Efficient Video Object Segmentation via Modulated Cross-Attention Memory

Supplementary Material

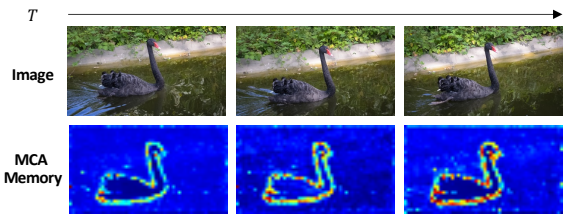

Figure 6. **Interpretation of the MCA memory**. Interpreting the dynamic frame of the MCA memory reveals its ability to encode temporal smoothness over time along the boundaries.

We provide additional details regarding:

- Architecture Details of MAVOS
- Additional Ablation
- More Qualitative Results
- Limitations

1. Architecture Details of MAVOS

The baseline DeAOT [63] is designed with four network variants. DeAOT-T/S/B are tailored for short videos, they consider only the reference frame as the long-term memory, leading to consistent FPS and memory but poor accuracy on long videos because the temporal context is limited. DeAOT-L is designed for both short and long-term videos, it updates long-term memory by appending a new memory frame representation for each δ number of frames (set to 2/5 for training/testing). Since our motivation is to propose an efficient method for long-term videos, we introduce a single efficient network, called MAVOS, equivalent to DeAOT-L in terms of all hyper-parameters and the number of LSTT/E-LSTT blocks, which are set to three blocks.

We propose three variants of MAVOS based on three visual encoders (MobileNet V2 [45], ResNet-50 [15], and Swin-Base [27]). Identification assignment (ID) of [63] is used to transfer the target masks into an identification embedding. Both visual and identification embeddings are propagated to the two branches of the proposed Efficient Long Short-Term Transformer (E-LSTT) block. The visual branch matches objects and propagates visual features from previous frames. The ID branch reuses the attention maps of the visual branch to propagate the ID embedding from past frames to the current frame. The masks are predicted through the decoder using the same Feature Pyramid Network (FPN) [25], as in [63].

2. Additional Ablation

We ablate in Table 7 the effect of different long-term memory frames. The proposed MAVOS, based on MCA memory, achieves promising performance across all datasets compared to other memory frames. In our evaluation on the LVOS validation set [51] presented in Table 8, we also conduct an ablation to examine the impact of varying the number of focal stages in the proposed MCA memory. When employing two focal levels, R50-MAVOS achieves a performance of 63.3% with a processing speed of 37.1 FPS. Notably, reducing the number of focal levels to one results in a marginal increase in FPS (1.4 FPS); however, this improvement is accompanied by a 0.8% drop in performance. The introduction of a third focal level leads to a further decrease in FPS. This trend suggests that employing three focal levels may be less advantageous, potentially diverting attention towards high-level features at the expense of low-level features. This is less helpful here, since the MCA memory already encodes high-level features through the attention mechanism.

Memory Frames	$\mathcal{J}\&\mathcal{F}\uparrow$			FPS ↑	Mem (GB)↓
	DAVIS	LTV	LVOS		
Ref	81.1	79.5	54.6	42.5	3.2
Previous	80.2	74.6	37.3	42.9	3.2
Ref + Previous	82.5	81.2	57.2	39.1	3.3
MCA (Ref + Dynamic)	**84.4**	**87.4**	**60.9**	38.2	3.3

Table 7. Ablation with different memory frames on short and long-term benchmarks.

Focal Levels	$\mathcal{J}\&\mathcal{F}\uparrow$	FPS ↑
1	62.5%	**38.5**
2	**63.3%**	37.1
3	63.1%	35.4

Table 8. Ablation for the number of focal levels in MCA memory of R50-MAVOS on LVOS validation set. The largest value is in bold.

3. More Qualitative Results

To interpret our MCA memory, we show in Fig. 6 the visual representation of its dynamic frame. It illustrates that the MCA memory effectively encodes temporal smoothness over time along the boundaries of the black swan. We also

Figure 7. **Qualitative result for R50-MAVOS on the Long-Time Video dataset [24]**. Our R50-MAVOS demonstrates good segmentation performance for sequences with more than two thousand frames at 38 FPS, accurately segmenting the target despite the fast movement.

