Referring Video Object Segmentation using Spatial-Temporal Propagation

Mingqi Gao[1,2], Jungong Han[1,3], Feng Zheng[2], James J.Q. Yu[2], Giovanni Montana[1]
[1]WMG, University of Warwick
[2]Department of Computer Science and Engineering, Southern University of Science and Technology
[3]Department of Computer Science, Aberystwyth University

Abstract

This technical report introduces our submission to the referring video object segmentation task of YouTube-VOS Challenge 2022. Our approach explores a fusing strategy to combine the existing advances in Referring Video Object Segmentation (RVOS) and Semi-supervised Video Object Segmentation (SVOS). Given an input video sequence and a text describing the target object, we first perform RVOS to generate the initial segmentation results. Next, we compute the relevance between the input text and each video frame. Finally, the confident results from the relevant frames are selected and propagated to the remaining frames. This approach can improve the existing implementations in temporal consistency and achieve competitive results on Ref-YouTube-VOS test set [17].

1. Introduction

Referring Video Object Segmentation (RVOS) aims to segment the object described by a text from the input video sequence. This task is formulated initially to segment the object with specific actions [6]. Therefore, the text descriptions in the early RVOS datasets are action-oriented. More recently, RVOS has been reformulated to segment general objects with unconstrained descriptions [9, 17].

Analogously to Semi-supervised Video Object Segmentation (SVOS) [15], RVOS focuses on segmenting specific objects, rather than general objects from videos. The main difference between these tasks lies in the prompt format for the objects to segment: In SVOS, the prompt is the human-annotated object mask on one video frame. By contrast, RVOS methods only infer object masks under the guidance of language expressions. Therefore, the prompt in RVOS is more user-friendly and requires less effort than SVOS during inference, making RVOS valuable in human-computer interaction applications.

The most frequently used paradigm for RVOS is first to encode visual and language features from input videos and texts. Then segmentation models could focus on the text-referred object by interacting between multi-modal features. Based on different ways to process video frames, the existing RVOS methods could be grouped into two categories: Sequential methods [6, 8, 10, 17, 19, 20, 23] and parallel methods [1, 13, 21]. The former performs segmentation frame-by-frame, where cross-modal attention and dynamic convolution are the frequently used techniques to interact between visual and language features. By contrast, The latter infers the entire input video with one feed-forward pass. To do so, they reformulate RVOS as the task of sequence prediction, where a sequence of binary masks is generated to cover the text-referred object on all video frames. Therefore, they can achieve better efficiency and consider more global context than the sequential methods.

Among the parallel methods, the transformer-based ones [1, 21] reach the state-of-the-art performance. They both build their models based on the DETR architecture [3]. The main difference between them lies in the decoding process: MTTR [1] infers all objects from the video and then filters out the irrelevant ones based on the text information. On the other hand, ReferFormer [21] directly segments the text-referred object by considering texts as queries. In addition to RVOS datasets, ReferFormer trains the model on the RIS datasets (Referring Image Segmentation, e.g., Ref-COCO series [12, 24]), further improving the performance.

Although achieving good results, there is still a problem that the existing methods (including both sequential and parallel ones) cannot handle well: Most methods consider each frame individually for multi-modal inference, resulting in inconsistent segmentation results. For example, the segmented masks switch between the target object and other background objects due to scene changes or less optimal language understanding. To mitigate this problem, we explore a fusing strategy on top of the methods for RVOS and SVOS. In particular, given an input video and text, we perform RVOS first on all video frames to generate the initial results. Next, the relevance scores between each frame and the text are measured. Finally, confident results are selected from the high-scored frames and propagated to the remaining frames. To some extent, our proposed approach can

avoid false-positive results since they usually come from the low-relevant frames to the text. After the initial inference, these results would be replaced with more temporally consistent ones propagated from high-relevant and confident object masks. The competitive results on Ref-YouTube-VOS test sets [17] demonstrate the effectiveness of our proposed approach.

