# TubeR: Tubelet Transformer for Video Action Detection
## *Supplementary Material*

Jiaojiao Zhao[1]*, Yanyi Zhang[2]*, Xinyu Li[3]*, Hao Chen[3], Bing Shuai[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

**Losses.** During TubeR training, we first produce an optimal bipartite matching $\delta$ between predictions and ground truth tubelets. $\delta(i)$ is the index of the prediction matched with the $i$-th ground-truth tubelet. We need to calculate the losses between a set of ground-truth tubelets $\mathbf{Y}=(Y_{\text{coor}}, Y_{\text{switch}}, Y_{\text{class}})$ and the matched predictions $\mathbf{y}=(y_{\text{coor}}, y_{\text{switch}}, y_{\text{class}})$.

We utilize four losses: an action classification loss, a box matching loss, a generalized IoU [3] loss and an action switch loss to train TubeR. The total loss is a linear combination of the four losses:

$$\begin{aligned}\mathcal{L} = \lambda_1 \mathcal{L}_{\text{switch}}(y_{\text{switch}}, Y_{\text{switch}}) + \lambda_2 \mathcal{L}_{\text{class}}(y_{\text{class}}, Y_{\text{class}}) \\ +\lambda_3 \mathcal{L}_{\text{box}}(y_{\text{coor}}, Y_{\text{coor}}) + \lambda_4 \mathcal{L}_{\text{iou}}(y_{\text{coor}}, Y_{\text{coor}}).\end{aligned} \tag{1}$$

$$\begin{aligned}\mathcal{L}_{\text{class}} = -\sum_{i=1}^{N}\sum_{j=1}^{L} [Y_{\text{class}}(i,j) \log y_{\text{class}}(\delta(i),j) \\ +(1-Y_{\text{class}}(i,j)) \log(1-y_{\text{class}}(\delta(i),j))] .\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{\text{switch}} = -\sum_{i=1}^{N}\sum_{j=1}^{T_{\text{out}}} [Y_{\text{switch}}(i,j) \log y_{\text{switch}}(\delta(i),j) \\ +(1-Y_{\text{switch}}(i,j)) \log(1-y_{\text{switch}}(\delta(i),j))] .\end{aligned} \tag{3}$$

$$\mathcal{L}_{\text{box}} = \sum_{i=1}^{N}\sum_{j=1}^{T_{\text{out}}} \|Y_{\text{coor}}(i,j) - y_{\text{coor}}(\delta(i),j)\|_1 . \tag{4}$$

$$\mathcal{L}_{\text{iou}} = \sum_{i=1}^{N}\sum_{j=1}^{T_{\text{out}}} \mathfrak{G}_{\text{iou}}(Y_{\text{coor}}(i,j), y_{\text{coor}}(\delta(i),j)), \tag{5}$$

$$\mathfrak{G}_{\text{iou}}(b,\hat{b}) = 1 - \left( \frac{|b \bigcap \hat{b}|}{|b \bigcup \hat{b}|} - \frac{|B(b,\hat{b}) \backslash b \bigcup \hat{b}}{B(b,\hat{b})} \right). \tag{6}$$

Here $\mathfrak{G}_{\text{iou}}(b,\hat{b})$ is the generalized IoU [3] loss between two given boxes $b$ and $\hat{b}$. We empirically set the scale parameter as $\lambda_1$=1, $\lambda_2$=5, $\lambda_3$=2, $\lambda_4$=2.

*Equally contributed.

| Model | f-mAP@IoU=0.5 |
|---|---|
| Baseline (using per-frame boxes) [1] | 22.8 |
| Only with tubelets | 27.7 |
| Long-term context without tubelets | 25.8 |
| Long-term context + tubelets | **28.8** |

Table 1. **Comparisons between a two-stage baseline and TubeRs**. All TubeRs performs significantly better than the baseline.

