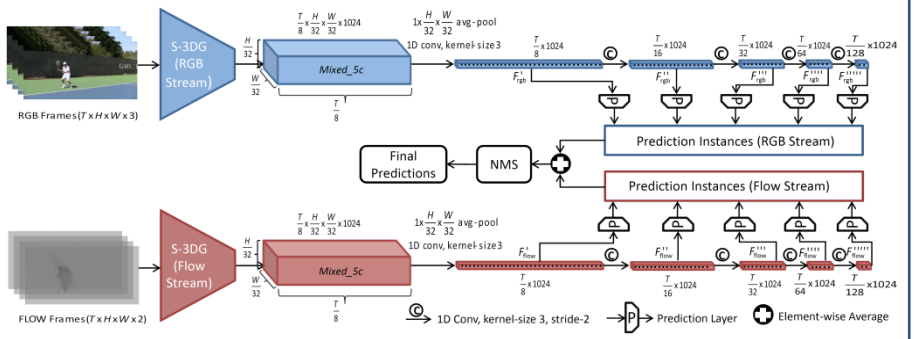
Single-Stage End-to-End Temporal Activity Detection in Untrimmed Videos

Md Atiqur Rahman and Robert Laganière
School of Electrical Engineering and Computer Science, University of Ottawa, Canada

Introduction: In this paper, we address the problem of human activity detection in temporally untrimmed long video sequences, where the goal is to classify and temporally localize each activity instance in the input video. Inspired by the recent success of the single-stage object detection methods, we propose a single-stage end-to-end trainable framework that learns task-specific spatio-temporal feature representations of a video sequence followed by a multi-scale detection architecture for direct classification and localization of the activities. Our proposed approach sets new state-of-the-art on the highly challenging THUMOS'14 benchmark – up from 44.2% mAP to 49.0% mAP .

Proposed Approach: We closely follow the architecture of the leading single-stage object detector called SSD. Though there are existing works that follow the SSD architecture, these methods cannot afford end-to-end learning of spatio-temporal features. Below is the architecture of our proposed approach.



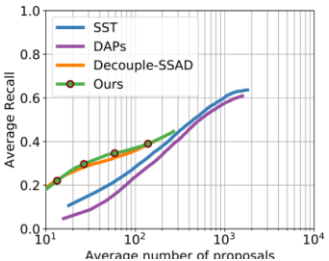
Dataset and Experiments: We conducted experiments on the THUMOS'14 dataset. Training is performed on the 200 untrimmed validation videos and results are reported on the 210 untrimmed test videos. We evaluate our proposed approach on two related tasks – Temporal Activity Detection and Temporal Activity Proposal.

Evaluation on End-to-End Feature Learning: The below table shows the performance of the proposed approach as we train different number of layers of the feature extractor 3D CNN. This results clearly validates our design choice for end-to-end feature learning, something the existing single-stage approaches cannot afford.

Fixed Training	Train the last layer	Train the last two layers
46.42 mAP(%)	47.48 mAP(%)	**48.97 mAP(%)**

State-of-the-Art Comparisons: The table on the left compares our proposed approach with the existing state-of-the-art approaches to temporal activity detection on THUMOS'14. On the right-top, we plot Average Recall against number of proposals per video, while on the right-bottom, Recall for top 100 predictions are plotted against higher tIoU thresholds. Our proposed approach outperforms the other methods.

Stage	Method	mAP @tIoU (%)						
		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Two-Stage	SCNN [27]	47.7	43.5	36.3	28.7	19.0	10.3	5.3
	DAP [33]	-	-	-	-	13.9	-	-
	SST [26]	-	-	41.2	31.5	20.0	10.9	4.7
	CDC [15]	-	-	40.1	29.4	23.3	13.1	7.9
	TURN [32]	54.0	50.9	44.1	34.9	25.6	-	-
	TCN [31]	-	-	-	33.3	25.6	15.9	9.0
	R-C3D [16]	54.5	51.5	44.8	35.6	28.9	19.1	9.3
	SSN [28]	66.0	59.4	51.9	41.0	29.8	19.6	10.7
	CBR [30]	60.1	56.7	50.1	41.3	31.0	19.1	9.9
	BSN [41]	-	-	53.5	45.0	36.9	28.4	20.0
	TAL-net [10]	59.8	57.1	53.2	48.5	42.8	**33.8**	**20.8**
Single-Stage	Yeung et. al. [34]	48.9	44.0	36.0	26.4	17.1	-	-
	SSAD [11]	50.1	47.8	43.0	35.0	24.6	15.4	7.7
	SS-TAD [9]	-	-	45.7	-	29.2	-	9.6
	G-TAN [12]	69.1	63.7	57.8	47.2	38.8	-	-
	Decouple-SSAD [13]	-	-	60.2	54.1	44.2	32.3	19.1
	Ours	**69.5**	**68.4**	**66.1**	**61.1**	**49.0**	32.9	16.7



Qualitative Results: Results on two test videos for two consecutive activity instances are visualized below.