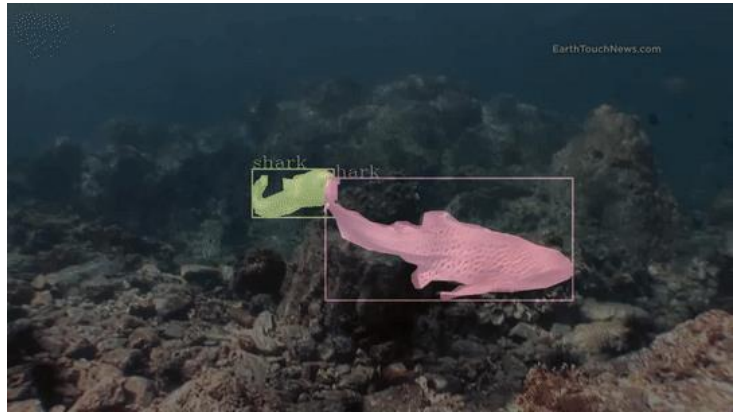
Learning from Video Segmentation Challenge

Ning Xu | Research Scientist, Adobe Research

Outline

- Video Object Segmentation
- Video Instance Segmentation

Video Object Segmentation



Video Instance Segmentation

Video Object Segmentation

- Problem Definition

Initial Object Mask

Automatic Propagation

Full Sequence

Adobe

Video Object Segmentation

- Evaluation Metrics

Region Similarity $\mathcal{J} = \frac{|M \cap G|}{|M \cup G|}$

Contour Accuracy $\mathcal{F} = \frac{2P_c R_c}{P_c + R_c}$

Comparison to VOT

- No-reset setting
- No restriction on model-free
- No restriction on causality
- Short-term clips, but could have reoccurrence objects
- Different evaluation metrics

Application

- Video Editing
- Video Inpainting

Progress in VOS (Before 2016)

- Dataset

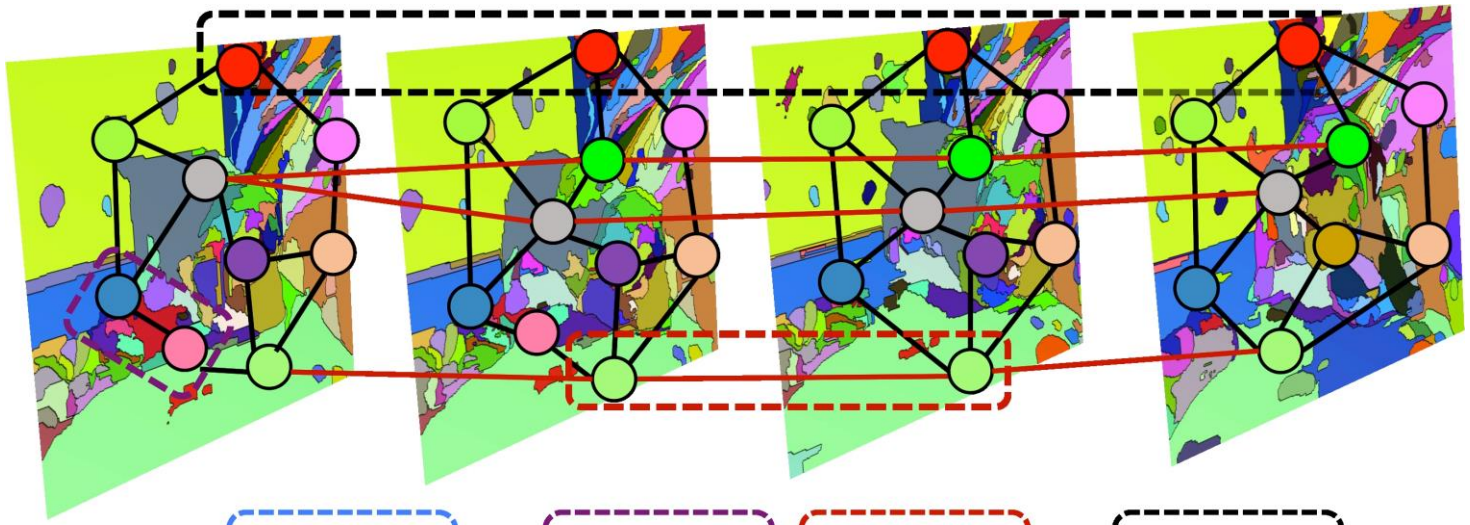
	Videos	Categories	Objects	Annotations	Duration (mins)
SegTrack v2	14	11	24	1475	0.59
JumpCut	22	14	22	6331	3.52
YouTube Objects	96	10	96	1692	9.01
FMBS	59	16	139	1465	7.7

Small scale, low resolution, short term

Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: ICCV (2013)
Fan, Q., Zhong, F., Lischinski, D., Cohen-Or, D., Chen, B.: Jumpcut:non-successive mask transfer and interpolation for video cutout. In: ACM Trans. Graph., 34(6) (2015)
Jain, S.D., Grauman, K.: Supervoxel-consistent foreground propagation in video. In: ECCV (2014)
Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. IEEE transactions on PAMI 36(6), 1187–1200 (2014)

Progress in VOS (Before 2016)

- Methods



Nagaraja, N.S., Schmidt, F.R., Brox, T.: Video segmentation with just a few strokes. In: ICCV. pp. 3235–3243 (2015)
Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: Computer Vision (ICCV), 2013 IEEE International Conference on. pp. 1777–1784. IEEE (20
Jain, S.D., Grauman, K.: Supervoxel-consistent foreground propagation in video. In: ECCV (2014)
Faktor, A., Irani, M.: Video segmentation by non-local consensus voting. In: BMVC (2014)

The DAVIS Dataset

- High Resolution
- Short Term
- Occlusion, Fast Motion, Camera Motion

	Videos	Categories	Objects	Annotations	Duration (mins)
DAVIS 2016	50	-	50	3440	2.88
DAVIS 2017	90	-	205	13543	5.17

Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., SorkineHornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (2016)
Pont-Tuset, J., Perazzi, F., Caelles, S., Arbel´aez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv:1704.00675 (2017)

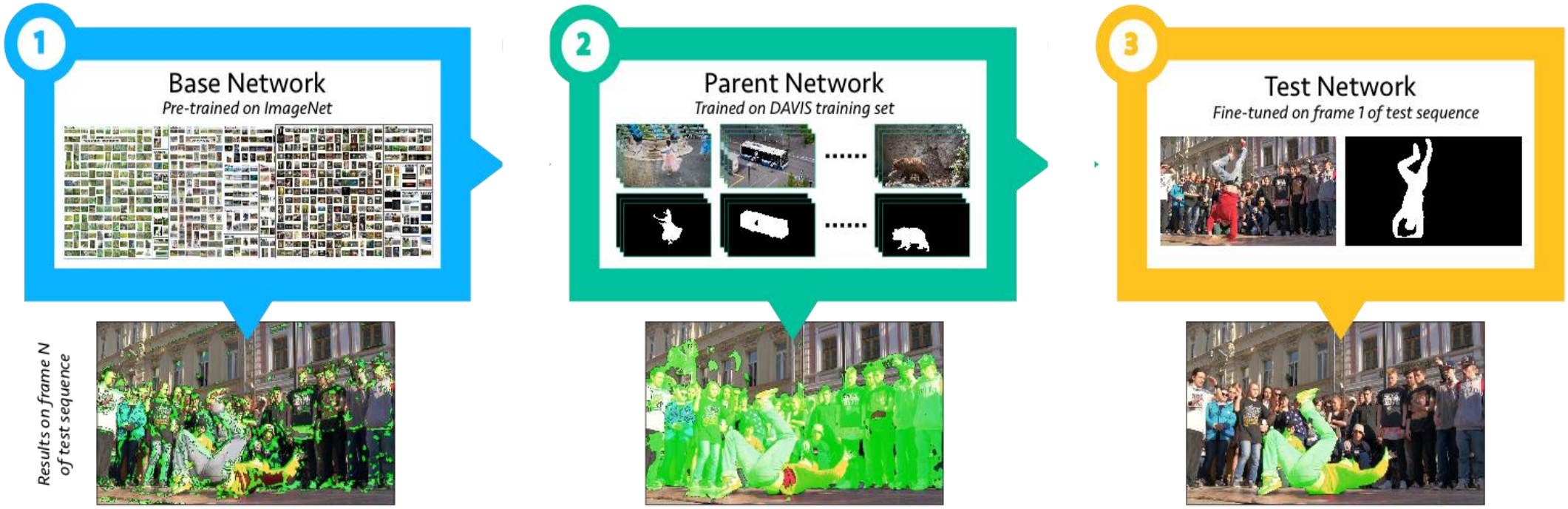
The DAVIS Dataset

- High-Quality Annotation

Adobe

Concurrent Methods

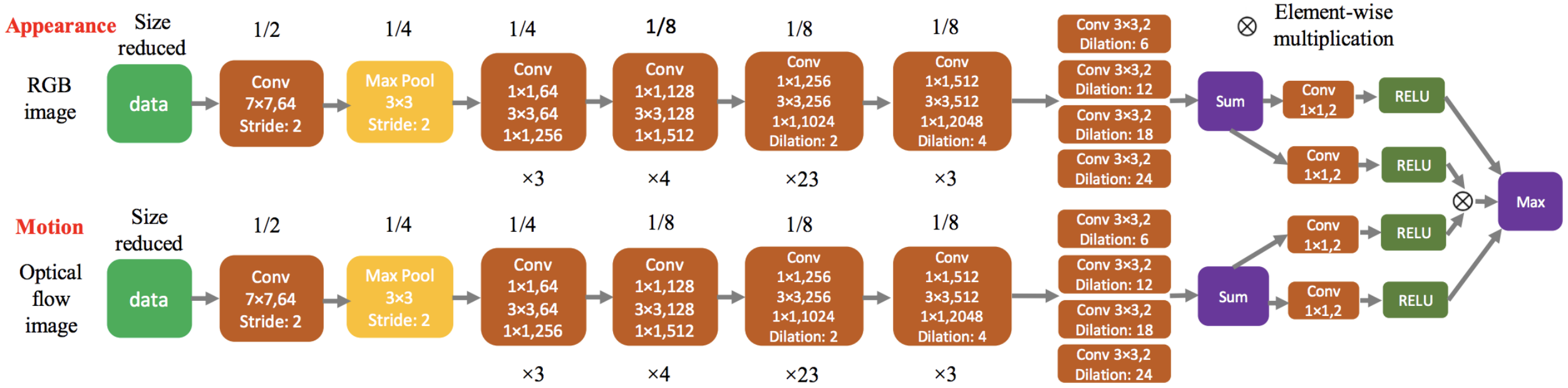
- Online Learning



Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taix´e, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. In: CVPR (2017)
Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., A.Sorkine-Hornung: Learning video object segmentation from static images. In: CVPR (2017)

