# The Second Place Solution for The 4th Large-scale Video Object Segmentation Challenge——Track 3: Referring Video Object Segmentation

Leilei Cao[1,2], Zhuang Li[1,3], Bo Yan[1], Feng Zhang[1],Fengliang Qi[1],Yuchen Hu[1], Hongbin Wang[1]
[1]Ant Group [2]Northwestern Polytechnical University [3]Tongji University
mcaoleilei@sina.com, jiangzi.lz@antgroup.com, hongbin.whb@antgroup.com

## Abstract

*The referring video object segmentation task (RVOS) aims to segment object instances in a given video referred by a language expression in all video frames. Due to the requirement of understanding cross-modal semantics within individual instances, this task is more challenging than the traditional semi-supervised video object segmentation where the ground truth object masks in the first frame are given. With the great achievement of Transformer in object detection and object segmentation, RVOS has been made remarkable progress where ReferFormer achieved the state-of-the-art performance. In this work, based on the strong baseline framework——ReferFormer, we propose several tricks to boost further, including cyclical learning rates, semi-supervised approach, and test-time augmentation inference. The improved ReferFormer ranks 2nd place on CVPR2022 Referring Youtube-VOS Challenge.*

## 1. Introduction

Referring Video Object Segmentation (RVOS) [11], a fundamental task in computer vision, aims to segment object instances in a given video referred by a language expression in all video frames. A wide range of video-related applications refer to RVOS, e.g., video editing, video surveillance, and human-object interactions. Comparing with referring image segmentation in which objects are referred to by their appearance, the objects in RVOS are referred to by the actions they are performing [1, 16]. This makes solving RVOS more complicated. RVOS is also more challenging than the traditional semi-supervised video object segmentation in which the ground truth object masks in the first frame are given [19], because it requires understanding cross-modal semantics within individual instances. In addition, objects appearance variation, occlusion, and complicated background also greatly challenge RVOS.

The traditional methods for RVOS can be divided into two categories: (1) Bottom-up methods. These methods fuse features extracted from videos and language, and then adopt a decoder [9] to produce object masks. (2). Top-down methods. These methods first use an instance segmentation model to generate all object masks in videos to form tracklet candidates, and then use the language as the grounding criterion to select the best-matched one [16]. With the advancement of Transformers [4, 7, 13] in object detection [2, 22] and segmentation [3, 5, 14, 15, 18, 21], RVOS has been made remarkable progress [1, 6, 16]. Botach et al. [1] proposed a multimodal Transformer model to process video and text together, which is end-to-end trainable and requires no additional mask-refinement post-processing steps. Wu et al. [16] proposed a unified framework termed ReferFormer to segment and track the referred object in all frames in an end-to-end manner, in which the language expression is viewed as queries and directly attend to the most relevant regions in the video frames. And ReferFormer achieved state-of-the-art results on Ref-Youtube-VOS dataset.

In this work, based on the strong framework of ReferFormer, we propose several tricks to improve ReferFormer on RVOS. Instead of monotonically decreasing the learning rate during training, we use the cyclical learning rates to finetune the trained model of ReferFormer on the Ref-Youtube-VOS dataset. This finetuned model has performed better than the baseline model, thus the predicted results on the validation set of Ref-Youtube-VOS dataset can be served as pseudo ground truth object masks of validation set. We then re-finetune the baseline model on the training set and validation set with pseudo labels. This semi-supervised approach is also employed on the testing set. The improved ReferFormer ranks 2nd place in the 4th Large-scale Video Object Segmentation Challenge (CVPR2022)——Track 3: Referring Video Object Segmentation, with an overall $\mathcal{J}\&\mathcal{F}$ of 68.6% and 61.7% on `test-dev` and `test-challenge`, respectively.

## 2. Method

Our model is finetuned based on ReferFormer, and the overall pipeline of ReferFormer is illustrated in Figure 1.

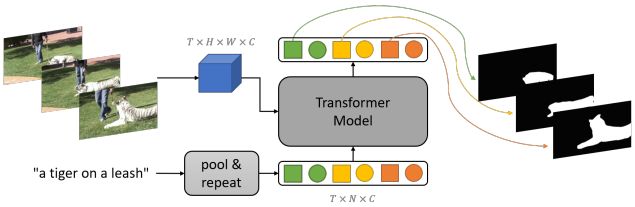


Figure 1. The overall pipeline of ReferFormer [16], where the language expression is served as conditional queries to focus on the referred object. The detailed architecture of ReferFormer can refer to [16].

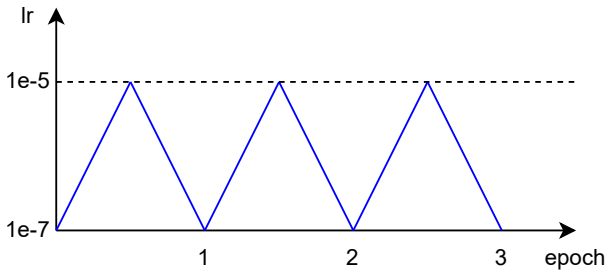


Figure 2. Cyclical learning rates.

