High Quality Video Object Tracking and Segmentation

-By E Gabriel Jomar & Krishi Thiruppathi

➔ Reference papers:

- Tracking Anything in High Quality : Jiawen Zhu , Zhenyu Chen , Zeqi Hao, Dalian University of Technology, China, DAMO Academy, Alibaba Group on HQTrack consisting of VMOS and MR.
  arXiv:2307.13974v1[cs.CV] 26 Jul 2023

- Segment and Track Anything : Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li on Computer Vision and Pattern recognition.
  https://arxiv.org/abs/2305.06558 May 2023

➔ Dataset Description and Link:

1. GIF datasets:

https://github.com/jiawen-zhu/HQTrack/tree/main/assets
The Object Tracking GIF Dataset is a curated collection of GIF videos designed specifically for the purpose of object tracking research and development. This dataset serves as a valuable resource for individuals and teams engaged in computer vision, machine learning, and artificial intelligence projects that involve object tracking algorithms.

2. CAMEL Dataset:

https://camel.ece.gatech.edu/
The Camel Dataset provides a benchmark for visual-infrared object detection.

3. CRCV Real-World Anomaly Detection Dataset:

https://www.crcv.ucf.edu/projects/real-world/
The CRCV Real-world Anomaly Detection Dataset consists of 1900 long and untrimmed real-world surveillance videos, with 13 realistic anomalies such as fighting, road accident, burglary, robbery, etc. as well as normal activities.

➔ Recent base architecture and explanation:

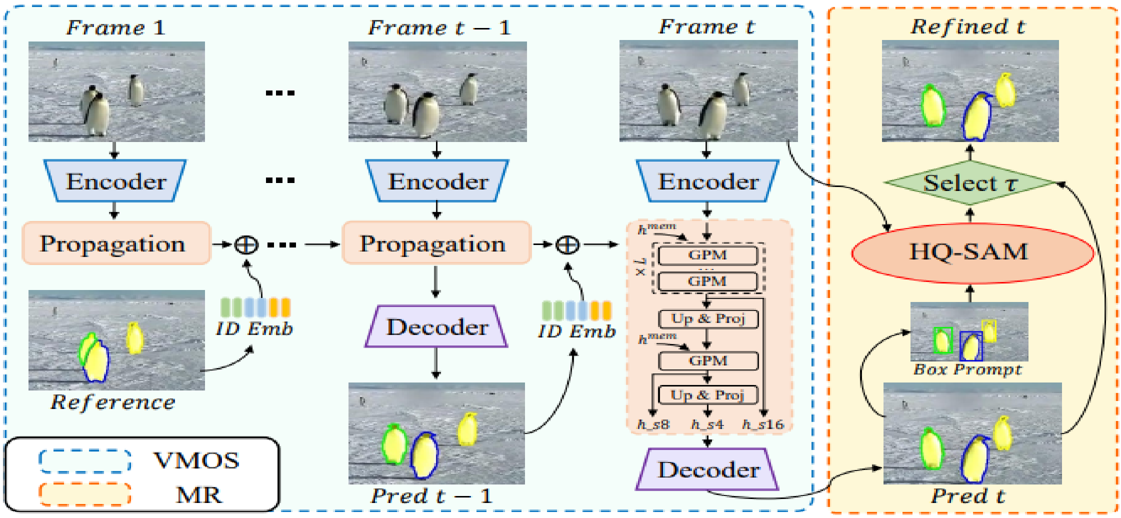


Figure 1. Overview of HQTrack. It mainly consists of a video multi-object segmenter (VMOS) and a mask refiner (MR).

