Supplementary Material for TSM: Temporal Shift Module for Efficient Video Understanding

Ji Lin
MIT
jilin@mit.edu

Chuang Gan
MIT-IBM Watson AI Lab
ganchuang@csail.mit.edu

Song Han
MIT
songhan@mit.edu

1. Uni-directional TSM for Online Video Detection

In this section, we show more details about the online video object detection with uni-directional TSM.

Object detection suffers from poor object appearance due to motion blur, occlusion, defocus, *etc*. Video based object detection gives chances to correct such errors by aggregating and inferring temporal information.

Existing methods on video object detection [4] fuses information along temporal dimension after the feature is extracted by the backbone. Here we show that we can enable temporal fusion in online video object detection by injecting our uni-directional TSM into the backbone. We show that we can significantly improve the performance of video detection by simply modifying the backbone with online TSM, without changing the detection module design or using optical flow features.

We conducted experiments with R-FCN [2] detector on ImageNet-VID [3] dataset. Following the setting in [4], we used ResNet-101 [1] as the backbone for R-FCN detector. For TSM experiments, we inserted uni-directional TSM to the backbone, while keeping other settings the same. We used the official training code of [4] to conduct the experiments, and the results are shown in Table 1. Compared to 2D baseline R-FCN [2], our online TSM model significantly improves the performance, especially on the fast moving objects, where TSM increases mAP by 4.6%. FGFA [4] is a strong baseline that uses optical flow to aggregate the temporal information from 21 frames (past 10 frames and future 10 frames) for offline video detection. Compared to FGFA, TSM can achieve similar or higher performance while enabling online recognition (using information from only past frames) at much smaller latency per frame. The latency overhead of TSM module itself is less than 1ms per frame, making it a practical tool for real deployment.

We visualize two video clips in Figure 1 and 2. In Figure 1, 2D baseline R-FCN generates false positive due to the glare of car headlight on frame 2/3/4, while TSM suppresses false positive. In Figure 2, R-FCN generates false positive surrounding the bus due to occlusion by the traffic sign on frame 2/3/4. Also, it fails to detect motorcycle on frame 4 due to occlusion. TSM model addresses such issues with the help of temporal information.

Table 1. Video detection results on ImageNet-VID.

Model	Online	Need Flow	Latency	mAP			
				Overall	Slow	Medium	Fast
R-FCN [2]	✓		1×	74.7	83.6	72.5	51.4
FGFA [4]		✓	2.5×	75.9	**84.0**	74.4	55.6
Online TSM	✓		1×	**76.3**	83.4	**74.8**	**56.0**

2. Video Demo

We provide more video demos of our TSM model in the following project page: https://hanlab.mit.edu/projects/tsm/.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[2] K. H. J. S. Jifeng Dai, Yi Li. R-FCN: Object detection via region-based fully convolutional networks. 2016. 1

