JUNE 18-22, 2023

CVPR

VANCOUVER, CANADA

FeatER: An Efficient Network for Human Reconstruction via Feature Map-Based TransformER

Ce Zheng[1], Matias Mendieta[1], Taojiannan Yang[1], Guo-Jun Qi[2,3], Chen Chen[1]

[1]Center for Research in Computer Vision, University of Central Florida

[2] OPPO Seattle Research Center, USA [3] Westlake University

{cezheng,mendieta,taoyang1122}@knights.ucf.edu; guojunq@gmail.com; chen.chen@crcv.ucf.edu

Introduction

Understanding human structure from monocular images is one of the fundamental topics in computer vision:

- Human Pose Estimation (HPE)
- Human Mesh Reconstruction (HMR)

In these tasks, feature maps are often extracted first from the image by a **CNN** backbone, and then further processed by **transformer** to predict the pose and mesh output.

Introduction

Limitations:

- Current transformer such as ViT can only deal with the flattened features when modeling attention. Feature maps with the shape of [n, h, w] need to be flattened as [n, d], forcing an unnatural flattening of the location-sensitive human structural information.

- Furthermore, large embedding dimension makes the transformer computationally expensive.

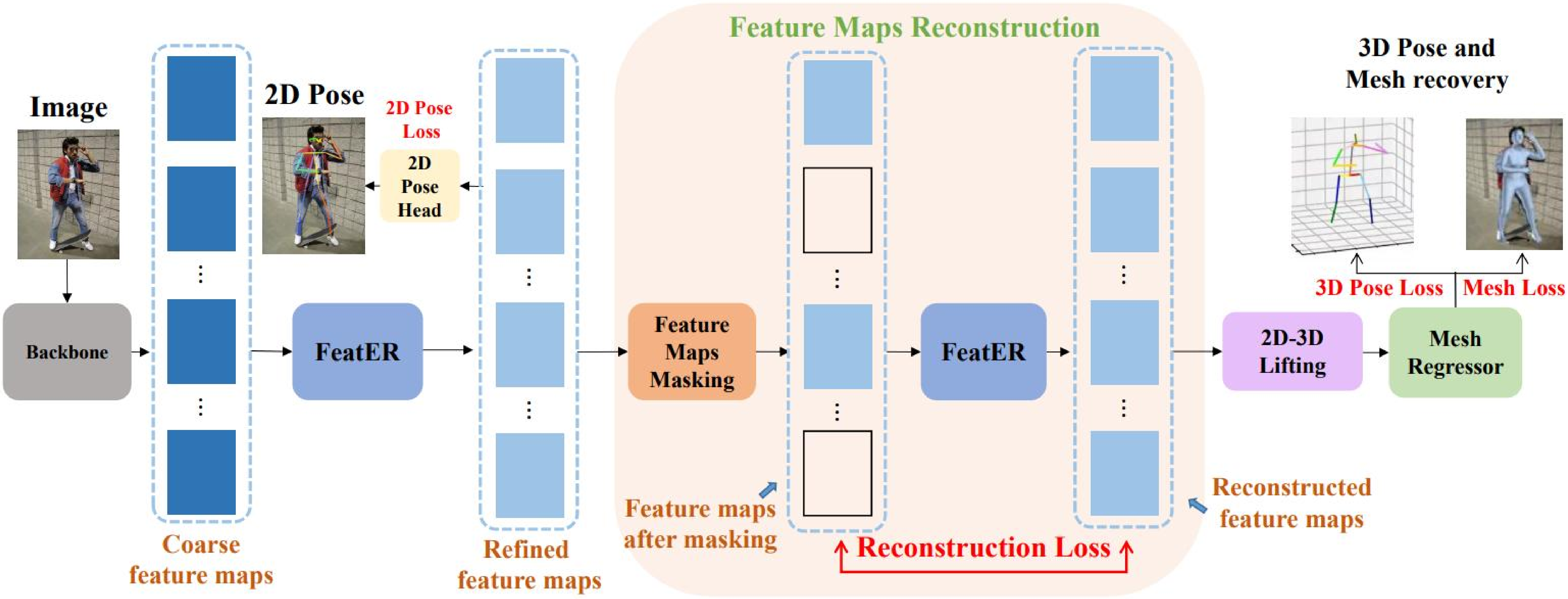
Introduction

Therefore, we propose a Feature map-based transformER **(FeatER)** architecture to properly model feature maps in a resource-friendly manner.

