Anchor Diffusion for Unsupervised Video Object Segmentation
SUPPLEMENTARY MATERIAL

Zhao Yang*
University of Oxford
zhao.yang@eng.ox.ac.uk

Qiang Wang*
CASIA
qiang.wang@nlpr.ia.ac.cn

Luca Bertinetto
Five AI
luca@robots.ox.ac.uk

Weiming Hu
CASIA
wmhu@nlpr.ia.ac.cn

Song Bai
University of Oxford
songbai.site@gmail.com

Philip H.S. Torr
University of Oxford
philip.torr@eng.ox.ac.uk

A. Global Comparison

Table 1 includes all metrics reported by the official DAVIS 2016 benchmark [13]. Our method substantially outperforms competing methods in the main evaluation metrics of mean region similarity $\mathcal{J}$ and mean contour accuracy $\mathcal{F}$. The small decay measure for both $\mathcal{J}$ and $\mathcal{F}$ shows AD-Net's long-term benefits on performance.

B. Per-sequence Comparison

Figures 1 and 2 compare the per-sequence $\mathcal{J}$ and $\mathcal{F}$ of AD-Net against top 7 competing methods on the leaderboard. Our method performs well on videos presenting a variety of challenges, such as large appearance changes (Car-Shadow, Parkour), cluttered background (Car-Roundabout, Scooter-Black), heavy occlusion (Libby, Bmx-Trees), fast motion (Bmx-Trees, Dog, Parkour), *etc*.

C. Qualitative Analysis on FBMS and ViSal

In Figures 3 and 4, we visualise segmentation results on videos from the test set of FBMS [11] and ViSal [19] respectively. The model is trained only with the DAVIS 2016 training set. We do not fine-tune it on the training set of FBMS or ViSal.

D. Foreground Correspondence Analysis

In Figure 5, we visualise more examples of foreground pixel correspondences to pixels in the anchor frame. Most pixels are randomly selected from the foreground area on the last frame of the video (except when foreground becomes too small in the last frame, in which case another frame is randomly chosen).

*Equal contribution.

E. Instance Pruning

Algorithm 1 details the instance pruning procedure. Function $SmallStatic$ returns a set of bounding boxes and the corresponding instance masks that represent small and nearly static instances determined from their trajectories. Then function $GetPruningMask$ takes these instances and the original masks as inputs, and generates a pruning mask per frame, which incorporates all small and static instances that are significantly smaller than the largest instance in the current frame. Finally, each original mask is multiplied element-wise with the corresponding pruning mask to output the final predictions.

References

[1] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *ICCV*, 2017. 2

[2] Alon Faktor and Michal Irani. Video segmentation by non-local consensus voting. In *BMVC*, 2014. 2

[3] Katerina Fragkiadaki, Geng Zhang, and Jianbo Shi. Video segmentation by tracing discontinuities in a trajectory embedding. In *CVPR*, 2012. 2

[4] Brent A. Griffin and Jason J. Corso. Tukey-inspired video object segmentation. In *WACV*, 2019. 2

[5] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *CVPR*, 2017. 2

[6] Margret Keuper, Bjoern Andres, and Thomas Brox. Motion trajectory segmentation via minimum cost multicuts. In *ICCV*, 2015. 2

[7] Yeong Jun Koh and Chang-Su Kim. Primary object segmentation in videos based on region augmentation and reduction. In *CVPR*, 2017. 2

[8] Dong Lao and Ganesh Sundaramoorthi. Extending layered models to 3d motion. In *ECCV*, 2018. 2