1. **Video Multi-Object Segmenter (VMOS):** The VMOS component is responsible for tracking multiple target objects in video frames. It is an improved variant of the DeAOT (Adaptive Object Tracking) method. The architecture of VMOS involves the following key components and steps:

   a. **Gated Propagation Module (GPM):** VMOS incorporates a Gated Propagation Module that allows the perception of small objects even in complex scenarios. In HQTrack, the GPM is cascaded with different scales (8× scale) to expand the propagation process.

   b. **Intern-T Backbone:** For feature extraction and enhancing object discrimination, the Intern-T backbone is employed. Intern-T is a large-scale CNN-based foundation model that uses deformable convolution as a core operator.

   c. **Multi-Scale Propagation and Encoder-Decoder:** The original DeAOT propagates visual and identification features at a 16× scale. In VMOS, multi-scale propagation is introduced, including up-sampling and linear projection to upscale the propagation features to 4× scale. These multi-scale propagation features are combined with encoder features in a decoder, which follows a Feature Pyramid Network (FPN) architecture.

d. **Long and Short-Term Memory:** VMOS employs long and short-term memory mechanisms to handle appearance changes in long-term video sequences. A fixed length of long-term memory is used, and a memory gap parameter is specified to control the range of frames retained in memory.

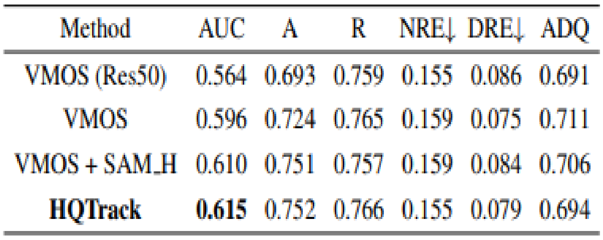
2. **Mask Refiner (MR):** The MR component utilizes a pre-trained HQ-SAM (High-Quality Segment Anything Model) to refine the segmentation masks generated by VMOS. The architecture of MR includes the following steps:

a. **Input Mask Prediction:** The segmentation masks predicted by VMOS serve as inputs to the MR component.

b. **HQ-SAM Refinement:** HQ-SAM is a variant of SAM (Segment Anything Model), which is trained on a high-quality annotated dataset containing a large number of masks. HQ-SAM refines the segmentation masks, providing high-quality and accurate results.

c. **Mask Selection:** The refined masks from HQ-SAM are compared with the masks predicted by VMOS using an IoU (Intersection over Union) threshold. If the IoU score is above the threshold, the refined mask is selected as the final output. This selective approach ensures that HQ-SAM focuses on refining masks that can benefit from the refinement process.

➔ **Result Analysis (as given in research paper) & Scope for improvement:**



Method	AUC	A	R	NRE↓	DRE↓	ADQ
VMOS (Res50)	0.564	0.693	0.759	0.155	0.086	0.691
VMOS	0.596	0.724	0.765	0.159	0.075	0.711
VMOS + SAM_H	0.610	0.751	0.757	0.159	0.084	0.706
HQTrack	**0.615**	0.752	0.766	0.155	0.079	0.694

Table 5. Performance on VOTS2023 test set.

HQTrack achieves an AUC of 0.615, outperforming VMOS + SAM H by 0.9%.

Incorporation of advanced techniques, such as joint tracking, mask refinement, and optimised encoders will contribute in improving the results.

➔ Applicability

Surveillance and Security: This project can be employed in surveillance systems to track individuals or objects of interest within a monitored area. Its ability to handle occlusions and maintain accurate tracking enhances the overall security and monitoring process.

Autonomous Vehicles: In the realm of self-driving cars and drones, this project can play a pivotal role in identifying and tracking pedestrians, vehicles, and obstacles. Accurate tracking is crucial for ensuring safe navigation and collision avoidance.

Sports Analysis: This project can be utilised in sports analysis to track players' movements during games. This data can provide valuable insights for coaches, analysts, and viewers.

Robotics: Robots can navigate complex environments, interact with objects, and collaborate with humans more effectively. This has applications in areas like warehouse automation, logistics, and manufacturing.

Healthcare: This project can be used in medical imaging to track and segment anatomical structures or medical instruments in real-time during surgical procedures, aiding surgeons and improving patient outcomes.

Entertainment and Filmmaking: The framework can enhance special effects and post-production processes by enabling accurate object tracking and manipulation within video footage.

Augmented Reality (AR) and Virtual Reality (VR): This project can contribute to immersive AR and VR experiences by accurately tracking and interacting with virtual objects in real-time.

Human-Computer Interaction: In interactive systems, this project can track hand gestures, body movements, and facial expressions, enabling more natural and intuitive interactions.