STMixer: A One-Stage Sparse Action Detector Supplementary Material

Tao Wu[1,*] Mengqi Cao[1,*] Ziteng Gao[1] Gangshan Wu[1] Limin Wang[1,2,✉]
[1] State Key Laboratory for Novel Software Technology, Nanjing University [2] Shanghai AI Lab
{wt,mg20370004}@smail.nju.edu.cn, gzt@outlook.com, {gswu, lmwang}@nju.edu.cn

1. More Experimental Results

In this section, we provide more detailed experimental results and analysis to investigate on which categories our STMixer shows more significant performance improvements, and on which categories the long-term classifier has a greater impact.

1.1. AP on Each Action Class Comparison

We provide detailed comparisons of the performance of STMixer and former state-of-the-art TubeR [10] on each action class of AVA v2.2 [4] in Figure 1. For a fair comparison, both STMixer and TubeR models use the CSN-152 backbone [8] and do not use long-term features. Out of all 60 classes, our method achieves higher AP on 47 classes, which makes the overall detection mAP of our STMixer higher by 1.7 than TubeR. We observe significant performance gaps on some interaction-related classes, such as interactions with objects (drive (*e.g.* a car, a truck) +12.9, shoot +12.0), and interactions with other people (sing to (*e.g.* self, a person, a group) +9.9, take (an object) from (a person) +3.8), which indicates our STMixer is more capable of modeling the relationship between the action performer and the surrounding objects and people.

1.2. Impact of Long-term Classifier

We provide the performance of STMixer with a short-term classifier or a long-term classifier on each action class of AVA v2.2 in Figure 2 to show the benefit of the long-term classifier. When using a long-term classifier, STMixer achieves better performance on the vast majority of action classes than using a short-term classifier, which demonstrates the importance of long-term information for action instance recognition. The experimental results also demonstrate that the action queries in our STMixer contain rich spatiotemporal information, and our design of long-term query bank and long-term query operation is effective. For action instances of some classes, the action performer sometimes interacts with objects or people appearing in other temporal segments. For example, for an action instance of "sing to (*e.g.* self, a person, a group)", the singer and the listener are often in the different temporal segments of the video. We observe remarkable improvements on these classes (sing to (*e.g.* self, a person, a group) +5.0). STMixer with a long-term classifier can attend action queries to the long-term query bank for information of the listener. The long-term query bank and sampled features of the current video clip are complementary to each other and are both important for action detection [9]. This paper mainly focuses on the exploration of adaptive feature sampling from the feature space of the current video clip and adaptively feature mixing to enhance the representations, yet our simple design of long-term query back also yields good performance.

1.3. Inference Speed Comparison

Method	Extra Dect.	Input	Backbone	GFLOPs	mAP		FPS
					v2.1	v2.2	
SlowFast [3]	✓	32× 2	SF-R101-NL	365	28.2	29.0	5.8
WOO [2]	✗	32× 2	SF-R101-NL	252	28.0	28.3	6.9
TubeR [10]	✗	32× 2	CSN-152	120	29.7	31.1	12.3
STMixer	✗	32× 2	SF-R101-NL	135	29.8	30.1	11.6
STMixer	✗	32× 2	CSN-152	126	31.7	32.8	11.9

Table 1. **Inference speed comparison on AVA dataset.** ✓of column "Extra Dect." denotes an extra human detector Faster RCNN-R101-FPN [6] is used. For a fair comparison, the resolution of input frames is set to 256×256 for all models, and all models are tested on a GeForce RTX 3090 GPU.

We compare the inference speed of our STMixer with former state-of-the-art methods in Table 1. Methods like AIA [7] and ACARN [5] follow the typical two-stage framework proposed by SlowFast [3] but use a more complicated classification head for context modeling, so their complexity is higher than SlowFast. As AIA and ACARN do not report their complexity in their paper and these data are not available for us, we consider SlowFast as a lower bound of complexity for these methods here. For a fair comparison, long-term features are not used in all methods. As shown in Table 1, because an extra human detection

*: Equal contribution. ✉: Corresponding author.