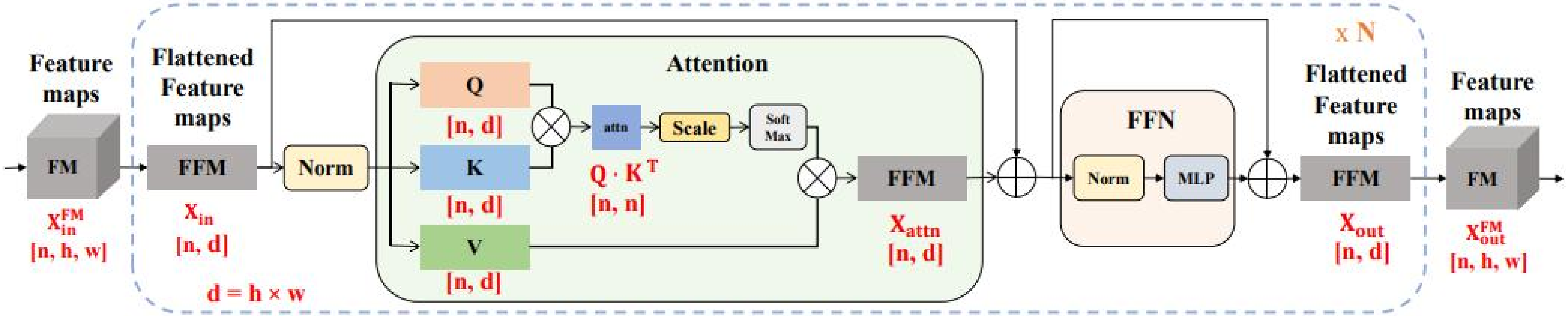
- FeatER preserves the feature map representation in the transformer encoder when modeling self-attention, which is naturally adherent with the HPE and HMR tasks.

- The decompositional design simultaneously provides a significant reduction in computational cost compared with the vanilla transformer. This makes FeatER more suitable for the needs of real-world applications.

