Directional Temporal Modeling for Action Recognition

Xinyu Li, Bing Shuai, and Joseph Tighe

Amazon Web Service
{xxnl,bshuai,tighej}@amazon.com

1 Visualization

We visualized the activation map generated by our CIDC network and R3D-NL network under 4 scenarios:

1. when there is no camera motion and slight action movement (Figure 1);
2. when there is slight camera motion and slight action movement (Figure 2);
3. when there is slight camera motion and significant action movement (Figure 3);
4. when there is significant camera motion and significant action movement (Figure 4).

The results show that our proposed network is able to capture the activity related features and ignore the irrelevant background information compared with R3D-NL.

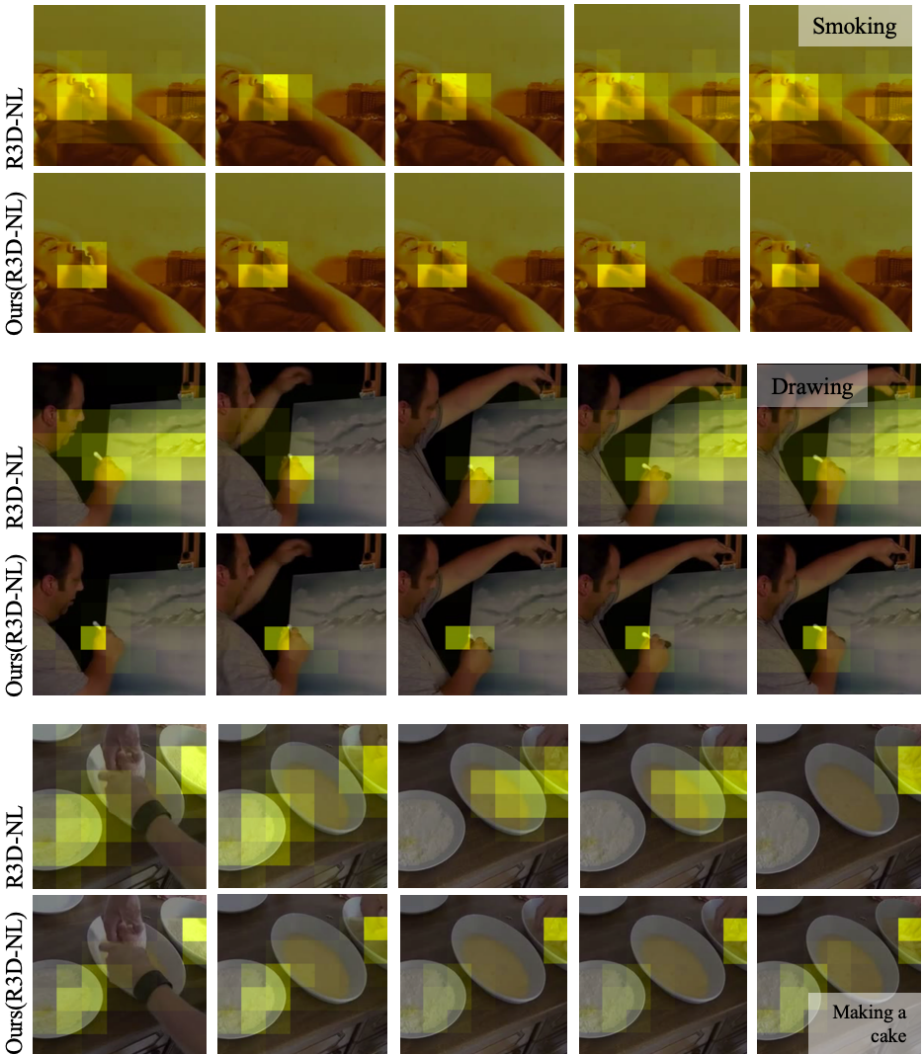

Fig. 1: We show a video clips **with no camera motion, slight action movement** (32 frames) and the spatial activation maps for the representative frame of every sub-clip (8 frames). Our proposed CIDC network is able to better capture the action related regions compared with R3D-NL.

