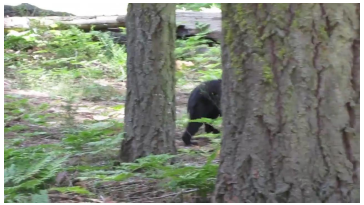
Revisiting Sequence-to-Sequence Video Object Segmentation with Multi-Task Loss and Skip-Memory

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

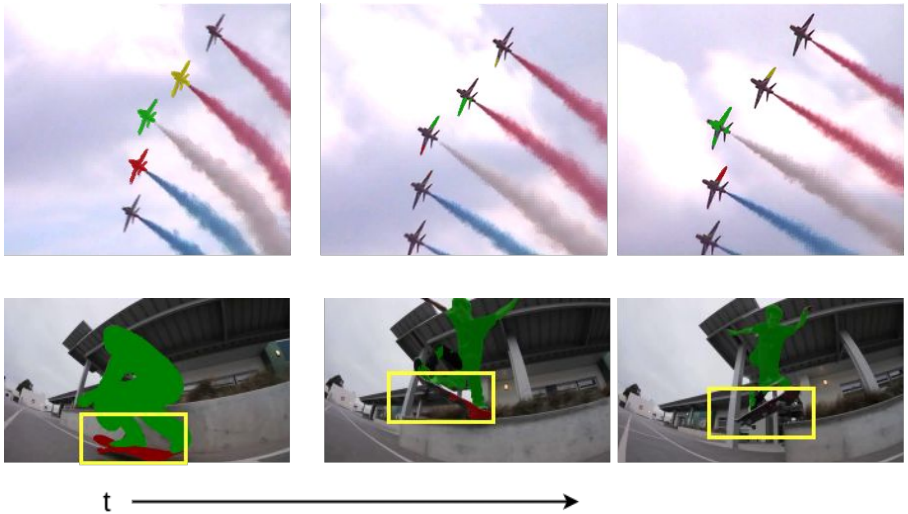
Different Approaches for VOS

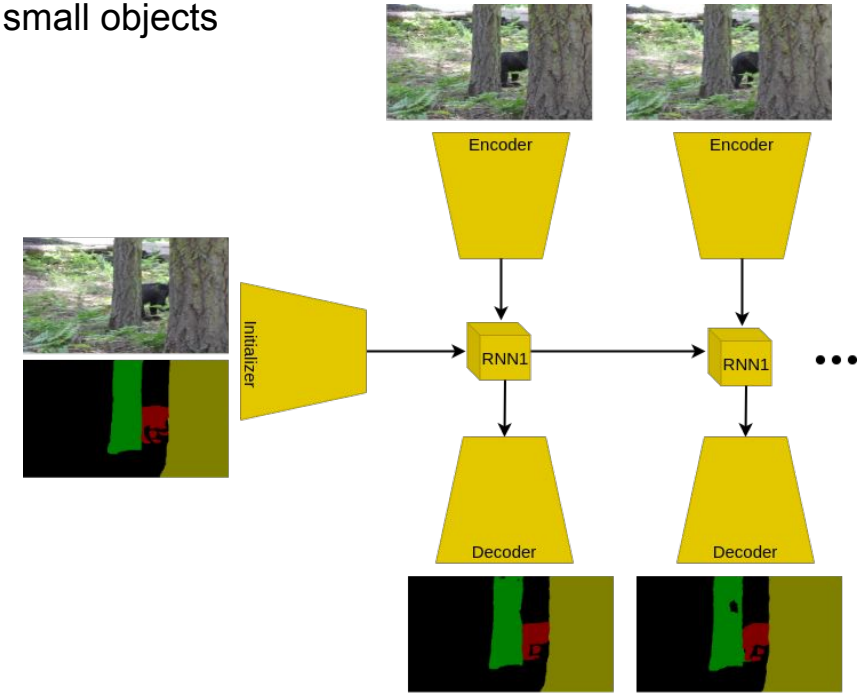
- Recurrent Neural Networks
  - Learns the spatio-temporal model of the target object

