Unseen Object Segmentation in Videos via Transferable Representations

Yi-Wen Chen[1,2], Yi-Hsuan Tsai[3], Chu-Ya Yang[1],
Yen-Yu Lin[1], Ming-Hsuan Yang[4,5]

[1]Academia Sinica, [2]National Taiwan University, [3]NEC Laboratories America,
[4]University of California, Merced, [5]Google Cloud

1 Analysis of Transferring Visual Information

We analyze the proposed method for transferring visual information by investigating the weights of the transferable layer. Table 1 presents the learned weights of the transferable layer on the DAVIS dataset for unseen object categories. For each target video, the source categories with higher weights are similar to the target video in appearance, which gives reasonable transform of visual information.

Table 1. Learned weights of the transferable layer on the DAVIS dataset for transferring knowledge from seen/source categories (rows) to unseen/target object categories (columns). For each unseen category, the largest weight over all seen categories is marked in bold.

Sequence	bear	bswan	camel	eleph	goat	malw	rhino
aero	0.286	0.419	0.381	0.412	0.279	0.430	0.325
bike	0.317	0.372	0.393	0.423	0.358	0.309	0.432
bird	0.624	**0.891**	0.538	0.572	0.614	**0.780**	0.595
boat	0.392	0.419	0.358	0.460	0.323	0.474	0.428
bottle	0.401	0.336	0.307	0.410	0.349	0.387	0.368
bus	0.392	0.262	0.266	0.440	0.306	0.200	0.327
car	0.488	0.317	0.469	0.559	0.379	0.292	0.508
cat	**0.756**	0.436	0.417	0.574	0.609	0.398	0.492
chair	0.507	0.314	0.406	0.528	0.466	0.362	0.450
cow	0.701	0.409	0.715	0.748	0.618	0.346	**0.846**
table	0.341	0.310	0.186	0.301	0.291	0.504	0.257
dog	0.700	0.476	0.534	0.603	**0.788**	0.417	0.576
horse	0.547	0.330	**0.898**	**0.770**	0.692	0.260	0.776
mbike	0.301	0.287	0.346	0.408	0.371	0.287	0.355
person	0.504	0.429	0.731	0.639	0.554	0.366	0.629
plant	0.463	0.418	0.364	0.437	0.428	0.451	0.474
sheep	0.721	0.525	0.491	0.662	0.616	0.348	0.605
sofa	0.366	0.309	0.366	0.447	0.404	0.291	0.412
train	0.298	0.260	0.343	0.488	0.320	0.204	0.419
tv	0.369	0.252	0.277	0.425	0.271	0.248	0.303

2 Runtime Performance

The runtime performance is shown in Table 2. All the timings are measured on a machine with 2.5 GHz Intel Xeon CPU. We compute the optical flow [7] and utilize the minimum barrier distance [13] to generate motion prior using MATLAB implementations. In the proposed formulation, feature extraction, segment mining, and CNN model training are implemented using Python and Caffe library on a GPU of NVIDIA GTX 1080 Ti with 11 GB memory. The CNN model is fine-tuned for 2000 iterations. Note that, we report the timings for each component during iterative optimization averaged on all the frames.

Table 2. Runtime performance on the DAVIS dataset.

Stage	Time (second)
Motion prior computing (per pair of frames)	0.01
Feature extraction (per frame)	1.72
Segment mining (per frame)	0.01
CNN model training (per frame)	7.31

3 Per-video Results on the DAVIS 2016 Dataset

In Table 3, we present the results of each video from the DAVIS 2016 dataset under weakly-supervised and unsupervised settings. We show that the proposed algorithm achieves better performance than the state-of-the-art methods in most videos.

Table 3. Per-video results on the DAVIS 2016 dataset.

	Weak Supervision			No Supervision				
Methods	SPFTN [12]	FCN [8]	Ours	MSG [10]	FST [9]	NLC [1]	FSEG [4]	Ours
bear	74.8	80.3	**89.8**	85.1	89.8	90.7	91.5	**91.8**
bswan	**87.6**	75.6	76.7	52.6	73.2	87.5	89.5	**90.3**
bumps	29.7	29.9	**36.2**	35.3	24.1	**63.5**	38.8	42.1
trees	35.0	29.2	**40.5**	18.8	18.0	21.2	34.7	**38.9**
boat	35.9	63.4	**67.0**	14.4	36.1	0.7	63.8	**63.8**
bdan	37.1	14.6	**46.0**	23.6	46.7	**67.3**	14.2	13.1
bdanF	70.0	51.4	**80.0**	15.7	61.6	**80.4**	54.9	62.7
bus	**81.5**	61.1	81.2	**88.5**	82.5	62.9	80.4	80.5
camel	**76.2**	70.9	72.0	75.6	56.2	76.8	76.4	**77.5**
carR	76.8	71.0	**88.8**	63.0	**80.8**	50.9	74.8	79.6
carS	78.1	87.1	**92.5**	88.0	69.8	64.5	88.4	**93.3**

