Rethinking Cross-modal Interaction from a Top-down Perspective for Referring Video Object Segmentation

Chen Liang[1,3], Yu Wu[3,4], Tianfei Zhou[2], Wenguan Wang[2], Zongxin Yang[3,4], Yunchao Wei[4] and Yi Yang[1]

[1]Zhejiang University   [2]ETH Zurich   [3]Baidu Research   [4]University of Technology Sydney

leonnnop@gmail.com, wenguanwang.ai@gmail.com, yangyics@zju.edu.cn

Abstract

Referring video object segmentation (RVOS) aims to segment video objects with the guidance of natural language reference. Previous methods typically tackle RVOS through directly grounding linguistic reference over the image lattice. Such bottom-up strategy fails to explore object-level cues, easily leading to inferior results. In this work, we instead put forward a two-stage, top-down RVOS solution. First, an exhaustive set of object tracklets is constructed by propagating object masks detected from several sampled frames to the entire video. Second, a Transformer-based tracklet-language grounding module is proposed, which models instance-level visual relations and cross-modal interactions simultaneously and efficiently. Our model ranks 1st place on CVPR2021 Referring Youtube-VOS challenge.

1. Introduction

Referring video object segmentation (RVOS) targets at segmenting video objects referred by given language expressions. RVOS is a challenging task as it requires not only comprehensive understanding the semantics within individual modalities, but also pixel-level cross-modal reasoning. Existing RVOS models [17, 9] typically work in a bottom-up fashion (Fig. 1-(a)), *i.e.*, perform grid-level alignment between visual and linguistic modalities. Thus they lack explicit knowledge about visual objects, leading to unreliable cross-modal reasoning and inaccurate segmentation.

In this work, we rethink RVOS from a top-down perspective (Fig. 1-(b)), by comprehensively exploring cross-object relations and conducting object-level cross-modal grounding. With a similar spirit of [12], our approach mainly consists of two stages: object tracklet generation and tracklet-language grounding. In the first stage, we generate a set of high-quality object tracklets from input videos. Then, in the second stage, we ground the reference over the detected tracklets and select the best-matched one as the final output.

More specifically, in the object tracklet generation stage, a lot of object candidate masks are first generated by apply-
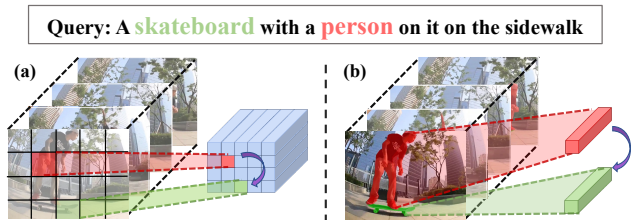

Figure 1: **An illustration of our motivation**. Previous bottom-up methods **(a)** perform cross-modal interaction at grid level, and fail to capture crucial object-level relations as top-down approach **(b)**.

ing instance segmentation over several sampled key frames. We further propagate the detected candidate masks to the whole video sequence, and generate an exhaustive set of object tracklets. After that, a tracklet-NMS mechanism is designed to remove redundant tracklets and select the high-quality ones as candidates for language-guided segmentation. In the tracklet-language grounding stage, we build a Transformer-based grounding module. Benefiting from the powerful self-attention computation within the Transformer blocks, the within-modal relations among objects and inter-modal interactions between tracklets and language can be comprehensively and efficiently modeled.

Our model ranked 1st place in 3rd Large-scale Video Object Segmentation Challenge (CVPR2021): Referring Video Object Segmentation track [1], with an overall $\mathcal{J}\&\mathcal{F}$ of 61.4% and 60.7% on `test-dev` and `test-challenge`, respectively.

