VidTr: Video Transformer Without Convolutions

Anonymous ICCV submission

Paper ID 6737

1. Fast VidTr

As a common practice, 3D ConvNets are usually tested on 30 crops per video clip (3 spatial and 10 temporal) that show performance boost while greatly increase the computation cost. The VidTr has been proved that learn long-term global spatio-temporal features better in a video clip, thus we propose to sample the data in TSN style (segment video into N chunks and randomly pick one frame from each chunk). During testing, we uniformly sample N frames from the video regardless the length of the video, and perform single-pass inference (center crop). Such design significantly reduce the inference computation and latency caused by the dense sampling with a very small performance drop (about 2%, see Table 1). Note that the R2D and I3D based methods do not work well with sparsely sampled frames, mainly because the convolution kernel has limited receptive field and can only aggregate features slowly. If adjacent frames are too far away from each other, the temporal convolution will not be able to establish the temporal relations well. We compare our fast VidTr model with pre-

Model	Input	Res.	GFLOPs	Latency(ms)	Top1
TSM [4]	$8fTSN$	256	330	170	74.1
3DEffi-B4 [3]	16×5	224	69	NA	72.4
TEINet [5]	16×4	256	990	1080	74.9
X3D-M [3]	16×5	224	47	1100	74.6
F-VidTr-S	8×8	224	1×39	37	72.9
F-VidTr-M	16×4	224	1×59	53	74.7

Table 1: Comparison of VidTr to other fast networks. All results from previous methods except for TEINet (30-crops) are based on 10 temporal crop and center spatial crop. The VidTr was achieved by uniformly sample 8/16/32 frames temporally and center-crop spatially.

vious SOTA light-weight models including TSM, TEINet and models from architecture search such as X3D on Kinetics 400 dataset and report the FLOPs, the latency and top1 accuracy with 10 center crops (Table 1) The results show that our proposed one-pass inference significantly outperforms the competitors with less FLOPs, lower latency and higher accuracy. The Fast VidTr (16 frames) is able to outperform TSM (+0.6% accuracy, 70% less FLOPs, 68% less latency); TEINet (-0.2% accuracy, 94% less FLOPs, 95% less latency), also note that the reported TEINet score is based on 30 crop evaluation; and X3D-M (+0.1% accuracy, 24% more FLOPs, 96% less latency). The results proves that the VidTr is able to aggregate long-term spatio-temporal features more effectively comparing the 3D ConvNets. It is worth mentioning that: 1. Even without considering the 10-crop evaluation required for ConvNets to achieve reported scores, the VidTr is still able to inference roughly at same speed comparing with TEINet and significantly faster than X3D. 2. X3D has low FLOPs but high latency mainly due to the heavily use of depth convolution.

2. More Ensemble Results

We provide additional ensemble results on Kinetics 400 (Table 2) and charades (Table 3), showing that the VidTr and 3D convolution based models can be complementary to each other, ensemble VidTr and 3D convolution based network significantly outperform the ensemble of any two 3D convolution based models. Our results show that the result level ensemble of I3D-101 and SOTA 3D model TPN-101 lead to about 1% accuracy boost and result level ensemble of VidTr-S with TPN-101 lead to about 3% performance boost. The similar conclusion can be draw from Charades on multi-label activities, where the ensemble of I3D-101 and CSN-152 only gives 2.8%mAP boost, while ensemble of VidTr-L with CSN-152 lead to SOTA (4.8%mAP boost over CSN-152) performance on Charades datasets.

3. Error Analysis Details

We show the top 5 classes that gains performance boost from VidTr and top 5 classes that got reduced performance from VidTr. The results (Table 4) show that the I3D generally performance well on local and fast action while the VidTr works well on actions require long-term temporal information. For example, the VidTr achieved 21.2 % accuracy improvement over I3D on "catching fish" that re-

Model	input	Ensemble	input	Top1	Top5
I3D50 [6]	16×4	-	-	75.0	92.2
I3D101 [6]	16×4	-	-	77.4	92.7
TPN101 [6]	16×4	-	-	78.2	93.4
I3D50 [6]	16×4	I3D101	16×4	77.7	93.2
TPN101[6]	16×4	I3D50	16×4	78.5	93.3
TPN101 [6]	16×4	I3D101	16×4	79.3	93.8
VidTr-S	8×8	I3D50	16×4	79.4	94.0
VidTr-S	8×8	I3D101	16×4	80.3	94.6
VidTr-S	8×8	TPN101	16×4	80.5	94.8

Table 2: More ensemble results on Kinetics-400 dataset. We report top 1 and top5 accuracy (%) on validation set.

