Supplementary Material
Dynamic Inference: A New Approach Toward Efficient Video Action Recognition

In this supplementary material, we provide more information about our dyamic inference frameworks, including the distribution of videos early stopped at each checkpoint and visualization of videos at different checkpoints.

1. Distribution of Videos

As shown in Table 1 and Table 2, we report the number of videos stopped at each checkpoint for MSDNet-38 [3] and ResNet-50 [2] on Kinetics-400 [4], UCF-101 [6], HMDB-51 [5], Someting-Something V1 & V2 [1].

Checkpoint	1	2	3	4	5	6	Total
Kinetics-400	1993	2392	2871	3445	4134	4961	19796
UCF-101 Split1	807	727	654	588	530	477	3783
HMDB-51 Split1	74	110	166	248	373	559	1530
Sth-Sth v1	1404	1755	2194	2742	3427	-	11522
Sth-Sth v2	4058	4464	4911	5402	5942	-	24777

Table 1. Number of videos which stop at each checkpoint of MSDNet-38 [3] on different datasets.

Checkpoint	1	2	3	4	Total
Kinetics-400	3433	4292	5365	6706	19796
UCF-101 Split1	1874	1030	567	312	3783
HMDB-51 Split1	265	332	415	518	1530
Sth-Sth v1	276	842	2569	7835	11522
Sth-Sth v2	1194	2745	6315	14523	24777

Table 2. Number of videos which stop at each checkpoint of ResNet-50 [2] on different datasets.

2. Visualization

To help us better understand how videos differentiate from each other in terms of their distinguishability for action recognition, we visualize the video instances which exit at different checkpoint of our method. We adopt dynamic inference with MSDNet-38 [3] and show six randomly sampled test videos from Kinetics-400 [4] validation set in Figure 1, the visualization illustrates the ability of our approach to reduce the computational requirements for recognizing "easy" videos. The top row Fig. 1(a) shows two videos that were correctly classified and exited by the first checkpoint. The middle row Fig. 1(b) shows two videos that were correctly classified and exited at the third checkpoint. The bottom row Fig. 1(c) shows two "hard" examples that would have been incorrectly classified by the first few checkpoints but were passed on the last checkpoint. The figure suggests that early checkpoint recognizes prototypical class examples, whereas the last classifier recognizes non-typical videos.

References

[1] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 1

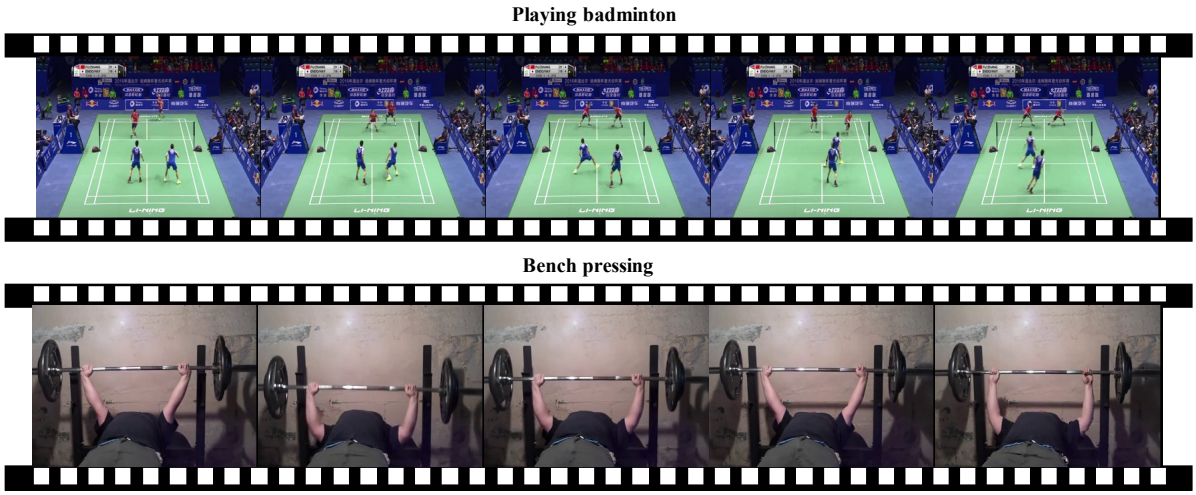
[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[3] Gao Huang and Danlu Chen. Multi-scale dense networks for resource efficient image classification. *ICLR*, 2018. 1, 2