2. Related Work

RVOS Datasets. The task of RVOS is proposed in [9], which mainly focuses on actor behavior understanding within an limited predefined action categories. Recently, Seo *et al.* introduced a new large-scale dataset [17], *i.e.*, Refer-Youtube-VOS (RVOS-D), derived from Youtube-VOS [24]. RVOS-D provides more complex language descriptions from a broader object categories (∼90 categories) within relatively longer video sequences (∼ 6 s). Thus it poses more challenges for RVOS methods. Referring Youtube-VOS challenge [1] is built upon RVOS-D [17].
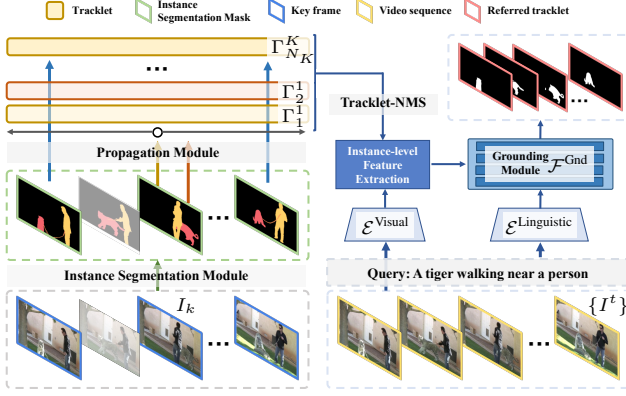
Figure 2: **Pipeline of our proposed method**, which contains two major stages, *i.e.*, object tracklet generation (left column) and tracklet-language grounding (right column).

RVOS Methods. Current studies in the filed of RVOS are made mainly around the theme of building effective multimodal feature representations. Existing methods typically make use of dynamic convolutions [21, 4] to adaptively generate convolutional filters that better respond to the referent, or leverage cross-modal attention [22, 16] to compute the correlations among input visual and linguistic embeddings. However, these methods only approach RVOS on the grid level, ignoring the importance of object-level visual cues.

3. Methodology

Overview. Given a video sequence $\mathcal{I} = \{I^t \in \mathbb{R}^{W \times H \times 3}\}_{t=1}^T$ with T frames, and corresponding referring expression $Q = \{q^l\}_{l=1}^L$ with L words, a set of referred object segmentation masks $\{S^t \in \{0, 1\}^{W \times H}\}_{t=1}^T$ is requested for RVOS. As illustrated in Fig. 2, we design a two-down approach with two stages for object tracklet generation and tracklet-language grounding, respectively. In the first stage, we construct a comprehensive set of object candidate tracklets from $\mathcal{I}$ and propose a sequence-nms module for reducing redundant candidates. In the second stage, the referred target is selected from the tracklets under the guidance of Q.

Object Tracklet Construction. We first uniformly sample K frames from $\mathcal{I}$. For each key frame I^k, a set of mask candidates, *i.e.*, $\mathcal{O}^k$, are generated through an image instance segmentation model $\mathcal{F}^{\text{Seg}}$:

$$\mathcal{O}^k = \{O_n^k\}_{n=1}^{N_k} = \mathcal{F}^{\text{Seg}}(I^k), \qquad (1)$$

where N_k refers to the number of the candidates in I^k, and, for each mask candidate $O_n^k \in \mathcal{O}^k$, we have $O_n^k \in \{0, 1\}^{W \times H}$. Then, a video mask propagation model $\mathcal{F}^{\text{Prop}}$ is applied for each $O_n^k \in \mathcal{O}^k$, to forward and backward propagate the mask to the entire video and get corresponding object tracklet Γ_n^k:

$$\Gamma_n^k = \mathcal{F}^{\text{Prop}}(O_n^k, \mathcal{I}) \in \{0, 1\}^{T \times W \times H}. \qquad (2)$$

Thus each tracklet is a sequence of masks, *i.e.*, $\Gamma_n^k = \{M^{k,t} \in \{0, 1\}^{W \times H}\}_{t=1}^T$, corresponds to the object candidate O_n^k in key frame I^k. And we define $\mathcal{T}^k = \{\Gamma_n^k\}_{n=1}^{N_k}$ as the set of all the tracklets generated from $\mathcal{O}^k$.

