Supplementary Material - Video Shuffle Networks

In this document, we provide additional materials to supplement our main submission. As mentioned in Section 3.4, we first present the training details for each dataset in Table 1. Taking these hyper-parameters, we have produced the experimental results in Figure 1 in main submission. In Table 2, the inference latency is also provided.

1. Training details for each dataset

Dataset	#Epochs	Init LR	LR scheduler	LR milestones	LR decay factor	Warm-up epochs	Dropout
Kinetics	80	0.001	cosine	-	-	5	0.5
UCF101	15	0.001	multistep	[5, 10, 15]	0.1	1	0.8
HMDB51	30	0.001	multistep	[10, 20, 30]	0.1	1	0.8
Moments	30	0.001	multistep	[15, 25, 30]	0.1	3	0.5
Jester	25	0.001	multistep	[10,15,25]	0.1	0	0.8
Charades	30	0.001	cosine	-	-	5	0.6
Something-V1	25	0.001	multistep	[10,15,25]	0.1	0	0.8
Something-V2	25	0.001	multistep	[10,15,25]	0.1	0	0.8

Table 1. We provide the training hyper-parameters for each dataset. Both TSN and VSN takes the same setting.

We present the description of each column here. #Epochs: total training epochs; Init LR: the initial learning rate; LR scheduler: the learning rate schedule; LR milestones: the learning rate decaying steps; LR decay factor: the learning rate decaying factor; Warm-up epochs: number of epochs for warming up; Dropout: the value of dropout rate.

2. Inference Latency

Model	Modality	#Frame	#Param	FLOPs	Latency	Throughput	Kinetics(%)	Something-V1(%)
VSN-R50	RGB	8	24.3M	33G	165.0ms	86.5vps	71.5	44.5
VSN-R101	RGB	8	42.9M	63G	287.9ms	52.2vps	72.8	59.4
VSN-R50	Flow	8×1	24.3M	33G	163.8ms	87.1vps	53.0	33.7
VSN-R50	Flow	8×5	24.3M	33G	170.3ms	82.7vps	56.7	37.5
VSN-R101	Flow	8×1	42.9M	63G	286.6ms	52.4vps	56.0	36.1
VSN-R101	Flow	8×5	42.9M	63G	293.3ms	50.7vps	60.1	41.4

Table 2. The inference latency of video shuffle networks with different backbones and modalities.

To measure the latency and throughput, we perform inference on one NVIDIA Tesla P100 GPU and use the average value of 500 times batch inference with batch size of 16. We provide the latency of VSN-R50 and VSN-101 with RGB and optical flow modalities. Optical flows are pre-extracted and saved as image format. The *vps* indicates the videos per second. We also report the accuracy on Kinetics and Something-Something-v1 using the standard center cropping. We can observe that our models have not only low FLOPs but also low latency and high throughput.