[Supplementary] Video Self-Stitching Graph Network for Temporal Action Localization

Chen Zhao Ali Thabet Bernard Ghanem
King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
{chen.zhao, ali.thabet, bernard.ghanem}@kaust.edu.sa

1. More Localization Results using Different Features

We provide results of our VSGN on two more types of features R2+1d [5] and TSP [1] for ActivityNet-v1.3 in Table 1. To generate both features, videos at frame rate 15 fps are used as input of the video representation networks. Non-overlapped snippets of $L = 16$ frames are extracted from each video, such that each snippet covers a temporal receptive field of approximately one second. We only use RGB streams for both features. It is worth-mentioning that TSP is trained with temporal action localization (TAL) annotations and on the TAL dataset, so our performance on TSP reflects its potential on end-to-end training.

Table 1: **Action localization results on validation set of ActivityNet-v1.3 with different features.**

Features	Stream	Snippets per sec.	0.5	0.75	0.95	Avg.	Short
R2+1d [5]	RGB	1	51.89	35.87	8.56	34.82	19.6
TSP [1]	RGB	1	53.26	36.76	8.12	35.94	20.9

2. More Ablation Results

In the Video self-stitching (VSS) component, we generate two different types of clips: Clip O and Clip U, and stitch them into one sequence as input to xGPN. In order to verify the effectiveness of using both clips, we provide ablation study to show the performance of only using either of them in Table 2. When only using Clip O or Clip U, we fill zeros in the positions of the other clip in the sequence. We can see that using both clips obviously outperforms using only one type of clip, in terms of both overall performance and performance of short actions.

Table 2: **Effectiveness of Clip O and Clip U in VSS.**

	THUMOS-14				ActivityNet-v1.3				
xGPN input	0.3	0.5	0.7	Short	0.5	0.75	0.95	Avg.	Short
Clip O	62.31	44.47	20.31	48.1	52.00	35.29	8.22	34.44	19.7
Clip U	66.37	49.25	23.70	53.0	52.06	35.44	**9.20**	34.72	18.0
Clip O & U	**66.69**	**52.45**	**30.40**	**56.6**	**52.38**	**36.01**	8.37	**35.07**	**19.9**

In the cross-scale graph network (xGN), we use two branches: the temporal branch and the graph branch, and sum up the features of the two branches as output. In Table 3, we compare the results of using either branch and using both. We can see that using both branches obviously outperforms using only one branch, in terms of both overall performance and performance of short actions.

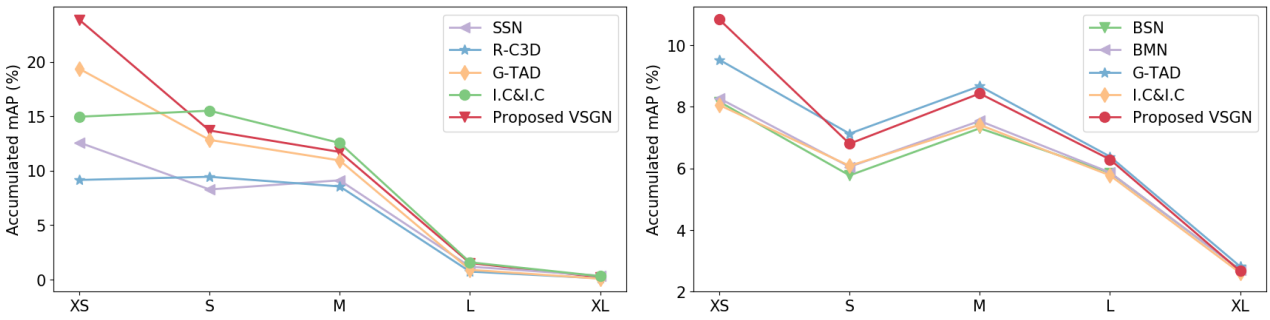
3. Performance at Different Temporal Scales

We illustrate the performance of VSGN on actions of different temporal scales in Fig. 1, where we evaluate the accumulated mAP of actions at different temporal scales. The accumulated mAP considers the number of action instances M_i and the average-mAP$_N$ [2] within each action-duration group, which is formulated as mAP$_{acc} = \frac{1}{\sum_{i=1}^{5} M_i} \sum_{i=1}^{5}$ average-mAP$_{N,i} M_i$. It signifies the contribution of different action duration to the overall performance. We can see that for all the methods the