Measure	ADNet	MotAdapt[14]	PDB[15]	ARP[7]	LVO[18]	FSEG[5]	LMP[17]	SFL[1]	TIS[4]	ELM[8]	FST[12]	CUT[6]	NLC[2]	MSG[10]	KEY[9]	CVOS[16]	TRC[3]
$\mathcal{J}$ Mean ↑	**81.7**	77.2	77.2	76.2	75.9	70.7	70.0	67.4	62.6	61.8	55.8	55.2	55.1	53.3	49.8	48.2	47.3
$\mathcal{J}$ Recall ↑	90.9	87.8	90.1	**91.1**	89.1	83.5	85.0	81.4	80.3	67.2	64.9	57.5	55.8	61.6	59.1	54.0	49.3
$\mathcal{J}$ Decay ↓	2.2	5.0	0.9	7.0	**0.0**	1.5	1.3	6.2	7.1	9.8	**0.0**	2.2	12.6	2.4	14.1	10.5	8.3
$\mathcal{F}$ Mean ↑	**80.5**	77.4	74.5	70.6	72.1	65.3	65.9	66.7	59.6	61.2	51.1	55.2	52.3	50.8	42.7	44.7	44.1
$\mathcal{F}$ Recall ↑	**85.1**	84.4	84.4	83.5	83.4	73.8	79.2	77.1	74.5	65.4	51.6	61.0	51.9	60.0	37.5	52.6	43.6
$\mathcal{F}$ Decay ↓	0.6	3.3	**-0.2**	7.9	1.3	1.8	2.5	5.1	6.4	8.8	2.9	3.4	11.4	5.1	10.6	11.7	12.9
$\mathcal{T}$ (GT 8.8) ↓	36.9	27.9	29.1	39.3	26.5	32.8	57.2	28.2	33.6	25.1	36.6	27.7	42.5	30.1	26.9	**25.0**	39.1

Table 1. Detailed evaluation results on the DAVIS 2016 validation set. We analyse region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and temporal stability $\mathcal{T}$ in terms of mean, recall, and decay, and compare with state-of-the-art methods from the DAVIS 2016 leaderboard.

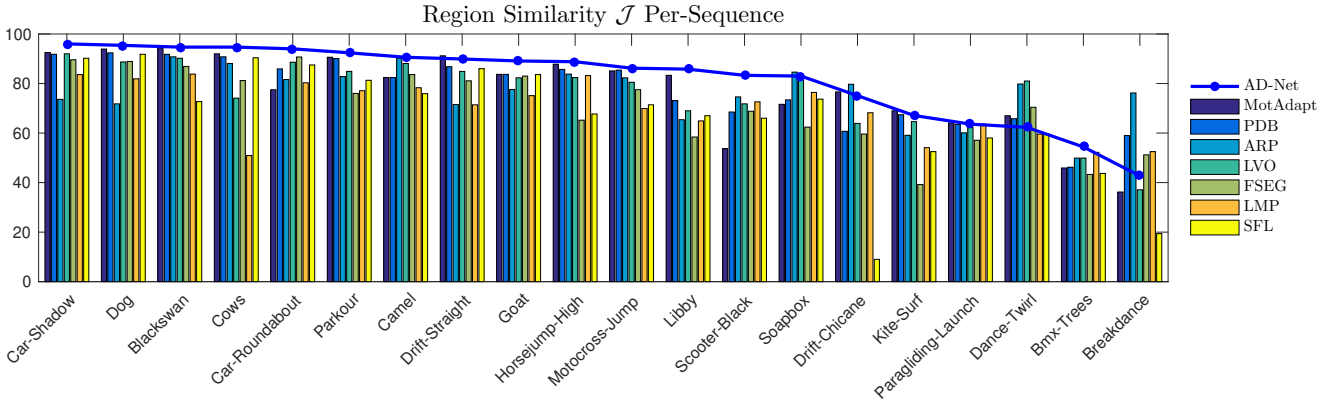

Figure 1. Per-sequence results of mean region similarity $\mathcal{J}$ against top 7 methods on the public leaderboard of DAVIS 2016. The blue line indicates AD-Net, while bars represent other methods. Sequences are organised in descending order of the performance of our method.

