The Second Place Solution for YouTube-VOS Challenge 2023: Referring Video Object Segmentation

Yan Li
Zhejiang University of Technology
Hangzhou, China

Qiong Wang
Zhejiang University of Technology
Hangzhou, China

Abstract

Referring Video Object Segmentation aims to produce segmentation masks of objects referred to by the referring expressions for each frame in a video sequence. Previous methods mainly take as input one referring expression to identity and segment the referred object. Due to the randomness of referring expressions provided by users, it is still challenging for existing models with one expression guidance to distinguish which object is referred to in a complex scenario. In this work, inspired by an early work [1] that reports the influence of components of referring expressions, we present a simple and effective approach "Ranking Referring Segmentation (RankRefSeg)" that takes advantage of the available multiple referring expressions to predict the referred object. The proposed approach RankRefSeg predicts the segmentation masks through ranking the probabilities of candidate indexes from multiple referring expressions, without the needs of re-training. Compared to the state-of-the-art methods, the proposed method achieves 71.2% the average value of the region similarity and the contour accuracy on the Refer-Youtube-VOS validation set, and 66% on the Refer-Youtube-VOS test set, ranking in the 2nd place on the Referring Youtube-VOS challenge 2023.

1. Introduction

Video Object Segmentation (VOS) is a fundamental task aiming at the segmentation of object instances across a video sequence. The unsupervised method segments the object without any guidance, which can make of ambiguity in identifying the target object. Great progress has been made in the exploration of various guidance that are beneficial to the disambiguation. The semi-supervised method involves one or more mask annotation of video frames, which is typically tedious and time-consuming. Scribble or click annotation is investigated in interactive video object segmentation, which allows user intervention during the inference. Recently, Referring Video Object Segmentation (RVOS) was introduced by the method [3], in which the target object is referred to by the given referring expression. Due to the availability and the high flexibility of language expression, RVOS has attracted much attention and has been applied to many practical problems involving human-computer interaction and video editing.

Up to now, Refer-Youtube-VOS [4] is one of the most popular used large-scale benchmark for RVOS. Regarding the language expression generation of this benchmark, Amazon Mechanical Turk is employed to annotate referring expressions. Specifically, given the original video and the mask-overlaid highlighted object, each Turker was asked to provide natural expressions to describe the target objects. As a result, almost two referring expressions are provided for each target instance in the Refer-Youtube-VOS dataset. Due to the high flexibility of the language expression, those available multiple referring expressions usually behave in various types [1], such as different length of language expression, different combinations of semantic attributes. Since the referred object could be identified with each referring expression by one of annotators, previous methods explored the alignment between language and vision with one referring expression input only. Despite progress achieved, understanding which object is referred to from single natural referring expression might still be insufficient for existing RVOS models, especially in a complex scenario (e.g. a scenario where multiple objects share similar appearance). Moreover, when some semantic attributes in the expression are poorly learned by neural networks, completely wrong segmentation mask might be produced. To address the above-mentioned concerns, we make attempt to explore the way of guidance using two provided referring expressions. Our method is built on top of the state-of-the-art method UNINEXT [6], which is fed with two referring expressions so as to accurately segment the referred object.

Corresponding author: wangqiong@zjut.edu.cn. This work was supported by the National Natural Science Foundation of China (Grant NO. 62002326)

