TDN: Temporal Difference Networks for Efficient Action Recognition Supplementary Material

Limin Wang Zhan Tong Bin Ji Gangshan Wu
State Key Laboratory for Novel Software Technology, Nanjing University, China
07wanglimin@gmail.com, tongzhan@smail.nju.edu.cn, binjinju@smail.nju.edu.cn, gswu@nju.edu.cn

A. Results on the UCF101 and HMDB51

Method	Pretrain	Backbone	UCF101	HMDB51
TSN [10]	ImageNet	Inception V2	86.4%	53.7%
P3D [5]	ImageNet	ResNet50	88.6%	-
C3D [7]	Sports-1M	ResNet18	85.8%	54.9%
I3D [1]	ImageNet+Kinetics	Inception V2	95.6%	74.8%
ARTNet [9]	Kinetics	ResNet18	94.3%	70.9%
S3D [11]	ImageNet+Kinetics	Inception V2	96.8%	75.9%
R(2+1)D [8]	Kinetics	ResNet34	96.8%	74.5%
TSM [4]	Kinetics	ResNet50	96.0%	73.2%
STM [2]	ImageNet + Kinetics	ResNet50	96.2%	72.2%
TEA [3]	ImageNet + Kinetics	ResNet50	96.9%	73.3%
TDN(Ours)	ImageNet + Kinetics	ResNet50	**97.4%**	**76.3%**

Table 1. Comparison with the state-of-the-art methods on **UCF101 and HMDB51**.

To further verify the generalization ability of TDN, we transfer the learned 16-frame TDN models from the Kinetics-400 dataset to the UCF101 and HMDB51. These two datasets are relatively small and the action recognition performance on them already saturates. We follow the standard evaluation scheme on these two datasets and report the mean accuracy over three splits. The results are summarized in Table 1. We compare our TDN with previous state-of-the-art methods such as 2D baselines of TSN [10], 3D CNNs of I3D [1] and C3D [7], R(2+1)D [8], and other temporal modeling methods [3, 2]. From the results, we can see that our TDN is able to outperform these methods, and the performance improvement is more evident on the dataset of HMDB51 by around 2.5%. The action classes in HMDB51 are more relevant with motion information, and thus temporal modeling is more important on this dataset.

B. Running time analysis

We report the inference time of our TDN with on Tesla V100 as follows. The testing batchsize is set as 16 and the running time include all evaluation, including loading data and network inference. The results are reported in Table 2. From these results, we see that our TDN is slower than previous method but still could run in real-time (i.e. $\geq$25 FPS).

Method	Frames$\times$Clips$\times$Crops	Time (ms/video)	Top1 (%)
TSN [10]	$8 \times 1 \times 1$	7.9	19.7
TSM [4]	$16 \times 1 \times 1$	16.7	47.2
STM [2]	$8 \times 1 \times 1$	11.1	47.5
I3D [1]	$32 \times 3 \times 2$	2095	41.6
S-TDM	$8 \times 1 \times 1$	12.3	49.5
L-TDM	$8 \times 1 \times 1$	15.8	48.9
TDN	$8 \times 1 \times 1$	22.1	52.3

Table 2. Running time analysis on a Tesla V100.

C. Visualization analysis

To further investigate the performance the TDN models, we use the technique of Grad-CAM [6] to visualize the feature representation of different models. Specifically, to better understand the effect of short-term TDM, we visualize the the features in Res2 stage of baseline model (corresponding to the first row in Table 1(e) of main article) and the TDM model only with S-TDM (corresponding to third row in in Table 1(e) of main article), and the results are shown in Figure 1. Note that, these visualizations only are performed on the center frame of 8-frame models. From these results, the models equipped with S-TDM focuses more on motion-relevant information. Then, we give more visualization examples of activation maps in Figure 2 and Figure 3. In these results, we give the visualization results on 8 frames and compare our TDM models with the baseline method (corresponding to the first row in Table 1(e) of main article). We could see that our TDN is able to yield more reasonable class activation maps than the baseline method.

