1st Place Solution for 5th LSVOS Challenge: Referring Video Object Segmentation

Zhuoyan Luo[1]†, Yicheng Xiao[1]†, Yong Liu[1,2]‡, Yitong Wang[2], Yansong Tang[1], Xiu Li[1], Yujiu Yang[1]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University [2]ByteDance Inc.

{luozy23, xiaoyc23, liu-yong20}@mails.tsinghua.edu.cn

Abstract

The recent transformer-based models have dominated the Referring Video Object Segmentation (RVOS) task due to the superior performance. Most prior works adopt unified DETR framework to generate segmentation masks in query-to-instance manner. In this work, we integrate strengths of that leading RVOS models to build up an effective paradigm. We first obtain binary mask sequences from the RVOS models. To improve the consistency and quality of masks, we propose Two-Stage Multi-Model Fusion strategy. Each stage rationally ensembles RVOS models based on framework design as well as training strategy, and leverages different video object segmentation (VOS) models to enhance mask coherence by object propagation mechanism. Our method achieves 75.7% $\mathcal{J}\&\mathcal{F}$ *on Ref-Youtube-VOS validation set and* 70% $\mathcal{J}\&\mathcal{F}$ *on test set, which ranks 1st place on 5th Large-scale Video Object Segmentation Challenge (ICCV 2023) track 3. Code will be available.*

1. Introduction

Referring Video Object Segmentation aims to segment and track the target object referred by the given text description in a video. This emerging field has garnered attention due to its potential applications in video editing and human-robot interaction.

The critical challenge in RVOS lies in the pixel-level alignment between different modalities and time steps, primarily due to the varied nature of video content and unrestricted language expression. Most early approaches [1, 7, 8] adopt multi-stage and complex pipelines that take the bottom-up or top-down paradigms to segment each frame separately, while recent works MTTR [2], Referformer [15] propose to unify cross-modal interaction with pixel-level understanding into transformer structure. For example, 2022 first winner [6] simply employs fine-tuned Referformer as backbone to generate a series of high quality masks. However, these methods may lose the perception of target objects for language descriptions expressing temporal variations of objects due to the lack of video-level multi-modal understanding. To address this issue, SOC [9], MUTR [17] efficiently aggregate inter and intra-frame information. Meanwhile, UNINEXT [16] proposes a unified prompt-guided formulation for universal instance perception, reuniting previously fragmented instance-level subtasks into a whole and achieve good performance for the RVOS task.

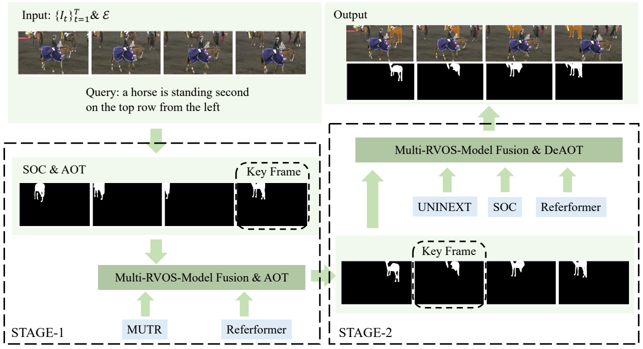

Input: $\{I_t\}_{t=1}^{T}$& $\mathcal{E}$
Query: a horse is standing second on the top row from the left
SOC & AOT
Key Frame
Multi-RVOS-Model Fusion & AOT
STAGE-1
MUTR
Referformer
Output
Multi-RVOS-Model Fusion & DeAOT
UNINEXT
SOC
Referformer
Key Frame
STAGE-2

Figure 1. The overall architecture of our method.

In our work, we incorporate benefits of the previous mainstream works to provide an effective paradigm. By utilizing the model ensemble strategy as well as semi-supervised VOS approaches as post-process to enhance the masks quality in each stage, we develop a Two-Stage Multi-model Fusion strategy. Specifically, we select AOT [18] to preliminary improve the masks quality in the first stage, but with the increase of propagation layers and number of RVOS models that are processed by that, it will inevitably lead to loss of information unrelated to the object, which may weaken the effect of the consequent model fusion. Therefore, on the basis of the high quality mask sequences from the first stage, we further exploit the potential of multi-model fusion by utilizing DeAOT [19] in the second stage.

The final leaderboard shows that our method ranks 1st

†Equal Contribution

‡This work is done during internships at ByteDance

place in the 5th Large-scale Video Object Segmentation Challenge (ICCV 2023): Referring Video Object Segmentation track.

