A. Supplementary material – Overview

In this supplementary material:

- we justify design choices and assumptions made in the main paper (Sections B-F);
- we show implementation details of our approach (Section G);
- we provide additional failure cases (Section H);
- we provide more qualitative results on DAVIS2016, SegTrack v2 and FBMS-59 (Section I).

B. Justifying only using the flow between adjacent frames

Our formulation, in Eq. (1) of the main paper, only involves the optical flow between adjacent frames (forward and backward). As discussed in Section 3 of the main paper, it can be related to the tridiagonal affinity matrix $\mathcal{A}$ in Eq. (4). We remark at the end of Section 3.3 that we could have used a denser matrix correlating more faraway frames, but that the optical flow between frames that are distant in time is less reliable.

To validate our choice of only using the optical flow between adjacent frames, we consider here the following variant of our objective function, where we introduce warps between more distant frames (up to some constant T):

$$\mathcal{L}(\{\mathbf{x}_p\}_p) = \sum_p \lambda \, \mathrm{CE}(\widehat{\mathbf{x}}_p, \mathbf{x}_p) + \sum_{t=1}^{T} \mathrm{CE}(\mathbf{x}_{p+t}, w_p^{p+t}(\mathbf{x}_p)) + \mathrm{CE}(\mathbf{x}_p, w_{p+t}^{p}(\mathbf{x}_{p+t}))\,, \tag{10}$$

Max. frame distance for optical flow	DAVIS2016 $\mathcal{J}\uparrow$	DAVIS2016 $\mathcal{F}\uparrow$
$T=1$	**76.8**	**77.0**
$T=2$	74.0	71.3
$T=3$	67.1	61.8

Table 3: **Study of the distance between frames for flow consistency enforcement.** We consider different values of T in Eq. (10) and evaluate on DAVIS2016 using our best configuration (DINO [ViT] + ARFlow + Opt) without CRF post-processing. The best performance is achieved for $T=1$, coinciding with the tridiagonal matrix configuration.

Table 3 shows that using a time horizon of a single frame is not only enough but actually better than considering the optical flow between more distant frames. In fact, using the flow regarding only the preceding and the succeeding frames already ties together all frames in the sequence. Additional terms with optical flows between more distant frames may actually introduce noise because of worse estimations due to larger displacements, deformations and occlusions.

C. Using the cross-entropy vs the dot product

In Section 3.4 of the main paper, we replace the dot products

$$(\widehat{\mathbf{x}}_p)^T \mathbf{x}_p\,, \quad \mathbf{x}_{p+1}^T w_p^{p+1}(\mathbf{x}_p)\,, \text{ and } \quad \mathbf{x}_p^T w_{p+1}^{p}(\mathbf{x}_{p+1})$$

by cross-entropies, respectively:

$$\mathrm{CE}(\widehat{\mathbf{x}}_p, \mathbf{x}_p), \mathrm{CE}(\mathbf{x}_{p+1}, w_p^{p+1}(\mathbf{x}_p)), \text{ and } \mathrm{CE}(\mathbf{x}_p, w_{p+1}^{p}(\mathbf{x}_{p+1})).$$

This was motivated empirically, as we observed that the cross-entropy was providing a better performance. Table 4 reports the quantitative results of this experiment.

Measurement of mask deviation	DAVIS2016 $\mathcal{J}\uparrow$	DAVIS2016 $\mathcal{F}\uparrow$
Cross-entropy	**76.8**	**77.0**
Dot product	62.1	60.4

Table 4: **Dot product vs cross-entropy.** Using the cross-entropy between two vectors in our objective function rather than their dot product leads to a significant improvement of **+14.6%** in $\mathcal{J}$ and **+16.6%** in $\mathcal{F}$.

D. On the Constant Norm Constraint in Eq. (3)

We estimate the second largest eigenvector of $\mathcal{W}$ via a maximization problem over a vector $\mathbf{X}$ under the constraint that $\|\mathbf{X}\|_2$ is constant, as stated in Eq. (3) in the main paper. At the end of Section 3.4, we claim that since the $\mathbf{x}_p$ vectors remain close to the $\widehat{\mathbf{x}}_p$ vectors, $\|\mathbf{X}\|_2 = \sqrt{\sum_p (\|\widehat{\mathbf{x}}_p\|_2)^2}$ remains approximately constant during optimization, thus satisfying the constraint in Eq. (3) up to a constant factor of $\sqrt{N}$.

Table 5 shows empirically that this constraint is indeed approximately met at each stage of the global optimization process.