Model	Input	Res.	Ensemble	Chad
I3D-Inception [2]	64×1	256	-	32.9
SlowFast-101-NL*	32×4	256	-	44.7
CSN-152*	32×4	256	-	46.4
En-I3D-101	32×4	256	I3D-50	42.1
En-I3D-101	32×4	256	SF-101	47.9
En-I3D-101	32×4	256	CSN-152	49.2
En-VidTr-L	32×4	224	I3D-101	47.3
En-VidTr-L	32×4	224	SF-101	48.9
En-VidTr-L	32×4	224	CSN-152	51.2

Table 3: Results on Charades dataset. The evaluation metrics are mean average precision (mAP) in percentage. * denotes the result that we re-produced.

quires long-term information from the status when the fish is in water to the final status after the fish is caught (Figure 1a). The VidTr performs worse than I3D on the activities that rely on slight motions (e.g., playing guitar, and shaking head, Figure 1b)

4. Visualization Details

We visualized the VidTr's separable-attention with attention roll-out method [1]. We multiplied all the affinity matrices between every two encoder layers and get $mask_t \in \mathbb{R}^{(WH+1)\times(T+1)\times(T+1)}$ for the temporal roll-out attention and $mask_s \in \mathbb{R}^{(T+1)\times(WH+1)\times(WH+1)}$ for the spatial roll-out attention. We selected the rows of class token from the roll-out attention for visualization as:

$$mask_t^{'} = mask_t^{(1:,0,1:)} \in \mathbb{R}^{WH\times T} \quad (1)$$

$$mask_s^{'} = mask_s^{(1:,0,1:)} \in \mathbb{R}^{T\times WH} \quad (2)$$

We multiplied $mask_t^{'}$ and $mask_s^{'}$ to represent the spatial-temporal attention for visualize as:

$$mask_{st}^{'} = Re(mask_t^{'}) \times mask_s^{'} \quad (3)$$

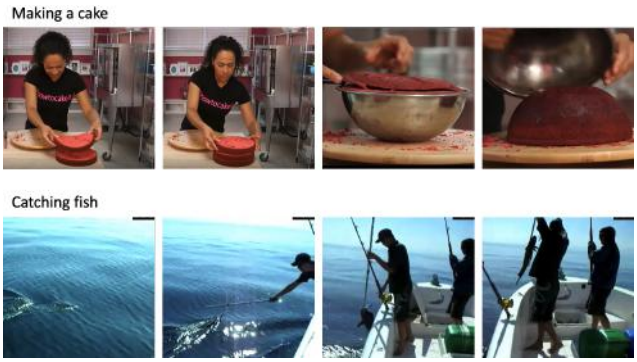
Top 5 (+)	Accuracy gain
making a cake	+26.0%
catching fish	+21.2%
catching or throwing baseball	+20.8%
stretching arm	+19.1%
spraying	+ 18.0 %

(a) Top 5 classes that VidTr works better than I3D.

Top 5 (-)	Accuracy gain
shaking head	-21.7%
dunking basketball	-20.8%
lunge	-19.9%
playing guitar	-19.9%
tap dancing	-16.3%

(b) Top 5 classes that I3D works better than VidTr.

Table 4: Quantitative analysis on Kinetics-400 dataset. The performance gain is defined as the disparity of the top-1 accuracy between VidTr network and that of I3D.



(a) video examples that VidTr performs better than I3D..

(b) video examples that VidTr performs worse than I3D.

Figure 1: Visualizations of video samples that VidTr works better and I3D works better.

where $mask_{st}^{'}$ is the spatio-temporal attention for visualize, and Re denotes a reshape function. We threshold $mask_s^{'}$

and $mask'_{st}$ by only highlighting the top 30% of values of them, and attached them onto the original frames for visualizing the spatio-only and spatio-temporal attentions.

4.1. More Visualizations

We first show more results of the VidTr's separable-attention with attention roll-out method [1] (Figure 2). We find that the spatial attention is able to focus on informative regions and temporal attention is able to skip the duplicated/non-representative information temporally.

