De-Supervision in Camouflaged Videos Supplementary Material

Luca Alessandrini, Antonino Maria Rizzo, Luca Magri, Giacomo Boracchi and Federica Arrigoni
DEIB – Politecnico di Milano, Italy

In this document we report additional qualitative results, that could not be included in the main paper due to space constraints. CIS [5] has not been included in this analysis, since, as shown in the main paper, it presents difficulties in the camouflaged domain: most of the sequences represent failure cases for this model.

Results on sample test sequences through consecutive frames are reported in Fig. 1, Fig. 2, Fig. 3 and Fig. 4. These visualizations confirm that the proposed network is able to generalize over the test set of MoCA-Mask [1], returning a reasonable segmentation of unseen examples. In particular, Fig. 1 emphasizes the difference in the retrieved masks between our framework and the starting network $\mathcal{T}$: DeSC-V produces good masks even when $\mathcal{T}$ alone predicts only noise. In general, the fact that no method is perfect underlines the challenge of segmenting camouflaged animals.

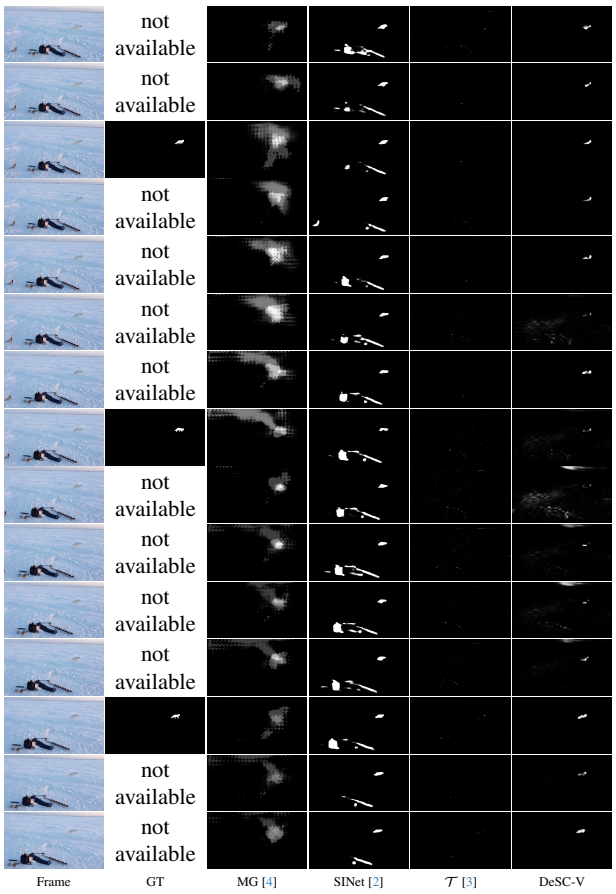

Figure 1. Sequence: arctic_fox_3 (frames 143 $\rightarrow$ 157) from MoCA-Mask [1]. As visible from the results, with respect to the starting network $\mathcal{T}$ (Pyramid Vision Transformer v2 [3] trained on COD10K [2]), DeSC-V is able to spot the fox in the majority of the cases. With respect to SINet [2], DeSC-V does not highlight the equipment as part of the subject. MG [4], being a motion-based method, shows some difficulties in highlighting solely the subject. Best viewed in coulor – zoom in for details –.

References

[1] Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In *CVPR*, pages 13864–13873, 2022.

[2] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE TPAMI*, 44(10): 6024–6042, 2022.

[3] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[4] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, pages 7177–7188, 2021.

[5] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *CVPR*, pages 879–888, 2019.