2. Related Works

Referring Video Object Segmentation Given a video sequence and text, Referring Video Object Segmentation (RVOS) aims to segment the text-referred object on all video frames. Therefore, the key to high-quality results is interacting well between multi-modal features. The early methods [6, 8, 10, 17, 19, 20, 23] encode visual and language features with CNNs and RNNs/linear layers, respectively. Then they generate text-related object masks via dynamic convolution or cross-modal attention. More recently, full-transformer models [1, 21] (including visual/language encoders and multi-modal modules) improve the RVOS performance in both accuracy and efficiency due to better feature interaction and parallel framework.

Semi-supervised Video Object Segmentation Given a video sequence and an annotated object mask, Semi-supervised Video Object Segmentation (SVOS) aims to propagate the mask to other video frames. The key to high-quality results is recognising visual and semantic features of the annotated object and maintaining consistent predictions throughout the sequence. Most early methods achieve this via online fine-tuning [2, 18]. To further improve the performance and save computation cost, current methods tend to perform SVOS via feature matching between the annotated/segmented frames and the frames to segment [4, 14].

3. Method

To achieve temporally consistent results, we explore a fusing strategy on top of existing RVOS and SVOS advances. Figure 1 illusrates the architecture of the proposed approach, which consists of three steps: (1) referring video object segmentation, (2) relevance score computation, and (3) confident result propagation.

Referring Video Object Segmentation Given an input video sequence and a text describing the object to segment, this stage aims to perform referring video object segmentation to generate initial results. We utilise the RVOS model ReferFormer [21] with Video-Swin-Transformer-Base backbone [11] to predict object masks on each frame.

Relevance Score Computation We improve the masks with the propagation-based method since temporal consistency is not fully considered during initial mask prediction. For each video, we select the most confident mask and propagate it to the remaining frames. The confidence comes from the instance-level probabilities by Referformer [21]. Apparently, confidence is critical to the final results since it indicates the "seed" object mask to propagate. Once the false-positive result (e.g., background object) is selected, it would degrade the overall results initialised from the corresponding video. To mitigate this issue, we incorporate the well-trained vision-language encoder [16] with ViT-L/14 backbone [5] to compute the relevance score between each frame and the text. To some extent, the video frames only containing background objects could be excluded to avoid considering false-positive results.

Table 1. Results on the 4th Large-scale Video Object Segmengation Challenge - Track 3: Referring Video Object segmentation.

#	Team	Overall ↑	$\mathcal{J}$ ↑	$\mathcal{F}$ ↑
1	Bo___	0.641	0.622	0.661
2	jiliushi	0.617	0.598	0.636
3	PENG	0.608	0.589	0.627
4	**ds-hohhot**	**0.596**	**0.579**	**0.612**
5	JQK	0.594	0.577	0.611
6	nero	0.580	0.561	0.599

Confident Mask Propagation For each video, we select the confident mask from the video frames with top 60% relevance scores. After this, the result can be propagated to the remaining frames to refine the overall performance in temporal consistency. The propagation is achieved by STCN [4] with ResNet-101 backbone [7], pretrained as in [4] and then fine-tuned on YouTube-VOS-2019 [22] only.

4. Experiments

Given an input video sequence (generally with the resolution of 1280×720 in YouTube-VOS datasets), we generate the initial results from downsampled videos (shorter side = 360, as in [21]). To compute relevance scores, we pad and resize the videos to 224×224 for visual feature embedding. With the initial results (original resolution) and scores, we perform mask propagation with full resolution.

As shown in Table 1, our approach achieves competitive performance on the RVOS task of YouTube-VOS Challenge 2022. We also show some qualitative results in Figure 2 to illustrate the effectiveness of the approach.

5. Conclusion

This technical report explores a fusing strategy for RVOS, where the advances in SVOS and vision-language

