# Deep Learning for Automated Shark Detection and Biometrics Without Keypoints

Jaden Clark[†1], Chinmay Lalgudi[†1], Mark Leone[†1], Jayson Meribe[1], Sergio Madrigal-Mora[2,3], and Mario Espinoza[2]

[1] Stanford University, Stanford CA, USA
[2] Universidad de Costa Rica, San Jose, CR
[3] California State University Long Beach, Long Beach CA, USA
jvclark@cs.stanford.edu
**

**Abstract.** This study introduces a pipeline designed to aid biologists via automatic detection and biometric analysis of marine animals. Our approach uses detection transformer (DETR) to detect subjects in an image, then generates a segmentation mask over the animal. We also introduce a new method to measure the center line of segmentations, which can be used to assess length during tail movement of animals in images. We test our system on a new dataset of aerial drone imagery of Pacific Nurse Sharks (*Ginglymostoma unami*). The detection model was trained on a dataset of drone-captured images under diverse environmental conditions of varying water clarity and lighting conditions, achieving a recall of 0.96 and precision of 0.80 at an IOU of 0.35. Notably, our method does not require labeled segmentations or keypoints in the dataset, as we find Segment Anything Model (SAM) has strong zero-shot performance. The efficiency of the pipeline was benchmarked against non-expert human annotators, showing a 91% decrease in data analysis time.

**Keywords:** Object Detection · Biometrics · Ecology

## 1 Introduction

The physical attributes of individual marine animals, such as body length, age, and body condition are critical indicators of their overall ecology and physiology [4, 42]. Studying biometrics at the individual level can be key to estimating population health and how it changes over time [10,35,40]. Long-term monitoring of these metrics is key to understanding how animal populations are responding to potential environmental shifts such as climate change or human disturbance [4, 40]. However, measuring biometrics of large marine animals is challenging as they can be difficult to approach and physically measure in their natural habitats. In this study, we present a new method for automatically estimating shark length, width, and age from aerial imagery. The pipeline was tested on

** [†]These authors all contributed to this work equally and are all considered first authors

aerial imagery of Pacific Nurse Sharks *Ginglymostoma unami* in Santa Elena Bay, Costa Rica.

Santa Elena Bay is a marine protected area (MPA) in northwestern Costa Rica (Figure 1) harboring several endangered marine species including sea turtles, whales, and black coral [11]. The recently described and endangered Pacific Nurse Sharks *Ginglymostoma unami* are also known to aggregate in the bay, and as a newly described species their behavior is particularly poorly understood [13]. Pacific Nurse Sharks are challenging to study because they are highly mobile and elusive. Works such as [29] and [17] use a variety of remote sensors including underwater cameras and acoustic tags to track and study these cryptic animals throughout their habitats. However, collecting and analyzing this data is laborious and requires significant manual effort. We aim to supplement this research through automatic analysis of manually collected drone imagery.

This research has important ramifications for the MPA and biodiversity in the region. Pacific Nurse Sharks are sometimes found in aggregations of up to 50 individuals, and these aggregations are of particular biological interest since they are potentially relevant for regulation of the social behavior, and courtship of the sharks [29]. This work will help assess factors affecting the unique aggregation behavior, how climate change could potentially affect the species, and the efficacy of the MPA.

In this paper, we present a pipeline for biologists to automatically detect marine animals from aerial drone imagery and compute the length, width, mass, and age of each animal. To do so, we benchmark several SoTA object detectors, which prompt a downstream SAM mask [26]. Then we use a custom heuristic algorithm to calculate shark biometrics. We test our pipeline on a new dataset of Pacific Nurse Shark drone imagery.

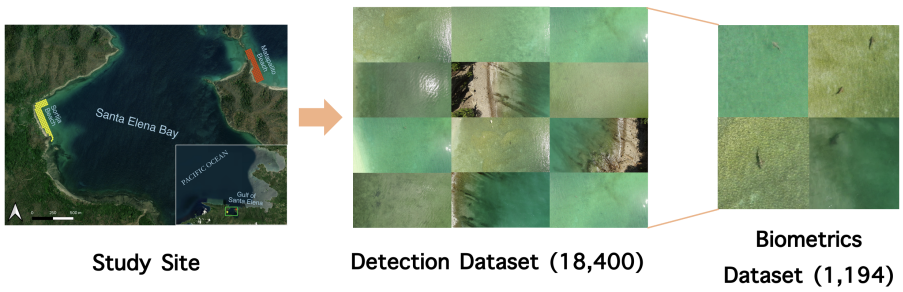


**Fig. 1:** We curated a dataset for detection and biometric analysis of Pacific Nurse Sharks from aerial imagery. The dataset contains 18,400 images for detection. This includes 1,194 images suitable for biometric analysis where the shark is fully visible. Note: images in the biometrics dataset figure are cropped for clarity.

## 2 Related Work

### 2.1 Object Detection

Convolutional Neural Networks (CNNs) have been the foundation of most advances in image recognition and object detection over the past decade [28]. CNNs automatically learn hierarchical feature representations from raw image data, which has proven effective for a wide range of image recognition tasks [1]. The initial layers capture basic image features like edges and textures, while deeper layers learn more complex features specific to the objects within the images.

Two-Stage Object Detectors, built on the architecture of CNNs, have been pivotal in advancing object detection. These models, such as R-CNN and its more advanced successors, Fast R-CNN and Faster R-CNN, prioritize detection accuracy [6, 19, 32]. The two-stage process involves first generating region proposals where objects might exist and then classifying each proposal into object categories while refining bounding box coordinates. Models like Faster R-CNN integrate these stages by introducing a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, significantly speeding up the process while maintaining high accuracy. This architecture is particularly beneficial in scenarios requiring precise localization and detailed contextual understanding.