2. Related Work

Semi-supervised Video Object Segmentation The objective of Semi-supervised Video Object Segmentation (VOS) is to achieve accurate object segmentation throughout the entire video sequence by utilizing provided one (generally at the first frame) or more mask annotation. The works [3, 10] focus on improving run-time efficiency and matching feature correlation between the target and other potential objects in the sequence to enhance the object tracking process. In addition, FEELVOS [13] extends the pixel-level matching mechanism by additionally doing local matching with the previous frame. STM [11] leverages a memory network to store past-frame predictions and apply attention mechanism to propagate the mask information. Recently, AOT [18] introduces hierarchical propagation into VOS and employs an identification mechanism to associate multiple targets. Furthermore, DEAOT [19] decouples object-agnostic and object-specific features in hierarchical propagation. In this work, we include two methods mentioned above as post-process.

Referring Video Object Segmentation. Referring Video Object Segmentation (RVOS) is first proposed by Gavrilyuk *et.al.* [5], which aims to generate a series of binary segmentation masks of the instance referred by the natural language description across a video clip. URVOS [12] introduces a large-scale RVOS benchmark and a unified framework that leverages attention mechanisms and mask propagation with a semi-supervised VOS method. Similar to URVOS [12], some approaches [1, 7] process each frame of the video clip separately through an image-level model. Meanwhile, compared to [14, 20] rely on complicated pipelines, MTTR [2] and Referformer [15] first adopt end-to-end framework modeling the task as the a sequence prediction problem, which greatly simplifies the pipeline. Currently, SOC [9] and MUTR [17] achieve excellent performance by efficiently aggregating intra and inter-frame information. What's more, UNINEXT [16] reformulates diverse instance perception tasks into a unified object discovery and retrieval paradigm. In this work, we combine the advantages of the above methods to obtain high-quality mask sequences.

3. Method

Given T frames of video clip $\mathcal{I} = \{I_t\}_{t=1}^{T}$, where $I_t \in \mathbb{R}^{3\times H_0\times W_0}$ and a referring text expression $\mathcal{E} = \{e_i\}_{i=1}^{L}$, where e_i denotes the i-th word in the text. RVOS task is to generate a series of binary segmentation masks $\mathcal{S} = \{s_t\}_{t=1}^{T}$, $s_t \in \mathbb{R}^{1\times H_0\times W_0}$ of the referred object.

3.1. Backbone

We adopt SOC [9], MUTR [17], Referformer [15] and UNINEXT [16], the current prevalent RVOS models, as our backbones to respectively generate binary segmentation masks $\mathcal{S} = \{s_t\}_{t=1}^{T}$.

$$\mathcal{S}^n = \mathcal{F}^n\left(\mathcal{I}, \mathcal{E}\right) \quad n \in \{soc, mutr, ref, uninext\}, \tag{1}$$

where $\mathcal{F}^n$ indicates the corresponding backbone. We train SOC jointly on RefCOCO and Ref-Youtube-VOS datasets, while we directly use checkpoints with the highest performance without training for other models.

3.2. Post-process

The video object segmentation has been proved to improve the segmentation mask consistency by object propagation mechanism. Specifically, [6] adopts AOT [18] as post-process to enhance the quality of mask results generated by RVOS models, which brings a clear improvement in accuracy. The general procedure are first selecting the keyframe index of mask sequences probability $\mathcal{P}$ from RVOS model, then using VOS model to perform forward and backward propagation. It can be formulated as:

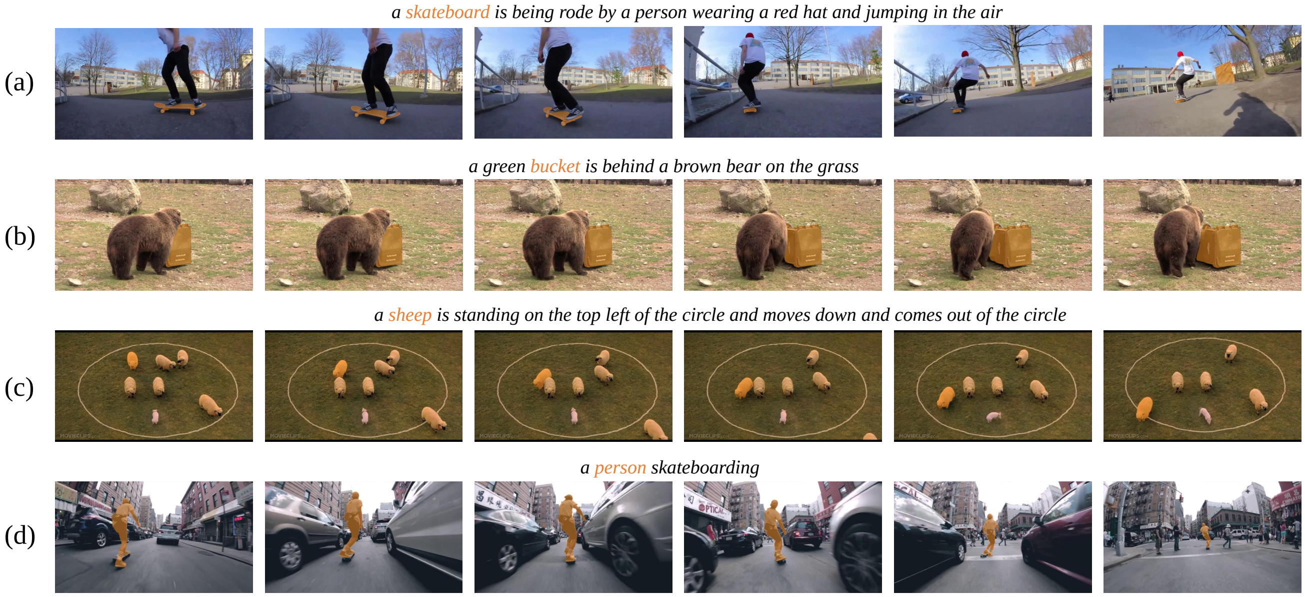
$$\begin{gathered}\mathcal{K}_{index} = argmax(\mathcal{P}),\\ \mathcal{M}^n = \left[\mathcal{G}\left(\{s_i^n\}_{i=K_{index}}^{0}\right), \mathcal{G}\left(\{s_j^n\}_{j=K_{index}}^{T}\right)\right],\end{gathered} \tag{2}$$

