ICPR 2026 · Lyon, France

TubeLite

Lightweight Multi-Actor Spatio-Temporal Action Detection

Ali Soltaninezhad, Melissa Cote, Alejandro Rico Espinosa, Tunai Porto Marques, Alexandra Branzan Albu

Department of Electrical and Computer Engineering, University of Victoria, BC, Canada

Stable action tubes from a small RGB model. No optical flow, no 3D convolutions, no space-time attention.

Funded by NSERC Alliance Grants, Archipelago Marine Research, and the Digital Research Alliance of Canada.

What is spatio-temporal action detection?

1 **Action recognition** *what?*

Take a trimmed clip, give it one label.

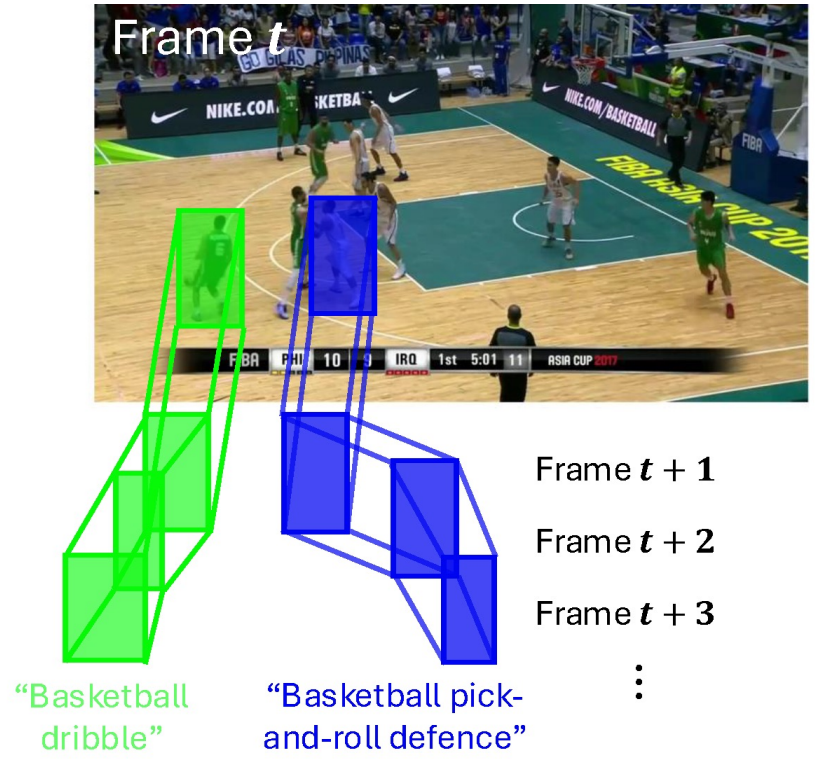
2 **Temporal action detection** *what and when?*

Find where the action starts and ends in a long video.

3 **Spatio-temporal action detection** *what, when, where, who?*

A box for every actor in every frame, plus the label and the time span.

We work on the third one.



One action tube: an actor's boxes over time, a label, and a start and end frame.

What makes it hard

1

Jittery boxes

A detector run frame by frame gives boxes that shift and sometimes drop out, especially under blur or occlusion.

2

Many actors at once

Actors get mixed up between frames, so a single tube breaks into pieces.

3

Fuzzy boundaries

Deciding the exact frame where an action starts and ends is hard, especially for short actions.

The usual answer is a bigger model

3D backbones such as SlowFast [12] and CSN-152 [35], optical flow with a second stream, or space-time transformers such as TubeR [41]. They work, but they cost 240 to 342 GFLOPs for a single clip.

We asked whether the problem is really visual capacity.

