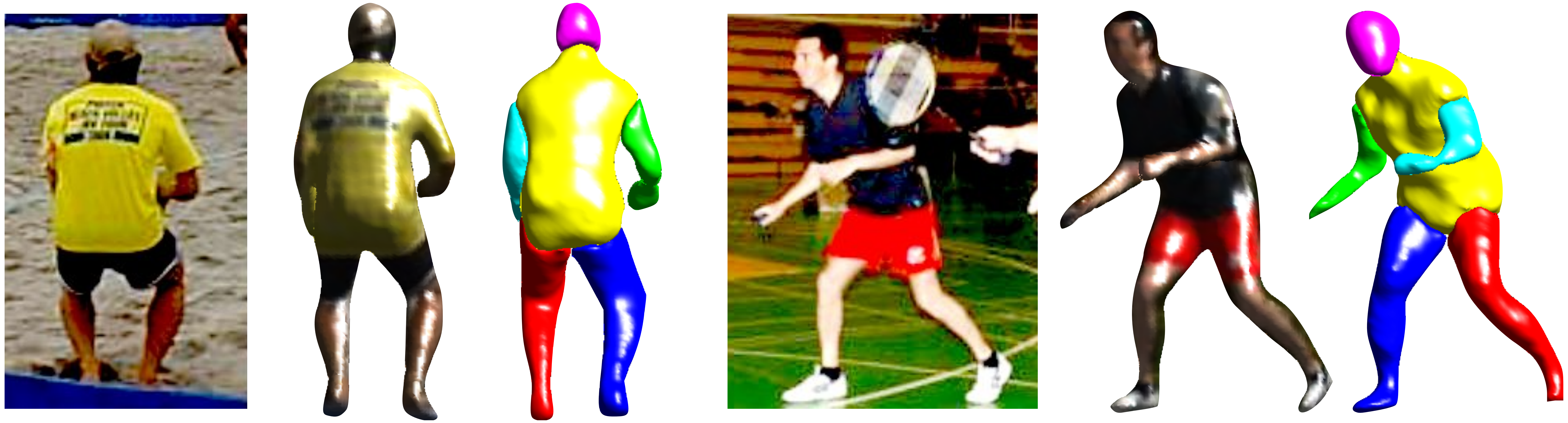
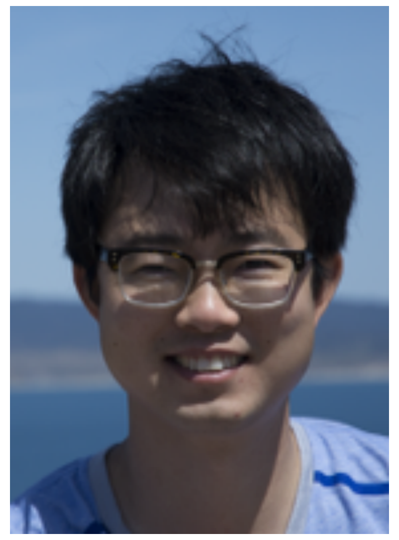


BodyNet: Volumetric Inference of 3D Human Body Shapes



Gul Varol[1]

Duygu Ceylan[2]

Bryan Russell[2]

Jimei Yang[2]

Ersin Yumer[2]

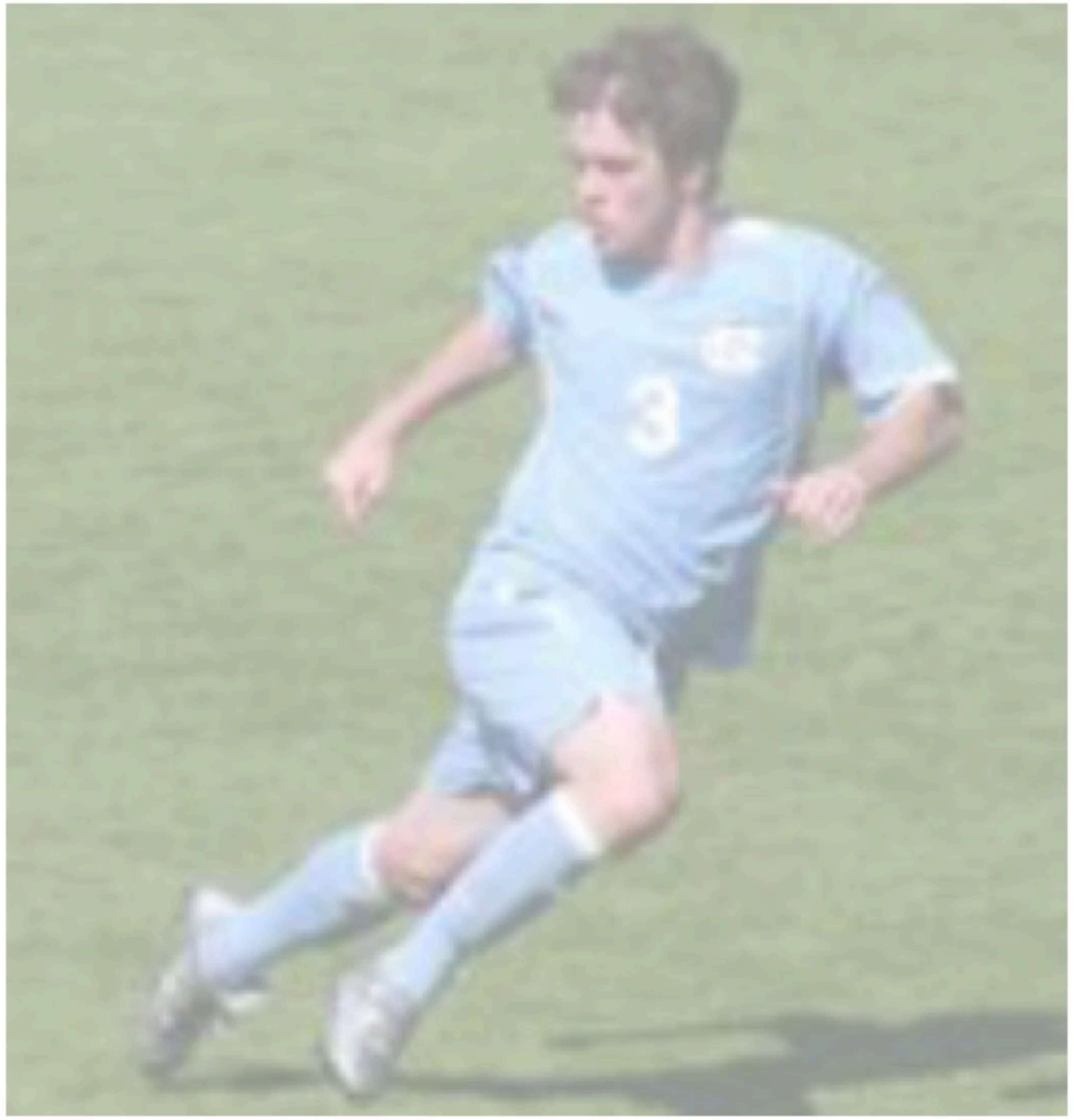
Ivan Laptev[1]

