Classifying, Segmenting, and Tracking Object Instances in Video with Mask Propagation

Supplementary Materials

Gedas Bertasius, Lorenzo Torresani
Facebook AI

1. Implementation Details

Training. We use a similar training setup as in [2]. The loss weights for each of the three cascade stages are set to $1, 0.5$, and 0.25 respectively. The loss weight for the semantic segmentation branch is set to 0.1. We train our model on pairs of frames, where the second frame in a pair is randomly selected with a time gap $\delta \in [-25, 25]$ relative to first frame. We use a multi-scale training approach implemented by resizing the shorter side of the frame randomly between 400 and 800 pixels. Our model is trained in a distributed setting using 64 GPUs, each GPU holding a single clip. The training is done for 20 epochs with an initial learning rate of 0.008, which is decreased by 10 at 16 and 19 epochs. We initialize our model with a Mask R-CNN pretrained on COCO for the instance segmentation. The hyperparameters of RPN and FPN are the same as in [2].

Inference. During testing, we run the bounding box prediction branch on 1000 proposals, apply non-maximum suppression, and use boxes with a score higher than 0.1 as input to the mask prediction and mask propagation branches. During inference, our MaskProp is applied to video clips consisting of 13 frames.

2. Additional Ablation Experiments

Importance of Frame-Level Instance Masks. As described in our main draft, we use frame-level instance masks for instance-specific feature computation. To investigate the contribution of these masks to the performance of our system, we experiment with masks obtained from two different Mask R-CNN models. These include Mask R-CNN with 1) ResNeXt-101-64x4d [4] and 2) Spatiotemporal Sampling Network (STSN) [1] ResNeXt-101-64x4d [4] backbones.

In Table 1, we present our results for this ablation. Our results indicate that frame-level instance masks obtained from stronger models allow us to achieve better video instance segmentation performance. Thus, we expect that future improvements in image instance segmentation will further benefit our method.

Mask R-CNN Model	mAP	AP@75
ResNeXt-101-64x4d [4]	44.3	48.3
STSN [1]-ResNeXt-101-64x4d [4]	**46.6**	**51.2**

Table 1: We study the effect of frame-level instance masks to our system's performance. We evaluate our method's accuracy when using instance masks obtained from Mask R-CNN with two different backbones. Our results indicate that frame-level instance masks obtained from stronger models lead to better video instance segmentation results.

3. Additional Qualitative Results

Comparison with MaskTrack R-CNN. In Figure 1, we compare our instance tracks (last row of predictions for each clip) with the MaskTrack R-CNN predictions (first row of predictions). We use different colors to depict different object instances. On these examples, MaskProp yields more robust and temporally coherent instance tracks than MaskTrack R-CNN. We observed that the differences in performance are especially noticeable when the video contains large object motion, occlusions, or overlapping objects.

Visualizing Propagated Instance Features. In Figure 2 we visualize instance-specific features propagated from frame t to other frames in the given video clip for two different object instances detected in frame t. Here we show activations from a randomly selected feature channel. Our results indicate that MaskProp reliably propagates features that are specific to each instance despite motion blur, object deformations and large variations in object appearance.

4. Supplementary Video

Due to space constraints, we could not include extensive qualitative results in the original draft. Furthermore, it is hard to judge robustness to occlusion and other nuisance effects from static figures. For these reasons, we enclose

