Master in Computer Vision Barcelona

[http://pagines.uab.cat/mcv/]

Xavier Giro-i-Nieto

@DocXavi

xavier.giro@upc.edu

Associate Professor

Universitat Politècnica de Catalunya

Module 6 - Day 9 - Lecture 2

Video Object Segmentation

5th April 2022

Acknowledgements

Carles Ventura

Miriam Bellver

Amaia Salvador

Andreu Girbau

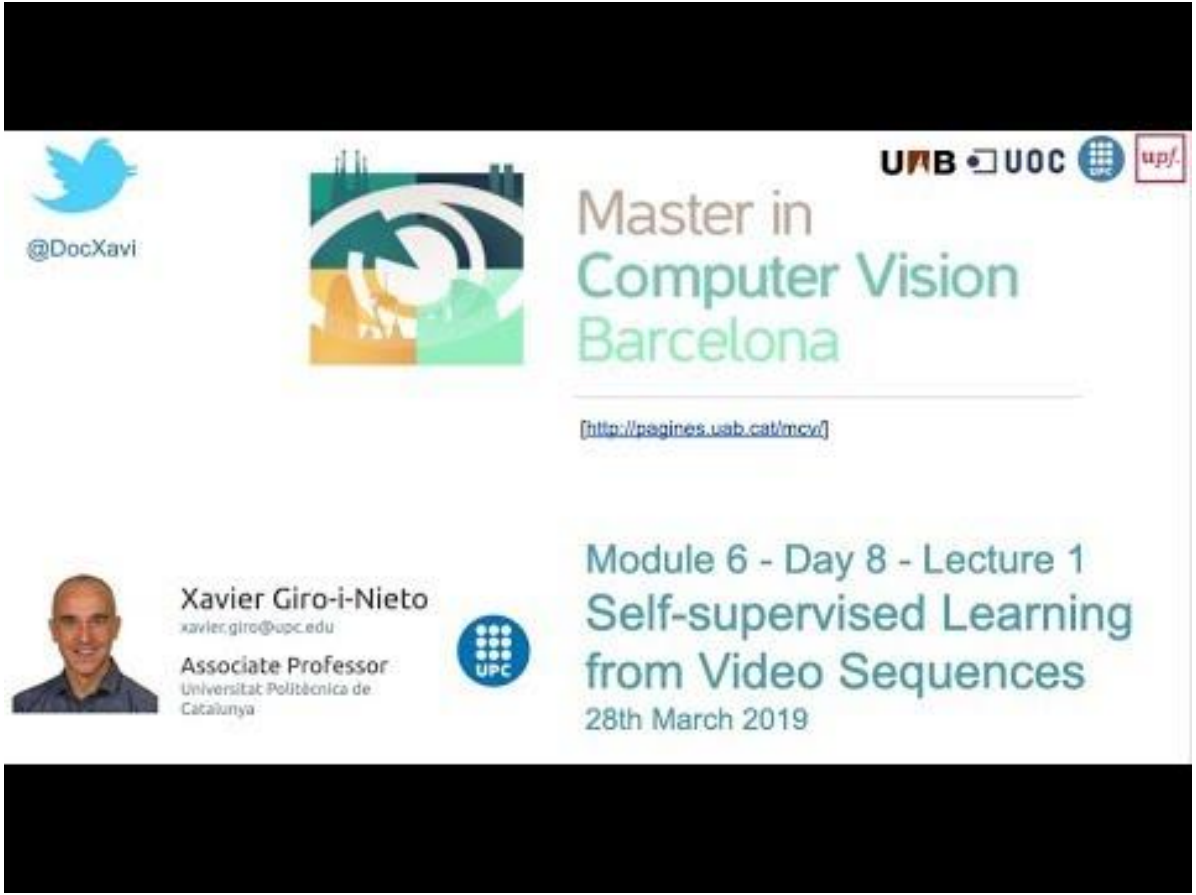
Universitat Oberta de Catalunya

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

UPC

MEDIAPRO

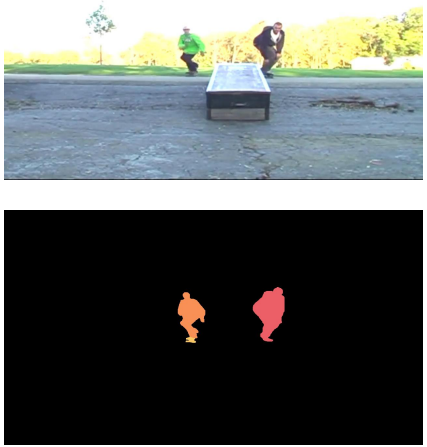
Videolecture



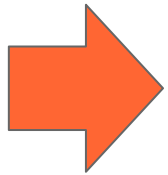
Outline

- **Motivation**
- Datasets & Benchmarks
- Online learning (Frame-based)
- Mask propagation
- Flow Propagation
- RNN

Video Object Segmentation (VOS)



VS



Semi-supervised ("One-shot") video object segmentation

Unsupervised ("zero-shot") video object segmentation

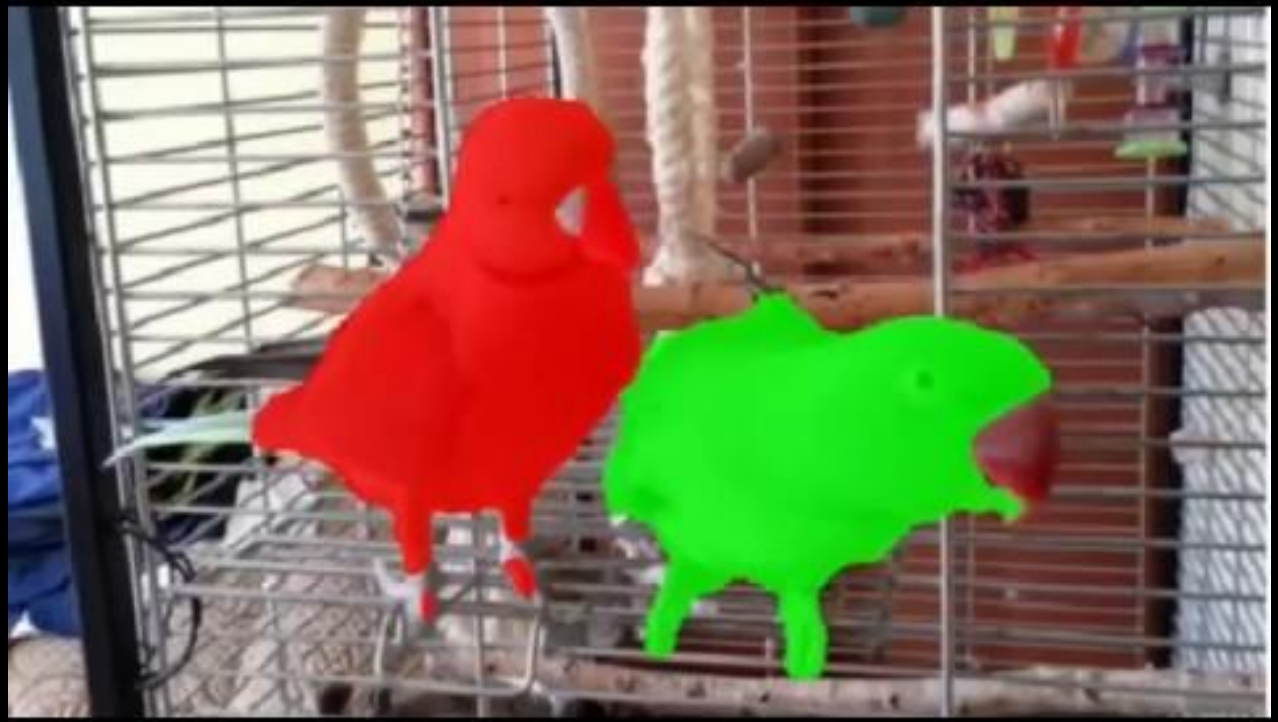
Semi-supervised VOS

One-shot VOS

#RVOS Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques and Xavier Giro-i-Nieto. "RVOS: End-to-End Recurrent Network for Video Object Segmentation", CVPR 2019.

Un supervised VOS

Zero-shot VOS



#RVOS Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques and Xavier Giro-i-Nieto. "RVOS: End-to-End Recurrent Network for Video Object Segmentation", CVPR 2019.

Outline

- Motivation
- **Datasets & Benchmarks**
- Online learning (Frame-based)
- Mask propagation
- Flow Propagation
- RNN

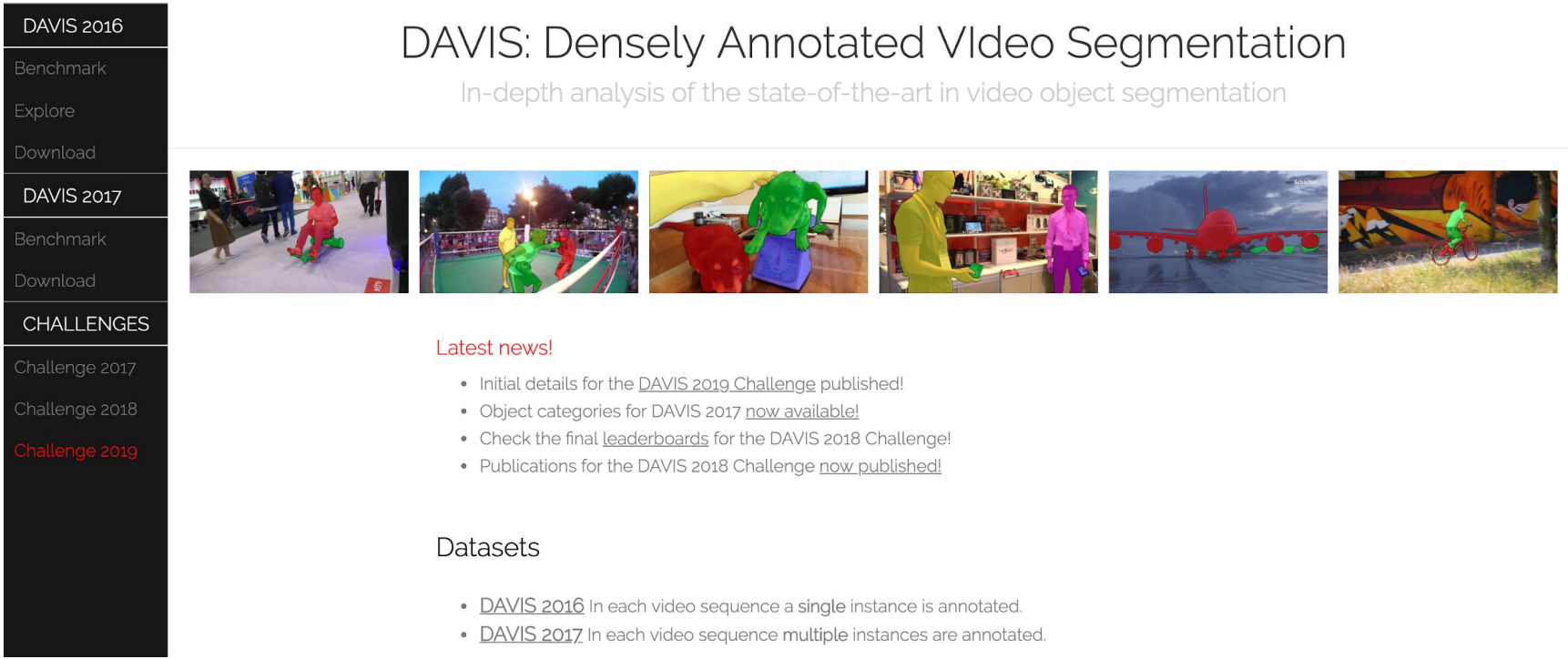
Datasets and Benchmarks

DAVIS-2017

- 90 training videos (train+val)
- 30 testing videos (test-dev set)

