Temporal Context Enhanced Referring Video Object Segmentation

Supplementary Material

Xiao Hu[1], Basavaraj Hampiholi[2], Heiko Neumann[2], and Jochen Lang✉[1]

[1]University of Ottawa, Canada, {xhu008, jlang}@uottawa.ca
[2]Ulm University, Germany, {basavaraj.hampiholi, heiko.neumann}@uni-ulm.de

A. Testcase Scenarios

In this supplementary material, we give further details on how we categorize testcase scenarios for the comparative evaluation of RVOS methods. We also give examples of videos and text expressions illustrating some categories along with example results for our TCE-RVOS and two comparison methods [1, 4].

A.1. Presence

In the category presence, we identify videos and text expressions where the referred instance is not in the field-of-view in some of the frames. This scenarios is due to the relative motion between the camera and the referred instance. Figure 1 shows an example of the scenario when the referred instance is not present in some frames of the video. The referred instance "white toilet" is not in the camera field-of-view in the first several frames, and appears only later due to the camera motion. TCE-RVOS (the bottom sequence in Figure 1) outperforms other methods by not only detecting that the referred instance is not present in the first two frames shown but also by generating more accurate masks in the remaining frames. The presence category is semantically different from occlusion as an occluded object is in the field-of-view of the camera but the view is (partially) obstructed by other foreground objects.

A.2. Object Motion

We use a simple yet effective script to classify object motion by evaluating the change in the bounding box location and size of the referred instance between frames. The pseudocode is shown in Algorithm 1.

Since Ref-Youtube-RVOS dataset [3] does not provide the ground truth for the validation set, we conduct experiments on the A2D dataset [2] with two different thresholds for the bounding box centre τ_c and size change τ_s. We use

Algorithm 1: Instance Motion Classification Script

Result: Return the classified motion status for each instance.
τ_c, τ_s: the pre-defined threshold for the bounding box center change and size change.;
w_i, h_i, c_i: the width, height, and center of the instance bounding box in the frame i;
$\hat{w} = w_i\text{-}w_{i-1}$, $\hat{h} = h_i\text{-}h_{i-1}$, $\hat{c}$ = distance(c_i,c_{i-1});
if $\hat{c} \geq \tau_c$ **then**
  The instance is in fast motion;
else
  if $\hat{w} \geq \tau_s$ *or* $\hat{h} \geq \tau_s$ **then**
    The instance is in slow motion;
  else
    The instance is in not in motion;
  end
end

$\tau_c = \tau_s = 25$ and 50 pixels, respectively. A comparison of the results is shown in Table 1. The results show that the improvement of TCE-RVOS over ReferFormer is mainly coming from the challenging scenarios. The incremental improvement for the slow and fast motion classes are higher than for the no motion class. A threshold of $\tau_c = \tau_s = 25$ makes this clearer in the case of the A2D dataset.

A.3. Interaction

In some of the video samples, the text expression describes the instance by its relationship with other objects in the scene. This scenario increases the difficulty of inference, since the model not only needs to detect multiple instances, but also model the relationship between them. Figure 2 shows an example of an interaction scenario. The target instance is referred by "an adult seal to the left of

Figure 1. Comparison between qualitative result of MTTR [1] (top sequence), ReferFormer [4] (middle sequence), and TCE-RVOS (bottom sequence) from the Ref-Youtube-RVOS dataset [3]. The expression is "the white toilet is between the white tub and green cabinet", and results are shown in purple masks. The example shows a partial presence scenario.

	Motion		
	No	Slow	Fast
$\tau_c = \tau_s = 25$ *pixels*			
Samples	220	264	811
ReferFormer	47.2	59.3	55.1
TCE-RVOS	47.2 (+0.0)	60.8 (+1.5)	56.0 (+0.9)
$\tau_c = \tau_s = 50$ *pixels*			
Samples	473	350	472
ReferFormer	51.3	60.6	54.0
TCE-RVOS	52.0 (+0.7)	61.8 (+1.2)	55.0 (+1.0)

Table 1. Effect of different thresholds for motion classification (see Algorithm 1). Results are with the VSwin-Base backbone on the A2D validation dataset [2].

another adult seal". The expression "left" is related to another seal on the right, and the expression "adult" is related to the baby seal in the front. As shown in Figure 2, all three compared methods understand the expression "left" and hence generate masks on one of the two seals on the left. MTTR (top sequence in Figure 2) cannot utilize the expression "adult", and generates the mask on the baby seal in the front. ReferFormer (middle sequence in Figure 2) is able to model the interaction between the different instances. However, the motion and occlusion of the referred instance leads to poor quality of the mask prediction. The result shows that TCE-RVOS not only better models the relative interaction between instances, but also generates more accurate masks compared with the two competitors.

