Multi-object video segmentation semi-supervised task

Haller Emanuela
ehaller@bitdefender.com

04 June 2019

DAVIS 2017

Bitdefender

Measure	PReMVOS	CINM	Lucid	FEELVOS	OSVOS-S	VOSwL	RGMP	OnAVOS	OSVOS	RVOS	FAVOS	SIAMMASK	OSMN
J&F Mean ↑	**71.6**	67.5	66.6	57.8	57.5	-	52.8	52.8	50.9	50.3	43.6	43.2	41.3
J Mean ↑	**67.5**	64.5	63.4	55.1	52.9	-	51.3	49.9	47.0	47.9	42.9	40.6	37.7
J Recall ↑	**76.8**	73.8	74.0	62.6	60.2	-	59.0	54.3	52.1	54.4	48.1	44.5	38.9
J Decay ↓	21.7	20.0	19.5	29.8	24.1	-	34.3	23.0	19.2	35.7	**18.1**	21.9	19.0
F Mean ↑	**75.8**	70.5	69.9	60.4	62.1	-	54.4	55.7	54.8	52.6	44.2	45.8	44.9
F Recall ↑	**84.3**	79.6	80.1	68.5	70.5	-	61.9	60.3	59.7	61.7	51.1	45.3	47.4
F Decay ↓	20.6	20.0	19.5	33.5	21.9	-	37.2	23.4	19.8	36.7	19.8	22.4	**17.4**

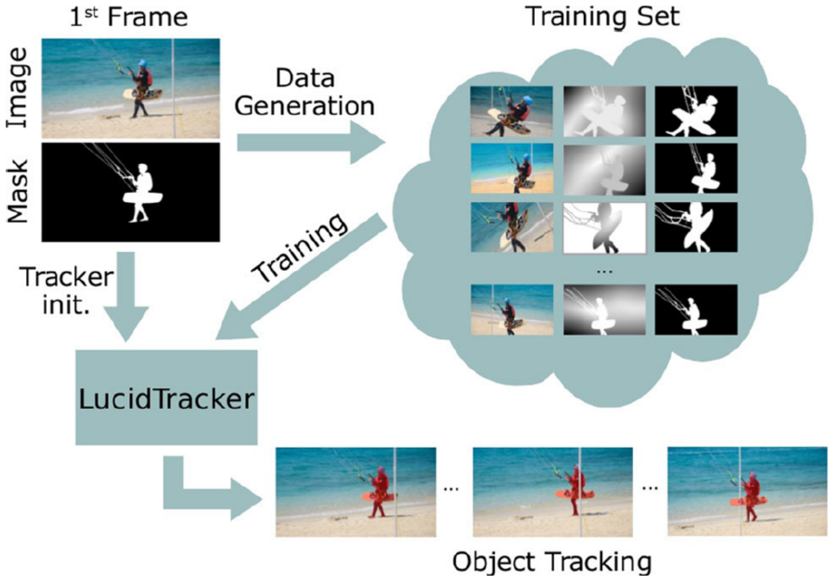
- PReMVOS* - ACCV 2018
- Lucid** - IJCV 2018

*Luiten et al. [2018], **Khoreva et al. [2018]

Lucid Data Dreaming for Video Object Segmentation

- reduce the necessity for large volumes of training data
- "lucid dreaming" - generate plausible future frames
- 2.5k in-domain samples better than thousands of samples from close-by domains

Lucid Data Dreaming for Video Object Segmentation



Lucid - Architecture (single object)

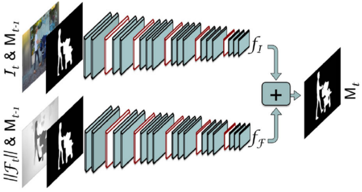
Two streams architecture

- $M_t = 0.5 \cdot f_I(I_t, w(M_{t-1}, F_t)) + 0.5 \cdot f_F(\|F_t\|, w(M_{t-1}, F_t))$

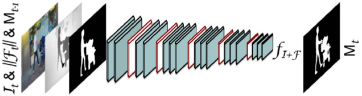
One stream architecture

- $M_t = f_{I+f}(I_t, \|F_t\|, w(M_{t-1}, F_t))$

- F_t - optical flow
- $\|F_t\|$ - optical flow magnitude
- M_t - object mask
- I_t - frame
- $w(M_{t-1}, F_t)$ - warped object mask