carT	75.4	86.7	**90.4**	62.1	85.1	83.3	90.7	**92.5**
cows	77.0	85.7	**88.1**	79.9	79.1	88.3	88.0	**88.3**
jump	34.2	33.6	**63.8**	6.5	59.8	**71.8**	10.3	11.2
twirl	46.1	27.8	**65.5**	36.6	45.3	34.7	**46.2**	41.0
dog	85.6	71.2	**89.1**	33.1	70.8	80.9	90.4	**91.6**
dogA	7.1	39.3	**72.9**	11.0	28.0	65.2	**68.9**	65.1
drtC	55.9	58.9	**67.1**	**75.8**	66.7	32.4	46.1	65.1
drtS	62.3	69.9	**79.4**	57.5	**68.3**	47.3	67.2	66.4
drtT	67.8	76.4	**80.6**	63.8	53.3	15.4	85.1	**89.7**
eleph	**75.6**	70.4	73.8	68.9	82.4	51.8	**86.2**	85.7
flamg	**38.1**	33.5	34.5	79.4	**81.7**	53.9	44.5	47.8
goat	72.8	83.1	**83.3**	73.5	55.4	1.0	84.1	**84.8**
hike	**89.3**	84.1	79.0	60.3	88.9	**91.8**	82.5	83.4
hockey	60.2	72.7	**73.1**	71.3	46.7	**81.0**	66.0	70.7
hjH	35.1	**77.6**	67.0	73.4	57.8	**83.4**	71.1	72.1
hjL	41.1	**79.5**	73.6	68.2	52.6	65.1	70.2	**76.5**
ksurf	**58.3**	55.8	46.5	41.9	27.2	45.3	47.7	**49.0**
kwalk	**73.3**	52.1	48.9	59.7	64.9	**81.3**	52.7	51.3
libby	50.8	49.5	**59.4**	5.0	50.7	63.5	67.7	**68.1**
lucia	83.3	**84.2**	78.9	41.7	64.4	**87.6**	79.9	81.0
malf	**70.8**	47.5	45.8	3.3	60.1	61.7	74.6	**75.2**
malw	**65.8**	40.9	41.6	4.5	8.7	76.1	83.3	**84.9**
motob	75.0	**77.7**	71.6	46.6	61.7	61.4	83.8	**85.2**
motoj	60.8	61.5	**65.5**	61.8	60.2	25.1	**80.4**	77.2
mbike	47.6	**78.5**	58.4	**73.8**	55.9	71.4	28.7	38.6
parag	**72.6**	30.9	28.1	**93.3**	72.5	88.0	17.7	5.5
paral	**62.8**	57.0	58.1	51.2	50.6	**62.8**	58.9	59.4
park	67.7	**84.0**	78.2	29.5	45.8	**90.1**	79.4	79.5
rhino	55.2	57.7	**71.0**	**90.2**	77.6	68.2	77.6	86.0
rolb	12.5	64.2	**73.2**	80.1	31.8	**81.4**	63.3	72.7
scbla	58.8	45.0	**72.1**	**57.9**	52.2	16.2	36.1	36.4
scgra	67.0	**73.7**	72.9	34.5	32.5	58.7	73.2	**75.7**
sobox	**57.8**	47.5	51.9	**67.2**	41.0	63.4	49.7	47.4
socB	49.0	**49.5**	46.3	37.0	**84.3**	82.9	29.3	28.3
strol	**65.4**	58.7	58.7	**67.8**	58.0	84.9	63.9	62.8
surf	**87.0**	78.4	79.1	77.0	47.5	77.5	88.8	**91.2**
swing	75.5	75.5	**76.4**	62.2	43.1	**85.1**	73.8	74.0
tennis	62.5	**78.2**	73.0	59.0	38.8	**87.1**	76.9	78.4
train	73.6	46.9	**77.3**	**88.7**	83.1	72.9	42.5	51.1
Avg.	61.2	61.6	**67.7**	54.3	57.5	64.1	64.7	**66.5**

4 Results on the SegTrack v2 Dataset

In Table 4, we provide experiments on the SegTrack v2 dataset [6] that contains many unseen objects. We use the ResNet-101 architecture and the training data from PASCAL VOC, which is the same setting as the appearance stream in FSEG [4]. We show that the proposed method performs better than FSEG [4], other unsupervised algorithms [9,5] and HVS [2] which includes human annotations in the procedure.

