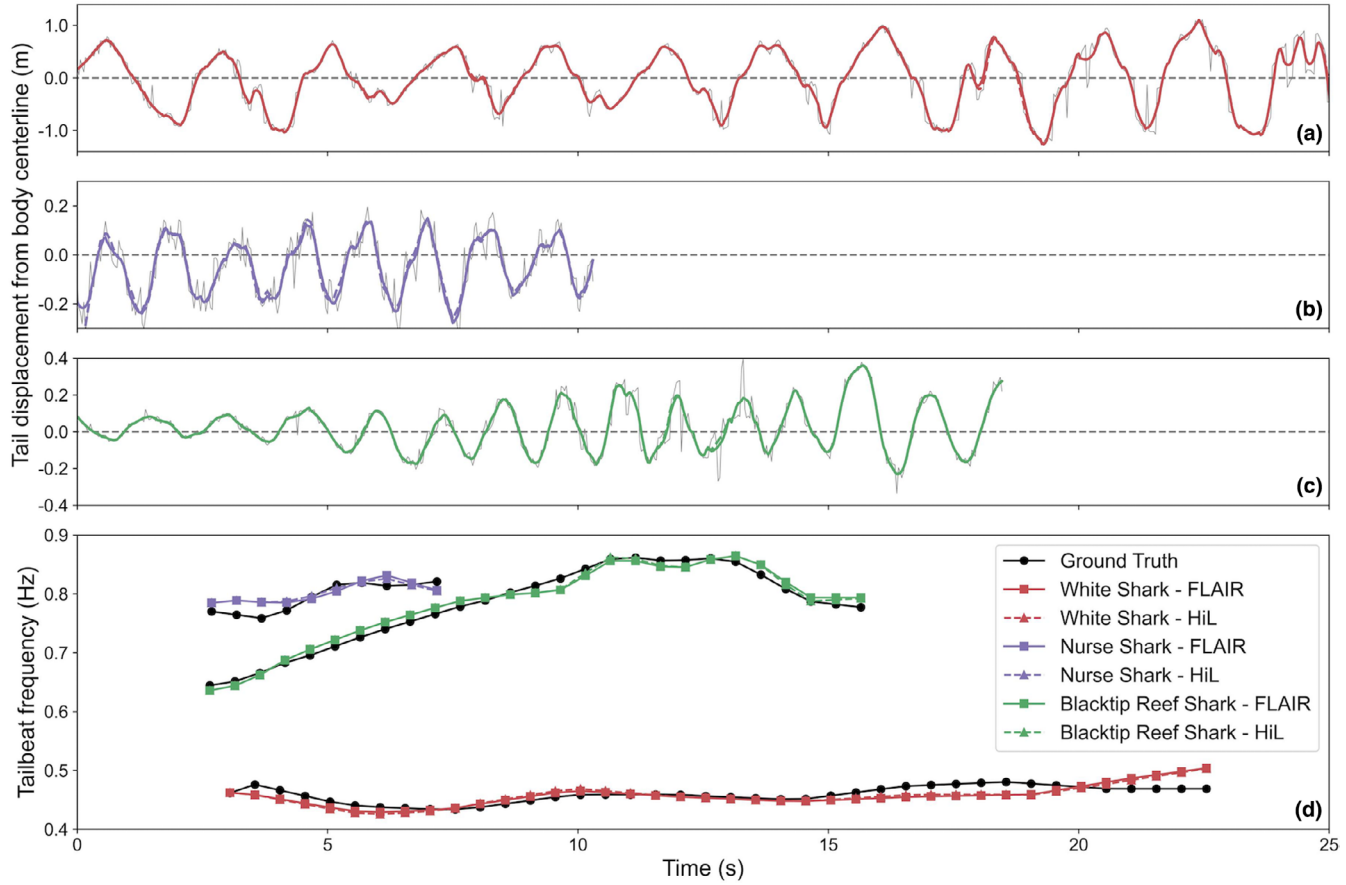

FIGURE 7 Tail displacement from body centreline over time for white shark (a), Pacific nurse shark (b) and blacktip reef shark (c). Thin grey lines denote the raw FLAIR signal. Thick lines denote the smoothed FLAIR (solid) and HiL (dashed) signals. (d) Comparison between automatically calculated tailbeat frequency (from FLAIR and HiL masks) and manually measured (ground truth) tailbeat frequency (black line) for all 3 species of shark.

across visually challenging videos in a zero-shot manner. Traditional object detection methods like YOLOv8 and DETR rely on large, diverse training datasets to generalize across videos with varying conditions—resources that are often unavailable for species lacking extensive datasets. Training these models also requires significant annotator and engineering effort. Despite the substantially reduced manual effort required, FLAIR and HiL-Tracking segment sharks accurately across a diverse set of videos containing Pacific nurse sharks, blacktip reef sharks and white sharks (Table S4). Given the scarcity of diverse aerial drone datasets for marine animal research, FLAIR offers an accessible solution for automated analyses (Weinstein, 2018).