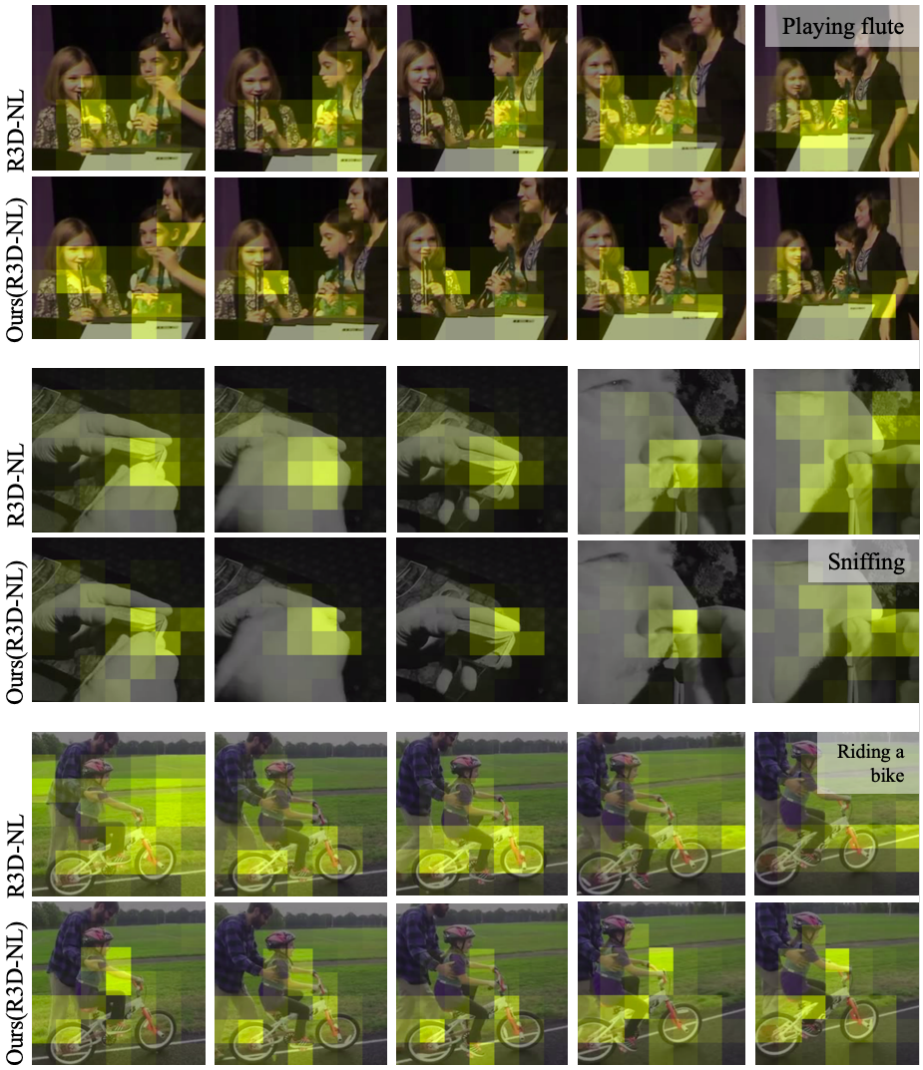

Fig. 2: We show a video clips **with slight camera motion, slight action movement** (32 frames) and the spatial activation maps for the representative frame of every sub-clip (8 frames). Our proposed CIDC network is able to better capture the action related regions compared with R3D-NL.