Table 4. Results on the SegTrack v2 dataset.

Methods	FST [9]	KEY [5]	HVS [2]	FSEG [4]	Ours
Avg. IoU	53.6	57.3	50.8	56.9	58.1

5 Segmentation Results

We show segmentation results compared to state-of-the-art approaches on the DAVIS dataset for unseen object categories in Fig. 1-2. In the supplementary video, we present more results for each sequence with unseen categories and compare our method with baseline settings and the state-of-the-art transfer learning approach [3]. In addition, we show results using the weakly-supervised setting on the DAVIS (Fig. 3-6) and YouTube-Objects (Fig. 7-8) datasets. Some failure cases are presented in Fig. 9.

References

1. Faktor, A., Irani, M.: Video segmentation by non-local consensus voting. In: BMVC (2014) 2
2. Grundmann, M., Kwatra, V., Han, M., Essa, I.: Efficient hierarchical graph-based video segmentation. In: CVPR (2010) 4
3. Hong, S., Oh, J., Lee, H., Han, B.: Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In: CVPR (2016) 4, 5, 6
4. Jain, S., Xiong, B., Grauman, K.: Fusionseg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos. In: CVPR (2017) 2, 4
5. Lee, Y.J., Kim, J., Grauman, K.: Key-segments for video object segmentation. In: ICCV (2011) 4
6. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: ICCV (2013) 4
7. Liu, C.: Beyond pixels: Exploring new representations and applications for motion analysis. PhD thesis, MIT (2009) 2
8. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015) 2, 7, 8, 9, 10, 11, 12
9. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: ICCV (2013) 2, 4, 5, 6, 7, 8, 9, 10
10. P.Ochs, T.Brox: Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions. In: ICCV (2011) 2, 5, 6, 7, 8, 9, 10

11. Taylor, B., Karasev, V., Soatto, S.: Causal video object segmentation from persistence of occlusions. In: CVPR (2015) 5, 6, 7, 8, 9, 10
12. Zhang, D., Yang, L., Meng, D., Xu, D., Han, J.: Spftn: A self-paced fine-tuning network for segmenting objects in weakly labelled videos. In: CVPR (2017) 2
13. Zhang, J., Sclaroff, S., Lin, Z., Shen, X., Price, B., Mech, R.: Minimum barrier salient object detection at 80 fps. In: ICCV (2015) 2

Fig. 1. Sample results on the DAVIS dataset for unseen object categories. Our final results contain less noisy segments and more details than other approaches and our baseline methods.

Fig. 2. Sample results on the DAVIS dataset for unseen object categories. Our final results contain less noisy segments and more details than other approaches and our baseline methods.