In contrast, One-Stage Object Detectors such as YOLO (You Only Look Once) [25] prioritize inference speed, making them suitable for real-time applications. These detectors simplify the detection pipeline: they do not require a separate region proposal step and instead treat object detection as a simple regression problem, predicting classes and bounding boxes for the entire image in a single pass. While traditionally less accurate than two-stage detectors, improvements in network design, training schemes, and integration of context have narrowed this gap, making them highly competitive.

More recently, the Detection Transformer (DETR) model integrates transformers into object detection [7]. DETR eliminates the need for many hand-engineered components of traditional object detectors by using a transformer to perform direct set prediction of object classes and bounding boxes. The use of global context by the transformer allows DETR to achieve impressive results rivaling non-transformer architectures [7]. This is particularly useful in scenarios like ours where objects are in close proximity in large groups of sharks or partially occluded via camera glare, high turbidity, or other sea life.

### 2.2 Marine Animal Detection from Aerial Imagery

There has been significant effort towards developing aerial UAV systems for studying wildlife [5, 24, 36]. Several works seek to plan optimal paths for one or multiple drones to perform aerial surveys in varying environmental conditions [9, 36], and others explore the efficacy of various sensor modalities for detecting wildlife [2, 34]. Instead, in this work, we focus on developing automated systems for analyzing existing drone imagery.

One of the first studies applying neural network to aerial imagery of marine life trained a vanilla CNN for sea turtle detection [21]. They achieved a precision of 16.3% and a recall of 76.5% using a threshold value of 0.93 equating to an f1-score of 0.134. A reason for its poor performance, as hypothesized by [21], was a lack of a rich internal feature representation caused by the fact that the model was trained from scratch via their dataset instead of using a transfer learning approach. This model had weaker performance than subsequent works, likely due to lack of pretraining and relatively simple model architecture [22,23,38]. In particular, transfer learning - using a pre-trained CNN with some skip connection has shown to improve model performance as in [15] where they trained Faster R-CNN with a ResNet-101 backbone pre-trained on the COCO dataset to achieve an f1-score of 0.91. Another recent work [37] fine-tuned VGG16, pretrained with ImageNet, to draw bounding boxes around a variety of different objects including sharks in the water. Through their approach, they achieved an average precision of 0.904.

In [44], they propose YOLOv7-sea, an object detector, built from YOLOv7, for maritime search and rescue missions using UAVs. YOLOv7-sea adds a prediction head to detect tiny objects and integrates the Simple, Parameter-Free Attention Module (SimAM) to find attention regions. On their dataset, they achieved 59% average precision across all selected thresholds. With this said, their focus seems to be on objects that humans use in the water like boats, jet skis, buoys, and lifesaving equipment.

Pavithra et al. [33] proposes a novel hybrid architecture called SwinConvMixerUNet for underwater image segmentation, combining Swin Transformer and ConvMixer, leveraging the Swin Transformer's ability to capture spatial information and the ConvMixer's channel-mixing capabilities to enhance feature extraction and segmentation accuracy. The SwinConvMixerUNet outperforms all existing models on the SUIM dataset, achieving 84.83% mean intersection of the union (mIOU). However, their task focuses on the segmentation of all objects in the scene from images captured underwater.

To the best of our knowledge, at this time there is no existing work applying recent object detection models to detecting nurse sharks from drone imagery. This presents an opportunity for increased predictive performance as models like DETR have achieved state-of-the-art in object detection and many other downstream applications [16, 27, 39, 45]. However, to our knowledge this model has not yet been applied to aerial shark imagery.

### 2.3 UAV-Based Animal Biometrics

Collecting marine animal biometrics from drone imagery is of great interest to biologists [31, 41]. Open source software packages such as [41] have provided a semi-automated biometrics workflow for whales, but still require human input to select keypoints. Bierlich et al. [3] proposed two deep learning models to automate data collection on whale body size and condition from UAV imagery. The first model, DeteX, uses the YOLOv5 as a detector to identify images containing whales. The second model, XtraX, uses key point detection and the Segment

Anything Model to extract body length and condition measurements from images selected by DeteX. The authors compared the automated measurements to manual measurements of 63 gray whales and found that the automated method was one-ninth as time-consuming and achieved a mean coefficient of variation of 1.46% between automatic and manual length measurements. In our study, we seek to collect biometrics of animals including length, width, mass, and age. Similar to Gray et al. [20], we use segmentation maps and photogrammetry to estimate these parameters. The aforementioned work used PCA to find the major axis of the whale, but this approach would not work for our dataset since while the tails of whales only deform through camber (dorsoventral undulation), sharks tails deform through torsion (lateral undulation). This lateral undulation requires a different approach, hence our design of a custom heuristic described in section 3.3.

