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

Fatemeh Azimi* [1,2], Benjamin Bischke* [1,2], Sebastian Palacio[1,2], Federico Raue[2], Joern Hees[2], Andreas Dengel[1,2]

[1]TU Kaiserslautern, [2]DFKI

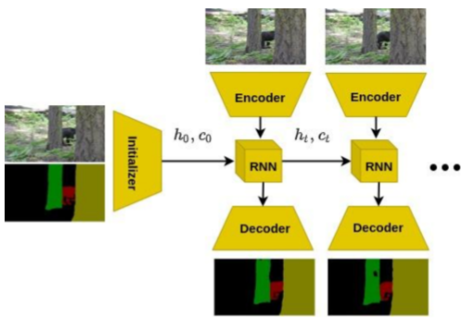
Introduction

One-shot Video Object Segmentation (VOS) aims to segment an object of interest in a video sequence, where the object mask in the first frame is provided. Amongst the different approaches proposed in the literature for solving VOS, we studied an RNN-based approach [6] due to their effectiveness in utilizing the Spatio-temporal information without requiring an additional external memory component. Our main contributions in this work are as follows.

- Identifying a limitation of this model in tracking smaller objects and addressing this with introducing the skip-memory connections
- Incorporating an auxiliary task, namely border distance classification which considerably improves the segmentation quality by providing fine-grained localization information to the model

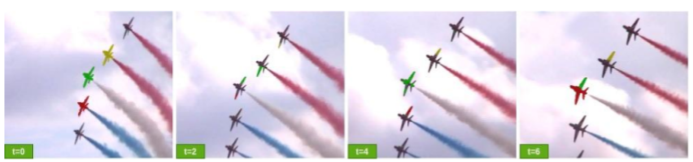
Baseline

- S2S [6] has an encoder-decoder architecture (VGG as backbone) with an RNN module in the bottleneck to memorize the target object
- An initializer network is used to process the first frame and segmentation mask and generate the initial hidden states of the RNN
- The training objective is binary cross-entropy loss



Incorporating Skip-Memory connections

We observed that the model has a much lower performance for smaller objects!

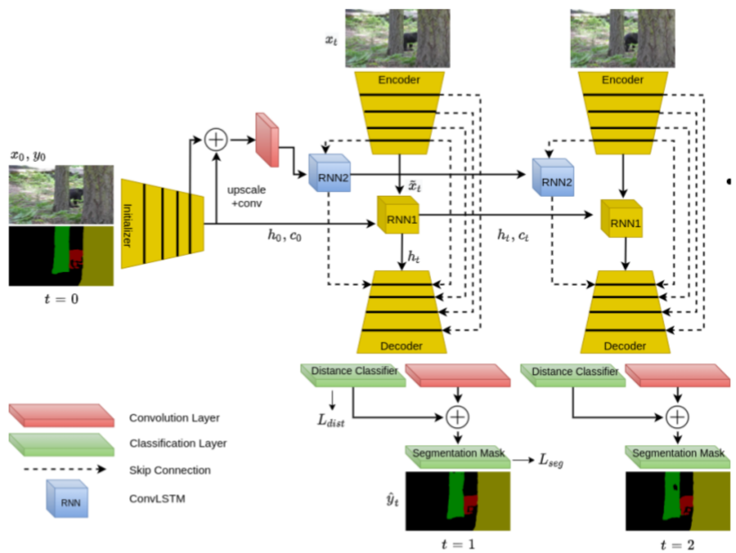


- We believe this issue is caused by losing the spatial information of the small object at the bottleneck, due to multiple pooling operations
- Inspired by skip-connections proposed for recovering the fine details in case od image segmentation, we propose an RNN-augmented connection called skip-memory
- Skip-memory connections enable the model to recover and track the fine details (small objects) in the scene