Our view, and what we built

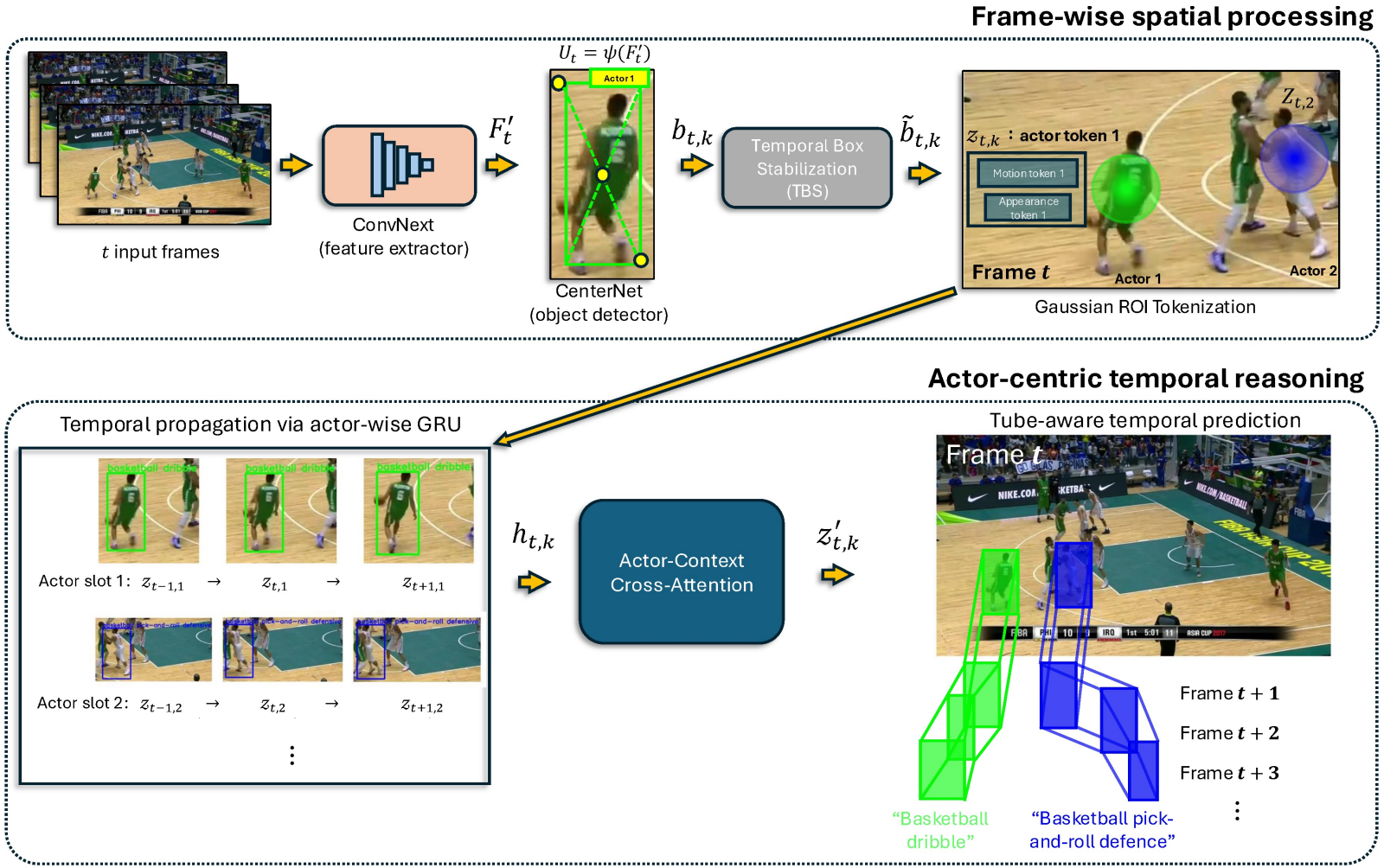
Most of the error comes from unstable tubes, not from weak visual features.

So we kept the backbone small and spent the effort on temporal structure instead.

1. **Small spatial detector** ConvNeXt-Tiny [27] with a CenterNet [10] head that shares one tower between both outputs.
2. **Gaussian actor tokens** One soft region per actor, pooled from features and from frame differences.
3. **A GRU per actor** Temporal propagation that is linear in clip length.
4. **One context token per frame** Actors look at a single learned summary.
5. **Tube-aware supervision** Losses on the boundaries and on how much a box in tube can change between frames.

No optical flow. No 3D convolutions. No global self-attention.