We also show that FLAIR segmentations enable biometric analysis of shark imagery, including computation of animal length—which is essential for understanding shark population demographics. Shark speed, tail beat frequency and other kinematic measurements can also be estimated using FLAIR. These biometric estimates from aerial imagery can be monitored across time and habitat changes, and studied in combination with biological metadata to assess fine-scale predator–prey interactions, cooperative behaviour and more (Andrzejaczek et al., 2019; Gleiss et al., 2009; Whitney et al., 2010).

The primary limitation of segmentation-based tracking methods is the accuracy of the segmentations themselves, which can impact downstream biometric estimates. When a shark is swimming near the seafloor in very shallow water, SAM 2 will occasionally segment the shadow of the shark along with the shark itself. In turbid water, the pectoral and caudal fins are occasionally left out of the segmentation, or the caudal lobe will be segmented separately from the rest of the body. In addition, obtaining precise biometrics from segmentation masks may be challenging in aerial videos where drone metadata is not available. Ultimately, the quality of the drone aerial imagery has a significant effect on the accuracy of detection and segmentation for any marine and terrestrial species (Ramos et al., 2022).

Approaches that rely on foundation models, like FLAIR, may suffer from increased computational costs at inference compared to object detection and instance segmentation models and may require higher-performance GPUs. We found that for a 5-min video, FLAIR took 3 times longer to generate segmentation masks than a fine-tuned Mask R-CNN. However, we found that FLAIR outperformed the Mask R-CNN on object detection and segmentation tasks. FLAIR also produced biometric results comparable to manual measurements, consistent with the findings of (Bierlich et al., 2024). Notably,

SAM does not always outperform fine-tuned Mask R-CNNs in segmenting marine megafauna from UAV imagery, as found in a recent study of right whales (Bagchi et al., 2025). However, frameworks centred on foundation models (such as FLAIR) can readily generalize in the face of shifting distributions and even when applied to new species. In instances where accuracy on a fixed, in-distribution dataset is prioritized over generalization, specialized segmentation models may outperform foundation model-based approaches. A hybrid approach could use FLAIR to generate masks from a small subset of video frames, eliminating the need for human annotation of a training set. These masks could then be used to train a segmentation model (e.g. Mask R-CNN) that enables efficient analysis of footage with minimal inference time and computational cost.

The core advantage of FLAIR is its potential to generalize to new marine and terrestrial aerial datasets. This pipeline is directly applicable to the tracking of any animal or object from aerial imagery. To demonstrate the generalizability of FLAIR beyond marine settings, we applied the framework to a dataset of aerial videos of Grévy's zebras in open plains (Price et al., 2023). Despite shifts in habitat and visual occlusions, FLAIR consistently tracked individual zebras across frames. Using aerial imagery to study wildlife allows for non-invasive tracking and observation, capturing dynamic information about animal biomechanics, interactions and behaviours (Bierlich et al., 2024; Gray, Bierlich, et al., 2019; Hansen et al., 2022; Torres et al., 2022). We hope that this framework will be applied across diverse ecosystems and species, providing a scalable solution for addressing conservation challenges, informing policy and fostering sustainable management practices in both marine and terrestrial environments (Buchelt et al., 2024; Stark et al., 2018).

Deep learning is transforming ecological research by enabling scientists to process and analyse massive datasets of wildlife imagery. Integrating state-of-the-art foundation model frameworks to derive biological conclusions and understand fine-grained changes in ecosystems should be a priority. The methods presented here, namely FLAIR, allow for automated detection and tracking of animals from aerial imagery, along with streamlined pipelines for downstream biometrics. As foundation models become increasingly powerful and efficient, we expect that methods like FLAIR will be scalable tools for understanding complex ecological interactions.

