RVOS: End-to-End Recurrent Network for Video Object Segmentation


Carles Ventura


Andreu Girbau


Ferran Marques


Miriam Bellver


Amaia Salvador


Xavier Giro

Universitat Oberta de Catalunya
Barcelona Supercomputing Center — Centro Nacional de Supercomputación
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Acknowledgements: Generalitat de Catalunya Departament d'Empresa i Coneixement; Gobierno de España Ministerio de Economía y Competitividad; NVIDIA.

https://imatge-upc.github.io/rvos/

Motivation

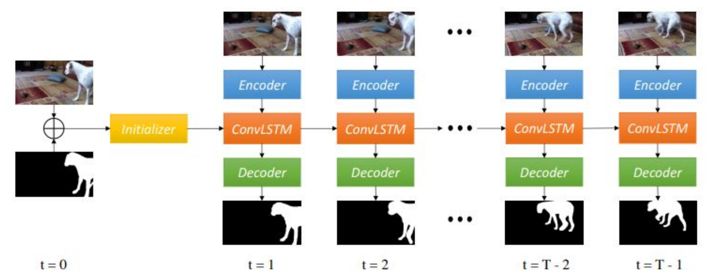
Multiple object video object segmentation is a challenging task, specially for the zero-shot case, when no object mask is given at the initial frame and the model has to find the objects to be segmented along the sequence. In our work, we propose a Recurrent network for multiple object Video Object Segmentation (RVOS) that is **fully end-to-end trainable**. Our model incorporates **recurrence** on two different domains: (i) the **spatial,** which allows to discover the different object instances within a frame, and (ii) the **temporal**, which allows to keep the coherence of the segmented objects along time. The contributions of our work are the following:

- First **end-to-end architecture** for video object segmentation that tackles multi-object segmentation without requiring any post-processing.
- The proposed model can easily be adapted to **one-shot** (or semi-supervised) and **zero-shot** (or unsupervised) video object segmentation problems.
- Our results for zero-shot video object segmentation have become the **baseline** for the new emerging DAVIS 2019 unsupervised challenge.
- We outperform previous VOS methods which do **not use online learning**. Our model achieves a remarkable performance without needing finetuning in inference, becoming the **fastest method**.

Related Work

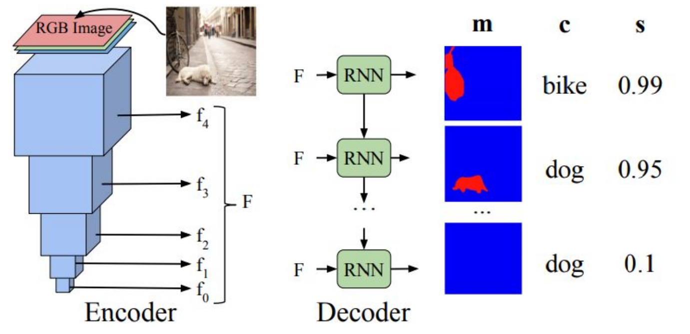
Sequence-to-Sequence Video Object Segmentation (S2S)

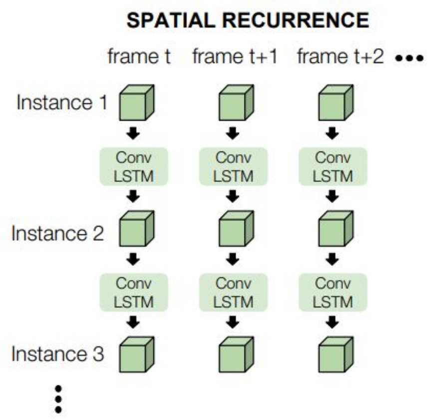
- Each instance is trained and segmented independently
- Designed only for one-shot video object segmentation



Recurrent Semantic Instance Segmentation (RSIS)