Concurrent Methods

- Motion or Optical Flow

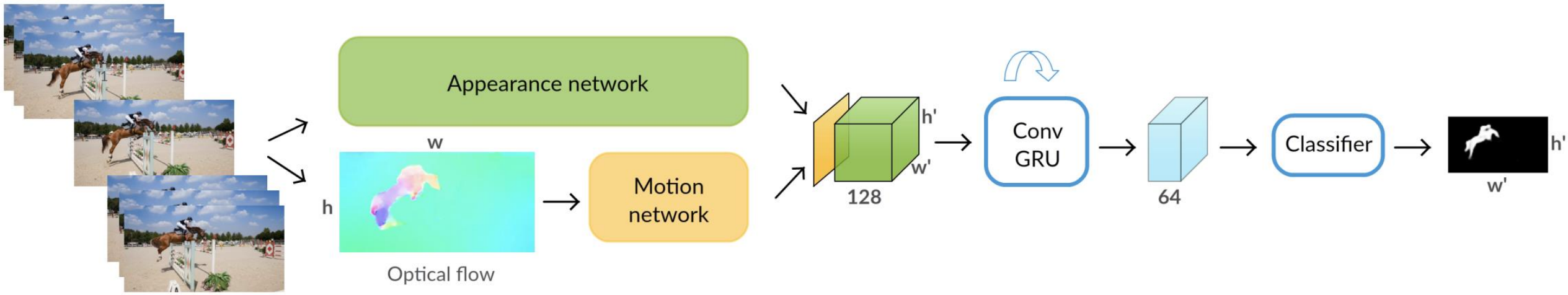


Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H.: Segflow: Joint learning for video object segmentation and optical flow. In: ICCV (2017)

Dutt Jain, S., Xiong, B., Grauman, K.: Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In: CVPR (201

Concurrent Methods

- Memory modules



Tokmakov, P., Alahari, K., Schmid, C.: Learning video object segmentation with visual memory. In: ICCV (2017)

Limitation

- Lack of data covering various motion, reoccurrence, appearance change etc.
- Methods have to use models pretrained on other datasets.
- Synthetic data could not synthesize realistic long-term movement.
- Evaluation can be improved.

Create A Large-Scale Dataset

- Select a comprehensive category list.
  - Animals
    - Eagle, Snail, Ant…
  - Vehicles
    - Airplane, Bicycle, Boat…
  - Humans in different activities
    - Tennis racket, Skating board, Motorcycle…
  - Accessories
    - Eyeglasses, Hat, Bag…
  - Other common objects
    - Potted plant, knife, umbrella…

Create A Large-Scale Dataset

- Collect video clips for each category.
  - Retrieval 100 videos per category from YouTube-8M.
  - Automatic video shot detection.
  - Randomly sample clips with length between 3s to 6s.
  - Manual filtering of bad-quality clips (scene transition, low resolution, blurry, no objects).

Create A Large-Scale Dataset

- Annotation of object masks
  - Up to 5 objects of proper sizes and categories per video clip.
  - 6 fps annotation rate.
  - Boundary tracing instead of polygons.

Previous Datasets

	Videos	Categories	Objects	Annotations	Duration (mins)
SegTrack v2	14	11	24	1475	0.59
JumpCut	22	14	22	6331	3.52
YouTube Objects	96	10	96	1692	9.01
FMBS	59	16	139	1465	7.7
DAVIS 2016	50	-	50	3,440	2.88
DAVIS 2017	90	-	205	13,543	5.17

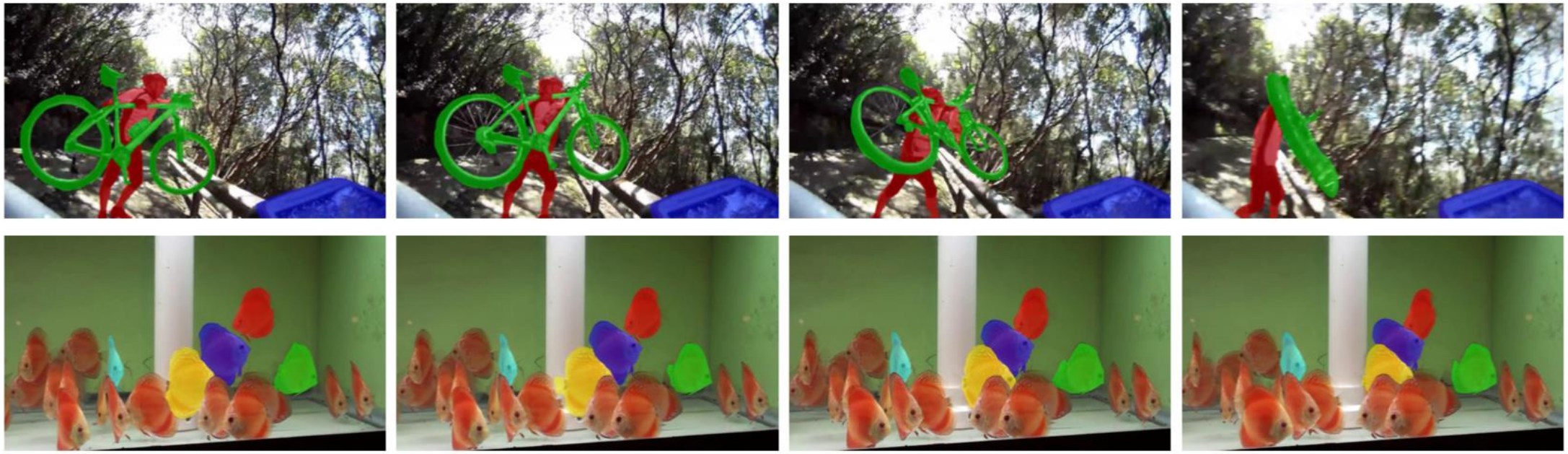
Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S. and Huang, T., 2018. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV* 2018.

Comparison to Previous Datasets

	Videos	Categories	Objects	Annotations	Duration (mins)
SegTrack v2	14	11	24	1475	0.59
JumpCut	22	14	22	6331	3.52
YouTube Objects	96	10	96	1692	9.01
FMBS	59	16	139	1465	7.7
DAVIS 2016	50	-	50	3,440	2.88
DAVIS 2017	90	-	205	13,543	5.17
YouTube VOS	**4453**	**94**	**7,755**	**197,272**	**334.81**

Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S. and Huang, T., 2018. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV* 2018.

