Instance Embedding Transfer to Unsupervised Video Object Segmentation Supplementary Materials

Siyang Li[1,2], Bryan Seybold[2], Alexey Vorobyov[2], Alireza Fathi[2], Qin Huang[1], and C.-C. Jay Kuo[1]
[1]University of Southern California, [2]Google AI Perception

1. Experiment Settings of Semi-supervised Video Object Segmentation

We extract the $N_S = 100$ seeds for the first frame (frame 0) and form image regions, as described in Sec. 3.2 and Sec. 3.3, respectively. Then we compare the image regions with the ground truth mask. For one image region A_j, if the ground truth mask G covers more than α of A_j, i.e.,

$$|A_j \bigcap G| \geq \alpha |A_j|, \tag{1}$$

where $|\cdot|$ denotes the area, the average embedding within the intersection is computed and added to the foreground embedding set S^0_{FG}. We set $\alpha = 0.7$.

For the background, if A_j does not intersect with G at all, i.e.,

$$A_j \bigcap G = \emptyset, \tag{2}$$

the average embedding in A_j is added to the background embedding set S^0_{BG}. A visual illustration is shown in Fig. 1.

Figure 1. **Left**: The image regions and the ground truth mask (in magenta) on frame 0. **Center**: The image regions (in magenta) whose average embeddings are used as foreground embeddings for the rest of frames. **Right**: The image regions (in blue) whose average embeddings are used as background embeddings for the rest of frames. Best viewed in color.

Then the foreground probability for a pixel on an arbitrary frame is obtained by Eqs. 13-15 and results are further refined by a dense CRF with identical parameters from the unsupervised scenario. We compare our results with multiple previous semi-supervised methods in Tab. 5 in the paper.

2. Per-video Results for DAVIS

The per-video result are shown for DAVIS *train* set and *val* set are listed in Tab. 1 and Tab. 2. The evaluation metric is the region similarity $\mathcal{J}$ mentioned in the paper. Note that we used the *train* set for ablation studies (Sec. 4.4), where the masks were *not* refined by dense CRF.

Sequence	ARP [2]	FSEG [1]	Ours	Ours + CRF
bear	0.920	0.907	0.935	**0.952**
bmx-bumps	0.459	0.328	0.431	**0.494**
boat	0.436	**0.663**	0.652	0.491
breakdance-flare	0.815	0.763	0.809	**0.843**
bus	**0.849**	0.825	0.848	0.842
car-turn	0.870	0.903	**0.923**	0.921
dance-jump	**0.718**	0.612	0.674	0.716
dog-agility	0.320	**0.757**	0.700	0.708
drift-turn	0.796	**0.864**	0.815	0.798
elephant	0.842	**0.843**	0.828	0.816
flamingo	**0.812**	0.757	0.633	0.679
hike	**0.907**	0.769	0.871	**0.907**
hockey	0.764	0.703	0.817	**0.878**
horsejump-low	0.769	0.711	0.821	**0.832**
kite-walk	0.599	0.489	0.598	**0.641**
lucia	0.868	0.773	0.863	**0.910**
mallard-fly	0.561	0.695	0.683	**0.699**
mallard-water	0.583	0.794	0.760	**0.811**
motocross-bumps	0.852	0.775	0.849	**0.884**
motorbike	**0.736**	0.407	0.685	0.708
paragliding	**0.881**	0.474	0.820	0.873
rhino	**0.884**	0.875	0.860	0.835
rollerblade	0.839	0.687	0.851	**0.896**
scooter-gray	0.705	**0.733**	0.686	0.655
soccerball	0.824	0.797	0.849	**0.905**
stroller	**0.857**	0.667	0.722	0.758
surf	**0.939**	0.881	0.870	0.902
swing	0.796	0.741	0.822	**0.868**
tennis	0.784	0.707	0.817	**0.866**
train	**0.915**	0.761	0.766	0.765
$\mathcal{J}$ mean	0.763	0.722	0.775	**0.795**

Table 1. Per-video results on DAVIS *train* set. The region similarity $\mathcal{J}$ is reported.