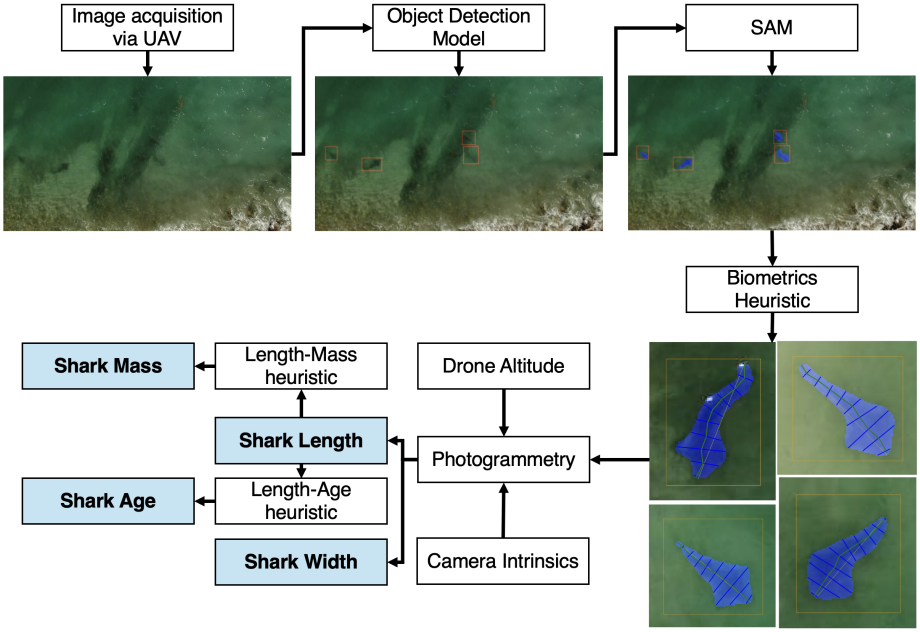


**Fig. 2:** Biometrics pipeline. Outputs are highlighted in blue.

## 3 Methods

### 3.1 Dataset

The imagery was collected from two field sites (Matapalito Beach and Sortija Beach) in the coastal waters of the Eastern Tropical Pacific Ocean, in Santa Elena

Bay (Figure 1). The dataset was collected from 2022-2024, over a period of 23 months, with varying water visibility (turbidity), illumination, and wind/wave conditions. Images were collected at each site by flying a DJI Mavic 2 drone on a preprogrammed path, recording a continuous video at 30FPS and 3840×2160 resolution. The drone stopped at predetermined waypoints for 3 seconds each (Figure 1, yellow and orange dots).

More than 6 hours of video was recorded in total, during 60 drone surveys, resulting in 648,000 total frames captured. For object detection, the dataset was pruned to include a diverse set of 8 videos from the two sites, across varying tidal, turbidity, water surface glare, and wave conditions. Ground truth bounding boxes were added for each Pacific Nurse by two of our team members with the Computer Vision Annotation Tool (CVAT) [12]. The images from 1 video were completely separated from the rest of the dataset, as a test for generalization. The rest of the dataset was time-blocked such that each 45 adjacent frames of a video were placed into a set, and that set was 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. Sharks were present on diverse conditions across videos and time segments, including variable turbidity and substrate. Prior to model training, images were rescaled from 3840 × 2160 pixels to 1080 × 1080 pixels. Standard data augmentation techniques, including random rotation, brightness, and hue adjustments, were also applied.

Excluding the images from the video held for generalization testing, our object detection dataset contained 9200 unique positive examples (images containing sharks) in our dataset, and a corresponding 27000 unique negative examples (images containing no sharks). We selected 9200 negative examples to remain in the dataset, such that our dataset contained an equal ratio of images containing sharks and images not containing sharks. The dataset was then randomly split into training, validation, and test sets, in a [80-10-10] ratio, maintaining the images in the 45-frame time blocks. This dataset contains 18400 total images.

We found that while all labeled sharks were suitable for training object detection models, many frames contained partially occluded sharks. These frames were not possible to extract ground truth biometrics from, so we therefore curated a smaller dataset containing 1194 clear shark images suitable for biometrics as seen in Figure 1.

### 3.2 Detection Models

In this research, we focused on evaluating a series of models for shark detection, predominantly leveraging the Detectron2 model library [43]. First, we employed several RCNN model architectures, a two-stage detector optimized for accuracy (as opposed to inference speed like many one-stage detectors) [19]. Faster RCNN architectures are composed of a feature extractor backbone (typically a pre-trained CNN) followed by a region proposal network to predict object boundaries and their "objectness" scores, and finally region-of-interest (ROI) pooling to

extract uniform-sized features. We trained several RCNN with various backbones (ResNet and ResNeXt architectures) and feature pyramid networks (FPN) to enhance the effect of objects of various scales in images. All of these models were pre-trained on the Imagenet dataset [14].

We also implemented DETR which uses a transformer architecture to reason about all objects in the scene globally, which to our knowledge is the first work using this model for shark detection from aerial imagery [7]. This setup is expected to be advantageous for detecting sharks in highly dynamic and complex environments. Finally, we trained a single-stage pre-trained YOLOv8 model which is particularly well suited for real-time detection tasks [25]. 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 obejcts detected.

### 3.3 Biometrics

The image and object detection bounding boxes were used to prompt the Segment Anything Model (SAM) mask predictor [26] to generate a mask containing shark pixels. We designed a custom algorithm to estimate the length of the shark along its centerline, robust to the orientation and deformation of the shark, as follows:

Given a binary mask $\in \{0,1\}^{m\times n}$, compute the largest contour $C$ and its convex hull $H$. Identify the furthest points $P_1, P_2 \in H$ by calculating the pairwise distances $D_{ij} = \|H_i - H_j\|_2$ and finding the maximum distance. Divide the segment $\overline{P_1P_2}$ into $n$ equal subsegments, and for each dividing point $S_i$, compute the center of mass of the mask $c_i$ along the line which is perpendicular to $\overline{P_1P_2}$ and intersects $S_i$. The shark's length $L$ is calculated as the sum of Euclidean distances between $P_1$, $c_i$, and $P_2$:

$$L = \|P_1 - c_0\| + \sum_{i=0}^{n-2} \|c_i - c_{i+1}\| + \|c_{n-1} - P_2\|. \tag{1}$$