Unlike the traditional methods, ReferFormer treats the language expression as conditional queries to focus on the referred object. And the queries are viewed as instance-aware dynamic kernels to filter out the segmentation masks. The details of ReferFormer can refer to [16]. To improve the performance of ReferFormer on Ref-Youtube-VOS dataset, we propose several tricks to finetune the model, including cyclical learning rates, semi-supervised approach and test time augmentation (TTA) inference.

### 2.1. Cyclical Learning Rates

In the original paper of ReferFormer, the model was first pretrained on the image referring segmentation datasets Ref-COCO [20], Ref-COCOg [20] and Ref-COCO+ [10]. Then, the model was finetuned on Ref-Youtube-VOS dataset. In the latest version of their released code, the model was joint trained with the image referring segmentations datasets, which achieved better results than pretraining. Since of strong performance of ReferFormer, we directly finetune the trained model on Ref-Youtube-VOS dataset. Instead of using fixed values of learning rates or monotonically decreasing the learning rates during finetuning, we use cyclical learning rates (CLR) [12] as illustrated in Figure 2. We use the triangular learning rate policy, in which the learning rate linearly increases to the maximum boundary and decreases to the minimum boundary. The length of a cycle is set to one epoch, and the minimum and maximum learning rate is set to 1e-7 and 1e-5, respectively.

### 2.2. Semi-Supervised Approach

We use a semi-supervised training method to finetune the model [17]. There are several steps to follow.

- Step 1: We use CLR to finetune the trained model of ReferFormer on Ref-Youtube-VOS dataset.
- Step 2: The finetuned model predicts object masks of validation-set, thus forming a validation-set composed of all videos frames and pseudo ground truth pairs.
- Step 3: We re-finetune the model on training-set and joint with validation-set with pseudo ground truth.
- Step 4: We re-predict the validation-set again to acquire better pseudo ground truth. The new pseudo ground truth and frames pairs can be used to complete a second-round re-finetuning.
- Step 5: The two-round finetuned model predicts testing-set to form pseudo ground truth labels and videos frames pairs. The validation-set and testing-set with pseudo ground truth labels and training-set are joint used to re-finetune the model.
- Step 6: Finally, the model predicts testing-set to get better results.

## 3. Experiments

### 3.1. Training Details

We follow the training details of original ReferFormer [1], except being specified. The baseline trained model of ReferFormer that we selected is the model with a Video-Swin-B backbone [7, 8] joint trained with Ref-COCO/+/g datasets, this model achieves the overall $\mathcal{J}\&\mathcal{F}$ of 64.9% on the validation-set. We finetune and evaluate our model with 8 A100 GPUs on Ref-Youtube-VOS dataset.

### 3.2. Components Analysis

**Test Time Augmentation Inference.** We evaluate validation-set based on the baseline trained model using test time augmentation inference, as shown in Table 1. We first evaluate using single-scale (the short side size of frames is set as 360) inference with horizontal flip, which boosts 0.6pt comparing with the baseline. Then we evaluate using multi-scale inference (the short side size of frames is set as 288,352,448,512,640), which boosts 0.9pt. Finally, we evaluate using multi-scale inference with horizontal flip to achieve the overall $\mathcal{J}\&\mathcal{F}$ of 66.6%.

**Cyclical Learning Rates.** As shown in Table 2, we finetune the baseline model using CLR for 4 epochs to achieve the best performance, which brings 1.0pt improvement using single-scale inference and 2.3pt improvement using TTA inference, respectively.

[1] https://github.com/wjn922/ReferFormer

Table 1. Experimental results of baseline ReferFormer on validation-set using test time augmentation inference.

| model | size of frames | $\mathcal{J}\&\mathcal{F}$ |
|---|---|---|
| baseline | 360 | 64.9 |
| + horizontal flip | 360 | 65.5 |
| + multi-scale inference | 288,352,448,512,640 | 65.8 |
| +horizontal flip + multi-scale inference | 288,352,448,512,640 | 66.6 |

Table 2. Experimental results of ReferFormer finetuned with cyclical learning rates.

| model | size of frames | $\mathcal{J}\&\mathcal{F}$ |
|---|---|---|
| baseline | 360 | 64.9 |
| + CLR | 360 | 65.9 |
| +CLR+TTA | 288,352,448,512,640 | 67.2 |

Table 3. Experimental results on validation-set of ReferFormer finetuned with semi-supervised approach on the first-round finetuning.

| model | size of frames | $\mathcal{J}\&\mathcal{F}$ |
|---|---|---|
| baseline | 360 | 64.9 |
| + semi-supervised | 360 | 66.8 |
| +semi-supervised+TTA | 288,352,448,512,640 | 68.0 |