Cordelia Schmid[1]

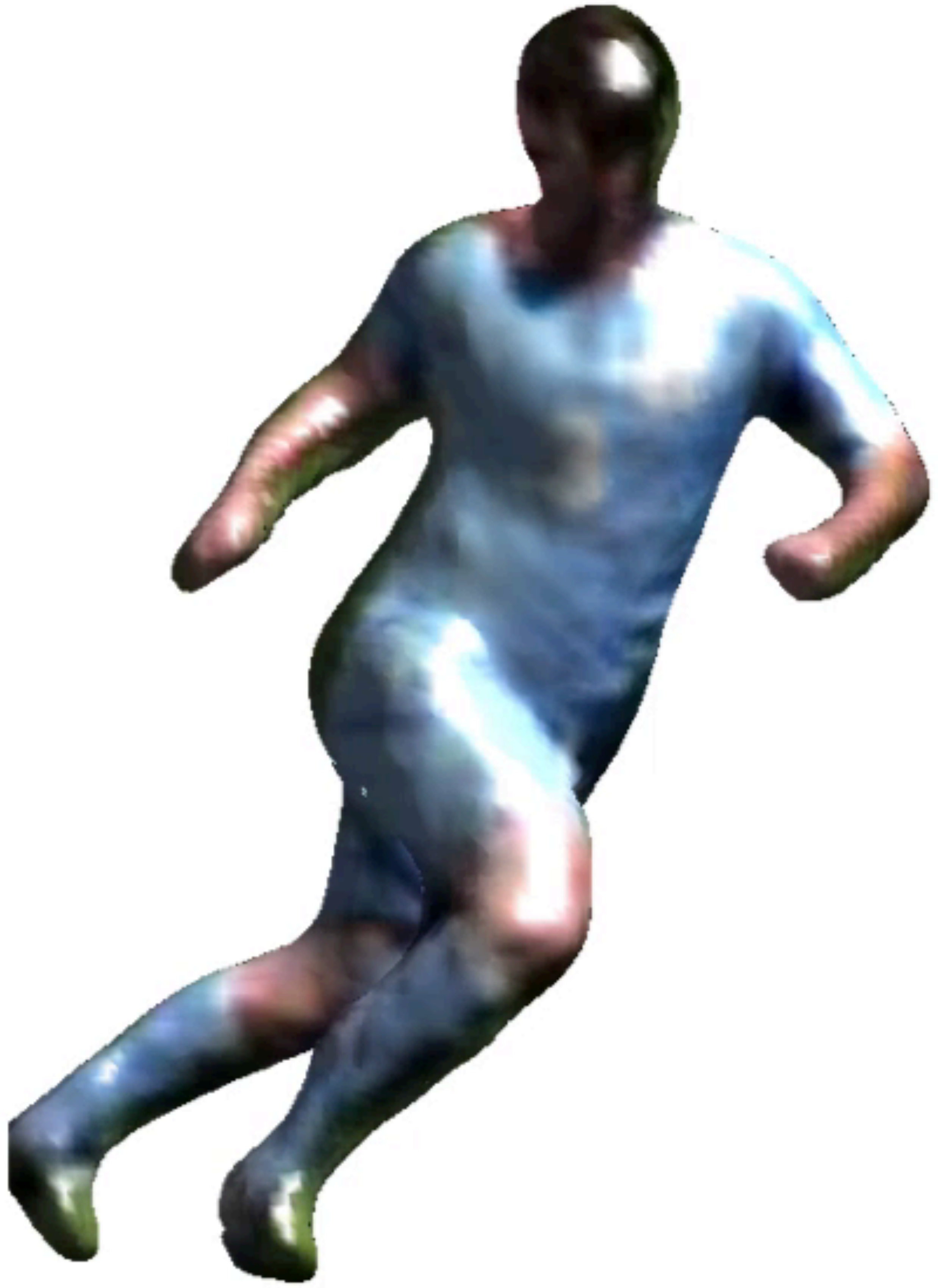
[1]Inria, [2]Adobe Research

Goal

RGB input

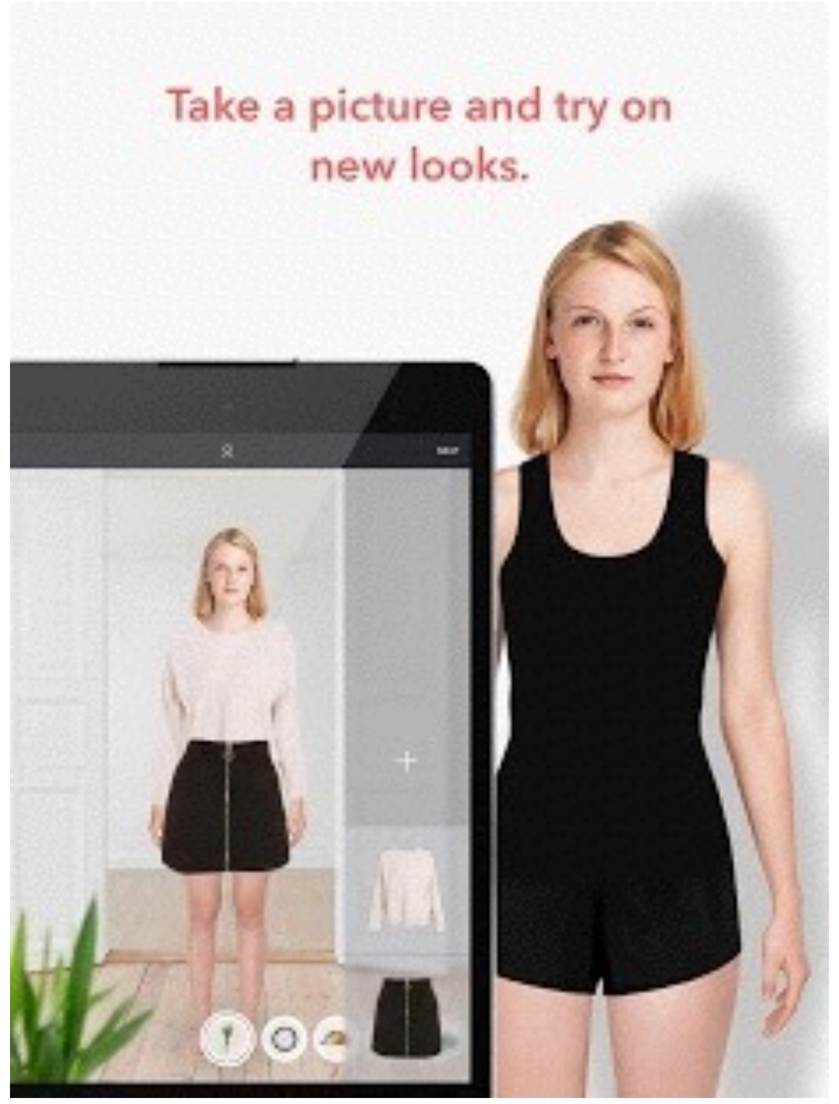


