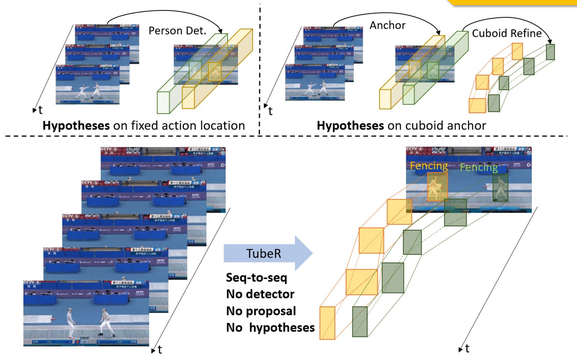
TubeR: Tubelet Transformer for Video Action Detection

Jiaojiao Zhao[1*], Yanyi Zhang[2*], Xinyu Li[3*], Hao Chen[3], BingShuai[3], Mingze Xu[3], Chunhui Liu[3], Kaustav Kundu[3], Yuanjun Xiong[3], Davide Modolo[3], Ivan Marsic[2], Cees G.M. Snoek[1], Joseph Tighe[3]

[1]University of Amsterdam [2]Rutgers University [3]AWS AI Labs

Introduction



TubeR takes as input a video clip and directly outputs **tubelets: sequences of bounding boxes and their action labels.**
TubeR runs **end-to-end** without person detectors, anchors or proposals.

Our contributions are:
1. TubeR: **a tubelet-level transformer** framework for action detection.
2. Tubelet query and attention-based formulation is able to **generate tubelets of arbitrary location and scale**.
3. Context aware classification head is able to aggregate **short-term and long-term contextual information**.
4. **State-of-the-art results** on three challenging action detection datasets.

TubeR

Encoder

The TubeR encoder is designed for processing information in the 3D spatio-temporal space.

Decoder

The decoder contains a tubelet-attention module and a cross-attention (CA) layer which is used to decode the tubelet-specific feature.

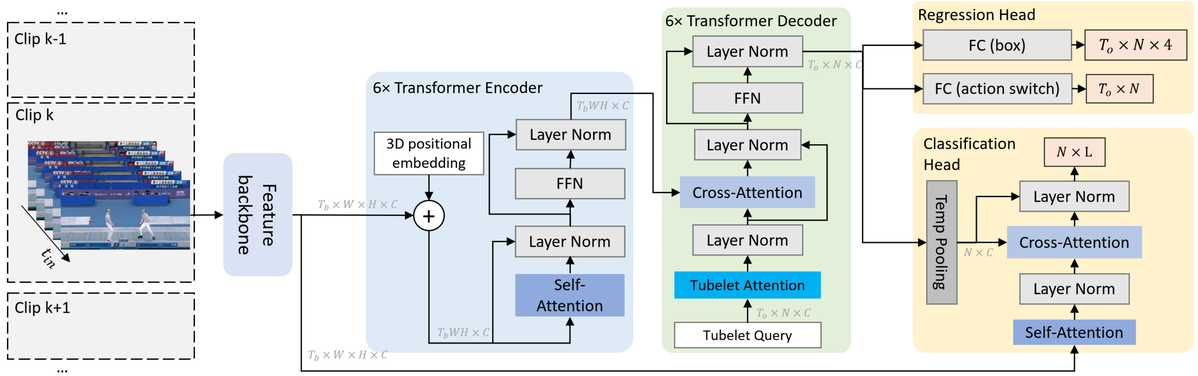
Tubelet Query: We propose to learn a small set of tubelet queries driven by the video data. TubeR uses a tubelet query to represent the dynamics of a tubelet, instead of hand-designing 3D anchors.

Tubelet Attention: Consists of a spatial-attention that learns the relative positional association between bounding boxes on each frame and a temporal-attention that links boxes for each action temporally.

Task-Specific Heads

Action switch regression head: The bounding boxes in a tubelet are simultaneously regressed with an FC layer. The action switch allows our method to generate action tubelets with a more precise temporal extent.

Context aware classification head. The classification can be simply achieved with a linear projection. We further propose to leverage spatio-temporal video context to help video sequence understanding.