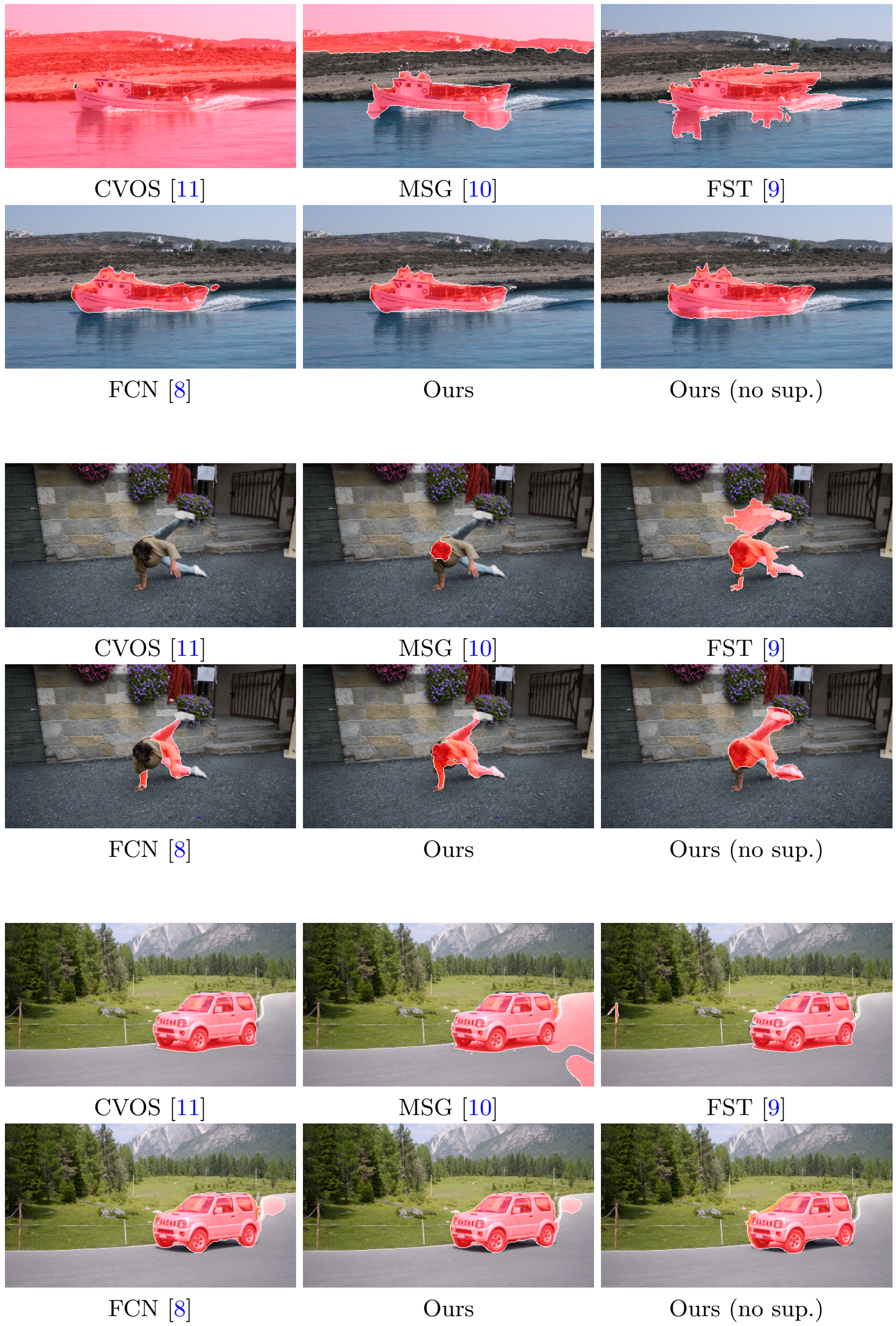


Fig. 3. Sample results on the DAVIS dataset with categories shared in the PASCAL VOC dataset. We show that our results with and without weak supervisions have more complete object segments with details.