To measure the width at each point, we calculated the perpendicular vector to the centerline at each center of mass point $c_i$. The perpendicular unit vector $\mathbf{u}_i$ is defined as:

$$\mathbf{u}_i = \frac{1}{\|\mathbf{v}_i\|}\begin{pmatrix} -v_{iy} \\ v_{ix} \end{pmatrix} \quad \text{where} \quad \mathbf{v}_i = \frac{(\mathbf{c}_{i+1} - \mathbf{c}_i) + (\mathbf{c}_i - \mathbf{c}_{i-1})}{2} \tag{2}$$

Here, $\mathbf{v}_i$ is the average direction vector at point $i$, calculated as the average of the direction vectors from $c_i$ to $c_{i+1}$ and from $c_{i-1}$ to $c_i$. The width at point $c_i$ is then measured by extending a line along $\mathbf{u}_i$ until it intersects the mask boundaries on both sides of the centerline.

The length $L$ and widths $W$ obtained from the mask, measured in pixels, were converted to centimeters using the following function, modified from [41]:

$$L_{\text{cm}} = \left(\frac{S_w}{I_w} \times \frac{A + D}{F} \times 100\right) \times L_{\text{pixels}} \tag{3}$$

where $S_w$ is the sensor width in mm (13.2 mm for a 1" CMOS sensor), $I_w$ is the image width in pixels (3840 pixels for 3840x2160p resolution), $A$ is the altitude of the drone in meters (37 meters), $D$ is the depth of the shark in meters, and $F$ is the focal length of the camera in mm (28 mm). We assumed a constant depth $D$ of 1.5 m for this analysis, since nurse sharks were typically observed at approximately 1.5 m depth in snorkeling surveys in Santa Elena Bay. In the future, benthic animal position extracted from camera metadata could be collocated with bathymetry data to improve the accuracy of the depth estimate. The pixel dimension in mm per pixel is calculated as the ratio of the sensor width to the image width. This is converted to meters per pixel, and then the Ground Sampling Distance (GSD) in meters per pixel is calculated. Finally, the GSD is converted from meters per pixel to centimeters per pixel to determine the length and widths in centimeters.

We used length to estimate shark age and mass. To estimate the age, we utilized the von Bertalanffy (vB) growth function with known coefficients for the nurse shark [8, 18]. Mass was estimated using a combined mass model for nurse sharks [30]. The full biometrics pipeline is visualized in Figure 2.

## 4 Experiments and Results

### 4.1 Object Detection

We tested our suite of models on an internal test dataset to assess the models' ability to learn representations, as well as a previously-unseen external holdout video taken on a different day to assess the generalizability of the model. Testing the models on both datasets gauges potential overfitting on previously-seen data, while also evaluating for practical usage on drone imagery from new distributions.

Among the models, YOLOv8 and DETR performed the best on the internal holdout test set across all metrics, including mean Average Precision and Average Recall from IOU thresholds between 0.5 to 0.95 and Average Precision and Average Recall at IOU thresholds of 0.35 and 0.5. The exact precision and recall values are shown in Table 1. YOLO had the highest mAP as well as the highest mAR, with DETR performing slightly better at lower IOU thresholds. Both of these performed better than the trained Faster R-CNN models with Feature Pyramid Network and Dilated-C5 backbones. It is important to note that although mAP averaged between 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 in order to obtain accurate results on downstream biometrics methods. Thus, although both YOLO and DETR had a relatively low mAP and mAR of 0.63 and 0.70, performance at a lower IOU threshold of 0.35 was significantly higher—with DETR achieving a near-perfect AP and AR of 0.94 and 0.99, respectively.

As expected, the performance of both YOLOv8 and DETR declined when tested on a completely separate video, with a mAP of 0.22 obtained for YOLOv8

and a mAR of 0.40 obtained for DETR, as seen in Table 2. Similarly, the precision and recall at lower IOU thresholds were significantly higher and proved to be sufficient for deriving downstream biometrics. The DETR model performed the best at an IOU threshold of 0.35, with an average precision of 0.80 and an average recall of 0.96. A higher recall was heavily prioritized for this shark detection task, with the identification of all sharks crucial given the ease of manual removal of potential false positives.

**Table 1:** Performance Metrics of Various Object Detection Models on a holdout test set. Best performance for each metric in bold, with larger values indicating higher average precision and recall.

| Model | mAP@[0.5:0.95] | AP@0.35 | AP@0.50 | mAR@[0.5:0.95] | AR@0.35 | AR@0.50 |
|---|---|---|---|---|---|---|
| YOLOv8 | **0.63** | 0.93 | **0.92** | **0.70** | 0.97 | **0.97** |
| Faster R-CNN X101-FPN | 0.48 | 0.63 | 0.61 | 0.70 | 0.99 | 0.60 |
| Faster R-CNN R101-DC5 | 0.29 | 0.43 | 0.43 | 0.35 | 0.43 | 0.41 |
| DETR | 0.43 | **0.94** | 0.90 | 0.58 | **0.99** | 0.97 |

**Table 2:** Performance Metrics of YOLOv8 and DETR on holdout video. Best performance for each metric in bold.

| Model | mAP@[0.5:0.95] | AP@0.35 | AP@0.50 | mAR@[0.5:0.95] | AR@0.35 | AR@0.50 |
|---|---|---|---|---|---|---|
| YOLOv8 | **0.22** | 0.77 | **0.64** | 0.34 | 0.79 | 0.73 |
| DETR | 0.14 | **0.80** | 0.53 | **0.40** | **0.96** | **0.85** |