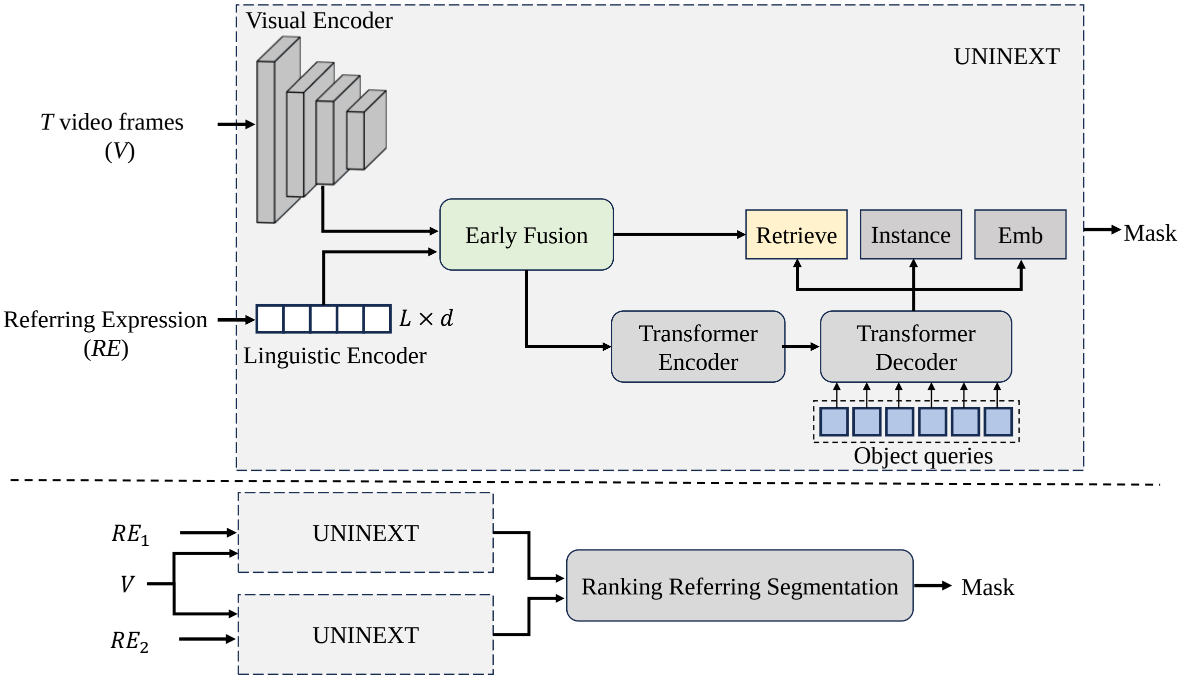


Figure 1. Overall pipeline of the proposed method. The upper part is used in the training, and the lower part used in the inference.

To summarize, we make the following contributions:

- A novel RVOS method is proposed to predict the target object with two referring expressions, without the needs of re-training. The proposed ranking mechanism ranks the probabilities of candidate indexes induced from two referring expressions to obtain the final prediction.
- We make evaluations of the proposed method on the Refer-Youtube-VOS [4] dataset. The proposed method achieves the gains 1.1% on the validation set than the baseline model UNINEXT [6], and achieves the 2nd place on the Referring YouTube-VOS Challenge 2023.

2. Method

We present a novel RVOS method that takes into account two referring expressions. Specifically, given a video clip V with a sequence of T video frames and two referring expressions for each referred object denoted by RE_1 and RE_2, V and RE_1, and V and RE_2 are input into UNINEXT [6] respectively. Then, a simple but effective approach Ranking Referring Segmentation (RankRefSeg) is proposed to produce the final mask prediction by ranking the maximum scores of RE_1 and RE_2. Figure 1 shows the overall pipeline of the proposed method.

2.1. Baseline: UNINEXT

Given a video sequence and a referring expression, the UNINEXT begins with the Visual Encoder and Linguistic Encoder for the feature extraction. The visual features and the text features are fused in Early Fusion and then fed into the Transformer Encoder. The output of the Transformer Encoder and object queries are then input to the Transformer Decoder. After the Instance Segmentation process, the outputs of model are fed into the proposed Ranking Referring Segmentation for final predictions.

The whole training of UNINEXT consists of three main stages: (1) general perception pretraining, (2) image-level joint training, (3) video-level joint training. To be more specific, in the first stage, UNINEXT is pretrained on the large-scale object detection dataset. Then based on the pretrained weights of the first stage, UNINEXT is fine-tuned jointly on the mixed image datasets. Finally, UNINEXT is further fine-tuned on video-level datasets. The training data in the third stage includes pseudo videos generated from mixed image datasets, Single Object Tracking (SOT) and Video Object Segmentation (VOS) datasets, Multiple Object Tracking (MOT) and Video Instance Segmentation (VIS) datasets, and RVOS dataset (i.e. Ref-Youtube-VOS). More details can be found in the paper [6].

