Supplementary Material of Weakly Supervised Temporal Action Localization through Contrast based Evaluation Networks

Ziyi Liu[1] Le Wang[1*] Qilin Zhang[2] Zhanning Gao[3] Zhenxing Niu[3] Nanning Zheng[1] Gang Hua[4]
[1]Xi'an Jiaotong University [2]HERE Technologies [3]Alibaba Group [4]Wormpex AI Research
liuziyi@stu.xjtu.edu.cn {lewang, nnzheng}@mail.xjtu.edu.cn
{samqzhang, zhanninggao, zhenxingniu, ganghua}@gmail.com

A. Supplementary Material

To further validate the superiority of our action localization over thresholding-based counterparts, we compare CleanNet with UntrimmedNet [1] in terms of the per-class localization performances. Besides, more qualitative examples are provided to demonstrate CleanNet can ensure the completeness of action proposals.

A.1. Numerical Comparison with UntrimmedNet

As presented in Table 1, we compare CleanNet with UntrimmedNet [1] in terms of the TAL mAP (%) for each class under different IoU thresholds on THUMOS14 test set. Thanks to the shared common backbone network of both CleanNet and UntrimmedNet [1], it is possible to isolate the effects of the action localization versus its thresholding-based counterpart. Overall, CleanNet significantly outperforms UntrimmedNet [1] under all IoU threshold settings with most action classes, with a few exceptions in classes such as *BasketballDunk*, *CricketBowling* and *SoccerPenalty*, where the action boundaries are intuitively ambiguous, as shown in Figure 1. It is difficult to distinguish the preparatory and follow-up phases from the real action phase with only video-level categorical labels available. We speculate that it might be necessary to incorporate temporal supervision (*i.e.* full supervision) to handle these challenging action classes.

A.2. Qualitative Comparison with UntrimmedNet

We show additional qualitative examples on THUMOS14 in Figure 2 to qualitatively compare CleanNet and UntrimmedNet [1].

Some challenging cases are illustrated in Fig. 2 with false negative error (Fig. 2a, *i.e.*, missing action instances) and false positive error (Fig. 2b, *i.e.*, producing spurious action instances). Such errors are more prominent with UntrimmedNet [1], which could be caused by the difficulty of adjusting proper localization thresholds in UntrimmedNet. On the contrary, CleanNet has an action proposal evaluator to facilitate proposal selection without relying on thresholding, which could be the justification for its better performances.

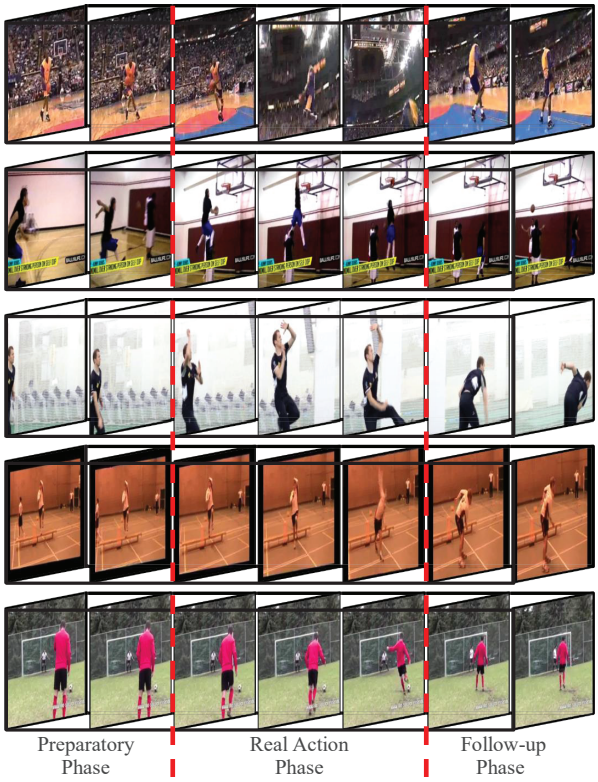


Figure 1: Action instances with ambiguous boundaries, such as the run-up as the preparatory phase. With the immediately preceding preparatory phase and the subsequent follow-up phase, it is challenging for algorithms to precisely locate the real action phase, especially with weakly supervised methods. The above five samples demonstrate such cases, where our proposed method misclassifies these transitional phases as part of the real action instance. The dashed red lines indicate the real temporal action boundaries provided by the groundtruth.

Over-segmentation (*i.e.*, breaking one action instance

*Corresponding author.