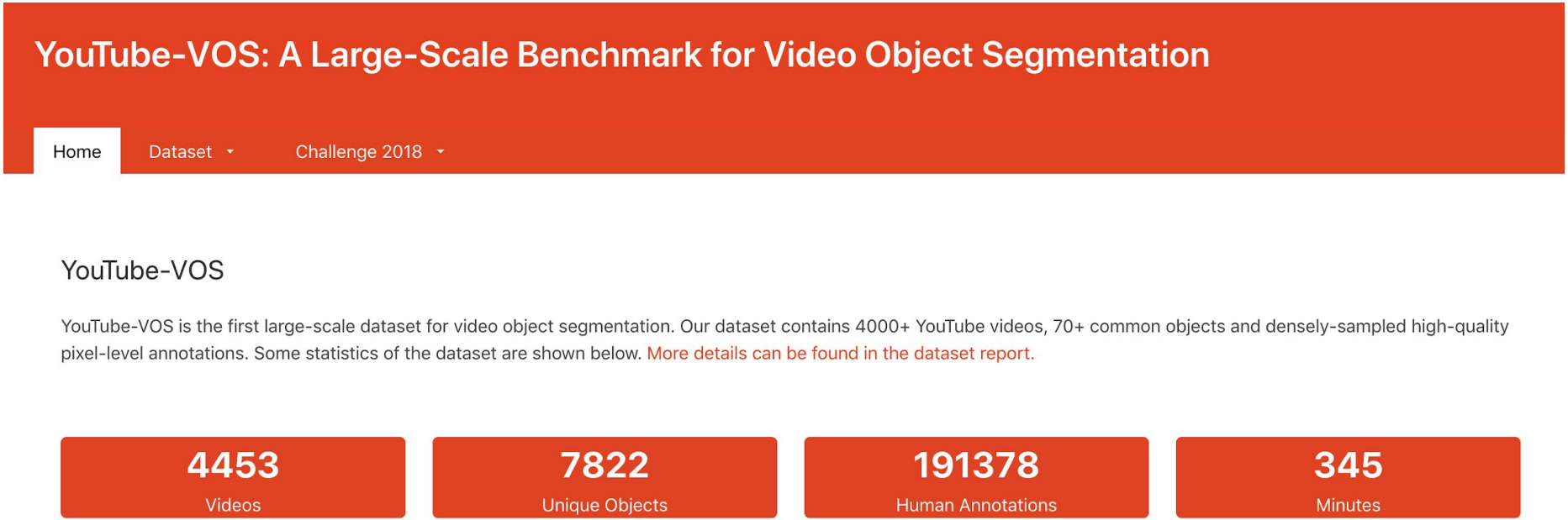
Perazzi, Federico, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. "A benchmark dataset and evaluation methodology for video object segmentation." CVPR 2016.

Datasets and Benchmarks



#DAVIS Perazzi, Federico, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. "A benchmark dataset and evaluation methodology for video object segmentation." CVPR 2016.

Datasets and Benchmarks



#YouTube-VOS Xu, Ning, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. "Youtube-vos: Sequence-to-sequence video object segmentation." ECCV 2018.

Datasets and Benchmarks

Scale	JC [21]	ST [22]	YTO [16]	FBMS [24]	DAVIS [15]	DAVIS [20]	**YouTube-VOS (Ours)**
Videos	22	14	96	59	50	90	**4,453**
Categories	14	11	10	16	-	-	**94**
Objects	22	24	96	139	50	205	**7,755**
Annotations	6,331	1,475	1,692	1,465	3,440	13,543	**197,272**
Duration	3.52	0.59	9.01	7.70	2.88	5.17	**334.81**

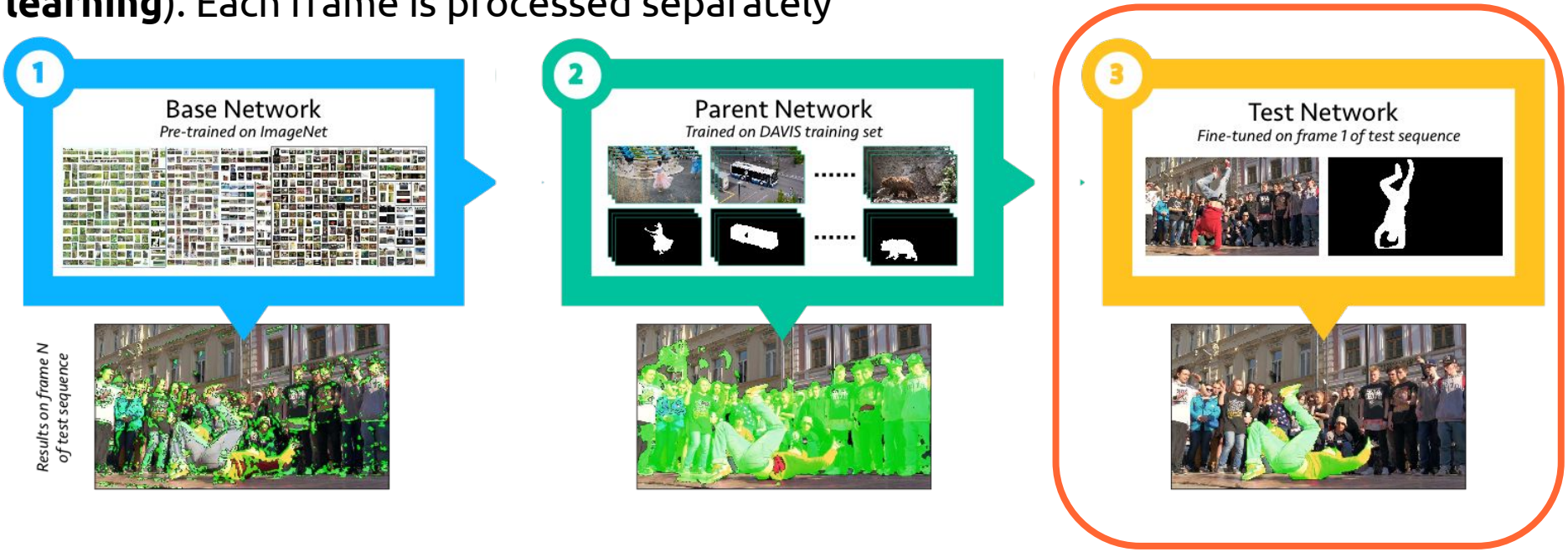
#YouTube-VOS Xu, Ning, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. "YouTube-VOS: A Large-Scale Video Object Segmentation Benchmark." arXiv preprint arXiv:1809.03327 (2018).

Outline

- Motivation
- Datasets & Benchmarks
- **Online learning (Frame-based)**
- Mask propagation
- Flow Propagation
- RNN

Online learning (frame-based)

A neural network is fine-tuned with the provided mask for the first frame (**online learning**). Each frame is processed separately



#OSVOS Caelles, Sergi, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. "One-shot video object segmentation." CVPR 2017. [Talk by Laura Leal-Taixé]

Online learning (frame-based)

Frame-based processing introduces temporal inconsistencies...

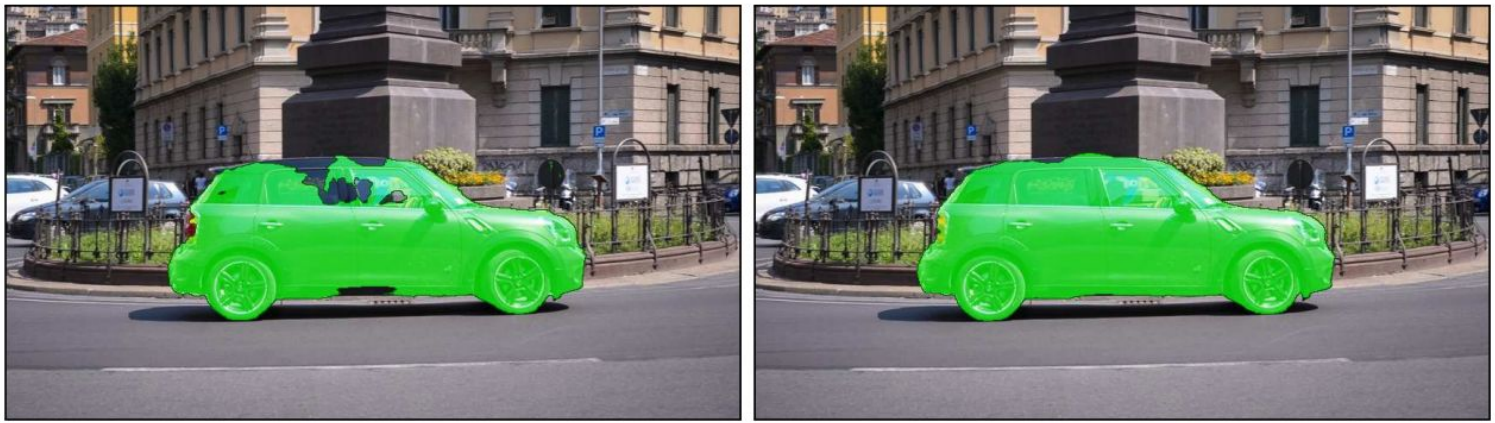
...but results are still very convincing.

#OSVOS Caelles, Sergi, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. "One-shot video object segmentation." CVPR 2017. [Talk by Laura Leal-Taixé]

Online learning (frame-based)

What are the limitations of online learning (OL) ?

Qualitative evolution of the fine tuning:
Results at 10 seconds and 1 minute per sequence.



#OSVOS Caelles, Sergi, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. "One-shot video object segmentation." CVPR 2017. [Talk by Laura Leal-Taixé]

Online learning (frame-based)

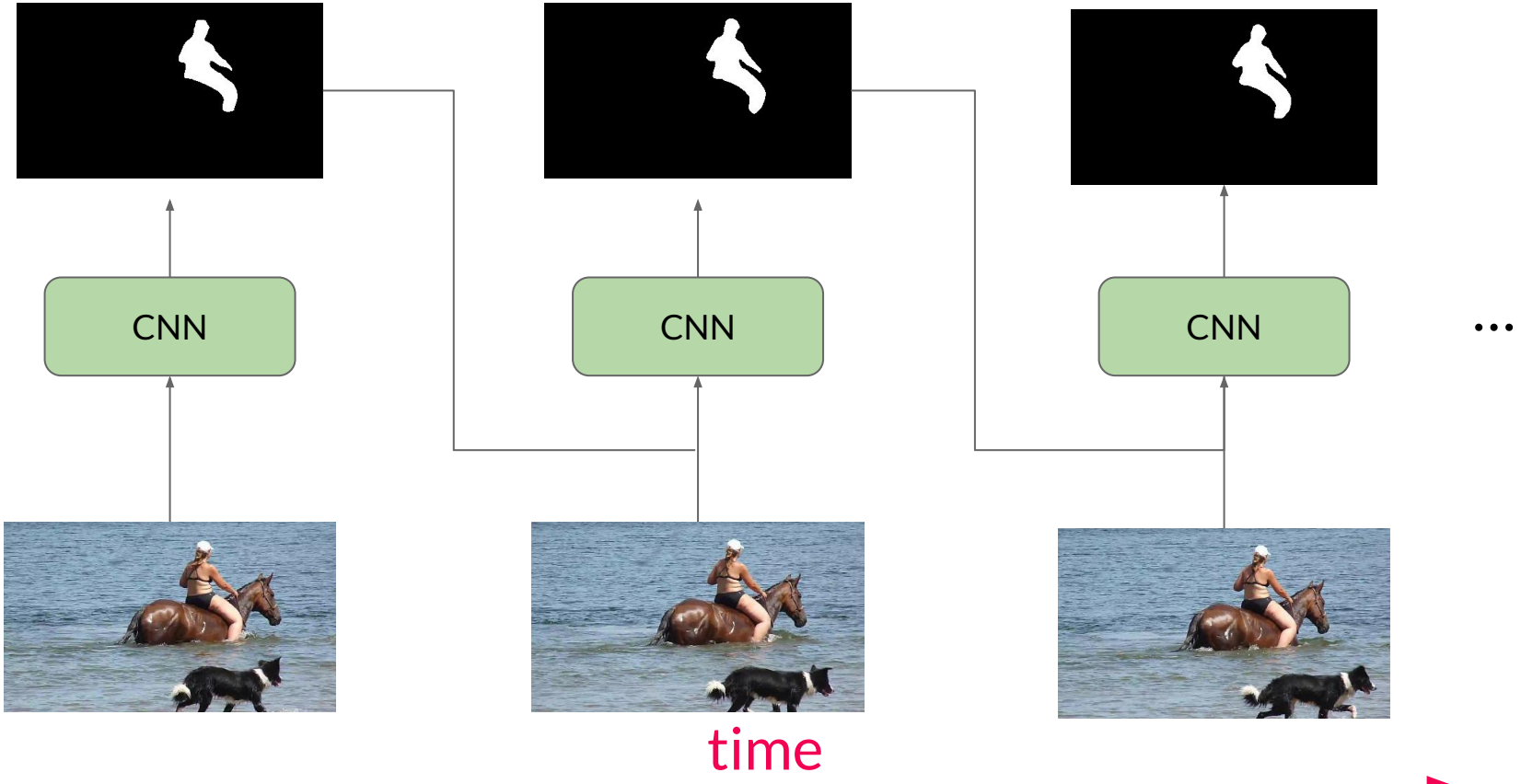
How is it possible to fine-tune a ConvNet with just a single frame ?