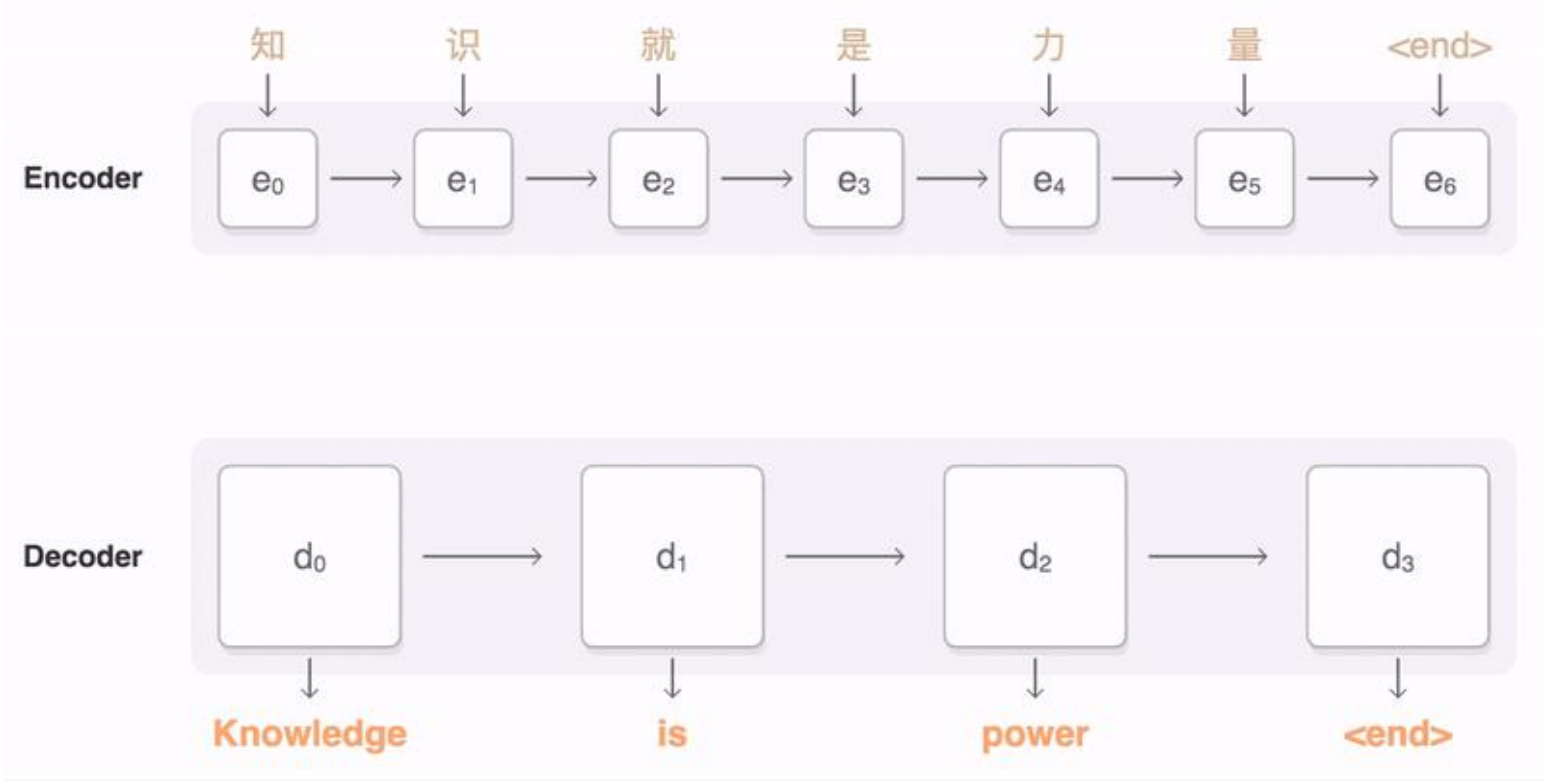
Annotation Samples



Evaluation Setting

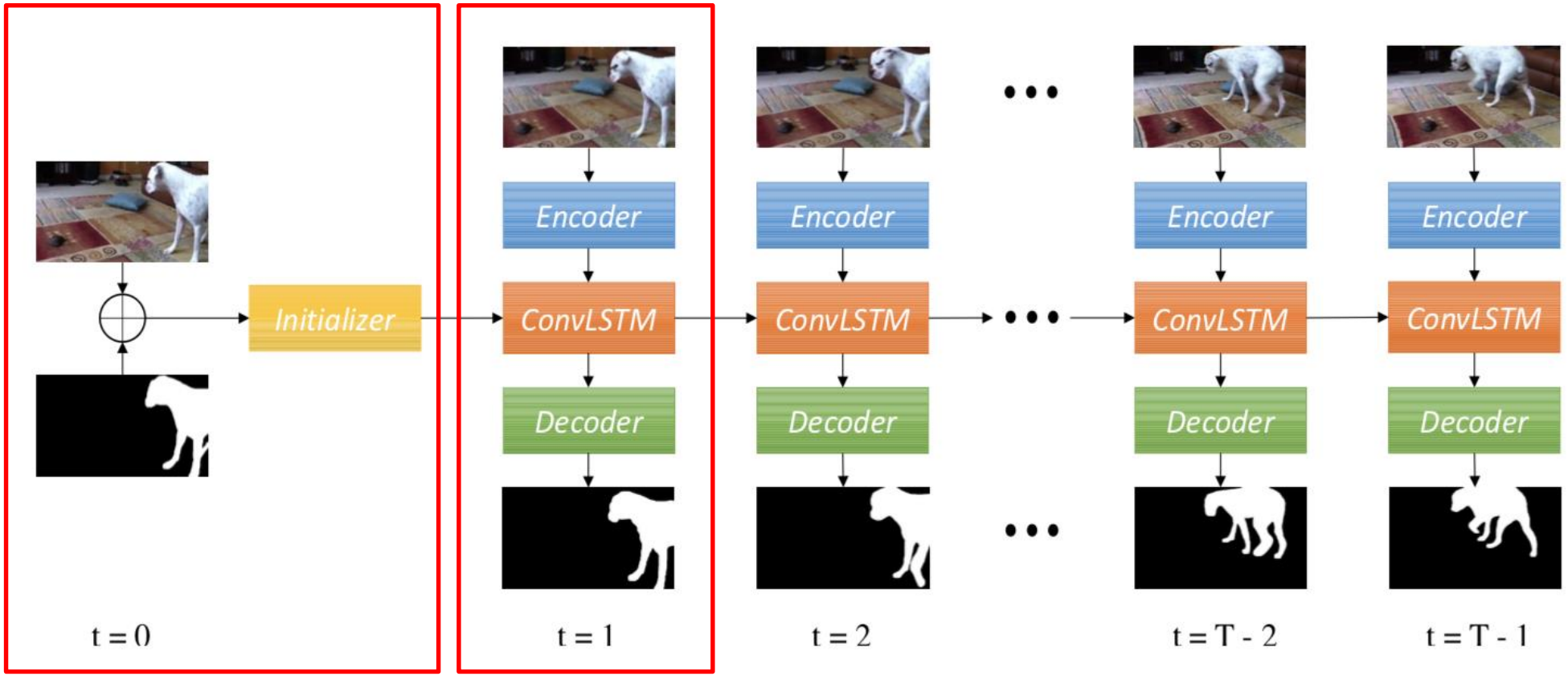
- Split 94 categories into 65 Seen and 29 Unseen.
- Training (3,471), Validation (474) and Test (508)
- Metrics
  - J Seen, J Unseen
  - F Seen, F Unseen
  - Overall

Machine Translation



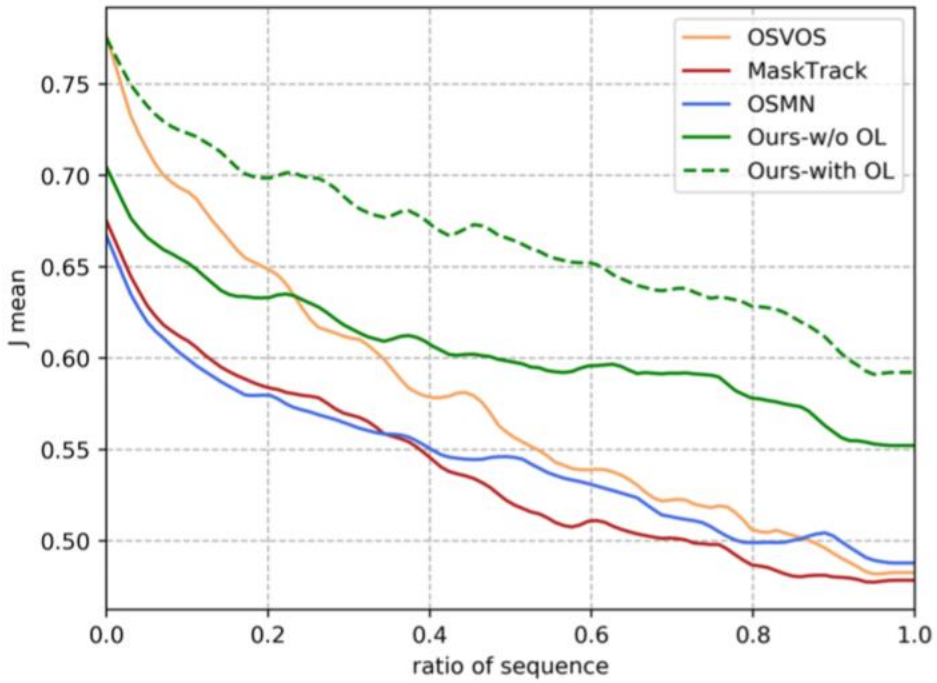
Sutskever, I., Vinyals, O. and Le, Q.V., 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104-3112).

S2S for VOS

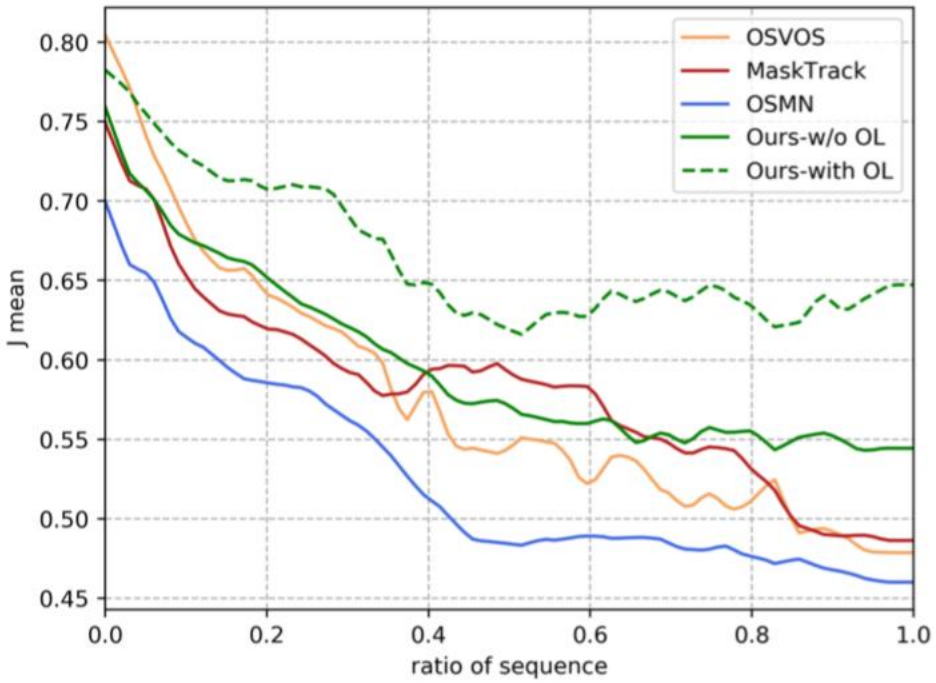


Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S. and Huang, T., 2018. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV* 2018.