E. Approximation in Eq. (8)

In the main paper, we assumed the following approximation:

$$\mathbf{x}_p^\top D_p^{-1} A_p \, \mathbf{x}_p \approx \widehat{\mathbf{x}}_p^\top \, \mathbf{x}_p\,. \tag{11}$$

L-BFGS iteration number	Theoretical norm of $\widehat{\mathbf{x}}_p$	Actual average norm of $\widehat{\mathbf{x}}_p$
1	1	1.003
2	1	1.001
3	1	1.007
4	1	1.022
5	1	1.022

Table 5: **Study of the constant norm constraint approximation.** We study the average of $\|\widehat{\mathbf{x}}_p\|_2$ over all frames of all sequences in the DAVIS2016 dataset at each iteration of our global optimization (using L-BFGS). We observe that the deviations to the theoretical norm of 1 are small, which in turns means that our approach of not using any explicit constraint is valid.

The derivation of this approximation is given below.

Since $W_p = D_p^{-1} A_p$ is a row-normalized stochastic matrix, the largest eigenvalue associated to its first eigenvector is 1. Besides, our initial mask estimate $\widehat{\mathbf{x}}_p$ is computed as the second largest eigenvector of W_p via Power Iteration Clustering (PIC) [44]. According to [52, 44], if K clusters are well-separated, then the significant eigenvalues of W_p, noted $\lambda_1 \geq \ldots \geq \lambda_K$, are such that $\lambda_i/\lambda_1 \approx 1$ for all $i \in \{1, ..., K\}$. Consequently, if the foreground object is well-separated from the background ($K \geq 2$), we may assume that $\lambda_2 \approx \lambda_1 = 1$. As $\widehat{\mathbf{x}}_p$ approximates the second largest eigenvector of W_p, we have:

$$W_p \, \widehat{\mathbf{x}}_p \approx \lambda_2 \, \widehat{\mathbf{x}}_p \approx \widehat{\mathbf{x}}_p \, . \tag{12}$$

Therefore, considering also that $\mathbf{x}_p$ deviates little from $\widehat{\mathbf{x}}_p$, *i.e.*, $\mathbf{x}_p \approx \widehat{\mathbf{x}}_p$, we have:

$$\begin{aligned} & \mathbf{x}_p^\top D_p^{-1} A_p \, \mathbf{x}_p \\ = \; & \mathbf{x}_p^\top W_p \, \mathbf{x}_p \\ \approx \; & \mathbf{x}_p^\top W_p \, \widehat{\mathbf{x}}_p \\ \approx \; & \mathbf{x}_p^\top \widehat{\mathbf{x}}_p \\ = \; & \widehat{\mathbf{x}}_p^\top \mathbf{x}_p \, . \end{aligned} \tag{13}$$

F. Dealing with Inaccurate Optical Flow

Our method relies on predicted optical flows. As they can be wrong or poor quality, they may introduce noise during the computation of the initial mask estimates and the optimization. In order to reduce the influence of this noise, we eliminate poor quality optical flow predictions. Given a predicted flow $\phi_{p,q}$, we first warp frame q to frame p. Next, we calculate the difference image between frame p and the reconstructed frame $\hat{p}$. The locations with high response on the difference image correspond to wrong or poor quality optical flow predictions. We use k-th percentile of the difference image as a threshold value to eliminate the poor quality optical flow predictions. The locations under the calculated threshold value indicate where optical flow fails to produce the accurate flows. We exclude these optical flow predictions from both the computation of the initial mask estimates and the optimization. We experimentally set k as 90-th percentile.

G. Implementation Details

All our experiments are implemented with PyTorch [58]. We use the L-BFGS [7] with learning rate of 1 to optimize our objective function. The weight λ in Eq. 1 is set to be 10.

We use DINO pretrained on ImageNet as appearance features. Due to their low resolution, we upscale the initial eigenvectors to the required resolution and run the full optimization pipeline. The optimized eigenvectors can be later either thresholded or clustered with K-means to obtain the final masks. We choose K-means as it is a more universal method that does not require finding threshold parameters. The final segments are then refined using CRF, as [90, 89].

We use the ARFlow model pretrained on the CityScapes [17] dataset in a self-supervised manner to predict optical flow. In ablation studies, we also use the RAFT model [73] for comparison, which is trained in a supervised manner with labeled data from the Sintel dataset [6].

H. Additional Failure Cases

In Figure 6, we show more failure examples of our approach. The first row shows another example of oversegmentation, where flowing particles are being segmented as foreground. The second row shows undersegmentation. The last row shows the inability of our approach to capture very fine details, such as the thin cables of the paraglider.

I. More Qualitative Results

On the next pages (Figures 7-19), we show more qualitative results of our approach, where we compare to the CIS [90] method, which is the third best self-supervised video object segmentation (VOS) method after ours, according to Table 2 in the main paper. (DyStab [89], which is the second best self-supervised VOS method, did not release code to rerun these experiments nor mask results).

Compared to CIS, our method is more successful at segmenting objects as a whole and capturing finer details of object boundaries.

J. Use of Existing Datasets and Codes

For the experiments, we used several datasets that are freely available for research purpose:

- DAVIS 2016[2] [60] is under license CC BY-NC 4.0.