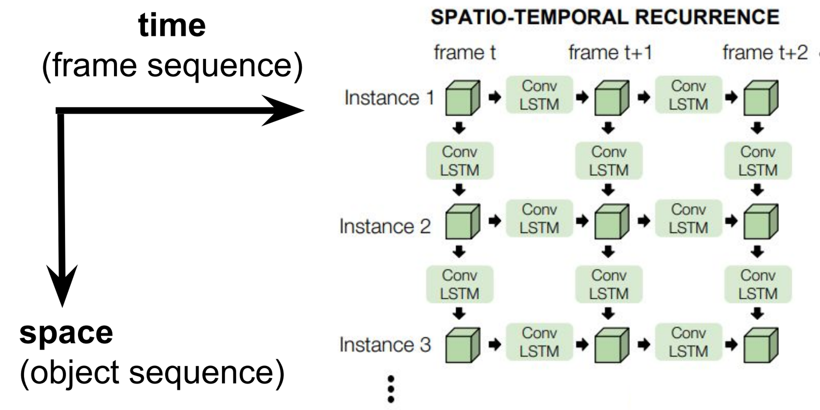
- Model based on spatial recurrence for instance segmentation in images
- If applied to videos, no temporal coherence is guaranteed along frames

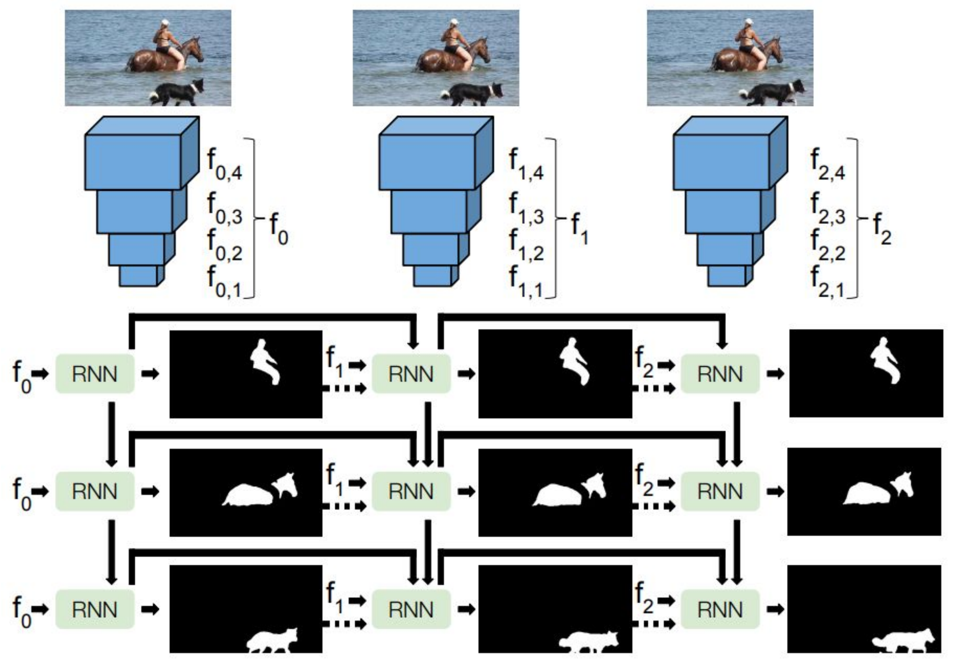




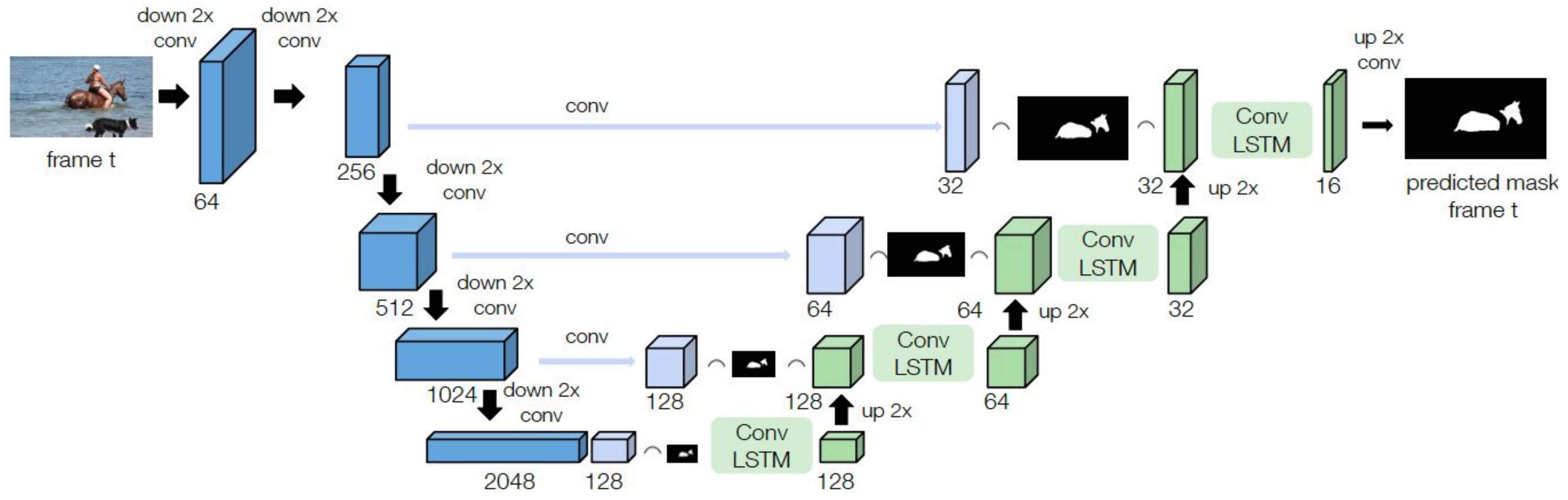
Proposed Model

We propose to extend RSIS (spatial recurrent model for images) to RVOS (spatio-temporal recurrent model for videos):

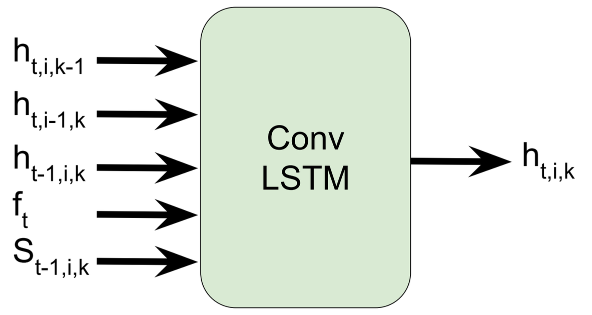




The details of the architecture for both the encoder and the decoder are shown in the following figure. While we need a forward pass of the decoder for each object instance, only a single forward pass of the encoder is required for the whole image.



In the next figure, we show the different input dependences of the *k*-th ConvLSTM layer for object *i* at frame *t*:



$h_{t,i,k-1}$: output from (*k-1*)-th ConvLSTM for object *i* at frame *t*
$h_{t,i-1,k}$: output from *k*-th ConvLSTM for the previous object (*i-1*) at same frame (*t*)
$h_{t-1,i,k}$: output from *k*-th ConvLSTM for the same object (*i*) at previous frame (*t-1*)
