Supplementary Material to ”Robust Unsupervised Video Foreground Segmentation via Context-Aware Bayesian Tensor Factorization”

Appendix A: Justification for Fig. 3

Fig. 3 in paper ”**Robust Unsupervised Video Foreground Segmentation via Context-Aware Bayesian Tensor Factorization**” illustrates the comparison of the performance consistency in terms of frame-wise F-measure, across 100 frames in each video. The performance of proposed CABTF^{TV}, CABTF^{P}, and all the state-of-the-art unsupervised methods are demonstrated in Fig. 3, where the proposed CABTF^{P} consistently outperforms all other methods with the highest F-measure for almost all frames in each video. The frame-wise F-measure of the supervised DeepLab is not shown in Fig. 3, because it contains zero F-measure that may change the vertical axis of Fig. 3, making the plots of the other methods crowded together and fail to show a clear comparison, as shown in Fig. A.1(d). The performance comparison, including all the unsupervised and supervised methods, is illustrated in Fig. A.1. In the video *Pedestrians*, there is a lady walking on the road. The challenge could be overcoming the random camera noise. In the video *Highway*, there are vehicles moving along the highway. The challenges could be focusing on the vehicles without being distracted by the dynamic noise caused by swaying trees on the left side of the highway. In the video *Canoe*, a canoe is moving on a lake, and the canoe is filled with people. The challenges could be overcoming the dynamic noise from rippling water. And in the video *Fountain02*, a car with a trail is driving through while partially blocked by the spraying fountain. The challenge could be overcoming the dynamic noise from the fountain.

Distinct F-measure variation of DeepLab indicates its performance inconsistency, as presented in Fig. A.1. DeepLab is pre-trained on the PASCAL-VOC dataset [1], which consists of more than 10000 images across 21 object classes, including vehicles, pedestrians, and boats, but not including swaying leaves, rippling water, and spraying fountains. Therefore, DeepLab is robust to such dynamic backgrounds, as it cannot recognize them. For unsupervised learning methods, the dynamic backgrounds such as rippling water in *Canoe* severely deteriorate their performance, as shown in the first sixty frames of Fig. A.1(c). However, the DeepLab fails to detect the people sitting in the canoe, and cannot identify the canoe if only part of it is presented in the frames, such as the first ten frames in Fig. A.1(c). Moreover, DeepLab produces considerable zero F-measures for *Fountain02* dataset, as shown in Fig. A.1(d). Because of the small size of the vehicle and potential blurring caused by the fountain, DeepLab fails to identify the vehicle. The limited segmentation accuracy of DeepLab can be potentially caused by the disparity between the training and testing datasets, which is avoidable in complex real-world scenarios. This suggests the limitation of supervised deep learning methods in terms of generalization. Overall, The proposed CABTF^P attains better performance, consistently, than all other methods consistently for almost all frames in each video.

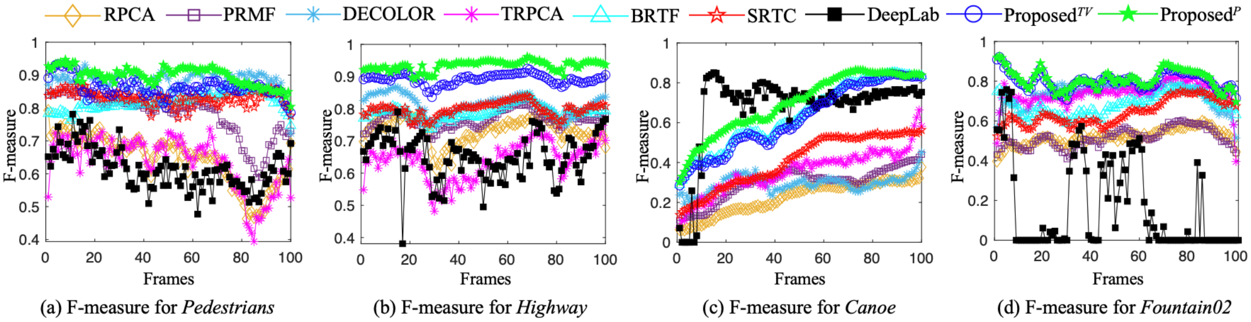

Figure A.1: Segmentation performance comparison by frames for different methods in four videos: (a) *Pedestrians*; (b) *Highway*; (c) *Canoe*; (d) *Fountain02*.

References

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. “The Pascal Visual Object Classes (VOC) Challenge”. In: *International Journal of Computer Vision* 88.2 (June 2010), pp. 303–338.