Supplementary material
D3S – A Discriminative Single Shot Segmentation Tracker

Alan Lukežič[1], Jiří Matas[2], Matej Kristan[1]

[1]Faculty of Computer and Information Science, University of Ljubljana, Slovenia

[2]Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic

alan.lukezic@fri.uni-lj.si

1. Qualitative examples

Due to the page limit of the paper we provide here additional qualitative examples of tracking and segmentation. Video sequences are collected from the VOT2016 [2], GOT-10k [1] and DAVIS [3, 4] datasets. Output of the D3S is segmentation mask and it is visualized with yellow color. A bounding box is fitted to the predicted segmentation mask and shown in red. Tracker reports binary segmentation mask for DAVIS, rotated bounding box for VOT sequences, while axis-aligned bounding box is required by the GOT-10k evaluation protocol. The following tracking and segmentation conditions are visualized:

- Figure 1 demonstrates the discriminative power of D3S by visualizing tracking in presence of distractors, i.e., visually similar objects.
- Figure 2 shows a remarkable segmentation accuracy and robustness of D3S on tracking of deformable objects and *parts* of objects.
- Figure 3 shows tracking in sequences we have identified as particularly challenging for the current state-of-the-art. It includes small objects and tracking parts of objects.
- Figure 4 shows (near real-time) video object segmentation results on DAVIS16 [3] and DAVIS17 [4] datasets.

References

[1] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 1

[2] Matej Kristan, Aleš Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka Čehovin, Tomas Vojir, Gustav Häger, Alan Lukežič, and Gustavo et al. Fernandez. The Visual Object Tracking VOT2016 challenge results. In *Proc. European Conf. Computer Vision*, 2016. 1

[3] 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 *Comp. Vis. Patt. Recognition*, 2016. 1

[4] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS Challenge on Video Object Segmentation. *arXiv:1704.00675*, 2017. 1

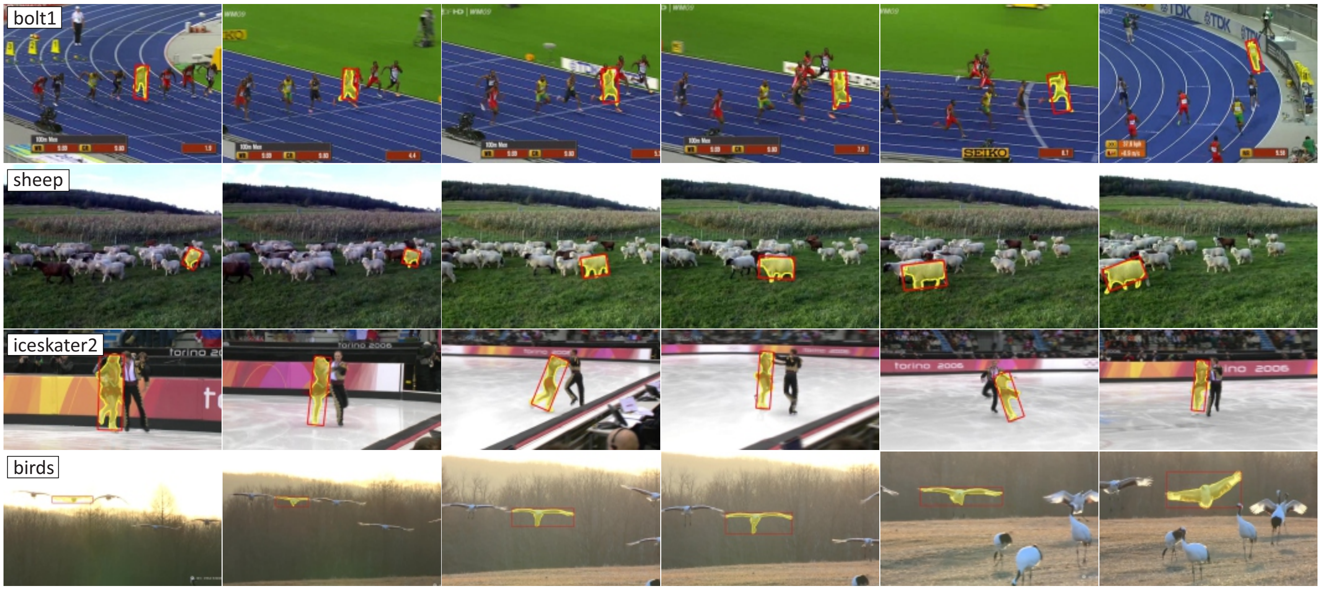

Figure 1. Sequences with distractors (similar objects in the target vicinity). D3S segments the correct target even though a similar target is close (or even overlapping). These examples show discriminative power of the proposed tracker achieved by the discriminative GIM and GEM.

