THU-Warwick Submission for EPIC-KITCHEN Challenge 2025: Semi-Supervised Video Object Segmentation

Mingqi Gao[1,2]   Haoran Duan[1]   Tianlu Zhang[1]   Jungong Han[1,*]

[1]Tsinghua University   [2]University of Warwick

mingqi.gao@warwick.ac.uk, tlzhang@mail.tsinghua.edu.cn, {haoranduan,jghan}@tsinghua.edu.cn

Abstract

In this report, we describe our approach to egocentric video object segmentation. Our method combines large-scale visual pretraining from SAM2 with depth-based geometric cues to handle complex scenes and long-term tracking. By integrating these signals in a unified framework, we achieve strong segmentation performance. On the VISOR test set, our method reaches a $\mathcal{J\&F}$ score of 90.1%.

1. Introduction

Egocentric visual understanding [9, 14, 15] enables intelligent systems to perceive the world from a human-centred perspective. It is vital in applications such as embodied AI, augmented reality, and assistive technologies. In these scenarios, most tasks are object-centric, focusing on recognising and tracking the objects the user sees or interacts with. Therefore, achieving pixel-level object perception and consistent tracking over time is crucial. This is the right goal of EPIC-KITCHENS VISOR [5]: to segment and track hands and active objects in egocentric videos.

Compared to traditional video object segmentation tasks that focus on third-person views [6, 10, 16, 22], VISOR presents unique challenges: 1) The egocentric viewpoint leads to rapidly changing and cluttered backgrounds, making it hard to distinguish target objects and 2) Frequent interactions between the user's hands and objects cause severe and dynamic occlusions, further complicating accurate segmentation. These challenges require methods that can robustly handle complex scenes with heavy occlusions and ambiguous object boundaries.

To address these challenges, our solution integrates depth to provide complementary geometric cues that help distinguish target objects from complex backgrounds and handle occlusions more effectively. Specifically, we leverage Depth Anything V2 [24], a large-scale depth model, and combine its backbone features with those from a visual

---

*Corresponding author.

backbone. This fusion is performed through a learnable, multi-scale module, enabling the model to utilise depth and RGB features to enhance the segmentation process.

Benefiting from large-scale data, SAM2 [18] has shown strong performance across various video object segmentation tasks. However, their use of segmentation history remains relatively shallow, relying more heavily on recent predictions. This makes it less effective in long-term, complex scenarios, which are common in VISOR. To address this, we adopt Cutie [4], a long-term segmentation framework, as our baseline and replace its visual backbone with SAM2's pre-trained weights. This allows us to combine the long-term temporal modelling capability of the baseline with SAM2's strong object perception and precise frame-to-frame correspondence learned from large-scale data.

In summary, we combine fine-grained visual and geometric foundation models to tackle the unique challenges of egocentric video object segmentation. This design highlights the benefit of jointly leveraging visual and depth representations. On the VISOR test set, our method achieves a strong $\mathcal{J\&F}$ score of 90.1%.

