

ZJU ReLER Submission for EPIC-KITCHEN Challenge 2023: Semi-Supervised Video Object Segmentation

Jiahao Li, Yuanyou Xu, Zongxin Yang, Yi Yang, Yueting Zhuang
ReLER, CCAI, Zhejiang University

{xljh, yoxu, yangzongxin, yangyics, yzhuang}@zju.edu.cn

Abstract

The Associating Objects with Transformers (AOT) framework has exhibited exceptional performance in a wide range of complex scenarios for video object segmentation [17, 19]. In this study, we introduce MSDeAOT, a variant of the AOT series that incorporates transformers at multiple feature scales. Leveraging the hierarchical Gated Propagation Module (GPM), MSDeAOT efficiently propagates object masks from previous frames to the current frame using a feature scale with a stride of 16. Additionally, we employ GPM in a more refined feature scale with a stride of 8, leading to improved accuracy in detecting and tracking small objects. Through the implementation of test-time augmentations and model ensemble techniques, we achieve the top-ranking position in the EPIC-KITCHEN VISOR Semi-supervised Video Object Segmentation Challenge.

1. Introduction

Video object segmentation is a vital task in computer vision that involves segmenting and isolating specific objects of interest in each frame of a video sequence. This task aims to provide pixel-level masks or contours delineating the target object’s boundaries in each frame.

Semi-supervised Video Object Segmentation (SVOS) has garnered significant attention in recent years, with a particular emphasis on learning-based methodologies. In this context, a prevailing approach involves pixel matching across frames to extract crucial insights regarding the target objects. Notably, PLM [12] emerges as a pioneering matching-based SVOS method. By performing multi-scale matching between the previous and target frames, PLM successfully segments the target object. To address the challenge of ambiguous backgrounds, the method specifically focuses on the region surrounding the object in the previous frame. FEELVOS [14] leverages both global and local matching based on pixel-wise embeddings to facilitate information transfer across frames. These matching results

are further integrated with semantic features to predict the final segmentation outcome. CFBI(+) [16, 18] employs the first frame and the previous frame as references, treating foreground and background regions with equal importance, implicitly enhancing the discriminability of encoded features. These pioneering techniques contribute to the ongoing development of SVOS, offering valuable insights and potential avenues for further exploration in the field.

AOT [17] stands as another prominent contribution in this field. Unlike previous methods which segment multiple objects one by one and merge results by post ensemble, AOT presents an innovative identification mechanism that encompasses the encoding, matching, and decoding of multiple objects. By incorporating transformers, AOT effectively associates objects across frames, fostering a deeper exploration and exploitation of inter-object relationships. Building upon this foundation, DeAOT [19], a variant of AOT, employs a hierarchical Gated Propagation Module (GPM) to independently propagate object-agnostic and object-specific embeddings from previous frames to the current frame. This novel approach effectively preserves object-agnostic visual information within the deep transformer layers.

Although the aforementioned methods have demonstrated remarkable performance on conventional datasets such as YouTube-VOS [15] and DAVIS [10], there remain several challenges that demand attention when dealing with egocentric VISOR [2, 3]. **i)** Egocentric videos often exhibit rapid camera movements, posing a significant obstacle. **ii)** In the VISOR dataset, frames are not uniformly sampled at a fixed frame rate, and the time intervals between frames can span up to 1 or 2 seconds. **iii)** Frequent object-hand interactions in egocentric scenarios introduce issues such as occlusion and motion blur, necessitating effective handling strategies.

To address above issues, we propose MSDeAOT, another AOT-based VOS model. MSDeAOT leverages the GPM [19] to propagate object masks from previous frames to the current frame. Specifically, we employ GPM in 2 feature scales with strides of 16 and 8, respectively. This multi-