- 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

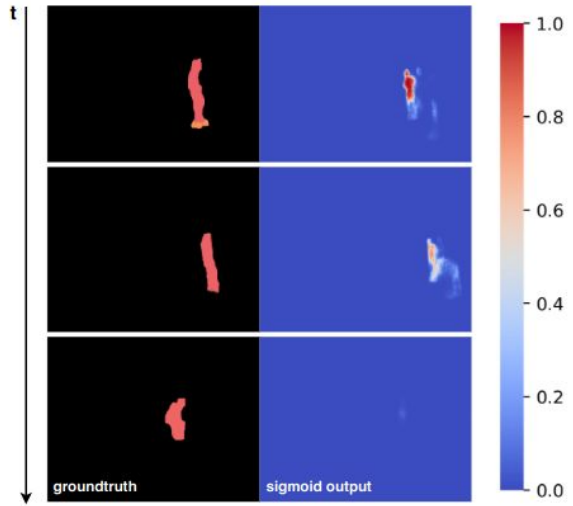


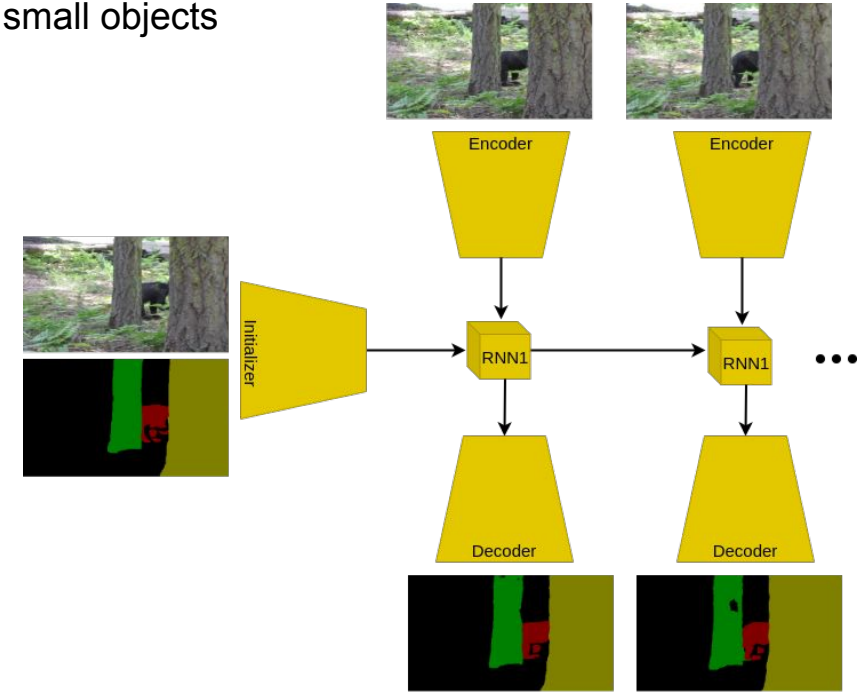


Xu, Ning, et al. "Youtube-vos: A large-scale video object segmentation benchmark." *arXiv preprint arXiv:1809.03327* (2018).