#OSVOS Caelles, Sergi, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. "One-shot video object segmentation." CVPR 2017. [Talk by Laura Leal-Taixé]

Outline

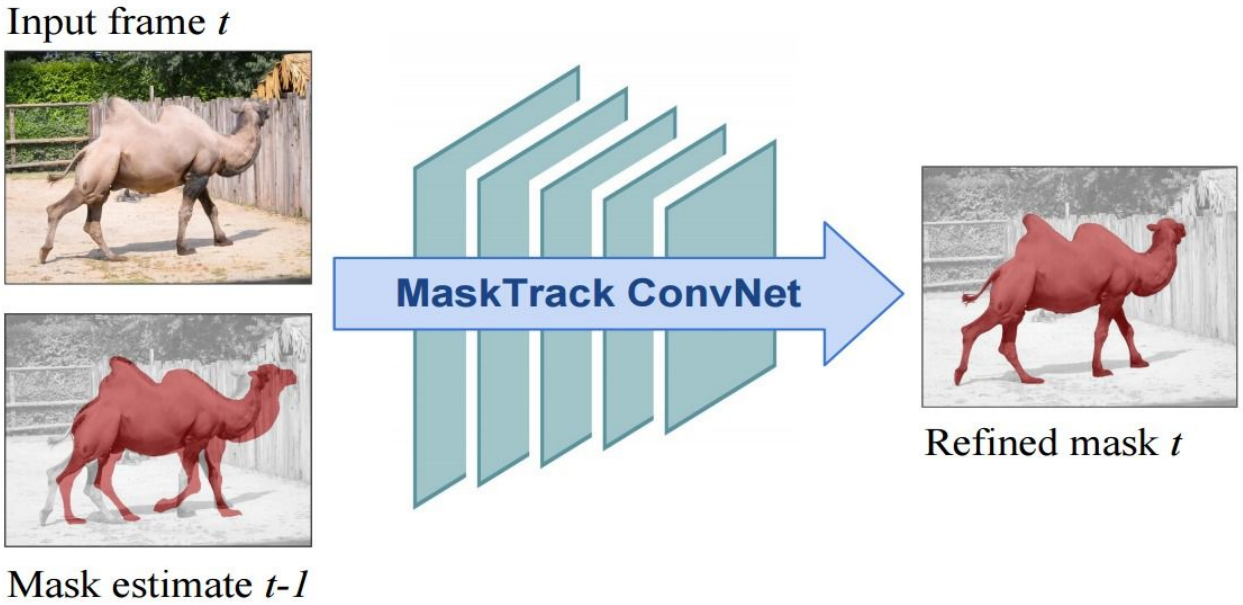
- Motivation
- Datasets & Benchmarks
- Online learning (Frame-based)
- **Mask propagation**
- Flow Propagation
- RNN

Mask Propagation



Mask Propagation

The ConvNet is trained to refine the previous mask to the current frame.

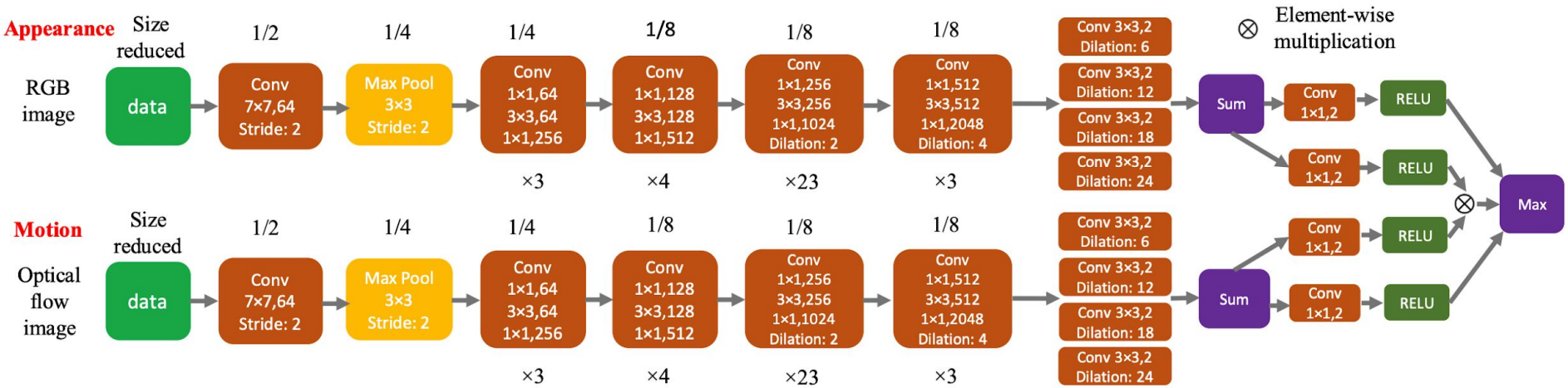


#MaskTrack Perazzi, Federico, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. "Learning video object segmentation from static images." CVPR 2017. [talk]

Outline

- Motivation
- Datasets & Benchmarks
- Online learning (Frame-based)
- Mask propagation
- **Flow Propagation**
- RNN

Flow Propagation



Jain, Suyog Dutt, Bo Xiong, and Kristen Grauman. "Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos." CVPR 2017.

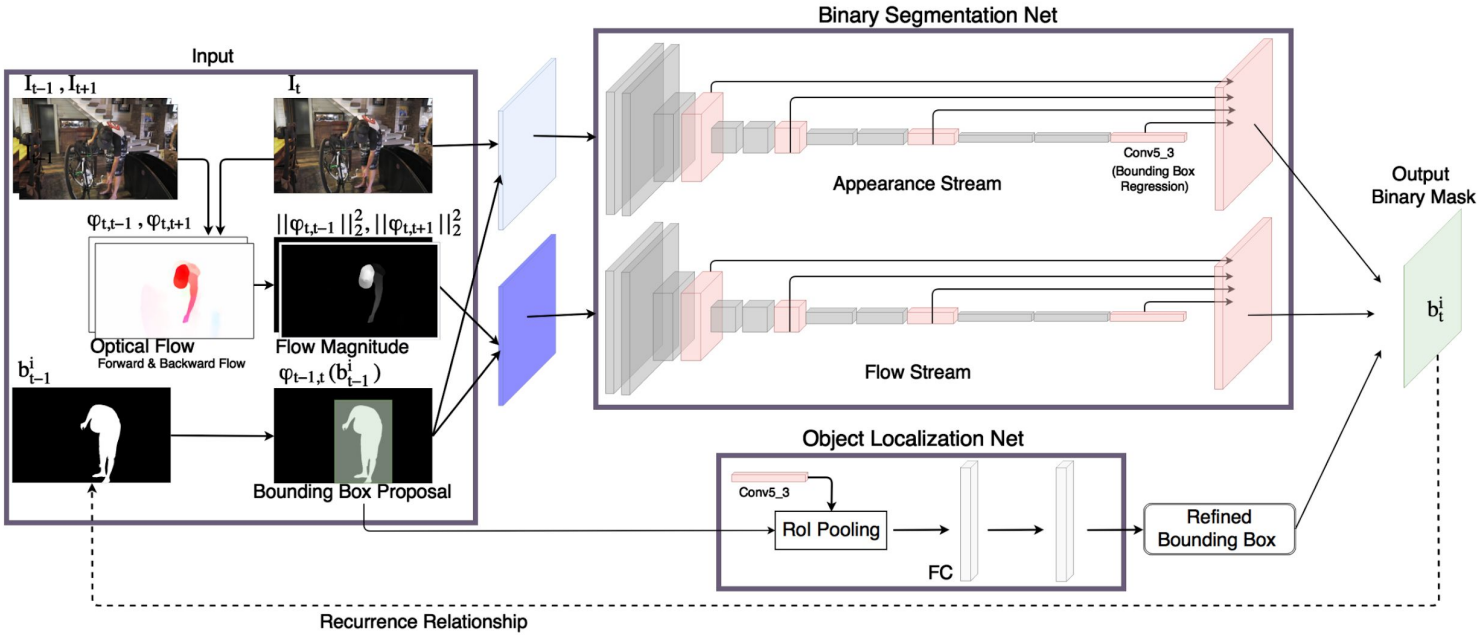
Jain, Suyog Dutt, Bo Xiong, and Kristen Grauman. "Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos." CVPR 2017.

Outline

- Motivation
- Datasets & Benchmarks
- Online learning (Frame-based)
- **Mask propagation**
- **Flow Propagation**
- RNN

Mask + Flow Propagation

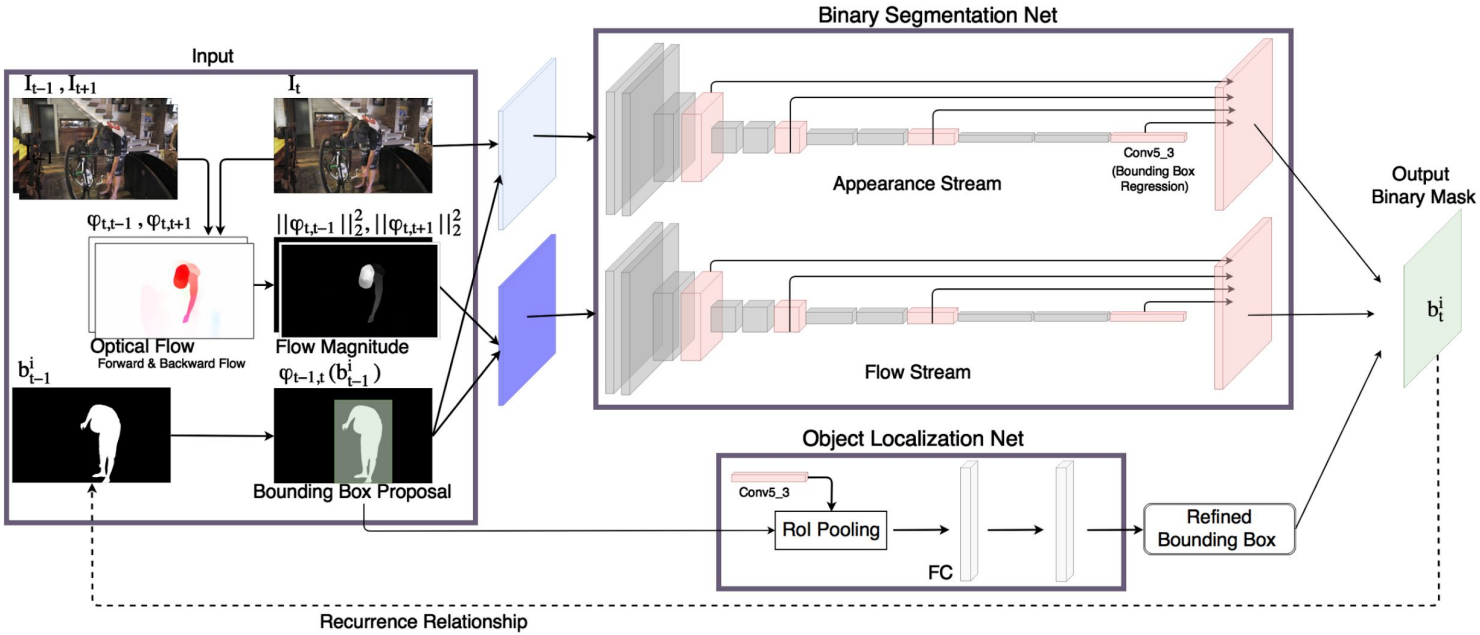
The masks of the N objects in the previous frame are warped with the optical flow.
Each mask is fed separately into another NN that detects & segments instances.



#MaskRNN Hu, Yuan-Ting, Jia-Bin Huang, and Alexander Schwing. "MaskRNN: Instance level video object segmentation." NIPS 2017.

Mask + Flow Propagation

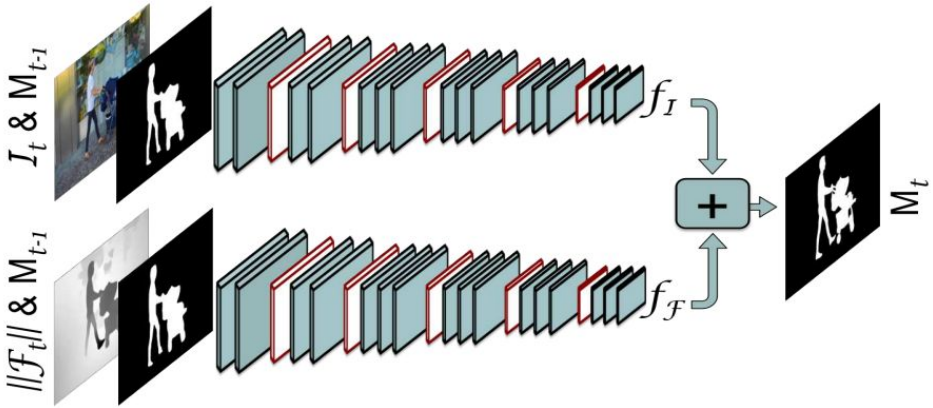
Where is the RNN in the MaskRNN architecture ?



