2nd Place Solution for the LSVOS Challenge 2023: Video Object Segmentation

Deshui Miao[1, 2 ‡], Xin Li[2 *], Zhenyu He[1 *], Yaowei Wang[2], Huchuan Lu[3], Ming-Hsuan Yang[4]
[1]Harbin Institute of Technology, Shenzhen [2] Peng Cheng Laboratory
[3] Dalian University of Technology [4] University of California at Merced

Abstract

For video object segmentation (VOS) tasks, Decouple Associating Objects with Transformer (DeAOT) has been widely used and achieves outstanding results. Based on DeAOT, we investigate a more powerful VOS model from the aspects of more robust feature representation and more effective long-term memory. A multi-scale gated propagation matching model (MGPM) is developed to take full use of multi-scale features. In addition, we design an adaptive memory bank to get more discriminative features for long-term target appearance modeling. The adaptive memory bank performs updating by merging the most similar memorized features based on the similarity between the features of new coming and old samples. Our solution achieves second place in the LSVOS Challenge 2023 Track1: Video Object Segmentation.

1. Introduction

The primary objective of video object segmentation is to delineate prominent objects in the foreground from the background. This practice holds significant promise across various applications, particularly as the prevalence of video content surges in domains like autonomous driving, augmented reality [5], and interactive video editing [4]. This study centers on the precise segmentation of objects within videos under a semi-supervised paradigm. In this approach, initial object segmentation is provided in the first frame, subsequently guiding the consistent segmentation of these identified objects throughout the ensuing frames. The persisting challenge lies in the dynamic transformations that objects undergo over time due to inherent traits and external occlusions.

There are several works studying the Semi-supervised Video Object Segmentation task. The most common way used by most methods is the matching similarity between the query frame and the past frames [1, 3, 6, 8]. Within these methods, the Space-Time Memory Network (STM) [6] gives great robustness towards the changes in the appearance of the targets. The Semi-VOS task [11, 13, 14] has made great progress since STM was proposed. However, STM suffers from memory consumption and heavy computation. The worse is that multi-target scenarios are general while STM can only process one target at once, which leads to repeated calculation. AOT [15] was developed by utilizing the provided mask as ID to perform multi-target processing. AOT associates the objects simultaneously with a long short-term transformer and a novel ID branch to get more accurate performance.

However, with the propagation of ID in AOT, the object-specific information is increased and this will inevitably lead to the loss of object-agnostic visual information in the later approach. To solve this problem, DeAOT [16] designed a dual branch for propagating both object-specific and object-agnostic information. The usage of a memory bank significantly boosts the performance in various Semi-VOS datasets [9]. The original DeAOT performs long-term memory writing every t frame, which lacks the distinction of the saved features. In order to better model the memory bank, Xmem [2] proposed a unified feature memory store by the Atkinson-Shirin memory model.

Inspired by the design of XMem, we proposed an adaptive memory bank for boosting the baseline of DeAOT. In a long-term video, the memory bank should save more discriminative features. As a result, our proposed method adaptively saves the appearance change of the targets and effectively saves the computation. Furthermore, the scale of an object in a video usually changes over time, which is also a common challenge in Semi-VOS task. Despite the effective design of DeAOT, only the smallest scale feature is fed to the GPM block. We argue that this design can not well suit the changes in the target scale. To tackle this problem, we propose a multi-scale GPM with a different scale decoder like HQTrack [17] and PAOT [12] to associate objects, which sequentially merges feature maps on multiple scales to decode the object. Compared with the original GPM, the proposed MGPM totally uses the abundant feature information from the encoder to propagate.

LSVOS challenge 2023 combines the classic YouTube-

‡Work done during research internships at Peng Cheng Laboratory.