2.2. Proposed Ranking Referring Segmentation

To overcome the challenging interaction between vision and language, a novel ranking segmentation mechanism concerning the probability of being referred is proposed. The class head, implemented as a feed forward network, predicts the classification probability for each query from the decoded hidden values of all frames, denoted by

$\mathbf{P} = [p_n^t]_{N \times T}$, $p_n^t \in \mathbb{R}$, where N is the total number of queries. To smooth out noises resulted from single frames, the average of probabilities over all video frames is computed for each queried instance sequence p_n.

$$p_n = \frac{1}{T} \sum_{t=1}^{T} (p_n^t) \tag{1}$$

Then, for each referring expression, S_P is calculated as the maximum score among all queries, which are obtained after ranking all p_n.

$$S_P = \max\{p_1, p_2, ..., p_N\} \tag{2}$$

As two referring expressions corresponds to the same referred object, one with a high confidence value should be selected. Consequently, the final index ind for the referred object is obtained by comparing the maximum scores of two referring expressions.

$$ind = \begin{cases} 1 & S_P^1 > S_P^2 \\ 2 & else \end{cases} \tag{3}$$

3. Experiments

3.1. Dataset and Metrics

Refer-Youtube-VOS [4] is a large-scale benchmark and covers 3,978 YouTube videos from YouTube-VOS dataset [5] with around 15k language expressions. 1,063 unique objects with 2,096 expressions are covered in 202 validation videos and 305 test videos. On average, each object has around 2 language expressions and each expression has 10.0 words. Referring Youtube-VOS challenge uses the test set of Refer-Youtube-VOS for the competition. As the annotations of validation and test set are not released publicly, both validation and test predictions needs to be submitted to the evaluation servers.

The standard metrics are used: region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$ and average value of $\mathcal{J}$ and $\mathcal{F}$. All compared results of Ref-Youtube-VOS are evaluated using the online validation server and test server of the challenge.

3.2. Results on YouTubeVOS Challenge 2023

As shown in Table 1, our approach using the ViT-H backbone [2] achieves 66% in $\mathcal{J}$ & $\mathcal{F}$, ranking in 2nd place on the Referring Youtube-VOS challenge 2023.

3.3. Ablation Study

To verify the effectiveness of the proposed Ranking mechanism on RVOS, the performance without and with the proposed RankRefSeg are compared, as shown in Table 2. The experimental results show that the proposed RankRefSeg improves the segmentation performance in all metrics.

Table 1. Performance comparison results on the Ref-Youtube-VOS test set.

Method	Ref-Youtube-VOS Test		
	$\mathcal{J}$ & $\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Robertluo	70	68	72
Ours	66	64	68
biubiubiubiu	60	59	62
MahouShoujo	60	58	61

Table 2. Performance comparison results on the Refer-Youtube-VOS validation set.

Method	backbone	Refer-Youtube-VOS Val		
		$\mathcal{J}$ & $\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Ours w/o Ranking	ViT-H	70.1	67.6	72.7
Ours	ViT-H	**71.2**	**68.7**	**73.7**

3.4. Visualization

Figure 2 and Figure 3 show the visual predictions of the proposed method on the Refer-Youtube-VOS validation set and test set respectively. The visualization results illustrate that our method is capable of predicting the correct referred objects in long-term videos. Moreover, the proposed method can produce temporarily consistent segmentation masks for video frames.

4. Conclusion

In this work, we present a simple but effective approach, called RankRefSeg, which is built upon the state-of-the-art RVOS model. This approach helps to alleviate the difficulties resulted from the randomness of referring expression in identifying the referred object. Evaluations are made on the Refer-Youtube-VOS benchmark, and RankRefSeg is ranked the 2nd place on the Referring YouTube-VOS Challenge 2023.

