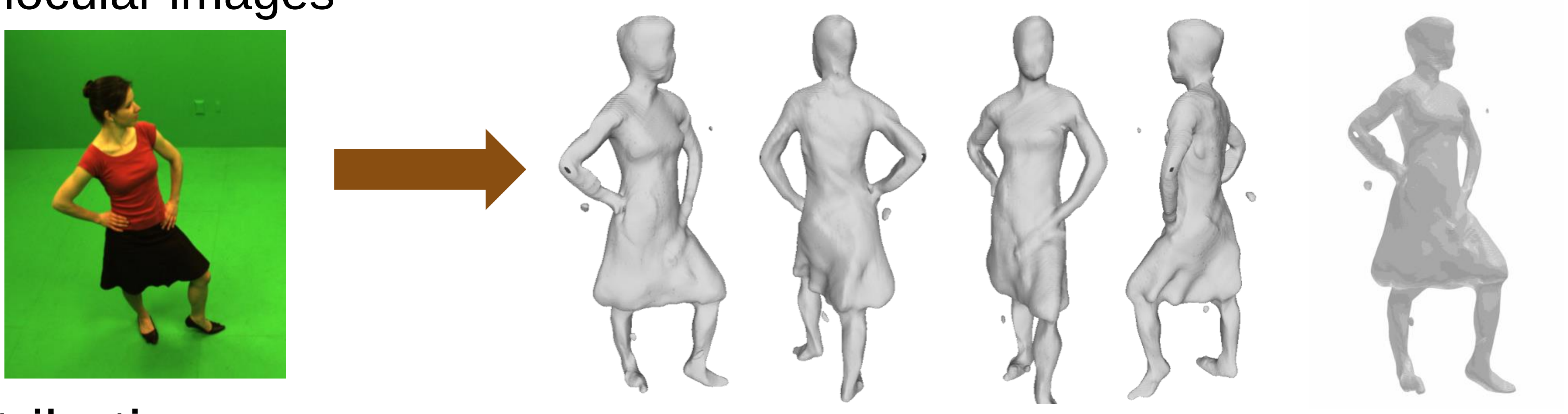
Learning to Implicitly Represent 3D Human Body From Multi-scale Features and Multi-view Images

Zhongguo Li, Magnus Oskarsson, Anders Heyden
Centre for Mathematical Sciences, Lund University

Introduction

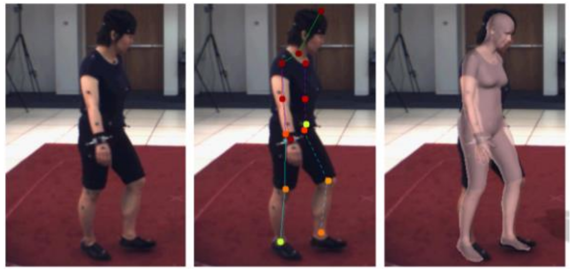
- Goal
  - Capturing and reconstructing detailed 3D human body models from monocular images



- Contribution
  - Estimate the shape details in a memory efficient way based on learning an implicit function
  - Multi-scale features encode both local and global information

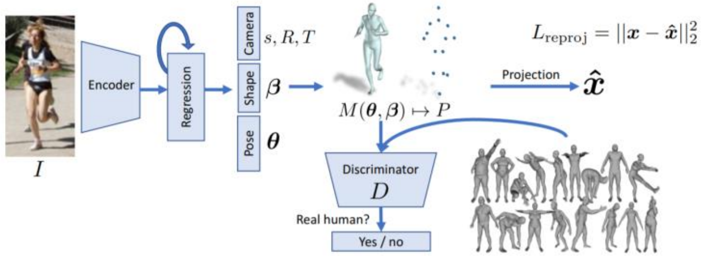
Related work

- Model based methods
  - Optimization based methods



Guan et al. ICCV 2019, Bogo et al. ECCV 2014, Huang et al. 3DV 2017, Xu et al. ACM ToG 2018.

  - Regression based methods

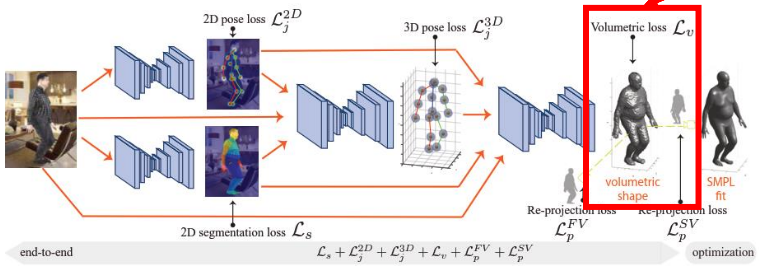


Kanazawa et al. CVPR 2018, Pavlakos et al. CVPR 2018, Kolotouros et al. CVPR 2019, Kolotouros et al. ICCV 2019.

