Supplementary: Post-Processing Temporal Action Detection

Sauradip Nag[1,2] Xiatian Zhu[1,3] Yi-Zhe Song[1,2] Tao Xiang[1,2]
[1] CVSSP, University of Surrey, UK [2] iFlyTek-Surrey Joint Research Center on Artificial Intelligence, UK
[3] Surrey Institute for People-Centred Artificial Intelligence, UK
{s.nag,xiatian.zhu,y.song,t.xiang}@surrey.ac.uk

1. Additional Results

Due to the nature of being a plug-and-play module, our proposed GAP can be readily used with any Temporal Action Detection (TAD) frameworks like [6–10] irrespective of the supervision setting.

1.1. GAP in Semi-Supervised Setting

We integrate the proposed GAP to state-of-the-art semi-supervised TAD approaches. In this experiment, we test on 10% unlabeled data setting on ActivityNet dataset using two representative semi-supervised approaches: SSTAP [12], and SPOT [8]. Since SPOT [8] has 2-stage training (pre-training then finetune), we use GAP once in pre-training and once during inference. It is to be noted that since When GAP is used for unsupervised pre-training, we apply the modulation on the pseudo-ground truth. From the results in Table 1 it is evident that GAP indeed brings improvement of 0.2∼0.4 % in avg mAP when used during inference. This indicates that in case of few-labeled data, the detection is inferior which can be improved to some extent using GAP. When used during training in SPOT [8], it shows further improved performance of 0.7% indicating that quantization error can be curbed during pre-training time.

Table 1. Effect of our GAP on semi-supervised methods on ActivityNet dataset using 10% labeled data setting.

Method	**mAP**			
	0.5	**0.75**	**0.95**	**Avg**
SSTAP [12]	40.7	29.6	9.0	28.2
SSTAP [12] + **GAP**	**41.5**	**30.2**	**9.1**	**28.6**
SPOT [8]	49.9	31.1	8.3	32.1
SPOT [8] + **GAP**	Training			
	52.8	**31.6**	**8.8**	**32.8**
	Inference			
	52.3	**31.4**	**8.5**	**32.3**

1.2. GAP in Weakly-Supervised Setting

We evaluate the effect of our GAP with top performing weakly-supervised TAD methods including popular approaches like DELU [1], CoLA [13] and ASL [5]. This test is done on THUMOS14 dataset. Similar to supervised TAD approaches, as shown in Table 2 weakly-supervised methods greatly benefit from GAP during post-processing.

Table 2. Effect of our GAP on weakly-supervised methods on THUMOS dataset.

Model	**mAP**					
	0.3	**0.4**	**0.5**	**0.6**	**0.7**	**Avg**
ASL [5]	51.8	-	31.1	-	11.4	32.2
ASL [5] + **GAP**	**53.0**	**-**	**31.7**	**-**	**11.5**	**32.4**
CoLA [13]	51.5	41.9	32.2	22.0	13.1	40.9
CoLA [13] + **GAP**	**51.8**	**42.2**	**32.4**	**22.2**	**13.2**	**41.0**
TS-PCA [4]	52.4	43.5	34.6	23.7	12.6	-
TS-PCA [4] + **GAP**	**52.9**	**44.0**	**34.9**	**24.0**	**12.8**	**-**
CO2-Net [3]	54.5	45.7	38.3	26.4	13.4	-
CO2-Net [3] + **GAP**	**54.9**	**46.0**	**38.8**	**27.1**	**14.0**	**-**
ASM-Loc [2]	57.1	46.8	36.6	25.2	13.4	45.1
ASM-Loc [2] + **GAP**	**58.1**	**47.5**	**37.1**	**25.6**	**13.8**	**45.5**
DELU [1]	56.5	47.7	40.5	27.2	15.3	46.4
DELU [1] + **GAP**	**57.0**	**48.1**	**40.9**	**27.6**	**15.5**	**46.6**

1.3. GAP in Few-Shot Setting

Our GAP can also be used in few-shot temporal action detection approaches. For this experiment, we evaluate our GAP using a recent few-shot TAD approach QAT [10]. We report 1/5-shot experiment result on ActivityNet. From Table 3, it is evident that GAP brings largest avg mAP improvement of 0.6% in 1-shot setting, indicating that the quantization error is high when there are very few labeled samples. This error reduces as we increase the number of shots, as expected.