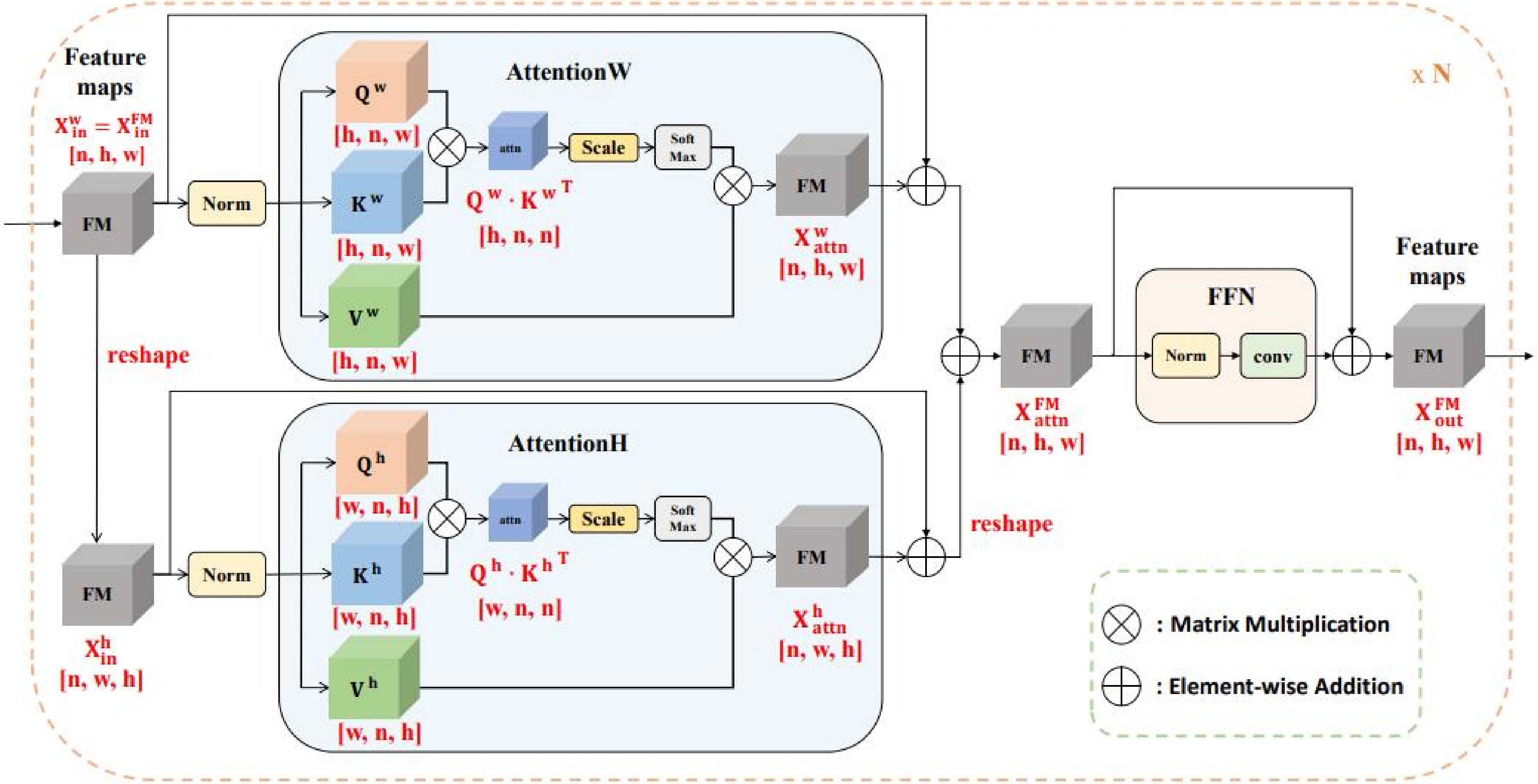
FeatER



An overview of our proposed network for 2D HPE, 3D HPE, and HMR tasks.



(a) Vanilla transformer



(b) FeatER

Experiment Results (2D HPE)

Table 1. 2D Human Pose Estimation performance comparison with SOTA methods on the COCO validation set. The reported Params and MACs of FeatER are computed from the entire pipeline.

Model	Year	Input size	Params (M)	MACs (G)	AP ↑	AP50 ↑	AP75 ↑	AP(M) ↑	AP(L) ↑	AR ↑
Compared with Small Networks										
DY-MobileNetV2 [3]	CVPR 2020	256×192	16.1	1.0	68.2	88.4	76.0	65.0	74.7	74.2
HRFormer_S [41]	NeurIPS 2021	256×192	7.8	2.8	74.0	**90.2**	81.2	70.4	80.7	79.4
Transpose_H_S [39]	ICCV 2021	256×192	8.0	10.2	74.2	-	-	-	-	78.0
Tokenpose_B [23]	ICCV 2021	256×192	13.5	5.7	74.7	89.8	81.4	**71.3**	81.4	**80.0**
FeatER		256×192	8.1	5.4	**74.9**	89.8	**81.6**	71.2	**81.7**	**80.0**
Compared with Large Networks										
SimpleBaseline [37]	ECCV 2018	256×192	34.0	8.9	70.4	88.6	78.3	-	-	76.3
HRNet_W32 [34]	CVPR 2019	256×192	28.5	7.1	74.4	**90.5**	**81.9**	-	-	78.9
PRTR [19]	CVPR 2021	384×288	57.2	21.6	73.1	89.4	79.8	68.8	80.4	79.8
PRTR [19]	CVPR 2021	512×384	57.2	37.8	73.3	89.2	79.9	69.0	80.9	**80.2**
FeatER		256×192	8.1	5.4	**74.9**	89.8	81.6	**71.2**	**81.7**	80.0

Experiment Results (3D HPE and HMR)

Table 2. 3D Pose and Mesh performance comparison with SOTA methods on Human3.6M and 3DPW datasets. The reported Params and MACs of FeatER are computed from the entire pipeline. † indicates video-based methods. The result of HybrIK* is with predicted camera parameters and ResNet34 is used as the backbone.

				Human3.6M		3DPW		
Model	Year	Params (M)	MACs (G)	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPVE↓
SPIN [16]	ICCV 2019	-	-	62.5	41.1	96.9	59.2	116.4
VIBE † [15]	CVPR 2020	-	-	65.6	41.4	82.9	51.9	99.1
I2LMeshNet [31]	ECCV 2020	140.5	36.6	55.7	41.1	93.2	57.7	-
TCMR † [6]	CVPR 2021	-	-	62.3	41.1	95.0	55.8	111.5
HybrIK* [18]	CVPR 2021	27.6	12.7	57.3	36.2	75.3	**45.2**	87.9
ProHMR [17]	ICCV 2021	-	-	-	41.2	-	59.8	-
PyMAF [45]	ICCV 2021	45.2	10.6	57.7	40.5	92.8	58.9	110.1
METRO [24]	CVPR 2021	229.2	56.6	54.0	36.7	77.1	47.9	88.2
MeshGraphormer [25]	ICCV 2021	226.5	56.6	51.2	34.5	74.7	45.6	87.7
DSR [9]	ICCV 2021	-	-	60.9	40.3	85.7	51.7	99.5
TCFormer [43]	CVPR 2022	-	-	62.9	42.8	80.6	49.3	-
FastMETRO [5]	ECCV 2022	48.5	15.8	53.9	37.3	77.9	48.3	90.6
FeatER		11.4	8.8	**49.9**	**32.8**	**73.4**	45.9	**86.9**