2. Related Works

Semi-Supervised Video Object Segmentation (SVOS) aims to segment objects of interest in videos, where the targets are indicated by the annotations of the first frame by humans [8]. Early works relied on propagation of labels through online fine-tuning [1] or matching strategies [21], to transfer the initial mask and previous predictions to the current frame. The introduction of memory-based methods significantly improved segmentation performance by enabling a more effective use of intermediate predictions. These methods selectively store features and masks from intermediate frames (between the first and current frames) as memory, and retrieve relevant information to guide segmentation in the current frame. Following the initial memory-based models [12], subsequent works have focused on improving memory management [11], frame-wise affinity [3], long-term memory [2], and object-aware memory design [4].
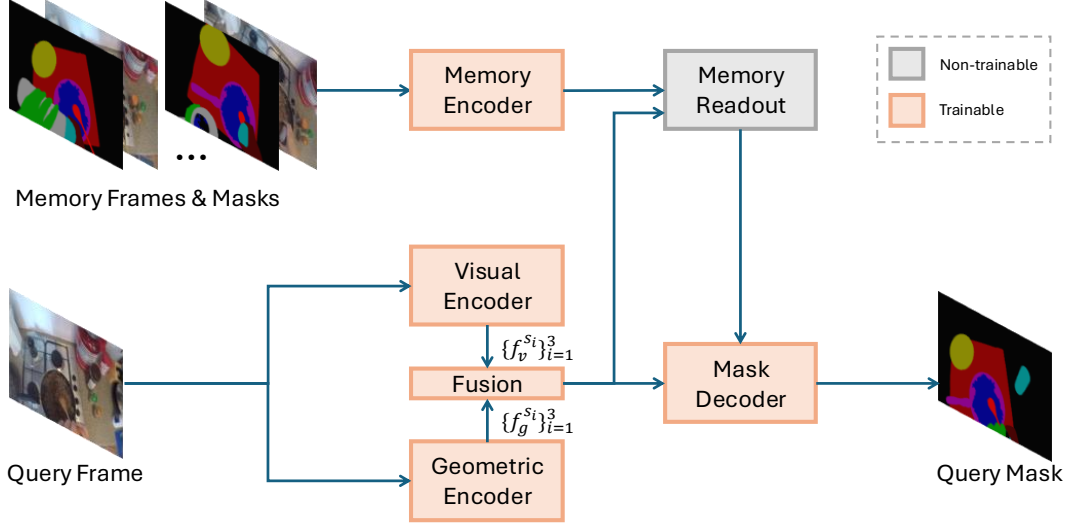
Figure 1. Overview of our solution.

The success of SAM2 [18] highlights that large-scale training can significantly improve segmentation performance, even with simple memory strategies that consider only the first and most recent frames. However, such limited memory usage is less effective for long-term videos. As a result, many follow-up works [7, 20, 23] have explored better memory selection and utilisation to boost performance further. Although these approaches improve the results of existing benchmarks, most remain focused on visual cues, with limited attention to geometric information despite its importance in complex, dynamic scenes.

3. Method

Fig. 1 shows our solution, where we consider Cutie [4] as the baseline, and we refer readers to the original paper for a comprehensive description. On top of the baseline, we consider Hiera-Large [19], with parameters pre-trained in SAM2 [18], as the visual encoder to enhance the object perception and frame-wise correspondence. In addition, we incorporate extra geometric cues to improve robustness against challenges in egocentric VOS. Specifically, the geometric encoder (DINOv2-Large [13] and a DPT decoder [17] for multi-scale embeddings) comes from Depth Anything V2 [24].

Given a query frame $F_{query} \in \mathbb{R}^{3 \times H \times W}$ of height H and width W, the visual and geometric encoders extract multi-scale features, denoted as $\{f_v^{s_i}\}_{i=1}^3$ and $\{f_g^{s_i}\}_{i=1}^3$, where s_1, s_2, and s_3 correspond to 1/4, 1/8, and 1/16 of the original resolution, respectively. Note that the patch sizes of DINOv2 and Hiera differ, which leads to mismatched feature resolutions between f_v and f_g. To address this issue, we resize the input frames fed into the geometric encoder, ensuring that the resulting feature maps have consistent res-

olutions across different patch sizes. After feature encoding, f_v and f_g are concatenated at matching scales and subsequently fused via a learnable MLP layer. The fused features could be seamlessly integrated into the baseline framework and generate a mask for the query frame (we keep other modules the same as the baseline).

4. Experiments

4.1. Main Results

Our training consists of two stages: 1) We freeze all encoders to initialise the fusion module and decoder on traditional VOS datasets, where we follow Cutie [4] to set data collections (the "Mega" version) and training coefficients. 2) To evaluate on the VISOR test set, we train our solution on the VISOR train and val sets for 100,000 iterations, with a batch size of 8, learning rate of 5e-5, and weight decay of 0.5. Due to sparse annotations and the long-term context in VISOR, we set the max skip as 1 when sampling frames as pseudo-training videos.

Table 1 shows our scores on the test set, where "Hiera-L (SAM2)" denotes the Hiera-Large and its initial parameters come from SAM2. "MS" indicates that we perform VOS under three different input frame sizes ($1.2\times$, $1.3\times$, and $1.4\times$), and "Flip" flags the use of flipped frames. In these cases of multiple inference, we sum all probabilities and generate the final predictions.