(a) Two streams architecture, where image $\mathcal{I}_t$ and optical flow information $\|\mathcal{F}_t\|$ are used to update mask M_{t-1} into M_t. See equation 1



(b) One stream architecture, where 5 input channels: image $\mathcal{I}_t$, optical flow information $\|\mathcal{F}_t\|$ and mask M_{t-1} are used to estimate mask M_t

Fig. 3 Overview of the proposed one and two streams architectures. See Sect. 3.1

Lucid - Architecture

- DeepLabv2* with VGG base network

*Chen et al. [2017]

Lucid - Optical flow

- FlowNet 2.0*
- $\|F_t\|$
  - subtract the median motion of each frame
  - average the magnitude of the forward and backward flow
  - scale the values, per-frame, to [0,255]

*Ilg et al. [2017]

Lucid - Architecture (multiple objects)

- $M_t = f_{I+F+S}(I_t, \|F_t\|, S_t, w(M^1_{t-1}, F_t), ..., w(M^N_{t-1}, F_t))$
  - M^i_t - mask of object i, in frame t
  - S_t - semantic segmentation
- ensemble
  - $M_t = 0.25 \cdot (f_I + f_{I+S} + f_{I+F} + f_{I+F+S})$

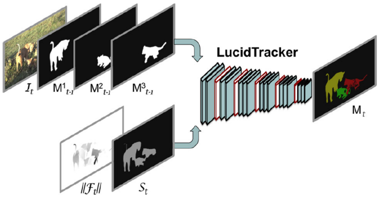
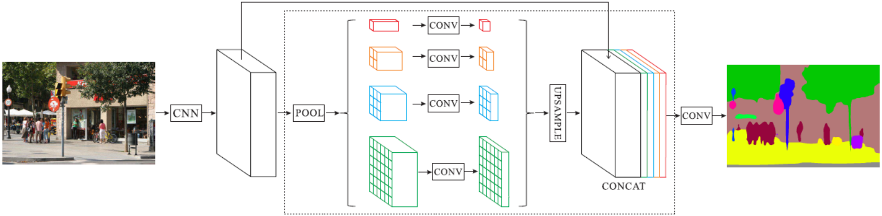


Fig. 4 Extension of LucidTracker to multiple objects. The previous frame mask for each object is provided in a separate channel. We additionally explore using optical flow $\mathcal{F}$ and semantic segmentation $\mathcal{S}$ as additional inputs. See Sect. 3.1

Lucid - Semantic Labels

- PSPNet* - Pyramid Scene Parsing Network
- trained on Pascal VOC12



(a) Input Image (b) Feature Map (c) Pyramid Pooling Module (d) Final Prediction

Figure 3. Overview of our proposed PSPNet. Given an input image (a), we first use CNN to get the feature map of the last convolutional layer (b), then a pyramid parsing module is applied to harvest different sub-region representations, followed by upsampling and concatenation layers to form the final feature representation, which carries both local and global context information in (c). Finally, the representation is fed into a convolution layer to get the final per-pixel prediction (d).

*Zhao et al. [2017]

Lucid - Temporal Coherency

- improve accuracy of warping step

- remove inconsistencies between M_{t-1} and M_{t-2}
- $\Rightarrow \widetilde{M}_{t-1}$ - pruned mask
- $w(\widetilde{M}_{t-1}, F_t)$

