ART: Actor-Related Tubelet for Detecting Complex-shaped Action Tubes SUPPLEMENTARY MATERIAL

Jiaojiao Zhao
Prime Video at Amazon
zhajiaoj@amazon.com

Frame-mAP reported. As shown in Tab. 1, although ART is primarily designed for tube detection, all its variants still achieve strong frame-mAP. With a ViT-B backbone, ART performs comparably to STAR on UCF101-24 and slightly trails STMixer on JHMDB51-21, despite STAR and STMixer being pretrained on larger datasets.

AVA [9] is not the target dataset for our work, as it lacks tube-level annotations. Following the main paper, Fig 1 illustrates the cumulative density function of the IoU for ground-truth bounding box pairs taken one second apart, plotted for the training sets of MultiSports, UCF, JHMDB, and AVA. On AVA, 90% of the box pairs have an IoU greater than 0.5, indicating that the motion in AVA is relatively small. Our ART is designed for complex-shape tubes. In Tab 2, most state-of-the-art methods rely on an offline person detector (typically Faster-RCNN) to first localize actors and then focus solely on action recognition. In contrast, our ART method operates end-to-end, simultaneously localizing actors and recognizing their actions. Using only Kinetics-400 pre-trained weights and without incorporating an additional detector, the pure transformer version of ART achieves 40.1 mAP. Although ART is specifically designed for complex-shaped tube detection, its architecture does not compromise performance on actions with small motion trajectories.

Temporal compensation. Tab 3 reflects the effectiveness of the temporal compensation module. Incorporating temporal information into actor-related tubelet queries accounts for changes in actors' poses and shapes over time, resulting in a 0.5 improvement in video-mAP@IoU=0.5 on UCF101-24.

More visualizations. We present additional action tube detection results on the MultiSports, UCF101-24, and JHMDB51-21 datasets in Fig 2. MultiSports features complex-shaped action tubes, including challenges such as camera motion, deformable shapes, and multiple actors as shown in Fig 2(a). UCF101-24 contains similarly complicated scenarios, such as intertwined actors and multiple actor interactions, as illustrated in Fig 2(b). ART effectively handles these intricate action tubes by leveraging actor information to construct tubelets. As noted in the main paper, JHMDB51-21 (Fig 2(c)) consists of simpler cases, characterized by short-length tubes, single actors, and small motion, as shown in the figure. As expected, ART performs well on this dataset.

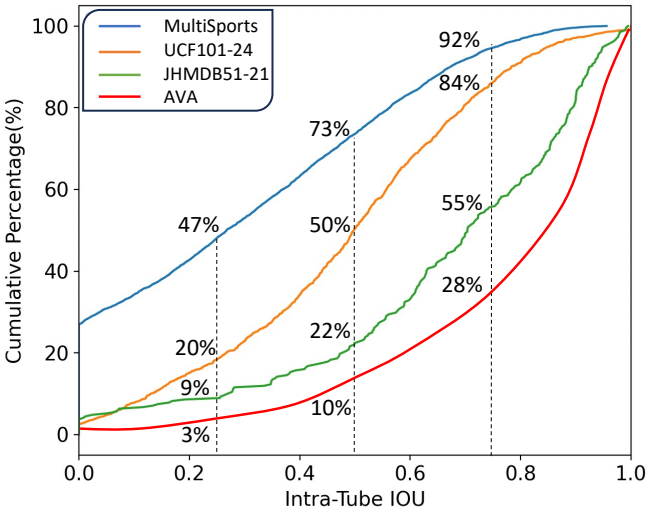

Figure 1. **Cumulative density function of intra-tube IoU** is presented for four action detection datasets: MultiSports, UCF101-24, JHMDB51-21, and AVA. Notably, only 10% of box pairs in AVA exhibit an IoU below 0.5, indicating that 90% of instances in this dataset experience small motion, with bounding boxes overlapping by more than 0.5.

Failure case. Our ART framework encounters challenges when handling extremely small actors, which complicates the extraction of actor-related information. An example of this issue is illustrated in Fig. 3. In particular, ART occasionally misses bounding boxes within a tube when actors are very tiny. We consider to apply multi-scale technology on both temporal and spatial dimensions to eliminate the issue. We will make it in the future work.

	models	pretrain	UCF101-24	JHMDB-51-21
ACT [11]	dual-model (rgb+flow)	IN1K	67.1	65.7
TacNet [17]	dual-model (rgb+flow)	-	72.1	65.5
MOC-DLA34 [13]	dual-model (rgb+flow)	IN1K→COCO	78.0	70.8
TubeR-I3D [22]	dual-model (rgb+flow)	K400	81.3	-
CFAD-I3D [12]	dual-model (rgb+flow)	K400	72.5	-
HIT-SF-R101 [5]	multi-model (person/object/keypoints)	K700	84.8	83.8
MOC-DLA34 [13]	single-model	IN1K→COCO	72.1	-
T-CNN-C3D [10]	single-model	K400	41.4	61.3
TAAD-SF101 [16]	single-model	-	81.5	-
TubeR-I3D [22]	single-model	K400	80.1	-
ART-I3D (ours)	single-model	K400	*81.4*	*76.8*
TubeR-CSN152 [22]	single-model	K400	83.2	-
ART-CSN152 (ours)	single-model	K400	*82.6*	*82.0*
STMixer-CSN [21]	single-model	IG65M	83.7	**86.7**
STAR-ViViT-B [8]	single-model	IN21K→K400	**87.3**	86.6
ART-ViT-B (ours)	single-model	K400	87.2	85.1

Table 1. **Comparison on UCF101-24 and JHMDB51-21** with frame-*mAP*@IoU=0.5. 'flow' means 'optical flow'. 'SF' denotes the slowfast network. Though ART is designed for tube detection, it achieves comparable frame-mAP with less pretraining data.