[2] https://davischallenge.org

Ours Ground Truth

Figure 6: **Failure cases.** Our approach has three main failure modes: over-segmentation in scenes with multiple objects, under-segmentation, and inability to capture very fine details.

- SegTrack-v2[3] [42] is under a custom non-commercial, research-only license, courtesy of Georgia Institute of Technology,
- FBMS-59[4] [55, 4] is under a custom non-commercial, research-only license, courtesy of University of Freiburg.

To compute appearance and flow, we experimented with the following methods, whose code is freely available for research purpose:

- DINO[5] [9] is under the Apache License 2.0.
- MoCov2[6] [12] is under the CC BY-NC 4.0 license.
- ARFlow[7] [46] is under the MIT License.
- RAFT[8] [73] is under the BSD 3-Clause License.

K. Societal Impact

We believe that our approach for the self-supervised discovery and segmentation of objects in videos has **very little potential for malicious uses** (including disinformation, surveillance, invasion of privacy, endangering security), in any case not more, *e.g.*, than the hundreds of previously published methods on supervised object detection and segmentation. Moreover, we are not bound nor promoting any dataset that would lead to unfairness in any sense.

Besides, the use of our method has a **very little environmental impact** as there is no training phase and as the optimization is relatively fast and in the same order of magnitude as other approaches.

[3] https://web.engr.oregonstate.edu/~lif/SegTrack2/dataset.html

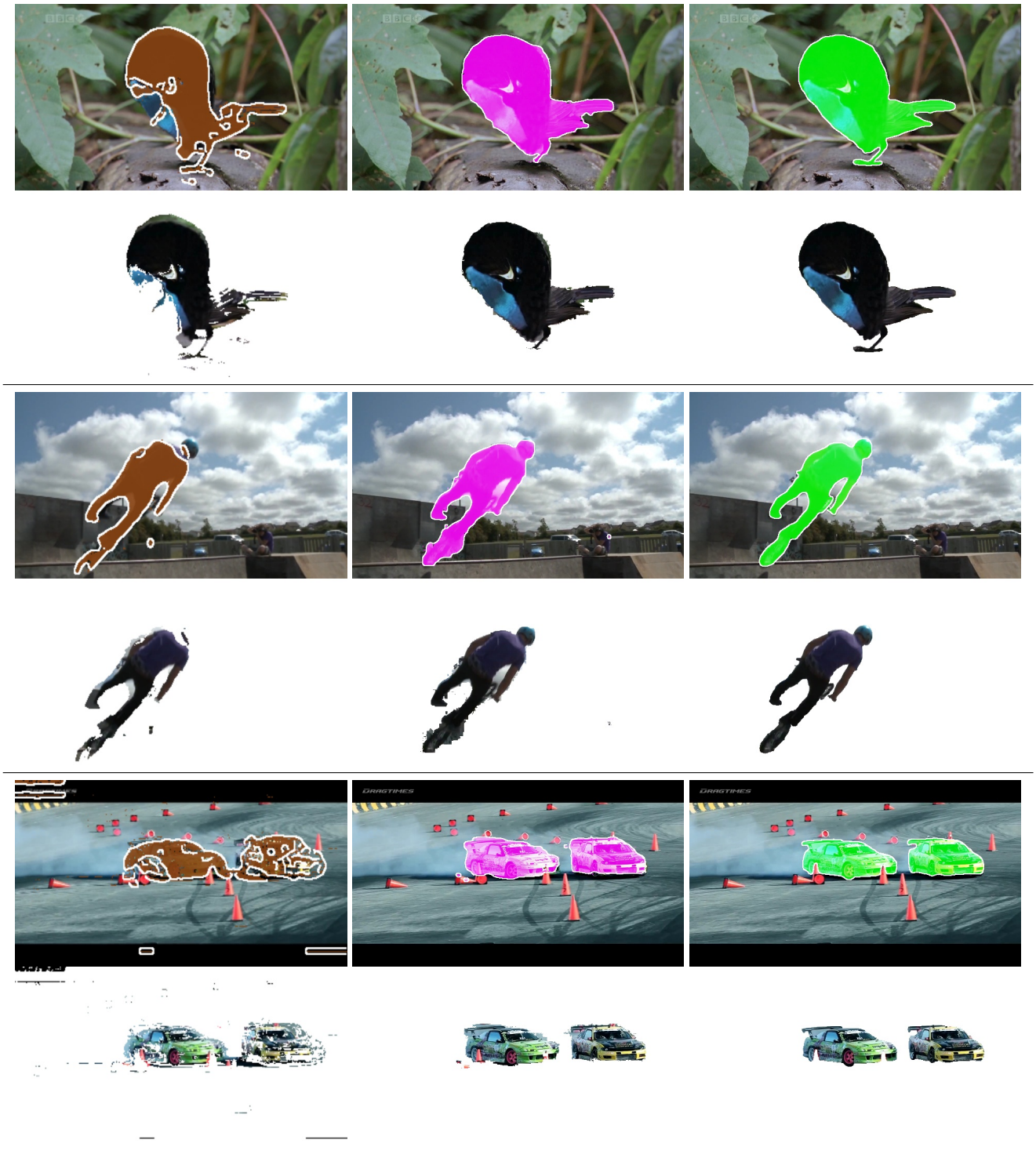
[4] https://lmb.informatik.uni-freiburg.de/resources/datasets/moseg.en.html