- applied during inference
- mitigates error propagation issues

Lucid - Post-processing

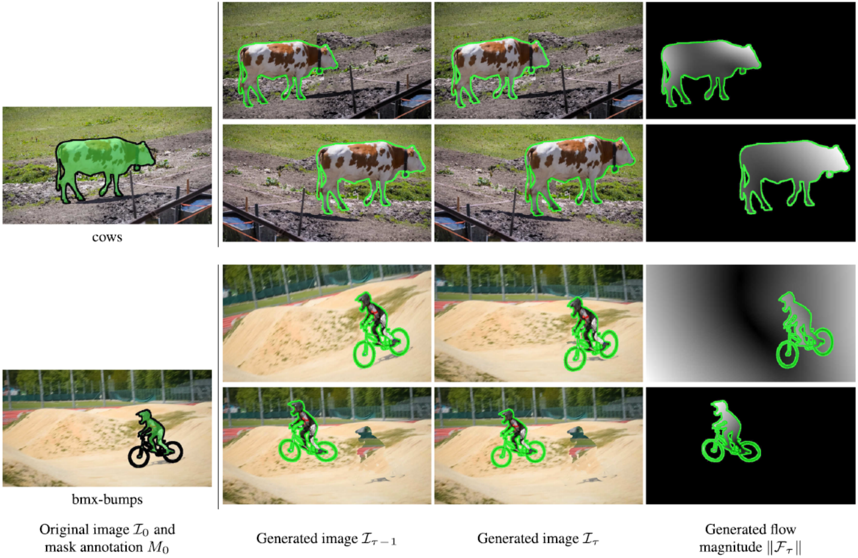
- DenseCRF*

*Krähenbühl and Koltun [2011]

Lucid Data Dreaming

- synthesizing samples from the provided annotated frame
- pairs of images (I_t, I_{t-1})
- "dream" the desired data
- $\approx$ 2500 pairs per annotation

Lucid Data Dreaming



Lucid Data Dreaming

- traditional small perturbations are insufficient to cover the expected variations
- targeted changes:
  - illumination
  - deformation
  - translation
  - occlusions
  - (different points of view)
  - dynamic background
- steps:
  - cut-out the foreground
  - inpaint the background
  - perturb both foreground and background
  - recompose the scene
- $\Rightarrow (I_{t-1}, I_t)$, (M_{t-1}, M_t) and F_t

Lucid Data Dreaming

- Illumination changes
  - randomly altering saturation S and value V (HSV color space)
  - $x' = a \cdot x^b + c$, $a \in 1 \pm 0.05, b \in 1 \pm 0.3, c \in \pm 0.07$
- Fg/Bg Split
  - remove foreground object
  - inpaint the cut-out area *
- Object Motion
  - simulate motion and shape deformation
  - random translation
    - for I_{t-1} - object placed at any location - uniform distribution
    - for I_t - translation of $\pm 10\%$ w.r.t. I_{t-1}
  - random rotation $\pm 30^0$
  - random scaling $\pm 15\%$
  - thin-plate splines deformations $\pm 10\%$ **

*Criminisi et al. [2004], **Bookstein [1989]

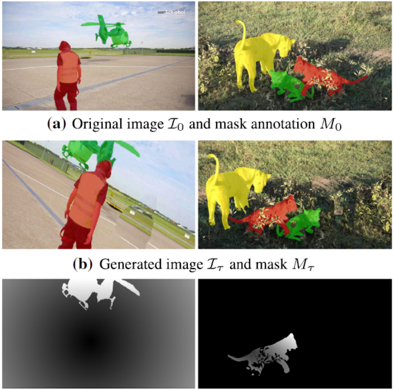
Lucid Data Dreaming

- Camera motion
  - affine deformation - simulate camera view changes
  - random translation
    - I_{t-1} - uniform distribution
    - I_t - s.t. $\pm 10\%$ w.r.t. I_{t-1}
  - random rotation $\pm 30^0$
  - random scaling $\pm 15\%$
- Fg/Bg Merge
  - blend the perturbed foreground with the perturbed background
  - Poisson matting *

*Sun et al. [2004]

Lucid Data Dreaming - multiple objects

- independent transformations for each object
- choose random depth ordering
- $\Rightarrow$ both partial and full occlusions



(a) Original image $\mathcal{I}_0$ and mask annotation M_0

(b) Generated image $\mathcal{I}_\tau$ and mask M_τ

(c) Generated flow magnitude $||\mathcal{F}_\tau||$

Fig. 6 Lucid data dreaming examples with multiple objects. From one annotated frame we generate a plausible future video frame ($\mathcal{I}_\tau$), with known optical flow ($\mathcal{F}_\tau$) and mask (M_τ)