Qualitative Results

Output 2D heatmaps

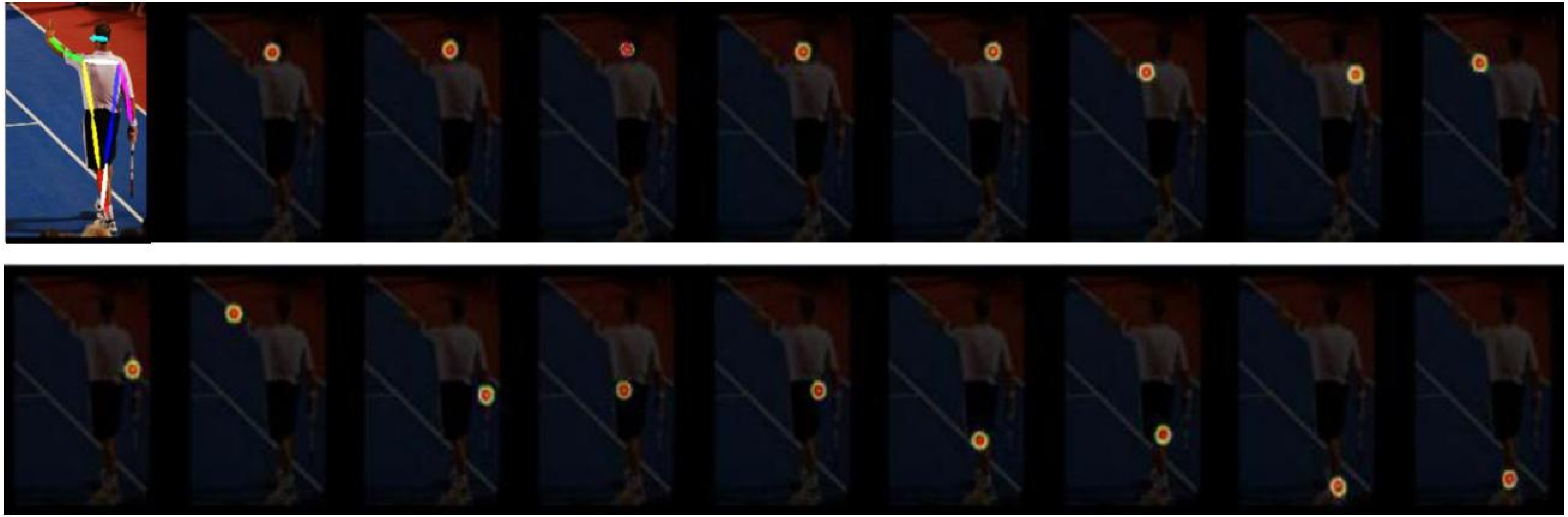
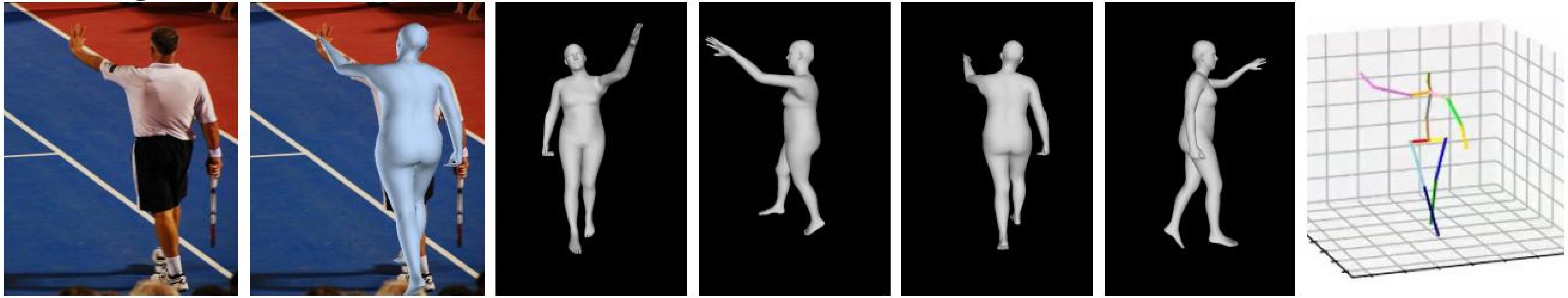
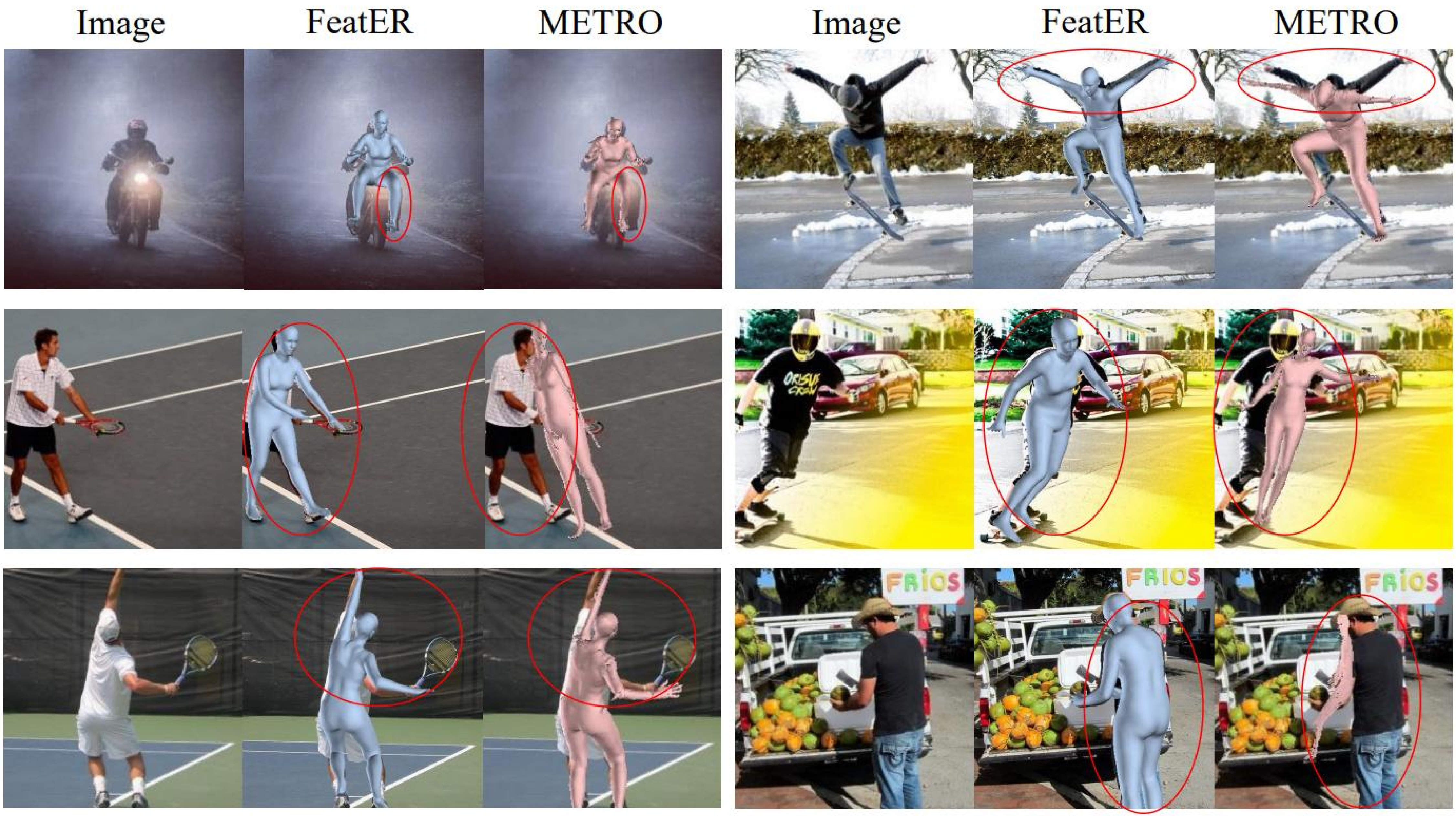


Image Mesh View 1 View 2 View 3 View 4 3D Pose



Qualitative comparison with SOTA method METRO (in-the-wild images)



Thanks for watching!