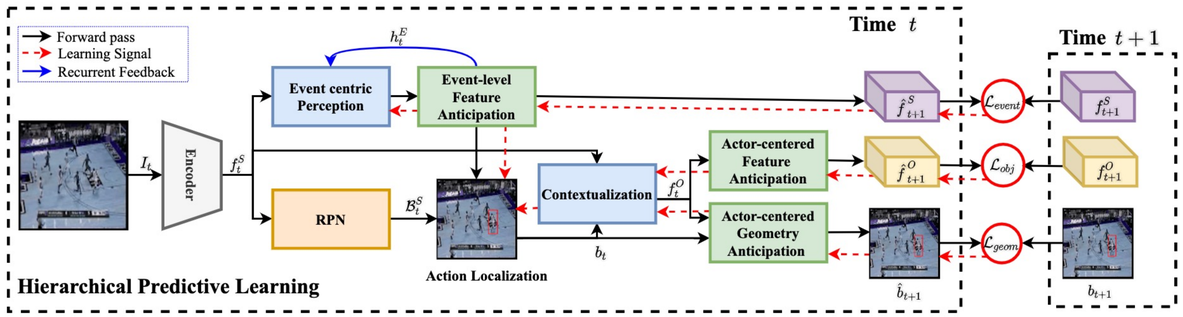
Actor-centered Representations for Action Localization in Streaming Videos

Motivation

- Event perception tasks such as action recognition and localization are important for visual understanding
- Progress has largely been driven by the use of large-scale, annotated training data in a supervised manner.
- **Goal:** Can we learn robust representations for video sequences for localizing the action in streaming videos?

Overall Framework



Contributions

- Introduce *hierarchical predictive learning* for unsupervised action localization in *streaming* videos
- Introduce a novel, attention-driven formulation for learning robust, actor-centered event features
- Demonstrate extension to to multi-actor group activity recognition and localization and generalization to data outside the training domain *without finetuning*

Quantitative Evaluation

Data: UCF Sports, THUMOS'13, JHMDB *Metrics*: mAP@IoU, Recall

Comparison to state-of-the-art approaches on action localization across different supervision needs, datasets and IOU thresholds

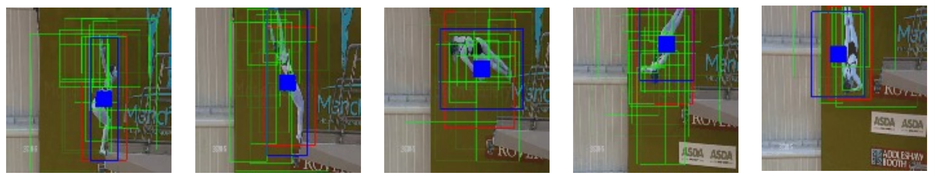
Approach	Supervision		UCF Sports		JHMDB		THUMOS'13	
	Spatial	Label	σ=0.2	σ=0.5	σ=0.2	σ=0.5	σ=0.2	σ=0.5
Tube CNN [12]	✓	✓	0.47	-	-	0.77	0.47	0.41
Action Tubelets [14]	✓	✓	0.53	0.27	-	-	0.48	-
Action Tubes [9]	✓	✓	0.56	0.49	0.55	0.45	-	-
MRSTL [48]	✓	✓	-	-	-	0.37	-	0.68
MENET [24]	✓	✓	-	-	-	**0.82**	-	**0.84**
HISAN [27]	✓	✓	-	-	-	0.77	-	0.73
ACAR-Net [26]	✓	✓	-	-	-	-	-	0.84
ALSTM [32]	✗	✓	-	-	-	-	0.06	-
VideoLSTM [21]	✗	✓	-	-	-	-	0.37	-
Actor Supervision [6]	✗	✓	-	**0.48**	-	**0.36**	**0.46**	-
Soomro *et al* [37]	✗	✗	0.46*	0.30*	**0.43***	**0.22***	0.21*	0.06*
PredLearn [3] (k=k_{gt})	✗	✗	0.55	0.32	0.30	0.10	0.31	0.10
AC-HPL (Ours, k=k_{gt})	✗	✗	**0.70**	**0.59**	**0.43**	0.15	**0.38**	**0.20**

Generalization to outside domain training data without finetuning

Test Data → / Train Data ↓	UCF Sports		JHMDB		THUMOS'13	
	AC-HPL	PredLearn	AC-HPL	PredLearn	AC-HPL	PredLearn
	σ=0.5		σ=0.2		σ=0.2	
UCF Sports	**0.59**	0.32	**0.39**	0.19	**0.38**	0.20
JHMDB	**0.48**	0.23	**0.43**	0.30	**0.35**	0.26
THUMOS'13	**0.50**	0.27	**0.40**	0.24)	**0.38**	0.31

Qualitative Results

Generic Single Actor localization



Multi-Actor localization