[5] https://github.com/facebookresearch/dino

[6] https://github.com/facebookresearch/moco

[7] https://github.com/lliuz/ARFlow

[8] https://github.com/princeton-vl/RAFT



CIS [90] Ours Ground Truth

Figure 7: Segmentation in sample frames from videos in SegTrack v2.

CIS [90] Ours Ground Truth

Figure 8: Segmentation in sample frames from videos in SegTrack v2.

CIS [90] Ours Ground Truth

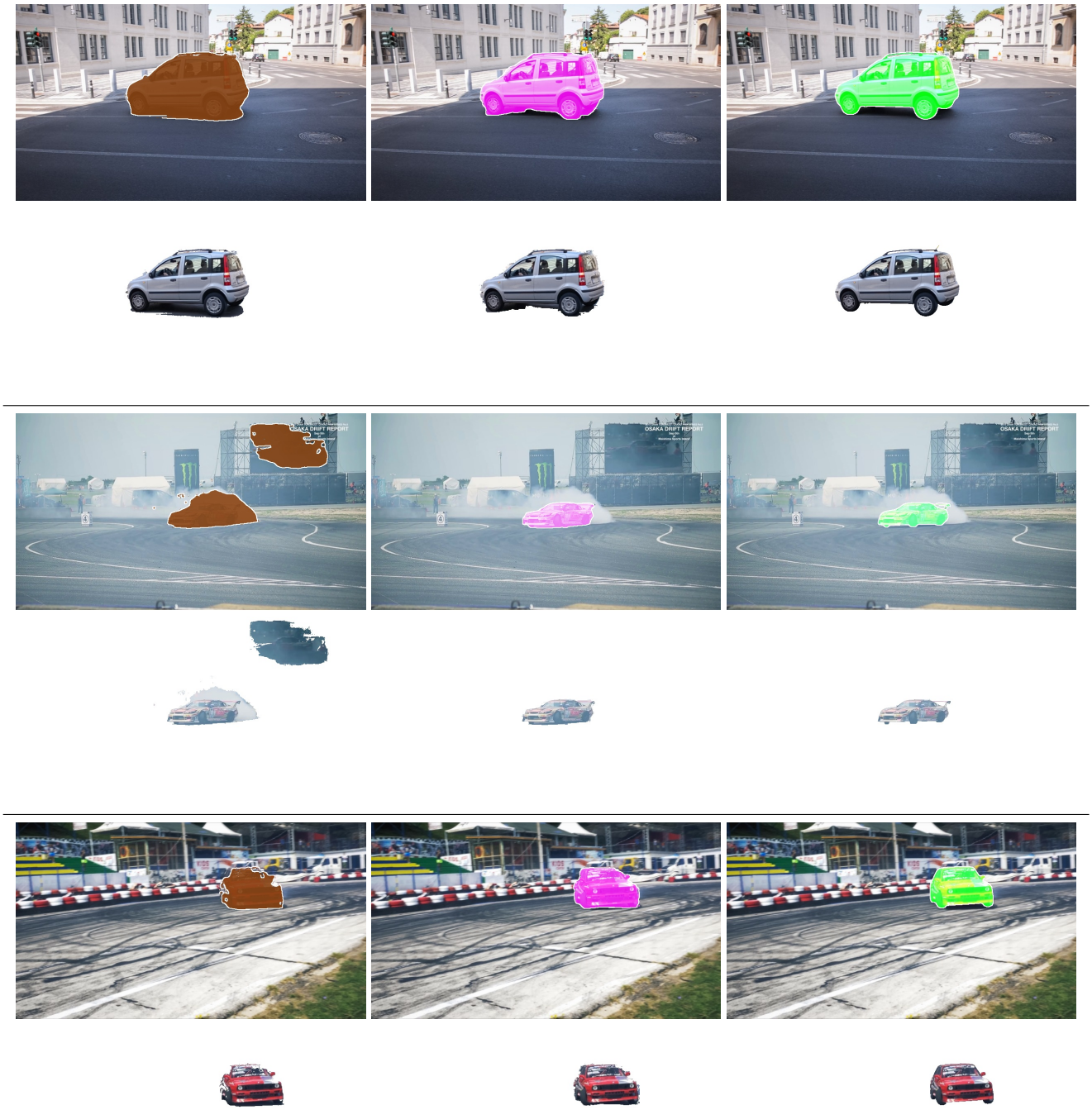
Figure 9: Segmentation in sample frames from videos in SegTrack v2.

