OpenHuman4D: Open-Vocabulary 4D Human Parsing

[1]Keito Suzuki, [1]Bang Du, [1]Runfa Li, [2]Kunyao Chen, [2]Lei Wang, [2]Peng Liu, [2]Ning Bi, and [1]Truong Nguyen

[1] UC San Diego [2] Qualcomm

Motivation

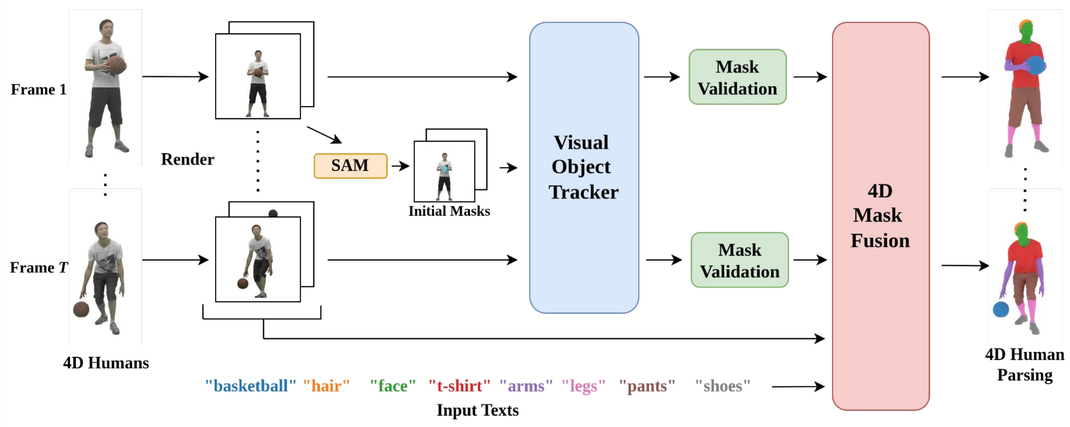
- Static 3D segmentation methods lack temporal information resulting in inter-frame inconsistencies
- 4D segmentation methods require long inference times and have limited segmentation categories

Contributions

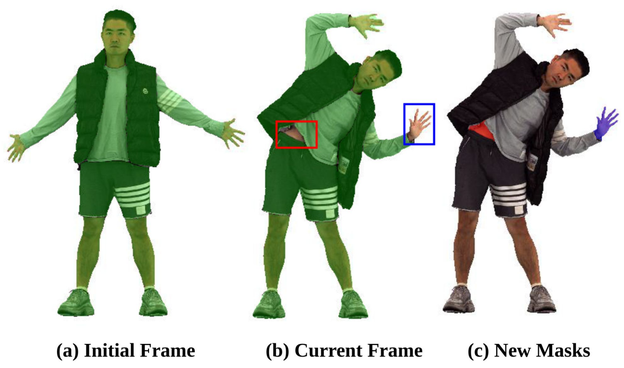
- **First open-vocabulary 4D human parsing method**
- Efficient 4D mask proposal generator using a visual object tracker
- Novel 4D MaskFusion module with memory of inter-frame correspondences

Method

Overall Framework



- Leverage SAM-2 to efficiently propagate mask proposals to each view/frame
- Apply Mask Validation module to check incomplete mask coverage



Mask Validation

4D Mask Fusion

- Compute embeddings for each mask using HumanCLIP
- Combine embeddings from each view/frame for consistency

Results

Numerical Comparison

Metric	Find3D [25]	CloSe-Net [2]	OpenHuman3D [39]	4D-DRESS [43]	Ours
OA	77.97	91.48	95.68	95.65	**96.88**
mAcc	78.09	89.01	93.20	93.08	**96.17**
mIoU	60.59	76.40	88.96	83.55	**92.78**

Comparison of Inference Time

Method	Direct 3D	SMPL	4D	One-time Inference	Avg. Inference
Find3D [25]	✓			7.43s	0.35s
CloSe-Net [2]	✓	✓		2m 50.00s	2m 50.00s
OpenHuman3D [39]				6h 51m 1.85s	4m 8.24s
4D-DRESS [43]			✓	4h 16m 53.99s	4h 16m 53.99s
Ours			✓	17m 17.90s	11.75s

Visual Comparison