We then show more results of the attention at 4th, 8th and 12th layer of VidTr (Figure 3), we found the spatial attention is getting to concentrate better when it goes to the deeper layer. The attention did not capture meaningful temporal instances at early stages because the temporal feature relies on the spatial information to determine informative temporal instances.

Finally we compared the I3D activation map and roll-out attention from VidTr (Figure 4). The I3D mis-classified the catching fish as sailing, as the I3D attention focused on the people sitting behind and water. The VidTr is able to make the correct prediction and the attention showed that the VidTr is able to focus on the action related regions across time.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, 2020.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[3] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.

[4] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019.

[5] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu. TEINet: Towards an Efficient Architecture for Video Recognition. In *The Conference on Artificial Intelligence (AAAI)*, 2020.

[6] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal Pyramid Network for Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

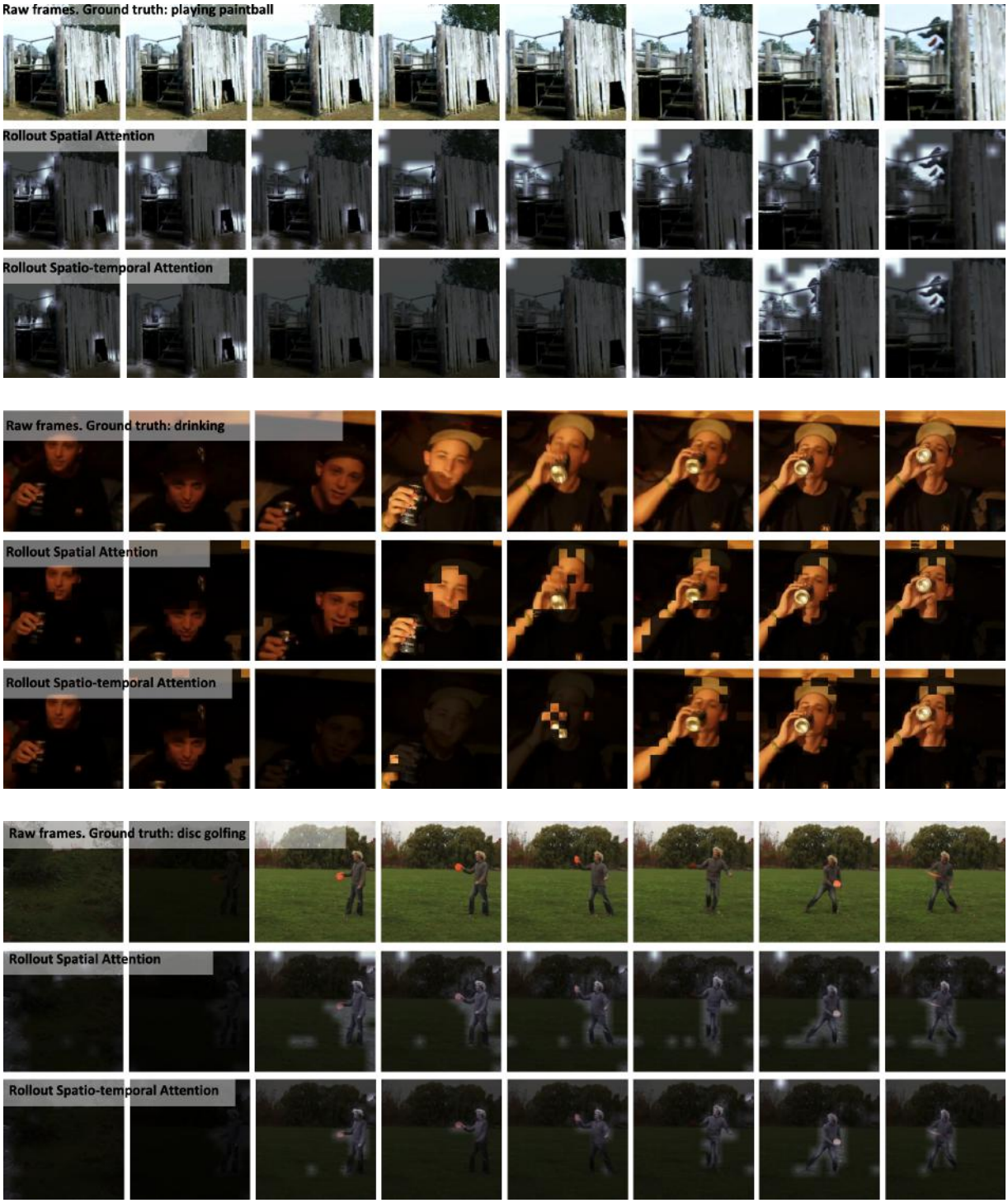


Figure 2: The spatial and temporal attention in Vidtr. The attention is able to focus on the informative frames and regions.