Temporal Consistency



(a) Seen categories

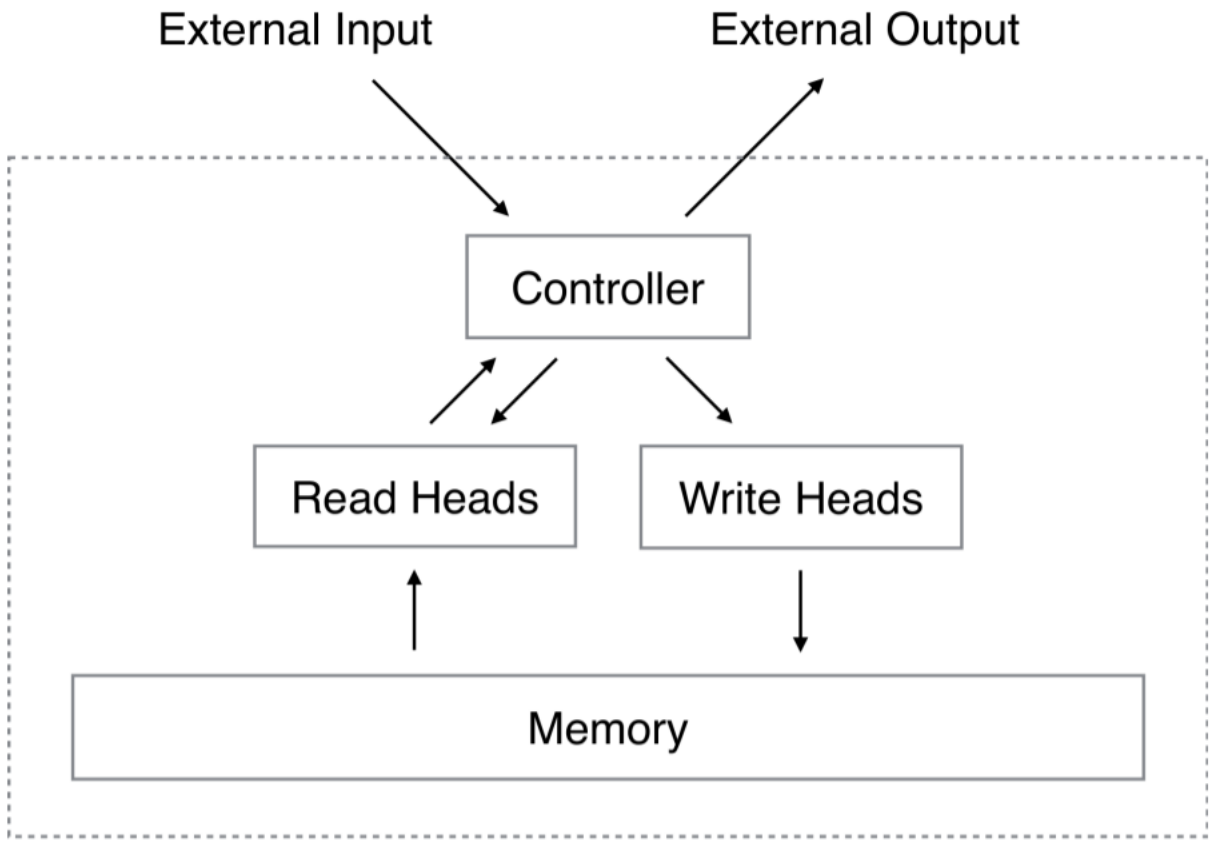


(b) Unseen categories

Quantitative Results

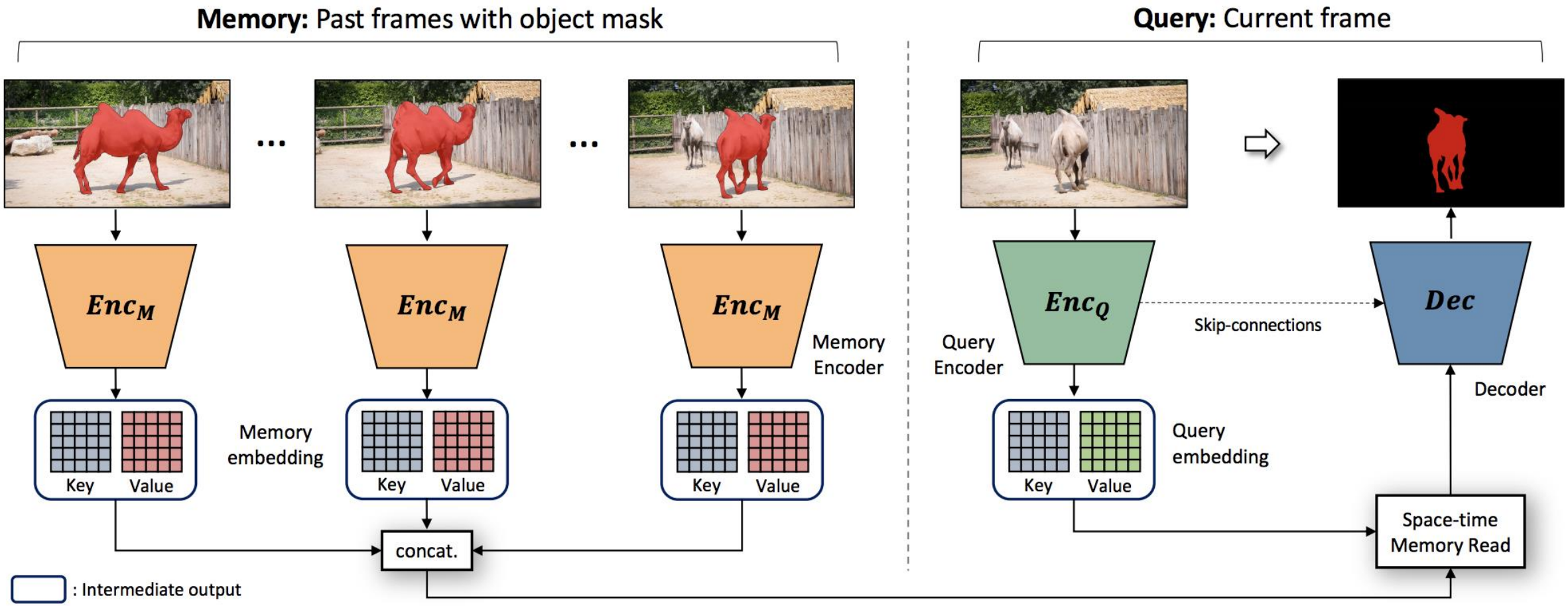
Method	$\mathcal{J}$ seen	$\mathcal{J}$ unseen	$\mathcal{F}$ seen	$\mathcal{F}$ unseen	Overall	Speed (s/frame)
OSVOS [7]	59.8%	54.2%	60.5%	60.7%	58.8%	10
MaskTrack [8]	59.9%	45.0%	59.5%	47.9%	53.1%	12
OSMN [9]	60.0%	40.6%	60.1%	44.0%	51.2%	**0.14**
OnAVOS [35]	60.1%	46.6%	62.7%	51.4%	55.2%	13
S2S (w/o OL) [34]	66.7%	48.2%	65.5%	50.3%	57.6%	0.16
S2S (with OL) [34]	**71.0%**	**55.5%**	**70.0%**	**61.2%**	**64.4%**	9