RNN-based Solution

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

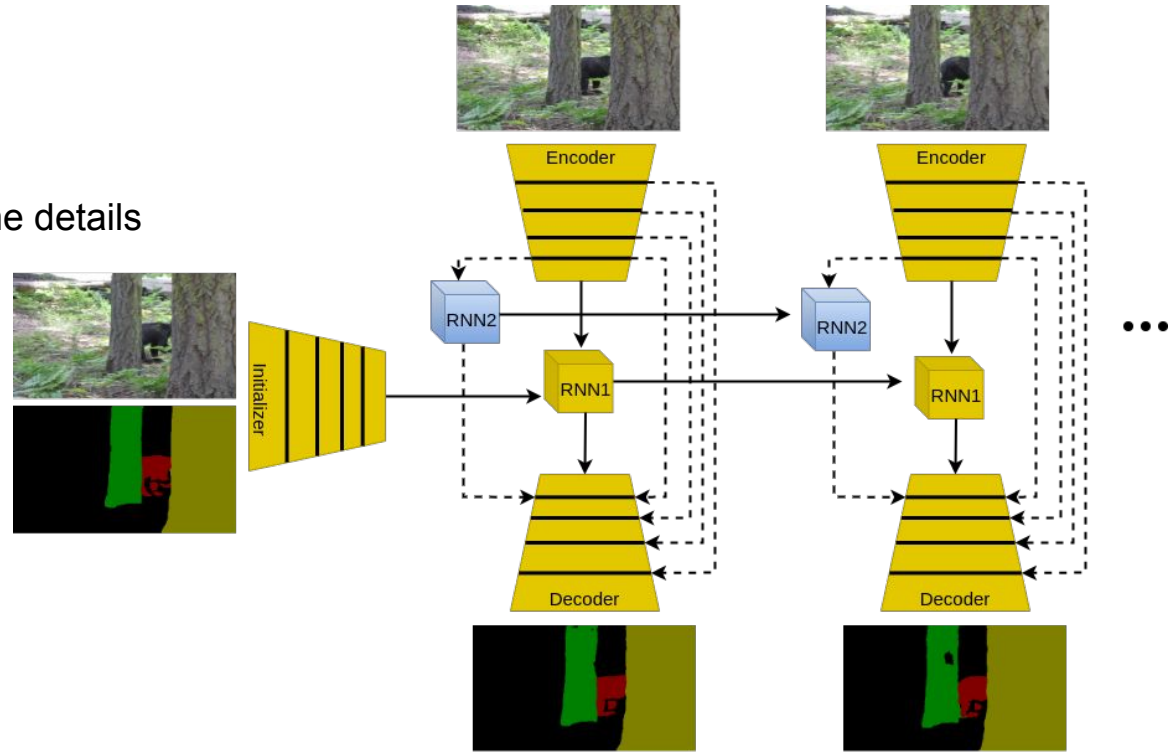




Xu, Ning, et al. "Youtube-vos: A large-scale video object segmentation benchmark." *arXiv preprint arXiv:1809.03327* (2018).

Memory Augmented Skip Connections

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



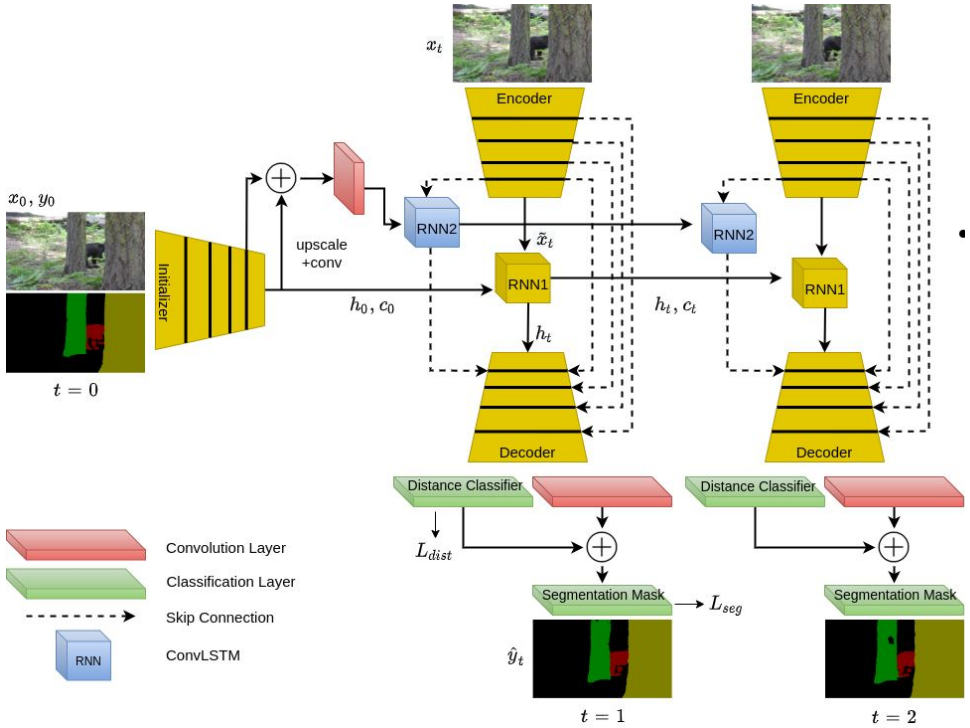
Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

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

Bischke, Benjamin, et al. "Multi-task learning for segmentation of building footprints with deep neural networks." *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019.