The whole model in one figure



Top: one frame at a time

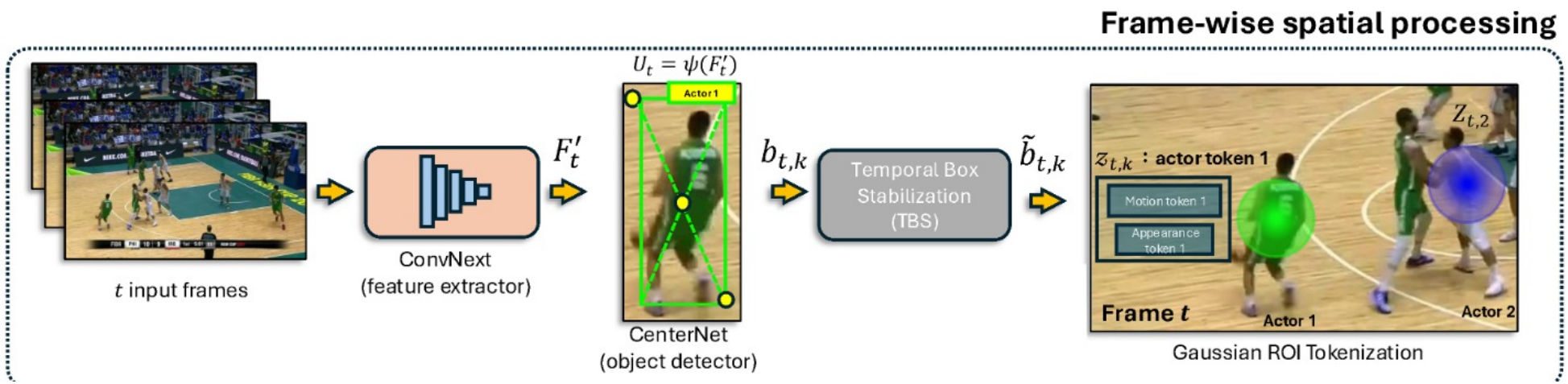
Find the actors, steady the boxes, turn each actor into one token.

Bottom: across the clip

Follow each actor through time, add scene context, then predict the action and its start and end.

Figure 1 from the paper. Everything is trained together in one pass.

One frame at a time



1

Backbone

ConvNeXt-Tiny on each frame. Small and fast.

2

Detector

A CenterNet head. Both outputs share one tower, so they cannot disagree.

3

Box smoothing

The box is averaged with its neighbours. Nothing is learned here.

4

Actor tokens

A soft Gaussian region per actor, pooled from features and from frame differences.

The result: steady boxes, and actor tokens that already carry motion. Optical flow is never computed.

Across the clip

A GRU for each actor

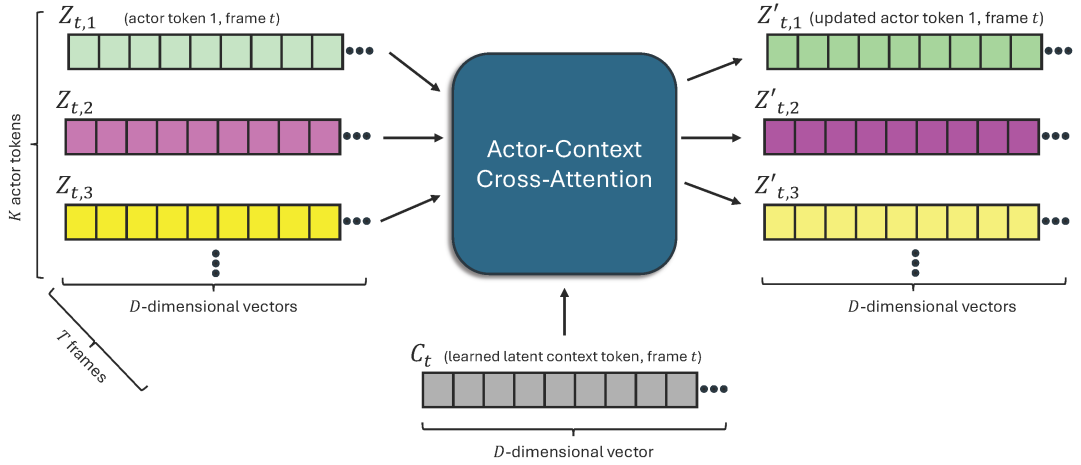
Each actor gets its own small recurrent unit that carries it through the clip.

Because actors stay separate here, a nearby actor cannot pull a token off track.

One context token per frame

A single learned token summarises the frame. It is a parameter, not an image feature.

Every actor talks to that one token instead of to every other actor.



This is what replaces clip level space-time attention, at a small fraction of the cost.

Figure 2 from the paper. Cross-attention runs inside a frame only.

Predicting the tube

At every frame, for every actor, the head predicts:

- the action label
- is an action happening?
- is this the start?
- is this the end?
- how far to the start and end
- how close to the middle

one clip, one action, from start to end

A loss on the whole segment

We compare the predicted start and end against the real ones as an overlap in time, so the model is judged on the interval, not only on single frames.

