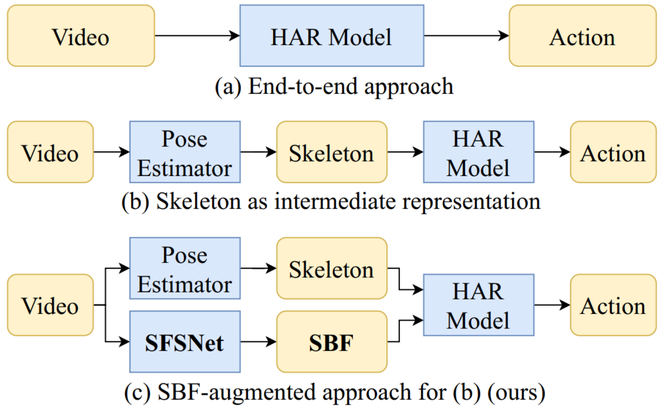
SBF: An Effective Representation to Augment Skeleton for Video-based Human Action Recognition

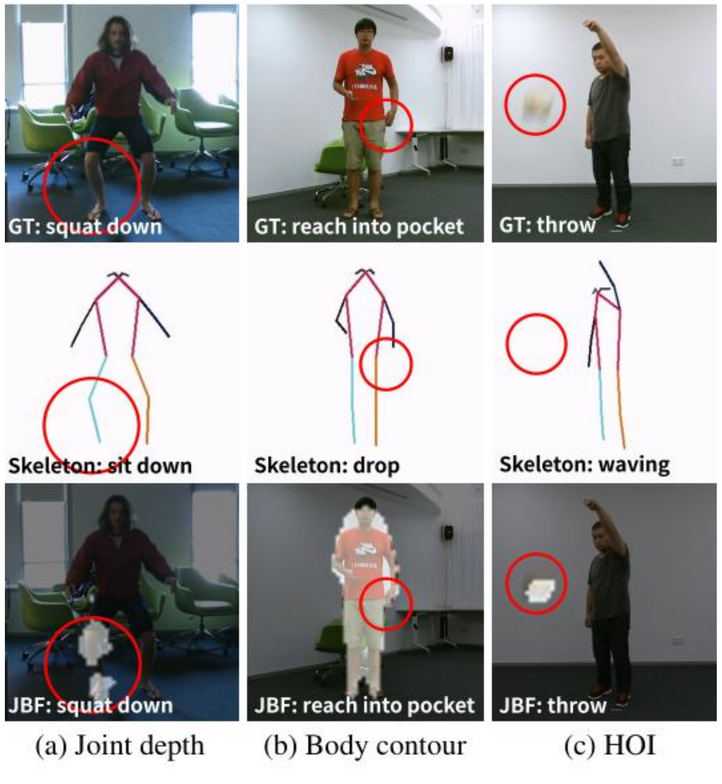
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Zhuoxuan Peng[1] · Yiyi Ding[2] · Lin Yang[1] · S.-H. Gary Chan[1]
[1]The Hong Kong University of Science and Technology
[2]The Hong Kong University of Science and Technology (Guangzhou)



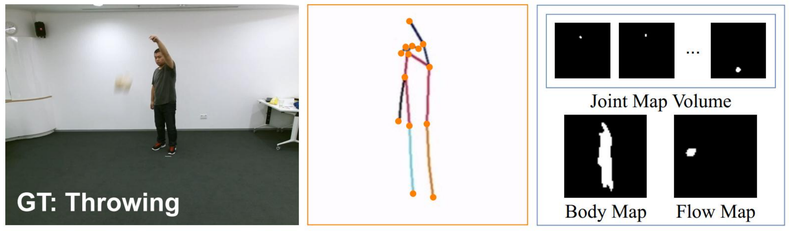
1. Background

- End-to-end video-based human action recognition (HAR) methods are computationally expensive
- 2D skeletons are compact but lack action-critical information



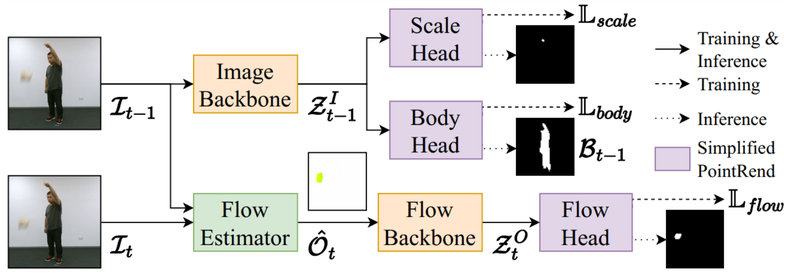
Failure modes addressed by SBF: depth ambiguity, missing body parts, and missing human-object interaction.