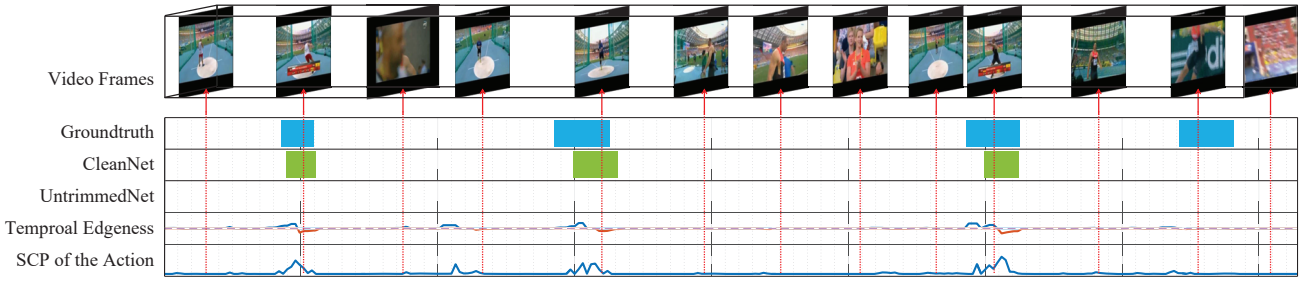
into multiple ones) and under-segmentation (*i.e.*, merging multiple instances into one segment) are generally more severe with UntrimmedNet [1], as illustrated in Fig. 2c and Fig. 2d, respectively. We speculate that such thresholding-based method accounts for the content of action proposals only without specific treatment of proposal boundaries and context information, thus it is less effective than CleanNet at ensuring the completeness of action proposals.

References

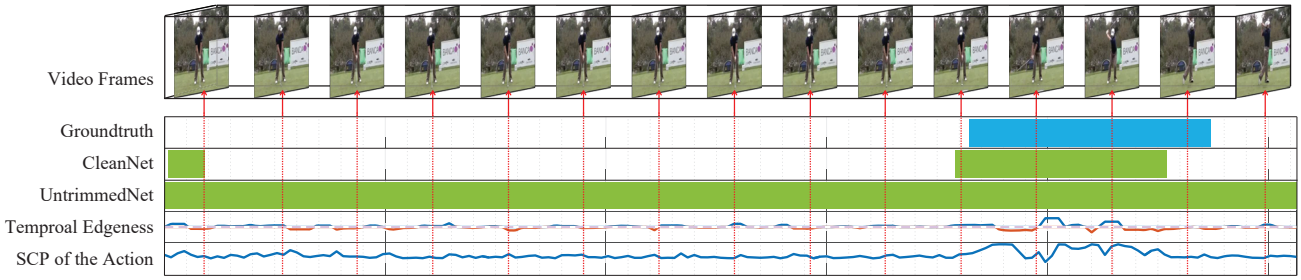
[1] L. Wang, Y. Xiong, D. Lin, and L. Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, pages 4325–4334, 2017. 1, 2, 3, 4

Table 1: TAL mAP (%) for each class under different IoU thresholds on THUMOS14 test set. UntrimmedNet [1] is selected to represent thresholding-based methods.