Neural Turing Machine



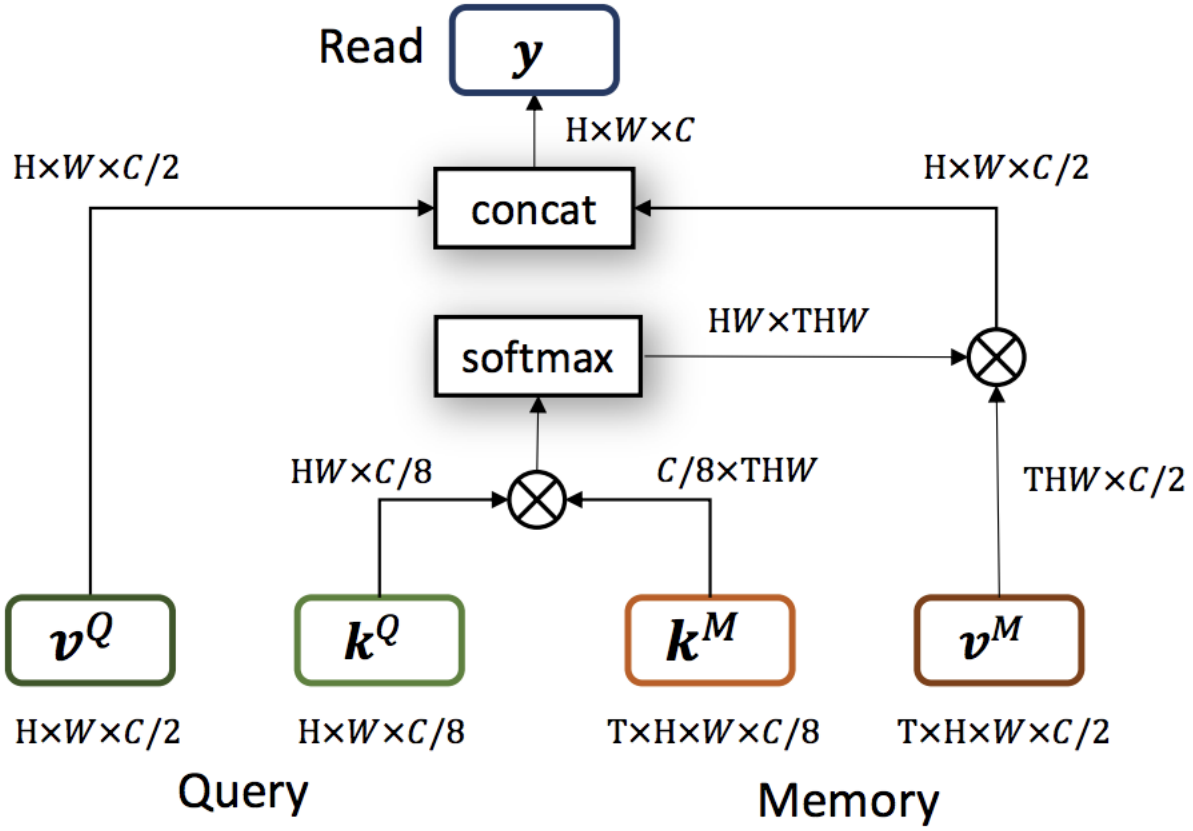
Graves, A., Wayne, G. and Danihelka, I., 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401*.

NTM for VOS



Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S. and Huang, T., 2018. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV* 2018.

Space-Time Memory Read



Wang, X., Girshick, R., Gupta, A. and He, K., 2018. Non-local neural networks. In *CVPR 2018* (pp. 7794-7803).

Quantitative Results

	Overall	Seen		Unseen	
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
OSMN [40]	51.2	60.0	60.1	40.6	44.0
MSK [26]	53.1	59.9	59.5	45.0	47.9
RGMP [24]	53.8	59.5	-	45.2	-
OnAVOS [34]	55.2	60.1	62.7	46.6	51.4
RVOS [32]	56.8	63.6	67.2	45.5	51.0
OSVOS [2]	58.8	59.8	60.5	54.2	60.7
S2S [38]	64.4	71.0	70.0	55.5	61.2
A-GAME [13]	66.1	67.8	-	60.8	-
PreMVOS [20]	66.9	71.4	75.9	56.5	63.7
BoLTVOS [35]	71.1	71.6	-	64.3	-
Ours	**79.4**	**79.7**	**84.2**	**72.8**	**80.9**

Qualitative Results

Effectiveness of Large-Scale Datasets

- Pretraining on images

Variants	Youtube-VOS	DAVIS-2017	
	Overall	$\mathcal{J}$	$\mathcal{F}$
Pre-training only	69.1	57.9	62.1
Main-training only	68.2	38.1	47.9
Full training	**79.4**	69.2	74.0
Cross validation	56.3	**78.6**	**83.5**

Effectiveness of Large-Scale Datasets

- Pretraining on images
- Finetuning on video

Variants	Youtube-VOS	DAVIS-2017	
	Overall	$\mathcal{J}$	$\mathcal{F}$
Pre-training only	69.1	57.9	62.1
Main-training only	68.2	38.1	47.9
Full training	**79.4**	69.2	74.0
Cross validation	56.3	**78.6**	**83.5**

Effectiveness of Large-Scale Datasets

- Pretraining on images
- Finetuning on video

Variants	Youtube-VOS	DAVIS-2017	
	Overall	$\mathcal{J}$	$\mathcal{F}$
Pre-training only	69.1	57.9	62.1
Main-training only	68.2	38.1	47.9
Full training	**79.4**	69.2	74.0
Cross validation	56.3	**78.6**	**83.5**

Effectiveness of Large-Scale Datasets

- Pretraining on images
- Finetuning on video
- Large-scale video dataset

Variants	Youtube-VOS	DAVIS-2017	
	Overall	$\mathcal{J}$	$\mathcal{F}$
Pre-training only	69.1	57.9	62.1
Main-training only	68.2	38.1	47.9
Full training	**79.4**	69.2	74.0
Cross validation	56.3	**78.6**	**83.5**

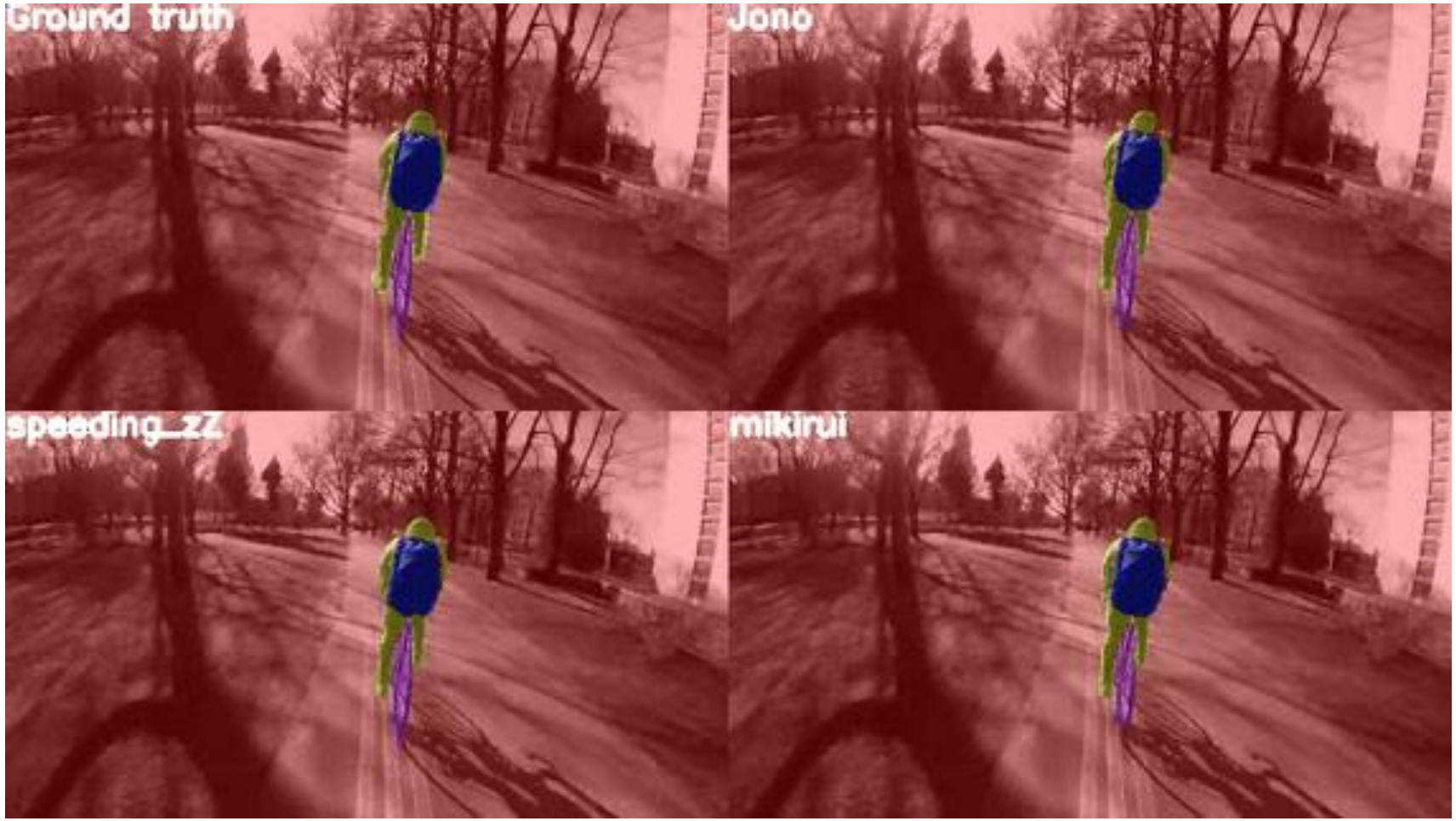
1st Large-Scale VOS Challenge

- Around 100 registered participants
- 19 teams/users submitted to validation server
- 10 teams/users submitted to test server