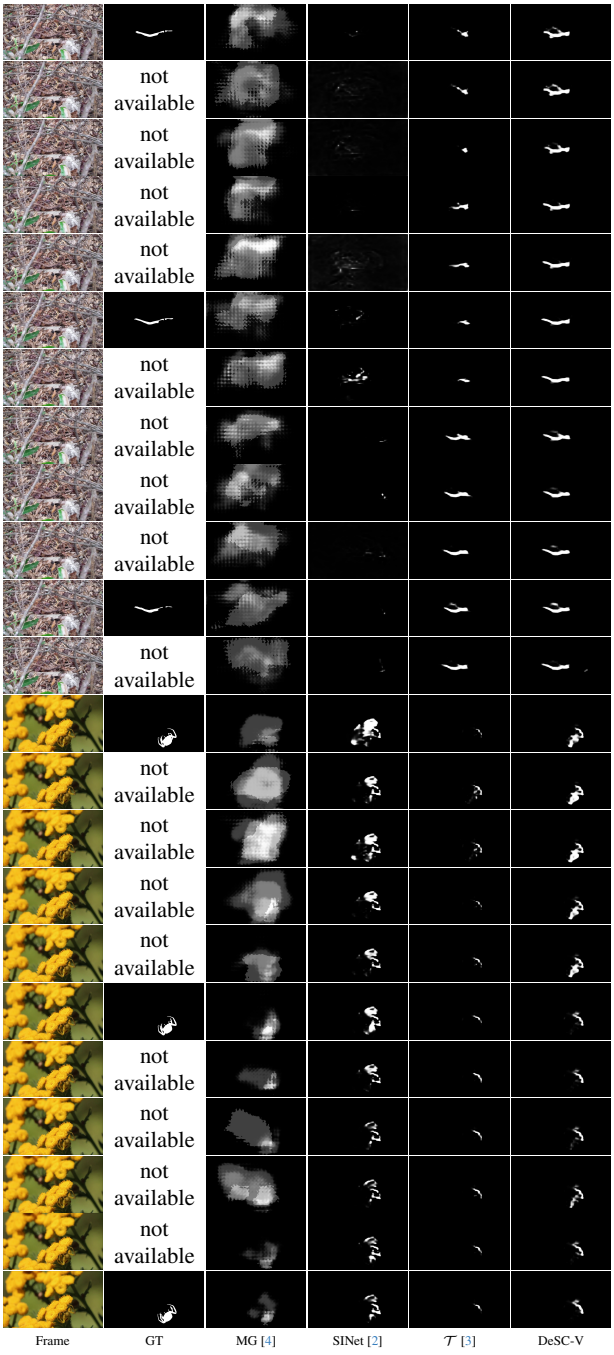

Figure 2. Sequences, from top to bottom: copperhead_snake (frames 65 $\rightarrow$ 76), flower_crab_spider_2 (frames 45 $\rightarrow$ 55). In both cases, MG [4] struggles in focusing solely on the subject, as its movement is not strongly evident. On the other hand, SINet [2] either fails in identifying the subject or tends to confuse other scenes' components with it. The starting network $\mathcal{T}$ [3] always performs a partial segmentation of the subject, which is often corrected by DeSC-V, resulting in the best qualitative segmentation, despite still not being perfect in the last part of the second sequence. Best viewed in colour – zoom in for details –.