Output 3D shape prediction



Why?

Virtual try-on



Pictofit

Creation of digital avatars

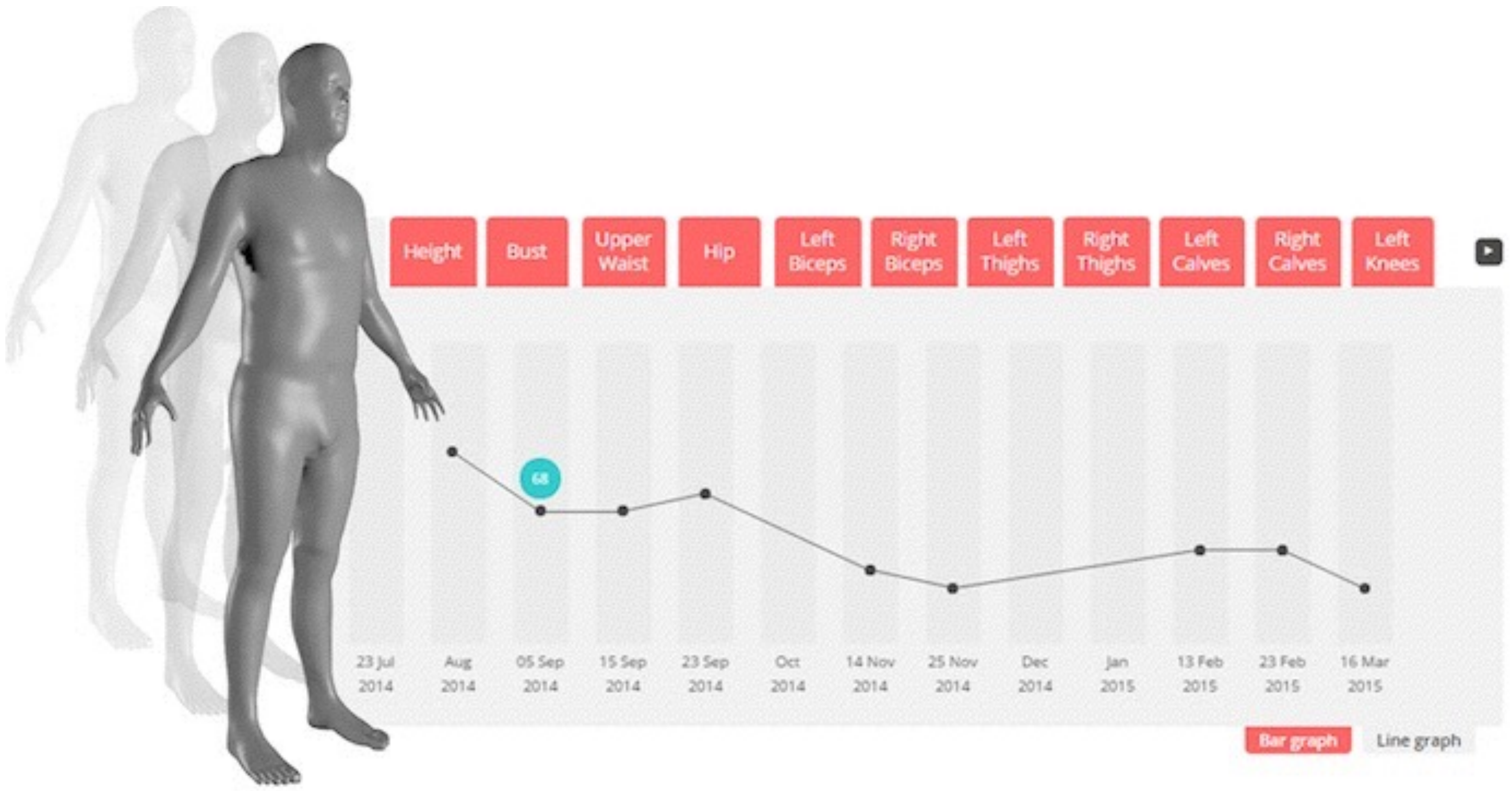
ToFit

Human-aware editing

Adobe

Why?