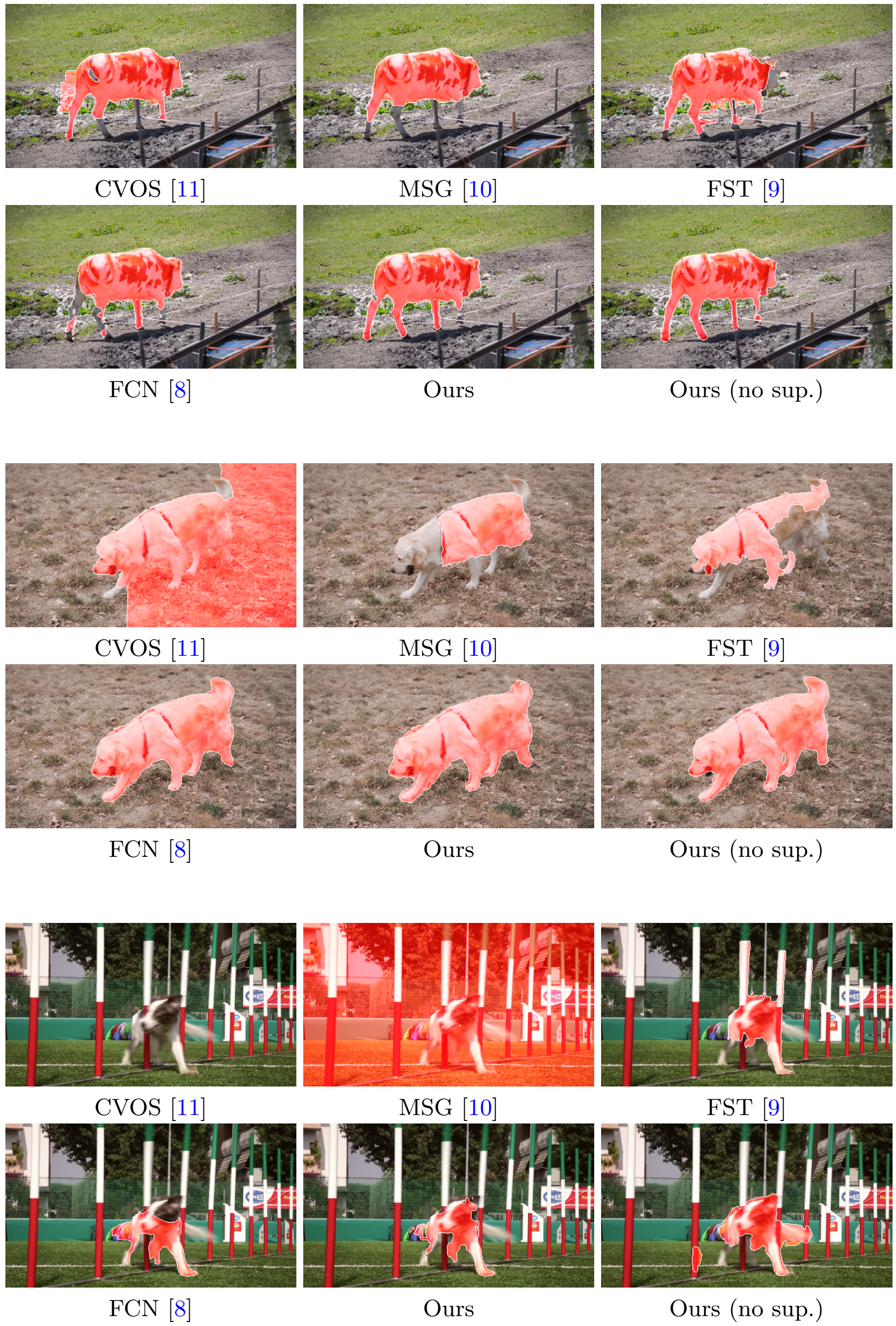


Fig. 4. Sample results on the DAVIS dataset with categories shared in the PASCAL VOC dataset. We show that our results with and without weak supervisions have more complete object segments with details.

