R^2-VOS: Robust Referring Video Object Segmentation via Relational Cycle Consistency

Anonymous Author(s)
Affiliation
Address
email

A Additional Experiments

In this section, we add additional experiments on a stronger backbone, ablation studies on dynamic kernel number and semantic alignment discrimination.

A.1 Stronger Backbone and Training Scheme

Method	Backbone	Refer-Youtube-VOS		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ReferFormer [8]	Video-Swin-B	64.9	62.8	67.0
Ours	Video-Swin-B	**69.5**	**67.5**	**71.4**

Table A: **Comparison to state-of-the-art R-VOS methods on Refer-Youtube-VOS val set with larger backbone and stronger training scheme.**

As shown in Table A, we report our result using a stronger Video-Swin-B [6] backbone. Notably, our method achieves 69.5 $\mathcal{J}\&\mathcal{F}$ on Refer-Youtube-VOS dataset.

A.2 Dynamic Kernel Number in Early Grounding Module

L_θ	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
1	56.4	55.1	57.7
2	57.0	55.3	58.0
3	**57.3**	**56.1**	**58.4**
4	57.1	55.9	58.2

Table B: **Impact of the dynamic filter number.**

As shown in Table B, we conduct experiments to investigate the impact of the dynamic filter number in the early grounding module. The dynamic convolution is extensively used to decode dense features in video instance segmentation [4, 5] and object detection [2] because of its strong ability to generate instance-specific filters to modify the feature maps. In our method, we use a text-guided dynamic convolution to ground referred object in the feature level. We notice that using a dynamic kernel number of 3 brings the best performance.

A.3 Semantic Alignment Discrimination

As shown in Table C, we conduct experiments without using the semantic alignment $\mathbb{1}(A)$ to filter out negative videos during inference. We notice that, even if $\mathbb{1}(A)$ is not applied to the final output,

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Method	Backbone	$\mathcal{J}\&\mathcal{F}$ & $\mathcal{R}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{R}$
ReferFormer [8]	ResNet-50	47.3	54.8	56.5	30.6
Ours	ResNet-50	**59.2**	**56.0**	**58.3**	**63.2**
MTTR [1]	Video-Swin-T	40.0	55.9	58.1	5.9
ReferFormer [8]	Video-Swin-T	49.1	58.0	60.9	28.5
Ours	Video-Swin-T	**62.7**	**59.4**	**62.9**	**65.5**

Table C: **Comparison to state-of-the-art R-VOS methods on R^2-Youtube-VOS without applying $\mathbb{1}(A)$ to filter out videos during inference.**

our model has a much higher $\mathcal{R}$ score compared to previous methods on R^2-Youtube-VOS. This indicates the consistency constraint can boost the model robustness to negative videos even without explicitly filtering out videos with semantic alignment discrimination.

B Visualization of Attentions in the Early Grounding Module

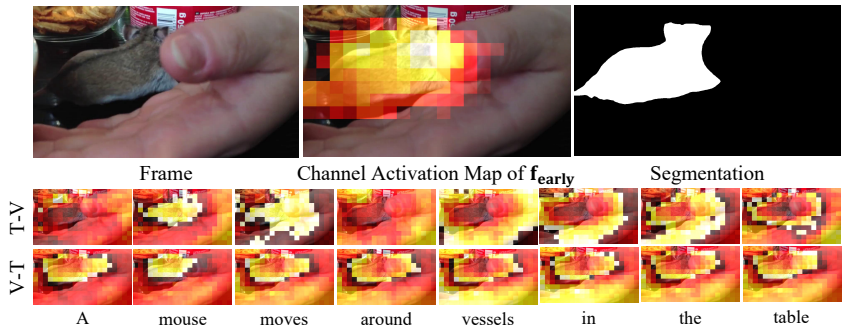

Figure A: **Visualization of cross-attention attentions and $\mathbf{f}_{\mathbf{early}}$ in the Early Grounding Module.**

C More Implementation Details

We pretrain our model on a combination of three image-level datasets, i.e., Ref-COCO, Ref-COCO+, and Ref-COCOg [10]. To be compatible with the image-level dataset, we set the window size to 1. We pretrain our model for 12 epochs, which takes about 1-2 days on 8 NVIDIA V100 32G GPUs depending on the backbones. We select the checkpoint with the best results on Ref-COCO val set as our pretrained weight for our main training.