Without detailed appearance

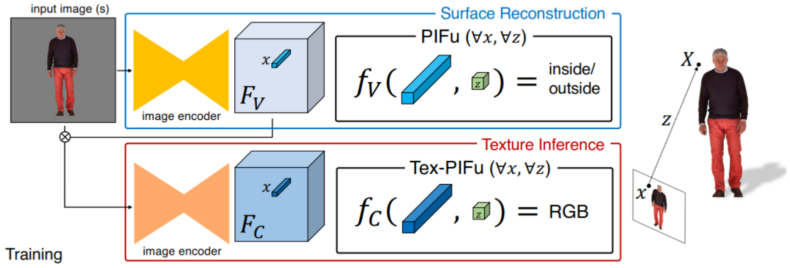
- Model free methods
  - Explicit representation

Volumetic Loss



Varol et al. ECCV 2018, Zheng et al. ICCV 2019, Natsume et al. CVPR 2019.

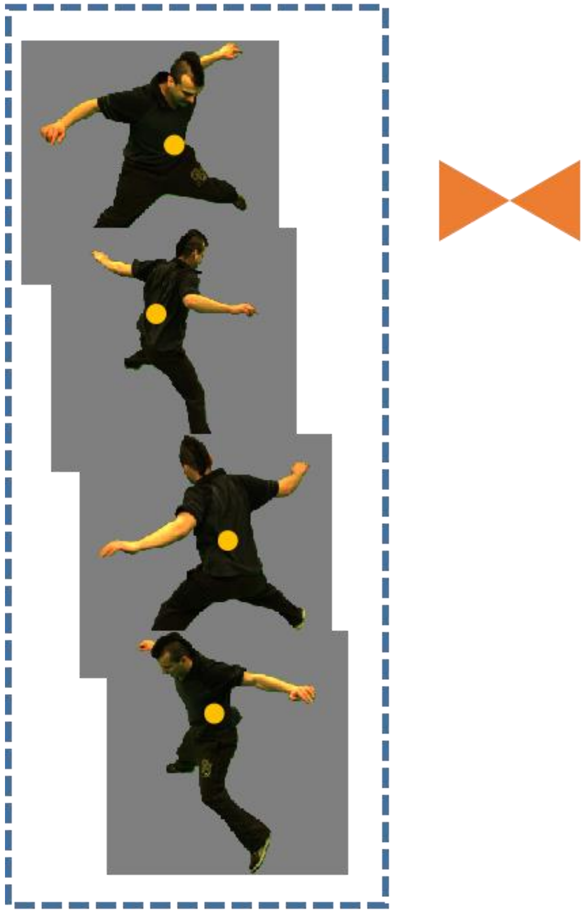
  - Implicit representation



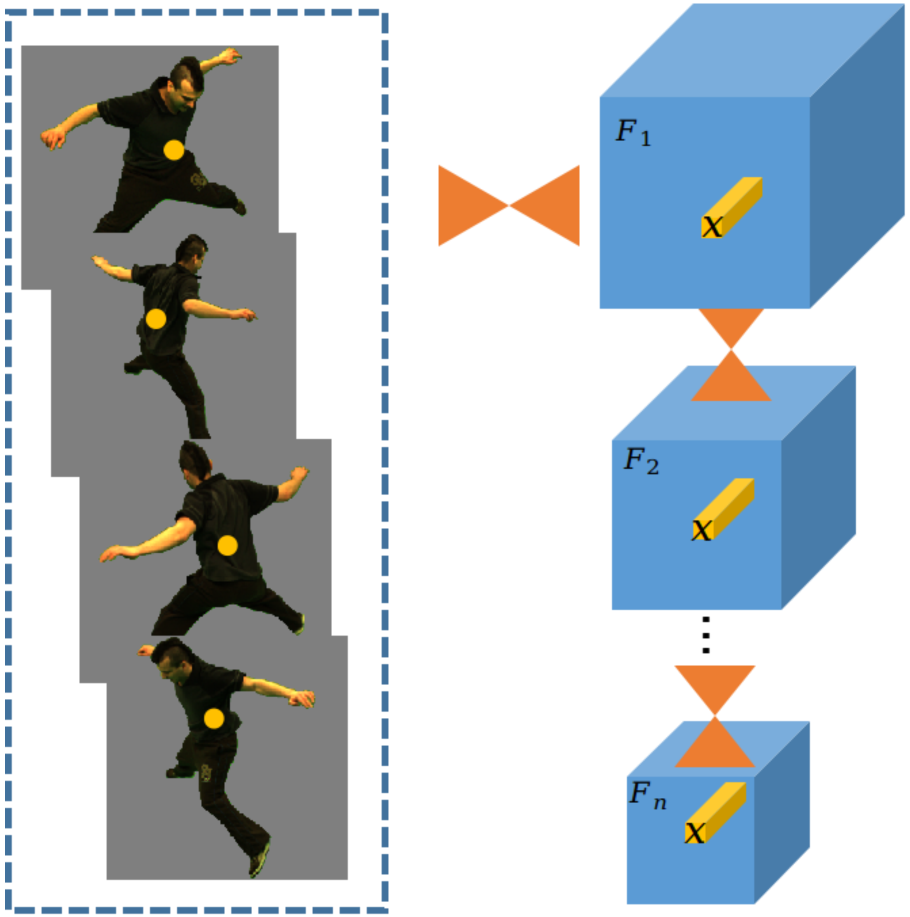
Saito et al. CVPR 2019. Chibane et al. CVPR 2020, Onizuka et al. CVPR 2020.