AUTHOR CONTRIBUTIONS

This study is a collaboration between authors from multiple countries, including scientists based in Costa Rica, where the study was conducted. Each author participated from the beginning of the research process, ensuring that the diverse perspectives they represented were considered. Chinmay K. Lalgudi, Mark E. Leone and Jaden V. Clark conceived the ideas and designed methodology; Sergio Madrigal-Mora and Mario Espinoza collected the data; Chinmay K. Lalgudi led development of software; Chinmay K. Lalgudi, Mark E. Leone and Jaden V. Clark analysed the data; Chinmay K. Lalgudi led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

Many thanks to M. Lara, S. Lara and A. Lara from Dive Center Cuajiniquil and C. Lowe for supporting data collection as a joint project between UCR CIMAR and CSULB Shark Lab. Our sincere thanks are extended to J. Meribe for his assistance with dataset labelling. We also gratefully acknowledge funding from the Save Our Seas Foundation (Project No. 664), the American Elasmobranch Society and the Richard D. Green Graduate Fellowship, which supported many of our drone survey trips. Support for data collection was also provided by the Mel Lane Student Grants Program from the Stanford Woods Institute for the Environment and Stanford ASSU.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.70116.

DATA AVAILABILITY STATEMENT

Data available via https://doi.org/10.5281/zenodo.15845961 (Lalgudi, 2025a). Code is available at https://github.com/conservation-technology-group/FLAIR and archived at: https://doi.org/10.5281/zenodo.16056059 (Lalgudi, 2025b).

ORCID

Chinmay K. Lalgudi https://orcid.org/0009-0003-0295-3171

REFERENCES

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*, 53. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00444-8

Andrzejaczek, S., Gleiss, A. C., Lear, K. O., Pattiaratchi, C. B., Chapple, T. K., & Meekan, M. G. (2019). Biologging tags reveal links between fine-scale horizontal and vertical movement behaviors in tiger sharks (*Galeocerdo cuvier*). *Frontiers in Marine Science*, *6*, 229.

Bagchi, C., Medina, J., Irschick, D. J., Maji, S., & Christiansen, F. (2025). Automated extraction of right whale morphometric data from drone aerial photographs. *Remote Sensing in Ecology and Conservation*.

Bierlich, K. C., Karki, S., Bird, C. N., Fern, A., & Torres, L. G. (2024). Automated body length and body condition measurements of whales from drone videos for rapid assessment of population health. *Marine Mammal Science*, *40*, e13137.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv Preprint* arXiv:2108.07258.

Buchelt, A., Adrowitzer, A., Kieseberg, P., Gollob, C., Nothdurft, A., Eresheim, S., Tschiatschek, S., Stampfer, K., & Holzinger, A. (2024). Exploring artificial intelligence for applications of drones in forest ecology and management. *Forest Ecology and Management*, *551*, 121530.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. European Conference on Computer Vision, 213–229.

Clark, J., Lalgudi, C., Leone, M., Meribe, J., Madrigal-Mora, S., & Espinoza, M. (2025). Deep learning for automated shark detection and biometrics without keypoints. European Conference on Computer Vision, 105–120.

CVATteam. (2020). *Computer vision annotation tool* (*CVAT*). https://github.com/opencv/cvat

Dedman, S., Moxley, J. H., Papastamatiou, Y. P., Braccini, M., Caselle, J. E., Chapman, D. D., Cinner, J. E., Dillon, E. M., Dulvy, N. K., Dunn, R. E., Espinoza, M., Harborne, A. R., Harvey, E. S., Heupel, M. R., Huveneers, C., Graham, N. A. J., Ketchum, J. T., Klinard, N. V., Kock, A. A., … Heithaus, M. R. (2024). Ecological roles and importance of sharks in the anthropocene ocean. *Science*, *385*(6708), adl2362.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255.

Desgarnier, L., Mouillot, D., Vigliola, L., Chaumont, M., & Mannocci, L. (2022). Putting eagle rays on the map by coupling aerial video-surveys and deep learning. *Biological Conservation*, *267*, 109494.

DiGiacomo, A., Abraham, A. M., Andrzejaczek, S., & Block, B. (2023). *Quantifying juvenile white shark swimming kinematics using unoccupied aircraft systems* (*UAS*) *and deep neural networks*. Abstract Book, White Sharks Global, 41.

Eikelboom, J. A., Wind, J., van de Ven, E., Kenana, L. M., Schroder, B., de Knegt, H. J., van Langevelde, F., & Prins, H. H. (2019). Improving the precision and accuracy of animal population estimates with aerial image object detection. *Methods in Ecology and Evolution*, *10*(11), 1875–1887.