Lucid - Training

- main: pretrained on ImageNet
- semantic segmentation: pretrained on PascalVOC
- 40k iterations per-video (160 epochs)

Lucid - Quantitative Results

Table 11 Comparison of video object segmentation results on DAVIS$_{17}$, test-dev set. Our LucidTracker shows top performance

Method	DAVIS$_{17}$, test-dev set							
	Rank	Global mean ↑	Region, J			Boundary, F		
			Mean ↑	Recall ↑	Decay ↓	Mean ↑	Recall ↑	Decay ↓
sidc	10	45.8	43.9	51.5	34.3	47.8	53.6	36.9
YXLKJ	9	49.6	46.1	49.1	22.7	53.0	56.5	22.3
haamooon (Shaban et al. 2017)	8	51.3	48.8	56.9	**12.2**	53.8	61.3	**11.8**
Fromandtozh (Zhao 2017)	7	55.2	52.4	58.4	18.1	57.9	66.1	20.0
ilanv (Sharir et al. 2017)	6	55.8	51.9	55.7	17.6	59.8	65.8	18.9
voigtlaender (Voigtlaender and Leibe 2017a)	5	56.5	53.4	57.8	19.9	59.6	65.4	19.0
lalalafine123	4	57.4	54.5	61.3	24.4	60.2	68.8	24.6
wangzhe	3	57.7	55.6	63.2	31.7	59.8	66.7	37.1
lixx (Li et al. 2017)	2	66.1	**64.4**	73.5	24.5	67.8	75.6	27.1
LucidTracker	1	**66.6**	63.4	**73.9**	19.5	**69.9**	**80.1**	19.4

Bold are the best numbers overall

Lucid - Quantitative Results

Bitdefender

Table 12 Comparison of video object segmentation results on DAVIS_{17}, test-challenge set. Our LucidTracker shows competitive performance, holding the second place in the competition

Method	DAVIS_{17}, test-challenge set							
	Rank	Global mean ↑	Region, J			Boundary, F		
			Mean ↑	Recall ↑	Decay ↓	Mean ↑	Recall ↑	Decay ↓
zwrq0	10	53.6	50.5	54.9	28.0	56.7	63.5	30.4
Fromandtozh (Zhao 2017)	9	53.9	50.7	54.9	32.5	57.1	63.2	33.7
wasidennis	8	54.8	51.6	56.3	26.8	57.9	64.8	28.8
YXLKJ	7	55.8	53.8	60.1	37.7	57.8	62.1	42.9
cjc (Cheng et al. 2017)	6	56.9	53.6	59.5	25.3	60.2	67.9	27.6
lalalafine123	6	56.9	54.8	60.7	34.4	59.1	66.7	36.1
voigtlaender (Voigtlaender and Leibe 2017a)	5	57.7	54.8	60.8	31.0	60.5	67.2	34.7
haamooon (Shaban et al. 2017)	4	61.5	59.8	71.0	21.9	63.2	74.6	23.7
vantam299 (Le et al. 2017)	3	63.8	61.5	68.6	**17.1**	66.2	79.0	**17.6**
LucidTracker	2	67.8	65.1	72.5	27.7	70.6	**79.8**	30.2
lixx (Li et al. 2017)	1	**69.9**	**67.9**	**74.6**	22.5	**71.9**	79.1	24.1

Bold are the best numbers overall

Lucid - Ablation study

Table 13 Ablation study of different ingredients. $DAVIS_{17}$, test-dev and test challenge sets