Table 3: **Effectiveness of Temporal Branch and Graph Branch in xGN.**

	THUMOS-14				ActivityNet-v1.3				
xGN branch	0.3	0.5	0.7	Short	0.5	0.75	0.95	Avg.	Short
Temporal	63.77	50.24	28.36	53.4	50.87	33.99	**9.09**	33.79	19.7
Graph	66.62	51.51	27.33	55.0	51.54	35.07	8.04	34.09	19.4
Temporal & Graph	**66.69**	**52.45**	**30.40**	**56.6**	**52.38**	**36.01**	8.37	**35.07**	**19.9**

shortest actions contribute the most to TAL performance. Our VSGN obviously outperforms the other methods at the shortest duration group while maintaining high ranks at longer ones.



a) THUMOS-14 b) ActivityNet-v1.3

Figure 1: **Performance at different temporal scales in terms of accumulated mAP.** For ActivityNet, We divide actions into 5 groups based on their duration in seconds: XS (0, 30], S (30, 60], M (60, 120], L (120, 180], and XL (180, inf) . For THUMOS, considering most of its actions fall into the shortest group based on the division above, we further explore the short actions by considering even finer division: XS (0, 3], S (3, 6], M (6, 12], L (12, 18], and XL (18, inf). These curves are obtained by DETAD analysis [2] on the detection results of each method. Our VSGN obviously outperforms the other methods at the shortest duration while maintaining a high rank for longer.

We also summarize mAP values at different action scales in Table 4 for clearer comparison. Our VSGN performs the best at the shortest scales, which have most instances, and reaches competitive scores for long ones.

Table 4: **Performance at different temporal scales in terms of average mAP$_N$.**

	THUMOS-14 (mAP@scale)					ActivityNet-v1.3 (mAP@scale)				
Scales (sec)	0-3	3-6	6-12	12-18	18-inf	0-30	30-60	60-120	120-180	180-inf
Instances (%)	48.4	25.3	21.7	3.4	1.1	54.4	16.3	15.9	9.4	3.9
SSN [9]	26.0	32.7	42.0	34.6	**31.0**	-	-	-	-	-
R-C3D [6]	18.9	37.3	39.4	21.3	9.4	-	-	-	-	-
BSN [4]	-	-	-	-	-	15.0	35.4	45.9	62.3	69.5
BMN [3]	-	-	-	-	-	15.2	37.1	47.4	62.4	69.4
G-TAD [7]	40.0	50.7	50.4	26.2	6.6	17.5	**43.7**	**54.5**	**67.9**	**72.2**
I.C&I.C [8]	30.9	**61.3**	**57.8**	46.4	27.5	14.8	37.3	46.6	61.4	66.9
VSGN (Ours)	**46.8**	55.6	47.3	**49.5**	25.4	**19.9**	41.7	53.0	66.9	68.4

4. Visualization of Localization Results

In Fig. 2, we visualize some examples of our localization results. In Fig. 3, we show the predicted actions compared to the ground-truth ones in an absolute time scale. We can see that our VSGN can accurately localize very short action instances as well as long ones, even when there are multiple consecutive instances in one video.

We also demonstrate cases where VSGN cannot generate precise boundaries. This also happens with other methods, when the model mistakes multiple consecutive short actions as one long action or the background is similar to the actions. These cases need further exploration for future work.