Healthcare



mPort

Action recognition / Surveillance

Challenges - Difficult without Scanner

- Existing methods recovering 3D body shape rely mostly on complex scanners

Max Planck Institute, Germany

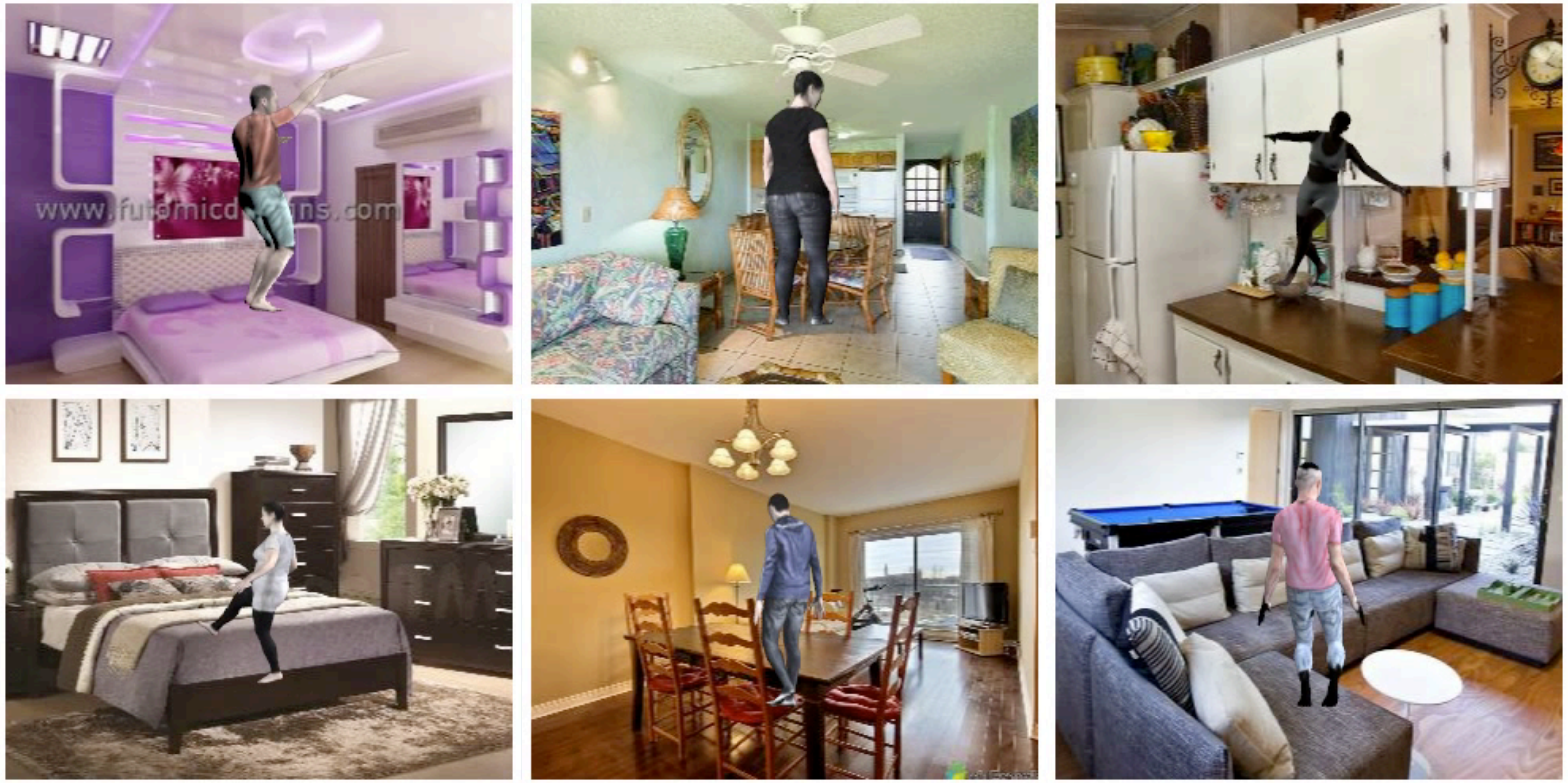
Challenges - Lack of Data

- Full 3D shape data is available either in **constrained settings…**

Ionescu et al. Human3.6M: Large Scale Datasets and Predictive Methods for
3D Human Sensing in Natural Environments, TPAMI 2014

Challenges - Lack of Data

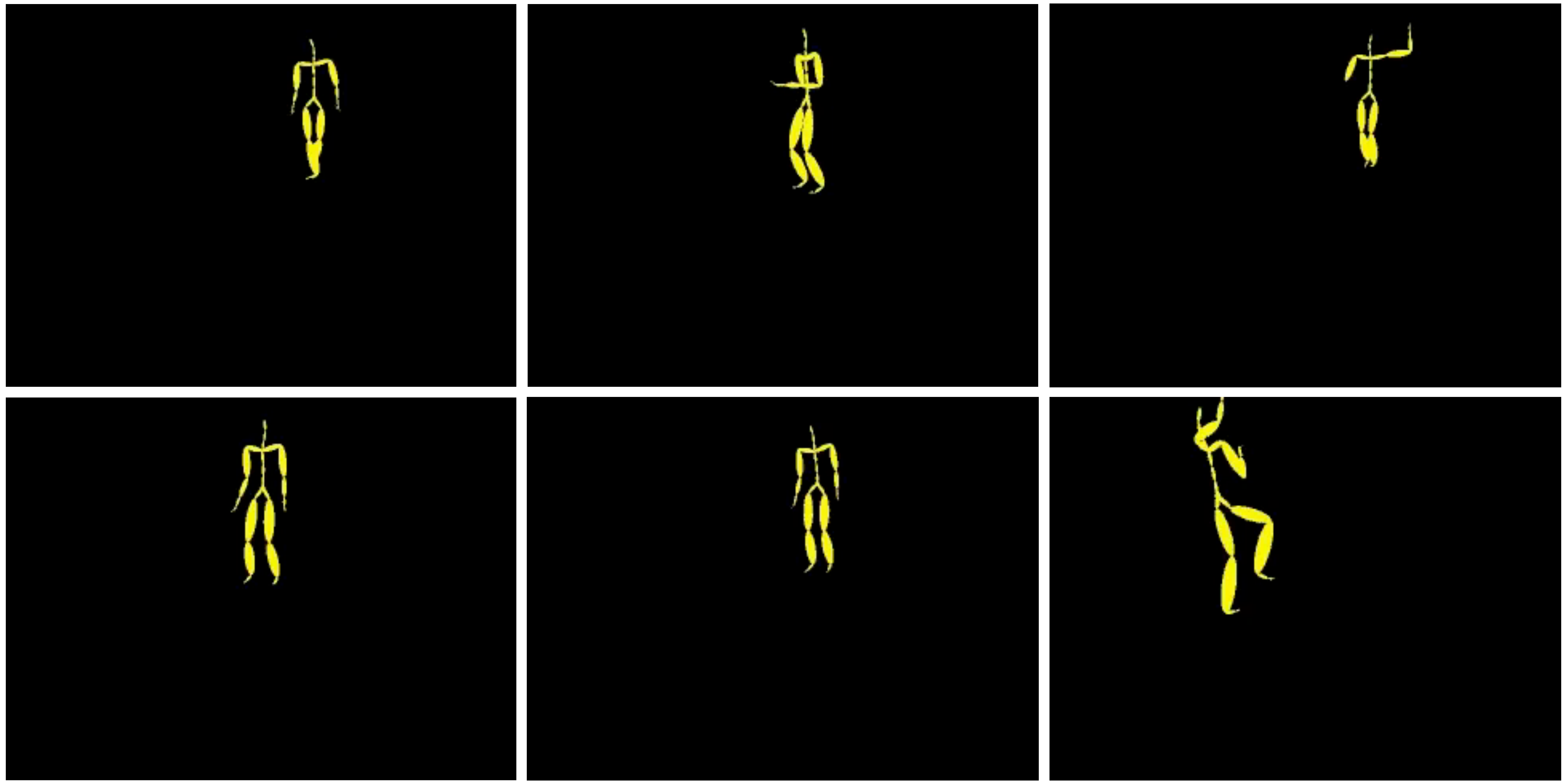
- Full 3D shape data is available either in constrained settings **or synthetic**