A.4. Ambiguity

The Ref-Youtube-RVOS dataset contains various complicated scenes in which the object is hard to describe with a single text expression. Figure 3 shows a good example. The supplied text expression is "a black bird flying among other birds to the left". This text expression is not able to uniquely identify a specific bird from all the birds in the video clip. Multiple instances satisfy the expression in the video clip. In Figure 3 only the original frames from the Ref-Youtube-RVOS dataset are shown because the ground truth is not provided.

A.5. Extra Examples

Two more comparison results are provided to show the improvements on challenging scenarios. Figure 4 shows a full occlusion scenario. The referred car is fully occluded by another foreground car at the beginning of the video. TCE-RVOS outperforms the competitors by not only inferring the desired object, but also by generating more precise masks. Figure 5 represents a scene with crowded objects. The expression refers to a person, and there are multiple people in the video. By enhancing the temporal context communication, TCE-RVOS is able to generate a stable sequence of masks on the same object through the video clip, while the competitor methods are influenced by other objects (e.g., other non-referred people in Figure 5) in the scene.

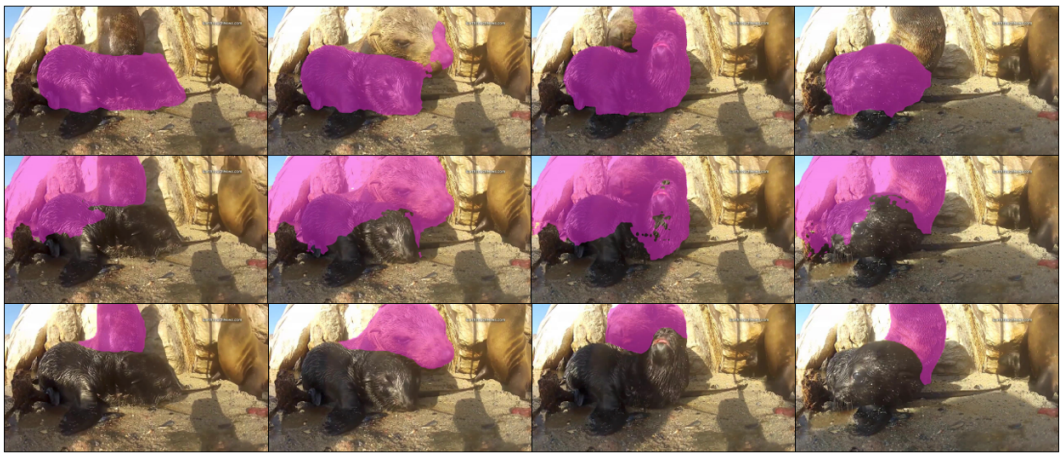

Figure 2. Comparison between qualitative result of MTTR (top sequence), ReferFormer (middle sequence), and TCE-RVOS (bottom sequence) from Ref-Youtube-RVOS dataset [3]. The expression is "an adult seal to the left of another adult seal", and results are shown in purple masks. The example shows an interaction scenario.

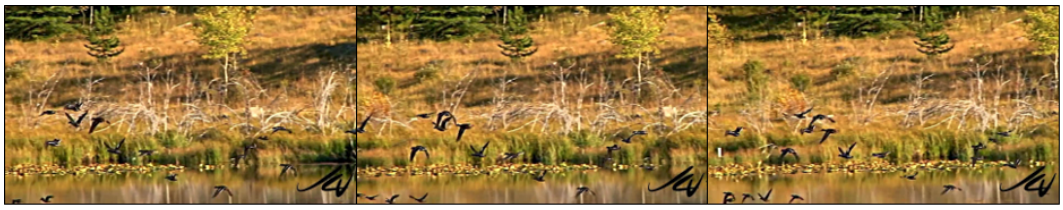

Figure 3. An example frame sequence with the expression "a black bird flying among other birds to the left". The example shows a scenario with an ambiguous text expression.

References

[1] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4985–4995, 2022. 1, 2

[2] 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. 1, 2

[3] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020. 1, 2, 3, 4

[4] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 1, 2

Figure 4. Comparison between qualitative result of MTTR (top sequence), ReferFormer (middle sequence), and TCE-RVOS (bottom sequence) from Ref-Youtube-RVOS dataset [3]. The expression is ”a white car on the left of another”, and results are shown in purple masks. The example shows a full occlusion scenario, since the referred car is fully occluded by another car in the front in the first frame.

Figure 5. Comparison between qualitative result of MTTR (top sequence), ReferFormer (middle sequence), and TCE-RVOS (bottom sequence) from Ref-Youtube-RVOS dataset [3]. The expression is ”a person on the left side of the road wearing a red shirt and grey pants”, and results are shown in purple masks. The example shows a scenario with a crowded scene, since there are multiple people in the video clip.