Reciprocal Transformations for Unsupervised Video Object Segmentation Supplementary Materials

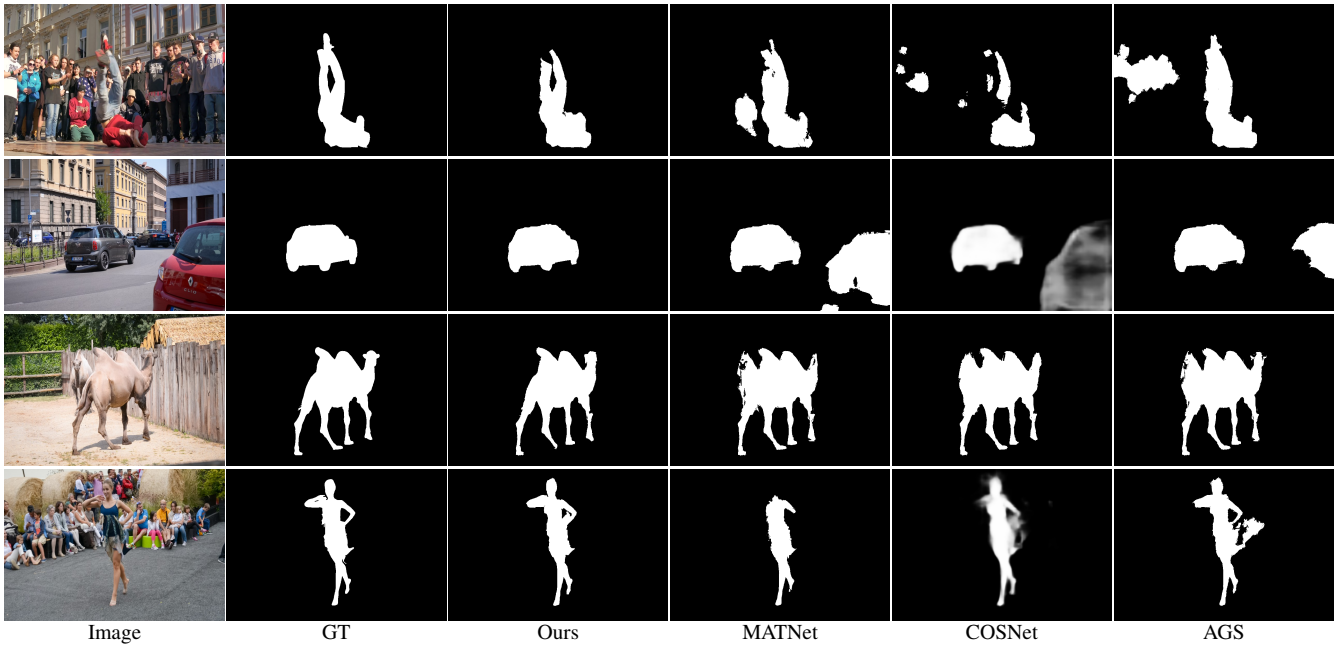

Figure 1: Comparison with the state-of-the-art methods

The supplementary materials contain: 1) More qualitative results; 2) Inference time; 3) Visual ablation on different modules.

1. Inference Time

The inference time of our lightweight models are 0.05s per frame using a RTX2080Ti GPU. Post-processing using CRF takes 0.2s. The timing performance of our method is comparable with the recent methods (e.g. MATNet).

2. More Qualitative Results

Fig. 1 shows the qualitative comparison with the state-of-the-art methods. Our method captures the primary object accurately. Taking the second row of Fig. 1 as an example, the red background car cannot be distinguished from the primary car using the comparison methods.

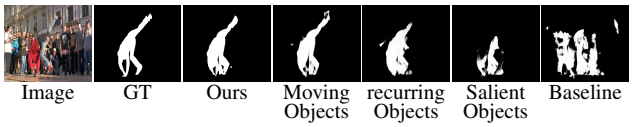

Figure 2: Example w/ moving, recurring, and salient objects

3. Visual Ablation

Fig. 2 shows an example involving moving, recurring, and salient objects, which severely confuses the baseline model. When we enhance the capability of detecting salient objects (the 6th column), recurring objects (the 5th column), and moving objects (the 4th column), respectively, our model segments the primary objects more and more accurately. When we transform all the information to the appearance stream to remove the ambiguity from the inconsistent appearance or the inaccurate motion information, our complete model can segment the primary object well (the 3rd column).