When Segment and Track Anything Meets Wildlife Videos

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

Master Thesis of Huiyi Wang

Supervisors: Prof. Dr. Andreas Weinmann, Prof. Dr. Elke Hergenröther

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbi
FACHBEREICH INFORMATIK

Abstract

- The increasing loss of endangered species is becoming one of the most concerning issues of the ongoing biodiversity and ecosystems crisis.
- The advancements in the field of computer vision have also contributed to the research on wildlife protection.
- The complex nature surroundings of wildlife pose challenges to research and analysis. For instance, the identification of wildlife might be complicated by its complex habitat.
- The goal of this thesis is to propose a general framework for the automatic segmentation of wildlife in video sequences that is not limited to only a few predetermined species, so that the effect of complex surroundings could be reduced.
- The framework is tested quantitatively with five high-resolution leopard video clips from the Pan African Programme[1] and achieves a score (Mask IoU between predicted masks and ground-truth masks) of over 85%. Moreover, the framework is tested qualitatively with two YouTube low-resolution videos that contain multiple overlapping animals. The results are reliable in the majority of cases.

Architecture

The framework contains three main components:

- **Detector:** The detector generates bounding boxes, which are used as box prompts for the segmentor. The YOLOV5-based "**MegaDetecor**"[2] is employed as detector given its efficient performance that has already been proven in previous research.
- **Segmentor:** Using box prompts received from the detector, the segmentor produces object masks, which are subsequently used as initialized masks in the tracker. "**SAM**"[3] is applied as the segmentor due to its impressive zero-shot capabilities as a segmentation foundation model.
- **Tracker:** The tracker uses the initialized segmentation masks to track the annotated objects. "**Cutie**"[4] serves as tracker since it has proven to be more effective than the previous VOS (Video Object Segmentation) models in both accuracy and performance time.

In addition, a **matching algorithm** and **post-processing** are implemented to address the overlapping issue.

Methodology

- **Initialization:** The first frame of an input video is fed into the framework. MegaDetector generates bounding boxes for any visible wildlife. Smaller animals are segmented first if overlapping exists, i.e., two bounding boxes have a Box IoU of over 0.9. Using sorted box prompts, "SAM" produces the segmentation masks, which are employed as the initialized reference masks to follow the objects in the 1*st* frame until the *nth* frame.
- **Update and Matching Algorithm:** Every *nth* frame, based on SAM segmentation mask and Cutie track mask, the reference masks are updated for three reasons: (1) Appearance of new targets that are to be tracked. (2) Segmentation masks of already existing targets have become imprecise and need to be refined. (3) Prevention of existing targets from being mis-detected.

  During the update phase, a **matching algorithm** using Mask IoU and inclusion rates is applied to guarantee tracking consistence and address the overlapping issue.
- **Post-process:** Retrack the targets in the reversed first n frames and correct the masks of the first n frames in case that animals are overlapping in the first initialized frame.

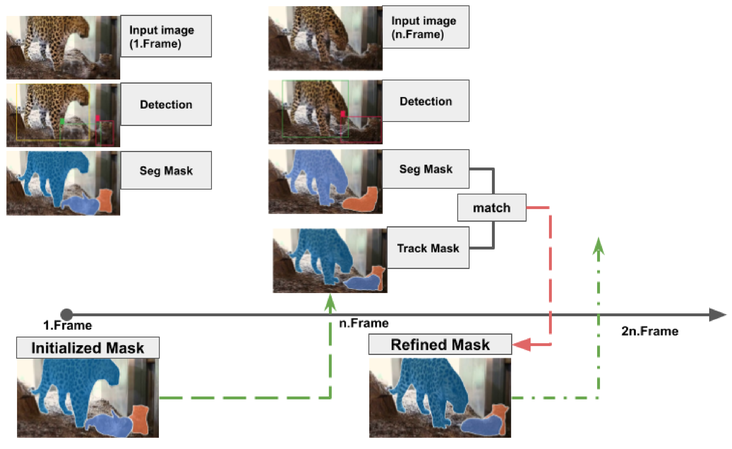
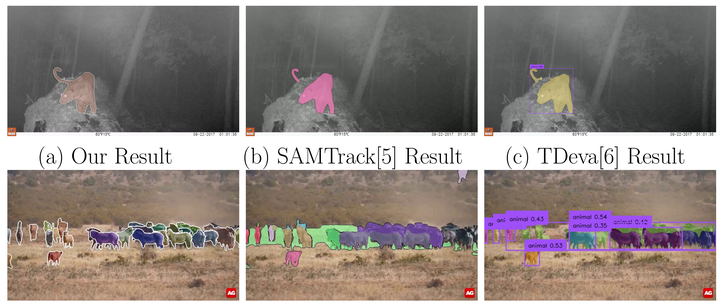


Figure 1: Framework

Results

Five high-resolution video clips of leopards, each with Ground-Truth masks, are used here to qualitatively evaluate the performance of the proposed framework. The test videos were filmed in a complex environment with dense forests and under challenging illumination conditions. A total of 49 frames are used for the evaluation. For all of the test frames, it achieves a score (Mask IoU between the predicted masks and the ground-truth masks) of over 85%. With regard to the average quantitative performance, the framework in this thesis is almost on a par with frameworks suggested in previous works. In the frames containing obstacles, the proposed framework in this thesis performs slightly better.

Besides high-resolution videos, low-resolution videos showing multiple overlapping animals are selected to test the robustness of the framework. The qualitative results meet the expectations.



(a) Our Result (b) SAMTrack[5] Result (c) TDeva[6] Result

(a) Our Result (b) SAMTrack[5] Result (c) TDeva[6] Result

Figure 2: Performance Comparison

Future Work

Domain adaption on "SAM" could be employed, so that the proposed framework might work more robustly for wildlife in camouflaged natural surroundings or nocturnal animals captured on night vision images.

References

[1] P. Programme, `http://panafrican.eva.mpg.de/`, [Online; accessed 16-November-2023], 2023.

[2] MegaDetector, `https://github.com/microsoft/CameraTraps/releases/tag/v5.0`, [Online; accessed 12-December-2023], 2022.

[3] A. Kirillov, E. Mintun, N. Ravi, *et al.*, *Segment anything*, 2023. arXiv: `2304.02643 [cs.CV]`.

[4] H. K. Cheng, S. W. Oh, B. Price, J.-Y. Lee, and A. Schwing, *Putting the object back into video object segmentation*, 2023. arXiv: `2310.12982 [cs.CV]`.

[5] Y. Cheng, L. Li, Y. Xu, *et al.*, *Segment and track anything*, 2023. arXiv: `2305.06558 [cs.CV]`.

[6] H. K. Cheng, S. W. Oh, B. Price, A. Schwing, and J.-Y. Lee, *Tracking anything with decoupled video segmentation*, 2023. arXiv: `2309.03903 [cs.CV]`.