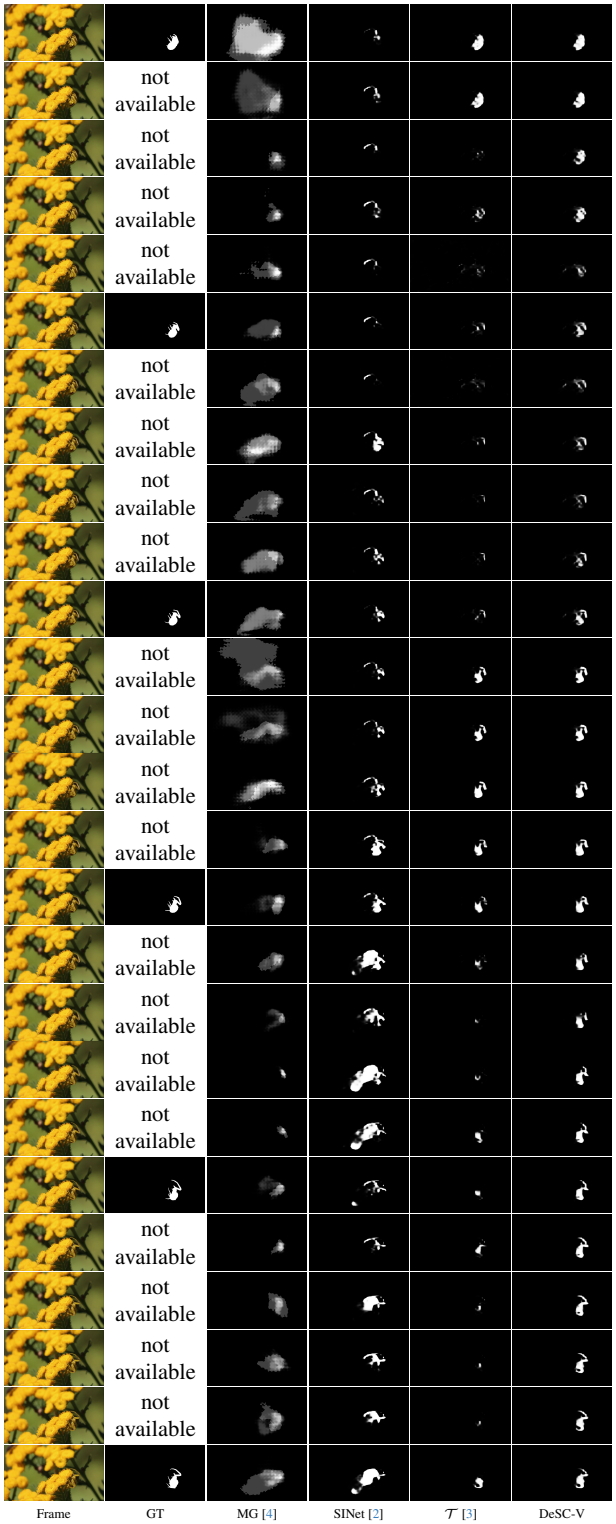

Figure 3. Sequence: flower_crab_spider_1 (frames 135 $\rightarrow$ 160). As visible, MG [4] fails to spot the spider due to scarce motion information, and SINet [2] also focuses on the flower. Thes tarting network $\mathcal{T}$ [3], instead, detects the subject only partially. Differently, DeSC-V returns the best result in most of the frames. Best viewed in colour – zoom in for details –.

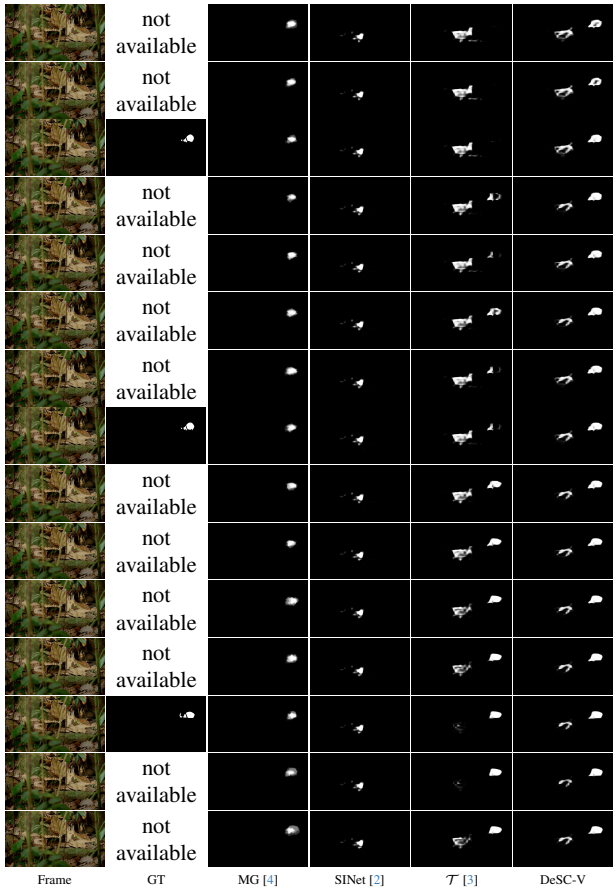

Figure 4. Sequence: rusty_spotted_cat_0 (frames 23 $\rightarrow$ 37). As visible from the results, this is a case in which MG [4], focusing solely on motion, is advantaged. However, it only partially segments the subject. SINet [2], on the other hand, is not able to focus on the subject, and gets fooled by other components of the scene. The starting network $\mathcal{T}$ [3] falls in the middle: it spots the cat better than [4] – especially in the last frames of the picture –, but also segments other components of the scene. On the other hand, DeSC-V performs an almost excellent identification of the subject, and even if it also segments other portions of the scene, it is significantly better than both [2] and [3], and, in contrast to [4], returns a complete subject segmentation. Best viewed in colour – zoom in for details –.