Our method

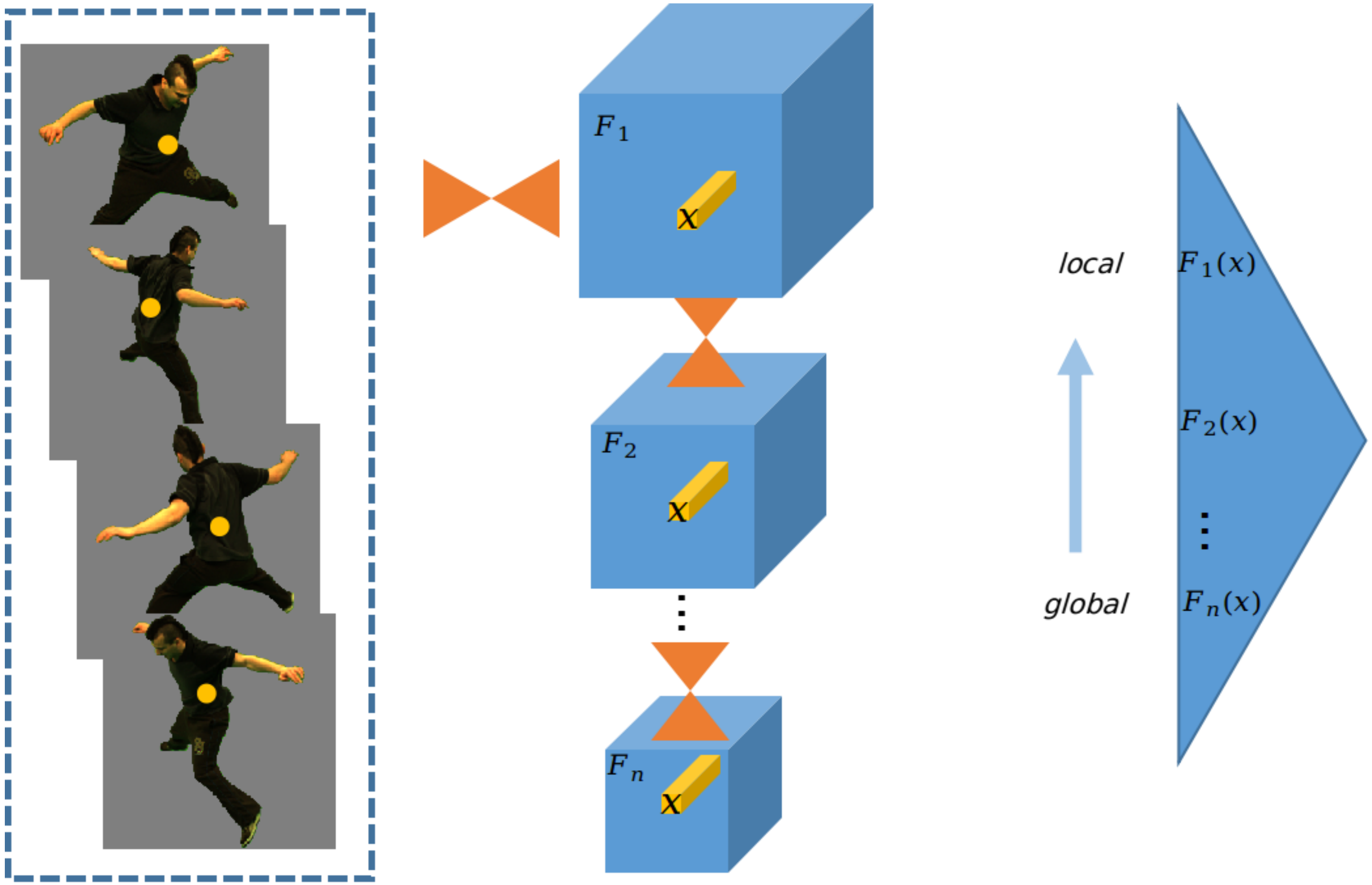
Our method



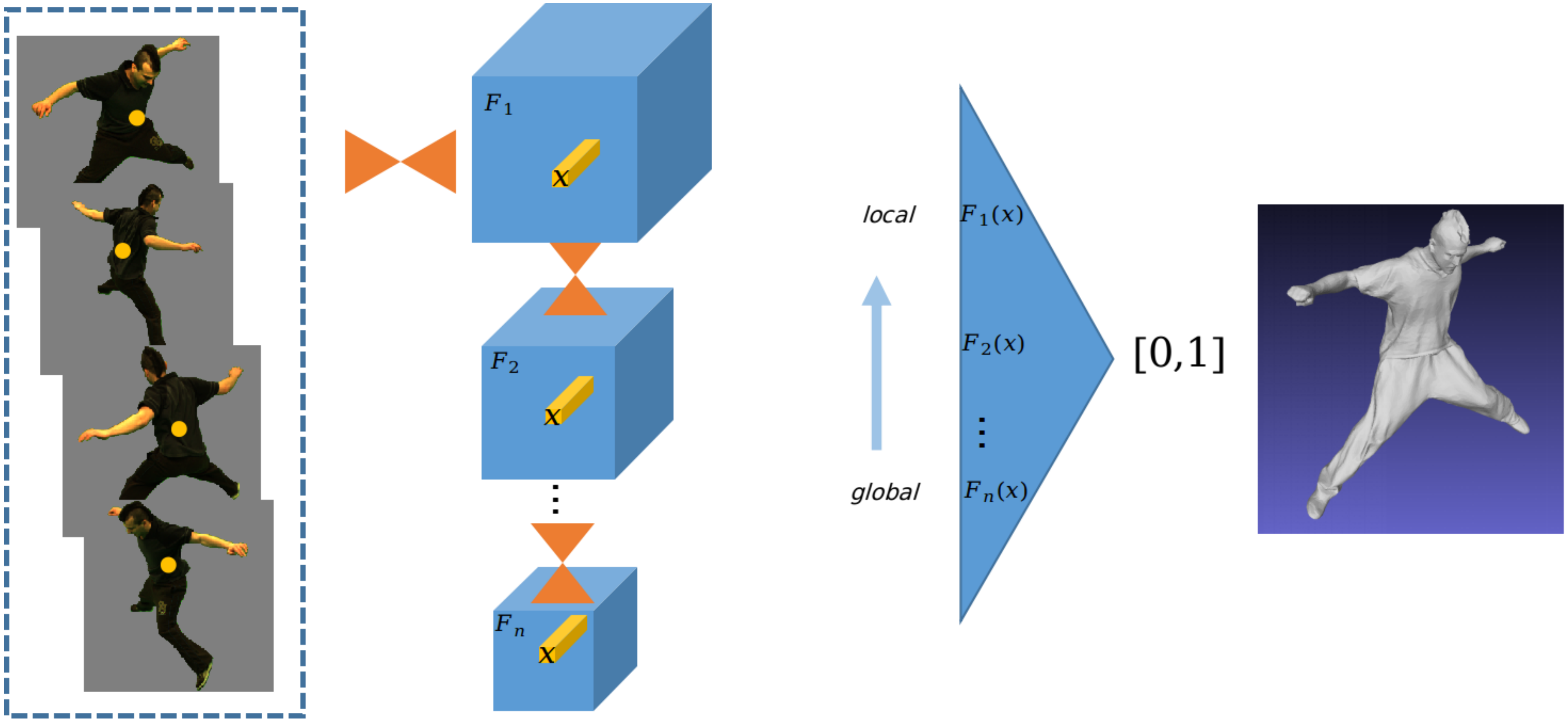
Our method