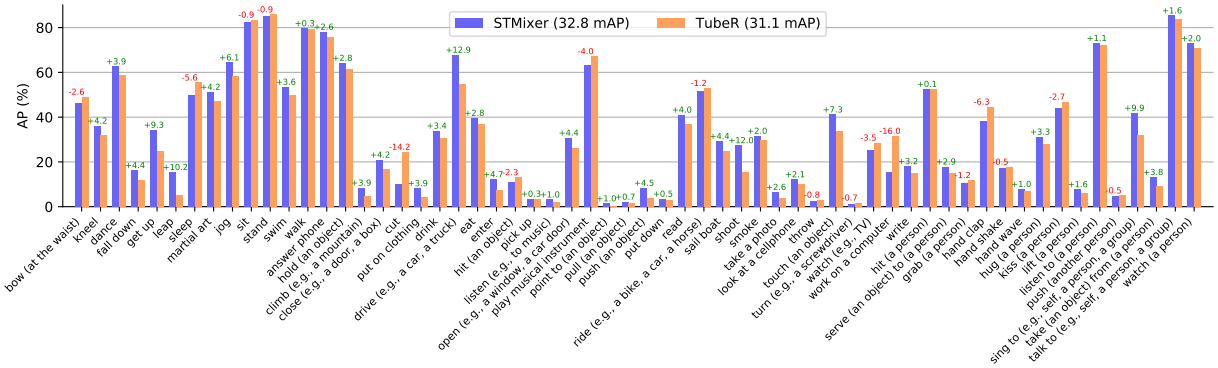

Figure 1. **AP of STMixer and TubeR on each action class of AVA v2.2.** We use the STMixer and TubeR models with the CSN-152 backbone for comparison. Both models do not use long-term features. When our STMixer has a higher AP, the difference is marked in green, otherwise, it is marked in red. Our STMixer has a higher AP on most action classes than TubeR.

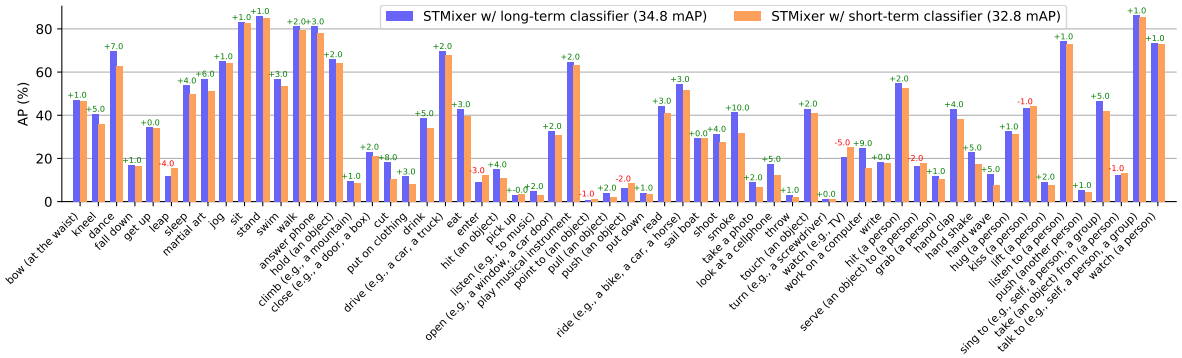

Figure 2. **AP of STMixer with a long-term or short-term classifier on each class of AVA v2.2.** When STMixer with a long-term classifier has a higher AP, the difference is marked in green, otherwise, it is marked in red.