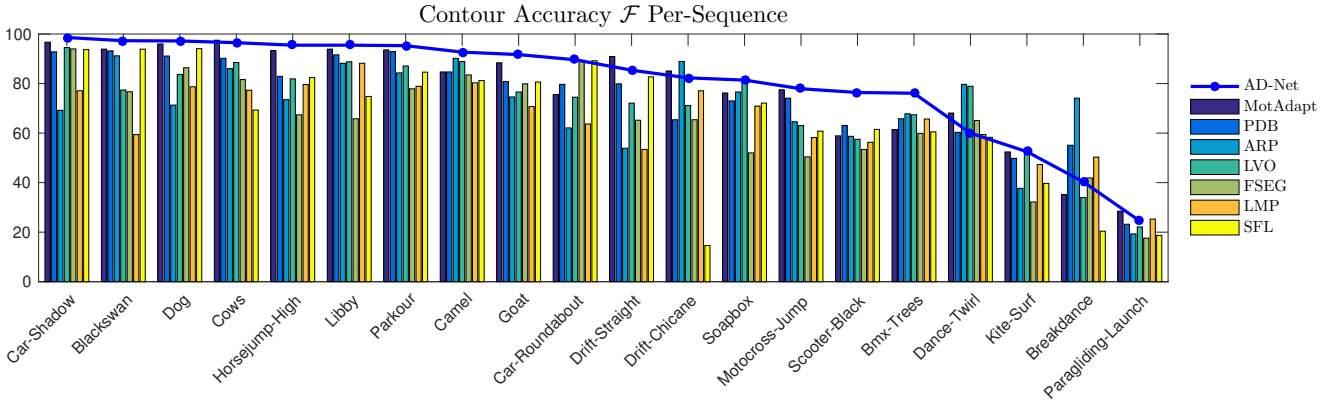

Figure 2. Per-sequence results of mean contour accuracy $\mathcal{F}$ against top 7 methods on the public leaderboard of DAVIS 2016. The blue line indicates AD-Net, while bars represent other methods. Sequences are organised in descending order of the performance of our method.

[9] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. Key-segments for video object segmentation. In *ICCV*, 2011. 2

[10] Peter Ochs and Thomas Brox. Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions. In *ICCV*, 2011. 2

[11] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *TPAMI*, 2014. 1

[12] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2013. 2

[13] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 1

[14] Mennatullah Siam, Chen Jiang, Steven Lu, Laura Petrich, Mahmoud Gamal, Mohamed Elhoseiny, and Martin Jagersand. Video object segmentation using teacher-student adaptation in a human robot interaction (hri) setting. In *ICRA*, 2019. 2

[15] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, 2018. 2

[16] Brian Taylor, Vasiliy Karasev, and Stefano Soatto. Causal video object segmentation from persistence of occlusions. In *CVPR*, 2015. 2

[17] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *CVPR*, 2017. 2

[18] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid.

Figure 3. Segmentation results on challenging videos from FBMS without fine-tuning.

Learning video object segmentation with visual memory. In *ICCV*, 2017. 2

[19] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *TIP*, 2015. 1

Algorithm 1 Instance Pruning

Input: original masks $X = [x_0, ..., x_{N-1}]$, bounding boxes/instance masks $E = [E_0, ..., E_{M-1}]$, for N frames and M total instances
Output: refined masks X'
$size\_low \leftarrow Area(Sort(E)[-N])$
$T \leftarrow SmallStatic(E, 0.6, 0.5N, size\_low)$
for $t = 1$ **to** N **do**
  Let b_t be instances on frame t from E
  $F \leftarrow GetPruningMask(x_t, T, size\_low)$
  $x_t \leftarrow x_t \odot F$
end for
return X
function $SmallStatic(b, iou, support, size)$
  $sm\_stat\_instances \leftarrow \emptyset$
  for b_i in b **do**
    for b_j in b **do**
      $count \leftarrow 0$
      if $IoU(b_i, b_j) > iou$ **then**
        $count \leftarrow count + 1$
      end if
    end for
    if $count > support$ and $Size(b_i) < size$ **then**
      Add b_i to $sm\_stat\_instances$
    end if
  end for
  return $sm\_stat\_instances$
end function
function $GetPruningMask(x_t, T, s)$
  $pruning\_mask \leftarrow \emptyset, target\_size \leftarrow -\infty$
  $T_t \leftarrow Sort(T_t, descending)$
  if $Size(T_t[0]) > s$ & $Len(T_t) > 0$ & $Size(T_t[0]) > 2Size(T_t[1])$ **then**
    $target\_size \leftarrow Size(T_t[0])$
  end if
  for all $T_t[i]$ in T_t **do**
    if $Size(T_t[i]) < \frac{target\_size}{3}$ **then**
      $pruning\_mask \leftarrow pruning\_mask \cup T_t[i]$
    end if
  end for
  return $pruning\_mask$
end function

Figure 4. Segmentation results on challenging videos from ViSal without fine-tuning.

Figure 5. Similarity scores of a foreground pixel on a later frame with pixels in the anchor. Left, middle, and right images for each video illustrate, respectively, the target frame and the sampled foreground pixel (blue cross), the anchor frame, and similarities overlaid on the anchor frame.