CIS [90] Ours Ground Truth

Figure 10: Segmentation in sample frames from videos in DAVIS 2016.

CIS [90] Ours Ground Truth

Figure 11: Segmentation in sample frames from videos in DAVIS 2016.



CIS [90] Ours Ground Truth

Figure 12: Segmentation in sample frames from videos in DAVIS 2016.

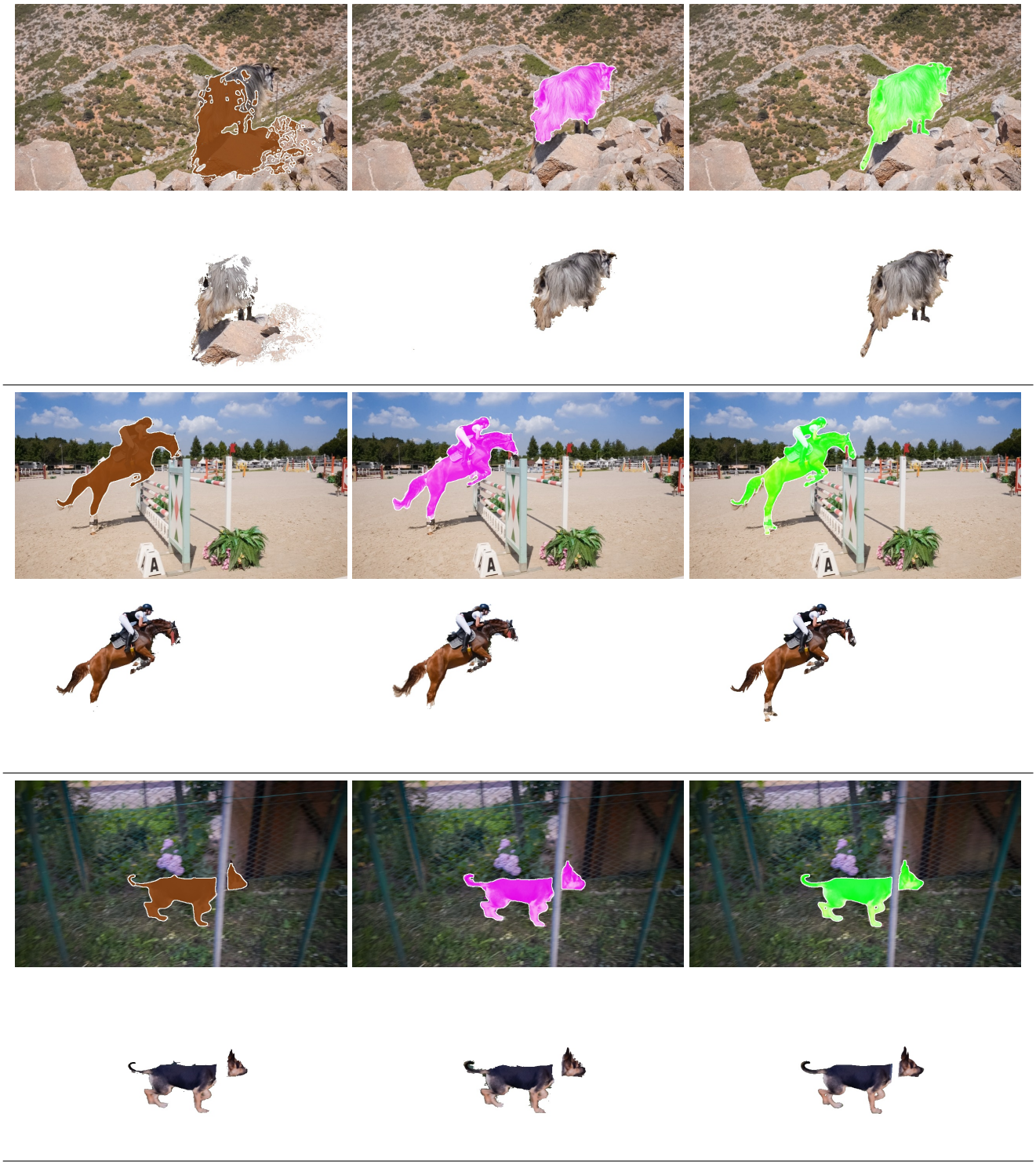


Figure 13: Segmentation in sample frames from videos in DAVIS 2016.