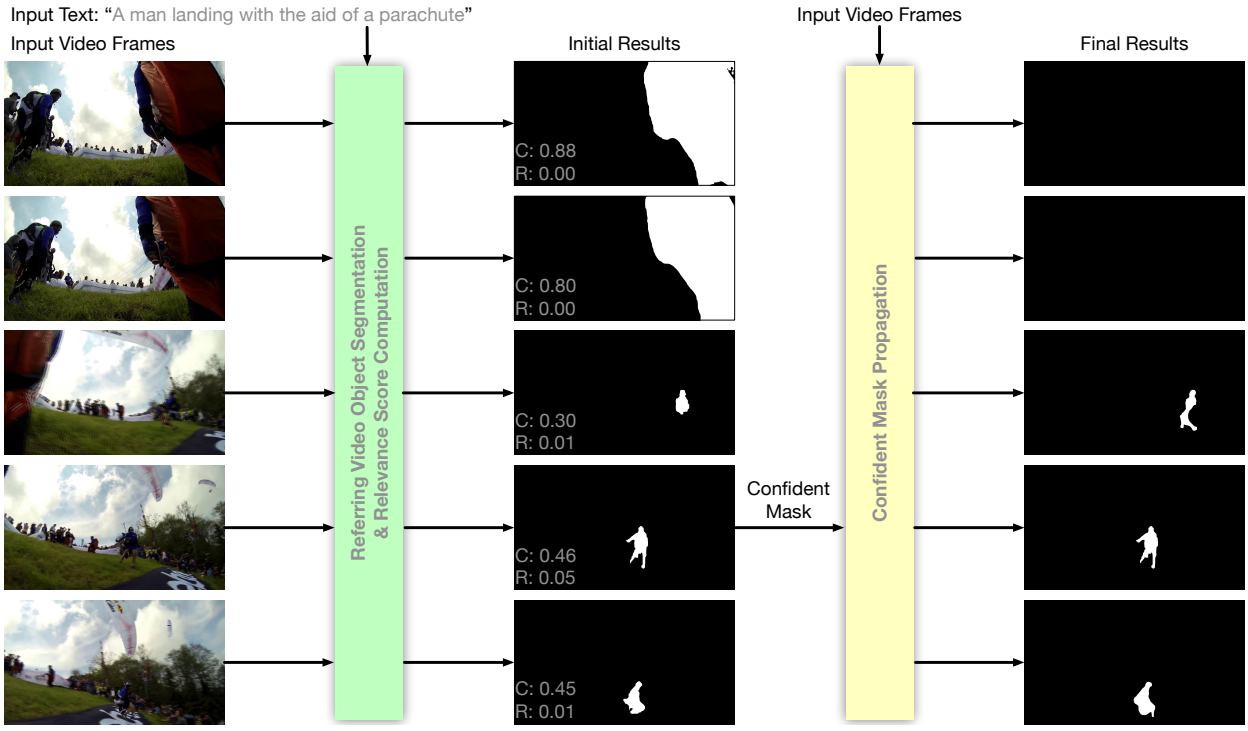

Figure 1. The framework of the proposed approach. C: confidence score, R: relevance score (after softmax across all video frames).

embedding are incorporated to refine the segmentation results in temporal consistency. The competitive results on the challenge dataset show the effectiveness of the proposed approach. In the future, we would explore more flexible and compact framework for end-to-end RVOS.

References

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

[2] 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

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 1

[4] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *NeurIPS*, 34, 2021. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[6] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *CVPR*, pages 5958–5966, 2018. 1, 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[8] Tianrui Hui, Shaofei Huang, Si Liu, Zihan Ding, Guanbin Li, Wenguan Wang, Jizhong Han, and Fei Wang. Collaborative spatial-temporal modeling for language-queried video actor segmentation. In *CVPR*, pages 4187–4196, 2021. 1, 2

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

[10] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *IEEE TPAMI*, 2021. 1, 2

[11] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022. 2

[12] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 1

[13] Bruce McIntosh, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. Visual-textual capsule routing for text-based video segmentation. In *CVPR*, pages 9942–9951, 2020. 1