Varol et al. Learning from Synthetic Humans, CVPR 2017

Challenges - Lack of Data

- Most methods focus on predicting a skeleton

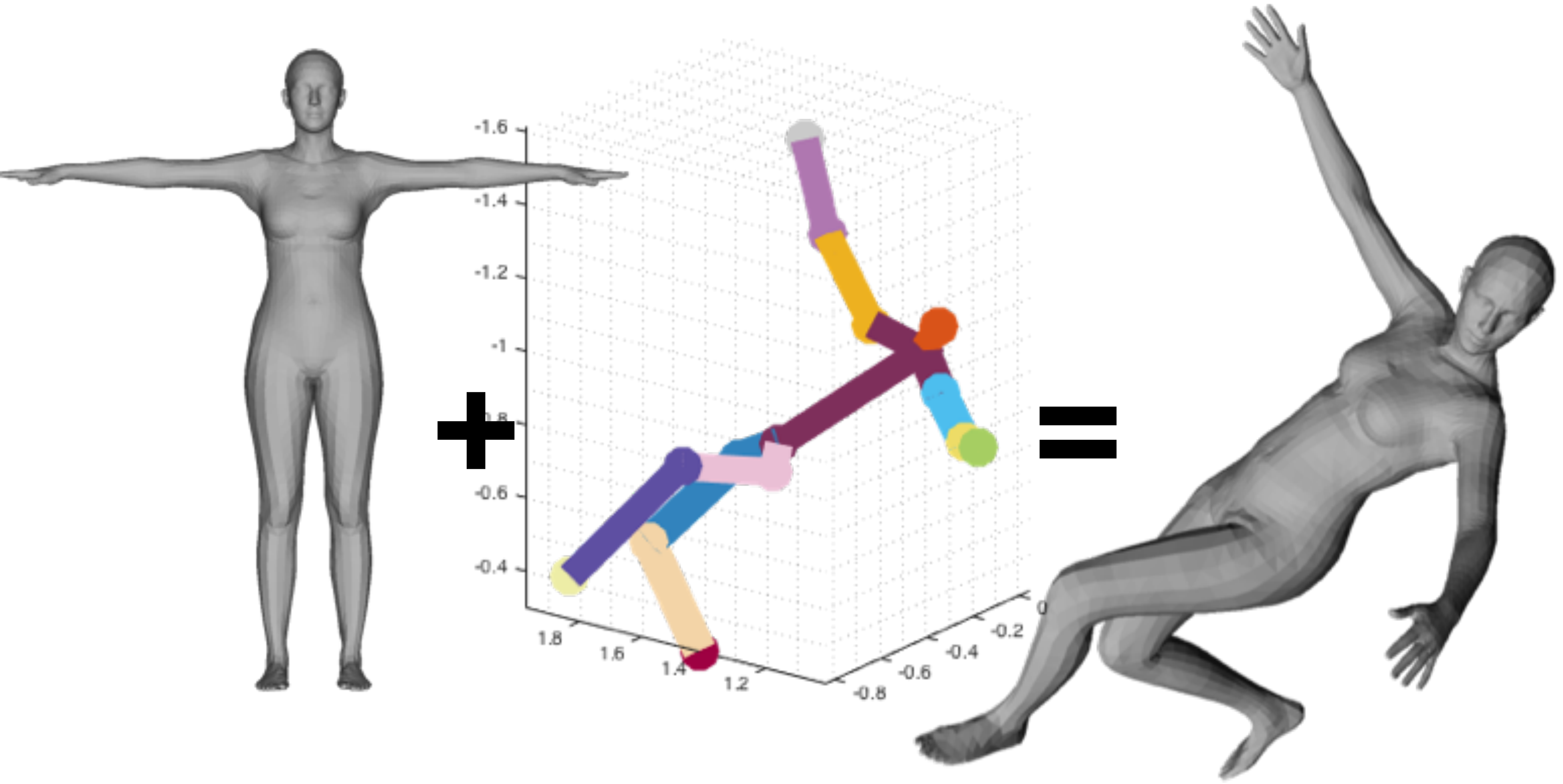


CMU MoCap Database

Challenges - Representation

10 + 72 = 82