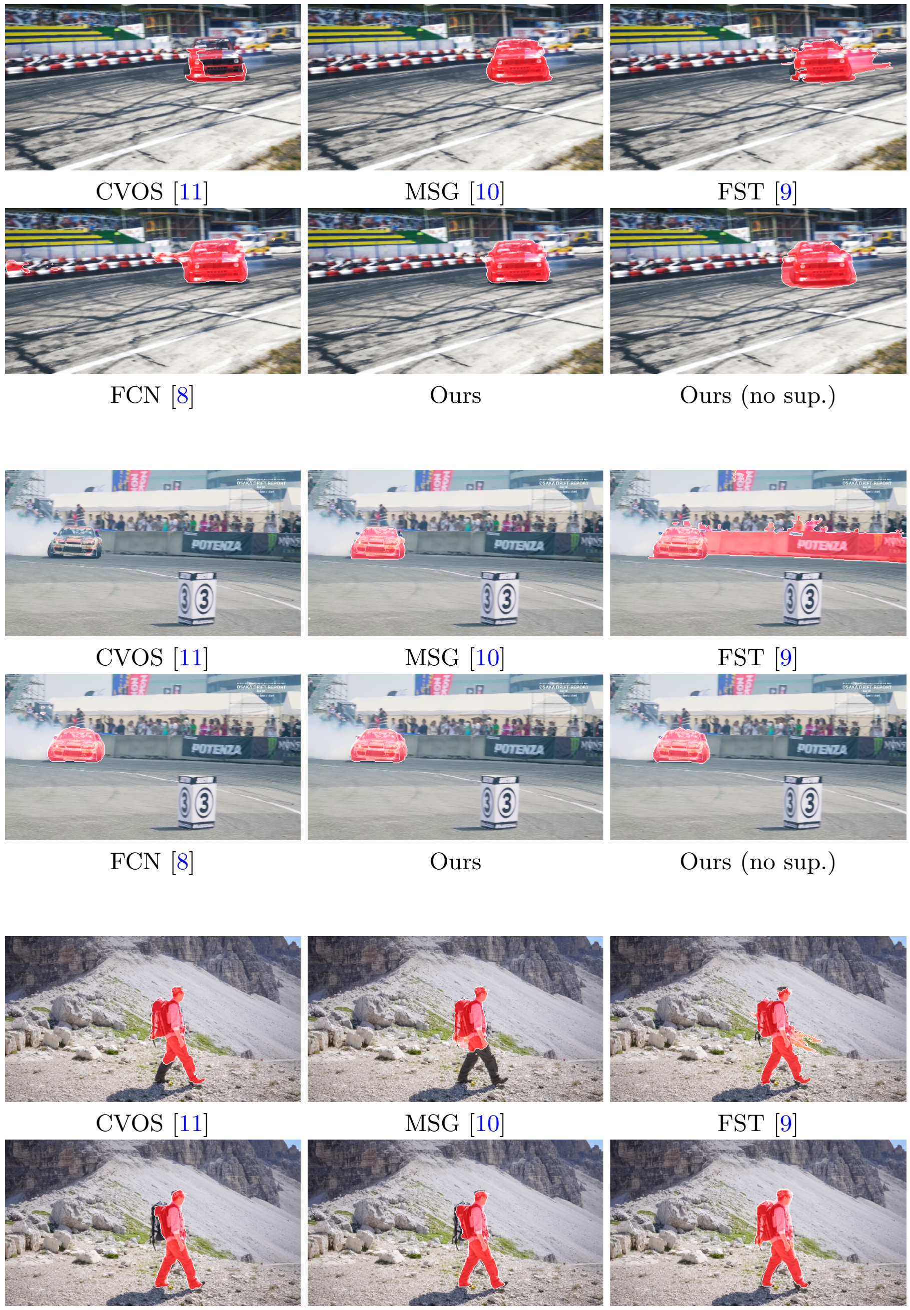


Fig. 5. Sample results on the DAVIS dataset with categories shared in the PASCAL VOC dataset. We show that our results with and without weak supervisions have more complete object segments with details.