Sequence	ARP [2]	FSEG [1]	Ours	Ours + CRF
blackswan	**0.881**	0.812	0.715	0.781
bmx-trees	**0.499**	0.433	0.496	0.496
breakdance	**0.762**	0.512	0.485	0.559
camel	0.903	0.836	0.902	**0.929**
car-roundabout	0.816	**0.907**	0.880	0.890
car-shadow	0.736	0.896	0.929	**0.930**
cows	0.908	0.869	0.910	**0.925**
dance-twirl	**0.798**	0.704	0.781	0.797
dog	0.718	0.889	0.906	**0.938**
drift-chicane	**0.797**	0.596	0.760	0.782
drift-straight	0.715	0.811	**0.884**	0.857
goat	0.776	0.830	0.861	**0.864**
horsejump-high	0.838	0.652	0.794	**0.851**
kite-surf	0.591	0.392	0.569	**0.653**
libby	0.654	0.584	0.686	**0.742**
motocross-jump	**0.823**	0.775	0.754	0.773
paragliding-launch	0.601	0.571	0.595	**0.625**
parkour	0.828	0.760	0.852	**0.910**
scooter-black	**0.746**	0.688	0.727	0.743
soapbox	**0.846**	0.624	0.668	0.670
$\mathcal{J}$ mean	0.762	0.707	0.758	**0.786**

Table 2. Per-video results on DAVIS *val* set. The region similarity $\mathcal{J}$ is reported.

3. Instance Embedding Drifting

In Sec. 4.4 of the paper, we mentioned the "embedding drift" problem. Here we conduct another experiment to demonstrate that the embedding changes gradually with time. In this experiment, we extract foreground and background embeddings based on the ground truth masks for every frame. The embeddings from the first frame (frame 0) are used as references. We compute the average distance between the foreground/background embeddings from an arbitrary frame and the reference embeddings. Mathematically,

$$d_{FG}(k,0) = \frac{1}{|FG_k|} \sum_{j \in FG^k} \min_{l \in FG^0} ||\mathbf{f}(j) - \mathbf{f}(l)||_2, \quad (3)$$

$$d_{BG}(k,0) = \frac{1}{|BG_k|} \sum_{j \in BG^k} \min_{l \in BG^0} ||\mathbf{f}(j) - \mathbf{f}(l)||_2, \quad (4)$$

where FG^k and BG_k denote the ground truth foreground and background regions, respectively, $\mathbf{f}(j)$ denotes the embedding corresponding to pixel j, and $d_{FG}(k,0)$/$d_{BG}(k,0)$ represent the foreground/background embedding distance between frame k and frame 0. Then we average $d_{FG}(k,0)$ and $d_{BG}(k,0)$ across sequences and plot their relationship with the relative timestep in Fig. 2. As we observe, the embedding distance is increasing with time elapsing. Namely, both objects and background become less similar to themselves on frame 0, which supports the necessity of online adaptation.

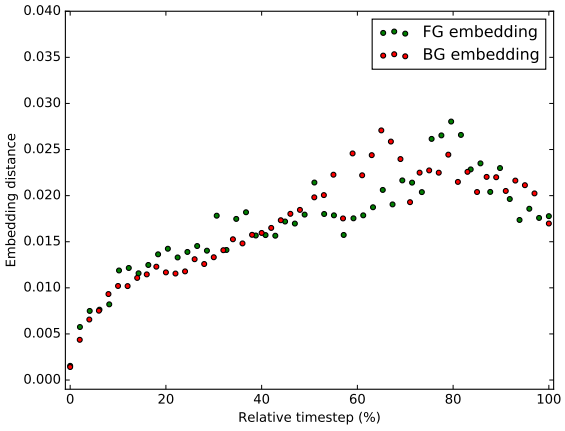

Figure 2. The FG/BG distance between later frames and frame 0. Both FG/BG embeddings become farther from their reference embedding on frame 0.

4. More visual examples

We provide more visual examples for the DAVIS dataset [4] and the FBMS dataset [3] in Fig. 3 and Fig. 4, respectively.