Variant	$\mathcal{I}$	$\mathcal{F}$	$\mathcal{S}$	Ensemble	CRF tuning	Temp. coherency	$DAVIS_{17}$					
							Test-dev			Test-challenge		
							Global mean	mIoU	mF	Global mean	mIoU	mF
LucidTracker (ensemble)	✓	✓	✓	✓	✓	✓	**66.6**	**63.4**	**69.9**	**67.8**	**65.1**	**70.6**
	✓	✓	✓	✓	✓	✗	65.2	61.5	69.0	67.0	64.3	69.7
	✓	✓	✓	✓	✗	✗	64.7	60.5	68.9	66.5	63.2	69.8
	✓	✓	✗	✓	✓	✗	64.9	61.3	68.4	–	–	–
	✓	✓	✗	✓	✗	✗	64.2	60.1	68.3	–	–	–
LucidTracker	✓	✓	✓	✗	✓	✗	62.9	59.1	66.6	–	–	–
$\mathcal{I}+\mathcal{F}+\mathcal{S}$	✓	✓	✓	✗	✗	✗	62.0	57.7	62.2	64.0	60.7	67.3
$\mathcal{I}+\mathcal{F}$	✓	✓	✗	✗	✗	✗	61.3	56.8	65.8	–	–	–
$\mathcal{I}+\mathcal{S}$	✓	✗	✓	✗	✗	✗	61.1	56.9	65.3	–	–	–
$\mathcal{I}$	✓	✗	✗	✗	✗	✗	59.8	63.1	63.9	–	–	–

Bold are the best numbers overall

Lucid - Qualitative Results

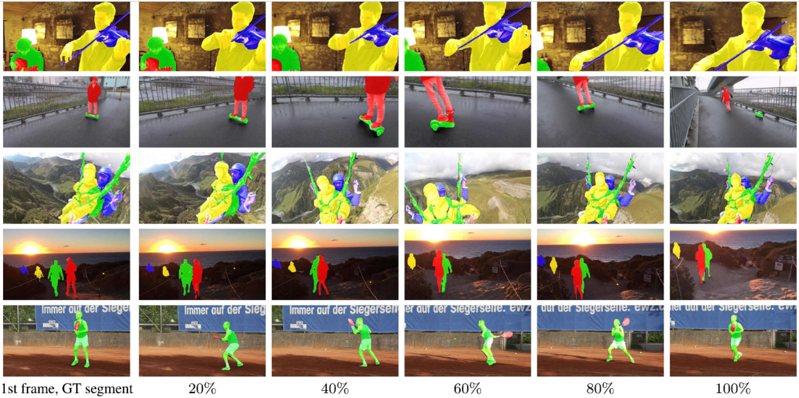


Fig. 11 LucidTracker qualitative results on $DAVIS_{17}$, test-dev set. Frames sampled along the video duration (e.g. 50%: video middle point). The videos are chosen with the highest mIoU measure

Lucid - Qualitative Results

1st frame, GT segment | 20% | 40% | 60% | 80% | 100%

Fig. 13 LucidTracker failure cases on DAVIS_{17}, test-dev set. Frames sampled along the video duration (e.g. 50%: video middle point). We show 2 results mIoU over the video below 50

PReMVOS

- Proposal generation, Refinement and Merging for Video Object Segmentation

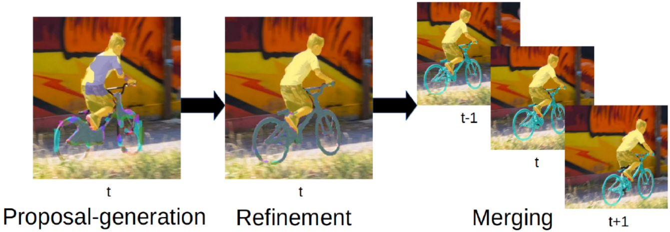


Fig. 1. PReMVOS overview. Overlay colours represent different object proposals.

PReMVOS

- Independent coarse object proposals
- Refined masks
- Merging strategy $\Rightarrow$ temporal consistency
  - objectness score
  - optical flow warping
  - Re-ID feature embedding vector
  - spatial constraints

PReMVOS - Components

Bitdefender

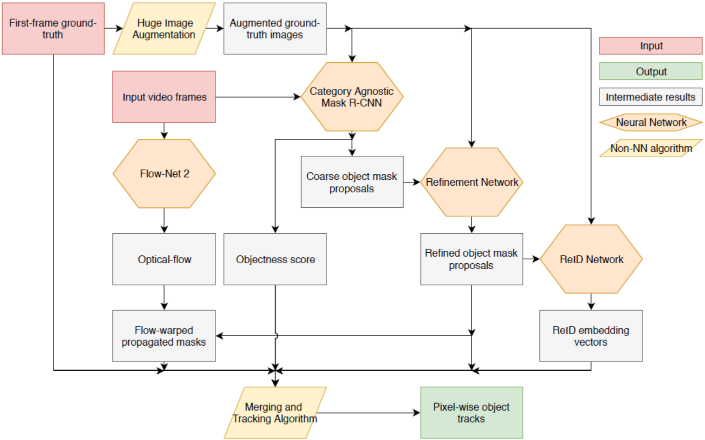


Fig. 2. Diagram showing the components of PReMVOS and their relationships.