4.2. Ablations

This section investigates the impact of different components of our solution on the segmentation performance. For simplicity and efficiency, all ablation variants are trained for 50,000 iterations with a batch size of 4, while keeping all

Backbone	Flip	MS	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Hiera-L (SAM2)	✓	✗	89.7%	87.5%	91.8%
Hiera-L (SAM2)	✓	✓	90.1%	88.1%	92.0%

Table 1. Ablations on the VISOR test set.

Backbone	Depth	Post	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Hiera-L (MAE)	✗	✗	86.9%	85.6%	88.2%
Hiera-L (SAM2)	✗	✗	88.3%	86.6%	89.9%
Hiera-L (SAM2)	✓	✗	88.7%	86.9%	90.5%
Hiera-L (SAM2)	✓	✓	88.9%	87.0%	90.8%

Table 2. Ablations on the VISOR val set. The highlighted scores and the bottom results in Table 1 come from the same setting.

other hyperparameters identical to those used in the main training setup. All results are shown in Table 2.

Impact of SAM2 Pretraining. We first examine the effect of initialising the visual backbone with SAM2. This change alone leads to a substantial performance boost, which we attribute to the strong object perception and inter-frame correspondence priors learned from large-scale training. In addition, we fine-tune the original SAM2 model directly on VISOR. While it achieves competitive results ($\mathcal{J}\&\mathcal{F}$: 87.8%), its performance remains below that of our full method w/o "Depth" and "Post", indicating that effective memory design for complex backgrounds and long-term dynamics is still crucial.

Benefit of Depth Integration. Next, we evaluate the contribution of the depth information. The results show performance improvements, confirming that depth provides valuable complementary cues.

Multi-Scale and Flip Fusion. Finally, we apply multi-scale inference and horizontal flip fusion during testing, which brings additional performance gains by improving prediction consistency.

5. Conclusion

In this report, we address the challenges of egocentric VOS by leveraging large-scale visual pre-training and geometric depth cues. Through extensive experiments and ablation studies, we demonstrate that combining high-quality training priors from SAM2 with depth-aware design significantly improves segmentation accuracy, especially in complex and long-term scenarios. Our results highlight the importance of data scale and spatial understanding in advancing egocentric video understanding.

References

[1] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *CVPR*, pages 221–230, 2017. 1

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

[3] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. 34:11781–11794, 2021. 1

[4] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *CVPR*, 2024. 1, 2

[5] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 1

[6] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *ICCV*, 2023. 1

[7] Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. *arXiv preprint arXiv:2410.16268*, 2024. 2

[8] Mingqi Gao, Feng Zheng, James JQ Yu, Caifeng Shan, Guiguang Ding, and Jungong Han. Deep learning for video object segmentation: a review. *Artificial Intelligence Review*, 56(1):457–531, 2023. 1

[9] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, pages 19383–19400, 2024. 1

[10] Lingyi Hong, Zhongying Liu, Wenchao Chen, Chenzhi Tan, Yuang Feng, Xinyu Zhou, Pinxue Guo, Jinglun Li, Zhaoyu Chen, Shuyong Gao, et al. Lvos: A benchmark for large-scale long-term video object segmentation. *arXiv preprint arXiv:2404.19326*, 2024. 1

[11] Yongqing Liang, Xin Li, Navid Jafari, and Jim Chen. Video object segmentation with adaptive feature bank and uncertain-region refinement. 33:3430–3441, 2020. 1

[12] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, pages 9226–9235, 2019. 1

[13] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou,

Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2

[14] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. In *CVPR*, 2025. 1

[15] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *IJCV*, 132(11):4880–4936, 2024. 1

[16] 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. 1

[17] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pages 12179–12188, 2021. 2

[18] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 2

[19] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*, pages 29441–29454. PMLR, 2023. 2

[20] Jovana Videnovic, Alan Lukezic, and Matej Kristan. A distractor-aware memory for visual object tracking with sam2. *arXiv preprint arXiv:2411.17576*, 2024. 2

[21] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, pages 9481–9490, 2019. 1

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

[23] Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv preprint arXiv:2411.11922*, 2024. 2

[24] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 1, 2