Border Distance Mask as Auxiliary Objective

- The standard loss used in the baseline is BCE, which provides only coarse localization information to the model (a pixel belongs to background or foreground)
- We utilize an additional objective of border classification for VOS
- By applying a distance transform to the mask, a border class is assigned to each pixel
- Via the auxiliary task of border classification, fine-grained localization information is provided to the model, resulting in improved segmentation accuracy

Final Architecture



Ablation Study

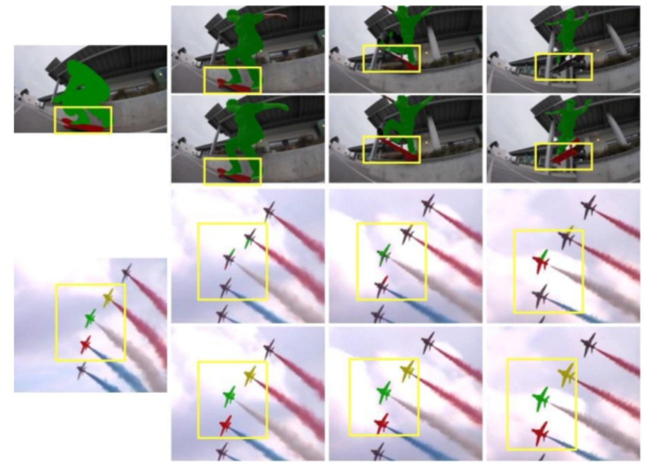
Method	J_{seen}	J_{unseen}	F_{seen}	F_{unseen}	overall
base model	65.36	43.55	67.90	47.50	56.08
base model + multi-task loss	67.65	44.62	70.81	49.84	58.23
base model + one skip-memory	66.89	46.82	69.22	50.08	58.25
base model + one skip-memory + multi-task loss	67.18	47.04	70.24	52.30	59.19
base model + two skip-memory + multi-task loss	**68.68**	**48.89**	**72.03**	**54.42**	**61.00**

The impact of different components of our model in the final performance

Comparison with SotA and Visual Samples

Method	Online training	J_{seen}	J_{unseen}	F_{seen}	F_{unseen}	overall
OSVOS [1]	yes	59.8	**54.2**	60.5	**60.7**	**58.08**
MaskTrack [2]	yes	59.9	45.0	59.5	47.9	53.08
OnAVOS [3]	yes	**60.1**	46.6	**62.7**	51.4	55.20
OSMN [4]	No	60.0	40.6	60.1	44.0	51.18
RVOS [5]	No	63.6	45.5	67.2	51.0	56.83
S2S [6]	No	66.7	48.2	65.5	50.3	57.68
S2S++(ours)	No	**68.68**	**48.89**	**72.03**	**54.42**	**61.00**

Comparison of our method with other state of the art models on YouTube-VOS dataset



Qualitative comparison between samples from S2S model in the top row and our S2S++ model in the bottom row

Conclusion

- In this work, we studied a limitation of S2S [6] model for tracking smaller objects and addressed this challenge with introducing skip-memory connections, a memory-augmented skip connection that enables the model to track the target at multiple scales.
- Furthermore, we incorporated the auxillary task of border distance classification which improves the segmentation quality by providing .
- We achieve considerable improvement in segmentation accuracy with only minimal changes to the S2S architecture.
- Our model does not rely on any form of external memory. This is advantageous since using external memory results in additional constraints in the inference phase.

References

[1] Maninis et.al. Video object segmentation without temporal information, TPAMI, 2018.
[2] Perazzi et. al. Learning video object segmentation from static images, CVPR 2017.
[3] Voigtlaender et. al. Online adaptation of convolutional neural networks for video object segmentation, BMVC 2017.
[4] Yang et. al. OSMN: one-shot modulation network for semi-supervised video segmentation, CVPR 2018.
[5] Ventura et. al.. Rvos: End-to-end recurrent network for video object segmentation. CVPR, 2019.
[6] Xu et. al. Youtube-vos:Sequence-to-Sequence video object segmentation, ECCV 2018.
[7] Ronneberger e. al. UNet: convolutional networks for biomedical image segmentation, MICCAI 2015