# THU-Warwick 参加 EPIC-KITCHEN 挑战赛 2025 的提交：半监督视频对象分割

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

在本报告中，我们描述了我们对于自我中心视频对象分割的方法。我们的方法结合了来自 SAM2 的大规模视觉预训练与基于深度的几何线索，以处理复杂场景和长期跟踪。通过在一个统一的框架中整合这些信号，我们实现了强大的分割性能。在 VISOR 测试集中，我们的方法达到了 90.1 的 $\mathcal{J}\&\mathcal{F}$ 分数%。

## 1. 引言

自我中心视觉理解 [8, 13, 14] 使智能系统能够从以人为中心的视角感知世界。这在诸如具身人工智能、增强现实和辅助技术等应用中至关重要。在这些场景中，大多数任务都是以物体为中心的，重点在于识别和跟踪用户看到或与之交互的物体。因此，实现像素级的物体感知和一致的时间跟踪是关键。这正是 EPIC-KITCHENS VISOR [5] 的目标：在自我中心视频中分割和跟踪手部和活动物体。

与传统的视频对象分割任务关注第三人称视角相比 [6, 9, 15, 20] ，VISOR 提出了独特的挑战：1）第一人称视角导致背景快速变化且杂乱，使得难以区分目标对象，以及 2）用户手与对象之间频繁的交互导致严重且动态的遮挡，进一步使精确分割复杂化。这些挑战需要能够稳健处理复杂场景中严重遮挡和模糊对象边界的方法。

为了解决这些挑战，我们的解决方案集成了深度信息，以提供互补的几何线索，帮助区分目标物体与复杂的背景，并更有效地处理遮挡。具体来说，我们利用了大规模深度模型 Depth Anything V2 [21] ，将其骨干特征与视觉骨干中的特征相结合。通过一个可学习的多尺度模块进行融合，使模型能够利用深度和 RGB 特征来增强分割过程。

得益于大规模数据，SAM2 [17] 在各种视频对象分割任务中表现出色。然而，它们对分割历史的利用仍然相对浅显，更依赖于最近的预测。这使得它在 VISOR 中常见的长期、复杂场景中效果不佳。为了解决这个问题，我们采用 Cutie [4] ，一个长期分割框架，作为我们的基线，并将其视觉主干替换为 SAM2 的预训练权重。这使我们能够结合基线的长期时间建模能力与 SAM2 从大规模数据中学习的强大对象感知和精确的帧间对应关系。

总之，我们结合细粒度的视觉和几何基础模型来应对第一人称视频对象分割的独特挑战。这一设计突显了联合利用视觉和深度表示的好处。在 VISOR 测试集上，我们的方法取得了强劲的 $\mathcal{J}\&\mathcal{F}$ 得分，达到了 90.1 %。

## 2. 相关工作

半监督视频目标分割（SVOS）旨在对视频中的感兴趣目标进行分割，其中目标由第一帧的人类注释指示 [7] 。早期的工作依赖于通过在线微调 [1] 或匹配策略 [19] 来传播标签，以将初始掩码和之前的预测转移到当前帧。基于记忆的方法的引入显著提高了分割性能，因为它们能够更有效地利用中间预测。这些方法选择性地存储来自中间帧（在第一帧和当前帧之间）的特征和掩码作为记忆，并检索相关信息以指导当前帧的分割。在最初的基于记忆的模型之后 [11] ，后续的工作集中在改进记忆管理 [10] 、逐帧亲和性 [3] 、长期记忆 [2] 和对象感知记忆设计 [4] 上。

SAM2 的成功突出表明，即使只使用简单的内存策略（仅考虑首次和最近的帧），大规模训练也可以显著提高分割性能。然而，这样的有限内存使用对长时间视频的效果较差。因此，许多后续研究探讨了更好的内存选择和利用方法，以进一步提高性能。虽然这些方法改进了现有基准的结果，但大多数仍然专注于视觉线索，对几何信息的关注有限，尽管在复杂、动态场景中几何信息很重要。

## 3. 方法

图 1 显示了我们的解决方案，其中我们将 Cutie [4] 视为基线，并建议读者参考原始论文以获得全面的描述。在基线之上，我们考虑使用 Hiera-Large [18] ，其参数在 SAM2 [17] 中进行了预训练，作为视觉编码器以增强对象感知和逐帧对应。此外，我们结合额外的几何线索以提高针对第一人称视频对象分割挑战的鲁棒性。具体而言，几何编码器（DINOv2-Large [12] 和 DPT

* Corresponding author.