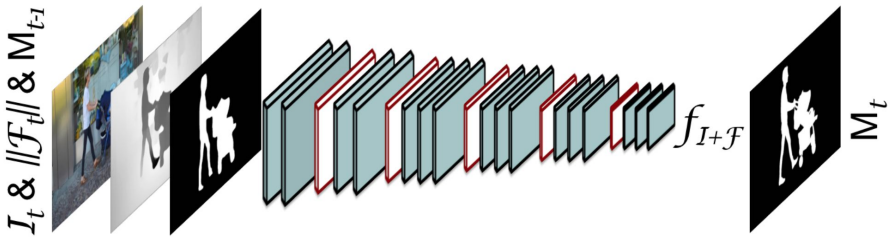
#MaskRNN Hu, Yuan-Ting, Jia-Bin Huang, and Alexander Schwing. "MaskRNN: Instance level video object segmentation." NIPS 2017.

Mask + Flow Propagation

Mask from previous frame is warped & concatenated with optical flow in two set ups:



Two streams

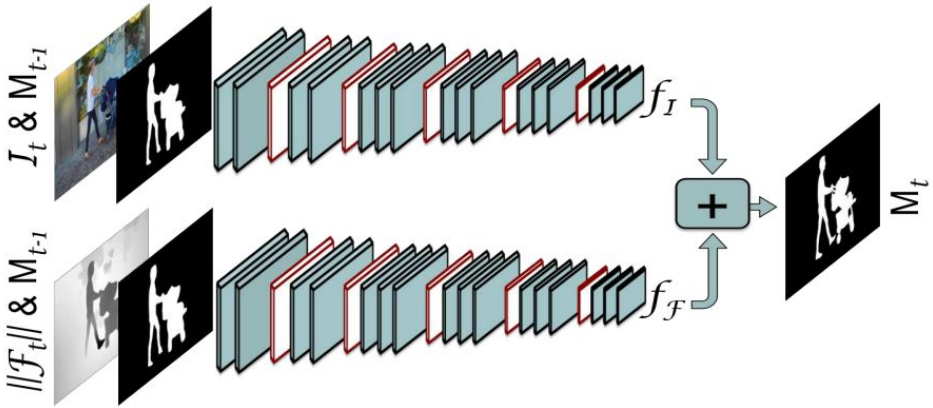


One stream

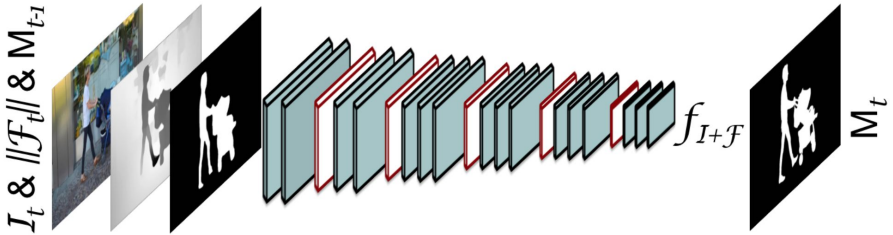
#LucidTracker Khoreva, Anna, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. "Lucid data dreaming for object tracking." IJCV 2019.

Mask + Flow Propagation

How could these architectures deal with multiple objects in a single pass ?



Two streams

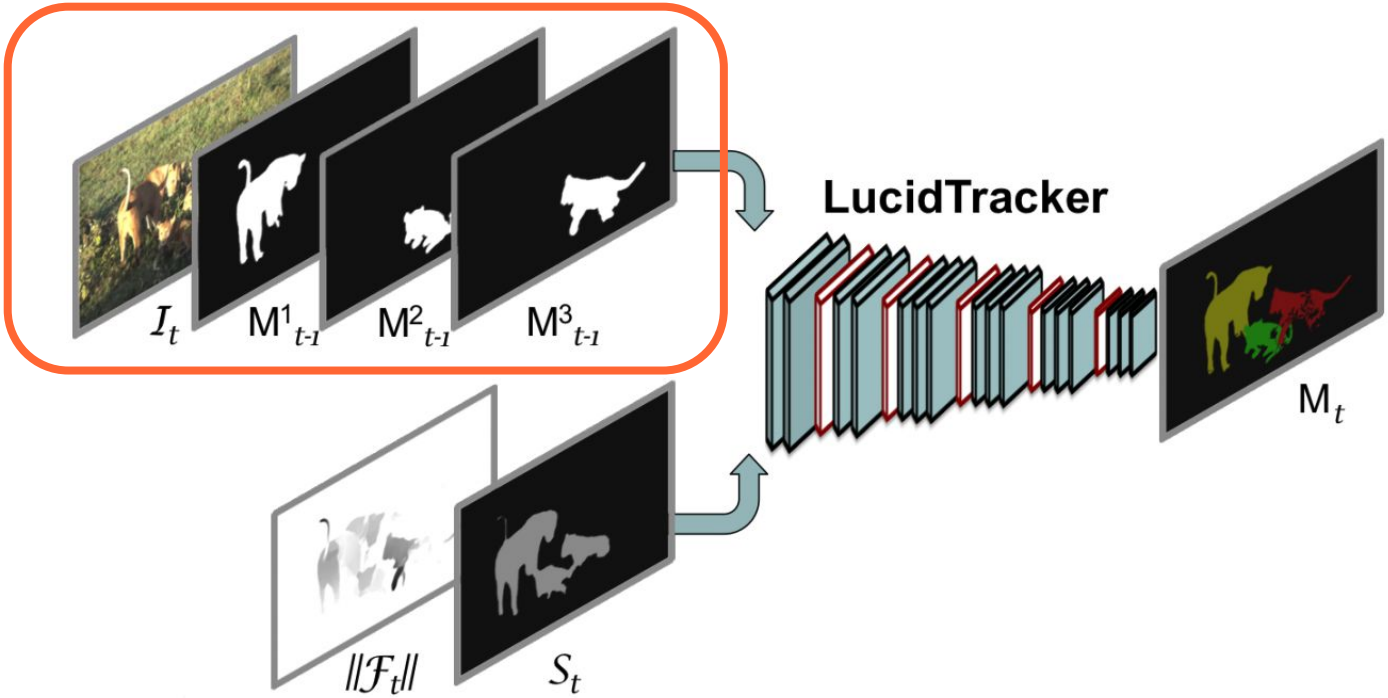


One stream

#LucidTracker Khoreva, Anna, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. "Lucid data dreaming for object tracking." IJCV 2019.

Mask + Flow Propagation

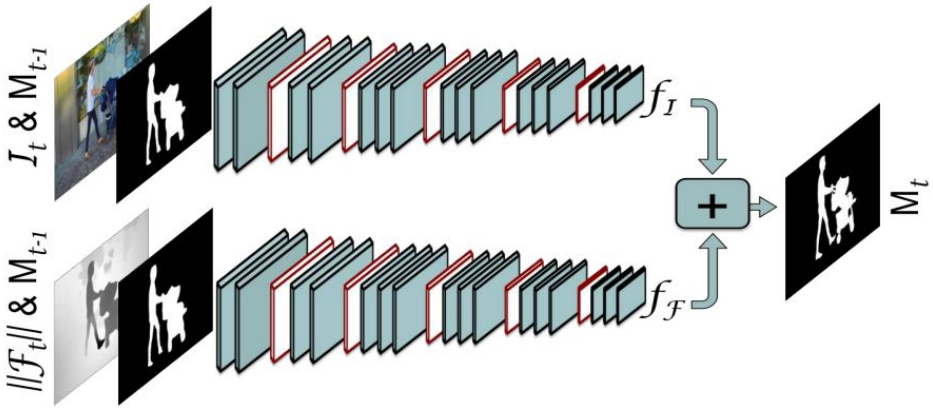
Multiple object tracking is handled by adding more mask channels.



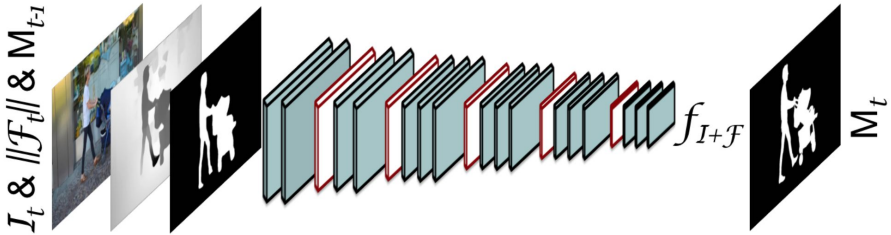
#LucidTracker Khoreva, Anna, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. "Lucid data dreaming for object tracking." IJCV 2019.

Mask + Flow Propagation

Which architecture do you think it will perform better ?



Two streams



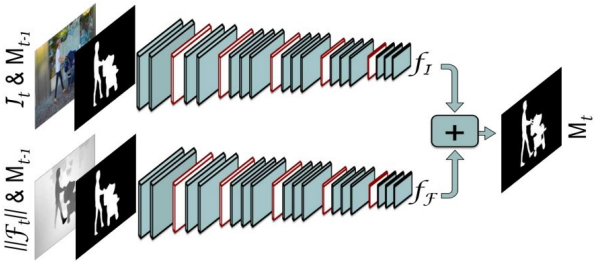
One stream

#LucidTracker Khoreva, Anna, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. "Lucid data dreaming for object tracking." IJCV 2019.

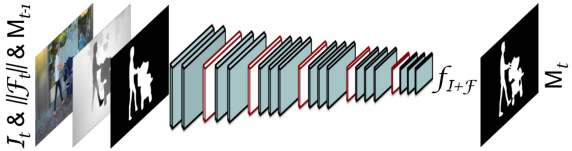
Mask + Flow Propagation

Which architecture do you think it will perform better ?

Two streams



One stream



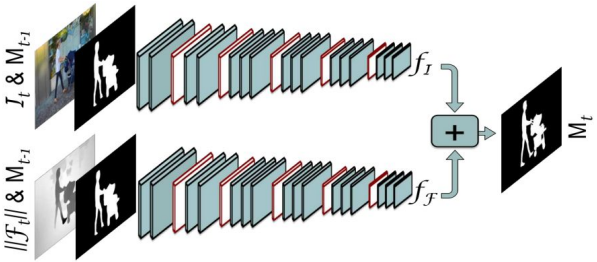
Architecture	ImgNet pre-train.	per-dataset training	per-video fine-tun.	$DAVIS_{16}$ mIoU
two streams	✓	✓	✗	**80.9**
one stream	✓	✓	✗	80.3

#LucidTracker Khoreva, Anna, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. "Lucid data dreaming for object tracking." IJCV 2019.

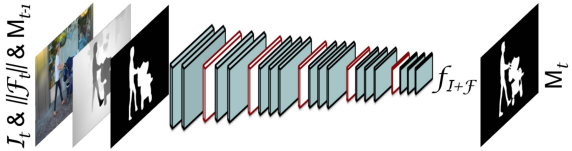
Mask + Flow Propagation

Which architecture would you use ?

Two streams



One stream



Architecture	ImgNet pre-train.	per-dataset training	per-video fine-tun.	$DAVIS_{16}$ mIoU
two streams	✓	✓	✗	**80.9**
one stream	✓	✓	✗	80.3

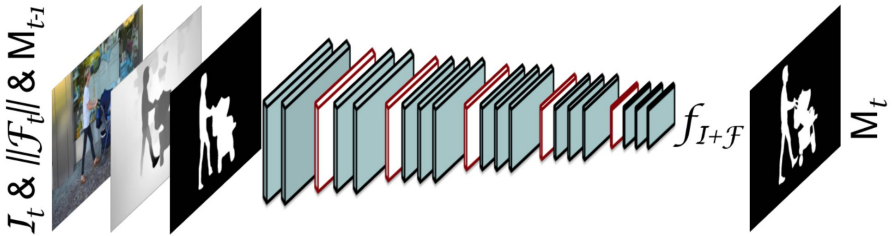
#LucidTracker Khoreva, Anna, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. "Lucid data dreaming for object tracking." IJCV 2019.

Mask + Flow Propagation

Which architecture would you use ?

"**One stream** network is more affordable to train and allows to easily add extra input channels, e.g. providing additional semantic information about objects."

