On the Benefits of 3D Pose and Tracking for Human Action Recognition

Jathushan Rajasegaran[1,2], Georgios Pavlakos[1], Angjoo Kanazawa[1], Christoph Feichtenhofer[2], Jitendra Malik[1,2]

[1]UC Berkeley [2]Meta AI, FAIR

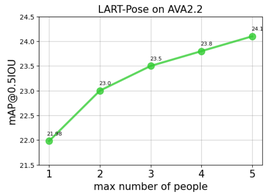
Introduction

- We propose LART for recognizing people monocular videos using 3D representations and contextual appearance .
- First, we track people in videos and use their SMPL 3D pose, shape for 3D representations.
- For each people we also capture their appearance and scene in a contextualized features extracted from MViT.
- We train a simple vanilla transformer on the tracks of multiple people, to learn actions of each actors and also learn to reason interaction between multiple actors.

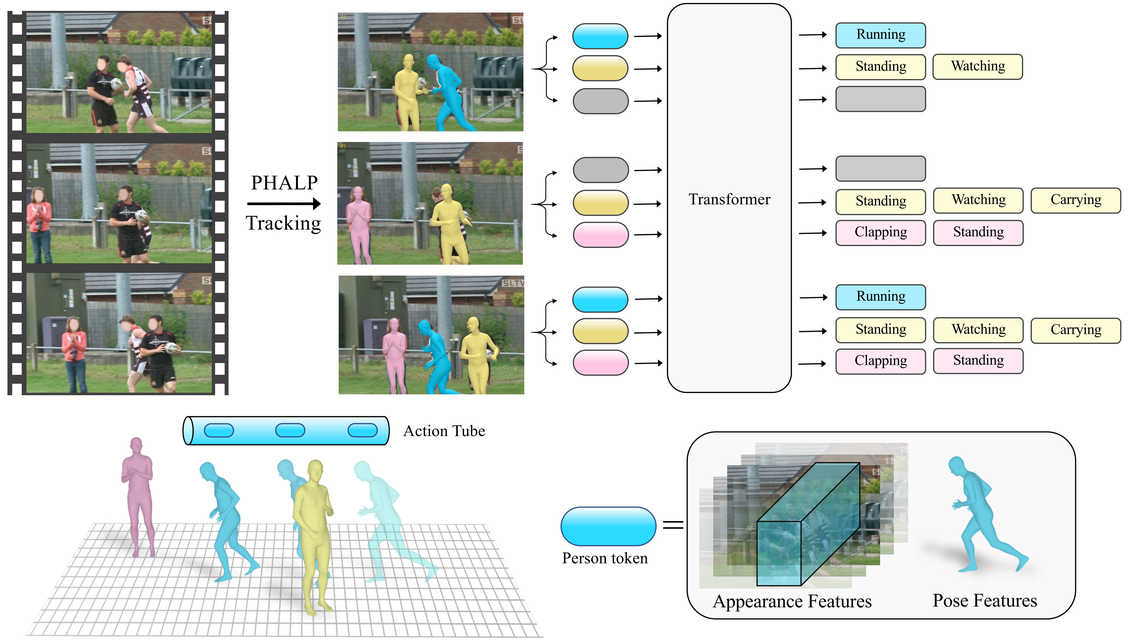
Experiments and Results

Model	Pretrain	mAP
SlowFast R101, 8×8 [18]	K400	23.8
MViTv1-B, 64×3 [14]		27.3
SlowFast 16×8 +NL [18]		27.5
X3D-XL [16]		27.4
MViTv1-B-24, 32×3 [14]	K600	28.7
Object Transformer [66]		31.0
ACAR R101, 8×8 +NL [43]		31.4
ACAR R101, 8×8 +NL [43]	K700	33.3
MViT-L↑312, 40×3 [37],	IN-21K+K400	31.6
MaskFeat [64]	K400	37.5
MaskFeat [64]	K600	38.8
Video MAE [17, 54]	K600	39.3
Video MAE [17, 54]	K400	39.5
LART	K400	42.3 (+2.8)

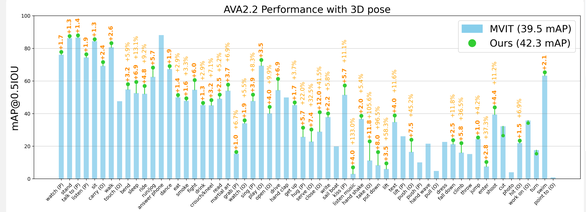
- LART achivie SOTA performance on AVA. AVA-Kinetics datasets.



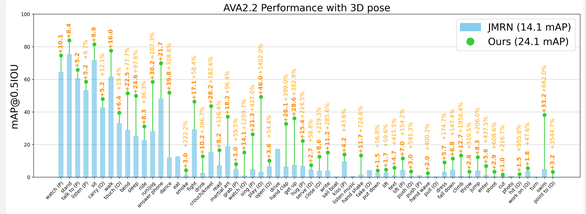
- LART-Pose scales with number of people in the scene to learn more complex interactions.



Experiments and Results



- Class wise comparison with our model and MViT, across multiple classes our lagrangian framework model perform better across multiple classes.



- Our model can also work only with 3D poses-tracks and having continuous tracks over time allow us to perform over 10 mAP better than previous STOA methods.

Qualitative Results

More results and Code