**TubeR vs. hypotheses-based method on UCF101-24.** We compare TubeR and [2], which depends on positional hypotheses to do detection on UCF101-24, with per-class *Video-mAP*@0.5 in Figure 1. For actions with multiple people, TubeR detects the action more precisely and produces higher video-mAP, like 44.83% for 'BasketballDunk' compared to [2] with video-mAP 1.19%. The tubelet attention mechanism better models the relations between the real action tubelets and surroundings. We note that [2] hardly works for 'Basketball' and 'TennisSwing' which have many transitional states. TubeR improves significantly for theses action categories. TubeR performs slightly worse for 'LongJump' in which actors may change scales along time. As [2] applies multiple scale anchors and multiple level features, it is more robust in this case. Incorporating multiple level features into TubeR will further help improve TubeR results.

**TubeR vs. two-stage method on AVA.** We use CSN-50 [4] as backbone with 1-view evaluation protocol unless specified otherwise. We report frame-mAP@IoU=0.5 for AVA v2.1. We compare the performance between a baseline [1] using offline person detection rather than a Region-Proposal-Network, and our variable TubeRs. We used the same input ($32\times2$). To make a fair comparison, the baseline is evaluated using bounding boxes generated by TubeR (93.3% AP for person localization). We clarify that the short-term context feature $F_{\text{b}}$ is the backbone feature, which is also used to generate tubelet-level feature $F_{\text{tub}}$. Together they belong to our tubelet design. The results are shown in Table 1. Only with tubelets achieves +5% frame-mAP compared to the baseline (using per-frame boxes) and improves more than

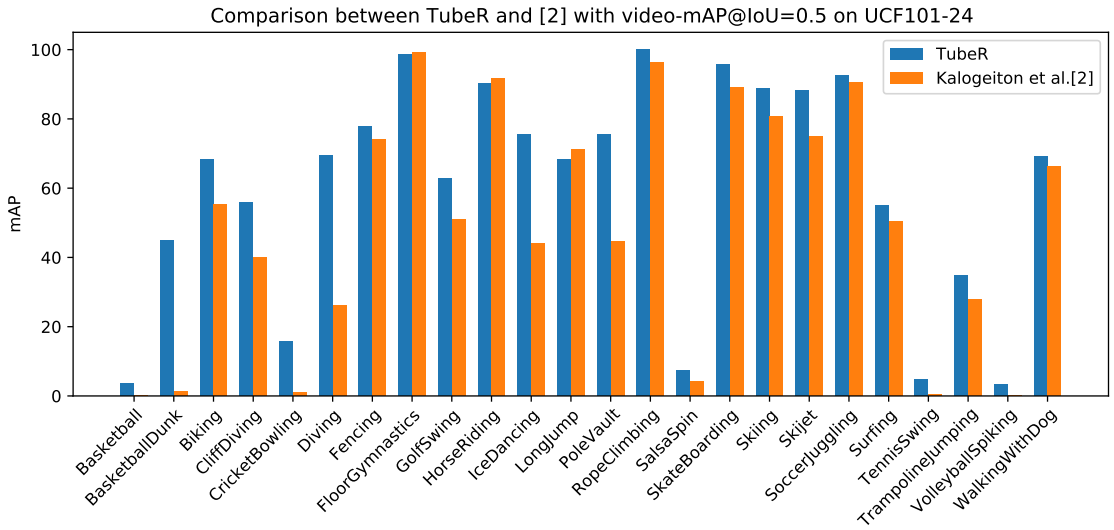


Figure 1. **Comparison between TubeR and a hypotheses-base detector on UCF101-24.** TubeR performs better on most of the action classes.