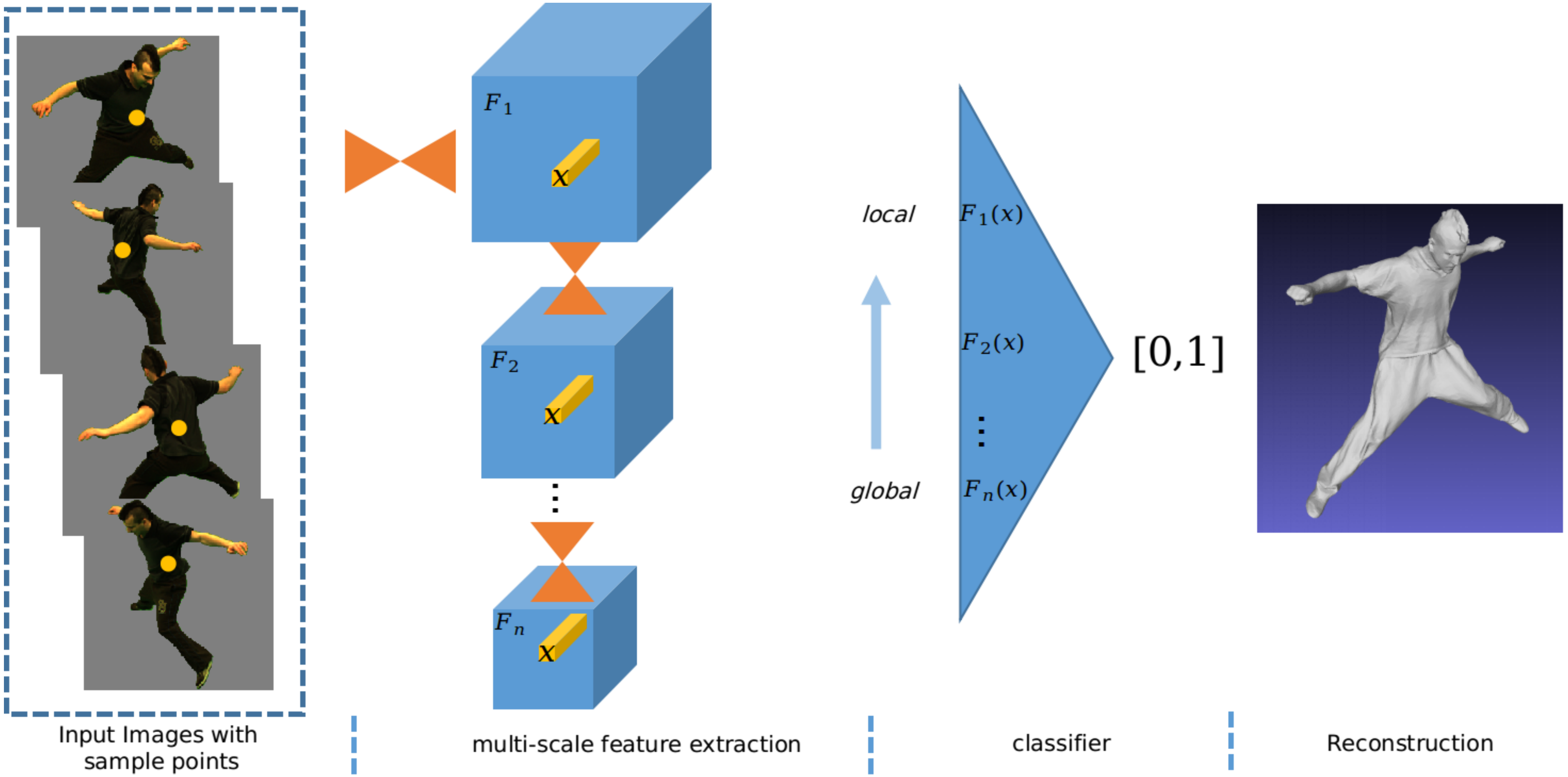
Our method



Our method

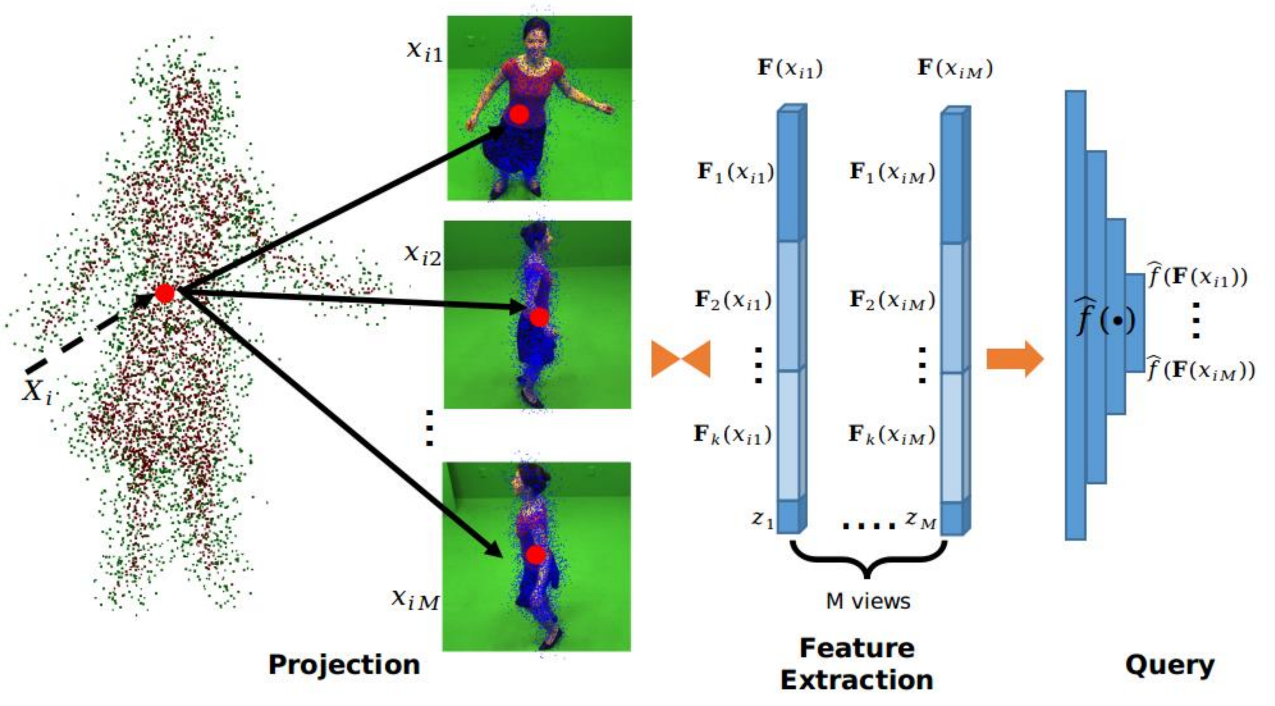


Our method



F_1
x
F_2
x
F_n
x
local
global
$F_1(x)$
$F_2(x)$
$F_n(x)$
[0,1]
Input Images with sample points
multi-scale feature extraction
classifier
Reconstruction