"The lighter one stream network performs as well as a network with two streams. We will thus use the **one stream** architecture"



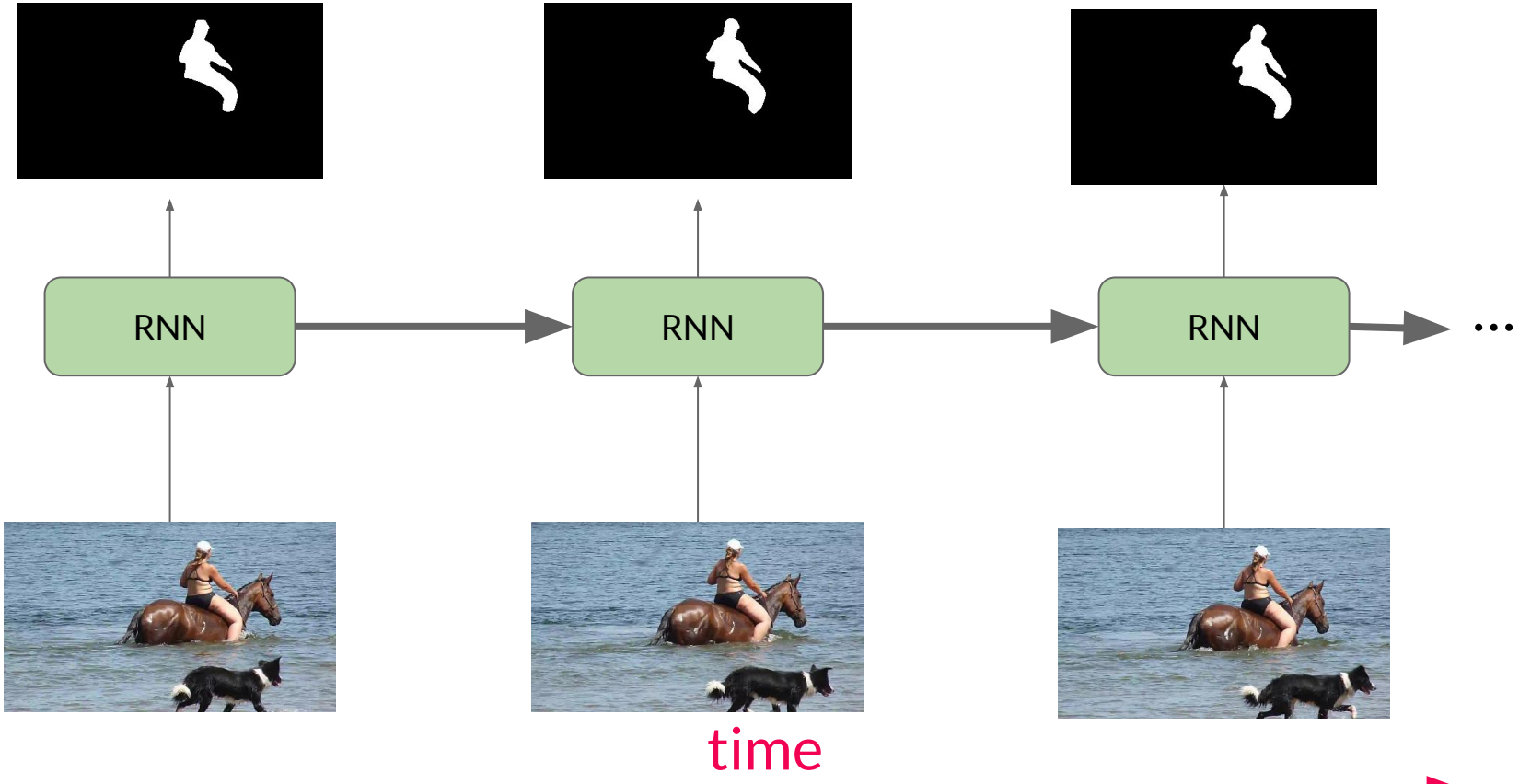
One stream

#LucidTracker Khoreva, Anna, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. "Lucid data dreaming for object tracking." IJCV 2019.

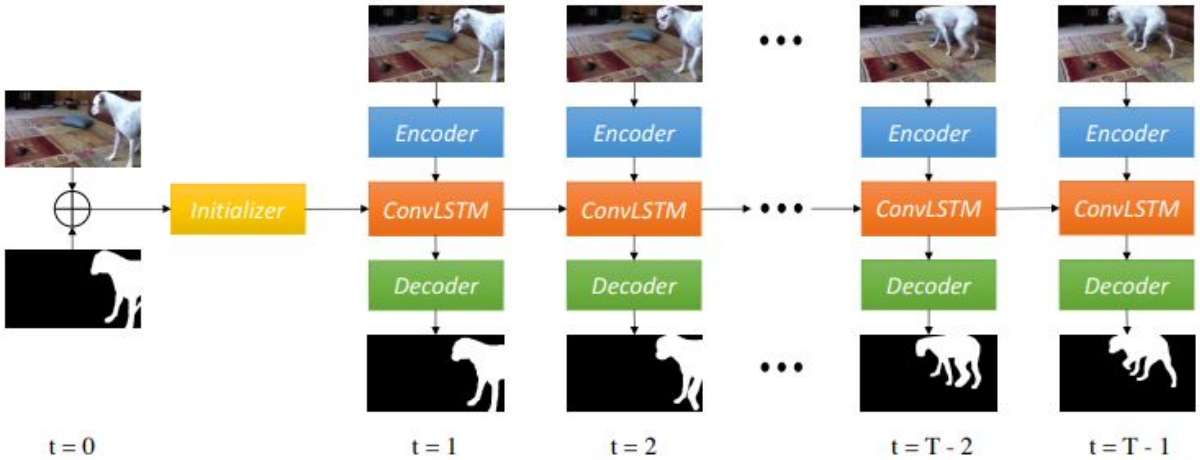
Outline

- Motivation
- Datasets & Benchmarks
- Online learning (Frame-based)
- Mask propagation
- Flow Propagation
- **RNN**

RNN



RNN (ConvLSTM)

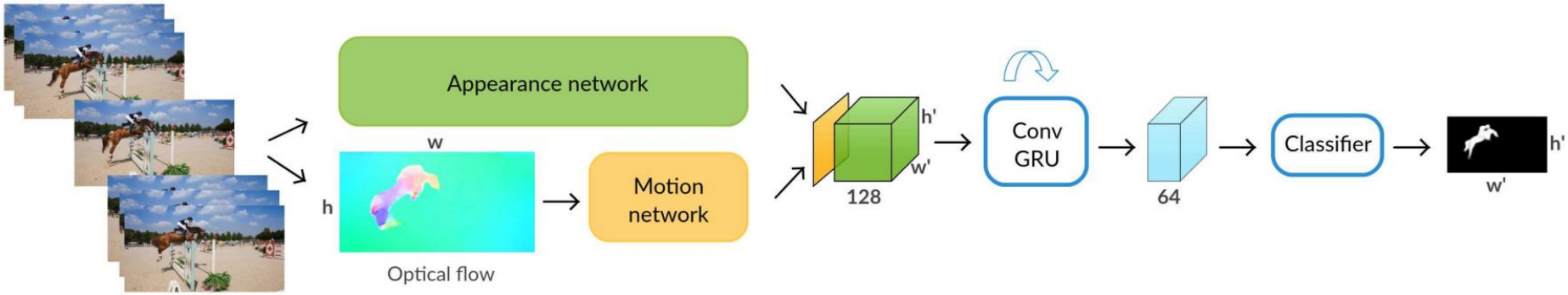


Limitations

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

#S2S Xu, Ning, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. "Youtube-vos: Sequence-to-sequence video object segmentation." ECCV 2018.

RNN



Limitations

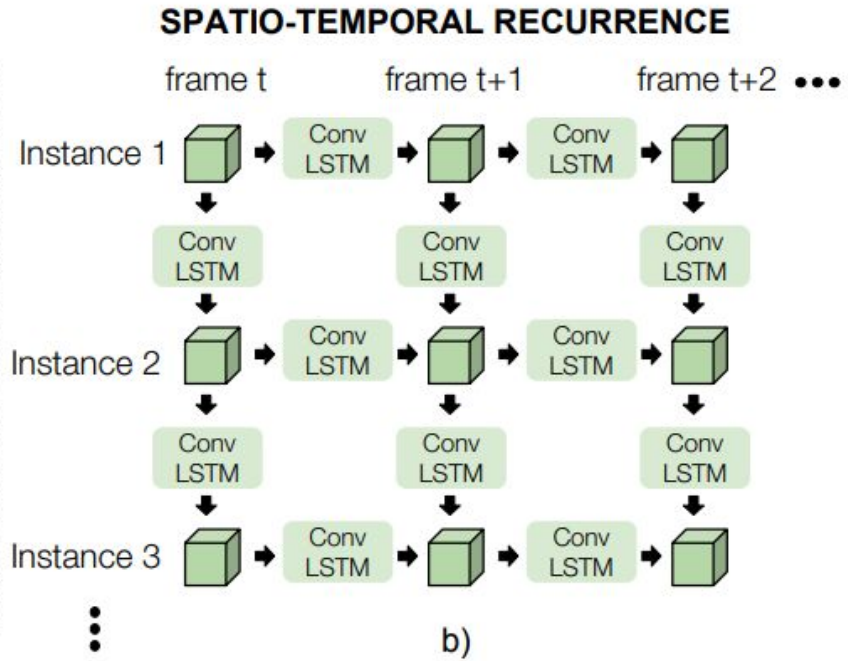
- Each instance is trained and segmented independently
- Optical flow depends on a network trained for another task: model is not end-to-end trainable

Tokmakov, Pavel, Karteek Alahari, and Cordelia Schmid. "Learning video object segmentation with visual memory." ICCV 2017. [talk]

RNN (Spatial + Temporal)

time
(frame sequence)

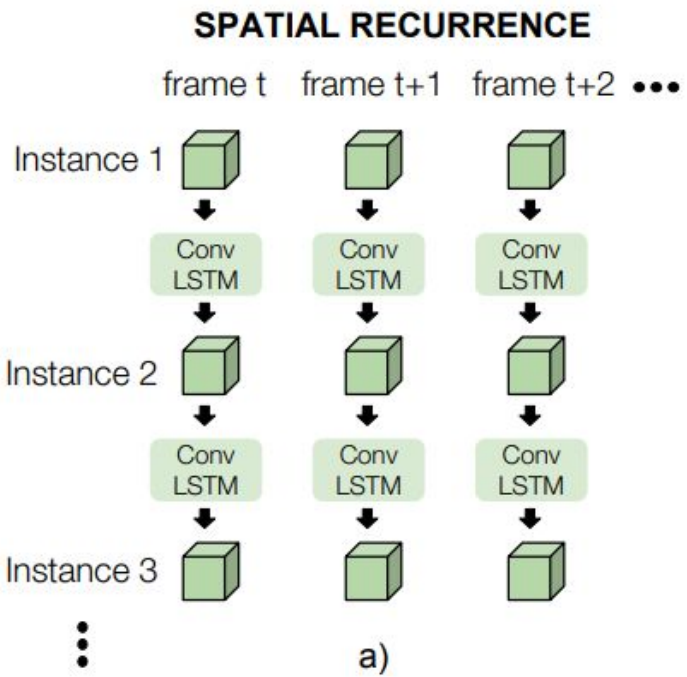
space
(object sequence)



#RVOS Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques and Xavier Giro-i-Nieto. "RVOS: End-to-End Recurrent Network for Video Object Segmentation", CVPR 2019.