Model	**Detector**	**Backbone**	**Pre-train**	**Inference**	***f-mAP***
SlowFast [7]	F-RCNN	R101	K600	6 views	29.8
ACAR-slowfast [15]	F-RCNN	R101	K600	6 views	33.3
AIA-slowfast [18]	F-RCNN	R101	K700	18 views	32.2
X3D-XL [6]	F-RCNN	X3D-XL	K700	1 view	27.4
Unified [1]	F-RCNN	R101	K400	1 view	28.8
WOO-slowfast [3]	✗	R101	K600	1 view	28.3
TubeR-CSN [22]	✗	R152	IG65M	1 view	31.1
MViTv1-24 [4]	F-RCNN	MViT-B-24	K600	1 views	28.7
MViTv2-L, 312^2 [14]	F-RCNN	MViT-L	IN21K+K700	1 views	34.4
STMixer-CSN [21]	✗	R152	IG65M	2 views	34.8
MemViT-24 [20]	F-RCNN	MViT-B-24	K700	1 views	35.4
EVAD [2]	✗	ViT-L	K700	NA	39.7
VideoMAE [19]	F-RCNN	ViT-L	K400	NA	37.0
ART-ViT-L (ours)	✗	ViT-L	K400	1 view	38.1
VideoMAE [19]	F-RCNN	ViT-H	K400	NA	39.5
ART-ViT-H (ours)	✗	ViT-H	K400	1 view	**40.1**

Table 2. **Comparisons on AVA v2.2** validation set. Detector shows if additional detector is required. Our ART performs best without an offline person detector.

	video-mAP@0.5
w/o tc	63.7
tc	**64.2**

Table 3. **Temporal compensation** (tc) helps improve video-mAP@0.5.

References

[1] Anurag Arnab, Chen Sun, and Cordelia Schmid. Unified graph structured models for video understanding. In *ICCV*, 2021. 2

[2] Lei Chen, Zhan Tong, Yibing Song, Gangshan Wu, and Limin Wang. Efficient video action detection with token dropout and context refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[3] 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*, 2021. 2

[4] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 2

[5] Gueter Josmy Faure, Min-Hung Chen, and Shang-Hong Lai. Holistic interaction transformer network for action detection. *arXiv preprint arXiv:2210.12686*, 2022. 2

[6] Christoph Feichtenhofer. X3D: Expanding architectures for efficient video recognition. In *CVPR*, 2020. 2