The effect of IOU between the predicted and ground truth boxes versus recall was studied for YOLOv8 on the holdout video, as shown in Figure 3. We see that the recall is stable through an IOU threshold of 0.4, then rapidly declines as IOU increases. This shows that we detect nearly 80% of the sharks in the holdout video at lower IOU values, thus motivating the non-standard AP and AR calculations at an IOU of 0.35. Precision-recall curves were generated for YOLOv8 and DETR to study the overall performance at this lower IOU threshold on the internal and external test datasets (see Figure 4). We see that both models performed better on the internal dataset, with high precision and recall achieved at lower confidence thresholds. Similarly, both YOLOv8 and DETR had relatively high performance on the holdout video, excluding low precision at lower confidence thresholds. These results show the large number of low-confidence predictions for both models under a confidence threshold of 0.2, with high precision and recall maintained at higher confidence values.

Although both models had high accuracy at lower IOU thresholds, DETR outperformed YOLOv8 on the holdout video with a nearly perfect recall of 0.96 on completely unseen data. Beyond a confidence threshold of 0.25, precision was maintained; thus all further biometric analysis was performed using YOLOv8 predicted bounding boxes with a confidence threshold of 0.25.

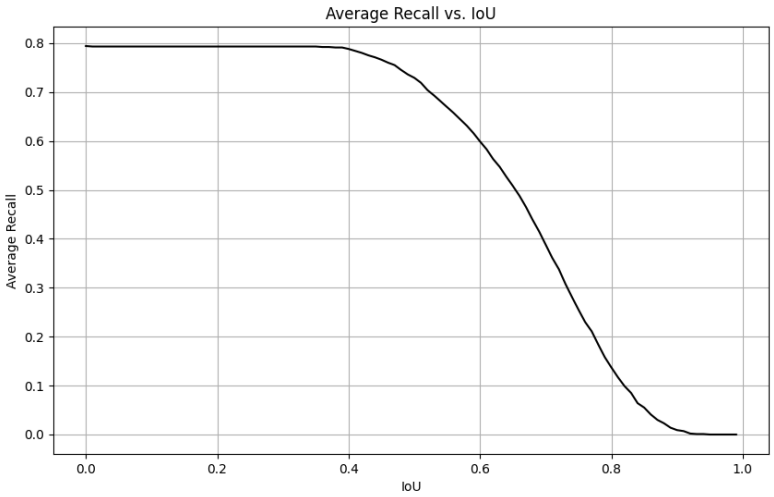


**Fig. 3:** Recall vs. IOU threshold for YOLOv8 on the holdout video

### 4.2 Biometrics

We compared the performance of our approach to manually measured shark lengths from a randomly selected subset of photos (n = 122) containing sharks from the holdout video. We benchmarked our approach based on percent error and coefficient of variation (CV, following Bierlich et al [3]) between the manually measured and automatically measured shark lengths for each image. In total, we found that SAM, when prompted with YOLOv8 predicted bounding boxes, successfully generated correct segmentations on 71% of the photos (n = 87), while it failed to generate a correct segmentation on 29% of the photos (n = 35). Examples of successful and failed segmentations are found in (Figure 5 (b) and (c)). The manually-sorted criteria for successful segmentation are that the shark is included in the segmentation mask symmetrically about its true centerline, and that there are no erroneous inclusions (shadows, surface glare) or exclusions (tip of the tail, a single pectoral fin) in the mask. We report results for the automated and manual measurements for the images that SAM generated correct shark segmentations. All length measurements had a CV $< 8\%$, with an average CV between automatic and manual length measurements of 2.71% (n = 87, median = 2.70%, SD = 1.54%, maximum = 5.55%) (Figure 5 (a)). Likewise, the length measurements had a percent error $< 12\%$, with an average percent error between manual and automatic measurements of 5.54% (n = 87, median = 5.56%, SD = 1.54%, maximum = 11.76%. This represents only a slight decrease in performance compared to ground truth bounding boxes (from which measurements had an average CV of 2.31%), showing that exact bounding box predictions are not necessary for prompting accurate segmentations from

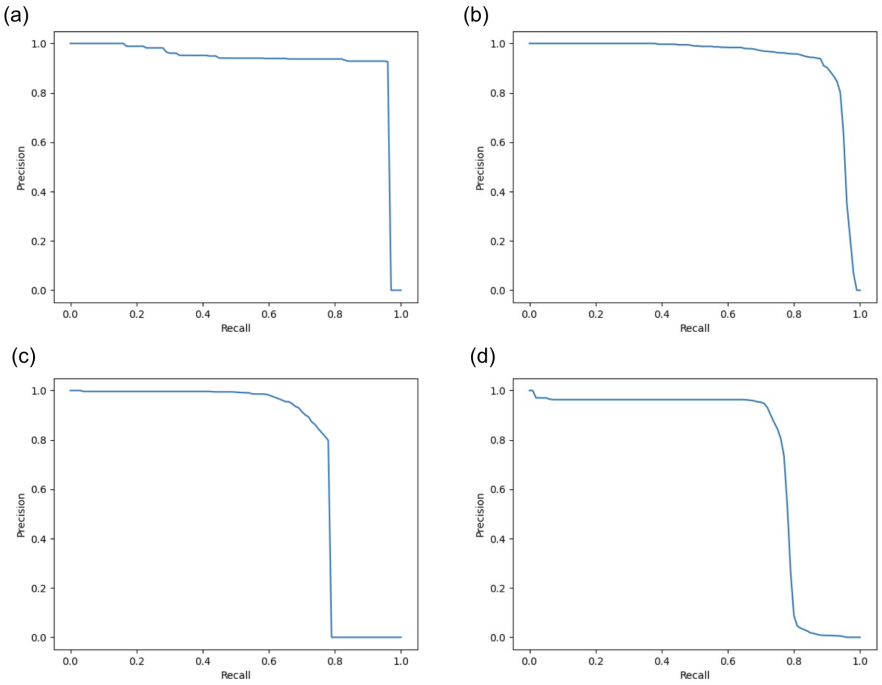


**Fig. 4:** Precision-Recall curves for models on internal and external test datasets. (a) YOLOv8 on internal test set. (b) DETR on internal test set. (c) YOLOv8 on holdout video. (d) DETR on holdout video.