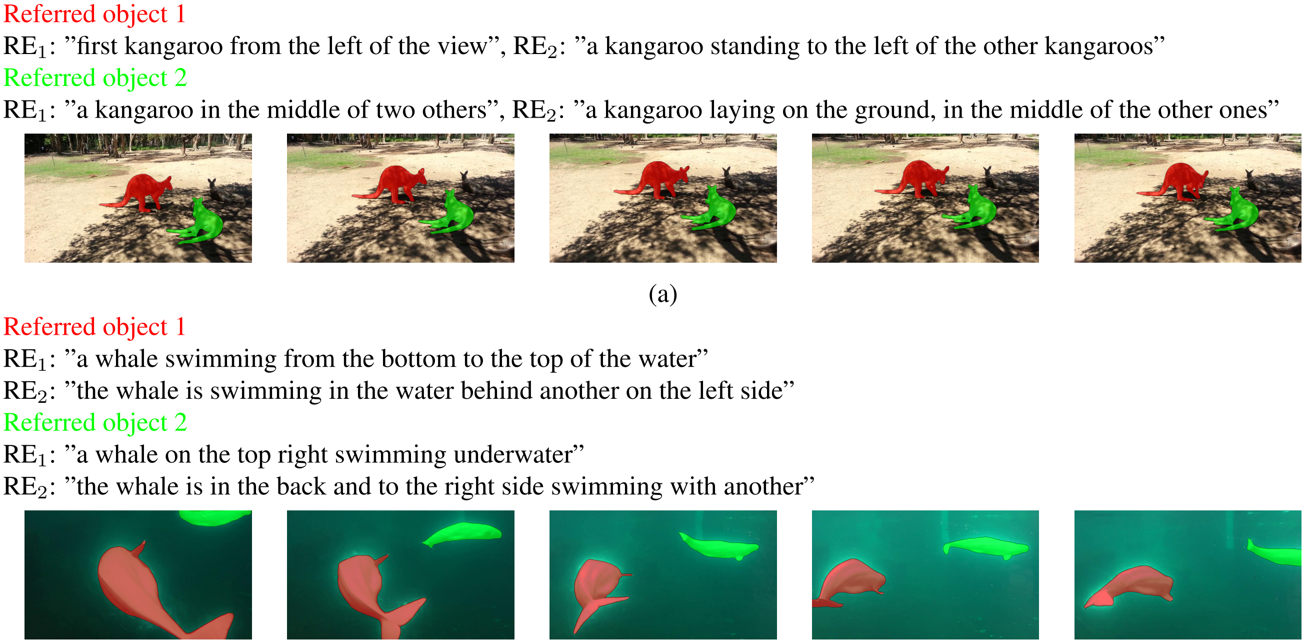
References

[1] Miriam Bellver, Carles Ventura, Carina Silberer, Ioannis Kazakos, Jordi Torres, and Xavier Giro-i Nieto. A closer look at referring expressions for video object segmentation. *Multimedia Tools and Applications*, 82(3):4419–4438, 2023.

[2] 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. *ArXiv*, abs/2010.11929, 2020.

[3] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5958–5966, 2018.

[4] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European Conference on Computer Vision*, pages 208–223. Springer, 2020.



Referred object 1

RE$_1$: ”first kangaroo from the left of the view”, RE$_2$: ”a kangaroo standing to the left of the other kangaroos”

Referred object 2

RE$_1$: ”a kangaroo in the middle of two others”, RE$_2$: ”a kangaroo laying on the ground, in the middle of the other ones”

(a)

Referred object 1

RE$_1$: ”a whale swimming from the bottom to the top of the water”

RE$_2$: ”the whale is swimming in the water behind another on the left side”