Results

Model	Detector	Input	Backbone	Pre-train	Inference	GFLOPs	mAP
Comparison to end-to-end models							
I3D [15]	✗	32 × 2	I3D-VGG	K400	1 view	NA	14.5
ACRN [35]	✗	32 × 2	S3D-G	K400	1 view	NA	17.4
STEP [45]	✗	32 × 2	I3D-VGG	K400	1 view	NA	18.6
VTr [13]	✗	64 × 1	I3D-VGG	K400	1 view	NA	24.9
WOO [5]	✗	8 × 8	SF-50	K400	1 view	142	25.2
TubeR	✗	16 × 4	I3D-Res50	K400	1 view	132	**26.1**
TubeR	✗	16 × 4	I3D-Res101	K400	1 view	246	**28.6**
Comparison to two-stage models							
Slowfast-50 [9]	F-RCNN	16 × 4	SF-50	K400	1 view	308	24.2
X3D-XL [8]	F-RCNN	16 × 5	X3D-XL	K400	1 view	290	26.1
CSN-152*	F-RCNN	32 × 2	CSN-152	IG + K400	1 views	342	27.3
LFB [43]	F-RCNN	32 × 2	I3D-101-NL	K400	18 views	NA	27.7
ACAR-NET [28]	F-RCNN	32 × 2	SF-50	K400	6 views	NA	28.3
TubeR	✗	32 × 2	CSN-50	K400	1 view	78	**28.8**
TubeR	✗	32 × 2	CSN-152	IG + K400	1 view	120	**31.7**
Comparison to best reported results							
WOO [5]	✗	8 × 8	SF-101	K400+K600	1 view	246	28.0
SF-101-NL [9]	F-RCNN	32 × 2	SF-101+NL	K400+K600	6 views	962	28.2
ACAR-NET [28]	F-RCNN	32 × 2	SF-101	K400+K600	6 views	NA	30.0
AIA [37]	F-RCNN	32 × 2	SF-101	K400+K700	18 views	NA	31.2
TubeR	✗	32 × 2	SF-101	K400+K700	1 view	240	**31.6**
TubeR	✗	32 × 2	CSN-152	IG + K400	2 view	240	**32.0**

Comparison on AVA v2.1 validation set.

Tuber is effective

TubeR outperforms the most recent **end-to-end works** WOO by 0.9% and VTr by 1.2%.
TubeR with CSN backbones outperforms **the two-stage model** with the same backbone by +4.4%.

TubeR is efficient

Our TubeR has 8% fewer FLOPs than the most recently **end-to-end model** WOO and is 4× more efficient than **the two-stage model** slowfast with noticeable performance gain.

	Backbone	f-mAP	*UCF101-24* 0.20	0.50	0.50:0.95	*JHMDB51-21* 0.20	0.50
RGB-stream							
MOC [26]	DLA34	72.1	78.2	50.7	26.2	-	-
TubeR	Res50	79.5	81.2	55.1	28.1	-	-
T-CNN [16]	C3D	41.4	47.1	-	-	78.4	76.9
TubeR	I3D	80.1	82.8	57.7	28.6	79.7	78.3
TubeR	CSN-152	**83.2**	**83.3**	**58.4**	**28.9**	**87.4**	**82.3**
Two-stream							
TacNet [33]	VGG	72.1	77.5	52.9	24.1	-	-
2in1 [49]	VGG		78.5	50.3	24.5	-	74.7
ACT [19]	VGG	67.1	77.2	51.4	25.0	74.2	73.7
MOC [26]	DLA34	78.0	82.8	53.8	28.3	77.3	77.2
STEP [45]	I3D	75.0	76.6	-	-	-	-
I3D [15]	I3D	76.3	-	59.9	-	-	78.6
*CFAD [25]	I3D	72.5	81.6	**64.6**	26.7	**86.8**	**85.3**
TubeR	I3D	**81.3**	**85.3**	60.2	**29.7**	81.8	80.7

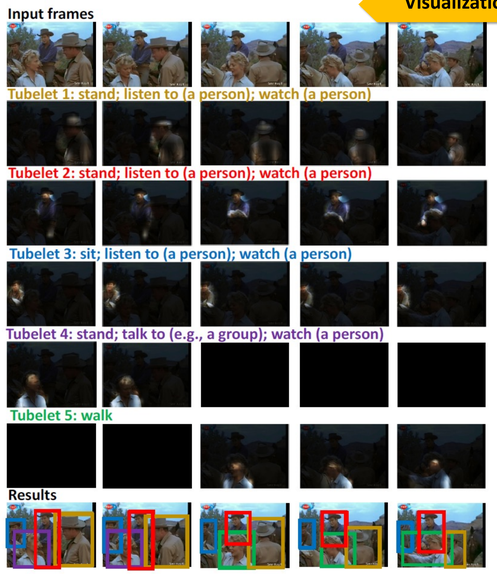
Comparison on UCF101-24 and JHMDB51-21 with video-mAP.

With same or comparable backbone: TubeR outperforms most of previous work with both RGB and two-stream inputs.

With stronger backbone: TubeR with RGB inoput outperforms previous works even with two-stream inputs.

Frame-mAP: TubeR also achieves the SOTA frame mAP comparing with previous works by a large margin.

Visualization



Visualization of **tubelet specific feature** with attention rollout.
1. TubeR can generate highly discriminative tubelet-specific features.
2. Action switch works as expected and initiates/cuts the tubelets.
3. TubeR generalizes well to scale changes (the brown tubelet).
4. Tubelets are tightly associated with tubelet specific feature.

Visualizations of **challenging** cases.
Top: shot changes;
Middle: actors moving with distance;
Bottom: multiple actors with small and large scales.