A loss on how fast things change

The box and the action score are not allowed to jump much between neighbouring frames. This is what keeps a tube in one piece.

Both losses are cheap, and neither needs any extra labels.

Boundary supervision follows BSN [25] and BMN [24], adapted so that it is per actor.

Results on both benchmarks

UCF101-24 [32]

Method	Stream	V@0.5	F@0.5
TacNet [31]	Two-stream	52.9	72.1
MOC [22]	RGB	50.7	72.1
I3D Det. [5]	Two-stream	59.9	76.3
TubeR [41]	RGB	58.4	83.2
CFAD [21]	Two-stream	64.6	72.5
TubeLite (ours)	**RGB**	**71.7**	**82.1**

+7.1 pp Video-mAP on UCF101-24, over the best method we compare with

MultiSports [20]

Method	Stream	V@0.5	F@0.5
MOC (K=11) [22]	RGB	0.62	25.22
MOC (K=7) [22]	RGB	0.77	22.51
YOWO [19]	Two-stream	0.87	9.28
SlowOnly Det. [12]	RGB	5.50	16.70
SlowFast Det. [12]	Two-stream	9.65	27.72
TubeLite (ours)	**RGB**	**14.20**	**20.07**

+4.5 pp Video-mAP on MultiSports, over the best method we compare with

Compare TubeLite with TubeR on UCF101-24: almost the same frame score (82.1 against 83.2), but a much better video score (71.7 against 58.4).

V@0.5 is Video-mAP@0.5, F@0.5 is Frame-mAP@0.5. TubeLite uses 16-frame RGB clips. Baseline numbers are from the original papers.

What it costs

15.65 M
parameters

61.96
GFLOPs per clip

273.6
frames per second

GFLOPs for one 16-frame clip

One 16-frame clip takes 58 milliseconds.

about 5 times
fewer operations than the video backbones this field usually builds on.

about 4 times
fewer parameters than TubeR, which has around 60 M.

Measured on one NVIDIA H100 with 16-frame clips. Baseline figures are as reported by their authors.

Which parts matter

Variant	V@0.5	Δ V	F@0.5	Δ F
TubeLite (full)	**71.7**	**—**	**82.1**	**—**
without the GRU	69.3	-2.4	79.7	-2.4
without the tube losses	69.1	-2.6	82.0	-0.1
without the context token	67.5	-4.2	78.7	-3.4

No single part carries the result. The three temporal pieces work together, and each one is cheap.