RNN (Spatial)

Previous work

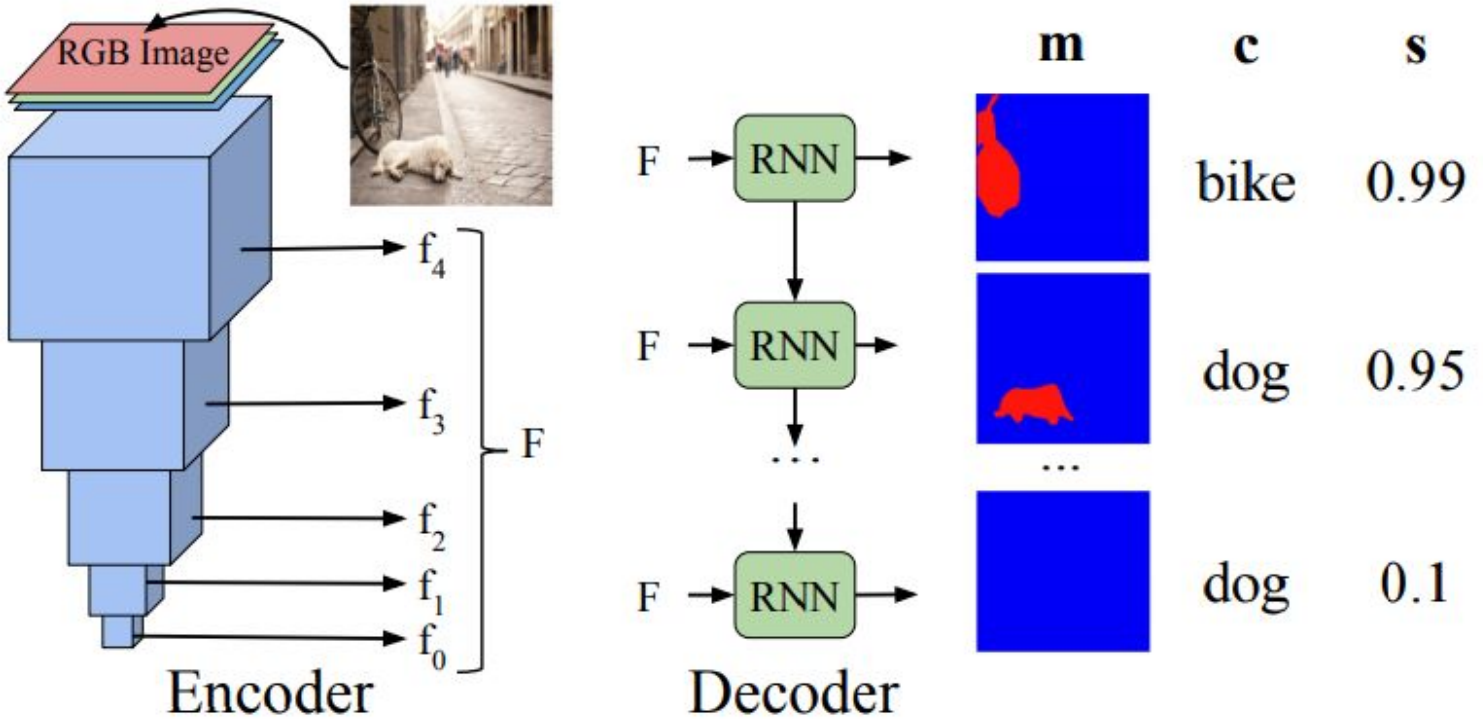


space
(object sequence)

#RSIS Salvador, Amaia, Miriam Bellver, Victor Campos, Manel Baradad, Ferran Marques, Jordi Torres, and Xavier Giro-i-Nieto. "Recurrent neural networks for semantic instance segmentation." CVPR Workshops 2018.

RNN (Spatial)

Previous work

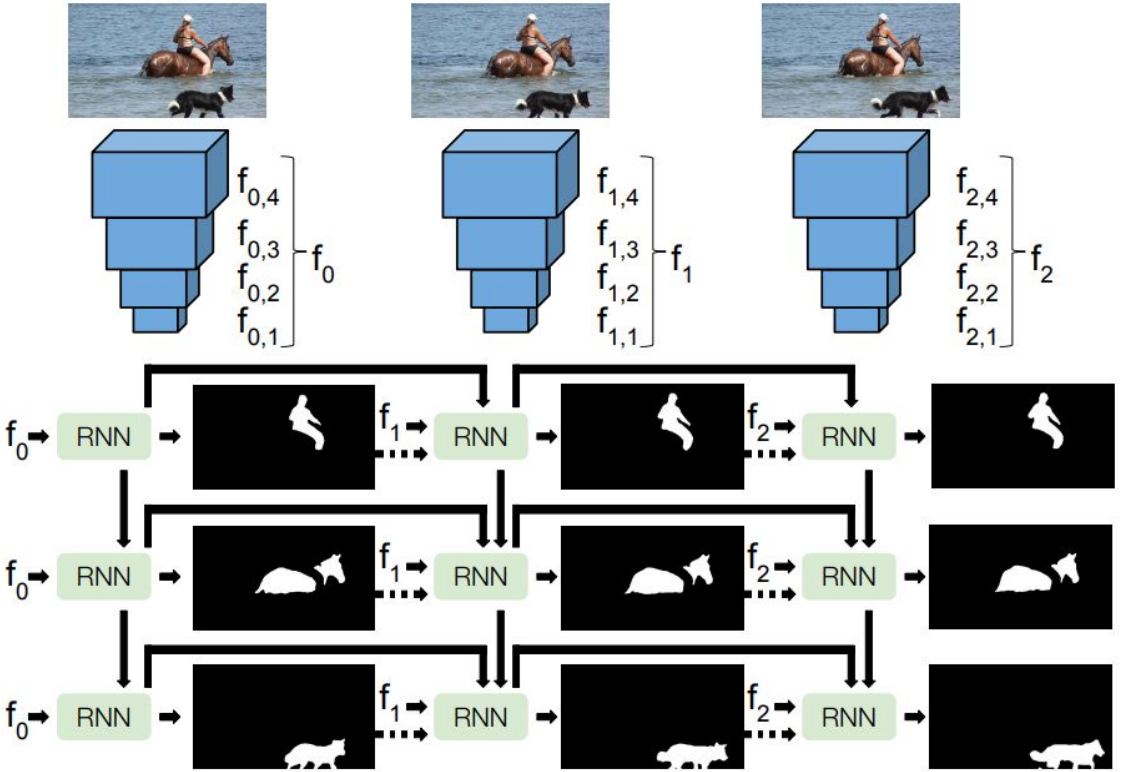


#RSIS Salvador, Amaia, Miriam Bellver, Victor Campos, Manel Baradad, Ferran Marques, Jordi Torres, and Xavier Giro-i-Nieto. "Recurrent neural networks for semantic instance segmentation." arXiv preprint arXiv:1712.00617 (2017).

RNN (Spatial + Temporal)

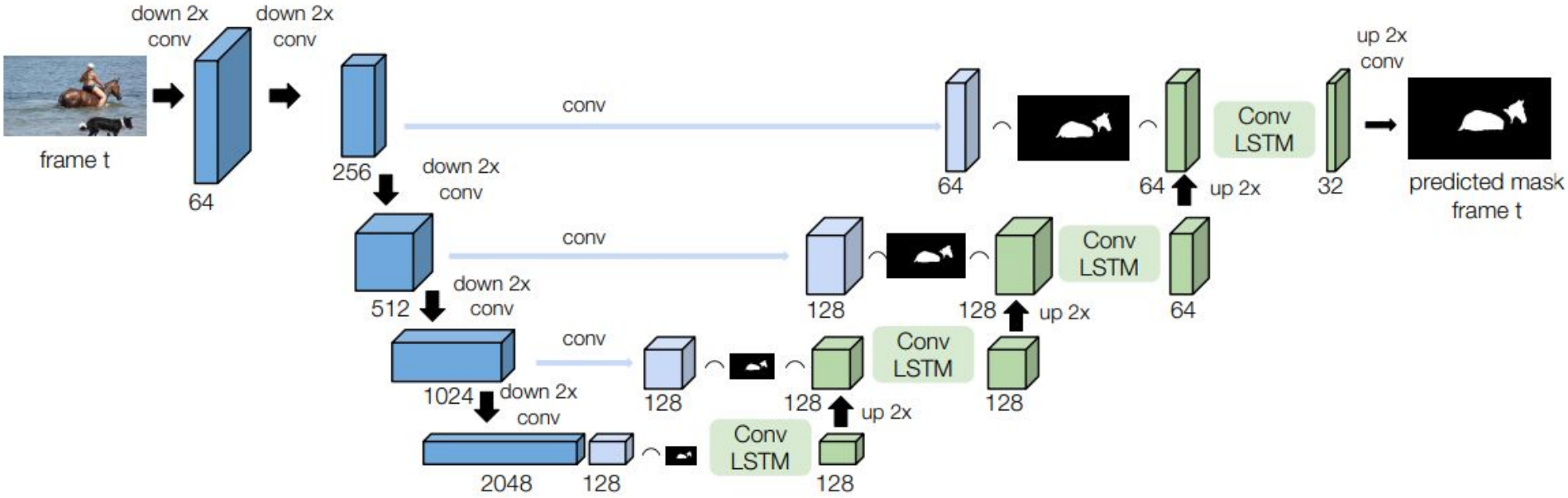
time
(frame sequence)

space
(object sequence)



#RVOS Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques and Xavier Giro-i-Nieto. "RVOS: End-to-End Recurrent Network for Video Object Segmentation", CVPR 2019.

RNN (Spatial + Temporal)



#RVOS Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques and Xavier Giro-i-Nieto. "RVOS: End-to-End Recurrent Network for Video Object Segmentation", CVPR 2019.

RNN (Spatial + Temporal)

One-shot Quality vs Inference Time for the Semi-supervised (one-shot) task

Speed values measured on a GPU K80 (*) and P100 (†), otherwise obtained from YouTube-VOS paper..

	OL	J_{seen}	J_{unseen}	F_{seen}	F_{unseen}	Speed (s/frame)
OSVOS [3]	✓	59.8	**54.2**	60.5	**60.7**	10
MaskTrack [17]	✓	59.9	45.0	59.5	47.9	12
OnAVOS [25]	✓	**60.1**	46.6	**62.7**	51.4	13
S2S w/o OL [28]	✗	**66.7**	**48.2**	65.5	50.3	0.16
OSMN [29]	✗	60.0	40.6	60.1	44.0	0.14 / 0.108* / 0.065†
RVOS-Mask-ST+	✗	63.6	45.5	**67.2**	**51.0**	**0.067* / 0.044†**

#RVOS Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques and Xavier Giro-i-Nieto. "RVOS: End-to-End Recurrent Network for Video Object Segmentation", CVPR 2019.

RNN (Spatial + Temporal)

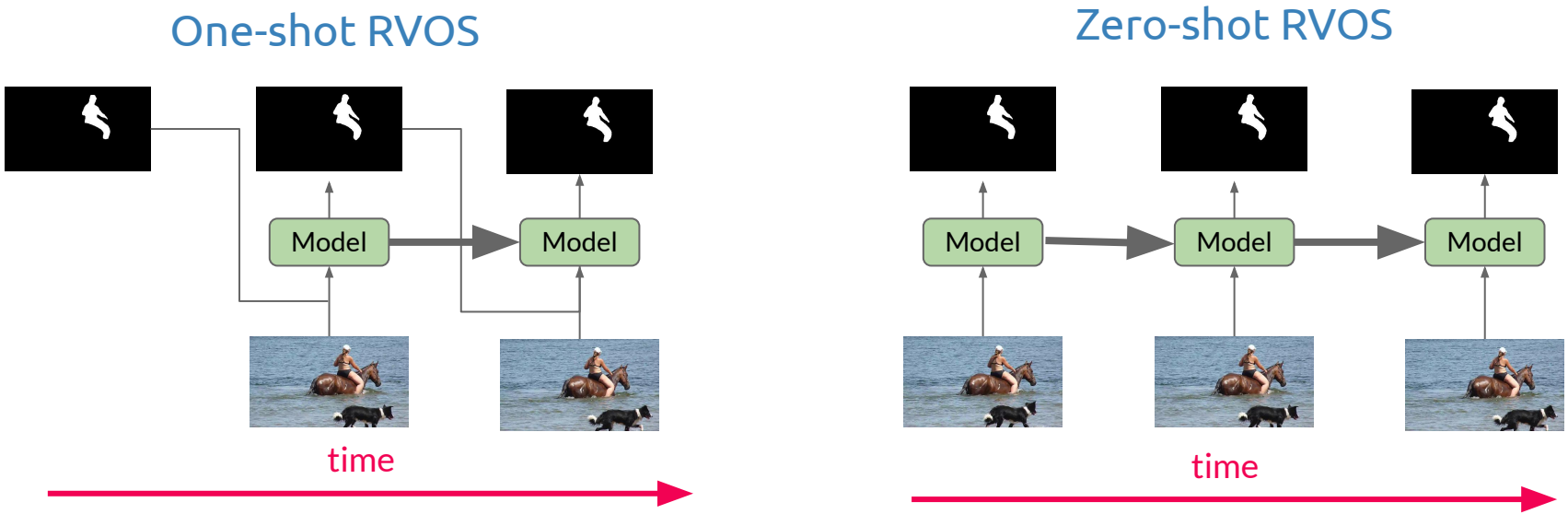
Why are techniques using online learning (OL) much slower than those that don't ?