We set the $\lambda_{text} = 0.1$, $\lambda_{cls} = 2$, $\lambda_{mask} = 2$, $\lambda_{align} = 1$, $\lambda_{angle} = 10$, $\lambda_{L1} = 5$, $\lambda_{giou} = 2$, $\lambda_{dice} = 2$ and $\lambda_{focal} = 5$ during all training process. $C_v = C_e = C_q = 256$ is utilized. The positional embedding added in the transformers is the standard triangle positional embedding used in [7]. We set the layer number to three for transformers decoders $\mathcal{D}_e$ and $\mathcal{D}_v$. The dynamic filter number K is set to 3. The data point to calculate the relational loss is selected within each batch for simplicity. The text encoder is frozen during the main training.

D Detailed Structure of Mask Decoding

As is shown in Fig. B, given the fused text embedding, we generate the instance query $\mathbf{z}_0$ by repeating the fused text embedding N times where N is the query number. After that, we generate instance embedding $\{z_t\}_{t=1}^{T}$ for each time step separately using a shared transformer decoder $\mathcal{D}_v$ with encoded memory $\{\mathbf{F}_t\}_{t=1}^{T}$ from visual encoder. The mask prediction M_t for each time step t is derived by a linear combination of $\mathbf{F}_t$ where weights are learned from instance embedding $\mathbf{z}_t$ by two fully connected layers. Note that, as positional embedding is added to the instance query $z_0 \in \mathbb{R}^{C_q \times N}$, each slot in the instance query is different.

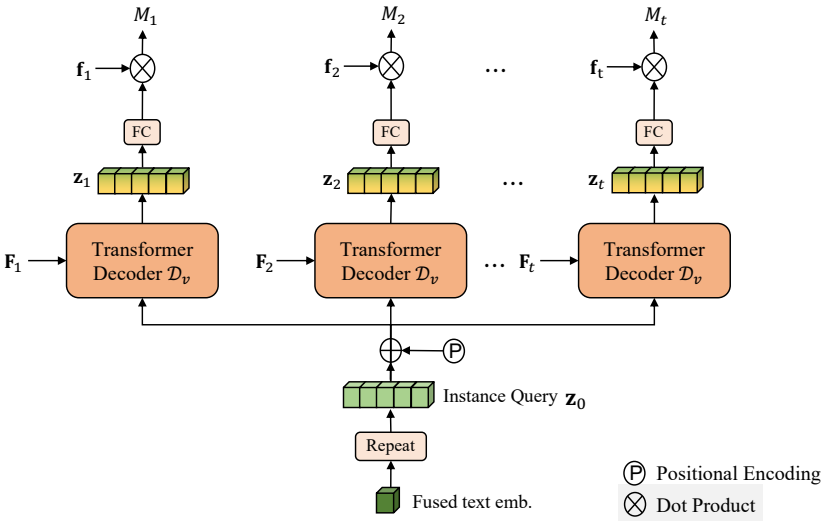

Figure B: **Illustration of mask decoding.**

Why use N instance queries for only one referred object in the video? Empirical, each slot in the instance query tends to focus on different visual features in the transformer decode $\mathcal{D}_v$ thus the N slots in the instance embedding are highly specialized. Each slot tends to represent an object with some specific properties. For example, slot 1 can always tend to predict an object located in the left of the frame. Slot 2 tends to predict objects belonging to "cat", "dog", etc., categories. By using more than one slot for the instance query, we can generate more specialized and accurate instance embedding, which is vital for mask decoding and confidence score, and box prediction.

E Broader Impact and Future Works

The false alarm problem in the RVOS task also exists in other referring prediction tasks, e.g., visual grounding [3] and referring image segmentation [9]. We consider our problem formulation that defines the negative and positive vision-language pairs can be extended to other tasks that require multi-modal semantic consensus.

References

[1] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. *arXiv preprint arXiv:2111.14821*, 2021.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[3] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. *arXiv preprint arXiv:2104.08541*, 2021.

[4] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *arXiv preprint arXiv:2106.03299*, 2021.

[5] Xiang Li, Jinglu Wang, Xiao Li, and Yan Lu. Hybrid instance-aware temporal fusion for online video instance segmentation. *arXiv preprint arXiv:2112.01695*, 2021.

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

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[8] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. *arXiv preprint arXiv:2201.00487*, 2022.

[9] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10502–10511, 2019.

[10] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.