References

[1] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733, 2017. 1

[2] Boyuan Jiang, Mengmeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. STM: spatiotemporal and motion encoding for action recognition. In *ICCV*, pages 2000–2009, 2019. 1

[3] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for

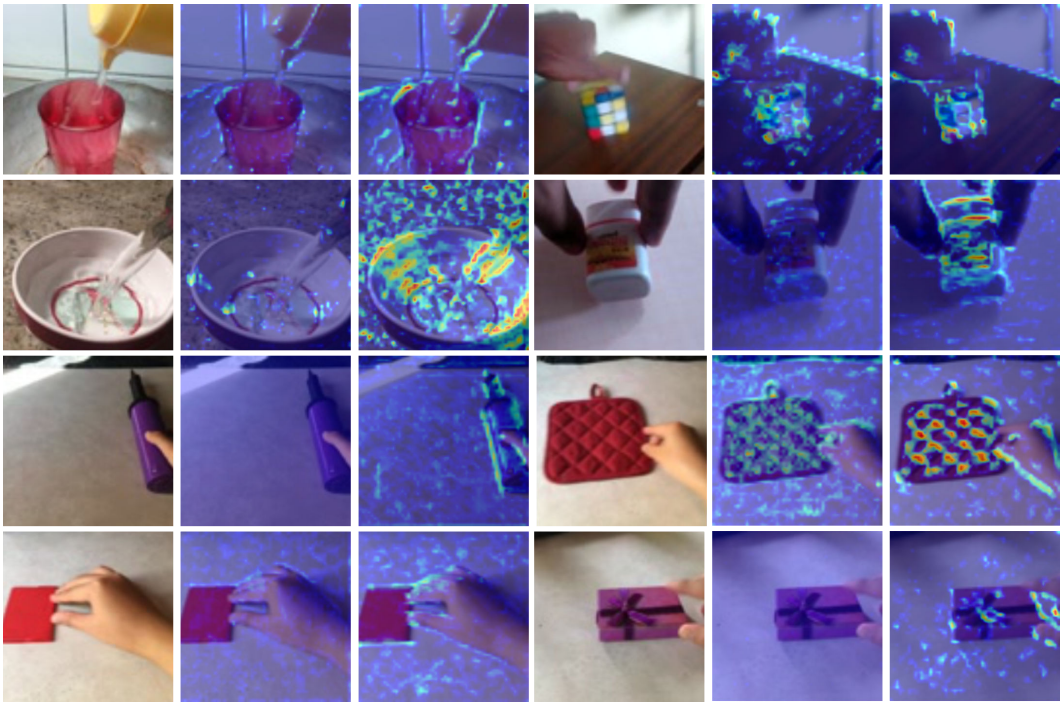

Figure 1. Visualization of Res2 features with Grad-CAM. We use 8-frame TDN models to visualize on the Something-Something V1 dataset. Left: video, Middle: baseline, Right: TDN with S-TDM. Note that we only show visualization on the center frame of sampled 8 frames.

action recognition. In *CVPR*, pages 909–918, 2020. 1

[4] Ji Lin, Chuang Gan, and Song Han. TSM: temporal shift module for efficient video understanding. In *ICCV*, pages 7082–7092, 2019. 1

[5] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. *ICCV*, pages 5534–5542, 2017. 1

[6] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IJCV*, 128(2):336–359, 2020. 1

[7] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 1

[8] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018. 1

[9] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *CVPR*, pages 1430–1439, 2018. 1

[10] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36, 2016. 1

[11] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, volume 11219, pages 318–335. Springer, 2018. 1