Based on above strategy, we generate a lot of tracklets, *i.e.*, $\{\mathcal{T}^k\}_k$, from the K key frames. This ensures that we can generate a complete object candidate set that covers object instances in $\mathcal{I}$ as many as possible, without the disturbance from object occlusion, and move-in/-out. We denote the set of all generated candidate tracklets as $\mathcal{T} = \cup_{k \in \{1 \cdots K\}} \mathcal{T}^k$.

Tracklet-NMS. As we sample several key frames, there exist a lot of similar tracklets that correspond to the same object instance. This would bring an extra challenge to the following tracklet-language grounding process. Inspired by [13], we introduce a tracklet-level NMS process that eliminates redundant candidates in $\mathcal{T}$ efficiently. We first define tracklet-IoU that measures the similarity between two tracklets, *i.e.*, $\Gamma_p, \Gamma_q \in \mathcal{T}$:

$$\text{tracklet-IoU}(\Gamma_p, \Gamma_q) = \frac{\sum_{t=1}^T |M_p^t \cap M_q^t|}{\sum_{t=1}^T |M_p^t \cup M_q^t|}, \qquad (3)$$

where $\Gamma_p = \{M_q^t\}_{t=1}^T$, and $\Gamma_q = \{M_q^t\}_{t=1}^T$. Each tracklet Γ is also assigned with a score, defined as the product of the confidence score of O_n^k (obtained from $\mathcal{F}^{\text{Seg}}$) and the mask propagation probability (obtained from $\mathcal{F}^{\text{Prop}}$), averaged over all the T frames. Based on the tracklet score and tracklet-IoU, traditional NMS algorithms [5, 6] is conducted. As at most 5 objects might be requested in our concerned experimental setting, we keep at most $P = 10$ tracklets with highest scores for each video after NMS. We refer the final tracklet set as $\hat{\mathcal{T}} = \{\Gamma_p\}_{p=1}^P$.

Tracklet-Language Grounding. We adopt per-frame reference grounding to determine the referred object from $\hat{\mathcal{T}}$. Each frame I^t and linguistic input Q are first fed into single-modality encoders, *i.e.*, separately for within-modal feature extraction:

$$\begin{aligned} \boldsymbol{I}^t &= \mathcal{E}^{\text{Visual}}(I^t) \in \mathbb{R}^{w \times h \times D}, \\ \boldsymbol{Q} &= \mathcal{E}^{\text{Linguistic}}(Q) \in \mathbb{R}^{L \times D}, \end{aligned} \qquad (4)$$

where $\boldsymbol{I}^t$ and $\boldsymbol{Q}$ are extracted visual and linguistic features, respectively.

For each tracklet $\Gamma_p \in \hat{\mathcal{T}}$, we extracted its corresponding feature at frame I^t through:

$$\boldsymbol{\tau}_p^t = \text{AvgPool}(\boldsymbol{I}^t \otimes M_p^t) \in \mathbb{R}^D, \qquad (5)$$

where $M_p^t \in \{0, 1\}^{W \times H}$ refers to the candidate mask of tracklet Γ_p in frame $\boldsymbol{I}^t$, and $\otimes$ denotes hadamard product. Note that the rescaling process for feature dimension alignment is omitted.

Given concatenated embeddings $\boldsymbol{\Gamma}^t = [\boldsymbol{\tau}_1^t, \cdots, \boldsymbol{\tau}_P^t] \in \mathbb{R}^{P \times D}$ for the candidate tracklets $\hat{\mathcal{T}} = \{\Gamma_p\}_{p=1}^P$ at

1. A person is showing his skate board skills on the road
2. A skateboard with a person on it on the sidewalk

1. A leopard is biting a neck of a injured deer   3. A deer is down and attacked by a leopard
2. A leopard with partial visibility at the left is watching another leopard
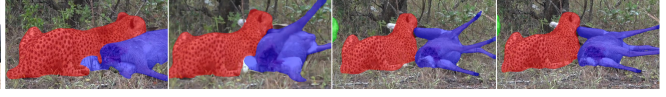
Figure 3: **Representative visual results** on RVOS-D `test-challenge` set. Each referent and the corresponding textual description are highlighted in the same color.