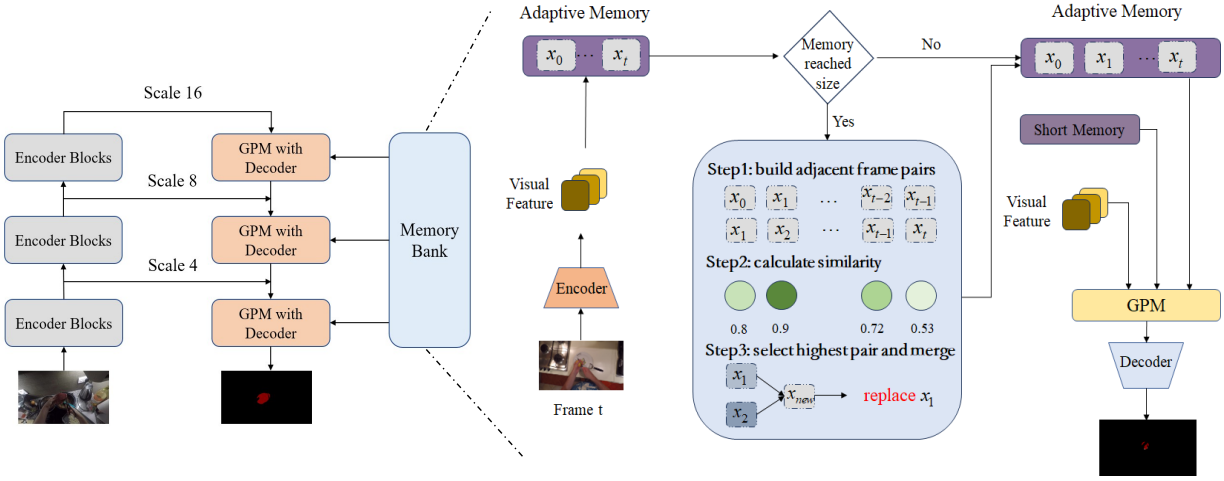

Figure 1. Overall framework of our approach. The left part shows the multi-scale GPM model and the right part depicts the adaptive memory component.

VOS benchmark with the newly proposed VOST [10] dataset. VOST focuses on complex object transformations and huge appearance changes. The proposed memory bank achieves great improvement on the VOST dataset due to the saved discriminative features. Moreover, our method shows its superior performance in the VOS track of the 5th LSVOS competition with rank 2nd place in the test stages.

2. Method

In this section, we will introduce the technical details of our method. Since the overall network is the same as DeAOT, a detailed explanation of DeAOT will not be repeated in this technical report. Please refer to DeAOT for detailed information. The overall architecture of the proposed method will be first shown in Section 2.1, while the detail will be introduced in Section 2.2.

2.1. Overall Framework

The overall framework illustrated in Fig. 1 follows the design of encoder-decoder blocks similar to the classical Segmentation network. The encoder consists of different down-sampling scale blocks to get features at different scales. The multiple-scale features provide abundant information for accurate tracking and segmentation.

In the decoder, unlike the FPN module employed in DeAOT, we propose the MGPM like HQTrack [17] to establish the multi-scale stages of MSDeAOT. The feature of each scale map from the encoder is fed into the corresponding stage. The GPM then takes charge of matching the current embedding with memory embeddings and aggregation mask information from memory frames. The decoder blocks finally decode the information from the GPM. In the memory bank, we design an adaptive update method for the saved memory embeddings. The proposed memory bank has a capacity threshold for saving memory. If the saved frame embeddings in the memory reach the threshold, the proposed strategy will fuse the similar embeddings and save the more discriminative embeddings in the memory.

2.2. Adaptive Memory

Long-term memory can effectively avoid the problem of knowledge forgetting, which is crucial for processing long videos. The long-term memory in DeAOT just saves the frame embedding every t frame. PAOT [12] proposed an updating method of memory bank by adopting the first-in-first-out strategy. However, the saved frame embeddings may exhibit significant similarity in the video. To this end, we propose a method to merge the most similar frame embeddings. This method transfers the traditional adjacent memory to sparse memory and stores more discriminated features.

In the adaptive memory, we execute memory consolidation by merging tokens that exhibit the highest similarity within neighboring frames, following the methodology introduced in ToMe. Our observation reveals that the frame embeddings within transformer models inherently encapsulate frame information, serving the purpose of computing cosine similarity (denoted as s)

$$s = \frac{1}{N}\sum_{j=1}^{N}\left[\cos\left(\mathrm{k}_i^j, \mathrm{k}_{i+1}^j\right)\right]. \quad (1)$$

Based on the baseline of DeAOT, we choose the saved key

Table 1. **Leaderboards of the 5th LSVOS challenge. We rank 2nd place in the competition.**

Method	Overall	J_{ytvos}	$J_{ytvos_{tr}}$	J_{vost}	$J_{vost_{tr}}$
ours	**0.651 (1)**	**0.849 (1)**	**0.834 (1)**	**0.527 (1)**	**0.394 (1)**
abcaaa	0.629 (2)	0.846 (2)	0.829 (2)	0.498 (2)	0.343 (4)
volkdown	0.629 (3)	0.844 (3)	0.827 (3)	0.488 (3)	0.355 (3)
warriors	0.627 (4)	0.842 (4)	0.824 (4)	0.484 (4)	0.357 (2)
lonelyqian	0.570 (5)	0.832 (6)	0.816 (6)	0.376 (7)	0.259 (7)
abcaaa	0.651 (1)	0.821 (2)	0.797 (3)	0.566 (1)	0.420 (1)
ours	**0.646 (2)**	**0.823 (1)**	**0.801 (1)**	**0.548 (2)**	**0.414 (2)**
Fayewong	0.630 (3)	0.814 (4)	0.798 (2)	0.519 (4)	0.389 (4)
uuyht	0.630 (4)	0.811 (6)	0.782 (7)	0.531 (3)	0.395 (3)
QSYCVTEAM	0.629 (5)	0.812 (5)	0.796 (4)	0.519 (4)	0.389 (4)

Table 2. Ablation results of different modules on the LSVOS test dataset.