f_t: image features from frame *t*
$S_{t-1,i,k}$: mask prediction for the same object (*i*) at previous frame (*t-1*)

Video Object Segmentation Tasks

ONE-SHOT (SEMI-SUPERVISED) VIDEO OBJECT SEGMENTATION

ZERO-SHOT (UNSUPERVISED) VIDEO OBJECT SEGMENTATION

Experimental Results

ABLATION STUDY

S: spatial
T: temporal
ST: spatio-temporal
ST+: spatio-temporal with two training stages

1. Using previous ground truth mask
2. Using previous inferred mask

	YouTube-VOS one-shot J_{seen}	J_{unseen}	F_{seen}	F_{unseen}
RVOS-Mask-S	54.7	37.3	57.4	42.4
RVOS-Mask-T	59.9	39.2	63.1	45.6
RVOS-Mask-ST	60.8	**44.6**	63.7	50.3
RVOS-Mask-ST+	**63.1**	44.5	**67.1**	**50.4**

- The spatio-temporal model outperforms both only spatial and only temporal models
- The two training stages using the previous inferred mask outperforms the model trained using only the previous ground truth mask

RUNTIME ANALYSIS

OL: Online Learning

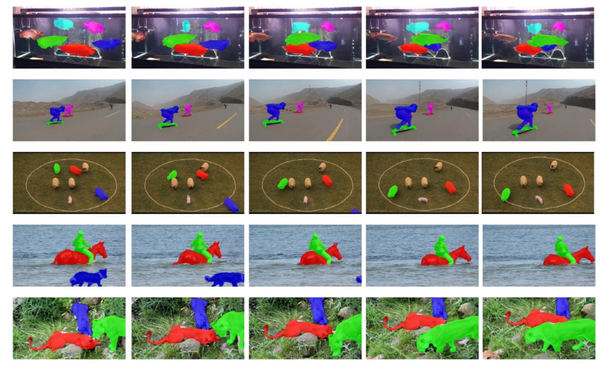
- RVOS does not use OL
- RVOS is the fastest method