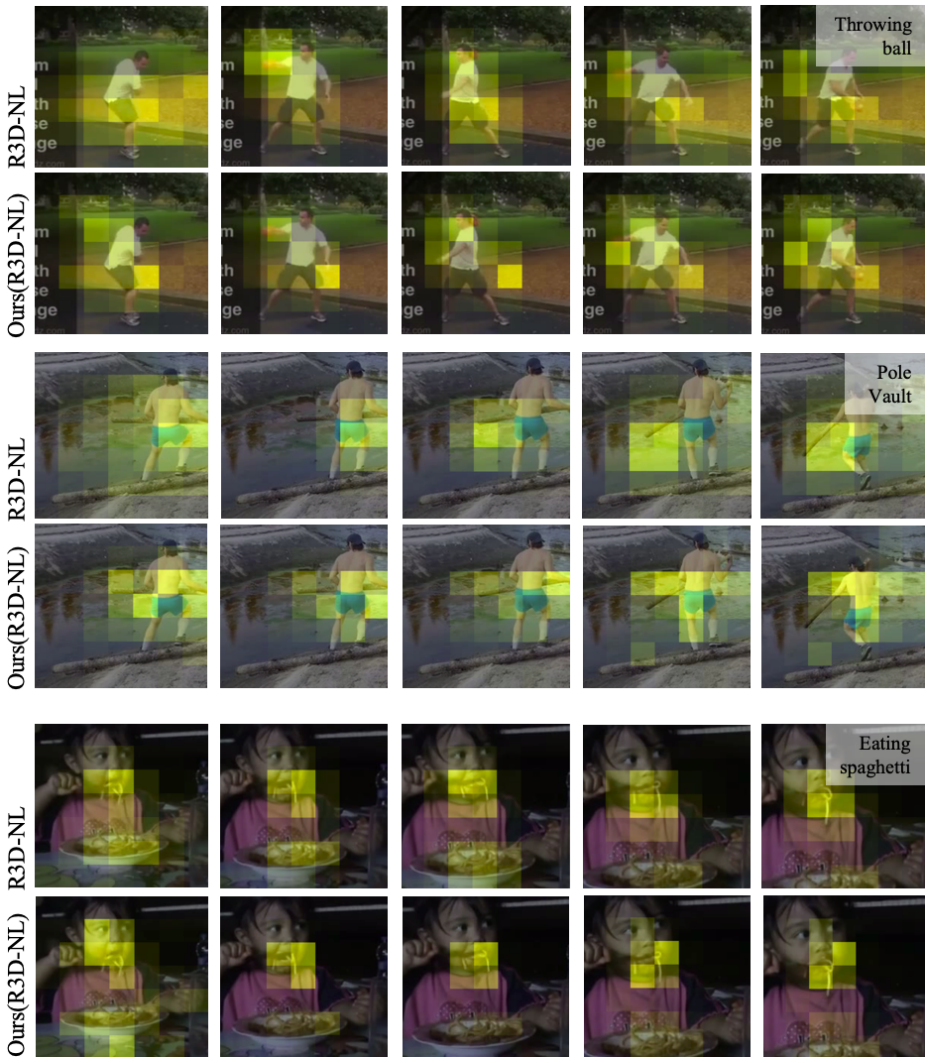

Fig. 3: We show a video clips **with slight camera motion, and significant action movement** (32 frames) and the spatial activation maps for the representative frame of every sub-clip (8 frames). Our proposed CIDC network is able to better capture the action related regions while R3D-NL is likely to capture the backgound features as well.

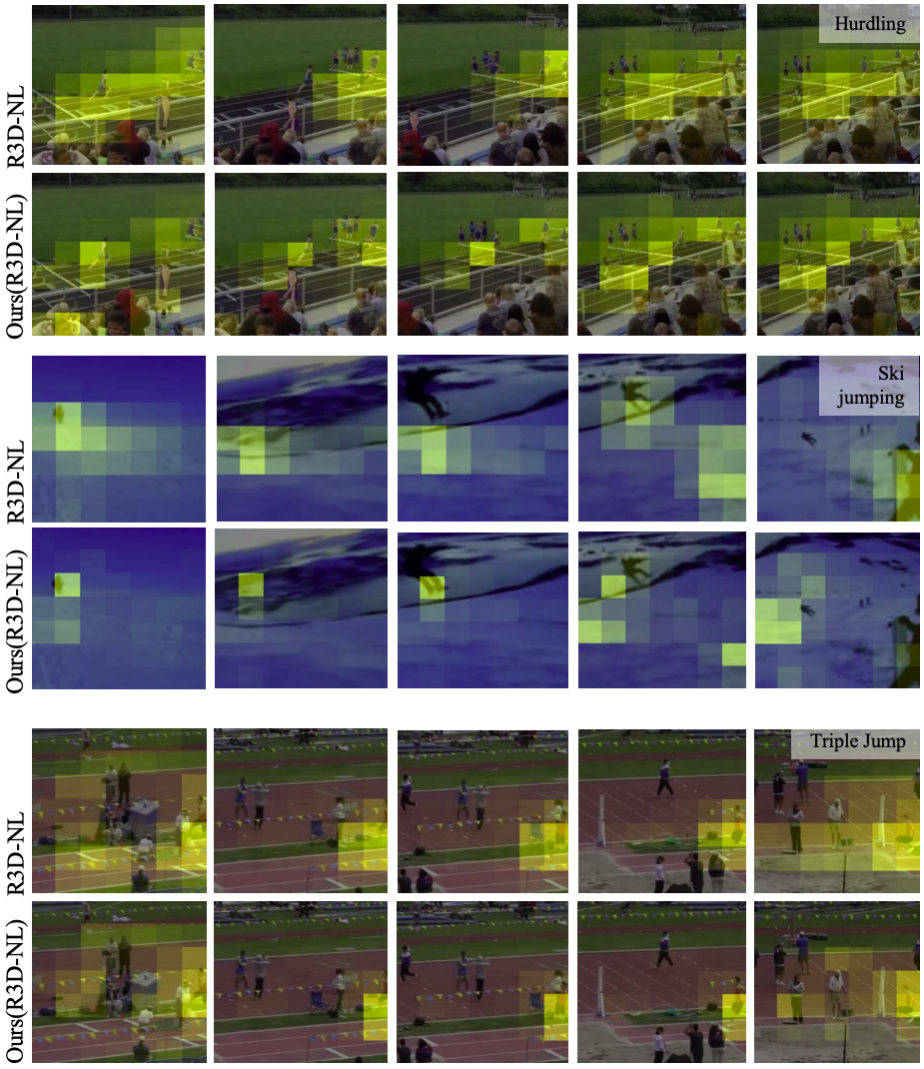

Fig. 4: We show a video clips **with significant camera motion, and significant action movement** (32 frames) and the spatial activation maps for the representative frame of every sub-clip (8 frames). Our proposed CIDC network is able to focus on activity related regions while the R3D-NL captures the background or irrelevant features under this scenario.