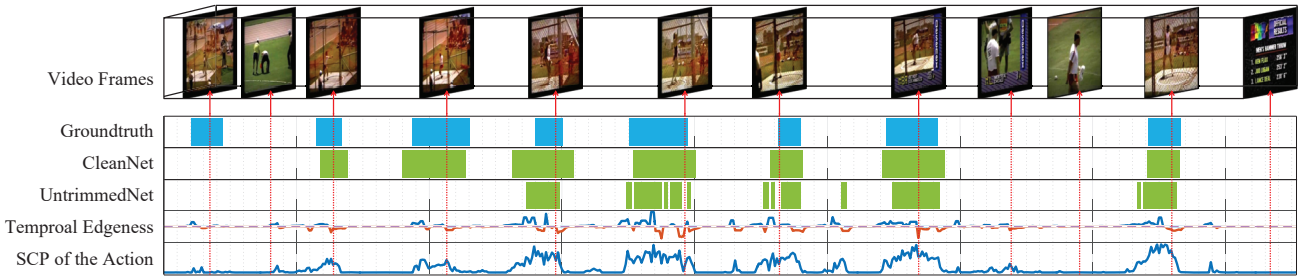
Class	mAP(%)@IoU									
	0.3		0.4		0.5		0.6		0.7	
	[1]	CleanNet	[1]	CleanNet	[1]	CleanNet	[1]	CleanNet	[1]	CleanNet
BaseballPitch	25.1	**30.0**	14.0	**21.9**	6.0	**16.1**	1.0	**2.9**	**1.0**	0.2
BasketballDunk	6.5	**8.5**	2.6	**2.8**	**1.6**	0.8	**0.6**	0.2	**0.1**	0.1
Billiards	3.3	**5.8**	1.2	**3.2**	0.6	**1.1**	0.3	**0.9**	0.1	**0.5**
CleanAndJerk	34.5	**51.5**	27.6	**45.8**	15.5	**35.5**	6.1	**18.9**	1.9	**7.1**
CliffDiving	47.4	**62.4**	40.0	**55.9**	28.2	**44.7**	17.2	**28.3**	10.3	**17.5**
CricketBowling	**17.4**	15.1	**9.4**	7.8	**4.7**	3.1	**3.0**	1.6	**1.4**	0.3
CricketShot	**12.0**	9.7	5.5	**5.7**	1.1	**2.2**	0.5	**0.7**	0.1	**0.3**
Diving	20.5	**45.0**	11.1	**37.5**	7.3	**27.2**	3.8	**15.8**	1.2	**7.2**
FrisbeeCatch	7.0	**11.8**	4.3	**8.2**	0.6	**5.6**	0.1	**0.7**	0.0	**0.1**
GolfSwing	34.1	**55.5**	27.9	**47.9**	23.6	**44.0**	16.8	**24.2**	7.3	**23.4**
HammerThrow	**64.3**	63.8	57.9	**58.9**	43.7	**48.2**	31.4	**36.4**	**23.8**	18.5
HighJump	29.9	**35.6**	19.8	**27.2**	13.6	**20.8**	6.7	**14.2**	2.6	**6.1**
JavelinThrow	45.1	**51.0**	38.6	**46.5**	24.4	**36.8**	12.8	**21.8**	5.5	**11.2**
LongJump	65.2	**67.0**	54.4	**65.3**	45.4	**59.7**	20.7	**36.3**	4.2	**14.1**
PoleVault	**64.8**	57.4	**50.3**	46.9	29.1	**32.6**	9.4	**16.4**	2.3	**5.3**
Shotput	39.7	**46.3**	25.8	**37.6**	17.6	**24.3**	10.6	**15.6**	**8.5**	6.8
SoccerPenalty	13.6	**18.2**	10.1	**14.5**	4.2	**8.8**	**2.1**	1.9	**2.1**	1.9
TennisSwing	0.8	**8.5**	0.4	**3.3**	0.1	**1.7**	0.1	**0.9**	0.0	**0.1**
ThrowDiscus	50.2	**63.3**	42.4	**58.4**	34.7	**50.1**	20.5	**34.1**	10.7	**20.0**
VolleyballSpiking	15.4	**26.8**	12.2	**20.1**	7.2	**12.4**	2.4	**5.5**	0.4	**2.1**
Avg.	29.8	**36.7**	22.8	**30.8**	15.4	**23.8**	8.3	**13.9**	4.2	**7.1**



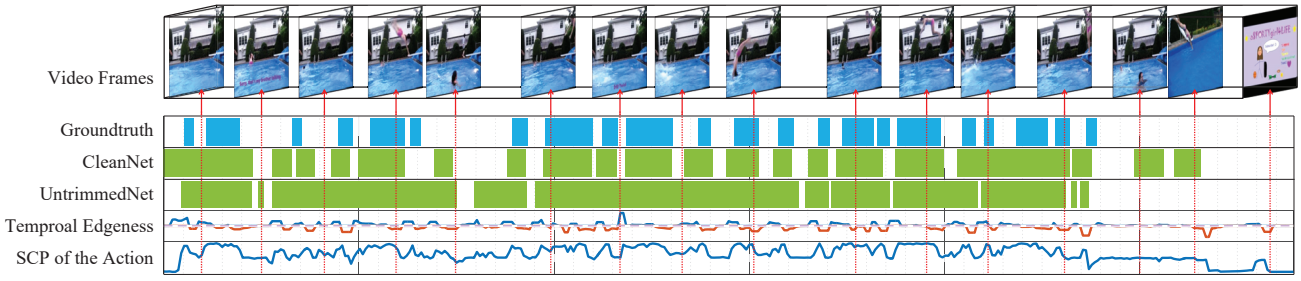
(a) An example from action *ThrowDiscus*



(b) An example from action *GolfSwing*



(c) An example from action *HammerThrow*



(d) An example from action *Diving*

Figure 2: Qualitative TAL examples between CleanNet and UntrimmedNet [1] on THUMOS14 test set. The ground truth temporal locations and predicted ones are illustrated with blue and green bars, respectively. Both the corresponding temporal edgeness and snippet-level classification classification predictions of the action are included. Specifically, for the edgeness score, a two-tone color scheme is used, with blue and orange colors representing positive and negative values, respectively. (a) An example video with false negative error (CleanNet missing the last action instance and UntrimmedNet missing all). (b) An example video with false postive error. (c) An example of over-segmentation (*i.e.*, breaking one instance into multiple segments). (d) An example of under-segmentation (*i.e.*, merging multiple instances into one segment).