The context token matters most

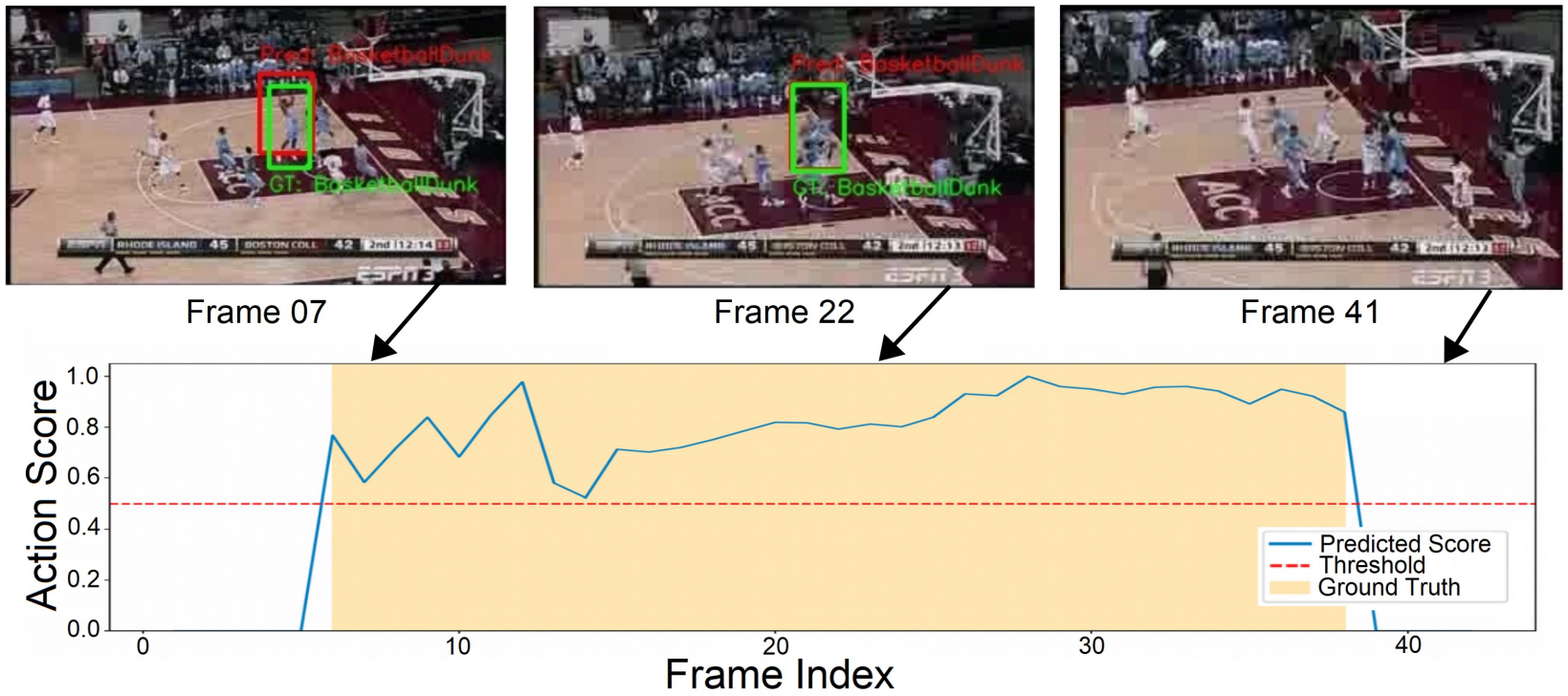
Taking it out costs 4.2 points of video score. Scene context helps a lot when several actors are close together.

The clearest result

Removing the tube losses changes the frame score by 0.1 but the video score by 2.6. Same boxes, worse tubes. That is exactly what those losses were meant to fix.

UCF101-24. Each row removes one part and retrains everything else the same way.

One example



Top row

The predicted box in red stays on the player, even with other players in the way and a moving camera.

Bottom plot

The action score goes up once and comes down once, instead of flickering across the threshold.

The orange band

The ground truth interval. The score follows it closely at both ends.

Figure 3 from the paper. BasketballDunk, UCF101-24.

Summary

1

A small model can win at the video level, if the temporal part is right.

Stable tubes, rather than a bigger backbone, are what moved the video score.

2

We did not need optical flow, 3D convolutions, or global attention.

A 2D backbone, one GRU per actor, and one learned token per frame were enough.

3

Best video score on both benchmarks, at a fraction of the compute.

Plus 7.1 points on UCF101-24 and plus 4.5 on MultiSports, with about 5 times fewer operations.

Where it stops, and what is next

We work on short clips with a fixed number of actor slots, so very crowded scenes and long actions are still open. Next: sparse long range attention, flexible actor slots, and a memory for identity.

Thank you

Happy to take questions.

71.7 / 14.20

Video-mAP@0.5 on UCF101-24 and MultiSports

15.65 M

parameters, RGB only, 16-frame clips

Contact

alisoltaninezhad@uvic.ca

Funded by NSERC Alliance Grants, Archipelago Marine Research, and the Digital Research Alliance of Canada.

