Online Adaptation of Convolutional Neural Networks for Video Object Segmentation: Supplementary Material

Paul Voigtlaender
voigtlaender@vision.rwth-aachen.de
Bastian Leibe
leibe@vision.rwth-aachen.de

Computer Vision Group
Visual Computing Institute
RWTH Aachen University
Germany

1 More Comprehensive Comparison to Other Methods

Table 1 shows a more comprehensive comparison of our results to the results obtained by other methods.

Method	DAVIS mIoU [%]	YouTube-Objects mIoU [%]
OnAVOS (ours), no adaptation	81.7 ± 0.2	76.6 ± 0.1
OnAVOS (ours), online adaptation	$\mathbf{85.7} \pm 0.6$	$\mathbf{77.4} \pm 0.2$
OSVOS [1]	79.8	72.5
MaskTrack [7]	79.7	72.6
LucidTracker [3] †	80.5	76.2
VPN [2]	75.0	-
FCP [5]	63.1	-
BVS [4]	66.5	59.7
OFL [8]	71.1	70.1
STV [9]	73.6	-

Table 1: Comparison to other methods on the DAVIS validation set and the YouTube-Objects dataset. Note that *MaskTrack* [7] and *LucidTracker* [3] report results on DAVIS for all sequences including the training set, but here we show their results for the validation set only. †: Concurrent work only published on arXiv.

2 Additional Evaluation Measures for DAVIS

Table 2 shows a more detailed evaluation on the DAVIS validation set using the evaluation measures suggested by Perazzi *et al.* [6]. The measures used here are the Jaccard index $\mathcal{J}$, defined as the mean intersection-over-union (mIoU) between the predicted foreground masks and the ground truth masks; the contour accuracy measure $\mathcal{F}$, which measures how well the segmentation boundaries agree; and the temporal stability measure $\mathcal{T}$, which measures the consistency of the predicted masks over time. For more details of these measures, we refer

the interested reader to Perazzi *et al*. [6]. Note that the results for additional measures for *LucidTracker* [5] are missing since they are only reported averaged over all 50 sequences of DAVIS and not on the validation set.

The table shows that each evaluation measure is significantly improved by the proposed online adaptation scheme. *OnAVOS* obtains the best mean results for all three measures. It is surprising that our result for the temporal stability $\mathcal{T}$ is better than the result by *MaskTrack* [7], although in contrast to our method, they explicitly incorporate temporal context by propagating masks.

Measure		*OnAVOS* (ours)		*OSVOS* [4]	*MaskTrack* [7]	*LucidTracker* [5]
		Un-adapted	Adapted			
$\mathcal{J}$	mean ↑	*81.7* ± 0.2	**85.7** ± 0.6	79.8	79.7	80.5
	recall ↑	92.2 ± 0.6	**95.4** ± 0.8	*93.6*	93.1	-
	decay ↓	11.9 ± 0.3	**7.1** ± 1.7	14.9	*8.9*	-
$\mathcal{F}$	mean ↑	*81.1* ± 0.2	**84.2** ± 0.8	80.6	75.4	-
	recall ↑	88.2 ± 0.3	*88.7* ± 1.3	**92.6**	87.1	-
	decay ↓	11.2 ± 0.5	**7.8** ± 1.8	15.0	*9.0*	-
$\mathcal{T}$	mean ↓	27.3 ± 2.2	**18.5** ± 0.1	37.6	*21.8*	-

Table 2: Additional evaluation measures on the DAVIS validation set. Best and second best results are highlighted with bold and italic fonts, respectively.

3 Per-Sequence Results for DAVIS

Table 3 shows mIoU results for each of the 20 sequences of the DAVIS validation set. On 18 out of 20 sequences, *OnAVOS* obtains either the best or the second best result.

4 Hyperparameter Study on DAVIS

As described in the main paper, we found $\alpha = 0.97$, $\beta = 0.05$, $d = 220$, $n_{online} = 15$, $n_{curr} = 3$, $\lambda = 10^{-5}$ and 15 for the erosion size to work well on DAVIS. Starting from these values as the operating point, we conducted a more detailed hyperparameter study by changing one hyperparameter at a time, while keeping all others constant (see Fig. 1). The plots show that the performance of *OnAVOS* is in general very stable with respect to the choice of most of its hyperparameters and for every configuration we tried, the result was better than the unadapted baseline (the dashed line in the plots). The single most important hyperparameter is the online learning rate λ, which is common for deep learning approaches. The online loss scale β and the positive threshold α have a moderate influence on performance, while changing the distance threshold d and the number of steps n_{online} and n_{curr} in a reasonable range only leads to minor changes in accuracy. For the erosion size, the optimum is achieved at 1, *i.e*. when no erosion is applied. This result suggests that the erosion operation is not helpful for DAVIS. The plots show that there is still some potential for improving the results by further tuning the hyperparameters. However, this study was meant as a characterization of our method rather than a systematic tuning.

The generalizability and the robustness of *OnAVOS* with respect to the choice of hyperparameters is further confirmed by the experiments on YouTube-Objects, which used the same hyperparameter settings as on DAVIS.