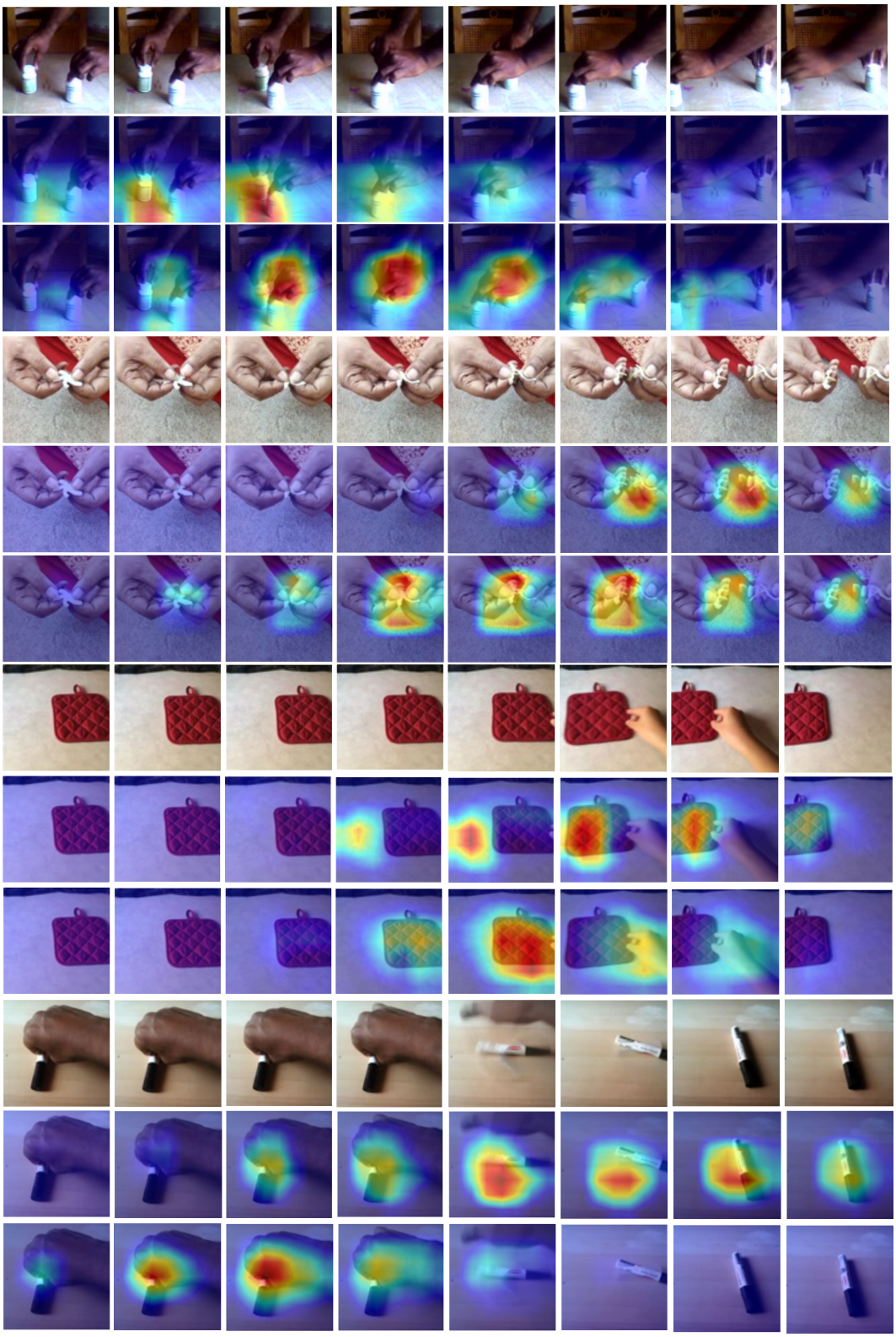

Figure 2. Visualization of activation maps with Grad-CAM. We use 8-frame TDN models to visualize on the Something-Something V1 dataset. In the first row, we plot the 8 RGB frames. In the second row, we plot the activation maps of the baseline method without temporal difference module (TDM). In the third row, we plot the activation maps of the TDN models.

