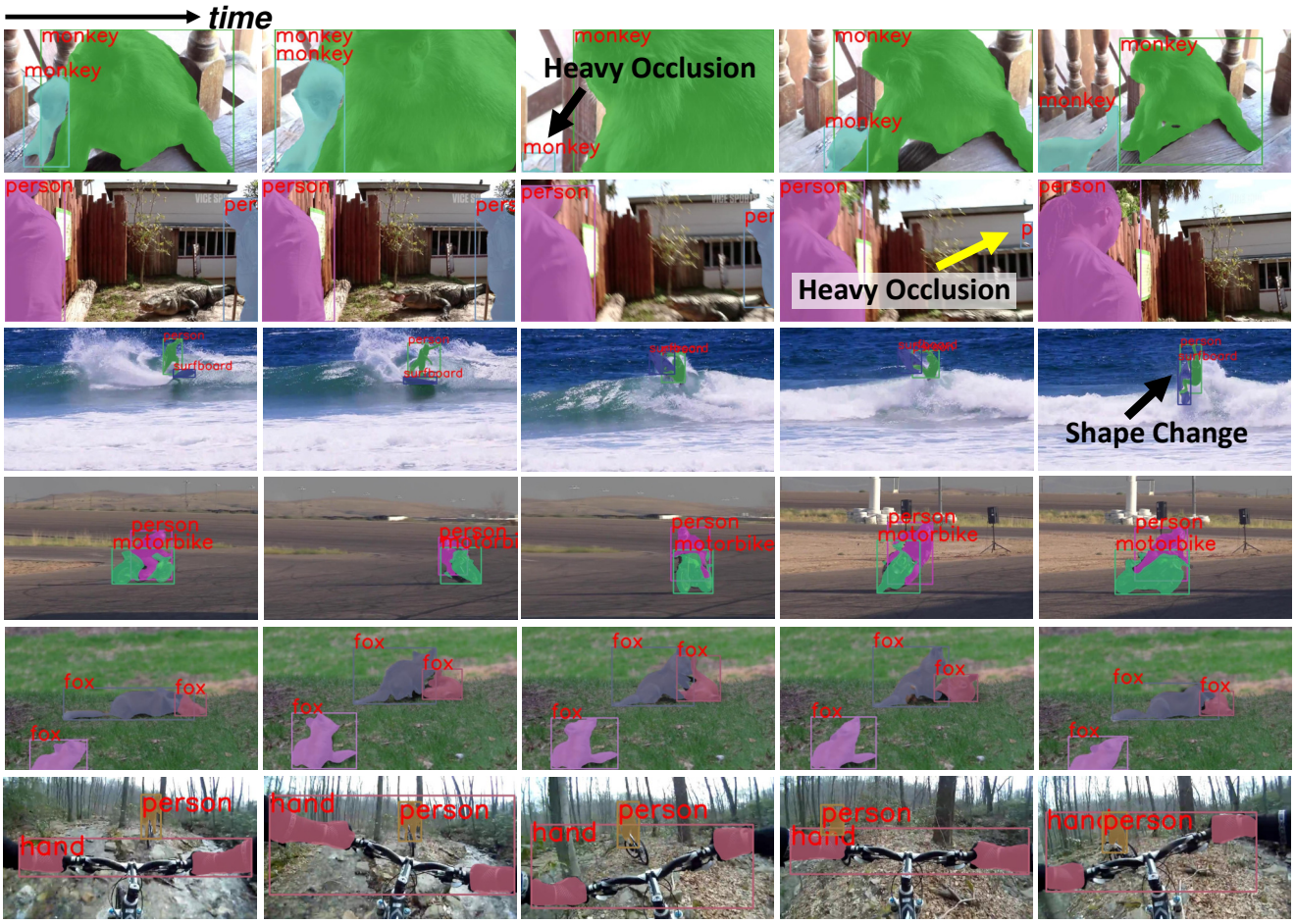

Figure 1. **Visualization on the YouTube-VIS dataset.** One color indicates one identity. The boxes denote our tracklet proposals.

Appendix

A. Qualitative Results on YouTube-VIS

As shown in Fig. 1, we visualize our video instance segmentation results on the YouTube-VIS 2019 dataset. We see from the figure that EfficientVIS can well recognize heavily occluded object instances. The reason is that the target instance information is propagated back and forth in a clip thanks to our clip-by-clip processing and *temporal dynamic convolution*. In this way, the non-occluded instance appearances from nearby frames are propagated to provide strong cues to recognize those heavily occluded instances. We also see in the figure that our tracklet proposal can successfully track target instance even its shape dramatically changes over time. This is because tracklet proposal is regressed conditioned on the target query representation, and we do not impose space-time constraints or smoothness. Thus, tracklet proposal is not limited by the target positions or shapes in nearby frames.