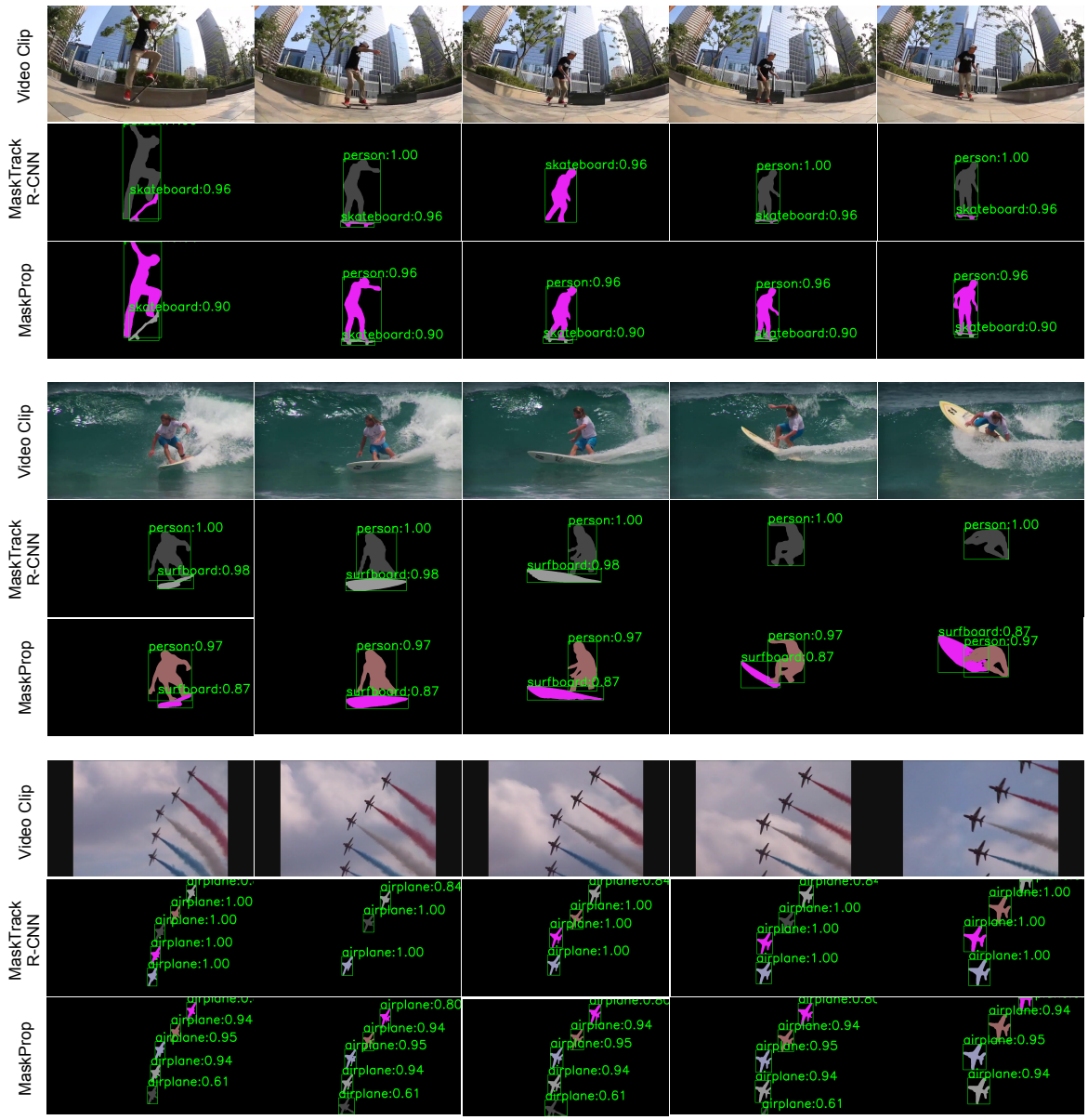

Figure 1: We compare our video instance segmentation results with those produced by MaskTrack R-CNN [5]. Different object instances are shown with different colors. The first row for each video shows the original frames. The second row illustrates the mask predictions of MaskTrack R-CNN and the third row those obtained with our MaskProp. Compared to MaskTrack R-CNN, on these sequences our MaskProp tracks object instances more robustly even when they are occluded or overlap with each other.

in supplementary video[1] the complete video instance segmentations produced by our MaskProp for several challenging input sequences involving occlusion, object instances with similar appearance, and small objects. For comparison we also include the results obtained with MaskTrack R-CNN [5]. Our video results suggest that MaskProp produces more robust instance tracks than MaskTrack R-CNN on these examples, particularly in scenarios involving motion blur, occlusions, or multiple nearby objects.