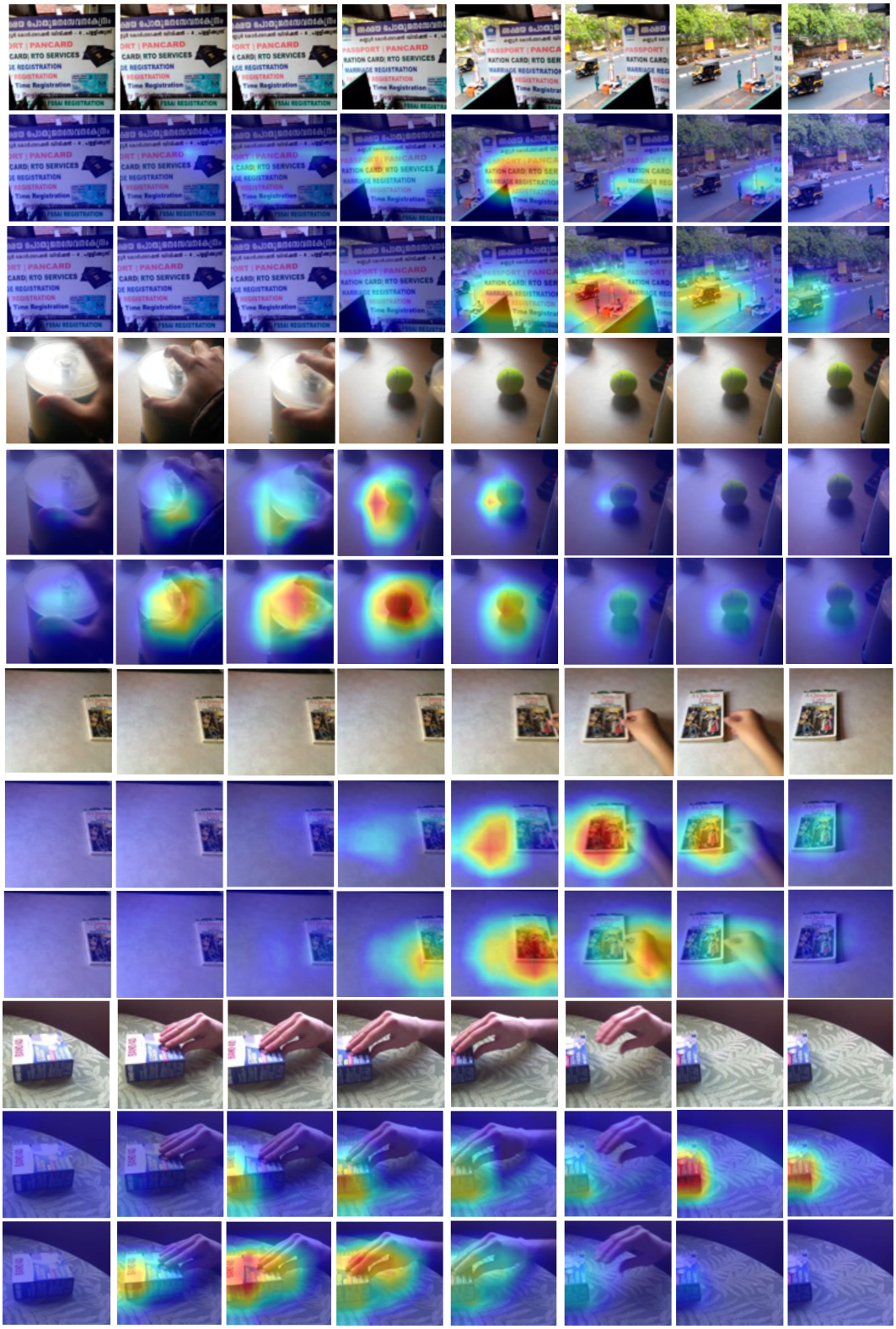

Figure 3. Visualization of activation maps with Grad-CAM. We use 8-frame TDN models to visualize on the Something-Something V1 dataset. In the first row, we plot the 8 RGB frames. In the second row, we plot the activation maps of the baseline method without temporal difference module (TDM). In the third row, we plot the activation maps of the TDN models.