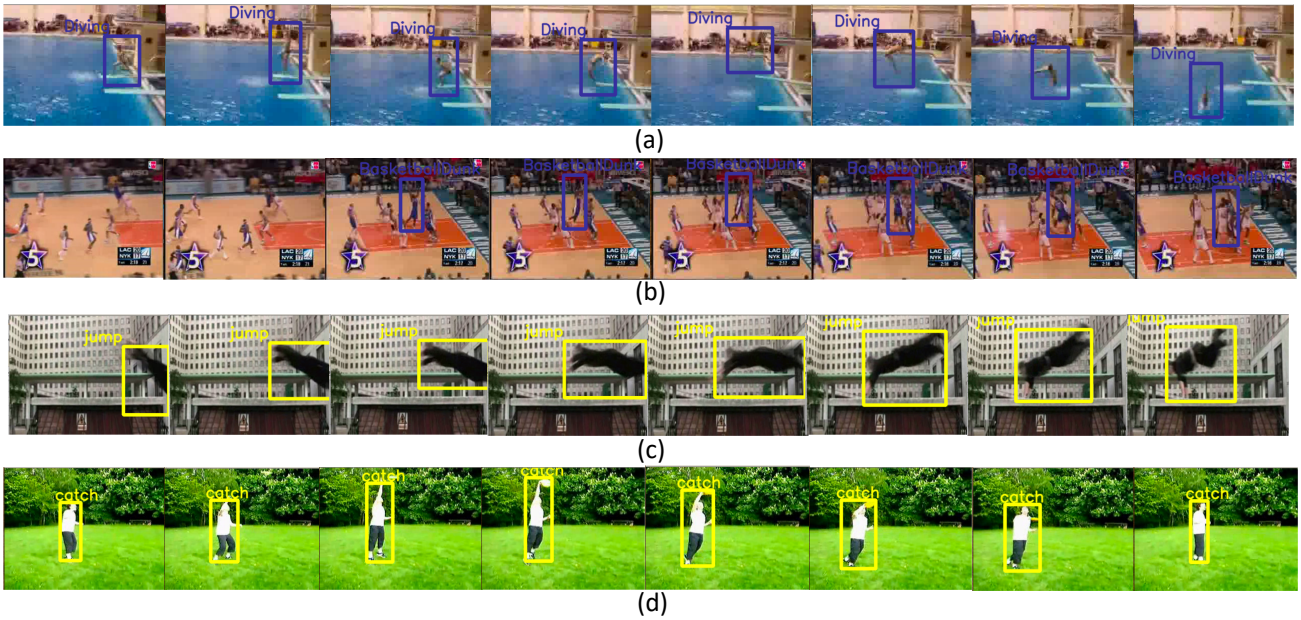


Figure 2. **Action tubelets visualization on UCF101-24 and JHMDB51-21.** Each action tubelet contains its action labels and boxes on each frame. (a-b) are from UCF101-24 to show the cases with deformable actors and crowded people. (c-d) are from JHMDB51-21 to show the fast action and interacted action.

long-term context without tubelets. Tubelet design not only brings performance gain, but also directly predicts tubelets without an offline linker. Our long-term context features are effective for long videos with shot changes. It results in a modest parameter increase from 70.1M to 84.3M, which is lower than most two-stage models.

**Visualization.** We show more action tubelets generated by TubeR in Figure 2. TubeR performs well in various cases. In Figure 2 (a-b), we show the cases with deformable actors and crowded people from UCF101-24. Figure 2 (c-d) present the fast action and interacted action from JHMDB51-21. Moreover, some challenging cases on AVA are visualized in Figure 3. All these cases show our TubeR is able to generate precise tubelets with various length.