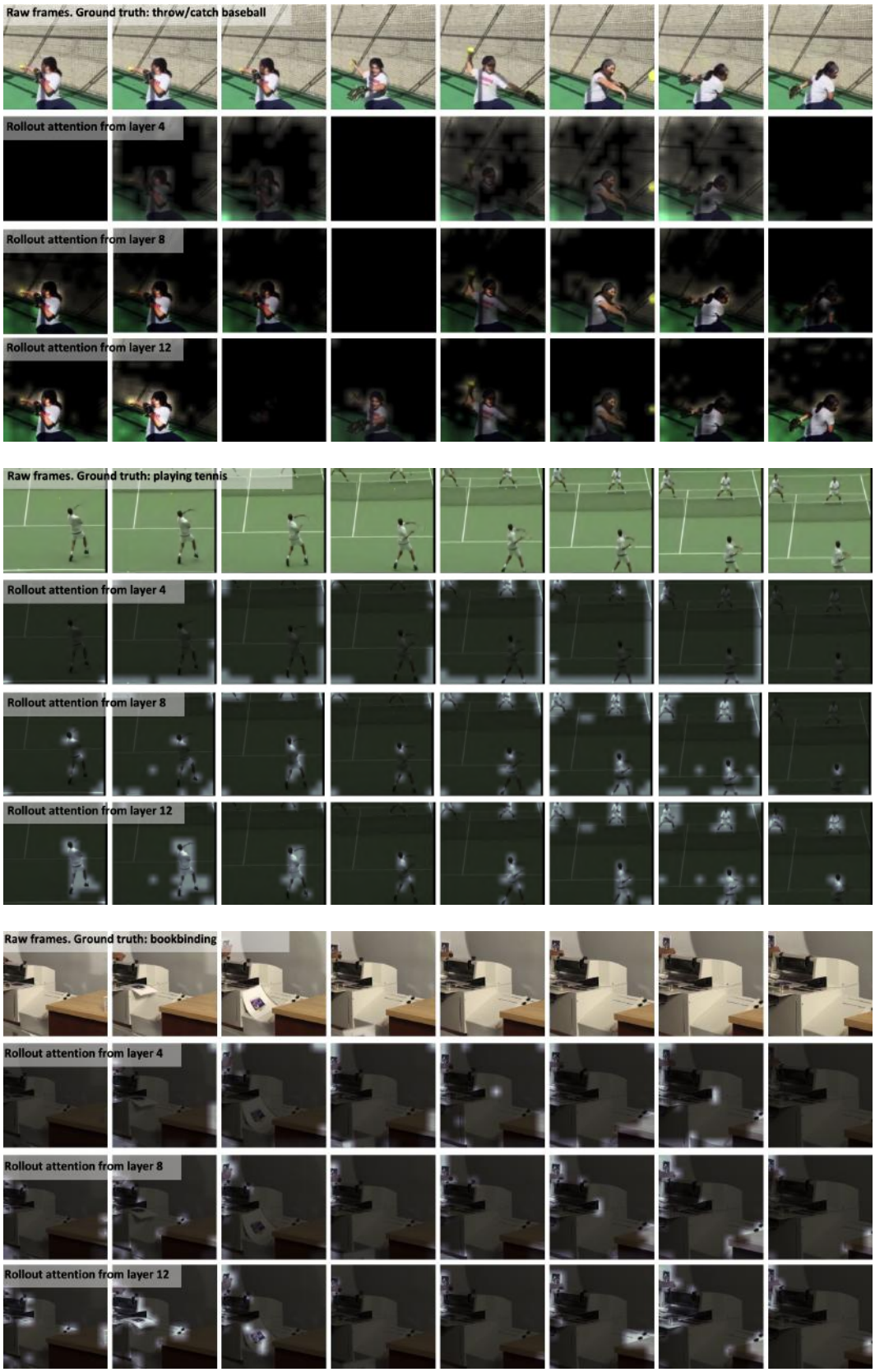


Figure 3: The rollout attentions from different layers of VidTr.