References

[1] S. D. Jain, B. Xiong, and K. Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2

[2] Y. J. Koh and C.-S. Kim. Primary object segmentation in videos based on region augmentation and reduction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2

[3] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1187–1200, 2014. 2

[4] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

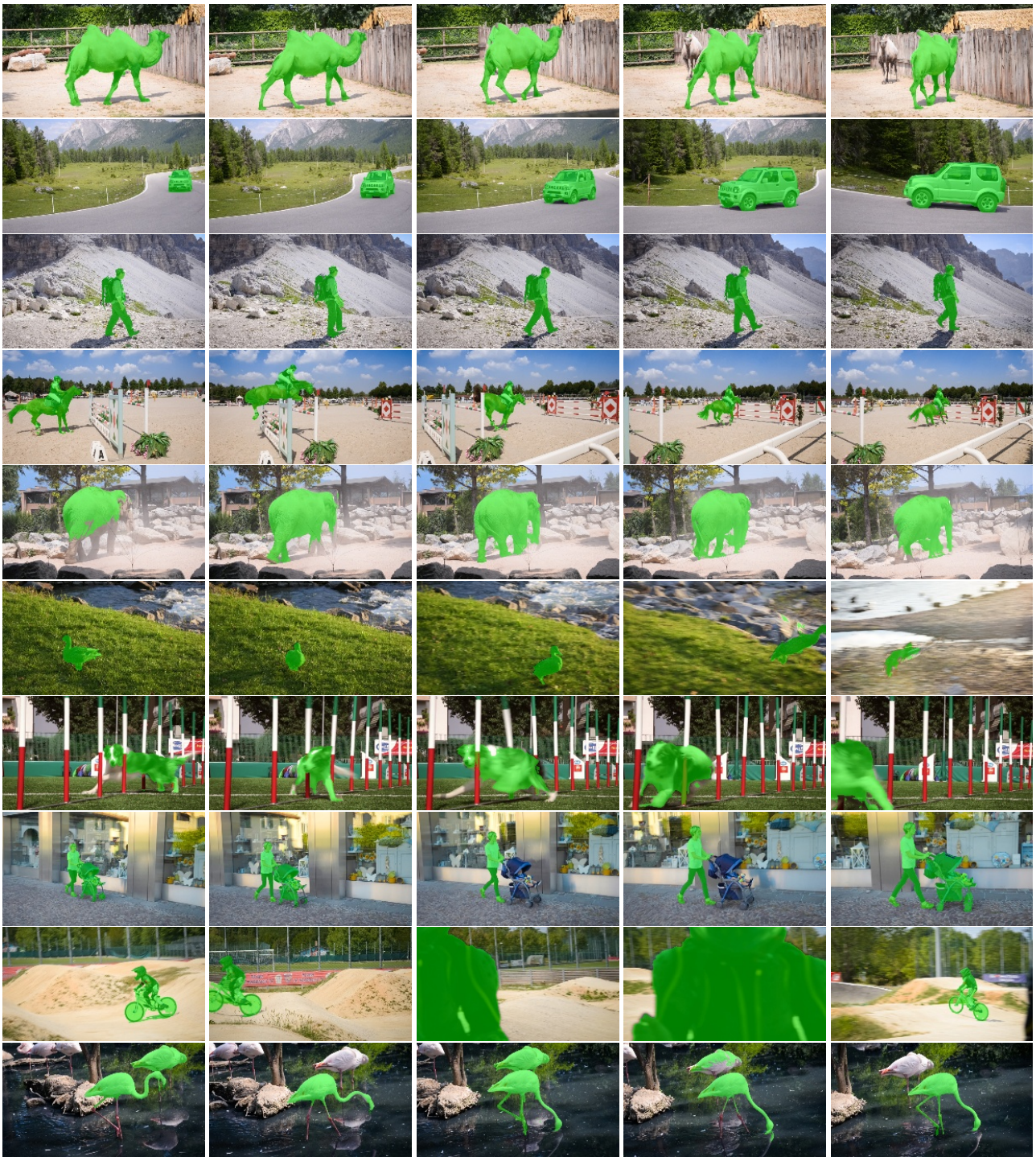

Figure 3. Visual examples from the DAVIS dataset. The “camel” sequence (first row) is mentioned as an example where the static camel (the one not covered by our predicted mask) acts as hard negatives because it is semantically similar with foreground while belongs to the background. Our method correctly identifies it as background from motion saliency. The last three rows show some failure cases. In the ”stroller” sequence (third last row), our method fails to include the stroller for some frames. In the ”bmx-bump” sequence (second last row), when the foreground, namely the rider and the bike, is totally occluded, our method wrongly identifies the occluder as foreground. The “flamingo” sequence (last row) illustrates a similar situation with the “camel” sequence, where the proposed method does less well due to imperfect optical flow (the foreground mask should include only the flamingo located in the center of each frame). Best viewed in color.