Team	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
leonnnop (Ours)	**61.4 (+6.6)**	**60.0 (+6.3)**	**62.7 (+6.7)**
nowherespyfly	54.8	53.7	56.0
seonguk	48.9	47.0	50.8
wangluting	48.5	47.1	49.9
Merci1	44.9	43.9	45.9

Table 1: **Benchmarking results** on the `test-dev` set of Referring Youtube-VOS challenge.

Team	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
leonnnop (Ours)	**60.7 (+11.3)**	**59.4 (+11.0)**	**62.0 (+11.7)**
nowherespyfly	49.4	48.4	50.3
feng915912132	48.2	47.4	49.0
Merci1	41.2	40.6	41.8
wangluting	40.7	39.5	41.8

Table 2: **Benchmarking results** on `test-challenge` set of Referring Youtube-VOS challenge.

Model	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
Image-level Baseline	40.9	40.5	41.3
+Video-level Propagation $\mathcal{F}^{\text{Prop}}$ (Eq. 2)	49.2	47.5	50.9
+Transformer-based Grounding $\mathcal{F}^{\text{Gnd}}$ (Eq. 6)	56.4	54.8	58.1
+Sequence-NMS (Eq. 3) & Model Ensemble	61.4	60.0	62.7

Table 3: **Ablation study** of essential components on `test-dev`.

frame I^t, and the linguistic representation $\mathbf{Q}$, we propose a Transformer-based [20] grounding module $\mathcal{F}^{\text{Gnd}}$ for tracklet-language grounding:

$$\{s_p^t\}_{p=1}^P = \mathcal{F}^{\text{Gnd}}([\mathbf{\Gamma}^t + \boldsymbol{e}^v, \mathbf{Q} + \boldsymbol{e}^l], \quad (6)$$

where $s_p^t \in [0,1]$ is the grounding score of tracklet Γ_p in I^t. Here $\boldsymbol{e}^v \in \mathbb{R}^{P \times D}$ and $\boldsymbol{e}^l \in \mathbb{R}^{L \times D}$ are learnable modal embeddings. Due to the self-attention mechanism in the Transformer, the interactions among different object tracklets and between different modalities are comprehensively captured, leading to promising grounding performance.

The final grounding score for each Γ_p is given as: $s_p = \text{Mean}(s_p^1, \cdots, s_p^T)$. The the segments $\{S^t\}_{t=1}^T$ are $\{M_{p*}^t\}_{t=1}^T$, where $p* = \arg\max_p(s_1, \cdots, s_P)$.

4. Experiment

Challenge Dataset and Evaluation Metrics. We test our model on Referring Youtube-VOS challenge [1], which is built upon the recently released RVOS-D dataset [17]. The challenge dataset has 3,978 videos with about 15K language reference sentences in total; 3,471 videos are released with annotations for training. The rest videos are split into 202/305 for constructing `test-dev`/`test-challenge` sets, whose annotations are preserved for benchmarking. We use standard metrics, *i.e.*, region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$, for evaluation.

Detailed Network Architecture. We employ two instance segmentation models, *i.e.*, HTC [2] and CondInst [19], for implementing $\mathcal{F}^{\text{Seg}}$ in Eq. 1. At each key frame I^k, the mask candidates $\mathcal{O}_k$ are a combination of all proposals generated from the two models. The mask propagation model $\mathcal{F}^{\text{Prop}}$ (Eq. 2) is implemented as CFBI+ [25]. We uniformly sample $K=7$ key frames for each video sequence. For tracklet-language grounding, we implement the visual encoder $\mathcal{E}^{\text{Visual}}$ as ResNet-101 [7] initialized from ImageNet-pretrained weights and linguistic encoder $\mathcal{E}^{\text{Linguistic}}$ as a standard BERT$_{\text{BASE}}$ model. The grounding module $\mathcal{F}^{\text{Gnd}}$ is

a 4-layer Transformer [20] with 12 heads in each layer, followed by a 2-layer MLP and a softmax layer for probability prediction. Input sentences are split by the WordPiece tokenizer [23] as in [3]. Both the hidden dimensions of Transformer and feature channel of within-modal representations are set to 768, *i.e.*, $D=768$.