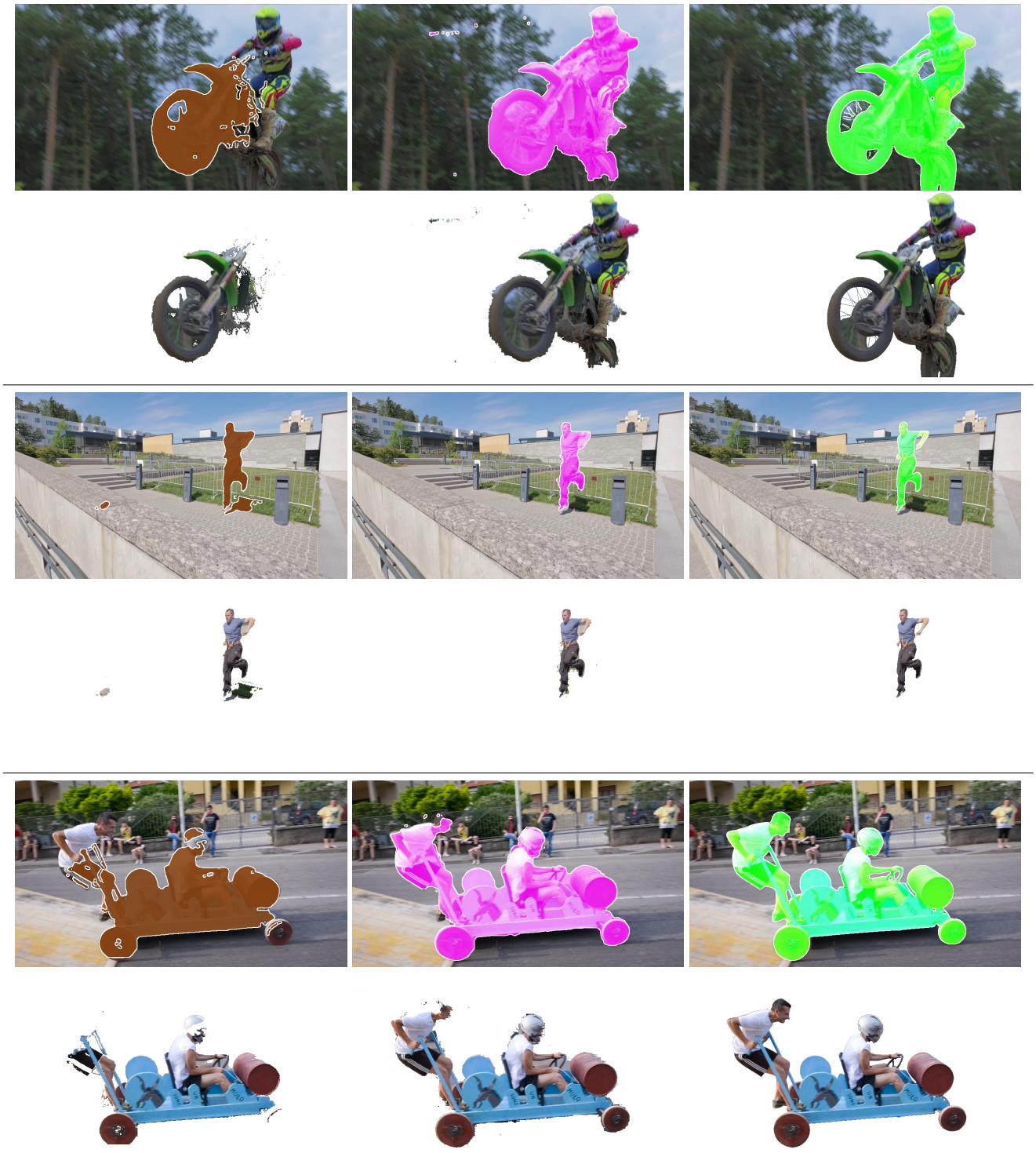
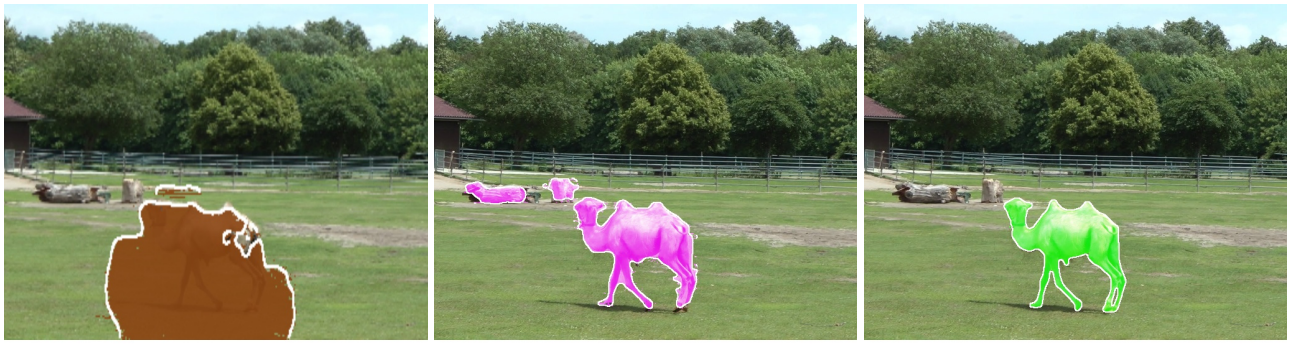