	OL	J_{seen}	J_{unseen}	F_{seen}	F_{unseen}	Speed (s/frame)
OSVOS [3]	✓	59.8	**54.2**	60.5	**60.7**	10
MaskTrack [17]	✓	59.9	45.0	59.5	47.9	12
OnAVOS [25]	✓	**60.1**	46.6	**62.7**	51.4	13
S2S w/o OL [28]	✗	**66.7**	**48.2**	65.5	50.3	0.16
OSMN [29]	✗	60.0	40.6	60.1	44.0	0.14 / 0.108* / 0.065†
RVOS-Mask-ST+	✗	63.6	45.5	**67.2**	**51.0**	**0.067* / 0.044†**

#RVOS Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques and Xavier Giro-i-Nieto. "RVOS: End-to-End Recurrent Network for Video Object Segmentation", CVPR 2019.

RNN (Spatial + Temporal)

RVOS can naturally solve both the semi-supervised (one-shot) & unsupervised (zero-shot) tasks:



#RVOS Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques and Xavier Giro-i-Nieto. "RVOS: End-to-End Recurrent Network for Video Object Segmentation", CVPR 2019.

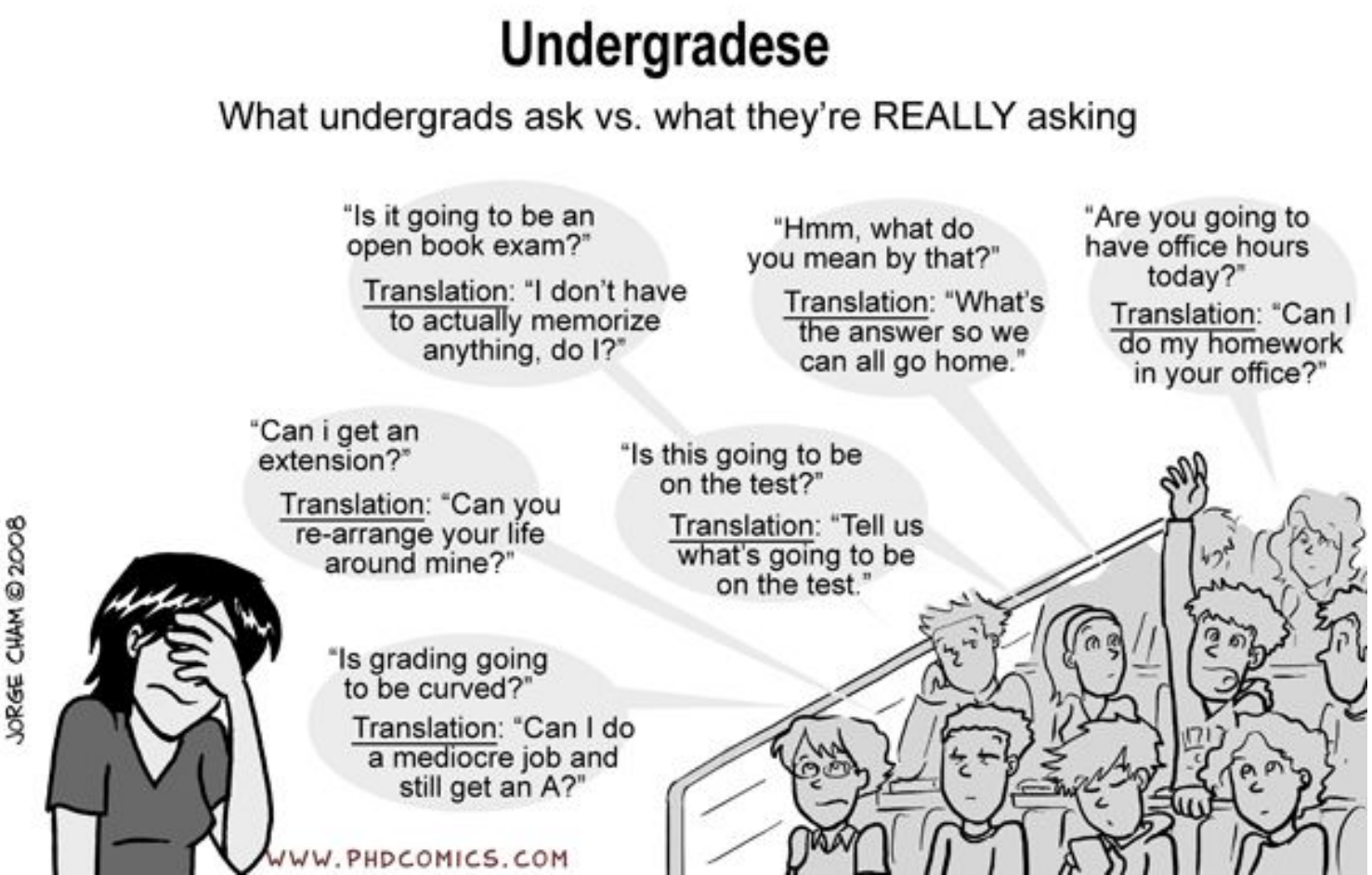
Outline

- Motivation
- Datasets & Benchmarks
- Online learning (Frame-based)
- Mask propagation
- Flow Propagation
- Transformer

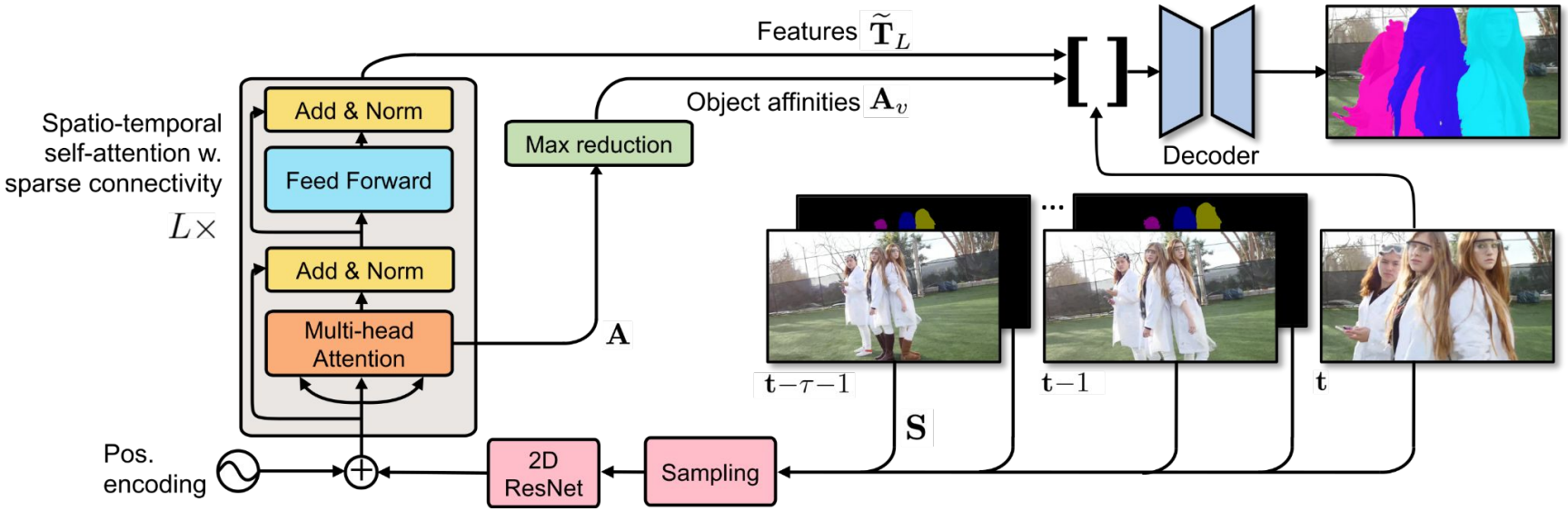
Spatio-Temporal Networks

#STM Oh, S. W., Lee, J. Y., Xu, N., & Kim, S. J. Video object segmentation using space-time memory networks. ICCV 2019. [code]

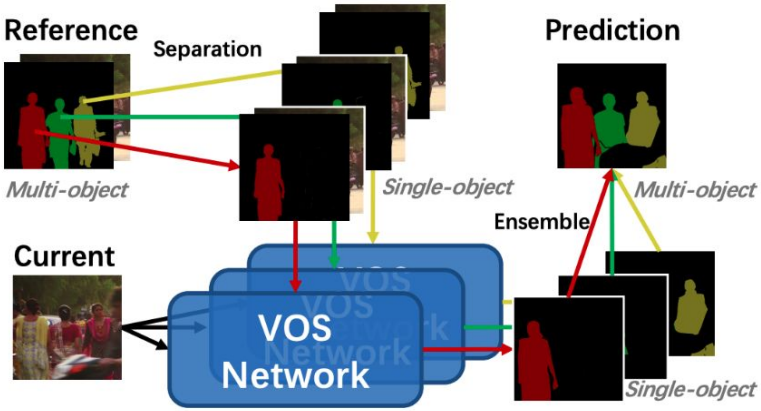
Questions ?


Undergradese
What undergrads ask vs. what they're REALLY asking
"Is it going to be an open book exam?"
Translation: "I don't have to actually memorize anything, do I?"
"Hmm, what do you mean by that?"
Translation: "What's the answer so we can all go home."
"Are you going to have office hours today?"
Translation: "Can I do my homework in your office?"
"Can i get an extension?"
Translation: "Can you re-arrange your life around mine?"
"Is this going to be on the test?"
Translation: "Tell us what's going to be on the test."
"Is grading going to be curved?"
Translation: "Can I do a mediocre job and still get an A?"
JORGE CHAM © 2008
WWW.PHDCOMICS.COM

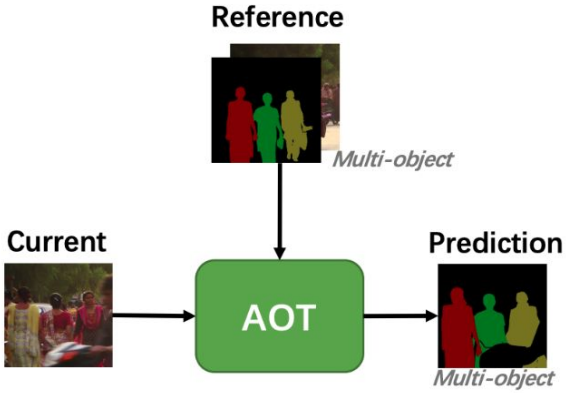
Sparse Transformers



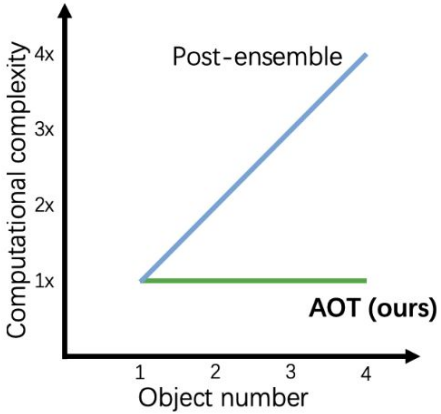
#SSTVOS Duke, B., Ahmed, A., Wolf, C., Aarabi, P., & Taylor, G. W. SSTVO: Sparse spatiotemporal transformers for video object segmentation. CVPR 2021. [code]