where $\mathcal{G}$ denotes the VOS model for post-process.

In our experiment, we find that although AOT can facilitate the temporal quality of mask results, the benefit decreases when conducting the Stage II fusions (which is elaborate on Sec. 3.3). It is hypothesized that AOT potentially lead to loss of object-agnostic visual information in deep propagation layer. Consequently, the advantage of ensemble is degraded due to the aggregated object information loss from different RVOS models that are post-processed by the the same VOS model. Intuitively, we propose to use two VOS models for post-processing in different stages to alleviate the problem.

3.3. Two-Stage Multi-model Fusion

We find that models with different frameworks process the object referred by the textual description in different perspectives. SOC unifies temporal modeling and cross modal alignment to achieve video-level understanding, which comprehends expressions containing temporal variations well. Due to the object discovery and retrieval paradigm, UNINEXT has strong ability of localizing and tracking same objects referred by different textual descriptions. MUTR introduces temporal transformer to facilitate objects interaction across frames, improving consistency of



a skateboard is being rode by a person wearing a red hat and jumping in the air

a green bucket is behind a brown bear on the grass

a sheep is standing on the top left of the circle and moves down and comes out of the circle

a person skateboarding

Figure 2. Visualization results on Ref-Youtube-VOS.

masks. In order to make full use of advantages of different frameworks, we propose two-stage multi-model fusion strategy. Similar to [6], we fuse the masks predicted by different referring expressions that describe the same target from different models, which is formulated as:

$$\hat{y} = \sum_{n}^{N} \sum_{q}^{Q} \mathcal{M}_q^n, \qquad \hat{y'}_i = \begin{cases} 0 & \hat{y}_i < thr \\ 1 & \hat{y}_i >= thr \end{cases}, \tag{3}$$

where Q denotes the number of different textual descriptions referred to the same object and N indicates the models. $i \in \{1, 2, \ldots, HW\}$ where H, W are the width and height of the mask respectively.

Stage I Referformer treats the task as sequence prediction problem and perform cross modal interaction in each frame. Its simple framework could serve as a baseline to segment the referred object but easily fails to capture the temporal variation of object across frames. We believe that SOC and MUTR can explicitly increase the inter-frame interaction, which is a reasonable compensation for Referformer. Therefore, in the first stage, we fuse three models and use AOT as post-process to enhance the mask quality. For clarity, the fused model is denoted as SMR.

Stage II UNINEXT is jointly trained with prevalent datasets of 10 instance perception tasks. It is capable of perceiving diverse objects referred by different descriptions, thanks to static object queries which absorb rich information from data in different domain. Although it achieve high performance with VIT-Huge backbone [4] by feeding large scale of data, the lack of global view of object may cause the inconsistency when generating masks across frames. Therefore, we solve this problem by two-fold. (1) Employ DeAOT to propagate the object information from the key frame to another. (2) Ensemble with the SMR fused model to integrate information from inter-frame interaction.

Model	$\mathcal{J}$ & $\mathcal{F}$ ↑
SOC	67.5
+AOT	69.5 (+2.0)
+Multi-model Fusion (Stage I) & AOT	72.4 (+2.9)
+Multi-model Fusion (Stage II) & DeAOT	75.7 (+3.3)

Table 1. Ablation study of each module on our model's performance on **validation set**.