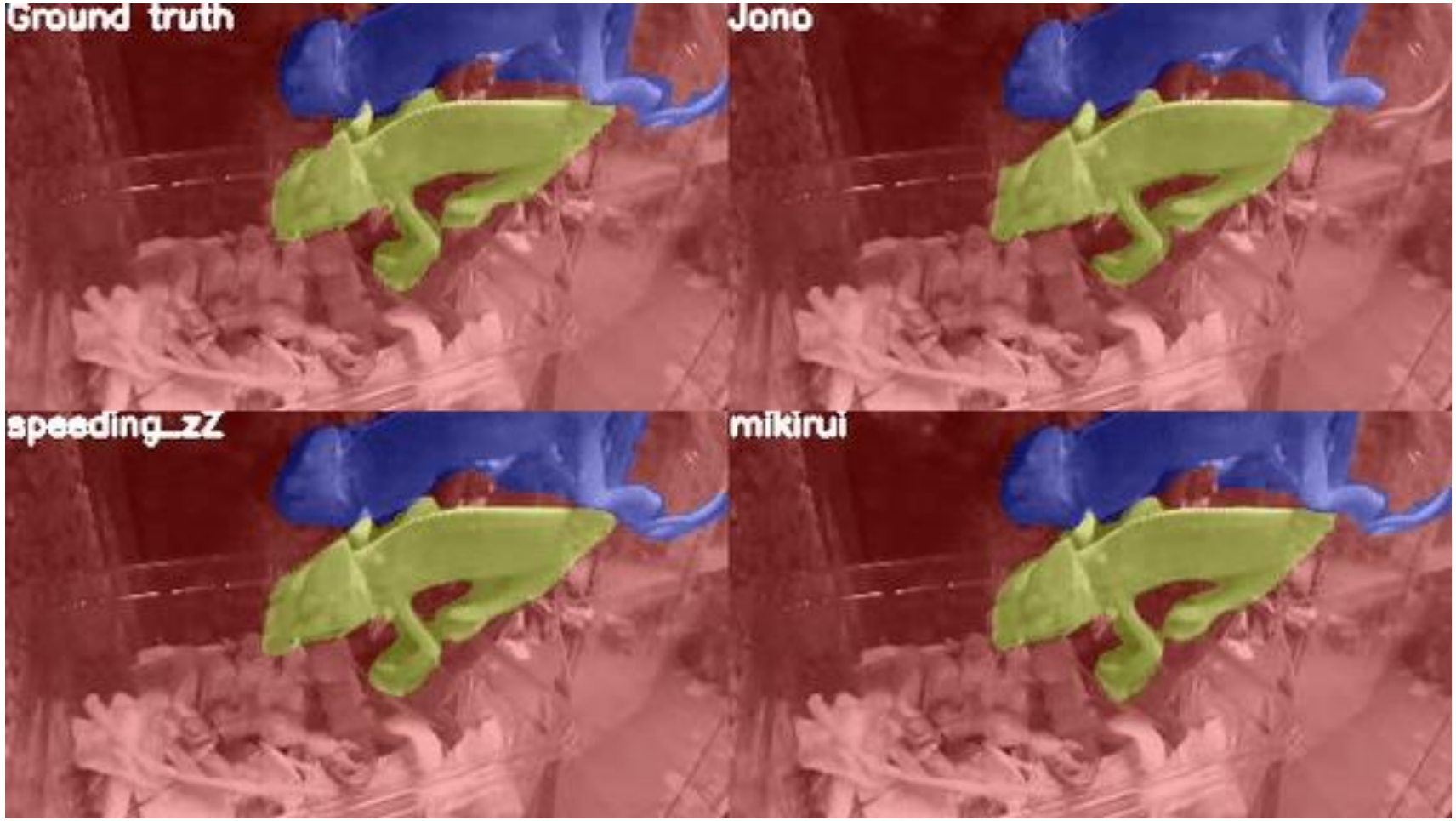
Top Results

#	User	Organization	Overall	J_seen	J_unseen	F_seen	F_unseen
1	Jono	RWTH Aachen University	0.722 (1)	0.737 (1)	0.648 (2)	0.778 (1)	0.725 (2)
2	speeding_zZ	Horizon Robotics	0.720 (2)	0.725 (3)	0.663 (1)	0.752 (3)	0.741 (1)
3	mikirui	CUHK & Tencent YouTu X-Lab	0.699 (3)	0.736 (2)	0.621 (4)	0.755 (2)	0.684 (4)

Fast Motion



Similar Objects



Small Objects

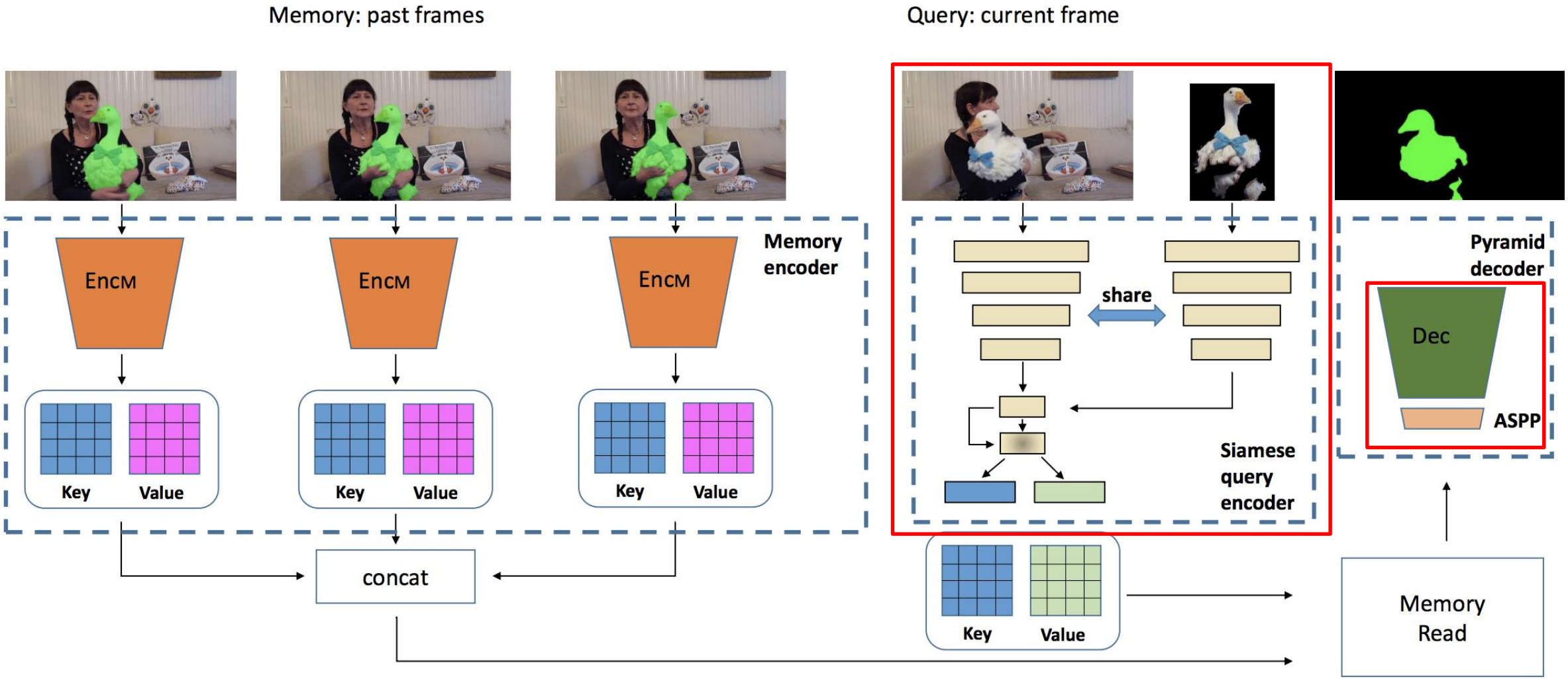
2nd Large-Scale VOS Challenge

- More annotated objects in the test set
- 250+ teams/users registered
- 38 teams/users submitted to validation server
- 12 teams/users submitted to test server

Top Results

Team Name	Overall	J_seen	J_unseen	F_seen	F_unseen	Ranking
zszhou	0.818 (1)	0.807 (1)	0.773 (2)	0.847 (1)	0.847 (2)	1
theodoruszq	0.817 (2)	0.800 (2)	0.779 (1)	0.833 (2)	0.855 (1)	2
zxyang1996	0.804 (3)	0.794 (3)	0.759 (4)	0.833 (3)	0.831 (4)	3
swoh	0.802 (4)	0.788 (4)	0.759 (3)	0.825 (4)	0.835 (3)	4
youtube_test	0.791 (5)	0.779 (5)	0.747 (5)	0.815 (5)	0.822 (5)	5
Jono	0.714 (7)	0.703 (10)	0.680 (7)	0.736 (10)	0.740 (8)	6
andr345	0.710 (8)	0.699 (11)	0.667 (8)	0.732 (11)	0.740 (7)	7

1st Ranked Method



Zhou, Z., Ren, L., Xiong, P., Ji, Y., Wang, P., Fan, H. and Liu, S., Enhanced Memory Network for Video Segmentation. In ICCVW 2019.

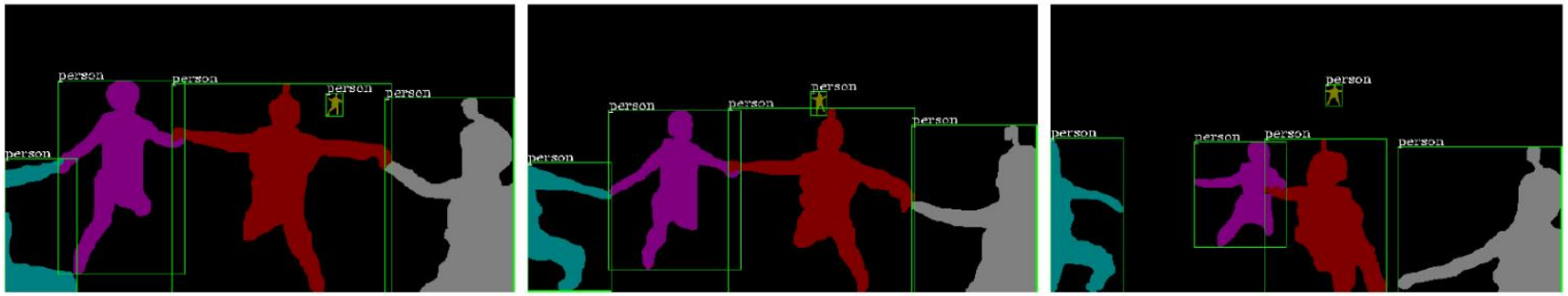
Summary of VOS

- Large-scale dataset is important.
- Spatial-Temporal consistency needs to be leveraged.
  - Borrow ideas from related fields such as NLP.
- Both pretraining on images and finetuning on videos are useful.