[14] 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

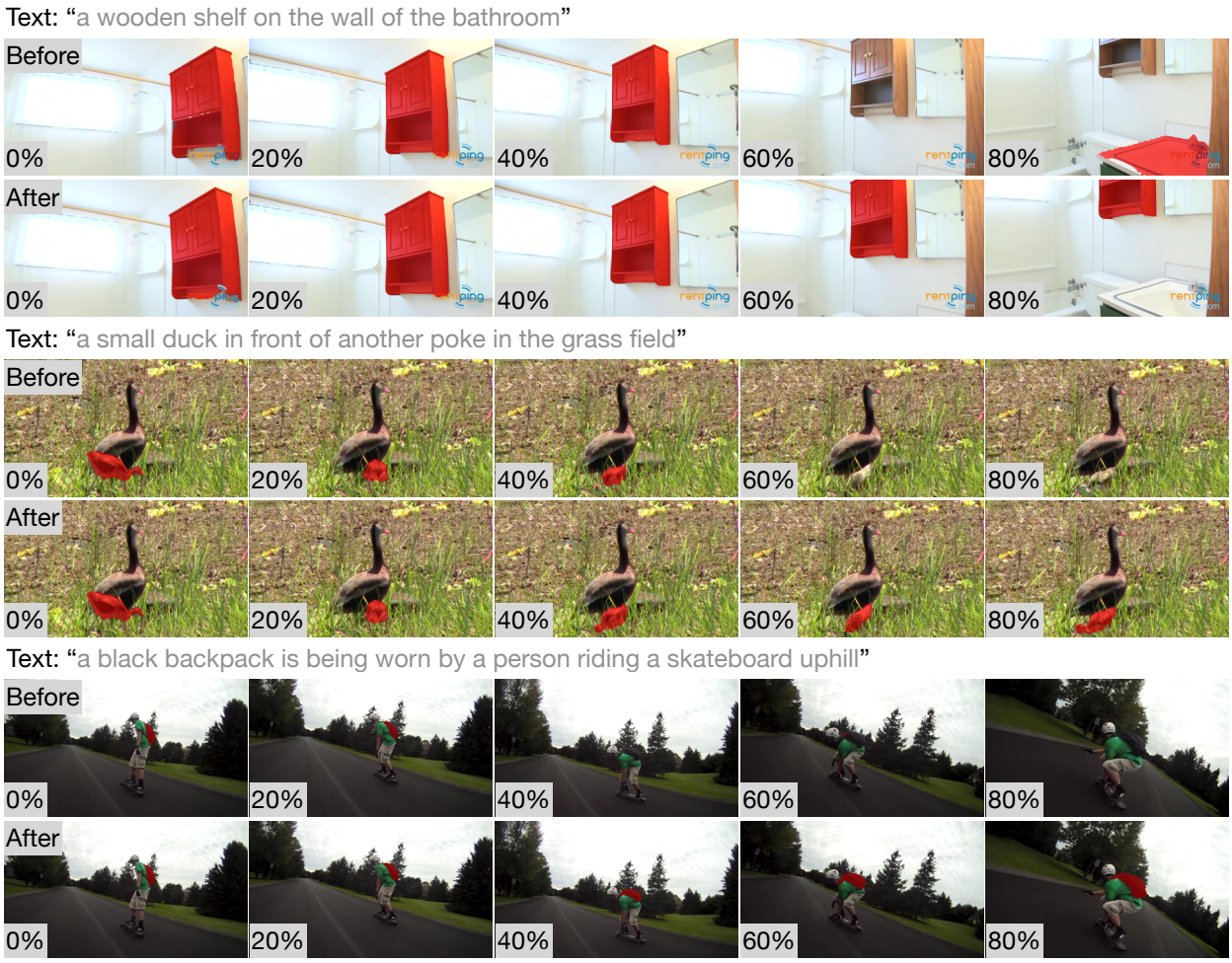

Figure 2. The qualitative results generated before and after propagation. The segmented objects are highlighted in red masks. The percentage indicates the relative position of each frame in the video.

[15] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016. 1

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2

[17] 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. Springer, 2020. 1, 2

[18] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. In *BMVC*, 2017. 2

[19] Hao Wang, Cheng Deng, Fan Ma, and Yi Yang. Context modulated dynamic networks for actor and action video segmentation with language queries. In *AAAI*, volume 34, pages 12152–12159, 2020. 1, 2

[20] Hao Wang, Cheng Deng, Junchi Yan, and Dacheng Tao. Asymmetric cross-guided attention network for actor and action video segmentation from natural language query. In *ICCV*, pages 3939–3948, 2019. 1, 2

[21] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, 2022. 1, 2

[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*, 2018. 2

[23] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, pages 10502–10511, 2019. 1, 2

[24] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85. Springer, 2016. 1