[1] https://gberta.github.io/maskprop/

References

[1] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. In *ECCV (12)*, volume 11216 of *Lecture Notes in Computer Science*, pages 342–357. Springer, 2018. 1

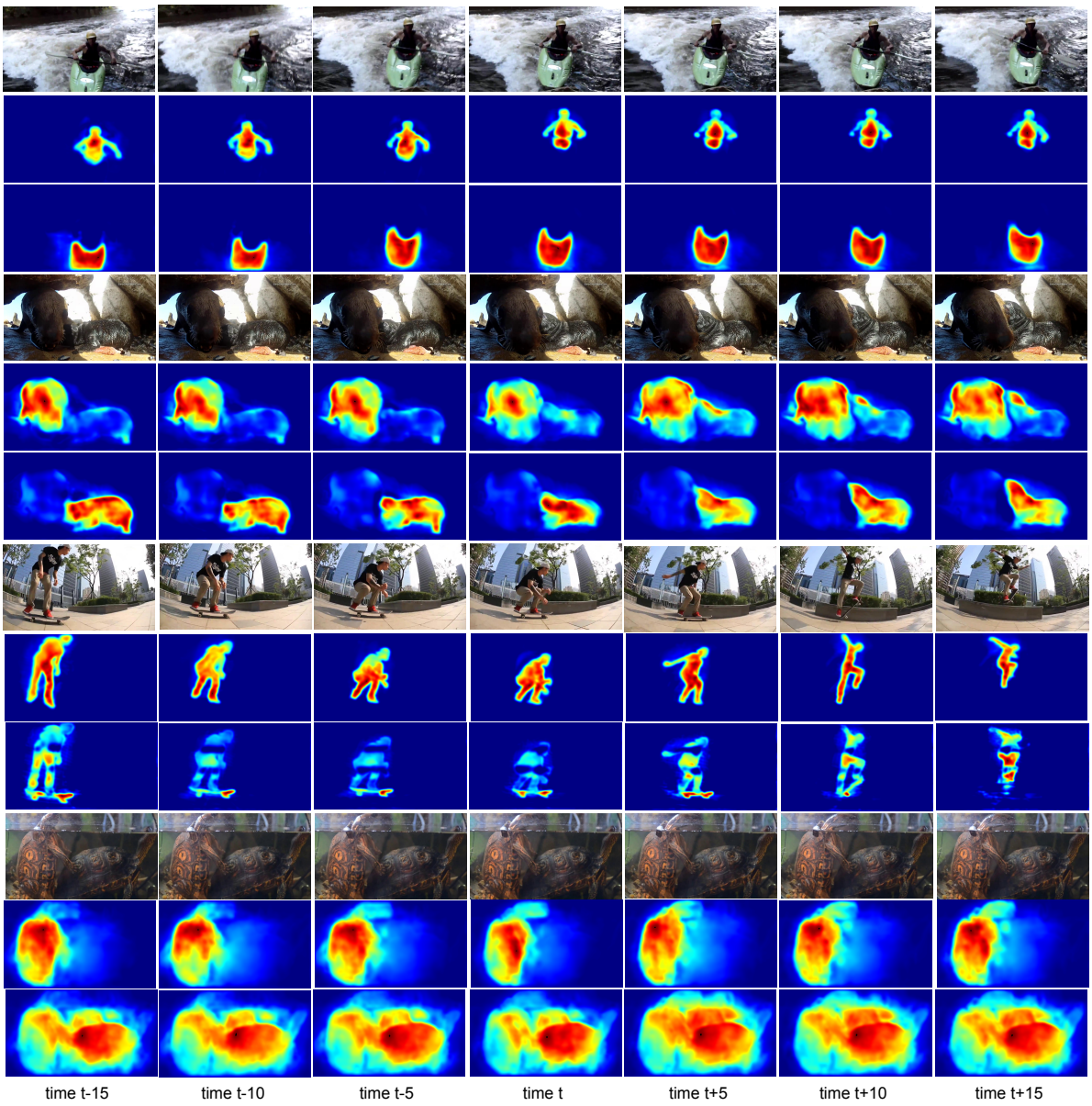

Figure 2: An illustration of instance-specific features propagated from frame t to other frames in the given video clip. Here, we visualize propagated activations from one randomly selected feature channel. The activations in the two rows correspond to two different object instances detected at time t. Our visualizations suggest that MaskProp reliably propagates features that are specific to each instance even when instances appear next to each other, and despite the changes in shape, pose and the nuisance effects of deformation and occlusion.

[2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4974–4983. Computer Vision Foundation / IEEE, 2019. 1

[3] Jonathon Luiten, Philip H.S. Torr, and Bastian Leibe. Video instance segmentation 2019: A winning approach for combined detection, segmentation, classification and tracking. In *2019 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, South Korea, October 27-November 2, 2019*. IEEE, 2019.

[4] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 1

[5] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 2