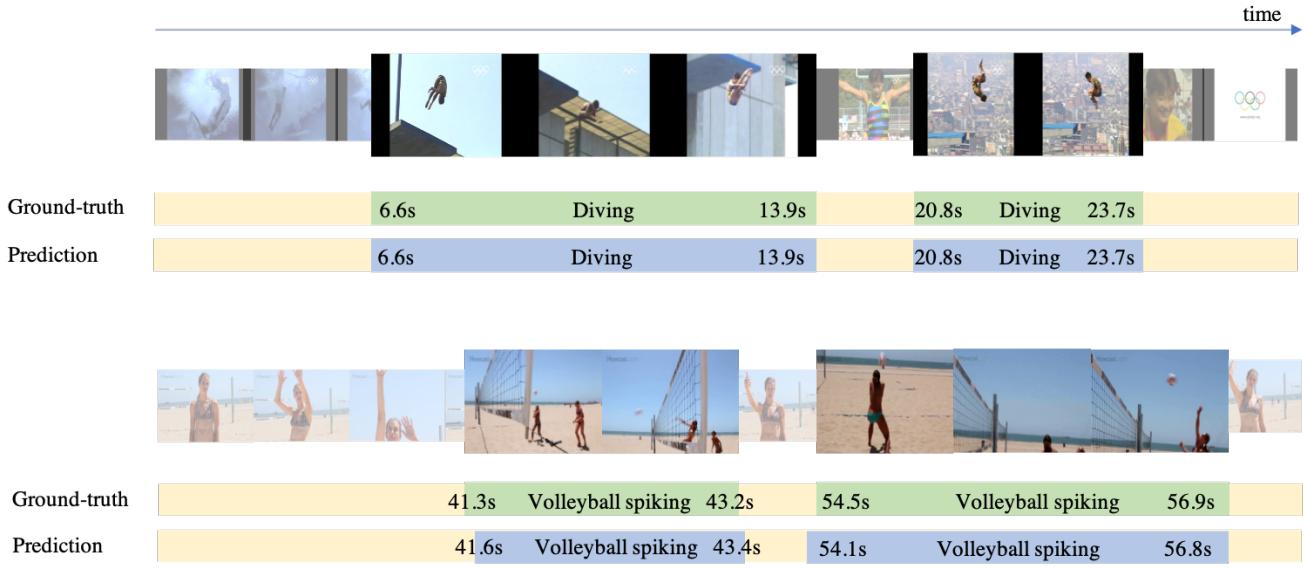


Figure 2: **Visualization of VSGN prediction results.**

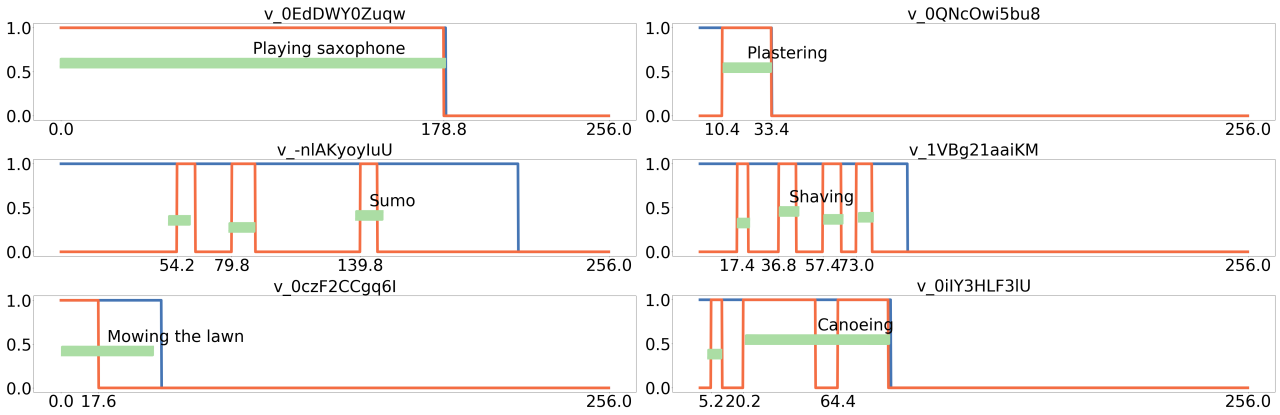


Figure 3: **Example localization results of different video duration and action duration.** In each sub-figure, the x-axis shows the time in seconds (the total input length of 1280 frames is 256.0 seconds in terms of temporal duration); the y-axis measures the confidence score. The blue curve areas are the original video; the red curves show the ground-truth actions, which have confidence scores 1.0 within their boundaries; the green segments are our predicted actions with their start/end time, confidence scores, and predicted labels. In the top two rows, we show the cases when our VSGN can successfully detect the actions with accurate boundaries, including very short actions such as those in the second row. In the third row, we also show the cases where VSGN fails to localize the actions.

References

[1] Humam Alwassel, Silvio Giancola, and Bernard Ghanem. Tsp: Temporally-sensitive pretraining of video encoders for localization tasks. *arXiv preprint arXiv:2011.11479*, 2020. 1

[2] Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem. Diagnosing error in temporal action detectors. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[3] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of The IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[4] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[5] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[6] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, page 5783–5792, 2017. 2

[7] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10156–10165, 2020. 2

[8] Peisen Zhao, Lingxi Xie, C. Ju, Y. Zhang, Yanfeng Wang, and Q. Tian. Bottom-up temporal action localization with mutual regularization. In *Proceedings of the European conference on computer vision (ECCV)*, 2020. 2

[9] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2914–2923, 2017. 2