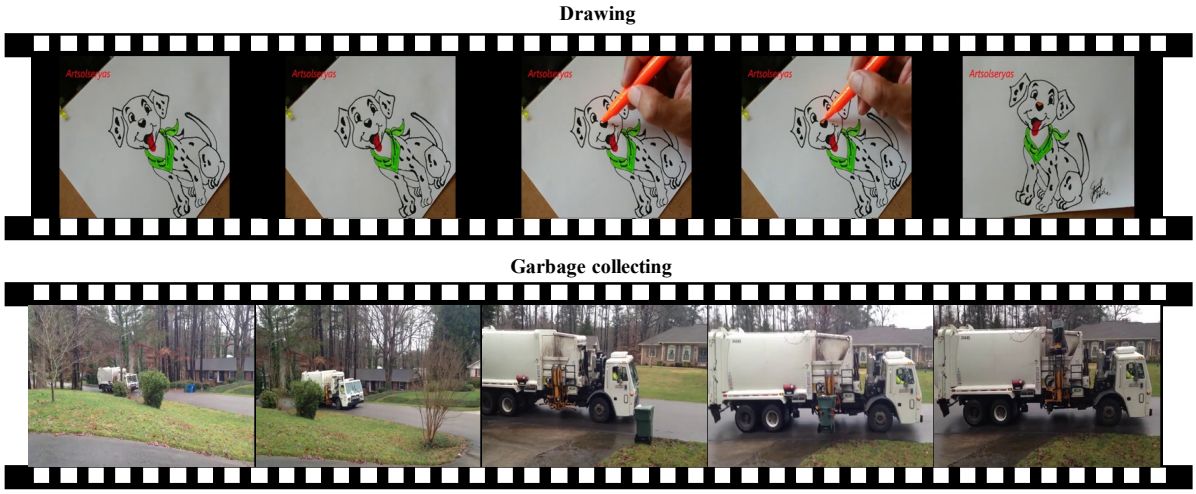
[4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[5] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 1

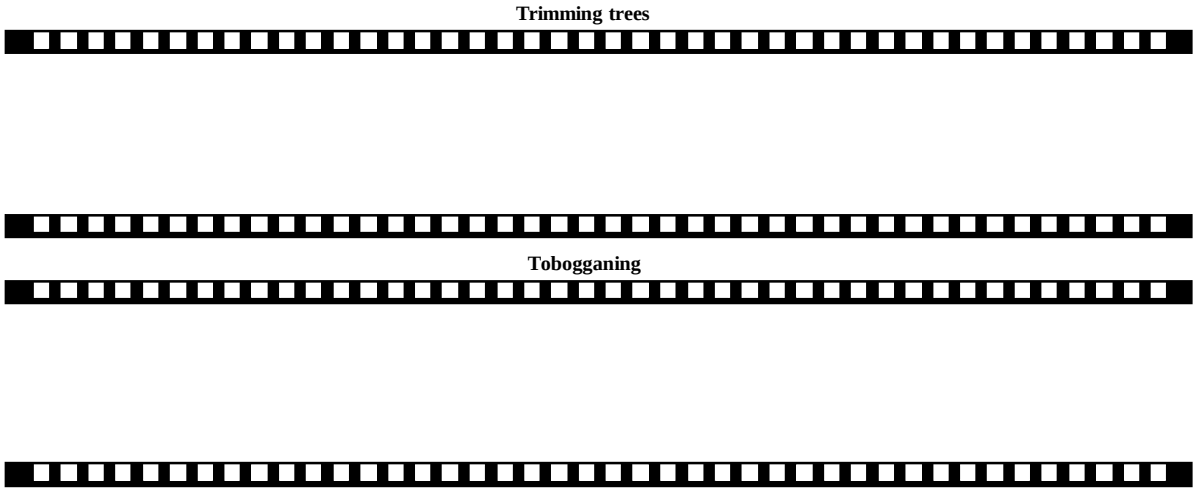
[6] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1



(a) Two video instances which stop at the **first checkpoint** of MSDNet-38 [3].



(b) Two video instances which stop at the **third checkpoint** of MSDNet-38 [3].



(c) Two video instances which stop at the **sixth checkpoint** of MSDNet-38 [3].

Figure 1. Video instances which stop at the different checkpoint of MSDNet-38 [3] on Kinetics-400 validation set.