Ferretti, F., Worm, B., Britten, G. L., Heithaus, M. R., & Lotze, H. K. (2010). Patterns and ecosystem consequences of shark declines in the ocean. *Ecology Letters*, *13*(8), 1055–1071.

Girshick, R. (2015). Fast r-cnn. Proceedings of the IEEE International Conference on Computer Vision, 1440–1448.

Gleiss, A. C., Gruber, S. H., & Wilson, R. P. (2009). Multi-channel data-logging: Towards determination of behaviour and metabolic rate in free-swimming sharks. In *Tagging and tracking of marine animals with electronic devices* (pp. 211–228). Springer Netherlands.

Gorkin, R., III, Adams, K., Berryman, M. J., Aubin, S., Li, W., Davis, A. R., & Barthelemy, J. (2020). Sharkeye: Real-time autonomous personal shark alerting via aerial surveillance. *Drones*, *4*(2), 18.

Gray, P. C., Bierlich, K. C., Mantell, S. A., Friedlaender, A. S., Goldbogen, J. A., & Johnston, D. W. (2019). Drones and convolutional neural networks facilitate automated and accurate cetacean species identification and photogrammetry. *Methods in Ecology and Evolution*, *10*(9), 1490–1500.

Gray, P. C., Fleishman, A. B., Klein, D. J., McKown, M. W., Bezy, V. S., Lohmann, K. J., & Johnston, D. W. (2019). A convolutional neural network for detecting sea turtles in drone imagery. *Methods in Ecology and Evolution*, *10*(3), 345–355.

Hansen, M., Krause, S., Dhellemmes, F., Pacher, K., Kurvers, R., Domenici, P., & Krause, J. (2022). Mechanisms of prey division in striped marlin, a marine group hunting predator. *Communications Biology*, *5*(1), 1161.

Heupel, M. R., Knip, D. M., Simpfendorfer, C. A., & Dulvy, N. K. (2014). Sizing up the ecological role of sharks as predators. *Marine Ecology Progress Series*, *495*, 291–298.

Hodgson, A., Kelly, N., & Peel, D. (2013). Unmanned aerial vehicles (UAVs) for surveying marine fauna: A dugong case study. *PLoS One*, *8*(11), e79556.

Jocher, G., Chaurasia, A., & Qiu, J. (2023). *Ultralytics YOLO* (Version 8.0.0) [Computer software]. https://github.com/ultralytics/ultralytics

Jorgensen, S. J., Micheli, F., White, T. D., Van Houtan, K. S., Alfaro-Shigueto, J., Andrzejaczek, S., Arnoldi, N. S., Baum, J. K., Block, B., Britten, G. L., Butner, C., Caballero, S., Cardeñosa, D., Chapple, T. K., Clarke, S., Cortés, E., Dulvy, N. K., Fowler, S., Gallagher, A. J., … Ferretti, F. (2022). Emergent research and priorities for shark and ray conservation. *Endangered Species Research*, *47*, 171–203.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment anything. *arXiv:2304.02643*.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., … Liang, P. (2021). Wilds: A benchmark of in-the-wild distribution shifts. International Conference on Machine Learning, 5637–5664.

Kohler, N. E., & Turner, P. A. (2001). Shark tagging: A review of conventional methods and studies. *Environmental Biology of Fishes*, *60*(1), 191–224.

Lalgudi, C. (2025a). Datasets used in "Zero-shot shark tracking and biometrics from aerial imagery". https://doi.org/10.5281/zenodo.15845961

Lalgudi, C. (2025b). FLAIR: Initial release (v1.0.0). https://doi.org/10.5281/zenodo.16056059

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 6999–7019.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part v 13, 740–755.

Madrigal-Mora, S., Chávez, E. J., Arauz, R., Lowe, C. G., & Espinoza, M. (2024). Long-distance dispersal of the endangered pacific nurse shark (*Ginglymostoma unami*, Orectolobiformes) in Costa Rica revealed through acoustic telemetry. *Marine and Freshwater Research*, *75*(2), MF23162.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, *21*(9), 1281–1289.

Matley, J. K., Klinard, N. V., Jaine, F. R., Lennox, R. J., Koopman, N., Reubens, J. T., Harcourt, R. G., Cooke, S. J., & Huveneers, C. (2024). Long-term effects of tagging fishes with electronic tracking devices. *Fish and Fisheries*, *25*, 1009–1025.