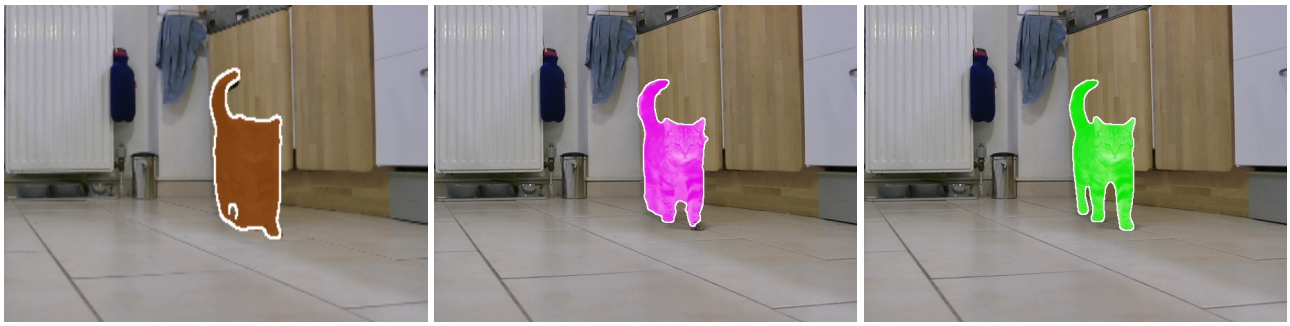
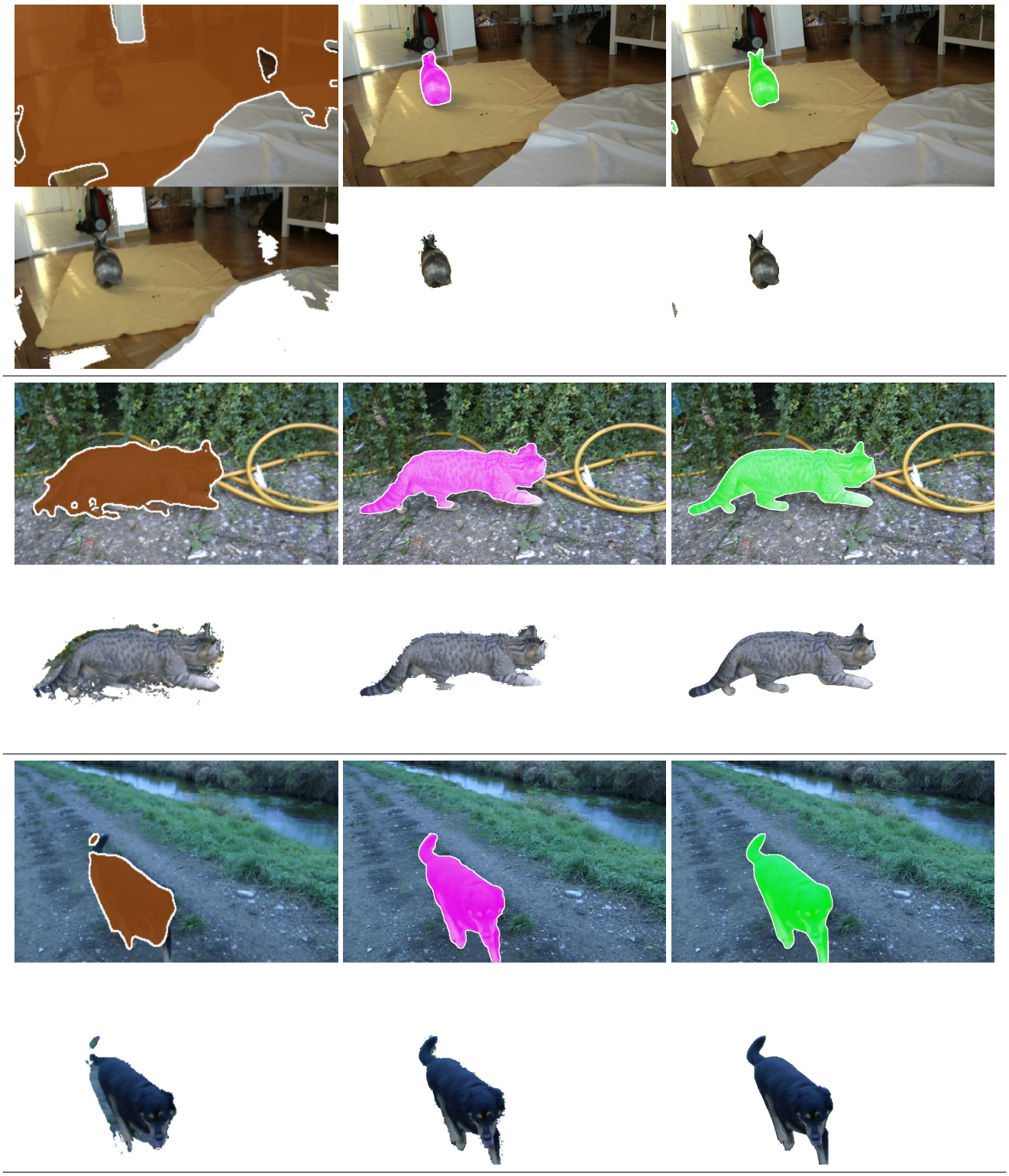