SAM. On average, the automatic measurements overestimated length by 2.47 cm (n = 87, median = 2.79 cm, SD = 3.91 cm, minimum = -8.48 cm, maximum = 7.69 cm). The average percent error between manual-length derived mass estimates and automatic-length derived mass estimates was 11.85% (n = 87, median = 11.81%, SD = 6.97%, maximum = 26.09%). The average CV for mass estimates was 5.64% (n = 87, median = 5.63%, SD = 3.20%, maximum = 11.54%). Similarly, the average percent error between manual-length derived age estimates and automatic-length derived age estimates was 9.23% (n = 87, median = 9.45%, SD = 5.36%, maximum = 19.91%). The average CV for age estimates was 4.45% (n = 87, median = 4.51%, SD = 2.53%, maximum = 9.05%).

### 4.3 User comparisons

To compare the accuracy and speed of our shark detector to a non-expert human, we randomly drew 100 samples from the holdout video dataset with an even ratio of positive and negative examples. After showing the person examples of negative and positive examples, and basics of CVAT labeling workflow, they labeled the dataset of 100 samples. We found that the human had a classification precision

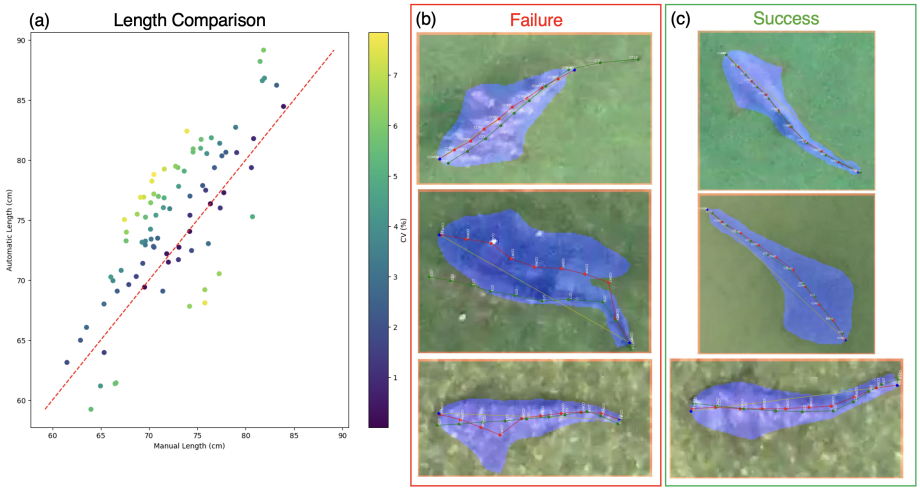


**Fig. 5:** (a) Manual and automatic length measurement results. The dashed red line represents a 1:1 ratio. Points are colored by coefficient of variation percentage (CV %), where lighter color points indicate a higher CV. (b) and (c) include manual (green) and automatic (red) polylines used to measure the length of each shark. (b) Example failure cases of SAM segmentations, from top to bottom: Missing the end of the tail, segmenting a shadow, and segmenting only one pectoral fin. (c) Examples of successful SAM segmentations and length measurement.

of 0.94 and a recall of 0.83 on the dataset. Our best model had a classification precision of 0.8 and a recall of 0.96 on the same dataset. The human took an average of 10.5 seconds per frame, while the model only takes an average of 0.25 seconds on an NVIDIA A10 GPU. The annotator also measured sharks in 100 images using the CVAT polyline tool in approximately 50 minutes, so on average, it took a human annotator 40.5 seconds on average to detect and measure the sharks in each frame. In contrast, the automatic biometrics pipeline takes an average of 3.75 seconds per image on an NVIDIA A10 GPU. These preliminary results show our model takes 91% less time than a human to perform inference and shark measurement, while achieving human-level performance on classification recall and measurement accuracy.

## 5 Conclusion

In this study, we develop a new pipeline which significantly improves the efficiency of studying Pacific Nurse Sharks (*Ginglymostoma unami*) through aerial drone imagery. We benchmark several SoTA object detection models such as YOLOv8, Faster R-CNN, and DETR, and conclude that DETR performed the best - enabling 91% faster data processing than manual annotation methods. To our knowledge, this is one of the largest datasets of nearshore shark aerial drone imagery, and the first shark detection work using DETR.

We also demonstrate the efficacy of the pre-trained SAM model on segmenting sharks when prompted by bounding boxes. Using a custom heuristic, we analyze SAM masks to compute marine animal length, width, mass, and estimated age—which are essential for understanding shark population demographics. We found that SAM generated successful segmentations 77% of the time. On 27 randomly sampled images from our biometrics dataset, our length heuristic had a 2.39% average CV from the manual labels. Our biometrics heuristic, when combined with the shark detection model, provides researchers a way to automatically estimate shark length to centimeter-scale accuracy from drone imagery. However, there are limitations and opportunities for improvement for the biometrics pipeline. The primary limiting factor is the accuracy of the segmentation, as that has downstream effects on the length and width prediction. In turbid water, often the pectoral fins of the shark and the tail will be left out of the segmentation, or the end of the tail will be segmented separately from the rest of the body. Also, when the shark is swimming above the bottom, SAM will sometimes segment the shadow of the shark instead of the shark itself (Figure 5 (b)). For simplicity, our photogrammetric analysis did not include the variable depth of water the sharks are swimming in, the variance in drone altitude, or corrections due to refraction at the air-water interface, but all of these considerations could feasibly be added if additional measurement accuracy is required.