Our method



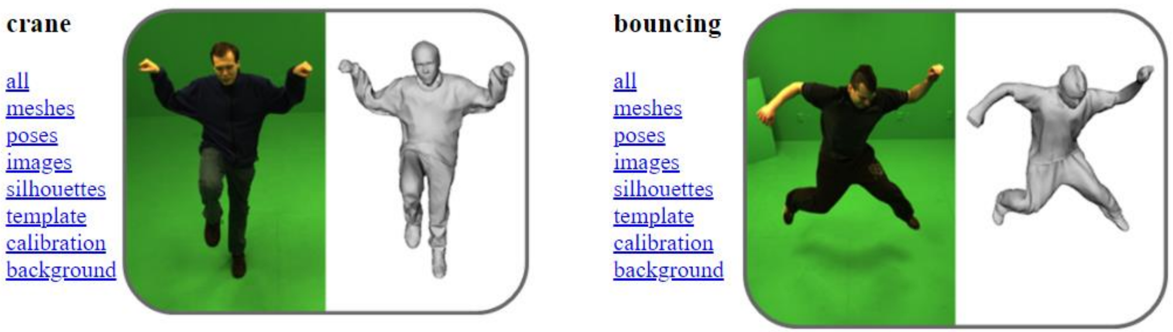
Loss Function:

$$\mathcal{L}_f = \sum_{i=1}^{N} \sum_{j=1}^{M} L(\underbrace{\hat{f}(\mathbf{F}^{(j)}(x_{ij}))}_{\text{Predicted occupancy value}}, \overbrace{o(\hat{X}_i)}^{\text{Ground truth occupancy value}})$$

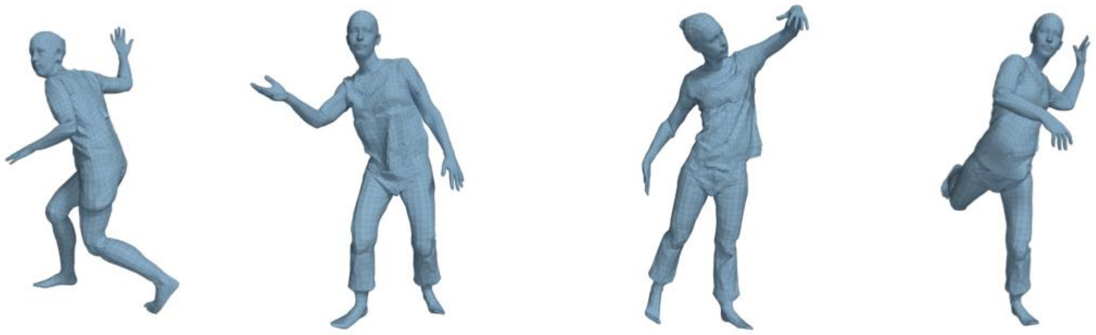
Extract multi-scale features from multi-view images

Experiments

- Datasets



Articulated dataset [1]



CAPE dataset [2]

Dataset	Synthetic?	Total Number	Train / Test
Articulated dataset	No	2000	80% / 20%
CAPE dataset	Yes	2910	80% / 20%

[1] Vlasic et al. Articulated Mesh Animation from Multi-view Silhouettes. ACM ToG 2008.
[2] Ma et al. Learning to Dress 3D People in Generative Clothing. CVPR 2020.

Experiments

- Metrics
  - Point-to-surface Euclidean distances (P2S) from the vertices on the predicted mesh to the ground truth mesh (Lower is better)
  - Volumetic intersection over union (IoU) (Higher is better)
  - Chamfer-L_2 (Lower is better)

$$\text{Chamfer-}L_2 = 0.5 \times \text{Completeness}^2 + 0.5 \times \text{Accuracy}^2$$

Completeness: Distance from the points of the GT mesh to the predicted mesh
Accuracy: Distance from the points of the predicted mesh to the GT mesh

Experiments

- Quantitative results

Quantitative comparison for the **Articulated dataset**

Methods	P2S ↓	Chamfer-L_2↓	IoU ↑
SPIN [1]	3.5206	0.2679	0.3506
DeepHuman [2]	3.9448	0.2675	0.3742
PIFu [3]	0.8194	0.0210	0.8255
Ours	**0.7332**	**0.0194**	**0.8484**

[1] Kolotouros et al. Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop. ICCV 2019.
[2] Zheng et al. DeepHuman: 3D Human Reconstruction From a Single Image. ICCV 2019.
[3] Saito et al. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. ICCV 2019.

