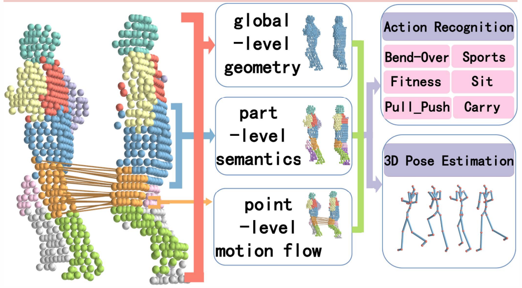
A Unified Framework for Human-centric Point Cloud Video Understanding

Yiteng Xu[1], Kecheng Ye[1], Xiao Han[1], Yiming Ren[1], Xinge Zhu[2], Yuexin Ma[1,*]
[1]ShanghaiTech University [2] The Chinese University of Hong Kong
{xuyt2023,mayuexin}@shanghaitech.edu.cn

CVPR
JUNE 17-21, 2024

Motivation



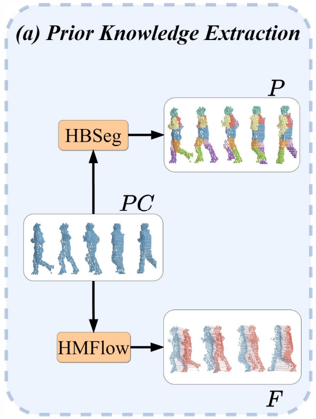
Considering that human has specific characteristics, including the structural semantics of human body and the dynamics of human motions, we propose a unified framework to make full use of the prior knowledge and explore the inherent features in the data itself for generalized human-centric point cloud video understanding.

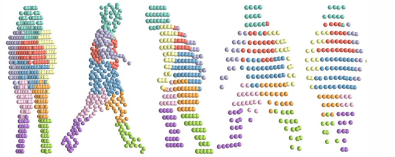
Contributions

- We propose the first framework for human-centric point cloud video understanding for various tasks.
- Containing semantic-guided spatio-temporal representation self-learning and hierarchical feature enhanced fine-tuning, our method takes advantage of prior knowledge of humans for human-centric representation learning.
- Our method achieves state-of-the-art performance on datasets for various human-centric tasks.

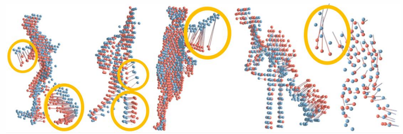
Prior Knowledge Extraction

- Build **HBSeg** and **HMFlow** networks and synthetic datasets to provide fine-grained geometric structure and motion information.



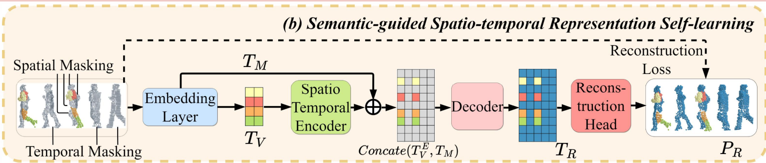


- **HBSeg**: Extract body structure semantic.



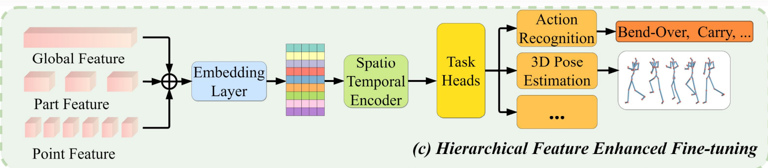
- **HMFlow**: Explore fine-grained motion infomation

Semantic-guided ST Representation Self-learning



- Based on structure semantics of human bodies, the model mines essential geometric and motion features from human point cloud video data itself by masking and predicting body part patches.

Hierarchical Feature Enhanced Fine-tuning



- Integrate global, part, and point level point cloud features to pre-trained STEncoder, therefore fully leveraging prior knowledge for effective and robust human-centric representation learning.

Experiment

	mAcc
PointNet	47.3
PointNet++	40.7
PointMLP	58.1
PointNeXt	50.0
PCT	57.6
HuCenLife	57.4
PointMAE†	58.0
MaST-Pre†	54.1
UniPVU-Human	**61.8**

	MPJPE(mm)↓
LiDARCap(PC)	69.4
LIP(PC)	60.1
UniPVU-Human(PC)	**58.8**
LIP(PC+IMU)	48.9
UniPVU-Human(PC+IMU)	**47.2**

↑ 3D Pose Estimation in LIP
← Action Recognition in HuCenLife

part division	Self-learning Mask		Hierarchical Feature		mAcc
	spatial	temporal	global token	motion flow	
✗	✗	✗	✗	✗	53.4
✗	✓	✗	✗	✗	56.1
✗	✓	✗	✓	✓	58
✓	✗	✗	✗	✗	54.1
✓	✗	✗	✓	✓	55.6
✓	✓	✗	✓	✓	59.9
✓	✗	✓	✓	✓	59.2
✓	✓	✓	✗	✗	58.9
✓	✓	✓	✗	✓	59.3
✓	✓	✓	✓	✗	61.3
✓	✓	✓	✓	✓	**61.8**

- Ablation Studies of Network Design

	proportion of fine-tuning dataset			
	20%	30%	50%	100%
MaST-Pre [25]	39.8(-14.3)	42(-12.1)	48.8(-5.3)	54.1
UniPVU-Human*	44.9(-10.9)	46.4(-9.4)	49.5(-6.3)	55.8
UniPVU-Human	51(**-10.8**)	53.8(**-8**)	57.3(**-4.5**)	61.8

- Effectiveness of Our Self-learning Mechanism in Semi-supervised Settings

Our project：
https://github.com/yiteng-xu/CVPR2024-UniPVU-Human