Method	Overall	J_{ytvos}	$J_{ytvos_{tr}}$	J_{vost}	$J_{vost_{tr}}$
PDeAOT	0.629	0.818	0.796	0.522	0.379
AdaMem-DeAOT	0.641	0.822	0.801	0.538	0.401
AdaMem-MSDeAOT	0.646	0.823	0.801	0.548	0.414

to perform frame embedding similarity. We repeat this process iteratively until the embedding count matches the predetermined value for each consolidation operation.

3. Experiments

In our experiments, we employ DeAOT as our baseline. SwinTransformer-Base is used as the backbone for the encoder. For the decoder, the MSDeAOT model incorporates GPM modules in multiple stages.

The training process comprises three phases, following the DeAOT framework. In the first phase, we pre-train our model using synthetic video sequences generated from static image datasets by augmentation methods. In the second stage, we finetuned our model in YTVOS and DAVIS [7] datasets as DeAOT. Then, the final model is gotten by fine-tuning the VOST dataset.

During training, we conducted 8 Tesla V100 GPUs with a batch size of 16. For pre-training, we use the same setting as DeAOT. The finetuning step on the VOST dataset is set to 20000.

3.1. Results

As shown in Table 1, our method achieves an overall score of 0.646 on the LSVOS 2023 challenge test set (Semi-VOS track) and ranks the second place.

3.2. Ablation Studys

We conduct extensive ablation studies on the LSVOS test set. As shown in Table 2, we first use the memory updating method in PAOT to improve DeAOT, which achieves significant performance gains. The proposed adaptive memory improves the performance on the VOST test datasets, which verifies the effectiveness of our method. Finally, the MGPM gets further improvement in overall performance. Experimental results demonstrate that the utilization of adaptive memory and MGPM leads to an increase of both J andJ_{tr}

4. Conclusion

In this report, we propose an adaptive memory bank to maintain long-term target appearance and use the multi-scale GPM for robust representation, which significantly improves the performance of DeAOT. We conduct experiments to investigate the effect of different memory strategies on the task of video object segmentation. By adopting the proposed memory bank and the MGPM, our method achieves 2nd place in the video object segmentation track of 5th LSVOS challenge.

References

[1] Emre Aksan and Otmar Hilliges. Stcn: Stochastic temporal convolutional networks. *arXiv preprint arXiv:1902.06568*, 2019. 1

[2] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 1

[3] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Videomatch: Matching based video object segmentation. In

Proceedings of the European conference on computer vision (ECCV), pages 54–70, 2018. 1

[4] King Ngi Ngan and Hongliang Li. *Video segmentation and its applications*. Springer Science & Business Media, 2011. 1

[5] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7376–7385, 2018. 1

[6] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 1

[7] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 3

[8] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. Hierarchical memory matching network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12889–12898, 2021. 1

[9] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023. 1

[10] Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the" object" in video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22836–22845, 2023. 2

[11] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *CoRR*, abs/1809.03327, 2018. 1

[12] Yuanyou Xu, Zongxin Yang, and Yi Yang. Video object segmentation in panoptic wild scenes. *arXiv preprint arXiv:2305.04470*, 2023. 1, 2

[13] Linjie Yang, Yuchen Fan, and Ning Xu. The 2nd large-scale video object segmentation challenge - video object segmentation track, Oct. 2019. 1

[14] Linjie Yang, Yuchen Fan, and Ning Xu. The 4th large-scale video object segmentation challenge - video object segmentation track, June 2022. 1

[15] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34:2491–2502, 2021. 1

[16] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. *Advances in Neural Information Processing Systems*, 35:36324–36336, 2022. 1

[17] Jiawen Zhu, Zhenyu Chen, Zeqi Hao, Shijie Chang, Lu Zhang, Dong Wang, Huchuan Lu, Bin Luo, Jun-Yan He, Jin-Peng Lan, et al. Tracking anything in high quality. *arXiv preprint arXiv:2307.13974*, 2023. 1, 2