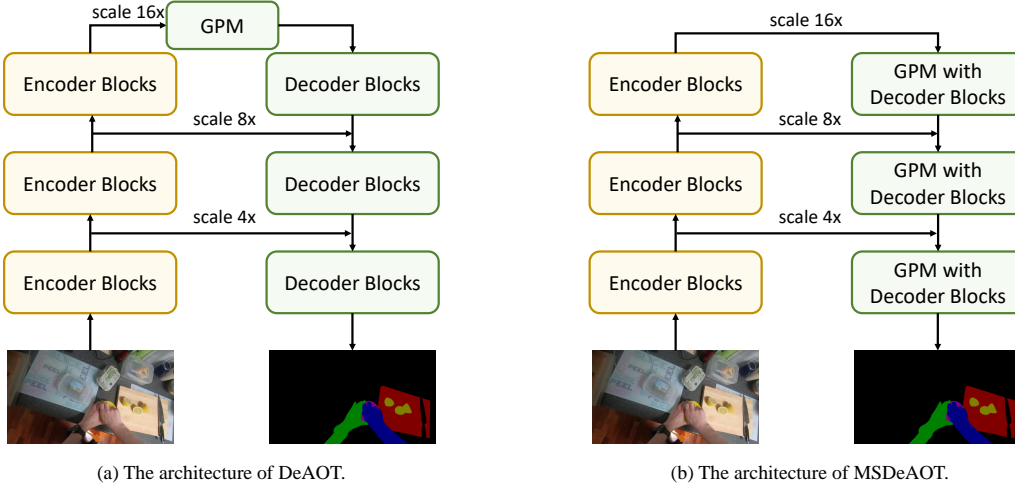


Figure 1. The above two pictures compare the architectures of DeAOT and MSDeAOT. The main difference is that the GPM module is combined with the decoder to form a multi-scale architecture in the MSDeAOT.

scale strategy effectively addresses the aforementioned challenges, enabling MSDeAOT to achieve $\mathcal{J}\&\mathcal{F}$ score of 89.0% in EPIC-KITCHENS VISOR Semi-supervised Video Object Segmentation Challenge.

2. Method

In this section, we present our main method in detail. We begin by providing a brief overview of the DeAOT model to familiarize the reader with the foundational concepts. Subsequently, we delve into the architecture design of our novel MSDeAOT model, outlining its key features and advancements. By elucidating these aspects, we aim to offer a comprehensive understanding of our approach and its contributions.

2.1. Revisiting DeAOT

DeAOT [19], an adaptation of AOT, preserves the fundamental IDentification mechanism of AOT [17]. Consequently, DeAOT exhibits the ability to effectively handle multiple objects concurrently. Diverging from AOT, which consolidates the visual (object-agnostic) and ID (object-specific) embeddings within a shared embedding space, DeAOT adopts a decoupled approach. This entails employing separate propagation processes for each embedding type while maintaining shared attention maps. By decoupling these embeddings, DeAOT achieves enhanced flexibility and discriminative power in handling object representations.

The Gated Propagation Module (GPM) plays a pivotal role within the DeAOT framework. In contrast to the LSTT block utilized by AOT, which leverages multi-head attention for propagation, GPM employs a Gated Propagation Function to seamlessly fuse and propagate both object-agnostic and object-specific embeddings. Additionally, GPM em-

loys single-head attention to match objects and effectively propagate information across frames. By incorporating these distinctive mechanisms, GPM enhances the overall performance of the DeAOT model, enabling more accurate and efficient object segmentation.

2.2. Multi-Scale DeAOT

The whole architecture of MSAOT as shown in Fig. 1b follows an encoder-decoder design similar to classical segmentation networks like U-Net. The encoder consists of multiple blocks that down-sample the input feature maps, yielding features at different scales. These encoder blocks provide multi-scale features that are crucial for accurate object tracking and segmentation.

In the decoder, unlike the FPN module employed in DeAOT (Fig. 1a), the Gated Propagation Module (GPM) is integrated with multiple decoder blocks to establish the multi-scale stages of MSDeAOT. Each scale’s feature maps from the encoder are fed into the corresponding stage, where the GPM module takes charge of matching the current frame with memory frames and aggregating mask information from the memory frames. The decoder blocks then decode this information.

This innovative design of multi-scale stages brings notable benefits. It effectively harnesses the potential of feature maps at different scales, in contrast to the FPN module used in DeAOT, where multi-scale feature maps solely serve as shortcut connections for residual structures. Specifically, in DeAOT, only the feature maps at the smallest scale are utilized for matching across memory frames using the GPM module. In contrast, MSDeAOT comprehensively engages feature maps from multiple scales during the matching process, thereby enhancing performance and enabling finer details of objects to be captured.

| Method | $\mathcal{J}\&\mathcal{F}$ | \mathcal{J} | \mathcal{F} |
|---------------|----------------------------|---------------|---------------|
| R50-MSDeAOT | 84.0 | 81.7 | 86.3 |
| SwinB-MSDeAOT | 85.6 | 83.2 | 87.9 |

Table 1. Comparisons on VISOR val set. Models are trained on VISOR train set.

| Method | Training data | $\mathcal{J}\&\mathcal{F}$ |
|-------------|-----------------------------|----------------------------|
| R50-DeAOTL | VISOR [3] | 83.1 |
| R50-DeAOTL | VISOR [3], YouTube-VOS [15] | 82.9 |
| R50-DeAOTL | VISOR [3], DENSE VISOR [3] | 82.5 |
| R50-MSDeAOT | VISOR [3] | 84.0 |
| R50-MSDeAOT | VISOR [3], VOST [13] | 83.7 |

Table 2. Ablation study of training data. Models are evaluated on VISOR val set.