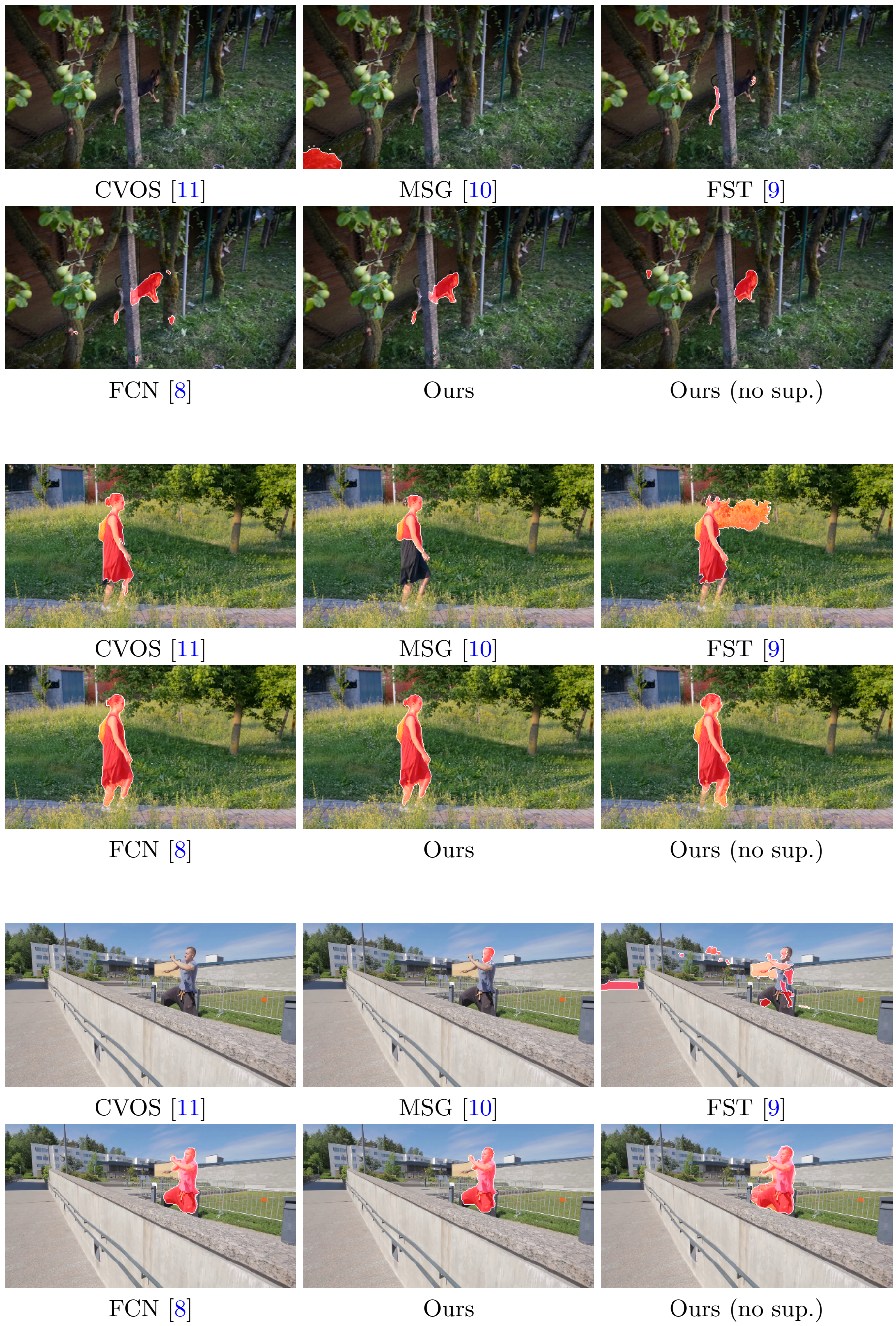
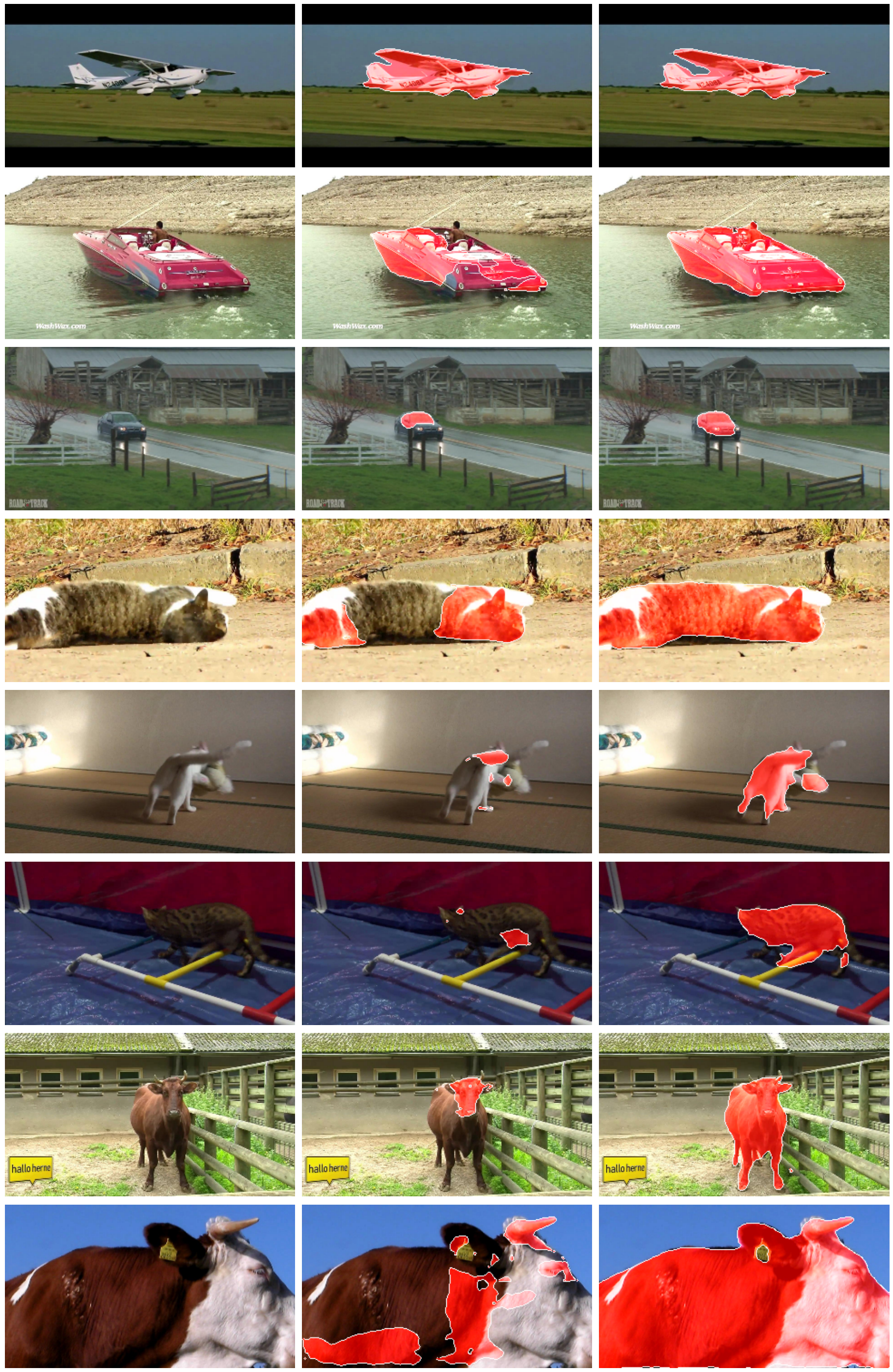
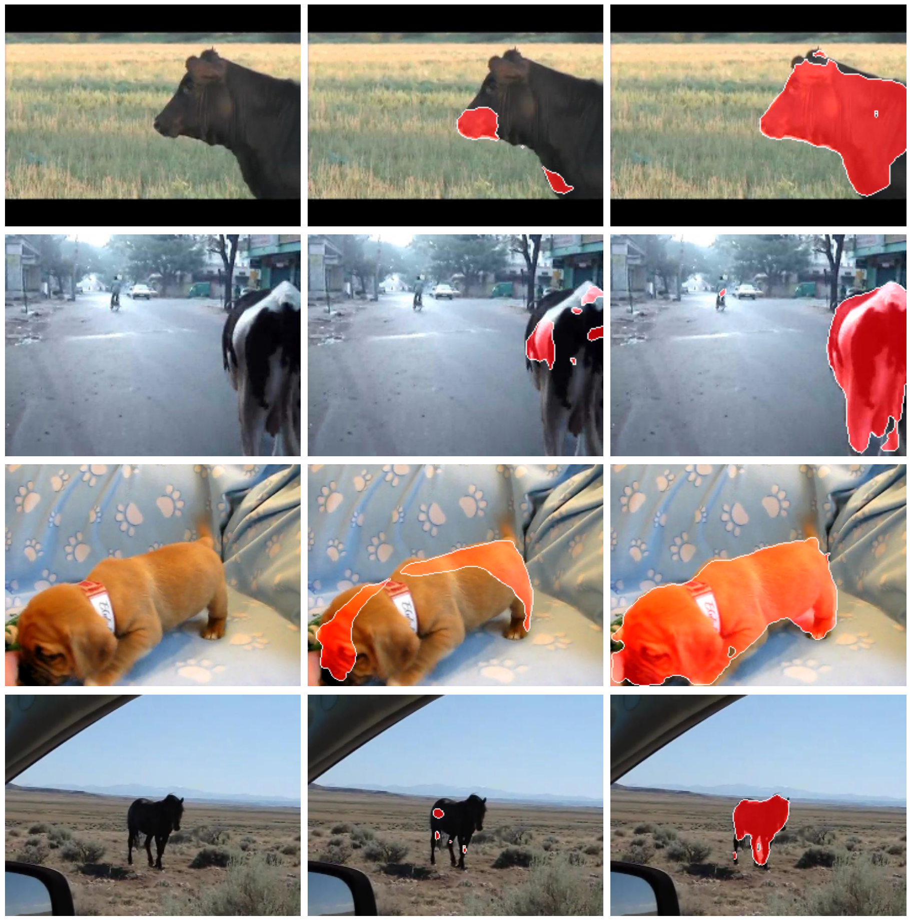