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1

[4] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 408–417, 2017. 1

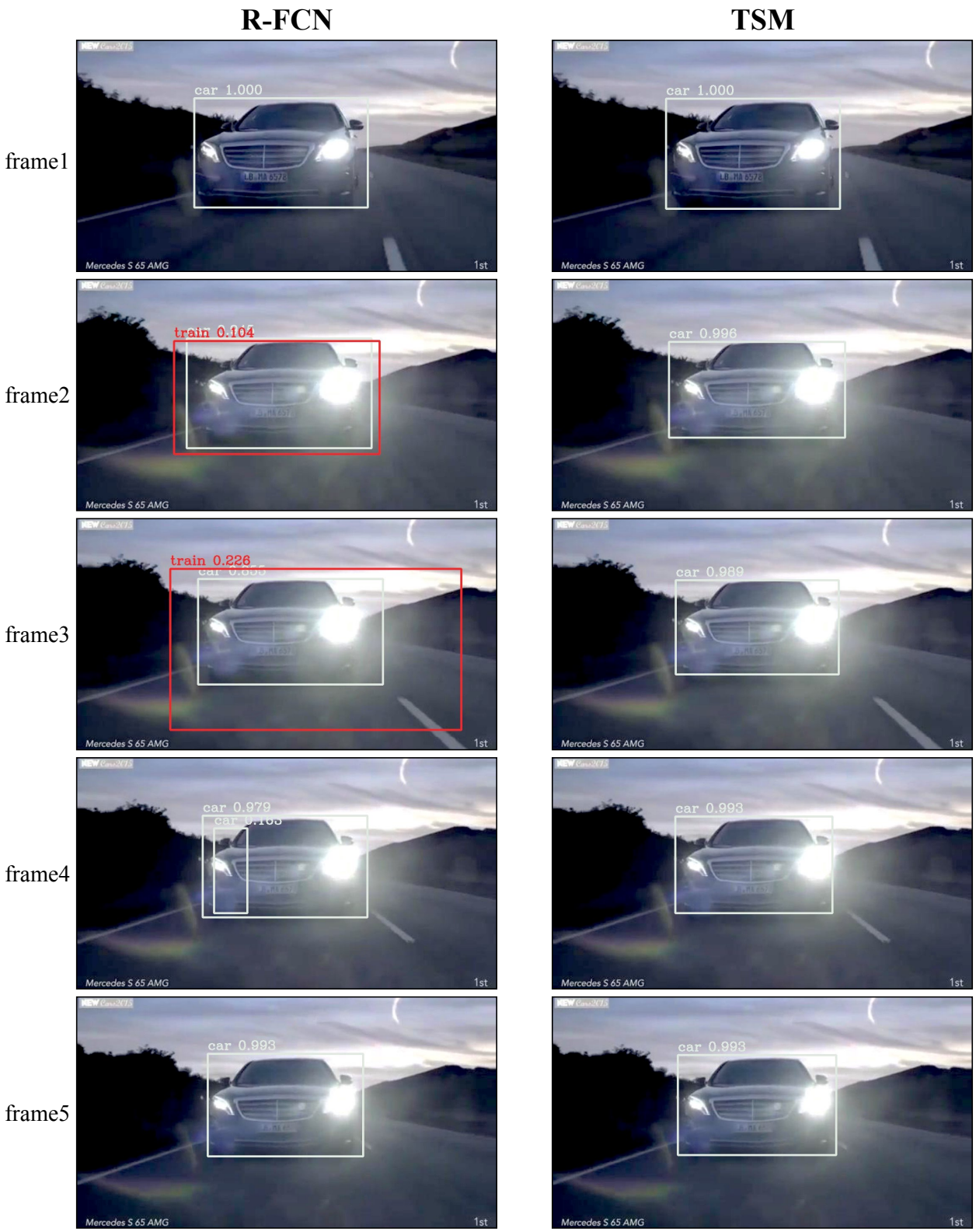

Figure 1. Comparing the result of R-FCN baseline and TSM model. 2D baseline R-FCN generates false positive due to the glare of car headlight on frame 2/3/4, while TSM does not have such issue by considering the temporal information.

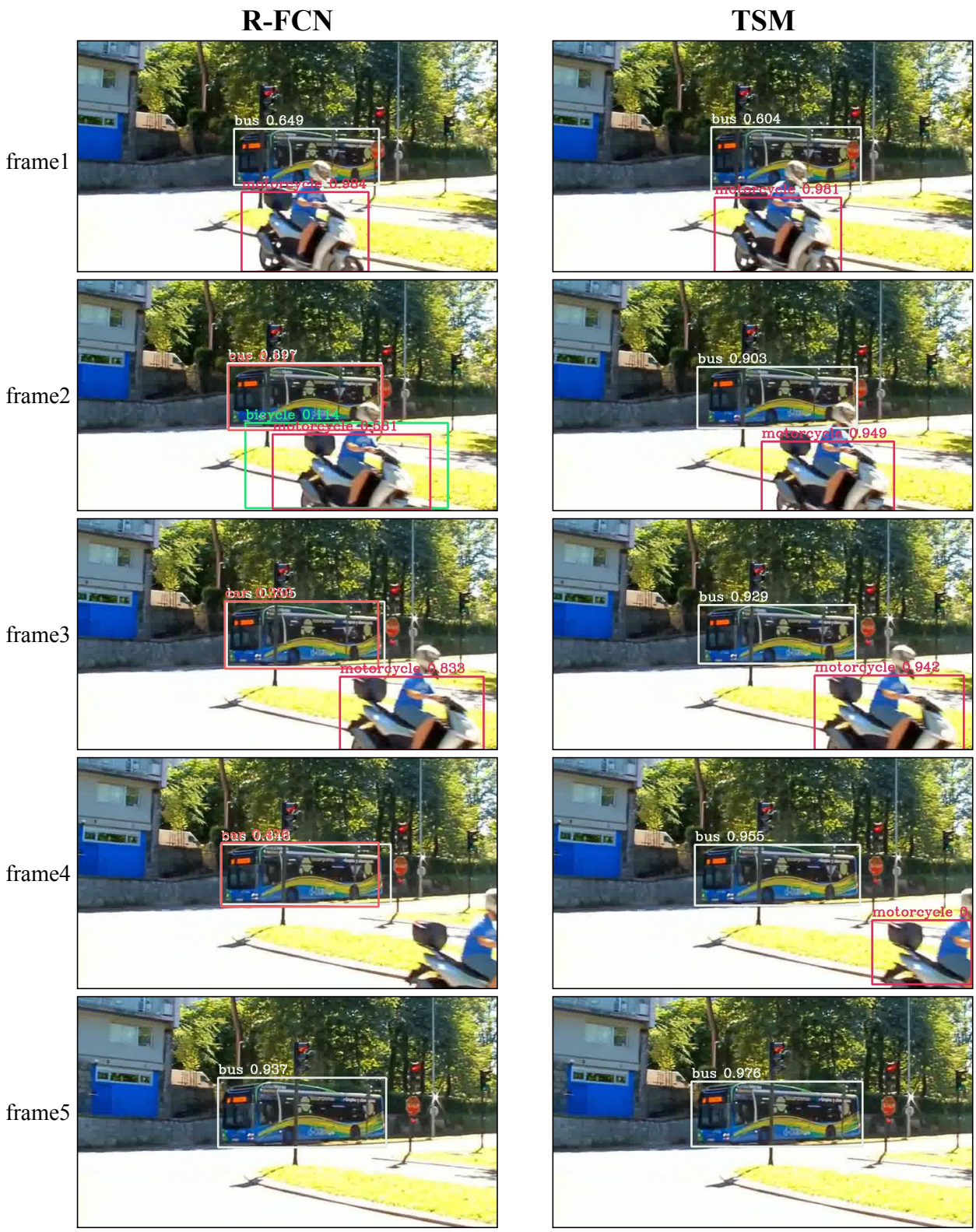

Figure 2. Comparing the result of R-FCN baseline and TSM model. 2D baseline R-FCN generates false positive surrounding the bus due to occlusion by the traffic sign on frame 2/3/4. Also, it fails to detect motorcycle on frame 4 due to occlusion. TSM model addresses such issues with the help of temporal information.