4. Experiment

4.1. Dataset and Metrics

Datasets. We evaluate our model on Ref-Youtube-VOS dataset of *2023 Referring Youtube-VOS challenge*. It contains 3,978 high-resolution YouTube videos with about 15K language expressions. These video are divided into 3,471 training videos, 202 validation videos and 305 test videos.

Metrics. we adopt standard evaluation metrics: region similarity ($\mathcal{J}$), contour accuracy ($\mathcal{F}$) and their average value ($\mathcal{J}$&$\mathcal{F}$) on Ref-Youtube-VOS.

4.2. Training Detail

We train SOC with pretrained Video Swin Transformer and RoBERTa as the encoder for 30 epochs. The model is optimized by Adam optimizer with the initial learning rate of 1e-4. During training, we apply RandomResize and Horizontal Flip for data augmentation. Specifically, all frames are downsampled to 360×640. In post-process, we follow [6] retrain DeAOT network with Swin-L backbone using default parameters in [19].

4.3. Main Results

Our method achieves 70% $\mathcal{J}\&\mathcal{F}$ on test set which outperforms the next team by 4% $\mathcal{J}\&\mathcal{F}$ and rank 1st place in Large-scale Video Object Segmentation Challenge (ICCV 2023): Referring Video Object Segmentation track.

4.4. Ablation Study

To validate the effectiveness of each module, we conduct simple ablation studies. As we mention above that we use SOC [9], MUTR [17], Referformer [15] and UNINEXT [16] as RVOS models to generate mask results for post-processing and fusion. It is noted that SOC is the main model that are included in two stages, and for simplicity, we set it as the baseline. As shown in Tab. 1, the preliminary model fusion and post-process with AOT in stage I, brings an improvement of 2.9% $\mathcal{J}\&\mathcal{F}$. While the fused model achieve 75.7% $\mathcal{J}\&\mathcal{F}$ with the second stage model ensemble, demonstrating the rationality and significance of our proposed two-stage multi models fusion.

4.5. Qualitative Results

Fig. 2 shows the prediction of our method for complex scenarios segmentation, *i.e.*, similar appearance, occlusion and large variations. It can be seen that our method precisely segments the referred object.

References

[1] Miriam Bellver, Carles Ventura, Carina Silberer, Ioannis Kazakos, Jordi Torres, and Xavier Giró-i-Nieto. Refvos: A closer look at referring expressions for video object segmentation. *CoRR*, abs/2010.00263, 2020. 1, 2

[2] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *CVPR*, pages 4975–4985, 2022. 1, 2

[3] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *NIPS*, pages 11781–11794, 2021. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[5] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees G. M. Snoek. Actor and action video segmentation from a sentence. In *CVPR*, pages 5958–5966, 2018. 2

[6] Zhiwei Hu, Bo Chen, Yuan Gao, Zhilong Ji, and Jinfeng Bai. 1st place solution for youtubevos challenge 2022: Referring video object segmentation. *arXiv preprint arXiv:2212.14679*, 2022. 1, 2, 3, 4

[7] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV*, volume 11364, pages 123–141. Springer, 2018. 1, 2

[8] Chen Liang, Yu Wu, Tianfei Zhou, Wenguan Wang, Zongxin Yang, Yunchao Wei, and Yi Yang. Rethinking cross-modal interaction from a top-down perspective for referring video object segmentation. *CoRR*, abs/2106.01061, 2021. 1

[9] Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. SOC: semantic-assisted object cluster for referring video object segmentation. *CoRR*, abs/2305.17011, 2023. 1, 2, 4

[10] 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. 2

[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. 2

[12] Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, pages 208–223, 2020. 2

[13] 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. 2

[14] Hao Wang, Cheng Deng, Fan Ma, and Yi Yang. Context modulated dynamic networks for actor and action video segmentation with language queries. In *AAAI*, pages 12152–12159, 2020. 2

[15] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, pages 4964–4974. IEEE, 2022. 1, 2, 4

[16] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023. 1, 2, 4

[17] Shilin Yan, Renrui Zhang, Ziyu Guo, Wenchao Chen, Wei Zhang, Hongyang Li, Yu Qiao, Zhongjiang He, and Peng Gao. Referred by multi-modality: A unified temporal transformer for video object segmentation. *CoRR*, abs/2305.16318, 2023. 1, 2, 4

[18] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *NIPS*, 34:2491–2502, 2021. 1, 2

[19] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. *NIPS*, 35:36324–36336, 2022. 1, 2, 4

[20] Yuting Zhang, Luyao Yuan, Yijie Guo, Zhiyuan He, I-An Huang, and Honglak Lee. Discriminative bimodal networks for visual localization and detection with natural language queries. In *CVPR*, pages 1090–1099, 2017. 2