2. Scale-Body-Flow Representation

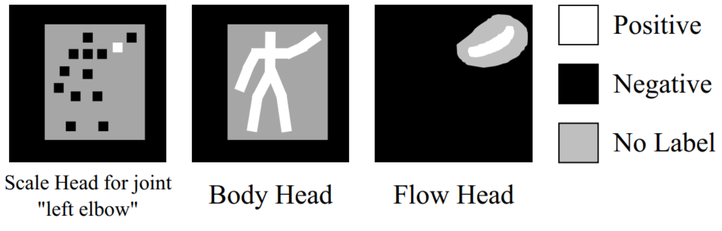


- We propose Scale-Body-Flow (SBF), a novel representation to augment 2D skeleton for video-based HAR.
- Each component is one or more binary maps
- **Scale map volume (S)** to capture joint depth: The scale of a joint on the camera plane provides valuable information on depth
- **Body map (B)** to capture body contour: contour of the entire human subject to cover body parts missing in skeleton
- **Flow map (F)** to capture human-object interaction: binarized optical flow values to reflect such interaction

3. Extracting SBF with SFSNet



- We propose SNSNet to predict SBF from video frames without additional annotations beyond existing human skeletons
- Point annotations are sampled as shown in the figure below
- Scale head and body head are supervised using point annotations derived from skeleton
- Flow head is supervised using unsupervised optical flow

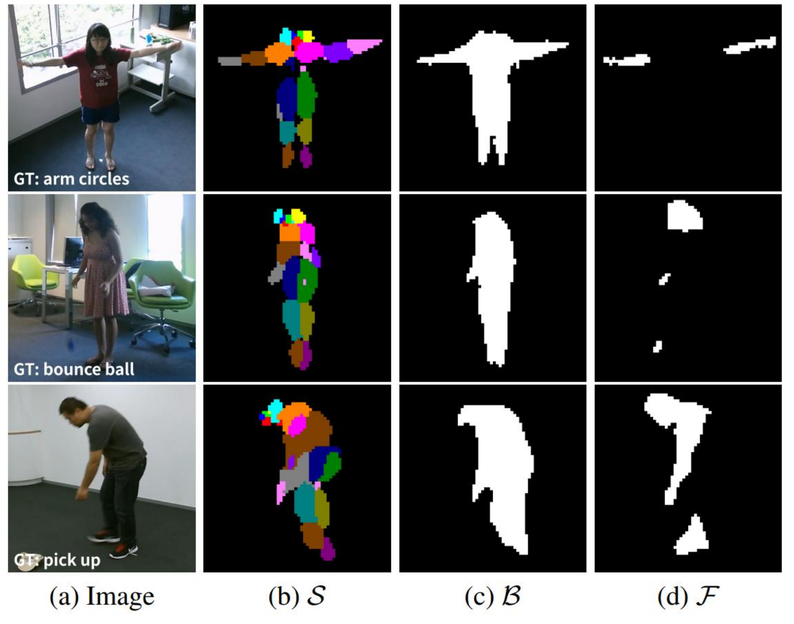


4. Experimental Results

Setting	PoseConv3D	Ours	Gain
NTU X-Sub	94.1	95.0	+0.9
NTU X-View	97.1	98.1	+1.0
NTU120 X-Sub	86.9	89.6	+2.7
NTU120 X-Set	90.3	92.3	+2.0
UCF101	87.0	87.5	+0.5
HMDB51	69.7	71.2	+1.5

Method (Extractor Size)	# Params	FLOPs	Acc
CTR-GCN (Base)	36.3M	1.1T	93.5
Proto-GCN (Large)	90.6M	3.6T	94.2
Ours (Small)	**25.7M**	**617G**	93.8
Ours (Base)	67.4M	3.5T	**94.8**

- SBFConv3D consistently improves over PoseConv3D while preserving compact downstream computation.



Visualization of SBF components predicted by SFSNet on NTU120 X-Sub. Each joint's scale map has a distinct color.

5. Conclusion

- SBF is a compact middle ground between pure skeleton and full RGB video: it keeps the efficiency of skeleton-based HAR while recovering depth, contour, and human-object interaction cues.