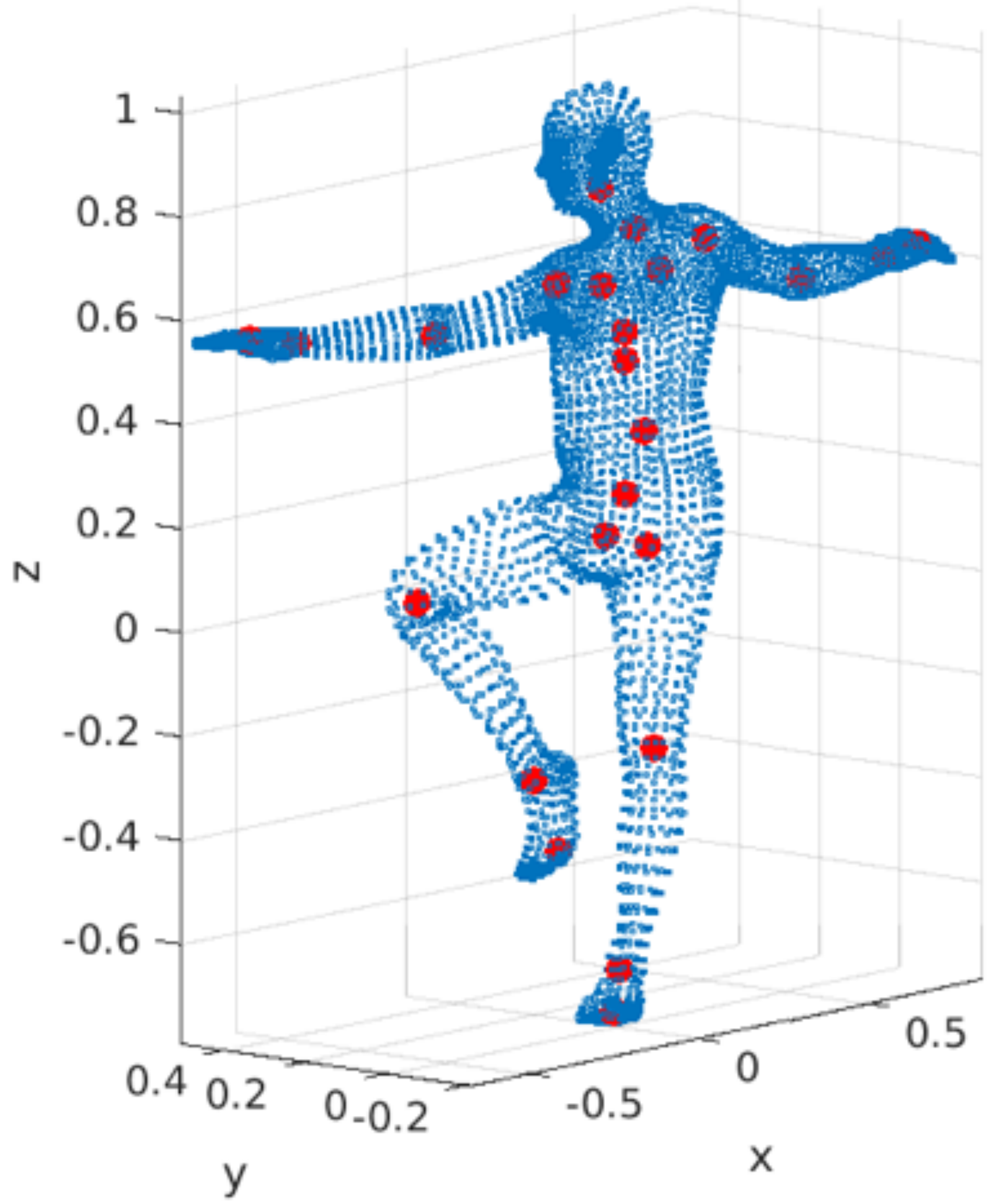
Parametric representation



Loper et al. [1]

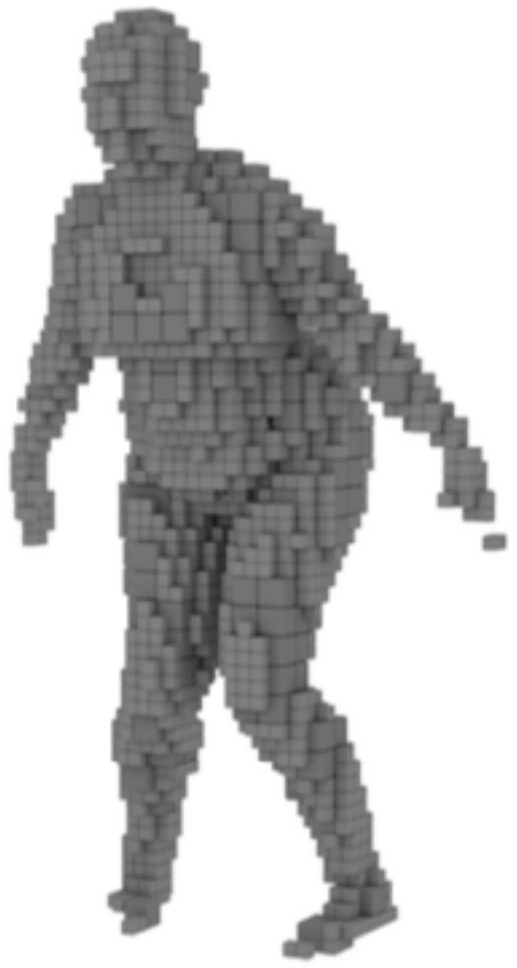
6890 x 3 = 20670

Point cloud representation



$64^3 = 262144$

Voxel representation



Tatarchenko et al. [2]

