Revisiting Sequence-to-Sequence Video Object Segmentation with Multi-Task Loss and Skip-Memory

Fatemeh Azimi, Benjamin Bischke, Sebastian Palacio, Federico Raue, Jörn Hees, Andreas Dengel
Video Object Segmentation (VOS)

- Track and Segment a target object given the first object mask

- Challenges:
  - Fast motion and motion blur
  - Tracking objects with various sizes
  - Occlusion
  - Error-propagation, appearance change
Different Approaches for VOS

- **Recurrent Neural Networks**
  - Learns the spatio-temporal model of the target object

- **Correspondence Matching**
  - Detects the target object via template matching with a reference frame
  - Further extensions possible with using external memory
RNN-based Solution

- Utilize the Recurrent Neural Network (RNN) to memorize the target object
- We observed this baseline struggles with tracking small objects
RNN-based Solution

- Utilize the Recurrent Neural Network (RNN) to memorize the target object
- We observed this baseline struggles with tracking small objects

Memory Augmented Skip Connections

- Skip connections
  - Recover the fine details
- Skip-Memory
  - Recover and track the fine details

Distance Loss

- Binary Cross-Entropy Loss
  - Is the pixel in the foreground object or the background?

- Distance Loss
  - What is the border class for the pixel with respect to the object boundary?
  - Utilize fine-grained location information of the pixels in the loss function

Final Architecture
Results on YouTube-VOS dataset

- More than 5pp improvement in F_score and 3.3pp in the overall segmentation accuracy
- Better segmentation of the small objects

<table>
<thead>
<tr>
<th>method</th>
<th>F_score</th>
<th>J_score</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2S[1]</td>
<td>57.9</td>
<td>57.45</td>
<td>57.68</td>
</tr>
<tr>
<td>S2S++(ours)</td>
<td>63.23</td>
<td>58.79</td>
<td>61.00</td>
</tr>
</tbody>
</table>

Visual Samples
Visual Samples
Thank You!