OnlineRefer: A Simple Online Baseline for Referring Video Object Segmentation

Supplementary Material

Dongming Wu\textsuperscript{1}, Tiancai Wang\textsuperscript{2}, Yuang Zhang\textsuperscript{3}, Xiangyu Zhang\textsuperscript{2,4}, Jianbing Shen\textsuperscript{5}†
\textsuperscript{1} Beijing Institute of Technology, \textsuperscript{2} MEGVII Technology, \textsuperscript{3} Shanghai Jiao Tong University, \textsuperscript{4} Beijing Academy of Artificial Intelligence, \textsuperscript{5} SKL-IOTSC, CIS, University of Macau

Figure 1: Pipeline of semi-online model of OnlineRefer, which associates the same referent object across different clips.

1. Semi-Online Model

We present more semi-online framework details of OnlineRefer in Fig. 1. Unlike the online model that follows a frame-by-frame pattern, the semi-online model propagates the target query across clips. Note that we use sharing query for multi-frame referring segmentation within each clip.

2. Additional Experiment Details

Refer-Youtube-VOS provides full-video expression by describing an entire video and first-frame expression based on the first frame, while we only use their full-video expression for training and validation.

Refer-DAVIS\textsubscript{17} also contains the full-video and first-frame expressions, which are developed by four annotators. Our final $J$ & $F$ scores are averaged from the four results.

3. Additional Ablation Study

We provide a thorough analysis of sampling length settings in Table 1. It is obvious that increasing frames from 2 to 3 brings performance improvement on ResNet-50, while it fails on Swin-L. When using the progressive sampling strategy (i.e., [2, 3]), the best results can be obtained on both two backbones. This indicates that the appropriate increment on sampling lengths is beneficial for guaranteeing model stability and improving model performance.

<table>
<thead>
<tr>
<th>Sampling Lengths</th>
<th>ResNet-50</th>
<th>Swin-L</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$J &amp; F$</td>
<td>$J$</td>
</tr>
<tr>
<td>[2]</td>
<td>56.5</td>
<td>55.1</td>
</tr>
<tr>
<td>[3]</td>
<td>56.7</td>
<td>55.2</td>
</tr>
<tr>
<td>[2, 3]</td>
<td>57.3</td>
<td>55.8</td>
</tr>
<tr>
<td>[2, 3, 4]</td>
<td>57.0</td>
<td>55.5</td>
</tr>
</tbody>
</table>

Table 1: The effect of sampling lengths on Refer-Youtube-VOS. ‘/’ means no results due to model divergence.

4. More Qualitative Results

Fig. 2 offers some qualitative comparison between the offline method ReferFormer [1] and our OnlineRefer. We can see that OnlineRefer performs better than ReferFormer under the situations of object occlusion and visually-similar background, approving the superiority of query propagation. Fig. 3 also shows OnlineRefer can deal well with other challenging situations, like size and appearance variation, small or missing objects, moving objects, etc.
Figure 2: Comparison between ReferFormer and OnlineRefer on Refer-Youtube-VOS.
Expression: a man behind another man in a harness

Expression: a large train racing down the tracks

Expression: a tennis racket on the left being held by a person

Expression: a small fox like dog on the right side of a sofa

Expression: the second giraffe from the right

Expression: a person on the far side of a tennis court serving a tennis ball

Expression: the palm of a person carrying a mouse

Expression: a black cat in the middle of the view

Figure 3: More qualitative results of our OnlineRefer on Refer-Youtube-VOS.

References