Experiments

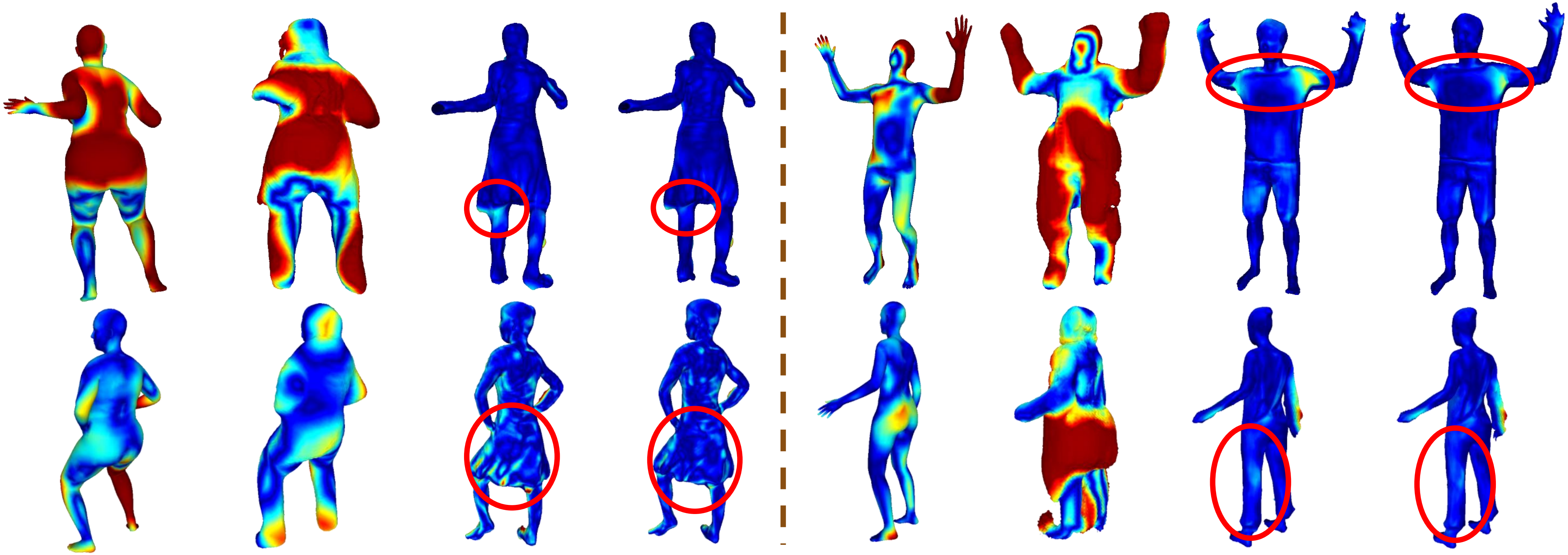
- Quantitative results

Quantitative comparison for the **CAPE dataset**

Methods	P2S ↓	Chamfer-L_2↓	IoU ↑
SPIN [1]	2.2134	0.1271	0.4044
DeepHuman [2]	3.4028	0.1850	0.3861
PIFu [3]	1.0330	0.0212	0.7571
Ours	**0.9482**	**0.0196**	**0.7829**

[1] Kolotouros et al. Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop. ICCV 2019.
[2] Zheng et al. DeepHuman: 3D Human Reconstruction From a Single Image. ICCV 2019.
[3] Saito et al. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. ICCV 2019.

Experiments



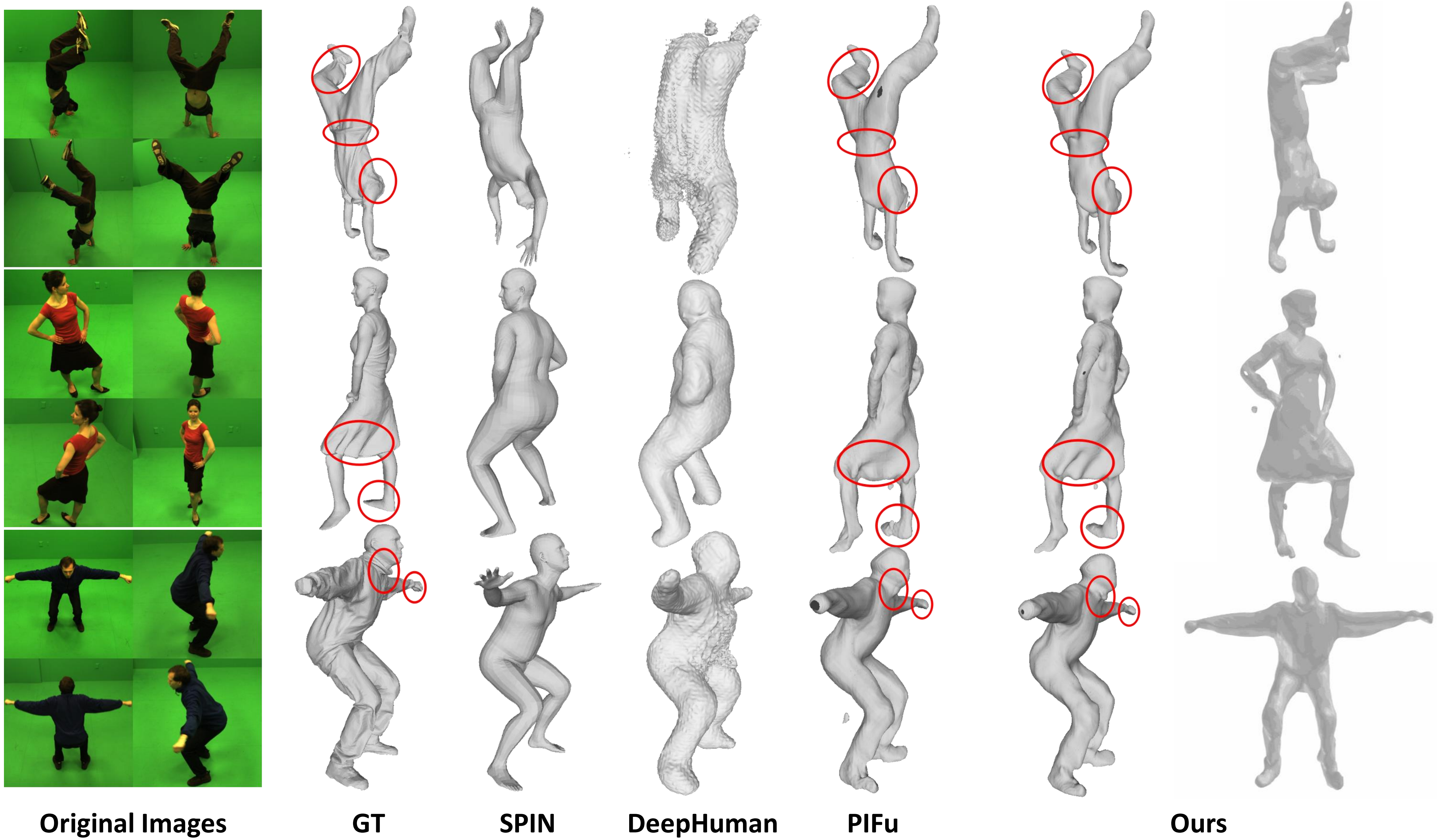
SPIN [1] DeepHuman [2] PIFu [3] Ours SPIN [1] DeepHuman [2] PIFu [3] Ours