ICPR 2026 International Conference on Pattern Recognition LYON

University of Victoria

archipelago

NSERC CRSNG

Digital Research Alliance of Canada

References

1. Alwassel, H., Giancola, S., et al.: TSP: Temporally-sensitive pretraining of video encoders for localization tasks. In: ICCV. pp. 3173–83 (2021)
2. Bertasius, G., Wang, H., et al.: Is space-time attention all you need for video understanding? In: ICLM (2021)
3. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: ICIP (2016)
4. Carion, N., Massa, F., et al.: End-to-end object detection with transformers. In: ECCV. pp. 213–29 (2020)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–308 (2017)
6. Chen, Y., Kalantidis, Y., et al.: A^2-Nets: Double attention networks. In: NeurIPS (2018)
7. Cho, K., Van Merriënboer, B., et al.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: EMNLP (2014)
8. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: NeurIPS (2013)
9. Dosovitskiy, A., Beyer, L., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020)
10. Duan, K., Bai, S., et al.: CenterNet: Keypoint triplets for object detection. In: ICCV. pp. 6569–78 (2019)
11. Feichtenhofer, C.: X3D: Expanding architectures for efficient video recognition. In: CVPR. pp. 203–13 (2020)
12. Feichtenhofer, C., Fan, H., et al.: Slowfast networks for video recognition. In: ICCV. pp. 6202–11 (2019)
13. Gkioxari, G., Girshick, R., et al.: Contextual action recognition with R* CNN. In: ICCV. pp. 1080–8 (2015)
14. Gu, C., Sun, C., et al.: AVA: A video dataset of spatio-temporally localized atomic visual actions. In: CVPR. pp. 6047–56 (2018)
15. He, K., Gkioxari, G., et al.: Mask R-CNN. In: ICCV. pp. 2961–9 (2017)
16. Hou, R., Chen, C., et al.: Tube convolutional neural network (T-CNN) for action detection in videos. In: ICCV. pp. 5822–31 (2017)
17. Jaegle, A., Gimeno, F., et al.: Perceiver: General perception with iterative attention. In: ICML. pp. 4651–64. PMLR (2021)
18. Kalogeiton, V., Weinzaepfel, P., et al.: Action tubelet detector for spatio-temporal action localization. In: ICCV. pp. 4405–13 (2017)
19. Köpüklü, O., Wei, X., et al.: You only watch once: A unified CNN architecture for real-time spatiotemporal action localization. arXiv preprint arXiv:1911.06644 (2019)
20. Li, Y., Chen, L., et al.: Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In: ICCV (2021)
21. Li, Y., Lin, W., et al.: CFAD: Coarse-to-fine action detector for spatiotemporal action localization. In: ECCV. pp. 510–27 (2020)
22. Li, Y., Wang, Z., et al.: Actions as moving points. In: ECCV. pp. 68–84 (2020)
23. Lin, J., Gan, C., et al.: TSM: Temporal shift module for efficient video understanding. In: CVPR. pp. 7083–93 (2019)
24. Lin, T., Liu, X., et al.: BMN: Boundary-matching network for temporal action proposal generation. In: ICCV. pp. 3889–98 (2019)
25. Lin, T., Zhao, X., et al.: BSN: Boundary sensitive network for temporal action proposal generation. In: ECCV. pp. 3–19 (2018)
26. Lin, T.Y., Goyal, P., et al.: Focal loss for dense object detection. In: ICCV. pp. 2980–8 (2017)
27. Liu, Z., Mao, H., et al.: A ConvNet for the 2020s. In: CVPR. pp. 11976–86 (2022)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
29. Rezatofighi, H., Tsoi, N., et al.: Generalized intersection over union: A metric and a loss for bounding box regression. In: CVPR. pp. 658–66 (2019)
30. Singh, G., Saha, S., et al.: Online real-time multiple spatiotemporal action localisation and prediction. In: ICCV. pp. 3637–46 (2017)
31. Song, L., Zhang, S., et al.: TACNet: Transition-aware context network for spatio-temporal action detection. In: CVPR. pp. 11987–95 (2019)
32. Soomro, K., Zamir, A.R., et al.: UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
33. Tan, M., Pang, R., et al.: EfficientDet: Scalable and efficient object detection. In: CVPR. pp. 10781–90 (2020)
34. Tian, Q., Miao, W., et al.: STCA: an action recognition network with spatio-temporal convolution and attention. Int. J. Multimed. Inf. Retr. **14**(1), 1 (2025)
35. Tran, D., Wang, H., et al.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR. pp. 6450–9 (2018)
36. Wu, C.Y., Li, Y., et al.: MeMViT: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In: CVPR. pp. 13587–97 (2022)
37. Xu, M., Zhao, C., et al.: G-TAD: Sub-graph localization for temporal action detection. In: CVPR. pp. 10156–65 (2020)
38. Yang, X., Yang, X., et al.: STEP: Spatio-temporal progressive learning for video action detection. In: CVPR. pp. 264–72 (2019)
39. Zhang, C.L., Wu, J., et al.: ActionFormer: Localizing moments of actions with transformers. In: ECCV. pp. 492–510. Springer (2022)
40. Zhao, J., Snoek, C.G.: Dance with flow: Two-in-one stream action detection. In: CVPR. pp. 9935–44 (2019)
41. Zhao, J., Zhang, Y., et al.: TubeR: Tubelet transformer for video action detection. In: CVPR. pp. 13598–607 (2022)