Porter, M. E., Ruddy, B. T., & Kajiura, S. M. (2020). Volitional swimming kinematics of blacktip sharks, carcharhinus limbatus, in the wild. *Drones*, *4*(4), 78.

Price, E., Khandelwal, P. C., Rubenstein, D. I., & Ahmad, A. (2023). A framework for fast, large-scale, semi-automatic inference of animal behavior from monocular videos. *bioRxiv*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. International Conference on Machine Learning, 8748–8763.

Ramos, E. A., Landeo-Yauri, S., Castelblanco-Martínez, N., Arreola, M. R., Quade, A. H., & Rieucau, G. (2022). Drone-based photogrammetry assessments of body size and body condition of Antillean manatees. *Mammalian Biology*, *102*(3), 765–779.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., & Feichtenhofer, C. (2024). Sam 2: Segment anything in images and videos. *arXiv Preprint* arXiv:2408.00714.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 28.

Sharma, N., Scully-Power, P., & Blumenstein, M. (2018). Shark detection from aerial imagery using region-based CNN, a study. AI 2018: Advances in artificial intelligence: 31st Australasian Joint Conference, Wellington, New Zealand, December 11-14, 2018, Proceedings 31, 224–236.

Stark, D. J., Vaughan, I. P., Evans, L. J., Kler, H., & Goossens, B. (2018). Combining drones and satellite tracking as an effective tool for informing policy change in riparian habitats: A proboscis monkey case study. *Remote Sensing in Ecology and Conservation*, *4*(1), 44–52.

Stevens, J., Bonfil, R., Dulvy, N. K., & Walker, P. (2000). The effects of fishing on sharks, rays, and chimaeras (Chondrichthyans), and the implications for marine ecosystems. *ICES Journal of Marine Science*, *57*(3), 476–494.

Torres, L. G., Bird, C. N., Rodríguez-González, F., Christiansen, F., Bejder, L., Lemos, L., Urban, R. J., Swartz, S., Willoughby, A., Hewitt, J., & Bierlich, K. C. (2022). Range-wide comparison of gray whale body condition reveals contrasting sub-population health characteristics and vulnerability to environmental change. *Frontiers in Marine Science*, *9*, 867258. https://doi.org/10.3389/fmars.2022.867258

Torres, W. I., & Bierlich, K. C. (2020). MorphoMetriX: A photogrammetric measurement GUI for morphometric analysis of megafauna. *Journal of Open Source Software*, *4*(44), 1825. https://doi.org/10.21105/joss.01825

Weinstein, B. G. (2018). A computer vision for animal ecology. *Journal of Animal Ecology*, *87*(3), 533–545.

White, J., Simpfendorfer, C., Tobin, A., & Heupel, M. (2013). Application of baited remote underwater video surveys to quantify spatial distribution of elasmobranchs at an ecosystem scale. *Journal of Experimental Marine Biology and Ecology*, *448*, 281–288.

Whitney, N. M., Pratt, H. L., Jr., Pratt, T. C., & Carrier, J. C. (2010). Identifying shark mating behaviour using three-dimensional acceleration loggers. *Endangered Species Research*, *10*, 71–82.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). Detectron2. https://github.com/facebookresearch/detectron2

Zhang, T. Y., & Suen, C. Y. (1984). A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, *27*(3), 236–239.

SUPPORTING INFORMATION

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

Table S1. Metadata for all videos used in this study.

Table S2. Performance Metrics (Average Precision and Recall at IOU thresholds) of various object detection models on the test set.

Table S3. Performance metrics (Dice Score) of segmentation masks predicted by GT + SAM2, YOLOv8 + SAM 2, DETR + SAM 2, and FLAIR on all videos.

Table S4. Summary of methods by human effort, inference time, and generalization performance.

Figure S1. (a) Total runtime for FLAIR as a function of window size in seconds. Longer window lengths, particularly N_{frames} (video length), have much longer inference time. (b) Dice scores from FLAIR-derived segmentations as a function of window size. There are diminishing returns beyond a certain window length threshold of 3.

Figure S2. Total loss and Mask R-CNN training accuracy across training iterations.

How to cite this article: Lalgudi, C. K., Leone, M. E., Clark, J. V., Madrigal-Mora, S., & Espinoza, M. (2025). Zero-shot shark tracking and biometrics from aerial imagery. *Methods in Ecology and Evolution*, *00*, 1–13. https://doi.org/10.1111/2041-210X.70116