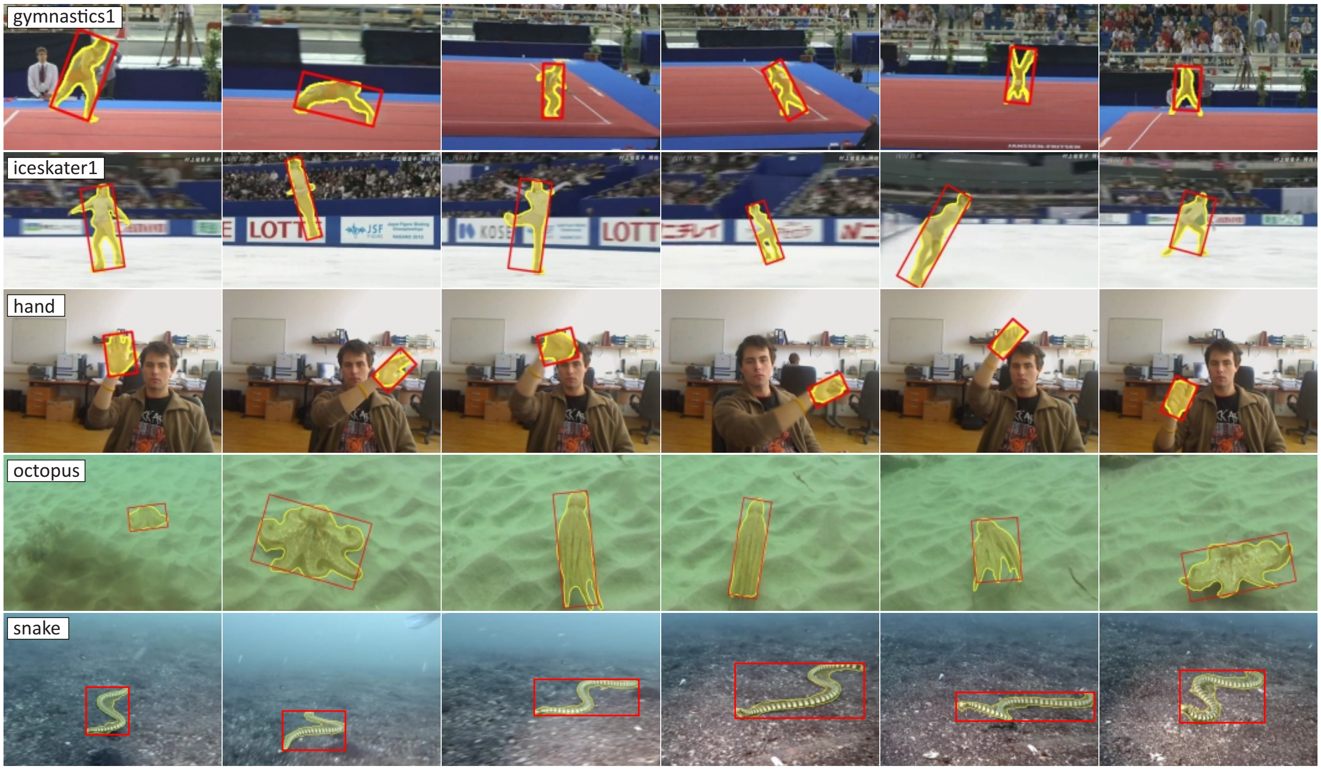

Figure 2. Examples of appearance changes and deforming targets. The geometrically invariant model (GIM) successfully segments the target due to geometrically unrestricted representation even under target rotation (*gymnastics1*), articulated (*iceskater1* and *octopus*) or significantly change its shape (*hand* and *snake*).