Video Instance Segmentation

- Simultaneous detection, segmentation and tracking across frames.
  - Extend image instance segmentation to video.

Video frames



Video instance annotations

Evaluation Metrics

- Temporal Intersection over Union (IoU)

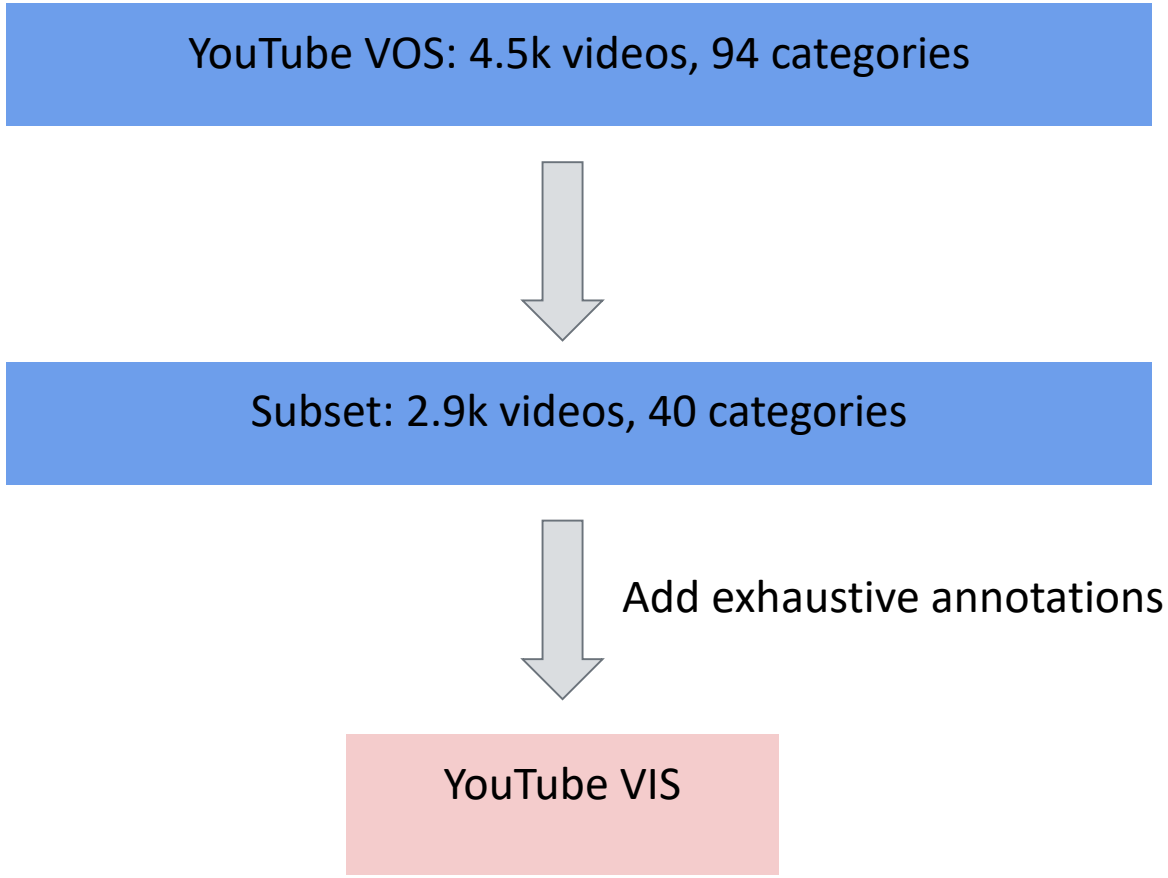
$$\mathrm{IoU}(i,j) = \frac{\Sigma_{t=1}^{T} |\mathbf{m}_t^i \cap \tilde{\mathbf{m}}_t^j|}{\Sigma_{t=1}^{T} |\mathbf{m}_t^i \cup \tilde{\mathbf{m}}_t^j|}$$

- Average Prevision (AP)
- Average Recall (AR)

Comparison to VOS

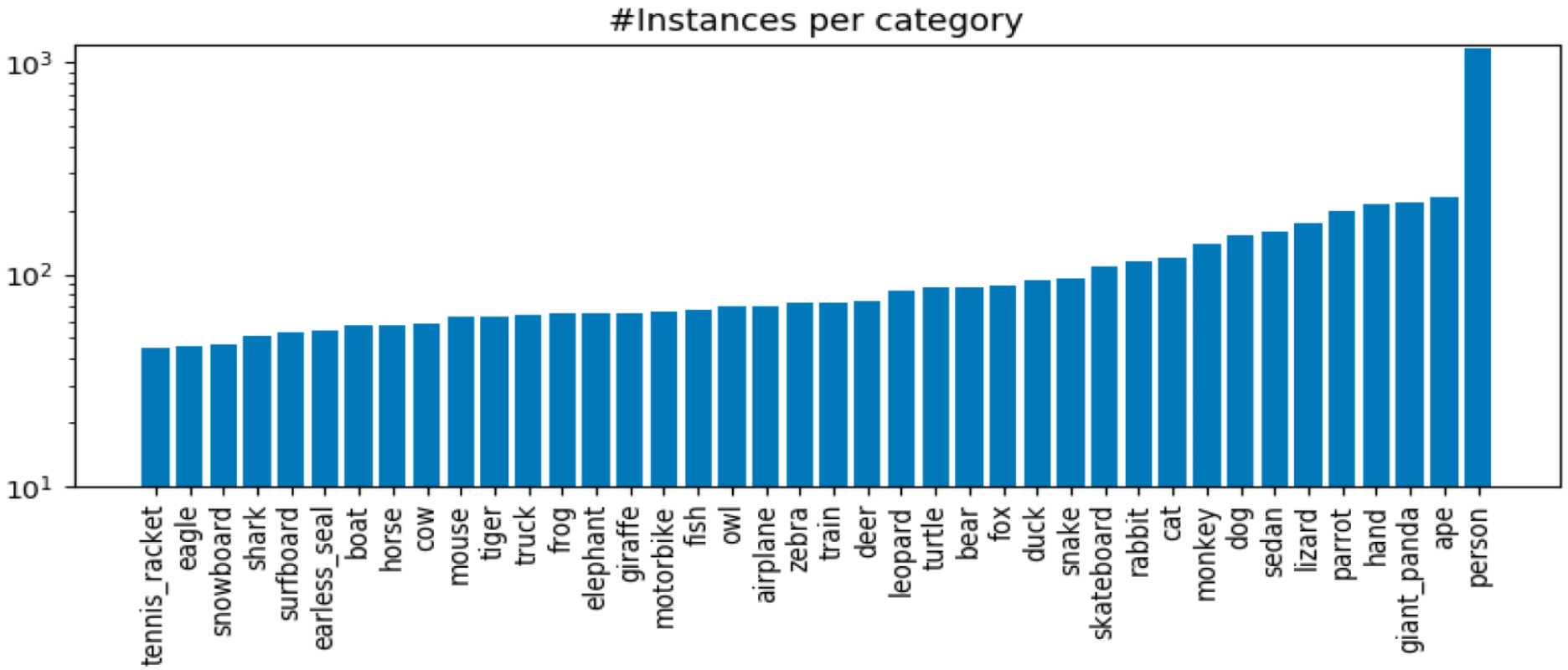
Tasks	First-frame mask	Recognize categories	Exhaustive annotations	Multiple object
unsupervised VOS	No	No	No	No
semi-supervised VOS	Yes	No	No	Yes
VIS	No	Yes	Yes	Yes

