Stop or Forward: Dynamic Layer Skipping for Efficient Action Recognition
Supplementary Materials

1. SoF-Net with other algorithms and datasets

Table A presents the performance of SoF-Net when it is integrated into TSM [19] on ActivityNet-v1.3 (val) while Table B shows the results on mini-Sports1M. Since the official download link of FCVID is broken, we cannot run on it.

Table A: Results of SoF-Net with TSM on ActivityNet-v1.3 (val). ET denotes early termination.

Method	mAP	GFLOPS/V	Frame Usage (%)
TSM [19]	73.0	40.3	100.0
SoF-Net$^{\text{TSM}}$ w/o ET	**74.1**	26.7	49.6
SoF-Net$^{\text{TSM}}$	73.9	**22.4**	**44.1**

Table B: Results on mini-Sports1M.

Method	Top 1	GFLOPS/V	Frame Usage (%)
TSN [32]	39.2	40.3	100.0
AR-Net	39.1	41.2	73.0
SoF-Net	**41.2**	**29.0**	**53.2**

2. Hyperparameters for early termination

Table C shows clear trade-off between accuracy and efficiency with the early termination threshold on ActivityNet-v1.3 (val).

Table C: Ablation study on the hyperparameters for early termination. We present mAP and GFLOPS/f for each case.

ρ	$T_{\text{min}} = 2$	$T_{\text{min}} = 3$	$T_{\text{min}} = 4$	$T_{\text{min}} = 5$
0.95	72.3 / 1.34	73.5 / 1.40	74.2 / 1.46	74.6 / 1.52
0.99	74.1 / 1.54	74.6 / 1.59	75.0 / 1.62	75.0 / 1.67
0.999	74.9 / 1.68	**75.3 / 1.71**	75.3 / 1.74	75.2 / 1.77

3. Ablation study of loss terms

Table D presents the analysis of our loss terms on ActivityNet-v1.3 (val) without early termination; the efficiency loss works as expected and the two terms in $\mathcal{L}_{\text{pg}}$ play their roles.

Table D: Ablation study on α and β of the loss function in Eq. (7) and (11). The hyphen (–) for β denotes that $\mathcal{L}_{\text{pg}}$ is not used.

(α, β)	(1.0, –)	(0.9, –)	(1.0, 0.0)	(1.0, 0.5)	(1.0, 1.0)	(0.9, 0.5)
mAP	75.5	75.0	75.2	75.9	74.7	75.3
GFLOPS/V	37.6	31.6	37.6	38.9	35.1	31.5

4. Qualitative Analysis

We present additional examples showing input frames and decision making for them in Figure A. In general, our method removes unimportant frames at an early stage no matter how it predicts correct label or not. On the other hand, early termination only happens for the correctly predicted videos. In the failure cases, the model predicts similar labels (e.g., assembling bicycle instead of fixing bicycle) but it successfully skips less important frames such as beginning and ending moments.

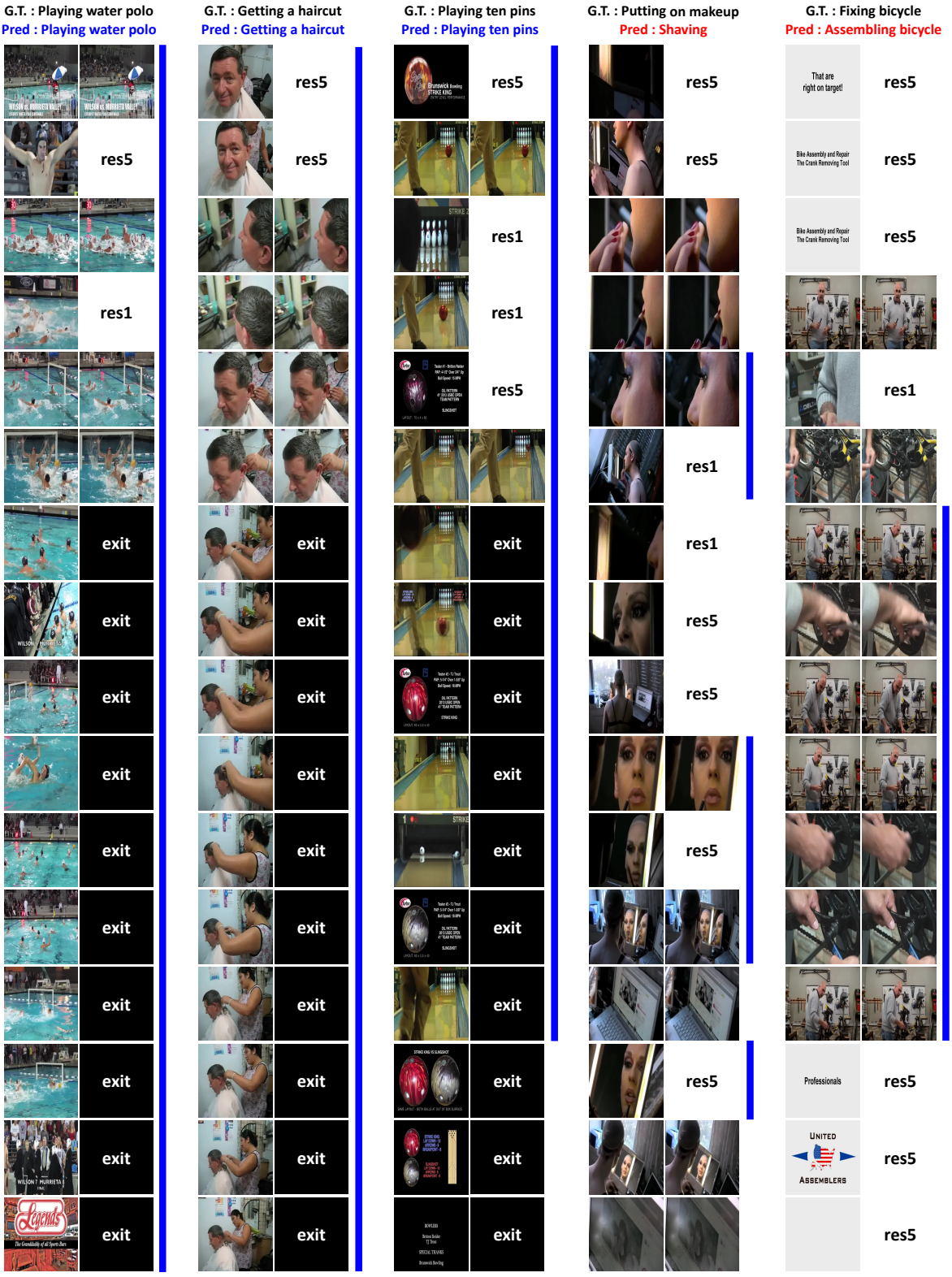

Figure A: Visualization of the decision making by SoF-Net. For each example, the left column shows the original input frames and the right column illustrates how the frames are processed in SoF-Net. The termination block (black box with a word 'exit') means the skipped frames due to early termination and the temporal action localization ground-truths indicate by the blue bar. Videos start from the top.