B. Qualitative Comparison

As shown in Fig. 2, we compare EfficientVIS with the VIS transformer (VisTR) [1]. Since VisTR produces an instance mask by segmenting the whole frame, one instance mask may easily contaminate other instances or background regions as shown in the figure. This suggests that it is hard to enforce the query representation in VisTR to be very discriminative to distinguish target object instance from the whole scene. In contrast to this frame-wise scheme, EfficientVIS filters out many irrelevant instances and regions by our tracklet proposals. Our method only needs to enforce tracklet query to distinguish target instance from the proposal region, which is easier for the model to achieve.

References

[1] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of*

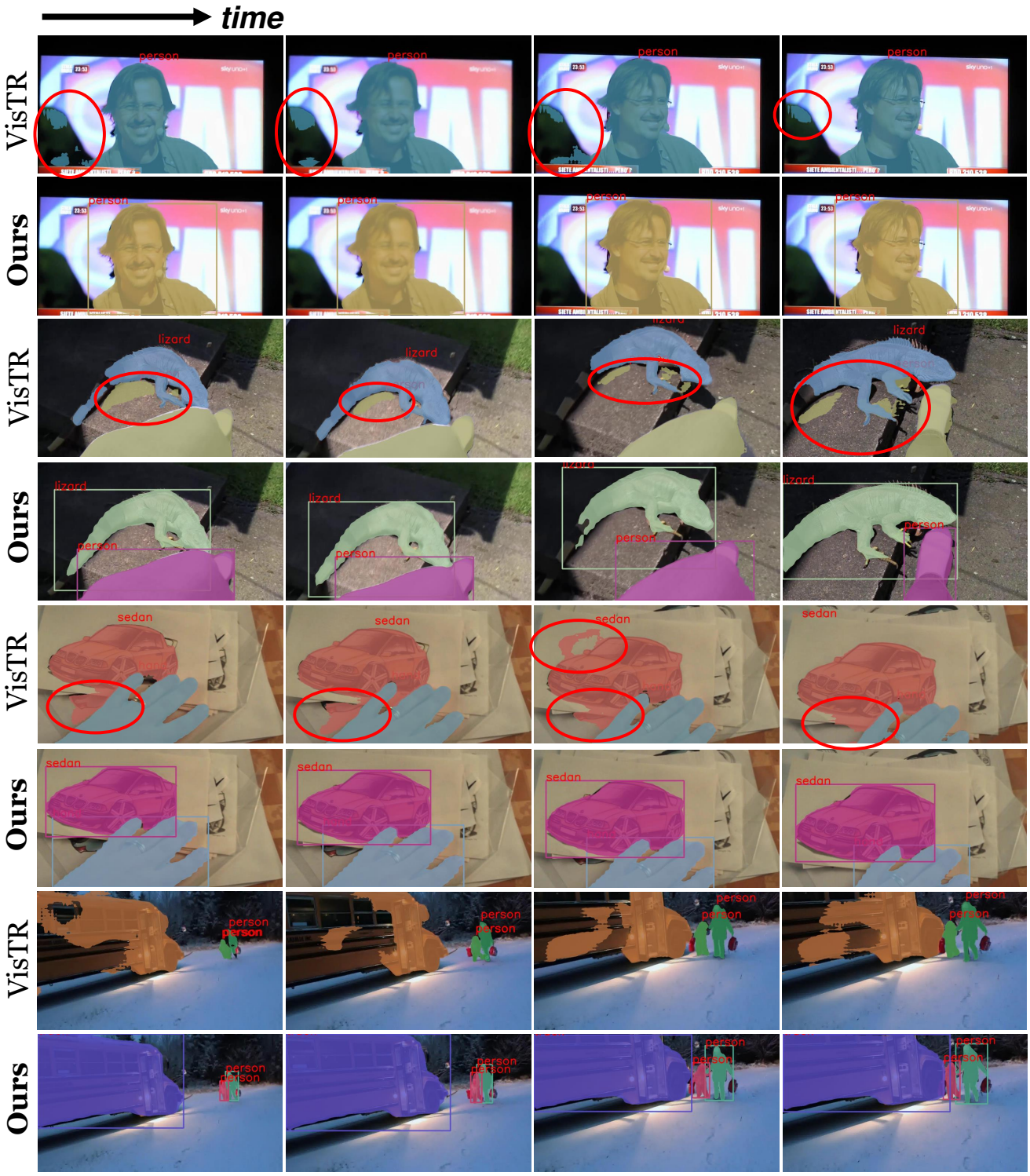

Figure 2. **Qualitative comparison** with the VIS transformer (VisTR) [1].

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8741–8750, 2021. 1, 2