	OL	YouTube-VOS one-shot J_{seen}	J_{unseen}	F_{seen}	F_{unseen}	Inference time (s/frame)
OSVOS [3]	✓	59.8	**54.2**	60.5	**60.7**	10
MaskTrack [20]	✓	59.9	45.0	59.5	47.9	12
OnAVOS [30]	✓	**60.1**	46.6	**62.7**	51.4	13
S2S w/o OL [33]	✗	**66.7**	**48.2**	65.5	50.3	0.160
OSMN [34]	✗	60.0	40.6	60.1	44.0	0.065
RVOS-Mask-ST+	✗	63.6	45.5	**67.2**	**51.0**	**0.044**

ONE-SHOT VIDEO OBJECT SEGMENTATION

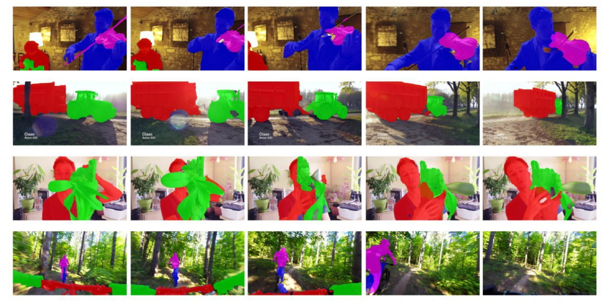
YOUTUBE-VOS

	OL	YouTube-VOS one-shot J_{seen}	J_{unseen}	F_{seen}	F_{unseen}
OSVOS [3]	✓	59.8	**54.2**	60.5	**60.7**
MaskTrack [20]	✓	59.9	45.0	59.5	47.9
OnAVOS [30]	✓	**60.1**	46.6	**62.7**	51.4
OSMN [34]	✗	60.0	40.6	60.1	44.0
S2S w/o OL [33]	✗	**66.7**	**48.2**	65.5	50.3
RVOS-Mask-ST+	✗	63.6	45.5	**67.2**	**51.0**



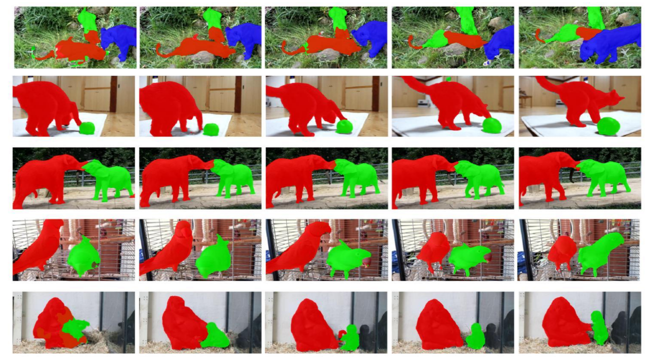
DAVIS 2017

	OL	DAVIS-2017 one-shot J	F
OSVOS [3]	✓	47.0	54.8
OnAVOS [30]	✓	49.9	55.7
OSVOS-S [17]	✓	52.9	62.1
CINM [2]	✓	**64.5**	**70.5**
OSMN [34]	✗	37.7	44.9
FAVOS [4]	✗	42.9	44.2
RVOS-Mask-ST+ (pre)	✗	46.4	50.6
RVOS-Mask-ST+ (ft)	✗	**48.0**	**52.6**



ZERO-SHOT VIDEO OBJECT SEGMENTATION

	YouTube-VOS zero-shot J_{seen}	J_{unseen}	F_{seen}	F_{unseen}
RVOS-S	40.8	19.9	43.9	23.2
RVOS-T	37.1	20.2	38.7	21.6
RVOS-ST	**44.7**	**21.2**	**45.0**	**23.9**



	DAVIS-2017 zero-shot J	F
RVOS-ST (pre)	21.7	27.3
RVOS-ST (ft)	**23.0**	**29.9**