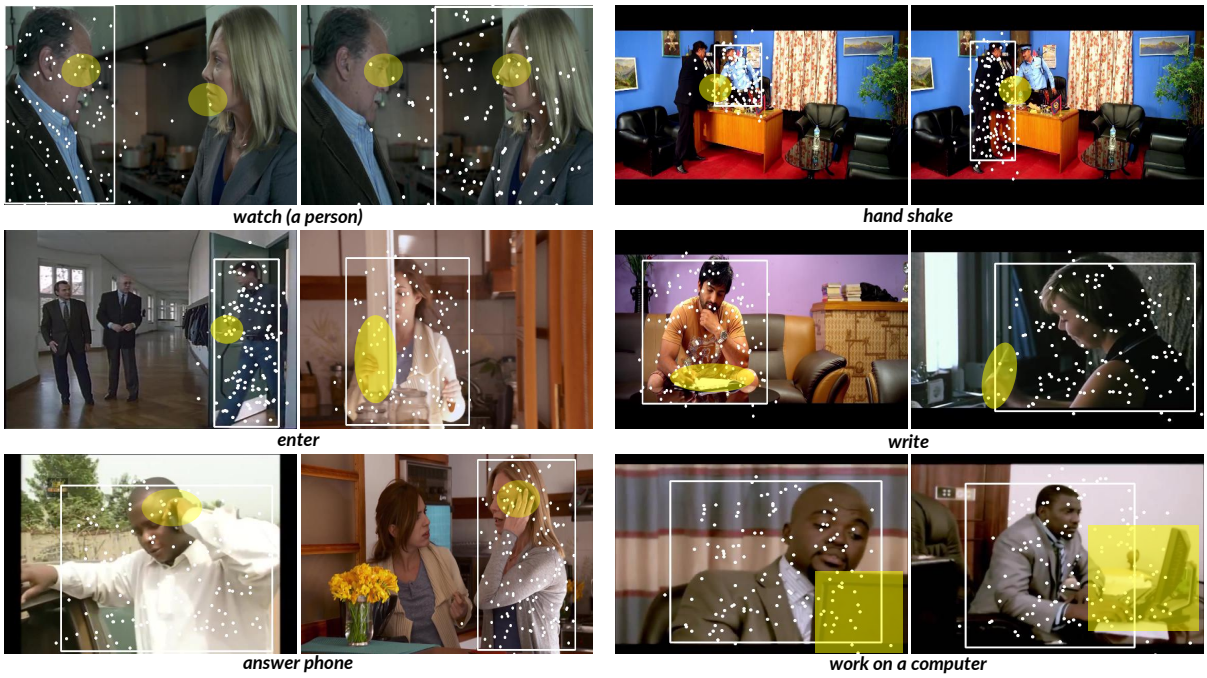

Figure 3. **Visualizations of sampled feature points for some action instances.** We show the sampled feature points of the last ASAM module. The yellow highlighted areas are considered to provide semantically relevant visual cues.

process is needed, two-stage methods like SlowFast have much lower inference speeds. Although training and inference are performed in an end-to-end manner, WOO [2] still adopts a two-stage pipeline while TubeR and STMixer perform actor localization and action localization in one stage. This simplified pipeline makes inference speeds of TubeR and STMixer much higher. Compared to TubeR, our STMixer has 2.0 and 1.7 points higher mAP on AVA v2.1 and v2.2 respectively, while the inference speed of STMixer is comparable to TubeR (11.9FPS versus 12.3FPS). The training overhead of STMixer is also smaller than TubeR. STMixer converges within 10 epochs, while TubeR needs to be trained for 20 epochs despite using DETR [1] initialization. The adaptive sampling and adaptive mixing mechanism proposed in our STMixer makes it easier to cast the action queries to action instances.

2. More Visualizations

As presented in the main paper, our proposed STMixer is a query-based framework for video action detection, which adaptively samples features from the spatiotemporal feature space without the restriction of human bounding boxes. In Figure 3, we provide more visualizations of sampled feature points for action instances. As shown in Figure 3, some of the sampling points go out of the human boxes and spread to other semantically related areas, which demonstrates the ability of our method to mine discriminative features from both the action performer itself and the surrounding context.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 3

[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 *ICCV*, pages 8178–8187, 2021. 1, 3

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 1

[4] 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 *CVPR*, pages 6047–6056, 2018. 1

[5] 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 *CVPR*, pages 464–474, 2021. 1

[6] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 1

[7] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *ECCV*, pages 71–87. Springer, 2020. 1

[8] Du Tran, Heng Wang, Matt Feiszli, and Lorenzo Torresani. Video classification with channel-separated convolutional networks. In *ICCV*, pages 5551–5560. IEEE, 2019. 1

[9] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, pages 284–293, 2019. 1

[10] Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, Ivan Marsic, Cees G. M. Snoek, and Joseph Tighe. Tuber: Tubelet transformer for video action detection. In *CVPR*, pages 13588–13597. IEEE, 2022. 1