Our contributions are:

2. TubeR query and attention-based formulation is able to generate action tublets with a more precise temporal extent.
3. Context aware classification head that allows the relative positional association between bounding boxes on each frame and a temporal-attention that links boxes for each action temporally.
4. Task-Specific Heads

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

Task-Specific Heads

4. Tubelets are tightly associated with tubelet specific feature.
5. TubeR generalizes well to scale changes (the brown tubelet).
6. TubeR can generate short-term and long-term contextual information.
7. State of the art results on three challenging action detection datasets. TubeR outperforms the most recent end-to-end model WOO by 0.9% and VfT by 1.2%.
8. TubeR with CNS backbones outperforms the two-stage model with the same backbone by +4.4%.

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.

Comparison on UCF101-24 and HMDB51-21 with video-motion.

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

With stronger backbone: TubeR with RGB input 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 of tubelet specific feature with attention rollout.

Action switch works as expected and initiates/cuts the tubelets.