Table 4. Experimental results on validation-set of ReferFormer finetuned with semi-supervised approach on the second-round finetuning.

| model | size of frames | $\mathcal{J}\&\mathcal{F}$ |
|---|---|---|
| baseline | 360 | 64.9 |
| + semi-supervised +flip | 720 | 68.3 |
| +semi-supervised+TTA | 352,512,720,896 | 68.6 |

Table 5. Experimental results on testing-set of ReferFormer finetuned with semi-supervised approach.

| model | size of frames | $\mathcal{J}\&\mathcal{F}$ |
|---|---|---|
| w/o semi-supervised + flip | 720 | 61.2 |
| + semi-supervised + flip | 720 | 61.7 |

Table 6. Testing results on CVPR2022 Referring-YouTube-VOS Challenge.

| Team | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ |
|---|---|---|---|
| Bo____ | 64.1 | 62.2 | 66.1 |
| **jiliushi (Ours)** | 61.7 | 59.8 | 63.6 |
| PENG | 60.8 | 58.9 | 62.7 |
| ds-hohhot | 59.6 | 57.9 | 61.2 |
| JQK | 59.4 | 57.7 | 61.1 |
| nero | 58.0 | 56.1 | 59.9 |

**Semi-supervised Approach.** We first predict object masks of validation-set as pseudo ground truth using the model which achieving the overall score of 67.2%. Then we re-finetune the baseline model on training-set and joint with validation-set with pseudo ground truth for 5 epochs using CLR. After that, we evaluate the re-finetuned model as shown in Table 3, which boosts 1.9pt with single-scale inference and 3.1pt with TTA comparing with the baseline, respectively. We then re-predict object masks of validation-set as new pseudo ground truth, and re-finetune again. In this second-round re-finetuning, all frames are downsampled so that the short side has the size of 608, 672 and 720 and the maximum size for long side is 1280. We re-finetune for 7 epochs using CLR and evaluate again as shown in Table 4. The performance of ReferFormer has been improved to 68.3% with single-scale (720) inference and horizontal flip, and it is been further improved to 68.6% with TTA inference. The second-round finetuned model can directly predict testing-set with single-scale (720) inference and horizontal flip and get the overall score of 61.2% as shown in Table 5. To improve the model on testing-set, the predicted object masks of testing-set can also be served as pseudo ground truth, and we re-finetune the baseline model using CLR for 7 epochs on training-set joint with validation-set and testing-set. Finally, we evaluate the last model on testing-set with single-scale (720) inference and horizontal flip, which achieves the overall $\mathcal{J}\&\mathcal{F}$ of 61.7%.

### 3.3. Challenge Results

We submit our results to CVPR2022 Referring Youtube-VOS Challenge, which ranks second place as shown in Table 6. It is worth noting that we did not use model ensemble and multi-scale inference since of time limited. And we only implement one-round re-finetuning on testing-set with pseudo ground truth. Using these tricks, the performance of RerferFormer may be improved further.

## 4. Conclusion

In this paper, we proposed several tricks to improve ReferFormer performing on Refer-YouTube-VOS dataset, including cyclical learning rate, semi-supervised approach, and test time augmentation inference. The boosted model achieves an overall $\mathcal{J}\&\mathcal{F}$ of 61.7% on testing-set, ranking second place on CVPR2022 Referring Youtube-VOS Challenge.

## References

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

[2] 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, 2020. 1

[3] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. 2021. 1

[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. 1

[5] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. 2021. 1

[6] Chen Liang, Wenguan Wang, Tianfei Zhou, Jiaxu Miao, Yawei Luo, and Yi Yang. Local-global context aware transformer for language-guided video segmentation. *arXiv preprint arXiv:2203.09773*, 2022. 1

[7] 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 *ICCV*, pages 10012–10022, 2021. 1, 2

[8] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021. 2

[9] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[10] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object description. 2

[11] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, page 208–223, 2020. 1

[12] Leslie N. Smith. Cyclical learning rates for training neural networks. In *WACV*, 2017. 2

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 1

[14] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. MaX-DeepLab: End-to-end panoptic segmentation with mask transformers. In *CVPR*, pages 5463–5474, 2021. 1

[15] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, pages 8741–8750, 2021. 1

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

[17] Jiahong Wu, Jianfei Lu, Xinxin Kang, Yiming Zhang, Yinhang Tang, Jianfei Song, Ze Huang, Shenglan Ben, Jiashui Huang, and Faen Zhang. Ainnoseg: Panoramic segmentation with high perfomance. *arXiv preprint arXiv:2007.10591*, 2020. 2

[18] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. 2021. 1

[19] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *NeurIPS*, 2021. 1

[20] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *ECCV*, 2016. 2

[21] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, pages 6881–6890, 2021. 1

[22] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 1