1.4. GAP in Zero-Shot Setting

Similar to few-shot approaches, we can use GAP in zero-shot TAD setting. We consider a very recent zero-shot method STALE [9] and a 2-stage baseline (similar to *Baseline-I* in [9]) on a challenging 50% seen data split on Activitynet dataset. Since the 2-stage baseline includes

Table 3. Effect of our GAP on few-shot action detection methods on ActivityNet dataset in 1-way multi-instance setting.

Shot	Models	mAP			
		0.5	0.7	0.9	Avg
1	QAT [10]	44.9	29.2	11.2	25.9
	QAT [10] + **GAP**	**45.8**	**30.0**	**11.8**	**26.5**
5	QAT [10]	51.8	32 .6	11.9	30.2
	QAT [10] + **GAP**	**52.2**	**32.9**	**12.1**	**30.4**

proposal-generation as an intermediate step, we can apply GAP during training of the CLIP [11] pre-trained classifier in the second stage. On the other hand, STALE is a single-stage approach hence we use GAP in the localization head during post-processing. From Table 4 we observe a higher improvement using GAP for the baseline, indicating 2-stage approaches have localization-error propagation which can be partially resolved by using GAP. This reveals another advantage of our model design.

Table 4. Effect of our GAP on zero-shot action detection methods on ActivityNet dataset in 50% seen data setting. † indicates GAP is used during training.

Models	mAP			
	0.5	0.75	0.95	Avg
Baseline	28.0	16.4	1.2	16.0
Baseline† + **GAP**	**28.7**	**16.8**	**1.7**	**16.5**
Baseline + **GAP**	**28.2**	**16.6**	**1.3**	**16.2**
STALE [9]	32.1	20.7	5.9	20.5
STALE [9] + **GAP**	**32.4**	**21.1**	**6.2**	**20.8**

2. Summary

From the experiments we have performed so far, we draw several conclusions regarding the usefulness of our proposed GAP. Besides being effective for fully-supervised setting (Table **??**), our GAP is also effective when there are **(i)** a large number of unlabeled training samples (refer to Table 1), **(ii)** unavailability of fine-grained annotation (refer to Table 2), **(iii)** only a few labeled samples (refer to Table 3), and **(iv)** no labeled samples (refer to Table 4). In all the above mentioned cases, the quantization error is more profound due to the design choice or problem setting which can be greatly reduced by using our GAP. This verifies the generic usefulness of our method across a variety of settings.

References

[1] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *European Conference on Computer Vision*, pages 192–208. Springer, 2022. 1

[2] Bo He, Xitong Yang, Le Kang, Zhiyu Cheng, Xin Zhou, and Abhinav Shrivastava. Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13925–13935, 2022. 1

[3] Fa-Ting Hong, Jia-Chang Feng, Dan Xu, Ying Shan, and Wei-Shi Zheng. Cross-modal consensus network for weakly supervised temporal action localization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1591–1599, 2021. 1

[4] Yuan Liu, Jingyuan Chen, Zhenfang Chen, Bing Deng, Jianqiang Huang, and Hanwang Zhang. The blessings of unlabeled background in untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6176–6185, 2021. 1

[5] Junwei Ma, Satya Krishna Gorti, Maksims Volkovs, and Guangwei Yu. Weakly supervised action selection learning in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7587–7596, 2021. 1

[6] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Temporal action localization with global segmentation mask transformers. 2021. 1

[7] Sauradip Nag, Xiatian Zhu, Yi-zhe Song, and Tao Xiang. Proposal-free temporal action detection via global segmentation mask learning. In *ECCV*, 2022. 1

[8] Sauradip Nag, Xiatian Zhu, Yi-zhe Song, and Tao Xiang. Semi-supervised temporal action detection with proposal-free masking. In *ECCV*, 2022. 1

[9] Sauradip Nag, Xiatian Zhu, Yi-zhe Song, and Tao Xiang. Zero-shot temporal action detection via vision-language prompting. In *ECCV*, 2022. 1, 2

[10] Sauradip Nag, Xiatian Zhu, and Tao Xiang. Few-shot temporal action localization with query adaptive transformer. *arXiv preprint arXiv:2110.10552*, 2021. 1, 2

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[12] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Changxin Gao, and Nong Sang. Self-supervised learning for semi-supervised temporal action proposal. In *CVPR*, pages 1905–1914, 2021. 1

[13] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16010–16019, 2021. 1