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *CVPR*, 2019. 2

[8] Alexey A Gritsenko, Xuehan Xiong, Josip Djolonga, Mostafa

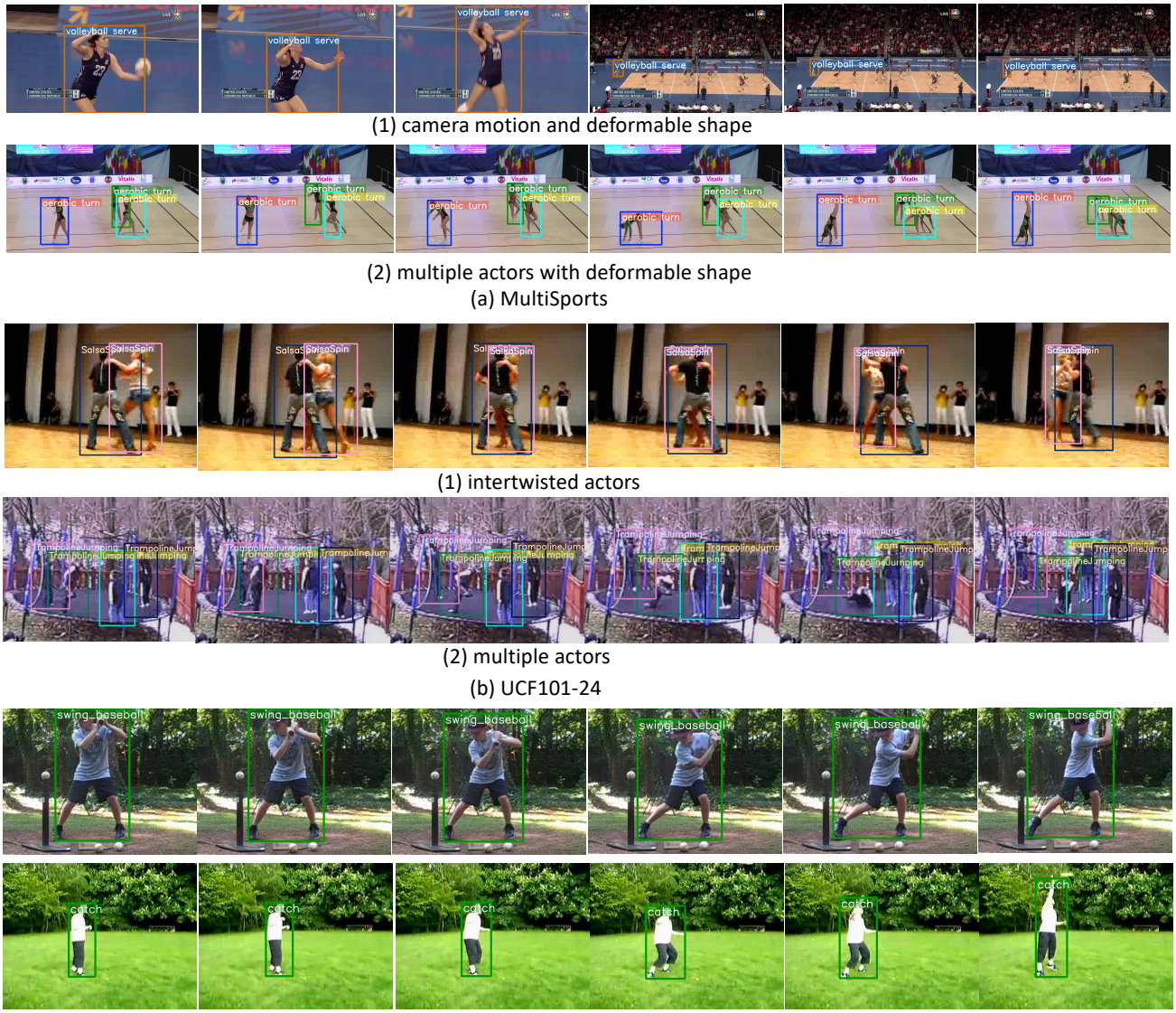

Figure 2. **Action tube visualization**.(a) Complex-shaped tubes involving camera motion and multiple actors in MultiSports. (b) Complex-shaped tubes with intertwisted actors and multiple actors in UCF101-24. (c) JHMDB51-21 has tubes characterized with single actor, small motion and short length. ART performs well for various cases.

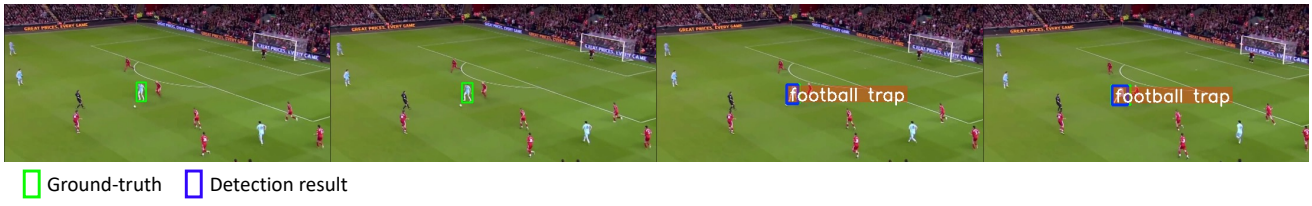

Figure 3. **Failure case**. ART faces challenges when dealing with extremely small actors, as it becomes difficult to incorporate precise actor information necessary for constructing actor-related tubelets.

Dehghani, Chen Sun, Mario Lucic, Cordelia Schmid, and Anurag Arnab. End-to-end spatio-temporal action localisation with video transformers. In *CVPR*, 2024. 2

[9] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 1

[10] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *ICCV*, 2017. 2

[11] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *ICCV*, 2017. 2

[12] Yuxi Li, Weiyao Lin, John See, Ning Xu, Shugong Xu, Ke Yan, and Cong Yang. Cfad: Coarse-to-fine action detector for spatiotemporal action localization. In *ECCV*, 2020. 2

[13] Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu. Actions as moving points. In *ECCV*, 2020. 2

[14] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022. 2

[15] 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*, 2021. 2

[16] Gurkirt Singh, Vasileios Choutas, Suman Saha, Fisher Yu, and Luc Van Gool. Spatio-temporal action detection under large motion. *arXiv preprint arXiv:2209.02250*, 2022. 2

[17] Lin Song, Shiwei Zhang, Gang Yu, and Hongbin Sun. Tacnet: Transition-aware context network for spatio-temporal action detection. In *CVPR*, 2019. 2

[18] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *ECCV*, 2020. 2

[19] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 2

[20] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *CVPR*, 2022. 2

[21] Tao Wu, Mengqi Cao, Ziteng Gao, Gangshan Wu, and Limin Wang. Stmixer: A one-stage sparse action detector. In *CVPR*, 2023. 2

[22] Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, et al. Tuber: Tubelet transformer for video action detection. In *CVPR*, 2022. 2