Figure 14: Segmentation in sample frames from videos in DAVIS 2016.

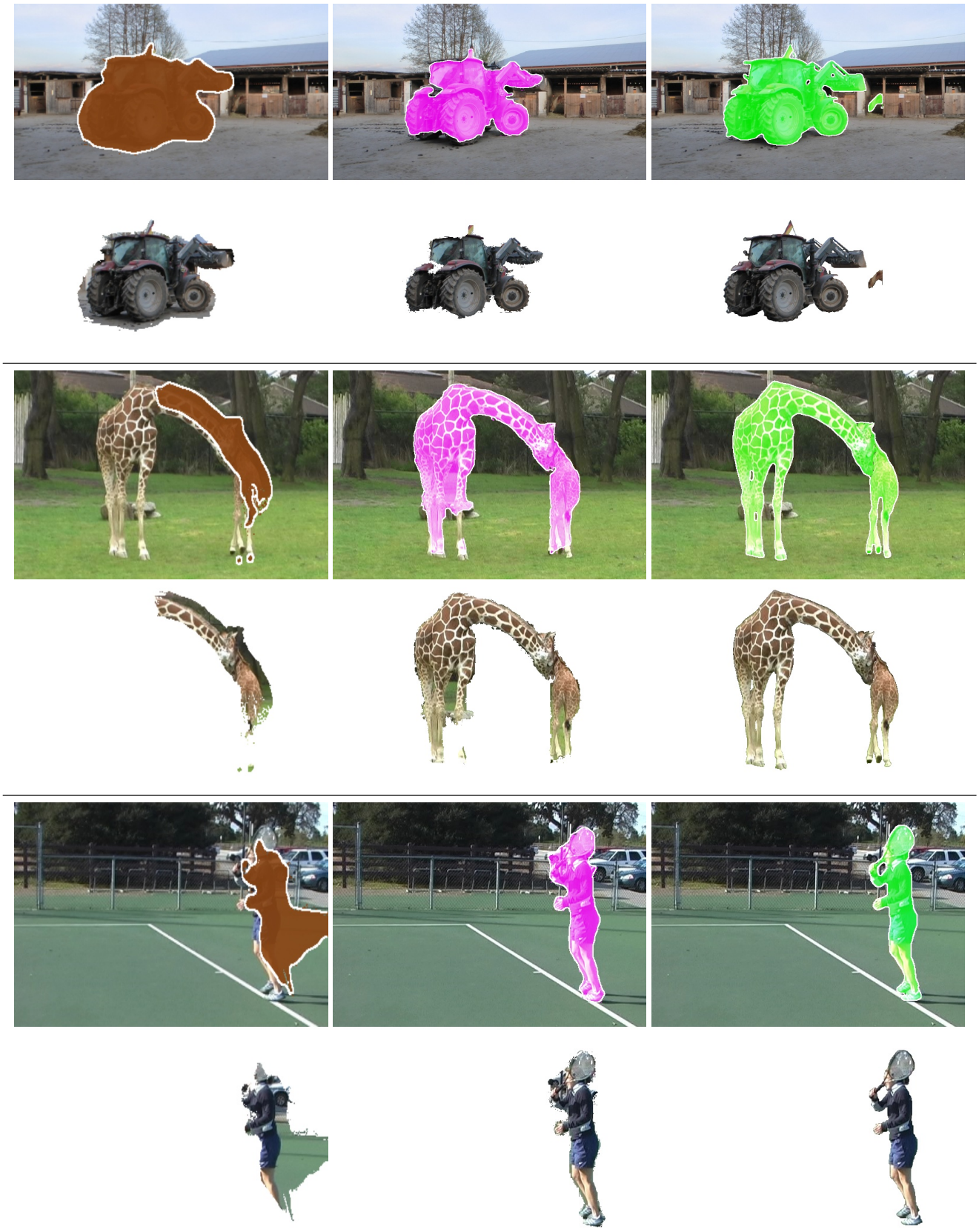






CIS [90] Ours Ground Truth

Figure 16: Segmentation in sample frames from videos in FBMS-59.



CIS [90] Ours Ground Truth

Figure 17: Segmentation in sample frames from videos in FBMS-59.

Figure 18: Segmentation in sample frames from videos in FBMS-59.

Figure 19: Segmentation in sample frames from videos in FBMS-59.