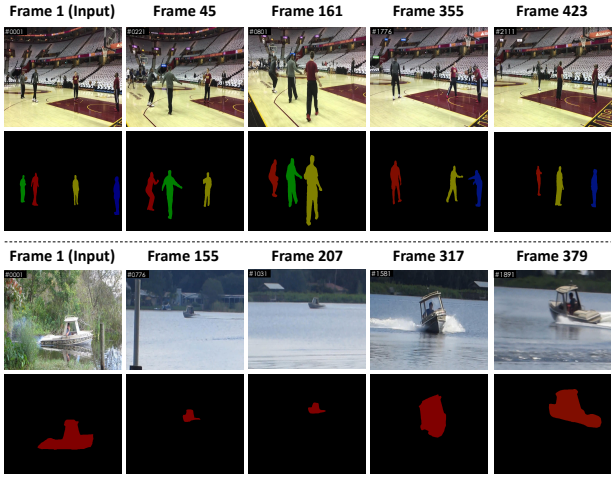

Figure 8. **Qualitative results for R50-MAVOS on two videos from LVOS validation set**. In the first two rows, the targets to segment are four basketball players in action. In the last two rows, the target is to segment the moving ship throughout the video. MAVOS showcases robust segmentation performance in both scenarios, accurately delineating targets despite occlusion and blocking in the first two rows, and coping with varying scaling factors in the last two rows.

show in Fig. 7 more qualitative results for R50-MAVOS on the Long-Time Video dataset [24]. Our MAVOS demonstrates favorable segmentation performance for a long sequence (more than two thousand frames), and runs at 37 FPS. MAVOS accurately segments the target despite the fast movement of the boy and occlusion with other objects between the frames. In addition, we present in Fig. 8 more qualitative results on the LVOS dataset [51]. In the first two rows, the given mask contains four different objects for segmentation throughout the video. Despite objects sometimes blocking each other and some disappearing and reappearing, Our MAVOS demonstrates promising performance for multi-object visual segmentation. In the last two rows, the goal is to segment a moving ship across the video. This is

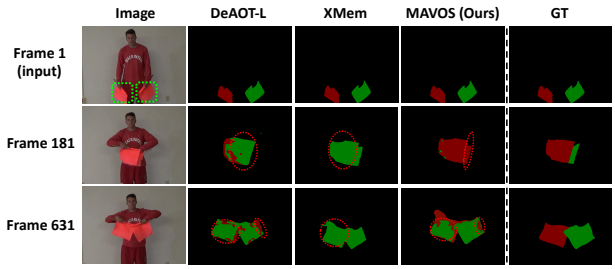

Figure 9. **Qualitative example for failure case from LVOS validation set**. MAVOS as well as the state-of-the-art methods fail to segment highly similar targets (almost identical) after severe occlusion. Both targets are marked with green dashed boxes, failure segmentations are marked with red dashed circles.

challenging because, firstly, the ship often moves far away, making it appear smaller. Secondly, the quality of this video is poor. Despite these challenges, our MAVOS is able to accurately segment the ship, whether it's close to the camera or far away, showing the effectiveness of our method.

4. Limitations

We observe that MAVOS often fails to segment targets when they are identical or highly similar after disappearance or severe occlusion occurs. We demonstrate this case in Fig. 9. This is a common problem not only for MAVOS but also for other state-of-the-art methods, including the baseline DeAOT-L [63] and XMem [5]. This is likely due to the lack of encoding sufficient discriminative features for the targets due to the high similarity between them. As shown in Fig. 9, in the first column, two masks of almost identical flags are given in the reference frame. At frame 181, the flag with the red mask overrides the flag with the green mask. DeAOT and XMem confuse both flags, while MAVOS partially segments them correctly. At frame 631, MAVOS as well as DeAOT-L and XMem fail to discriminate both flags after the severe occlusion between both of them due to high similarity between both flags. We argue that this is likely due to the lack of discriminative features from the visual encoder and the short-term memory.