A key feature of our work is the centerline estimation heuristic, which is novel for marine animals and has broad applicability beyond just measuring nurse shark biometrics. This method is not only valuable for length measurements but also for animal pose estimation, enabling researchers to compute important new metrics such as tailbeat frequency and amplitude.

Future work in object detection will involve testing different data augmentation approaches and exploring different DETR variants. There remains a large amount of unlabeled shark video data in our dataset, so including a wider swath of this data in the training will likely improve detection results. Further development on the biometrics pipeline will incorporate object tracking to maintain the identity of sharks between frames, which would allow biologists to observe a shark's speed, tailbeat frequency, tailbeat amplitude, and other kinematic observations from aerial imagery. These observations are particularly relevant for understanding shark behavior and physiology on a fine scale, and they would allow biologists to quantitatively study the movement patterns of sharks in aggregations. By finetuning SAM on labeled segmentation data, we can extract better masks to feed into the biometrics heuristic, which will improve the quality of results for kinematic observations, which otherwise would require laborious data labeling efforts.

## 6 Examples

An output video generated from our shark detection and biometrics pipeline can be found in on YouTube.

## 7 Acknowledgements

Many thanks to Chris Lowe for supporting data collection. Also, thank you to the participants of our user comparison study.

## References


1. 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.: Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. Journal of Big Data **8**(53) (2021), https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00444-8
2. Beaver, J.T., Baldwin, R.W., Messinger, M., Newbolt, C.H., Ditchkoff, S.S., Silman, M.R.: Evaluating the use of drones equipped with thermal sensors as an effective method for estimating wildlife. Wildlife Society Bulletin **44**(2), 434–443 (2020)
3. Bierlich, K.C., Karki, S., Bird, C.N., Fern, A., Torres, L.G.: Automated body length and body condition measurements of whales from drone videos for rapid assessment of population health. Marine Mammal Science **n/a**(n/a), e13137. https://doi.org/https://doi.org/10.1111/mms.13137, https://onlinelibrary.wiley.com/doi/abs/10.1111/mms.13137
4. Brock, P.M., Hall, A.J., Goodman, S.J., Cruz, M., Acevedo-Whitehouse, K.: Immune activity, body condition and human-associated environmental impacts in a wild marine mammal. PLoS ONE **8**(6), e67132 (Jun 2013). https://doi.org/10.1371/journal.pone.0067132, http://dx.doi.org/10.1371/journal.pone.0067132
5. Butcher, P.A., Colefax, A.P., Gorkin III, R.A., Kajiura, S.M., López, N.A., Mourier, J., Purcell, C.R., Skomal, G.B., Tucker, J.P., Walsh, A.J., et al.: The drone revolution of shark science: A review. Drones **5**(1), 8 (2021)
6. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
8. Castro, J.I.: The biology of the nurse shark, ginglymostoma cirratum, off the florida east coast and the bahama islands. Environmental Biology of Fishes **58**, 1–22 (2000)
9. Clark, J., Shah, K., Schwager, M.: Online path repair: Adapting to robot failures in multi-robot aerial surveys. IEEE Robotics and Automation Letters (2024)
10. Claro, F., Fossi, M., Ioakeimidis, C., Baini, M., Lusher, A., McFee, W., McIntosh, R., Pelamatti, T., Sorce, M., Galgani, F., Hardesty, B.: Tools and constraints in monitoring interactions between marine litter and megafauna: Insights from case studies around the world. Marine Pollution Bulletin **141**, 147–160 (2019)
11. Cortés, J., Jiménez, C.E., Fonseca, A.C., Alvarado, J.J.: Status and conservation of coral reefs in costa rica. Revista de Biología Tropical **58**, 33–50 (2010)
12. CVATteam: Computer vision annotation tool (cvat). https://github.com/opencv/cvat (2020)
13. Del Moral-Flores, L.: Ginglymostoma unami sp. nov. (chondrichthyes: Orectolobiformes: Ginglymostomatidae): a new species of nurse shark from the tropical eastern pacific. Revista mexicana de biodiversidad **86**, 48–58 (2015)
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