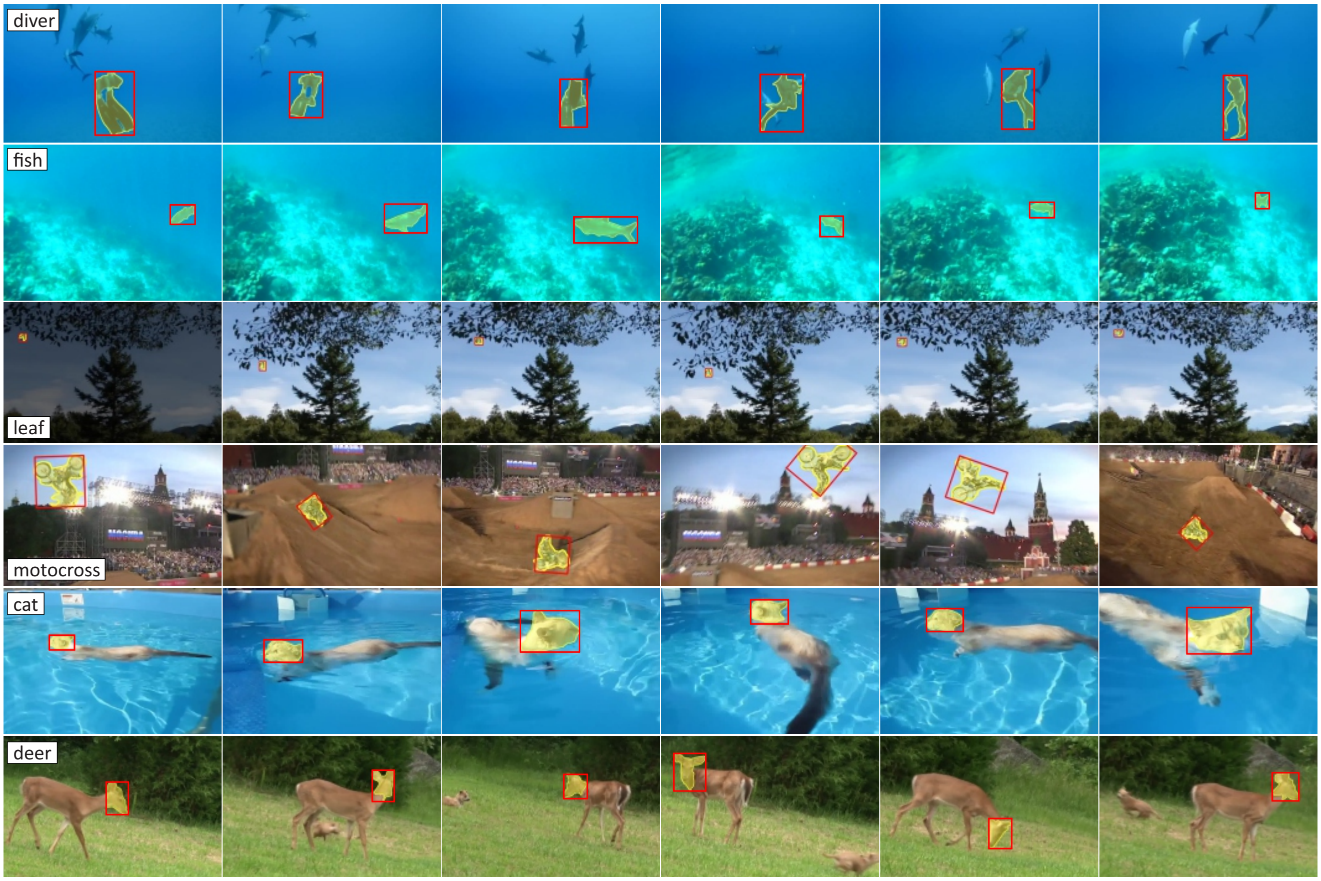

Figure 3. Difficult examples to track and segment. Underwater video sequences *diver* and *fish* are challenging due to the low contrast between the target and background – the D3S refinement pathway still produces an accurate segmentation. Small target in *leaf* sequence is successfully tracked and segmented due to the large search range (4-times of target size) and the discriminative architecture, even though several similar leaves are in the vicinity and all leaves undergo abrupt motion due to a high wind. Target rotation and scale change in *motocross* sequence are successfully addressed by the geometrically invariant model (GIM). A challenging scenario where only the head of the *cat* and *deer* is tracked. Foreground and background feature vectors in GIM and combination with GEM prevent segmenting the whole animal as the target.