Training Detail. For $\mathcal{F}^{\text{Seg}}$, HTC is trained on COCO [14] without finetuning. CondInst is pretrained on COCO and finetuned on the training split of RVOS-D with the standard training setting in [19] for about 15,000 steps. The propagation module $\mathcal{F}^{\text{Prop}}$, *i.e.*, CFBI+, is pretrained on COCO and finetuned over training split of VOS [24] track as a standard training setting in semi-supervised VOS task (see [25] for more details). For tracklet-language grounding module (Eqs. 4- 6), we pretrain it using the data from RefCOCO [26], RefCOCOg [26] and RefCOCO+ [15] for about 20 epochs. We use Adam [10] as the optimizer with a learning rate of 4e-5, batch size of 48 and weight decay of 1e-4. The module is further finetuned on the training split of RVOS-D for five epochs with a learning rate of 1e-5.

Model Ensemble. Model ensemble is also used in our final submission. We build five models with different implementations of the visual encoder $\mathcal{E}^{\text{Visual}}$, *i.e.*, ResNet101 [7], HRNet [18] and ResNeSt101 [27], and linguistic encoder $\mathcal{E}^{\text{Linguistic}}$, *i.e.*, Deberta$_{\text{BASE}}$ [8] and Bart$_{\text{BASE}}$ [11]. We use different hyperparameter settings to further promote model performance and simply average grounding probabilities from different models for final prediction.

Results on RVOS Challenge. Table 1 and Table 2 show the ranking result of top teams in `test-dev` and `test-challenge` sets respectively. Our approach achieves the best performance on both the two sets across

3

all the metrics and outperforms 2nd best team with a large margin, *i.e.*, **11.3%** in terms of overall $\mathcal{J}\&\mathcal{F}$ on `test-challenge`. Fig. 3 shows qualitative results of our proposed model on `test-challenge`. With the effective top-down model design, our approach generates robust predictions even in challenging scenes, *e.g.*, semantically similar instances, inconspicuous referent, complex linguistic description, *etc.*

Ablation Study. We start our ablation study with a simple image-level grounding pipeline (1st row in Table 3), which only contains an image instance segmentation module $\mathcal{F}^{Seg}$ (Eq. 1) for image-level object candidates generation and implements the grounding module $\mathcal{F}^{Gnd}$ (Eq. 6) as a naïve feature similarity operation without tracklet construction $\mathcal{F}^{Prop}$ (Eq. 2). Then we progressively add essential modules (2nd-4th rows in Table 3). With the fully exploration of intra- and inter-modal interactions, consistent performance improvements can be achieved.

References

[1] The 3rd large-scale video object segmentation challenge. https://youtube-vos.org/challenge/2021/. 1, 3

[2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 3

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2018. 3

[4] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *CVPR*, 2018. 2

[5] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 2

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[8] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*, 2021. 3

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

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 3

[11] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020. 3

[12] Chen Liang, Yu Wu, Yawei Luo, and Yi Yang. Clawcranenet: Leveraging object-level relation for text-based video segmentation. *arXiv preprint arXiv:2103.10702*, 2021. 1

[13] Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Jiaya Jia. Video instance segmentation with a propose-reduce paradigm. *arXiv preprint arXiv:2103.13746*, 2021. 2

[14] 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 *ECCV*, 2014. 3

[15] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 3

[16] Ke Ning, Lingxi Xie, Fei Wu, and Qi Tian. Polar relative positional encoding for video-language segmentation. In *IJCAI*, 2020. 2

[17] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, 2020. 1, 3

[18] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 3

[19] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, 2020. 3

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

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

[22] 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*, 2019. 2

[23] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. 3

[24] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, 2018. 1, 3

[25] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV*, 2020. 3

[26] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 3

[27] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020. 3