End-to-End Action Segmentation Transformer

Supplementary Material

6. Additional Results

Feature Extraction Time. As demonstrated in Table 14, offline feature extraction introduces significant time costs, ranging from hours (GTEA) to months (Assembly101). Therefore, our end-to-end method provides substantial reductions in both training and inference time compared to methods relying on pre-extracted features.

SOTA Methods with Low FPS Input. The impact of video downsampling on the SOTA methods and EAST is evaluated on Breakfast in Table 15. The SOTA methods are trained using I3D frame features sampled at 1 FPS, with their output framewise classification subsequently upscaled to the original 15 FPS, as per their reported evaluation setting. Table 15 shows significant performance degradation for all the SOTA methods when working with the low FPS at the input. As shown in Tables 15 and 9, EAST maintains the best performance at low FPS rates and improves as the frame rate increases, subject to memory and compute constraints.

Qualitative Results. Fig. 5 illustrates EAST detector's action segmentation on three example videos from 50Salads. For each video, frame labels and prediction scores are visualized in four rows, including (from top to bottom): (a) Ground-truth frame labels; (b) Predicted highest-scoring frame labels; (c) A hypothetical frame labeling guided by an oracle which replaces incorrect highest-scoring labels in (b) with the second-highest scoring class; and (d) maximum softmax score of the predicted class for each frame. The video title includes the video name and the accuracy of (b) and (c). By comparing (a) and (b) in Fig.5, we observe that labeling errors predominantly occur in frames with low soft-

Dataset	GTEA	50salads	Breakfast	Assembly101
Avg. Time/Video (min)	5.6	57.8	10.5	63.7
Videos	28	50	1712	6108
Total Time (h)	2.6	48.2	300.3	6430.2

Table 14. Feature extraction time for different datasets. "Avg. Time/Video" shows the average processing time per video. "Videos" denotes the number of videos. "Total Time" indicates the overall extraction time; times are based on an H100 GPU.

Method	FPS	F1@{10,25,50}			Edit	Acc
MSTCN [9]	15	52.6	48.1	37.9	61.7	66.3
('CVPR19)	1	72.5	65.8	49.8	71.1	67.9
ASFormer [39]	15	76.0	70.6	57.4	75.0	73.5
('BMCV21)	1	74.3	67.6	51.9	73.5	69.6
LTContext [3]	15	77.6	72.6	60.1	77.0	74.2
('ICCV23)	1	77.2	70.5	56.1	74.1	69.8
FACT [27]	15	81.4	76.5	66.2	79.7	76.2
('CVPR24)	1	76.3	70.7	56.8	74.4	70.9
EAST	3	**86.2**	**82.2**	**71.8**	**84.5**	**82.8**
	1	84.1	79.8	69.6	81.7	80.4

Table 15. Impact of FPS (frame per second) on SOTA and EAST performance on Breakfast.

Figure 5. EAST detector's framewise classification and softmax scores on example videos from 50Salads. The top two videos depict the same training video (original vs. augmented predictions), while the bottom video, representing one of the worst cases, is from the evaluation set. A color-coded legend for frame labels and softmax score ranges is shown below.

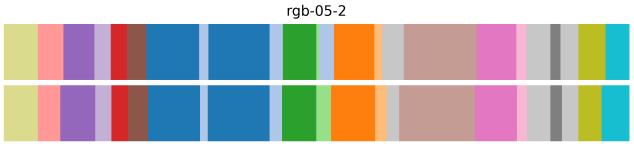

Figure 6. (top) Ground truth and (bottom) EAST's final framewise result for the video rgb-05-2 of 50Salads also considered in Fig. 5.

max scores, indicating that our detector is reliably trained. In (c), most errors can be corrected by replacing the incorrect highest-scoring class with the second highest, assuming access to an oracle. This highlights the potential for self-correction by refining predictions at frames with low softmax scores — the main purpose of the aggregator. Fig.6 demonstrates that EAST effectively refines the initially predicted action segments by the detector, enhancing their alignment with the ground-truth labels.