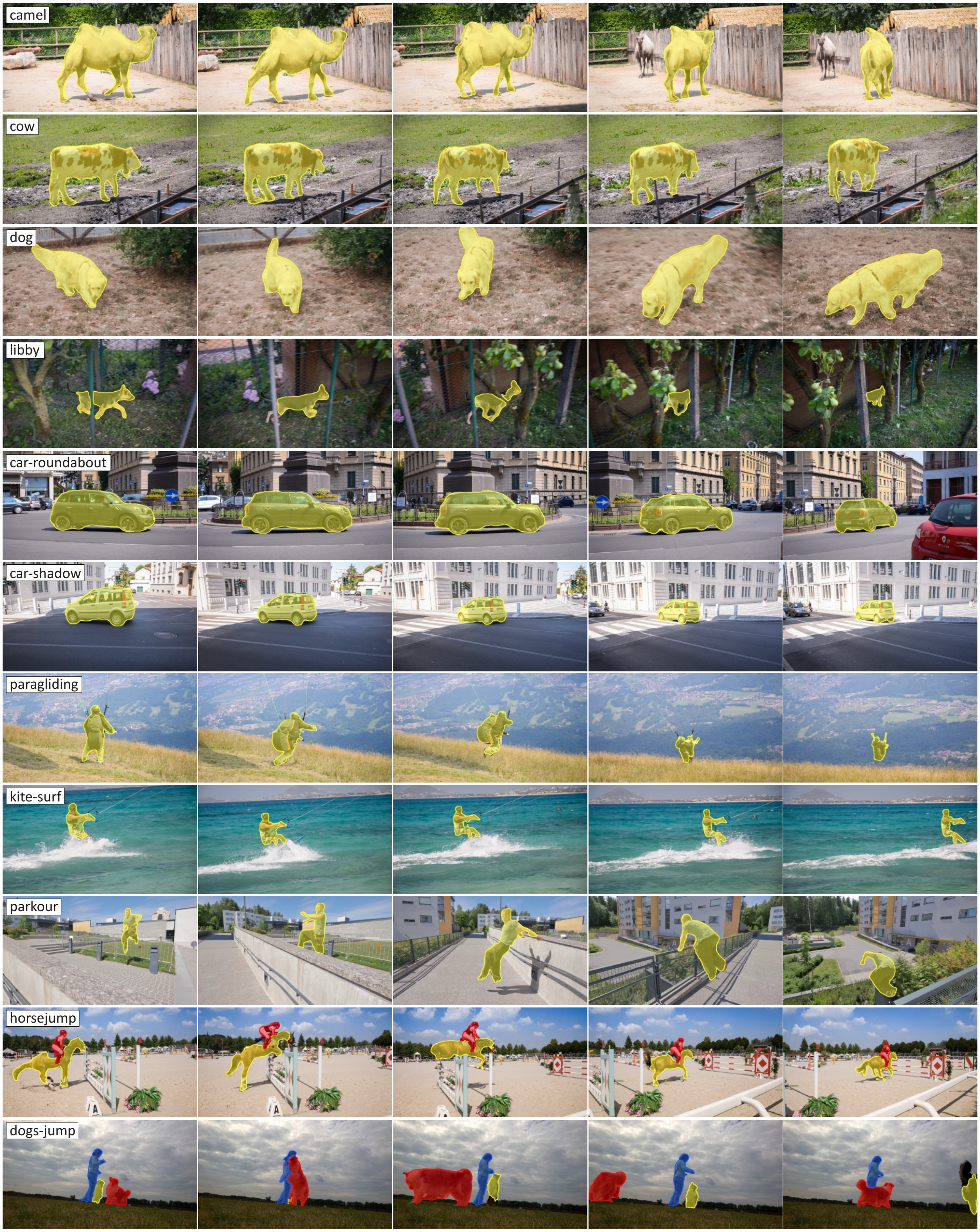

Figure 4. Video object segmentation on DAVIS datasets. D3S produces a highly accurate segmentation in near real-time. In sequences with multiple objects the tracker was run independently on each target.