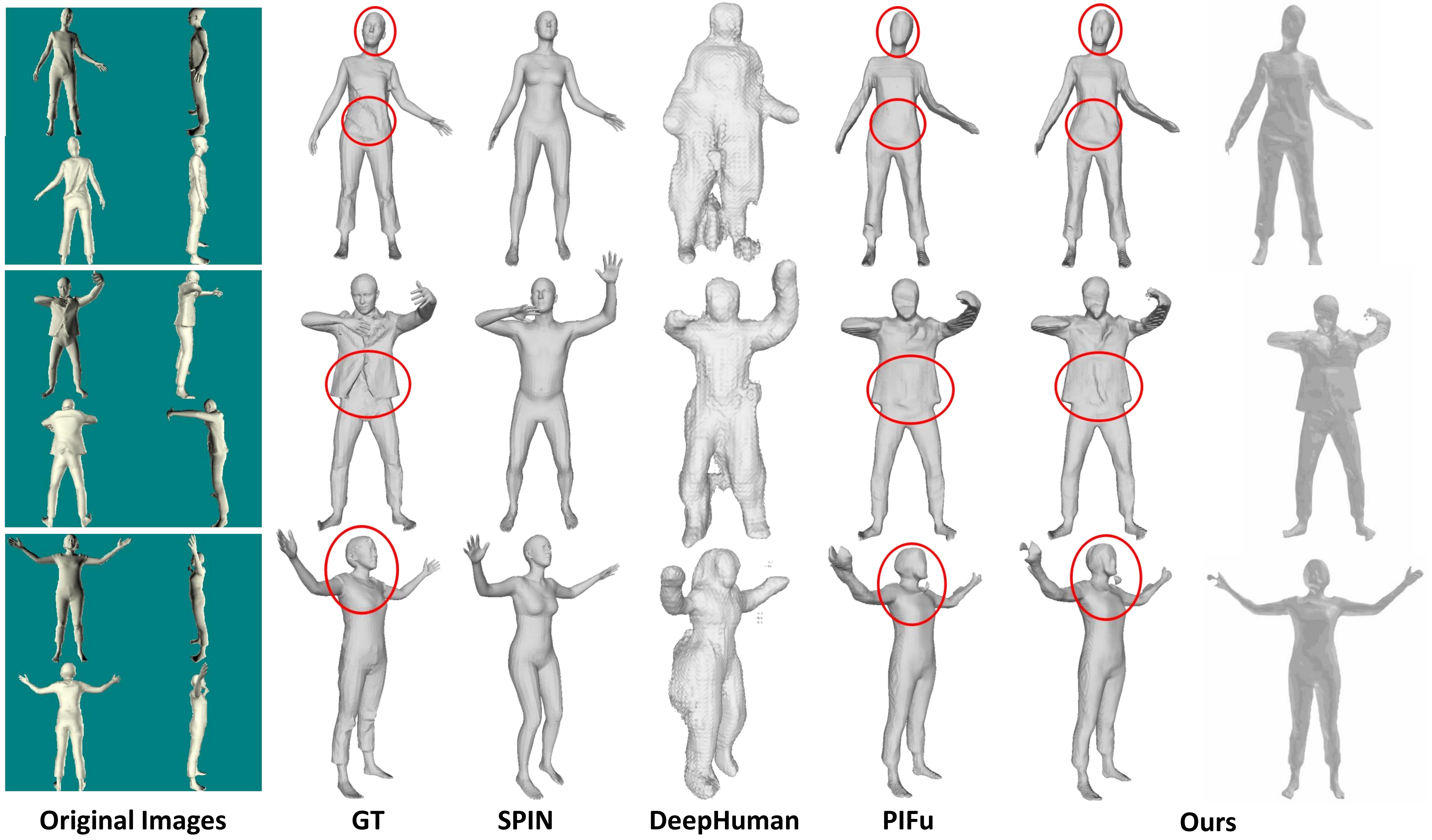
Visualization of P2S

[1] Kolotouros et al. Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop. ICCV 2019.
[2] Zheng et al. DeepHuman: 3D Human Reconstruction From a Single Image. ICCV 2019.
[3] Saito et al. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. ICCV 2019.

Experiments

- Qualitative results


Original Images
GT
SPIN
DeepHuman
PIFu
Ours


Original Images
GT
SPIN
DeepHuman
PIFu
Ours

Thank you!