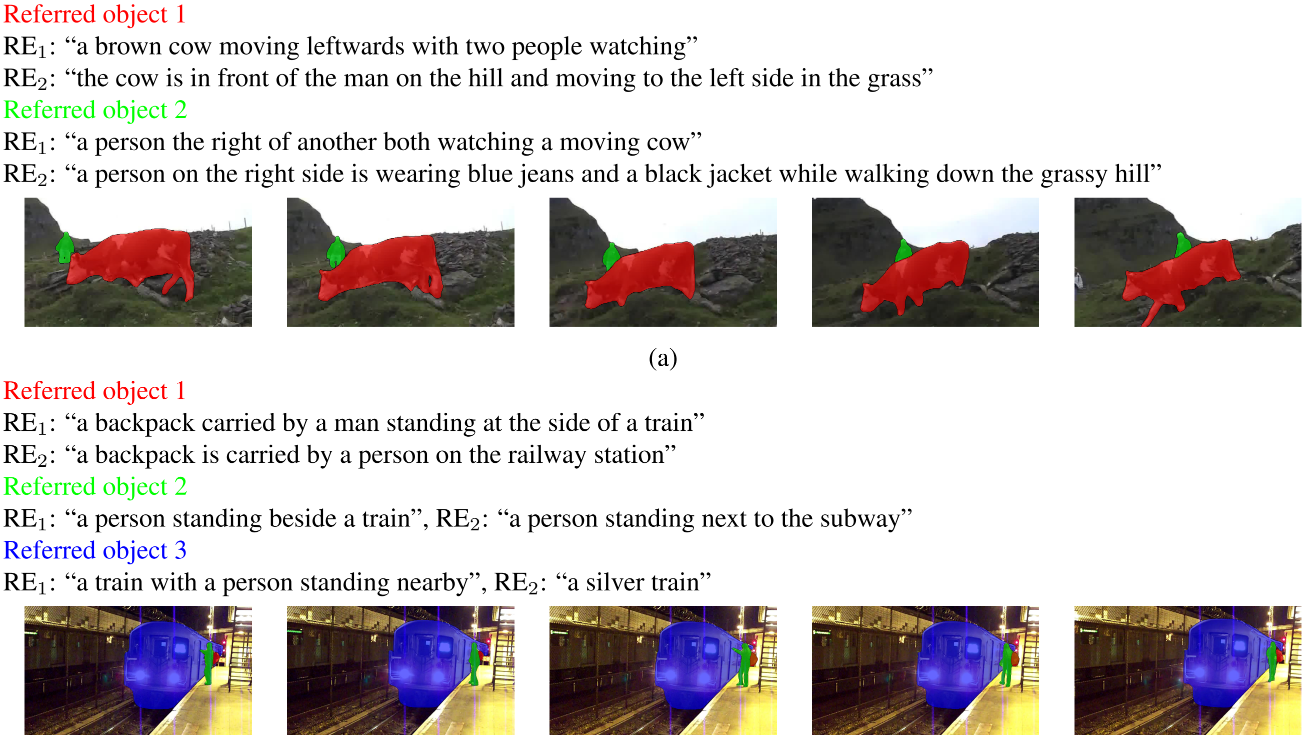
Referred object 2

RE$_1$: ”a whale on the top right swimming underwater”

RE$_2$: ”the whale is in the back and to the right side swimming with another”

(b)

Figure 2. Visualization results on Refer-Youtube-VOS validation set.



Referred object 1

RE$_1$: “a brown cow moving leftwards with two people watching”

RE$_2$: “the cow is in front of the man on the hill and moving to the left side in the grass”

Referred object 2

RE$_1$: “a person the right of another both watching a moving cow”

RE$_2$: “a person on the right side is wearing blue jeans and a black jacket while walking down the grassy hill”

(a)

Referred object 1

RE$_1$: “a backpack carried by a man standing at the side of a train”

RE$_2$: “a backpack is carried by a person on the railway station”

Referred object 2

RE$_1$: “a person standing beside a train”, RE$_2$: “a person standing next to the subway”

Referred object 3

RE$_1$: “a train with a person standing nearby”, RE$_2$: “a silver train”

(b)

Figure 3. Visualization results on Refer-Youtube-VOS test set.

[5] N. Xu, L. Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian L. Price, Scott D. Cohen, and Thomas S. Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *European Conference on Computer Vision*, 2018.

[6] 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.