## References

[1] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George

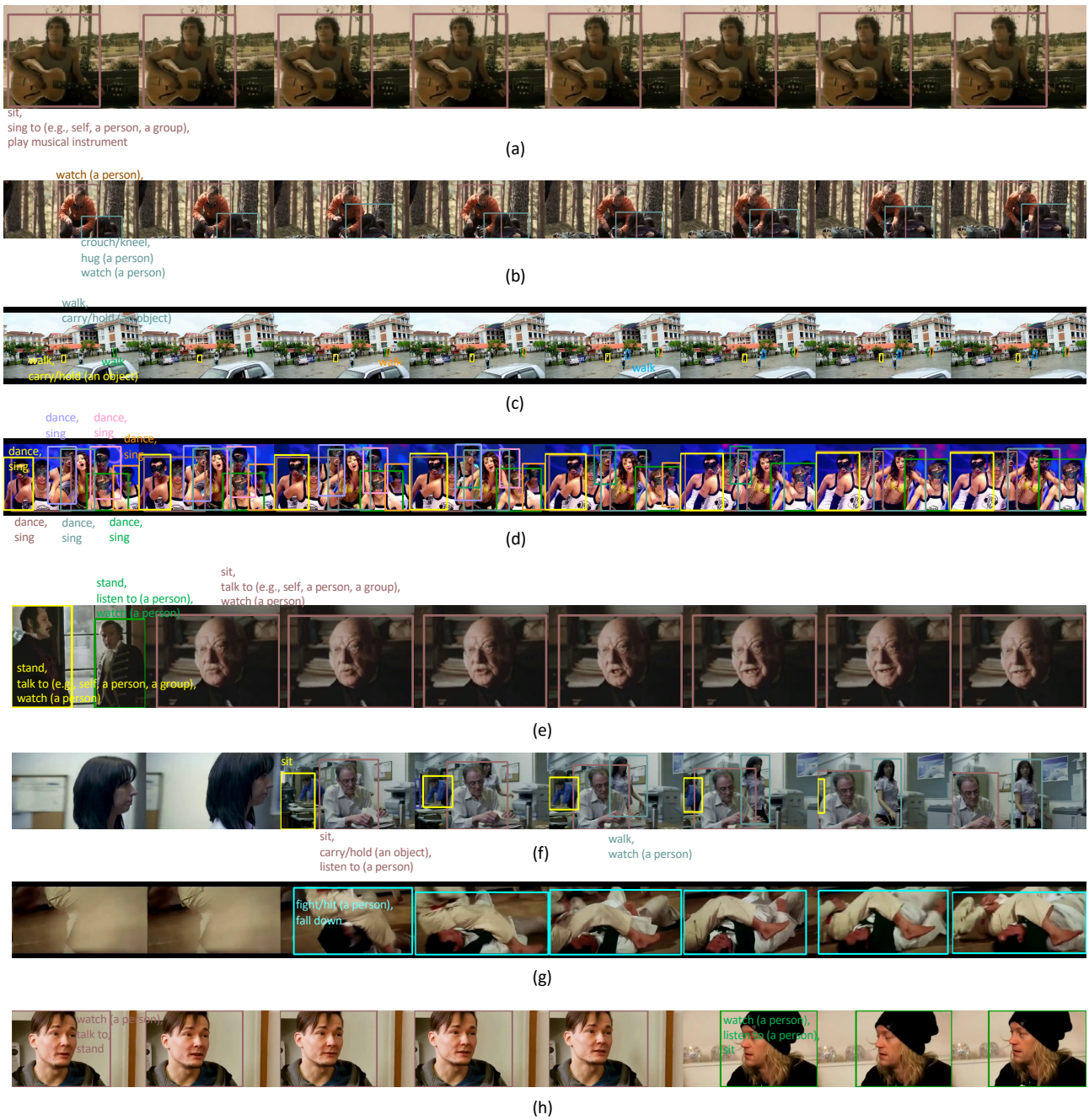


Figure 3. **Action tubelets visualization on AVA.** We use different colors to mark different tubelets. Each action tubelet contains its action labels and boxes on each frame. We only show the action labels on the first frame of an action tube. We show some challenging cases here. (a) and (b) Raw actions: "play musical instrument", "hug (a person)" . (c) Tiny actions. The actors are very tiny. (d) Crowded cases. (e-h) Shot cuts. All these cases show our TubeR is able to generate precise tubelets with various length.

Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 1

[2] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *ICCV*, 2017. 1

[3] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 1

[4] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *ICCV*, 2019. 1