Sequence	Method, mIoU [%]				
	OnAVOS (ours)		*OSVOS* [1]	*MaskTrack* [2]	*LucidTracker* [3]
	Un-adapted	Adapted			
blackswan	*96.1* ± 0.1	**96.2** ± 0.1	94.2	90.3	95.0
bmx-trees	48.2 ± 0.8	*57.0* ± 1.0	55.5	**57.5**	55.0
breakdance	62.6 ± 4.2	73.6 ± 3.8	70.8	*76.1*	**87.2**
camel	84.6 ± 0.1	*85.5* ± 0.1	85.1	80.1	**94.3**
car-roundabout	86.5 ± 0.2	**97.5** ± 0.0	95.3	*96.0*	*96.0*
car-shadow	*94.1* ± 0.1	**96.8** ± 0.1	93.7	93.5	90.3
cows	**95.4** ± 0.0	**95.4** ± 0.0	*94.6*	88.2	93.1
dance-twirl	78.4 ± 0.7	*85.6* ± 1.0	67.0	84.4	**88.6**
dog	**95.6** ± 0.1	**95.6** ± 0.1	90.7	90.8	*95.0*
drift-chicane	*87.4* ± 0.5	**89.2** ± 0.2	83.5	86.2	1.4
drift-straight	*81.3* ± 5.6	**93.7** ± 0.9	67.6	56.0	79.9
goat	*90.8* ± 0.1	**91.4** ± 0.1	88.0	84.5	88.9
horsejump-high	*89.3* ± 0.3	**90.1** ± 0.0	78.0	81.8	87.1
kite-surf	**70.1** ± 1.0	*69.1* ± 0.1	68.6	60.0	64.6
libby	*87.1* ± 1.0	**88.6** ± 0.1	80.8	77.5	85.5
motocross-jump	**89.7** ± 0.2	70.4 ± 11.9	*81.6*	68.3	75.1
paragliding-launch	**64.6** ± 0.1	*64.3* ± 0.1	62.5	62.1	63.7
parkour	92.4 ± 0.2	**93.6** ± 0.0	85.6	88.2	*93.2*
scooter-black	64.8 ± 7.1	**91.3** ± 0.1	71.1	82.4	*86.5*
soapbox	74.0 ± 4.6	*89.8* ± 1.2	81.2	*89.9*	**90.5**
mean	*81.7* ± 0.2	**85.7** ± 0.6	79.8	79.7	80.5

Table 3: Per-sequence results on the DAVIS validation set. Best and second best results are highlighted with bold and italic fonts, respectively.

References

[1] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *CVPR*, 2017.

[2] V. Jampani, R. Gadde, and P. V. Gehler. Video propagation networks. In *CVPR*, 2017.

[3] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for object tracking. In *arXiv preprint arXiv: 1703.09554*, 2017.

[4] N. Maerki, F. Perazzi, O. Wang, and A. Sorkine-Hornung. Bilateral space video segmentation. In *CVPR*, 2016.

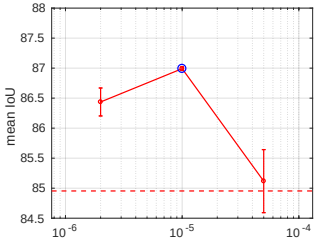
[5] F. Perazzi, O. Wang, M. Gross, and A. Sorkine-Hornung. Fully connected object proposals for video segmentation. In *ICCV*, 2015.

[6] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.

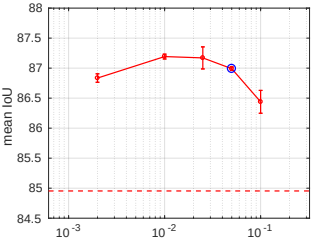
[7] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, 2017.

[8] Y-H. Tsai, M-H. Yang, and M. J. Black. Video segmentation via object flow. In *CVPR*, 2016.

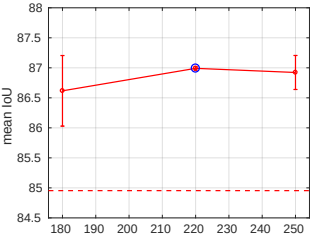
[9] Wenguan W. and Shenjian B. Super-trajectory for video segmentation. *arXiv preprint arXiv:1702.08634*, 2017.



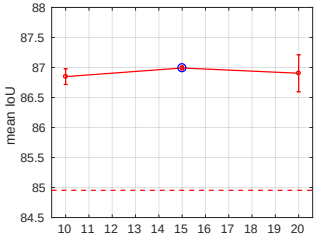
(a) online learning rate λ



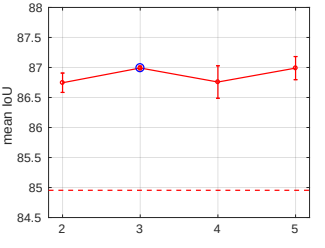
(b) online loss scale β



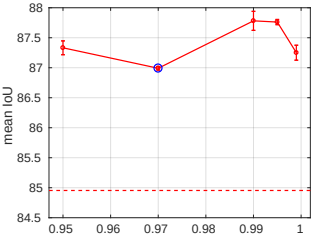
(c) distance threshold d



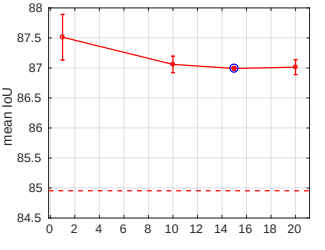
(d) total steps n_{online}



(e) update steps n_{curr}



(f) positive threshold α



(g) erosion size

Figure 1: Influence of online adaptation hyperparameters on the DAVIS training set. The blue circle marks the operating point, based on which one parameter is changed at a time. The dashed line marks the un-adapted baseline. The plots show that overall our method is very robust against the exact choice of hyperparameters, except for the online learning rate λ. The standard deviations estimated by three runs are shown as error bars. In some cases, including the operating point, the estimated standard deviation is so small that it is hardly visible.