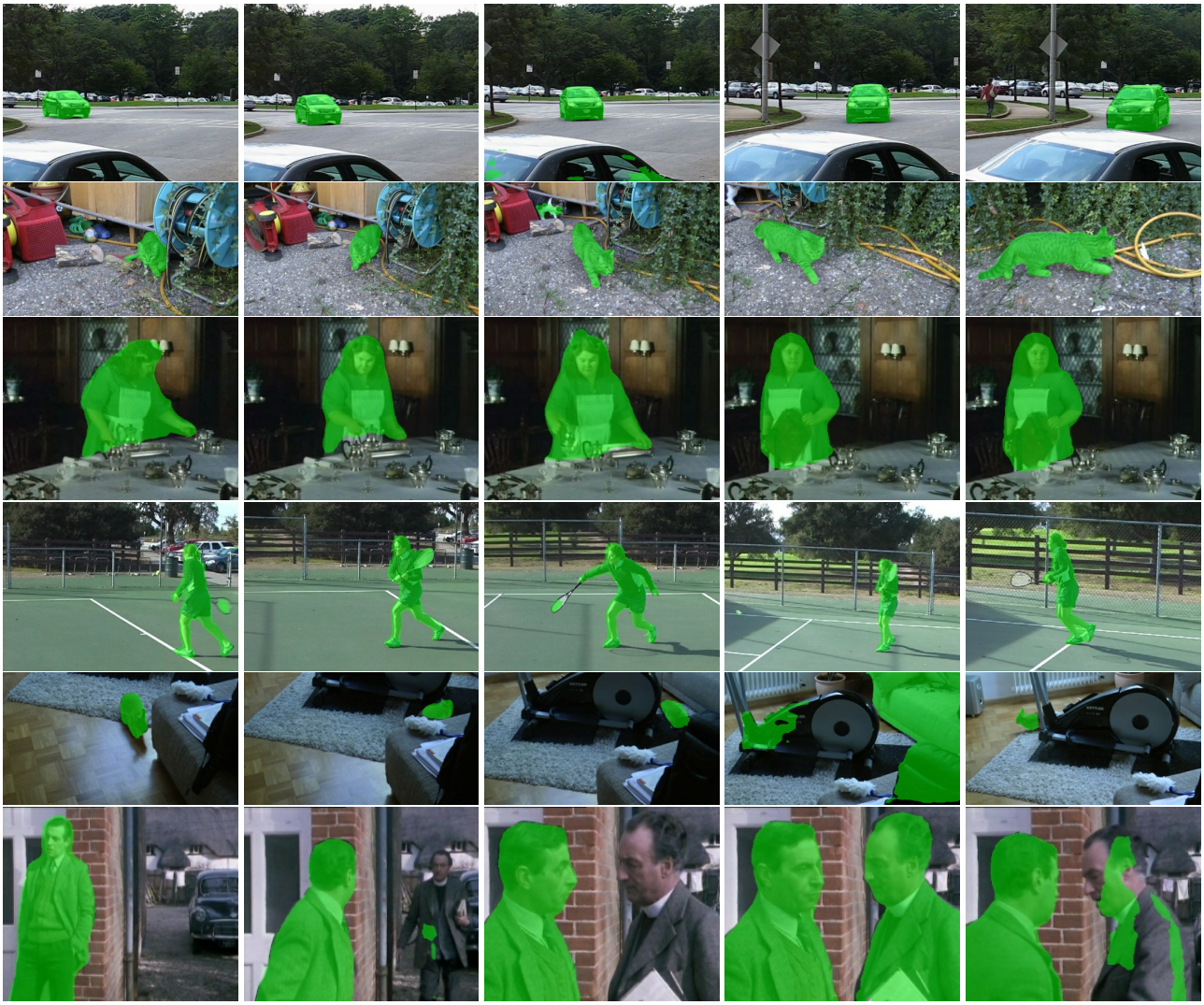

Figure 4. Visual examples from the FBMS dataset. The last two rows show some failure cases. In the "rabbits04" sequence (second last row), the foreground is wrongly identified when the rabbit is wholly occluded. In the "marple6" sequence (last row), the foreground should include two people, but our method fails on some frames because one of them demonstrates low motion saliency. Best viewed in color.