Final Architecture



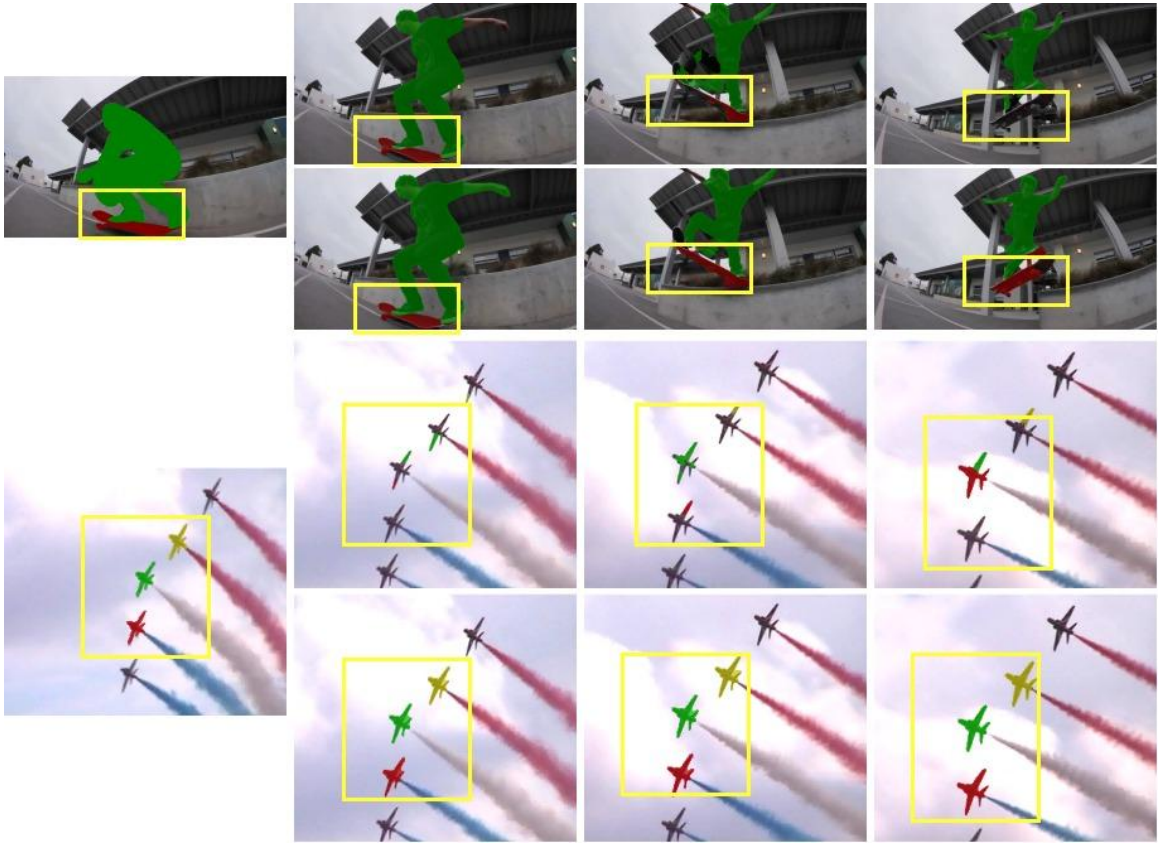
Results on YouTube-VOS dataset

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

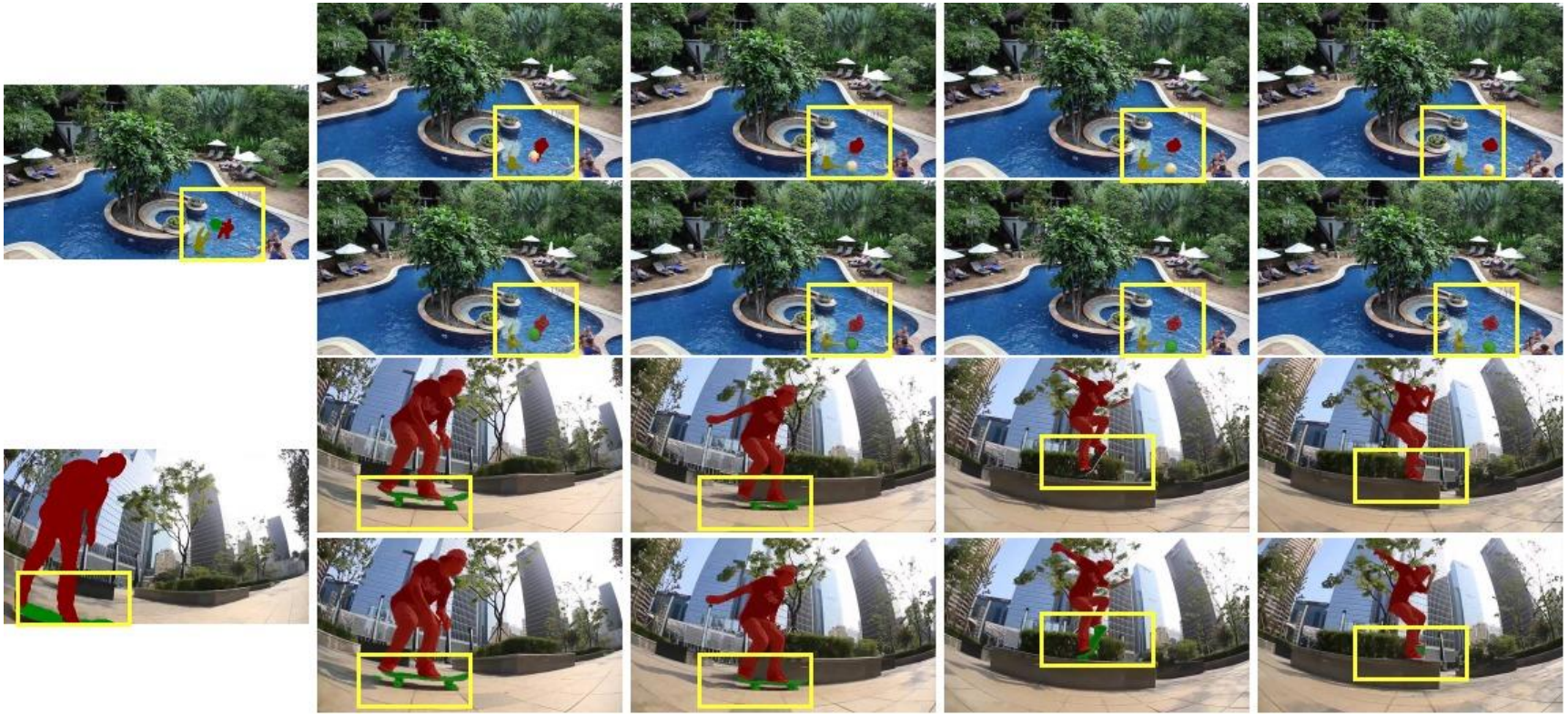
method	F_score	J_score	overall
S2S[1]	57.9	57.45	57.68
S2S++(ours)	**63.23**	**58.79**	**61.00**

Azimi, F.*, Bischke, B.*, Palacio, S., Raue, F., Hees, J. and Dengel, A., 2020. Revisiting Sequence-to-Sequence Video Object Segmentation with Multi-Task Loss and Skip-Memory. ICPR2020
[1] Xu, Ning, et al. "Youtube-vos: A large-scale video object segmentation benchmark." *arXiv preprint arXiv:1809.03327* (2018).

Visual Samples



Visual Samples



Thank You!