PReMVOS - Augmentation

- Lucid Data Dreaming
- single images
- 2500 augmented images for each video

PReMVOS - Proposal Generation

- Mask R-CNN* with ResNet101** backbone
- category agnostic - map all classes to a single foreground class
- train:
  - start from pretrained ImageNet weights
  - train on COCO and Mapillary datasets
  - fine tune for each video
- result:
  - coarse mask proposals
  - bounding boxes
  - objectness scores
- keep proposals with score > 0.05
- NMS

*He et al. [2017], **He et al. [2016]

PReMVOS - Proposal Refinement

- fully convolutional inspired by DeepLabv3+*
- input 385x385 image patch - bounding box around object
- train:
  - start from pretrained ImageNet, COCO and PASCAL weights
  - train on Mapillary
  - fine tune for each video

*Chen et al. [2018]

PReMVOS - Mask Propagation

- mask warping between frames
- FlowNet 2.0*

*Ilg et al. [2017]

PReMVOS - ReID Embedding Vectors

- triplet-loss based ReID embedding network (ResNet)
- differentiate between objects
- input: 128x128 image patch - bounding box around object
- train:
  - start from pretrained ImageNet
  - train on COCO
  - fine tune per dataset

PReMVOS - Proposal Merging

- score each proposal based on the likeliness of belonging to a particular object track
- hard decisions at each time step
- notations:
  - $s_{type,t,i,j}$
    - type - score type
    - time step
    - i^{th} proposal $(c_{t,i})$
    - j^{th} track
  - f_j
    - j^{th} object in the first frame
  - $r(x)$
    - ReID embedding vector of x

PReMVOS - Proposal Merging

Bitdefender

- Objectness score
  - $s_{obj,t,i,j}(c_{t,i}) = MaskObj(c_{t,i})$
    - confidence value provided by Proposal Generation network
- ReID score
  - $s_{reid,t,i,j}(c_{t,i}, f_j) = 1 - \frac{\left\| r(c_{t,i}) - r(f_j) \right\|}{max_{\tilde{t},\tilde{i}} \left\| r(c_{\tilde{t},\tilde{i}}) - r(f_j) \right\|}$
- Mask Propagation score
  - $s_{maskprop,t,,j}(c_{t,i}, p_{t-1,j}) = IOU(c_{t,i}, warp(p_{t-1,j}))$
- Inverse ReID score
  - $s_{inv\_reid,t,i,j} = 1 - \max_{k \neq j}(s_{reid,t,i,k})$
- Inverse Mask Propagation score
  - $s_{inv\_maskprop,t,i,j} = 1 - \max_{k \neq j}(s_{maskprop,t,i,k})$

PReMVOS - Proposal Merging

- Final score

  - $s_{comb,t,i,j} = \sum_{q \in obj,reid,maskprop,inv_reid,inv_maskprop} \alpha_q s_{q,t,i,j}$

  - $\sum_q \alpha_q = 1$
    - equal weights for single object, tuned weights for multiple objects
  - $\alpha_q \geq 0$