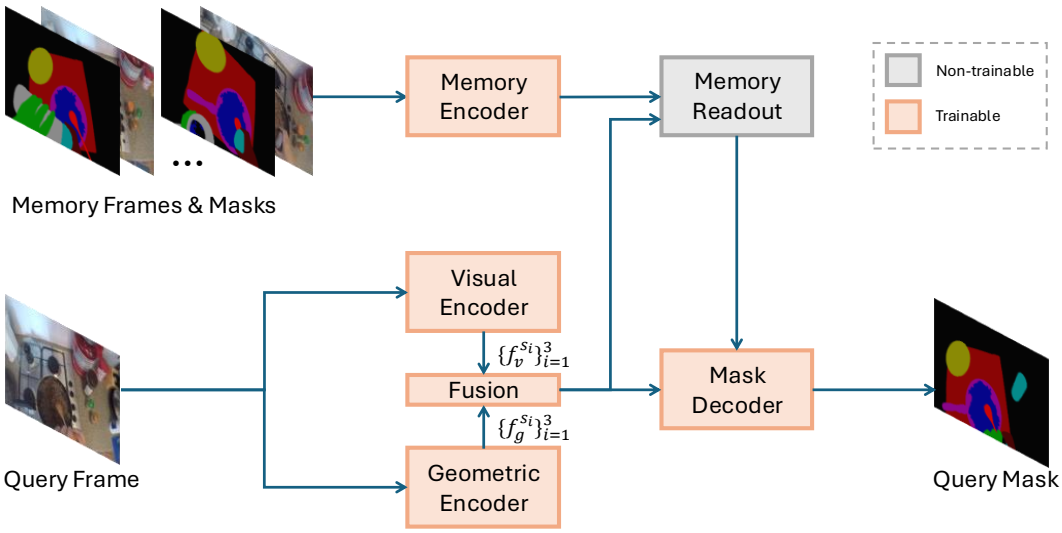


Figure 1. 我们解决方案的概述。

解码器 [16] 用于多尺度嵌入）来自深度任意 V2 [21] 。

给定一个高度为 $H$ ，宽度为 $W$ 的查询帧 $F_{\text{query}} \in \mathbb{R}^{3\times H\times W}$ ，视觉和几何编码器提取多尺度特征，表示为 $\{f_{\text{v}}^{s_i}\}_{i=1}^3$ 和 $\{f_{\text{g}}^{s_i}\}_{i=1}^3$ ，其中 $s_1$ 、$s_2$ 和 $s_3$ 分别对应于原始分辨率的 1/4、1/8 和 1/16。注意，DINOv2 和 Hiera 的 patch 大小不同，这导致了 $f_{\text{v}}$ 和 $f_{\text{g}}$ 之间的特征分辨率不匹配。为了解决这个问题，我们调整输入到几何编码器的帧的尺寸，确保生成的特征图在不同的 patch 大小下具有一致的分辨率。特征编码完成后，$f_{\text{v}}$ 和 $f_{\text{g}}$ 在匹配尺度下连接，并通过一个可学习的 MLP 层融合。融合后的特征可以无缝集成到基线框架中，并为查询帧生成一个掩码（我们保持其他模块与基线相同）。

## 4. 实验

### 4.1. 主要结果

我们的训练包括两个阶段：1）我们冻结所有编码器，以传统 VOS 数据集初始化融合模块和解码器，在此过程中，我们遵循 Cutie [4] 设置数据集合（“Mega” 版本）和训练系数。2）为了在 VISOR 测试集上进行评估，我们在 VISOR 的训练和验证集上进行训练，进行了 100,000 次迭代，批量大小为 8，学习率为 5e-5，权重衰减为 0.5。由于 VISOR 中的稀疏注释和长期上下文，在采样帧作为伪训练视频时，我们将最大跳过设为 1。

表格 1 显示了我们在测试集上的得分，其中 “Hiera-L (SAM2)” 表示 Hiera-Large，其初始参数来自 SAM2。“MS” 表示我们在三种不同输入帧尺寸 (1.2 × , 1.3 × , 和 1.4 × ) 下执行 VOS，而 “Flip” 表示使用翻转帧。在这些多重推理的情况下，我们将所有概率相加并生成最终预测。

| 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. 在 VISOR 测试集上的消融实验。

### 4.2. 消融实验

本节研究了我们解决方案中不同组件对分割性能的影响。为了简化和提高效率，所有消融变体均使用批量大小为 4 进行 50,000 次迭代训练，同时保持其他所有超参数与主训练设置中使用的一致。所有结果均在表 2 中展示。

我们首先检查用 SAM2 初始化视觉主干的效果。仅此更改就能显著提升性能，我们将其归因于从大规模训练中学习到的强大物体感知和帧间对应先验。此外，我们直接在 VISOR 上微调原始的 SAM2 模型。尽管其取得了具有竞争力的结果（$\mathcal{J}\&\mathcal{F}$ ：87.8 %），但其性能仍低于我们的完整方法去除 “Depth” 和 “Post” 后的效果，这表明对复杂背景和长期动态进行有效的内存设计仍然至关重要。

接下来，我们评估深度信息的贡献。结果显示性能有提升，证实了深度提供了有价值的补充线索。

最后，我们在测试过程中应用多尺度推理和水平翻转融合，通过提高预测一致性带来了额外的性能提升。

## 5. 结论

在本报告中，我们通过利用大规模视觉预训练和几何深度线索来应对以自我为中心的视频对象分割（VOS）的挑战。通过广泛的实验和消融研究，我们证明了将来自 SAM2 的高质量训练先验与深度感知设计相结合能

| 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. 关于 VISOR 验证集的消融实验。表 1 中的高亮分数和底部结果来自相同的设置。

够显著提高分割精度，尤其是在复杂和长期场景中。我们的结果强调了数据规模和空间理解对推进以自我为中心的视频理解的重要性。

## 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.

[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.

[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.

[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.

[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.

[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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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