15. Desgarnier, L., Mouillot, D., Vigliola, L., Chaumont, M., Mannocci, L.: Putting eagle rays on the map by coupling aerial video-surveys and deep learning. Biological Conservation **267**, 109494 (2022)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), https://arxiv.org/abs/2010.11929
17. Espinoza, M., Araya-Arce, T., Chaves-Zamora, I., Chinchilla, I., Cambra, M.: Monitoring elasmobranch assemblages in a data-poor country from the eastern tropical pacific using baited remote underwater video stations. Scientific Reports **10**(1), 17175 (Oct 2020). https://doi.org/10.1038/s41598-020-74282-8, [invalidURLremoved]
18. Fadool, B.A., Bostick, K.G., Brewster, L.R., Hansell, A.C., Carlson, J.K., Smukall, M.J.: Age and growth estimates for the nurse shark (ginglymostoma cirratum) over 17 years in bimini, the bahamas. Frontiers in Marine Science **11** (2024). https://doi.org/10.3389/fmars.2024.1265150, https://www.frontiersin.org/articles/10.3389/fmars.2024.1265150
19. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
20. Gray, P.C., Bierlich, K.C., Mantell, S.A., Friedlaender, A.S., Goldbogen, J.A., Johnston, D.W.: Drones and convolutional neural networks facilitate automated and accurate cetacean species identification and photogrammetry. Methods in Ecology and Evolution **10**(9), 1490–1500 (2019)
21. Gray, P.C., Fleishman, A.B., Klein, D.J., McKown, M.W., Bezy, V.S., Lohmann, K.J., Johnston, D.W.: A convolutional neural network for detecting sea turtles in drone imagery. Methods in Ecology and Evolution **10**(3), 345–355 (2019)
22. He, K., Sun, J.: Convolutional neural networks at constrained time cost (2014), https://arxiv.org/abs/1412.1710
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015), https://arxiv.org/abs/1512.03385
24. Hodgson, J.C., Mott, R., Baylis, S.M., Pham, T.T., Wotherspoon, S., Kilpatrick, A.D., Raja Segaran, R., Reid, I., Terauds, A., Koh, L.P.: Drones count wildlife more accurately and precisely than humans. Methods in Ecology and Evolution **9**(5), 1160–1167 (2018)
25. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO (Jan 2023), https://github.com/ultralytics/ultralytics
26. 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.: Segment anything. arXiv:2304.02643 (2023)
27. Li, C., Li, L., Geng, Y., Jiang, H., Cheng, M., Zhang, B., Ke, Z., Xu, X., Chu, X.: Yolov6 v3.0: A full-scale reloading (2023), https://arxiv.org/abs/2301.05586
28. Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J.: A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems **33**(12), 6999–7019 (2021)
29. Madrigal-Mora, S., Chávez, E.J., Arauz, R., Lowe, C.G., Espinoza, M.: 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), NULL–NULL (2024)
30. Natanson, L.J., McCandless, C.T., Passerotti, M.S., Belcher, C.N., Bowlby, H., Driggers III, W.B., Frazier, B.S., Gelsleichter, J., Gulak, S.J.B., Hendon, J.M.,

Hoffmayer, E.R., Joyce, W.: Morphometric conversions for 33 shark species from the western north atlantic ocean. Marine Fisheries Review **84**(3-4) (2022). https://doi.org/10.7755/MFR.84.3-4.1
31. Petso, T., Jamisola Jr., R.S., Mpoeleng, D., Bennitt, E., Mmereki, W.: Automatic animal identification from drone camera based on point pattern analysis of herd behaviour. Ecological Informatics **66**, 101485 (2021)
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
33. S, P., Denny J, C.M.: An efficient approach to detect and segment underwater images using swin transformer. Results in Engineering **23**, 102460 (2024). https://doi.org/https://doi.org/10.1016/j.rineng.2024.102460, https://www.sciencedirect.com/science/article/pii/S2590123024007151
34. Saunders, D., Nguyen, H., Cowen, S., Magrath, M., Marsh, K., Bell, S., Bobruk, J.: Radio-tracking wildlife with drones: a viewshed analysis quantifying survey coverage across diverse landscapes. Wildlife Research **49**(1), 1–10 (2022)
35. Sequeira, A.M.M., Thums, M., Brooks, K., Meekan, M.G.: Error and bias in size estimates of whale sharks: implications for understanding demography. Royal Society Open Science **3**(3), 150668 (Mar 2016). https://doi.org/10.1098/rsos.150668, http://dx.doi.org/10.1098/rsos.150668
36. Shah, K., Ballard, G., Schmidt, A., Schwager, M.: Multidrone aerial surveys of penguin colonies in antarctica. Science Robotics **5**(47), eabc3000 (2020)
37. Sharma, N., Scully-Power, P., Blumenstein, M.: Shark detection from aerial imagery using region-based cnn, a study. In: AI 2018: Advances in Artificial Intelligence: 31st Australasian Joint Conference, Wellington, New Zealand, December 11-14, 2018, Proceedings 31. pp. 224–236. Springer (2018)
38. Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway networks (2015), https://arxiv.org/abs/1505.00387
39. Srivastava, S., Sharma, G.: Omnivec: Learning robust representations with cross modal sharing (2023), https://arxiv.org/abs/2311.05709
40. 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.: Range-wide comparison of gray whale body condition reveals contrasting sub-population health characteristics and vulnerability to environmental change. Frontiers in Marine Science **9** (Apr 2022). https://doi.org/10.3389/fmars.2022.867258, http://dx.doi.org/10.3389/fmars.2022.867258
41. Torres, W., Bierlich, K.: Morphometrix: a photogrammetric measurement gui for morphometric analysis of megafauna. Journal of Open Source Software **4**(44), 1825 (2020). https://doi.org/10.21105/joss.01825
42. Webb, P.M., Crocker, D.E., Blackwell, S.B., Costa, D.P., Boeuf, B.J.L.: Effects of buoyancy on the diving behavior of northern elephant seals. Journal of Experimental Biology **201**(16), 2349–2358 (Aug 1998). https://doi.org/10.1242/jeb.201.16.2349, http://dx.doi.org/10.1242/jeb.201.16.2349
43. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
44. Zhao, H., Zhang, H., Zhao, Y.: Yolov7-sea: Object detection of maritime uav images based on improved yolov7. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). pp. 233–238 (2023). https://doi.org/10.1109/WACVW58289.2023.00029
45. Zong, Z., Song, G., Liu, Y.: Detrs with collaborative hybrid assignments training (2023), https://arxiv.org/abs/2211.12860