3. Implementation Details

In MSDeAOT, we employ ResNet-50 [6] and Swin Transformer-Base [8] as the backbones for the encoder. While ResNet-50 offers a lightweight option, Swin Transformer-Base achieves superior performance. For the decoder, the MSDeAOT model incorporates GPM modules in multiple stages. Specifically, we set the number of layers in the GPM to 2 for the $16\times$ scale stage and 1 for the $8\times$ scale stage. To save computational resources, we exclude the $4\times$ scale feature maps and instead duplicate the $16\times$ scale feature maps twice to form the feature pyramid.

The training process comprises two phases, following the AOT framework. In the initial phase, we pre-train the model using synthetic video sequences generated from static image datasets [1, 4, 5, 7, 11] by randomly applying multiple image augmentations [9]. In the subsequent phase, we train the model on the train and val sets of the VISOR dataset [3], incorporating random video augmentations [16].

During MSDeAOT training, we employ 8 Tesla V100 GPUs with a batch size of 16. For pre-training, we use an initial learning rate of 4×10^{-4} for 100,000 steps. For main training, the initial learning rate is set to 2×10^{-4} , and the training steps are 100,000. The learning rate gradually decays to 1×10^{-5} using a polynomial decay schedule [16].

4. EPIC-Kitchens Challenge: Semi-Supervised Video Object Segmentation

4.1. Ablation study on VISOR val set

We train our MSDeAOT on two backbones, ResNet-50 [6] and Swin Transformer-Base [8], and evaluate them on the VISOR val set. As shown in Tab. 1, SwinB-MSDeAOT achieves better performance than R50-MSDeAOT, with a $\mathcal{J}\&\mathcal{F}$ score of 85.6.

As for training data, we train R50-DeAOTL and R50-

| Ensemble Method | $\mathcal{J}\&\mathcal{F}$ | \mathcal{J} | \mathcal{F} |
|-----------------|----------------------------|---------------|---------------|
| Logits Meaning | 88.9 | 86.7 | 91.0 |
| Masks Voting | 89.0 | 86.9 | 91.2 |

Table 3. Results on VISOR test set. Models are trained on VISOR train and val set.

MSDeAOT on multiple datasets and evaluate them on VISOR val set. As shown in Tab. 2, R50-MSDeAOT trained on VISOR achieves the best performance, with a $\mathcal{J}\&\mathcal{F}$ score of 84.0 and more training data has no improvement.

4.2. Model Ensemble

To achieve the best performance on hidden test, we train 3 models, *i.e.*, SwinB-DeAOTL [19], R50-MSDeAOT, and SwinB-MSDeAOT on VISOR train and val set. We then ensemble the predictions of these 3 models to obtain the final results. As for test-time augmentations, both multi-scale test and flip test are used. The scales are $1.2\times$, $1.3\times$, $1.4\times$ and each scale includes non-flipped and flipped test.

Logits Meaning. During the inference process of each model, we refrain from storing the masks themselves. Instead, we retain the logits, which represent the probabilities of each pixel belonging to a specific object. Once inference is completed across multiple models, we acquire multiple sets of logits. These sets are subsequently weighted, averaged, and evaluated. The final label is determined by selecting the set with the highest probability value, thus consolidating the collective predictions from the ensemble of models. However, saving logits requires a large disk space, we can only apply two logits for ensemble.

Mask Voting. First, we adopt a multi-model approach to perform inference, wherein each model produces a mask for an image. As a result, each pixel is associated with multiple labels. To determine the ultimate label for each pixel, we employ a weighted voting scheme that aggregates the labels. This process ensures the derivation of a final label that captures the collective information from the multiple models. As saving masks requires much less disk space than saving logits, we can apply more results of different models for ensemble.

5. Conclusion

In this paper, we propose MSDeAOT, a variant of AOT, for semi-supervised video object segmentation. MSDeAOT leverages the hierarchical Gated Propagation Module (GPM) to independently propagate object-agnostic and object-specific embeddings from previous frames to the current frame. MSDeAOT shows remarkable performance on the EPIC-KITCHENS VISOR Semi-supervised Video Object Segmentation Challenge with $\mathcal{J}\&\mathcal{F}$ of 89.0% on the test set.

References

- [1] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582, 2014. 3
- [2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 1
- [3] 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, 3
- [4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 3
- [5] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011. 3
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3
- [7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3
- [8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3
- [9] 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. 3
- [10] 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
- [11] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):717–729, 2015. 3
- [12] Jae Shin Yoon, Francois Rameau, Junsik Kim, Seokju Lee, Seunghak Shin, and In So Kweon. Pixel-level matching for video object segmentation using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2167–2176, 2017. 1
- [13] 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. 3
- [14] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9481–9490, 2019. 1
- [15] 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, 3
- [16] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V*, pages 332–348. Springer, 2020. 1, 3
- [17] 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, 2
- [18] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by multi-scale foreground-background integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4701–4712, 2021. 1
- [19] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. In *Advances in Neural Information Processing Systems*. 1, 2, 3