  - $p_{t,j} = c_{t,k_j}$, where $k_j = \arg\max_i s_{comb,t,i,j}$

PReMVOS - Quantitative Results

Bitdefender

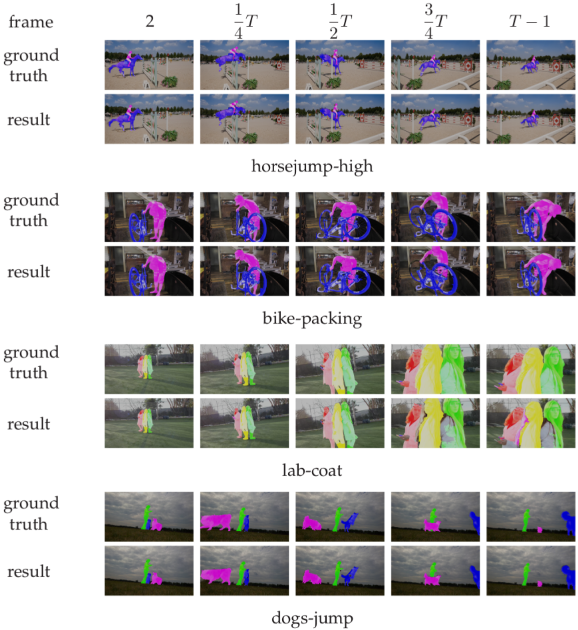
			Ours	DyeNet [17]	MRF [1]	Lucid [15]	ReID [18]	OSVOS-S [21]	OnAVOS [29] [30]	OSVOS [2]
17 T-D	$\mathcal{J}\&\mathcal{F}$	Mean	**71.6**	68.2	67.5	66.6	66.1	57.5	56.5	50.9
	$\mathcal{J}$	Mean	**67.5**	65.8	64.5	63.4	64.4	52.9	52.4	47.0
		Recall	**76.8**	-	-	73.9	-	60.2	-	52.1
		Decay	21.7	-	-	19.5	-	24.1	-	**19.2**
	$\mathcal{F}$	Mean	**75.7**	70.5	70.5	69.9	67.8	62.1	59.6	54.8
		Recall	**84.3**	-	-	80.1	-	70.5	-	59.7
		Decay	20.6	-	-	**19.4**	-	21.9	-	19.8
17 Val	$\mathcal{J}\&\mathcal{F}$	Mean	**77.8**	74.1	70.7	-	-	68.0	67.9	60.3
	$\mathcal{J}$	Mean	**73.9**	-	67.2	-	-	64.7	64.5	56.6
		Recall	**83.1**	-	-	-	-	74.2	-	63.8
		Decay	16.2	-	-	-	-	**15.1**	-	26.1
	$\mathcal{F}$	Mean	**81.7**	-	74.2	-	-	71.3	71.2	63.9
		Recall	**88.9**	-	-	-	-	80.7	-	73.8
		Decay	19.5	-	-	-	-	**18.5**	-	27.0

PReMVOS - Quantitative Results - Ensemble

		Ours (Ens)	Ours	DyeNet [16]	ClassAgno. VOS [35]	OnlineGen. VOS [8]	Lucid [14]	ContextBased VOS [28]
$\mathcal{J}\&\mathcal{F}$	Mean	**74.7**	71.8	73.8	69.7	69.5	67.8	66.3
$\mathcal{J}$	Mean	71.0	67.9	**71.9**	66.9	67.5	65.1	64.1
	Recall	**79.5**	75.9	79.4	74.1	77.0	72.5	75.0
	Decay	19.0	23.2	19.8	23.1	15.0	27.7	**11.7**
$\mathcal{F}$	Mean	**78.4**	75.6	75.8	72.5	71.5	70.6	68.6
	Recall	**86.7**	82.9	83.0	80.3	82.2	79.8	80.7
	Decay	20.8	24.7	20.3	25.9	18.5	30.2	**13.5**

Table 2. Our results (with and without ensembling) on the DAVIS `test-challenge` dataset compared with the top five other competitors in the 2018 DAVIS Challenge.

PReMVOS - Qualitative Results



PReMVOS - Ablation study - Proposal Refinement

Bitdefender

- oracle merging - choose proposal with best IOU

	$\mathcal{J}$ mean	$\mathcal{F}$ mean	$\mathcal{J}\&\mathcal{F}$ mean
Without Refinement	71.2	77.3	74.2
With Refinement	**77.1**	**85.2**	**81.2**
Boost	5.9	7.9	7.0

Table 3. Quantitative results of an ablation study on the `2017 val` dataset showing the effect of the Refinement Network on the accuracy of generated mask proposals. Presented results are calculated using *oracle merging* (see Section 4.1).