[1] Loper et al. SMPL: A Skinned Multi-Person Linear Model, SIGGRAPH Asia 2015

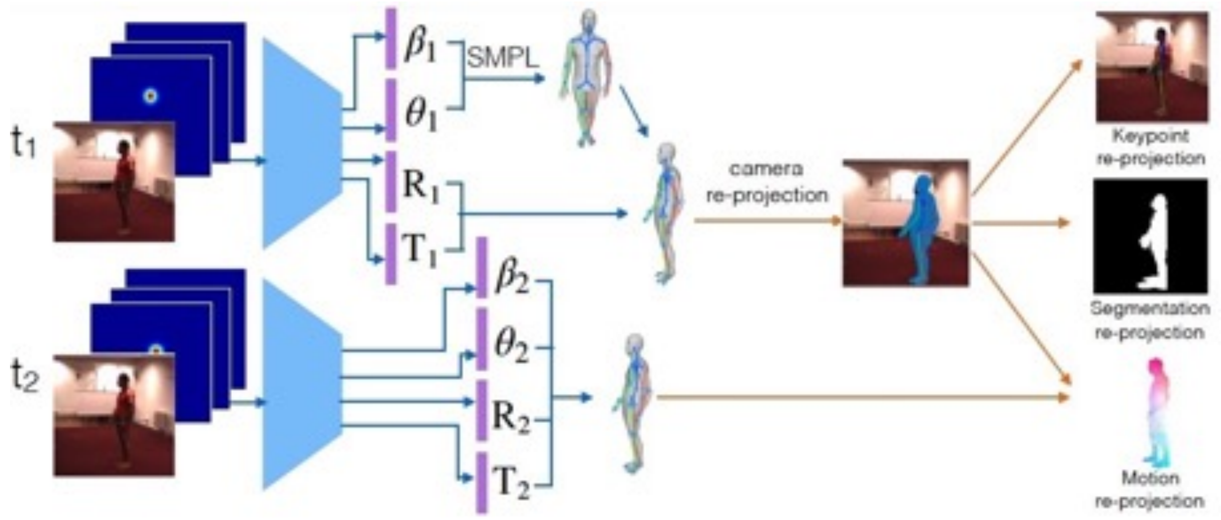
[2] Tatarchenko et al. Octree Generating Networks: Efficient Convolutional Architectures for High-resolution 3D Outputs, ICCV 2017

3D Human Body Shape Representation in the literature

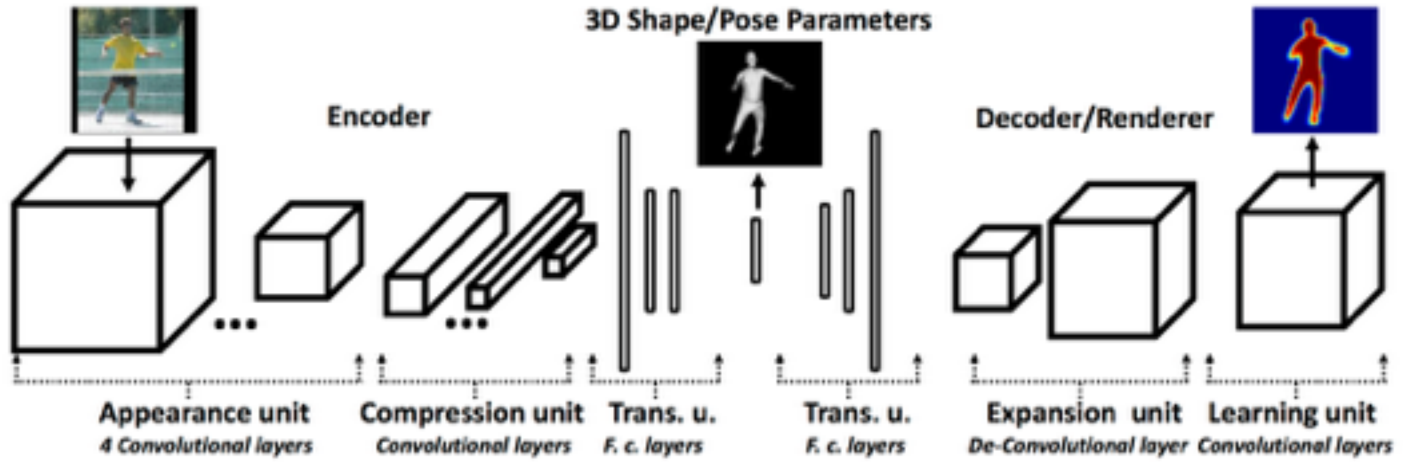
SMPL
(Skinned Multi-Person Linear Model)

SMPLify - Bogo et al. ECCV 2016

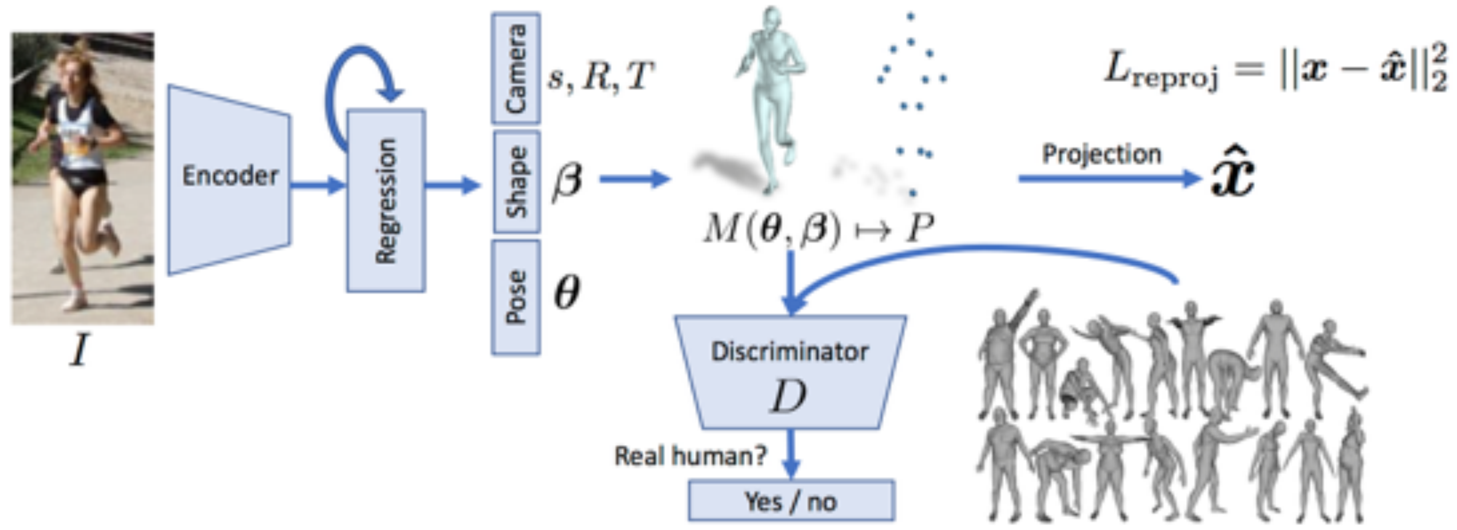
Unite the People - Lassner et al. CVPR 2017