(a) Post-ensemble

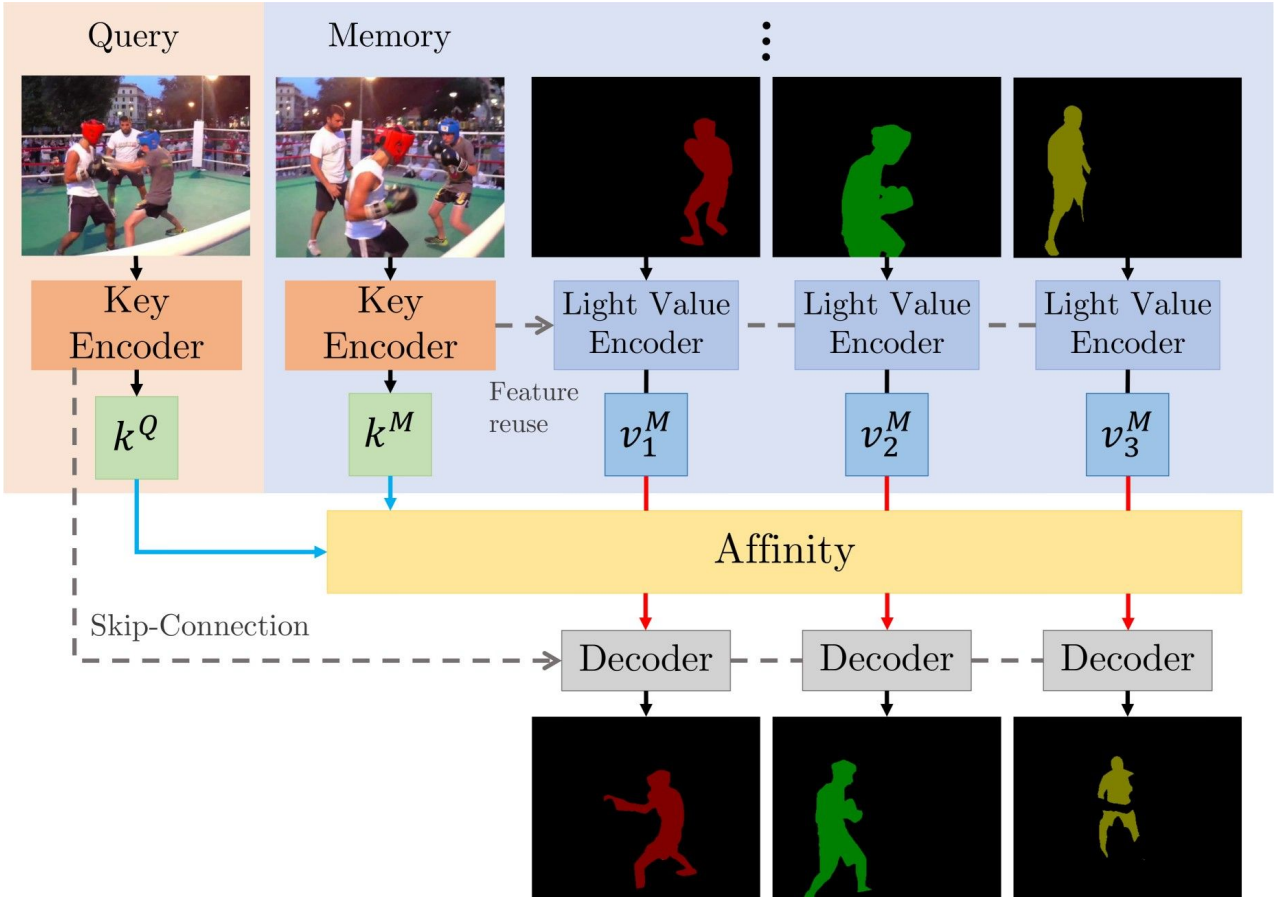


(b) Associating objects (ours)

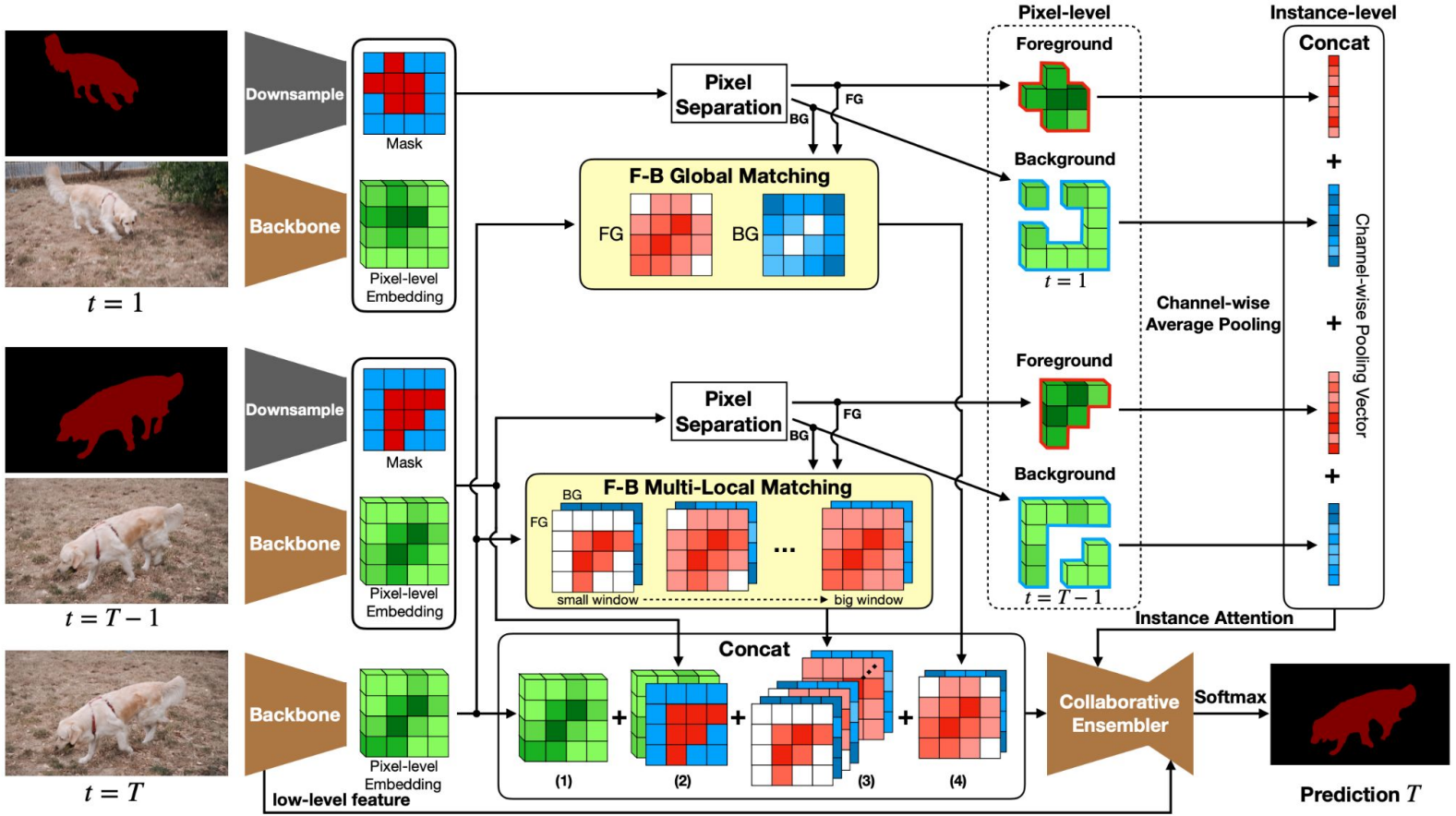


(c) Comparison

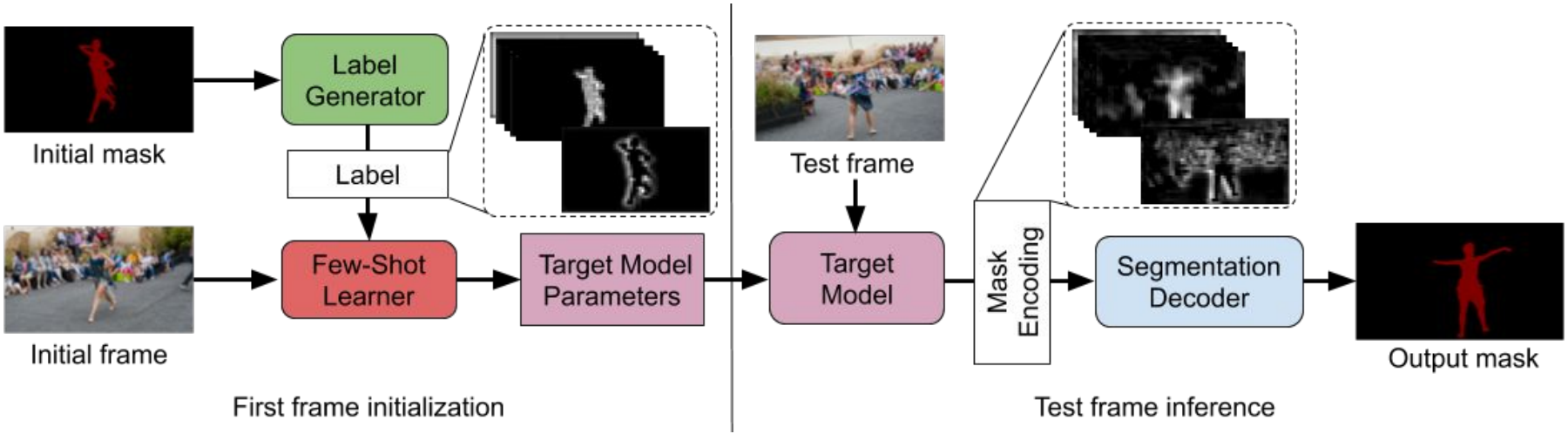
#AOT Yang, Zongxin, Yunchao Wei, and Yi Yang. "Associating Objects with Transformers for Video Object Segmentation." NeurIPS 2021.



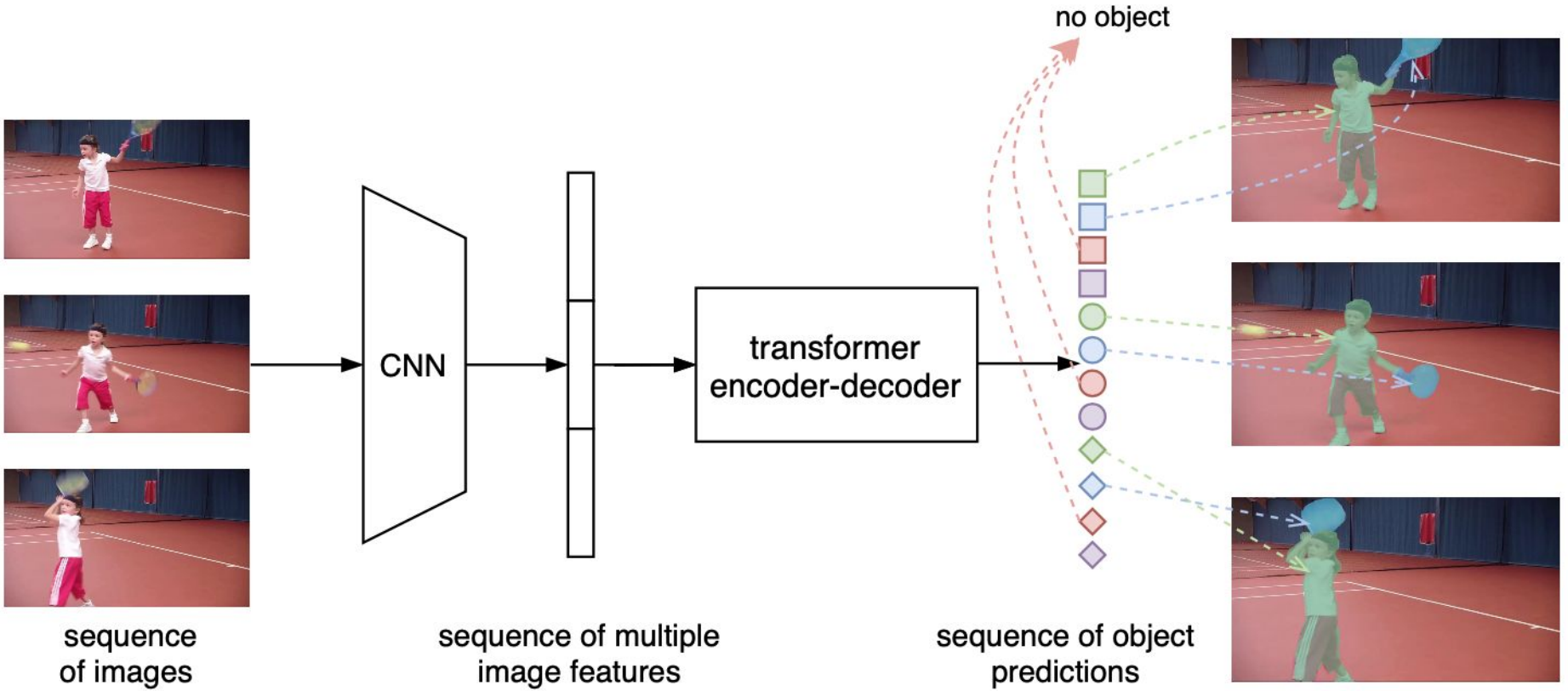
#STCN Cheng, Ho Kei, Yu-Wing Tai, and Chi-Keung Tang. "Rethinking Space-Time Networks with Improved Memory Coverage for Efficient Video Object Segmentation." NeurIPS 2021. [code]



#CFBI Yang, Z., Wei, Y., & Yang, Y. Collaborative video object segmentation by foreground-background integration. ECCV 2020. [code]



#LWL Bhat, G., Lawin, F. J., Danelljan, M., Robinson, A., Felsberg, M., Van Gool, L., & Timofte, R. Learning what to learn for video object segmentation. ECCV 2020. [code]



#VisTR Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., & Xia, H. End-to-end video instance segmentation with transformers. CVPR 2021. [code]