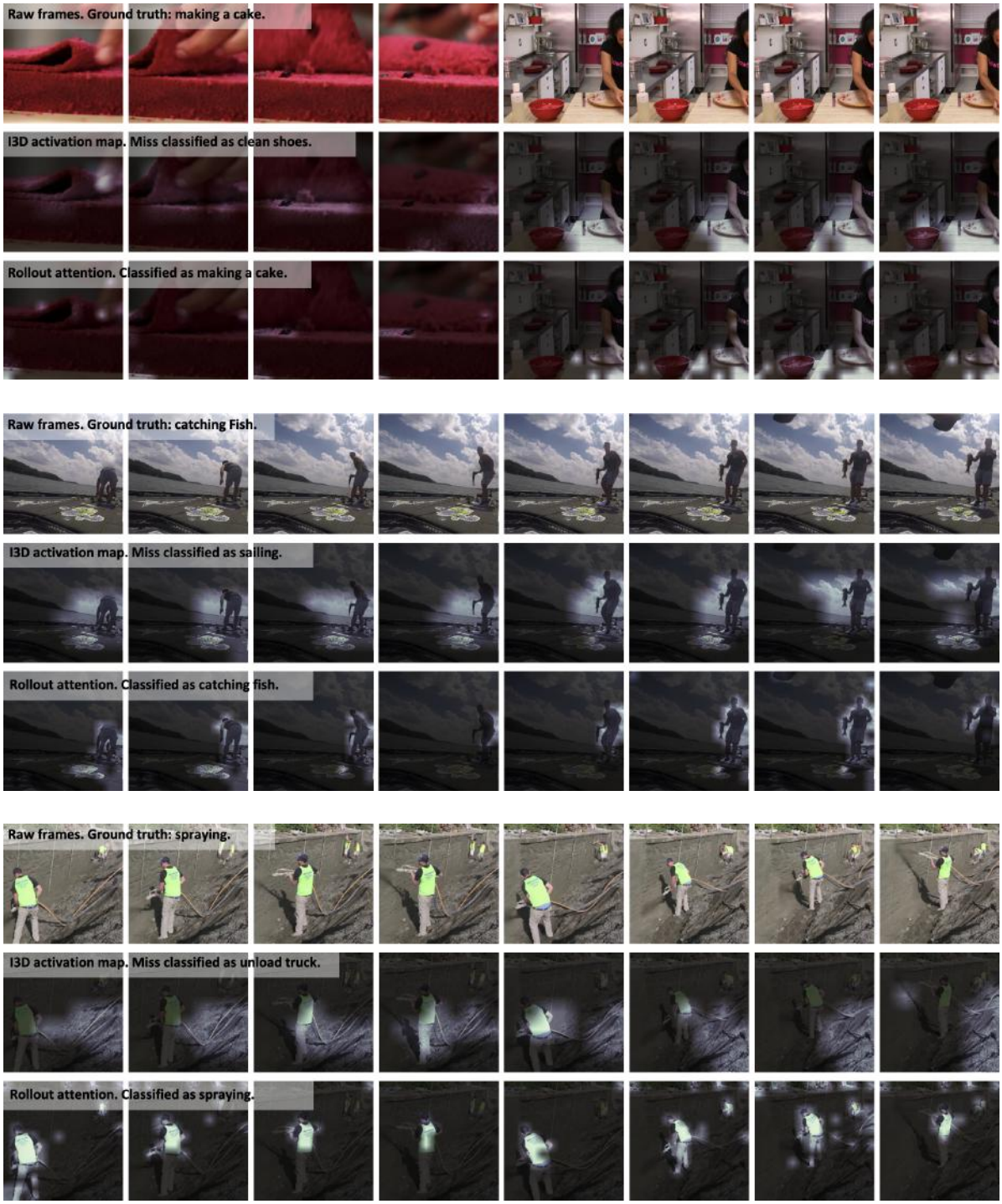


Figure 4: Comparison of I3D activations and VidTr attentions.