Fig. 6. Sample results on the DAVIS dataset with categories shared in the PASCAL VOC dataset. We show that our results with and without weak supervisions have more complete object segments with details.



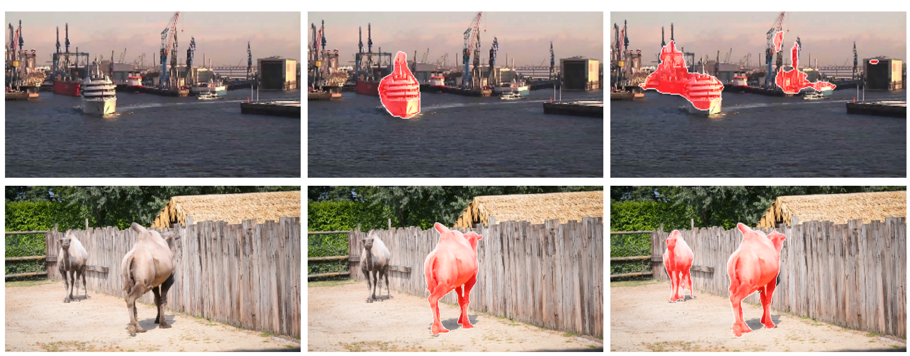
Input FCN [8] Ours

Fig. 7. Sample results on the YouTube-Objects dataset.



Input FCN [8] Ours

Fig. 8. Sample results on the YouTube-Objects dataset.



Input Ground Truth Ours

Fig. 9. Sample failure cases. Although our results differ from the ground truths, the segmented areas belong to the same semantic category.