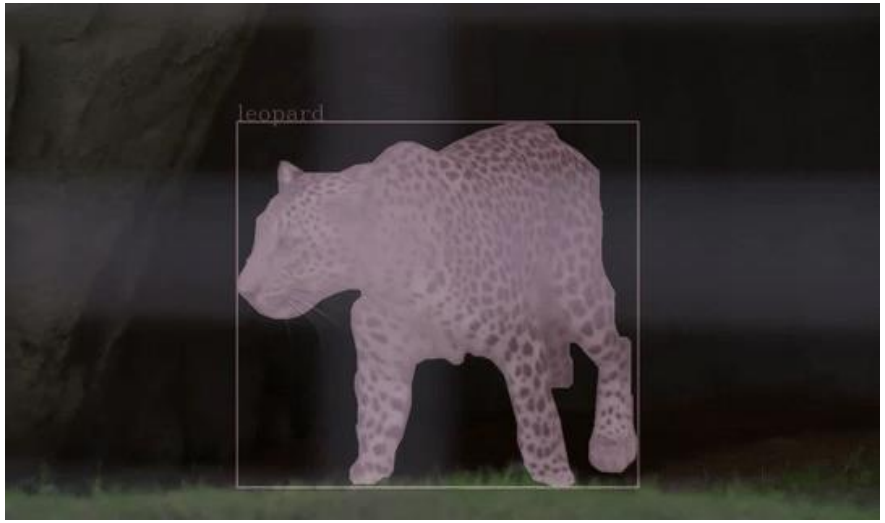
YouTube-VIS Dataset

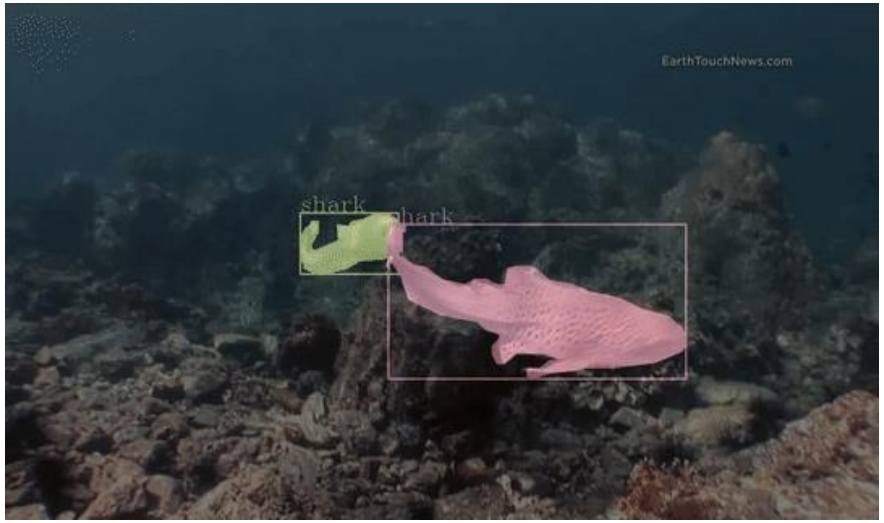


Yang, L., Fan, Y. and Xu, N., Video Instance Segmentation. In ICCV 2019.

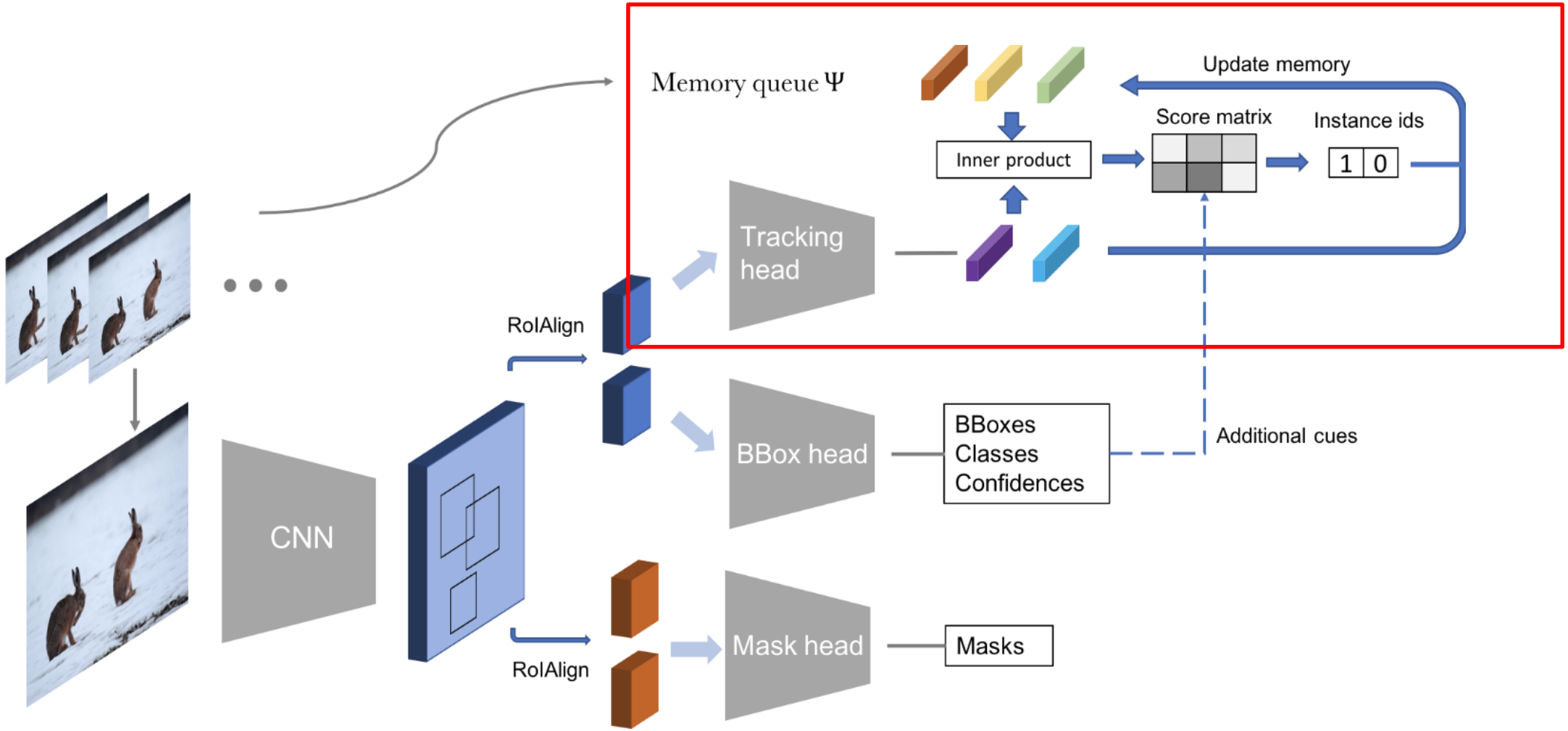
Dataset Statistics







MaskTrack R-CNN



Yang, L., Fan, Y. and Xu, N., Video Instance Segmentation. In ICCV 2019.

Quantitative Results

Methods		validation set					test set				
		AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
Mask propagation	OSMN [36]	23.4	36.5	25.7	28.9	31.1	27.3	44.4	28.0	28.8	34.0
	FEELVOS [31]	26.9	42.0	29.7	29.9	33.4	29.6	45.4	30.7	33.4	36.8
Track-by-detect	IoUTracker+	23.6	39.2	25.5	26.2	30.9	25.2	41.9	26.2	28.7	33.7
	OSMN [36]	27.5	45.1	29.1	28.6	33.1	27.3	44.4	28.0	28.8	34.0
	DeepSORT [33]	26.1	42.9	26.1	27.8	31.3	27.2	44.0	29.2	29.1	33.3
	SeqTracker	27.5	45.7	28.7	29.7	32.5	29.5	48.1	31.2	32.0	34.5
	MaskTrack R-CNN	**30.3**	**51.1**	**32.6**	**31.0**	**35.5**	**32.3**	**53.6**	**34.2**	**33.6**	**37.3**

Qualitative Results

1st Large-Scale VIS Challenge

- 180+ teams/users registered
- 26 teams/users submitted to validation server
- 19 teams/users submitted to test server

Adobe

Leaderboard

Team Name	mAP	AP50	AP75	AR1	AR10	Ranking
Jono	0.467 (1)	0.697 (1)	0.509 (1)	0.462 (1)	0.537 (2)	1
foolwood	0.457 (2)	0.674 (3)	0.490 (3)	0.435 (5)	0.507 (4)	2
bellejuillet	0.450 (3)	0.636 (6)	0.502 (2)	0.447 (3)	0.503 (5)	3
linhj	0.449 (4)	0.665 (4)	0.486 (5)	0.453 (2)	0.538 (1)	4
mingmingdiii	0.444 (5)	0.684 (2)	0.487 (4)	0.436 (4)	0.508 (3)	5
xiAaonice	0.400 (7)	0.578 (10)	0.449 (7)	0.396 (10)	0.452 (10)	6
guwop	0.400 (8)	0.608 (8)	0.439 (9)	0.412 (8)	0.491 (6)	7
exing	0.397 (9)	0.621 (7)	0.426 (10)	0.414 (6)	0.461 (9)	8

Top Ranked Method

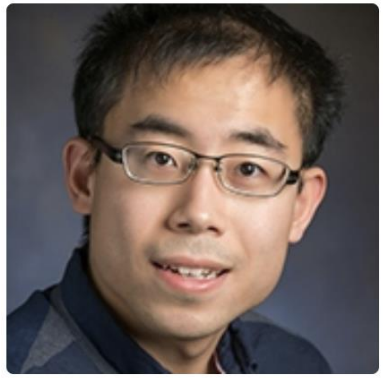
- Detection
  - Mask R-CNN trained on both video and image datasets (COCO, OpenImages)
- Classification
  - Pretrained image classification networks (ResNeXt-101)
- Segmentation
  - Separated segmentation network (DeepLabV3+)
- Tracking
  - Optical flow (PWC-Net) for short tracklets
  - ReID for long tracklets

Summary of VIS

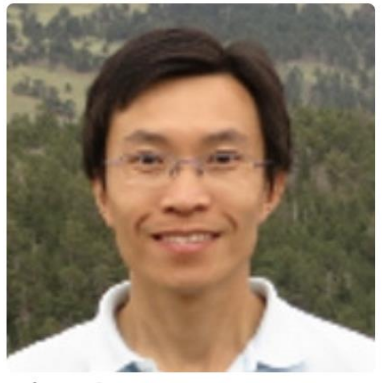
- A more complex task than VOS and VOT.
- Results have large room to be improved.
- Need more principled methods and frameworks.

Acknowledgement

Linjie Yang



Yuchen Fan



Jianchao Yang

Thomas Huang

Seoung Wug Oh

Joon-Young Lee

Scott Cohen

Brian Price


Adobe