Appendix to "A Simple and Efficient Pipeline to Build an End-to-End Spatial-Temporal Action Detector"

This appendix provides further details of the main paper. We present extra experiments and visualizations. (1) Our SE-STAD can cooperate with stronger action classification heads; (2) The ablation of feature pyramid architecture; (3) Additional visualizations of TLA.

1. Cooperating with stronger action classification heads

In this paper, we dive deep into building an end-to-end spatial-temporal action detector with minimum efforts. Contrary to [2, 6] which highly rely on extra attention mechanisms, no additional attention module is introduced in our proposed SE-STAD. In order to show that our SE-STAD could also benefit from attention modules, we adopt a simple ACRN head [8] as the action classification head. Table 1 shows the experiment results. We could find that the basic SE-STAD could get 1.5% gains from the introduced simple ACRN head. Such results demonstrate that we could further boost the performance with other fancy attention modules, such as [8, 7].

Method	Attention Module	val mAP
SE-STAD	None	25.5
SE-STAD	ACRN [8]	27.0

Table 1. **Ablation study on classification head**. We try to use ACRN [8] head as the classification head. Experiments are performed with SlowFast R50 backbone.

2. Cooperating with other backbones

For fair comparisons, we only perform experiments with SlowFast backbone in the main paper. To show the generalizability of our model, we further perform experiments with I3D [1] network. As shown in Table 2, we could reach 23.0 mAP with the I3D backbone which is still far better than AVA baseline [5] and result obtained with more pretraining data and heavy head [4].

3. Ablation of feature pyramid architecture

As shown in Fig. 2 in this paper, we build the feature pyramid (P3-P5) on top of features of keyframes from Res3 and Res4 layers. We adopt such an architecture after considering the balance of computation complexity and performance. We do extra experiments about adopting different feature pyramid architectures to perform actor localization. Experiment results are shown in Table 3. These results show that the architecture of the feature pyramid does not play the most essential role in SE-STAD. Although building P2-5 on Res2-4 could boost the performance from 25.5 to 26.0, it will introduce around 40% computation additionally. Hence, we select to build P3-P5 based on Res3 and Res4 to make a trade-off between the detector performance and computation complexity.

Method	Backbone	Pretrain	val mAP
AVA baseline [5]	I3D [8]	K400	15.8
Better baseline* [4]	I3D [8]	K600	21.9
SE-STAD	I3D [8]	K400	23.0
SE-STAD	SlowFast R50 [3]	K400	25.0

Table 2. **Ablation study on different backbones**. We try to use I3D as the backbone of SE-STAD. No additional attention module is introduced. * means the "Better baseline" method use heave I3D blocks to perform action classification.

Method	Source	Target	val mAP	GFLOPs
SE-STAD	Res3-4	P3-5	25.5	111.3
SE-STAD	Res3-5	P3-5	25.2	113.4
SE-STAD	Res2-4	P2-5	26.0	152.8

Table 3. **Ablation study on the architecture of feature pyramid**. We try different feature pyramid architectures with SlowFast R50 backbone.

4. FLOPs Analysis

We analyse the component of different spatial-temporal action detectors, including proposal-based methods and our SE-STAD. The analysis results are in Fig. 2. From Fig. 2, we can easily observe that: FCOS head only occupies around 12% percent of the whole SE-STAD, and backbone action classification network has the majority computational complexity in SE-STAD. In contrast, the majority computational burden of two-stage detectors lies in the person detector part, which is redundant.

Figure 1. **Visualization of results generated by TLA.** Top row: the former neighbour keyframe with ground-truth annotations. Medium row: the later neighbour keyframe with ground-truth annotations. Bottom row: the frame labeled by TLA. Last column: a failure case. The figure is best viewed when zoomed in. Numbers of person proposals indicate the entity ids in keyframes with ground-truth annotations and assigned entity ids in frames labeled by TLA.

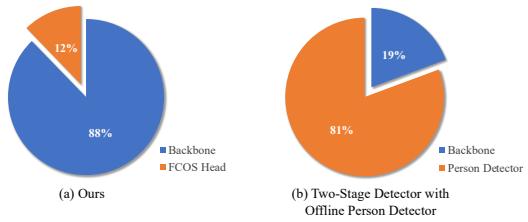

Figure 2. FLOPs pie chart with different parts in the spatial-temporal action detector. The left figure is our SE-STAD and the right figure is the proposal-based SlowFast. FLOPs of action classification head is ignored as the computation complexity of this part is too small. Both methods use the same SlowFast R50 backbone.

5. Visualization of TLA

In Sec. 3, we propose a novel labeling strategy, i.e., the temporal label assignment (TLA), to better utilize every possible piece of information in sparse annotated spatial-temporal action detection datasets. TLA which utilizes the temporal restriction could provide more clear temporal action boundaries and fine-grained information to the detector. In Fig. 1, we visualize some results produced by TLA. The visualization results illustrate that TLA produces relatively reliable pseudo labels successfully on unlabeled frames. In addition, we find that there are often missing labels in provided ground-truth annotations which deteriorate the performance of the spatial-temporal action detector. Whereas, thanks to the promising actor localization ability of SE-STAD, TLA could detect most of the actors and assign dependable labels.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6299–6308, 2017.

[2] Shoufa Chen, Peize Sun, Enze Xie, Chongjian Ge, Jiannan Wu, Lan Ma, Jiajun Shen, and Ping Luo. Watch only once: An end-to-end video action detection framework. In *Int. Conf. Comput. Vis.*, pages 8178–8187, 2021.

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *Int. Conf. Comput. Vis.*, pages 6202–6211, 2019.

[4] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. A better baseline for AVA. *arXiv preprint arXiv:1807.10066*, 2018.

[5] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6047–6056, 2018.

[6] Shentong Mo, Jingfei Xia, Xiaoqing Tan, and Bhiksha Raj. Point3D: tracking actions as moving points with 3d cnns. In *Brit. Mach. Vis. Conf.*, pages 1–14, 2021.

[7] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 464–474, 2021.

[8] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In *ECCV*, pages 318–334, 2018.