PReMVOS - Ablation study - Proposal Refinement

- oracle merging - choose proposal with best IOU

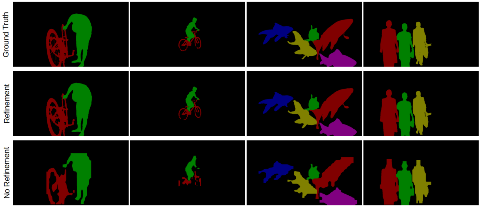


Fig. 4. Qualitative results showing the effect of the Refinement Network on the mask proposal accuracy. Results are calculated using *oracle merging* (Section 4.1).

PReMVOS - Ablation Study - Proposal Merging

Num.	Merging Sub-Score Components					$\mathcal{J}\&\mathcal{F}$
Comp.	Objectness	ReID	InvReID	MaskProp	InvMaskProp	Mean
0	Oracle merging					**81.2**
5(opt.)	19%	18%	14%	22%	27%	**78.2**
5	✓	✓	✓	✓	✓	**77.8**
	✓	✓	✓	✓	-	76.7
	✓	✓	✓	-	✓	*75.5*
4	✓	✓	-	✓	✓	**76.9**
	✓	-	✓	✓	✓	76.3
	-	✓	✓	✓	✓	75.9
	✓	✓	✓	-	-	74.2
	✓	✓	-	✓	-	75.0
	✓	✓	-	-	✓	74.2
	✓	-	✓	✓	-	73.5
3	✓	-	✓	-	✓	69.6
	✓	-	-	✓	✓	71.1
	-	✓	✓	✓	-	75.8
	-	✓	✓	-	✓	*69.3*
	-	✓	-	✓	✓	**75.9**
	-	-	✓	✓	✓	74.3
	✓	✓	-	-	-	72.7
	✓	-	✓	-	-	64.7
	✓	-	-	✓	-	69.1
	✓	-	-	-	✓	57.9
2	-	✓	✓	-	-	68.7
	-	✓	-	✓	-	**74.3**
	-	✓	-	-	✓	68.8
	-	-	✓	✓	-	74.0
	-	-	✓	-	✓	*47.3*
	-	-	-	✓	✓	73.6
	✓	-	-	-	-	*29.5*
	-	✓	-	-	-	67.4
1	-	-	✓	-	-	44.3
	-	-	-	✓	-	**72.8**
	-	-	-	-	✓	34.4

PReMVOS - Computation time

	Augm. Gen.	Fine-tuning	Prop. Gen.	Prop. Refine.	ReID	Optic. Flow	Warping	Merging	**Total**	Av. # Prop.	Mean $\mathcal{J}\&\mathcal{F}$
Original	23.4	12.3	0.41	1.04	0.05	0.14	0.32	0.02	37.4	17.52	77.8
Fast-finetuned	0.02	3.9	0.26	0.45	0.03	0.14	0.20	0.02	5.02	9.28	73.7
Not-finetuned	0.00	0.00	0.14	0.33	0.02	0.14	0.16	0.02	0.81	6.87	65.7

Table 5. Runtime analysis of the different components of the PReMVOS algorithm. Times are in seconds per frame, averaged over the DAVIS 2017 val set. Augmentation Generation is run on 48 CPU cores, and Fine-tuning is done on 8 GPUs. Otherwise, everything is run sequentially on one GPU / CPU core.

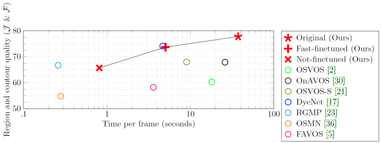


Fig. 5. Quality versus timing on the DAVIS 2017 val set. For methods that only publish runtime results on the DAVIS 2016 dataset, we take these timings as per object timings and extrapolate to the number of objects in the DAVIS 2017 val set.

Thank you!

References I

F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989.

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212, 2004.

References II

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for video object segmentation. *International Journal of Computer Vision*, pages 1–23, 2018.

References III

P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

J. Luiten, P. Voigtlaender, and B. Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. *arXiv preprint arXiv:1807.09190*, 2018.

J. Sun, J. Jia, C.-K. Tang, and H.-Y. Shum. Poisson matting. In *ACM Transactions on Graphics (ToG)*, volume 23, pages 315–321. ACM, 2004.

H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.