Self-supervised learning - Tung et al. NIPS 2017

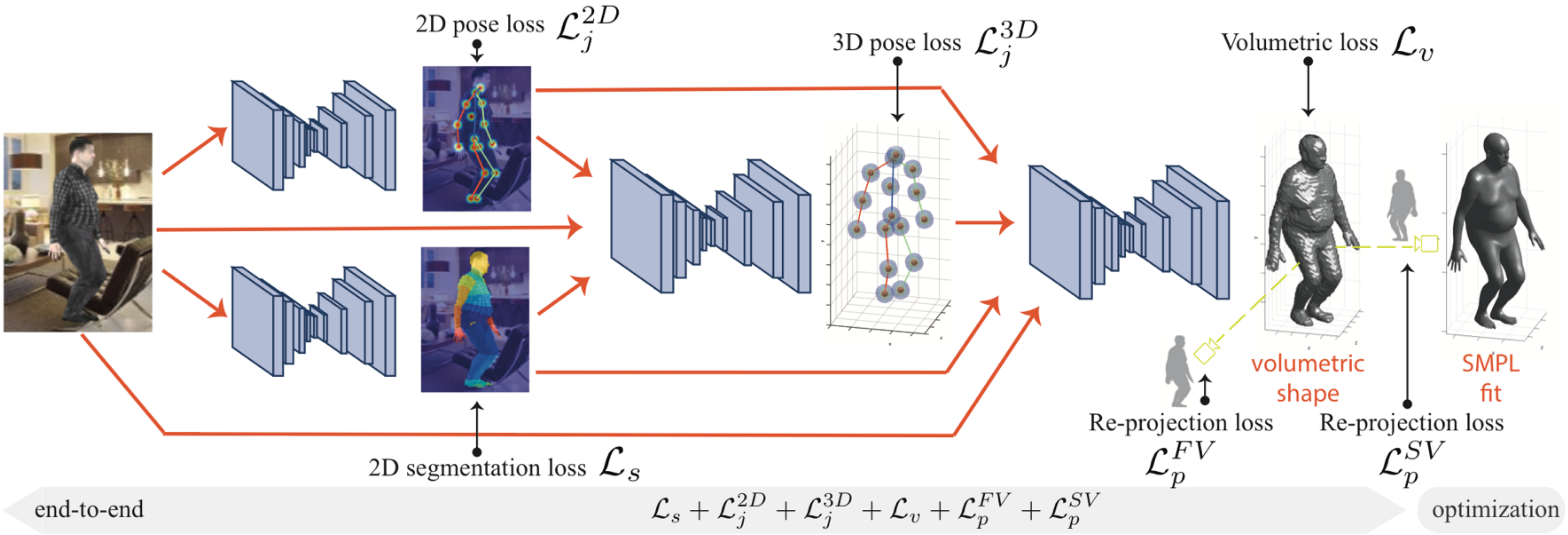


Indirect learning - Tan et al. BMVC 2017



HMR - Kanazawa et al. CVPR 2018

Proposed method: BodyNet



2D pose loss $\mathcal{L}_j^{2D}$

3D pose loss $\mathcal{L}_j^{3D}$

Volumetric loss $\mathcal{L}_v$

volumetric shape

SMPL fit

Re-projection loss $\mathcal{L}_p^{FV}$

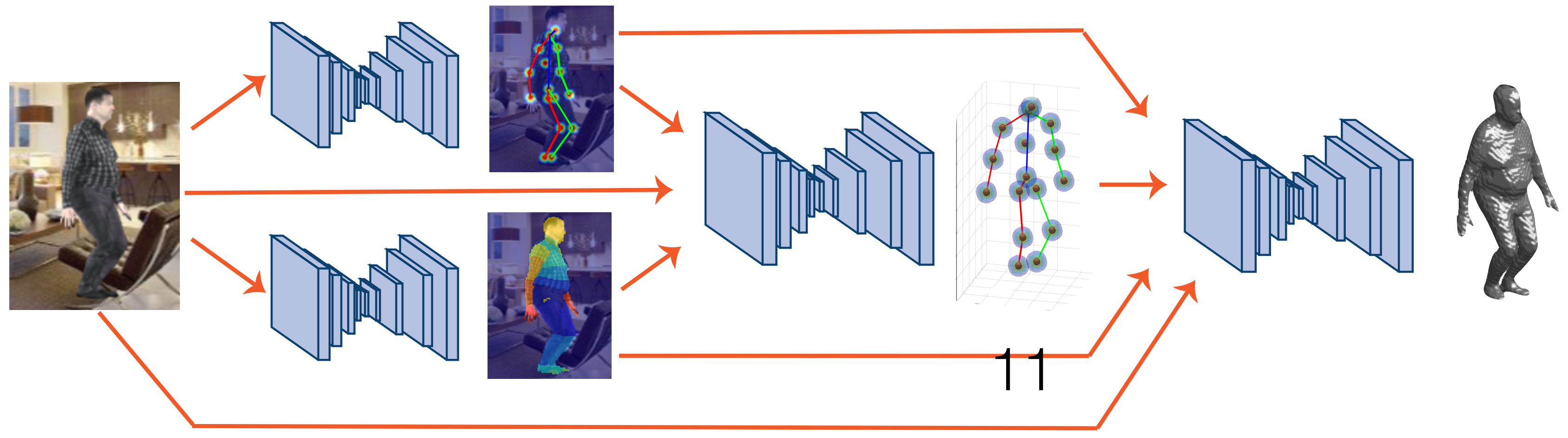
Re-projection loss $\mathcal{L}_p^{SV}$

2D segmentation loss $\mathcal{L}_s$

end-to-end

$$\mathcal{L}_s + \mathcal{L}_j^{2D} + \mathcal{L}_j^{3D} + \mathcal{L}_v + \mathcal{L}_p^{FV} + \mathcal{L}_p^{SV}$$

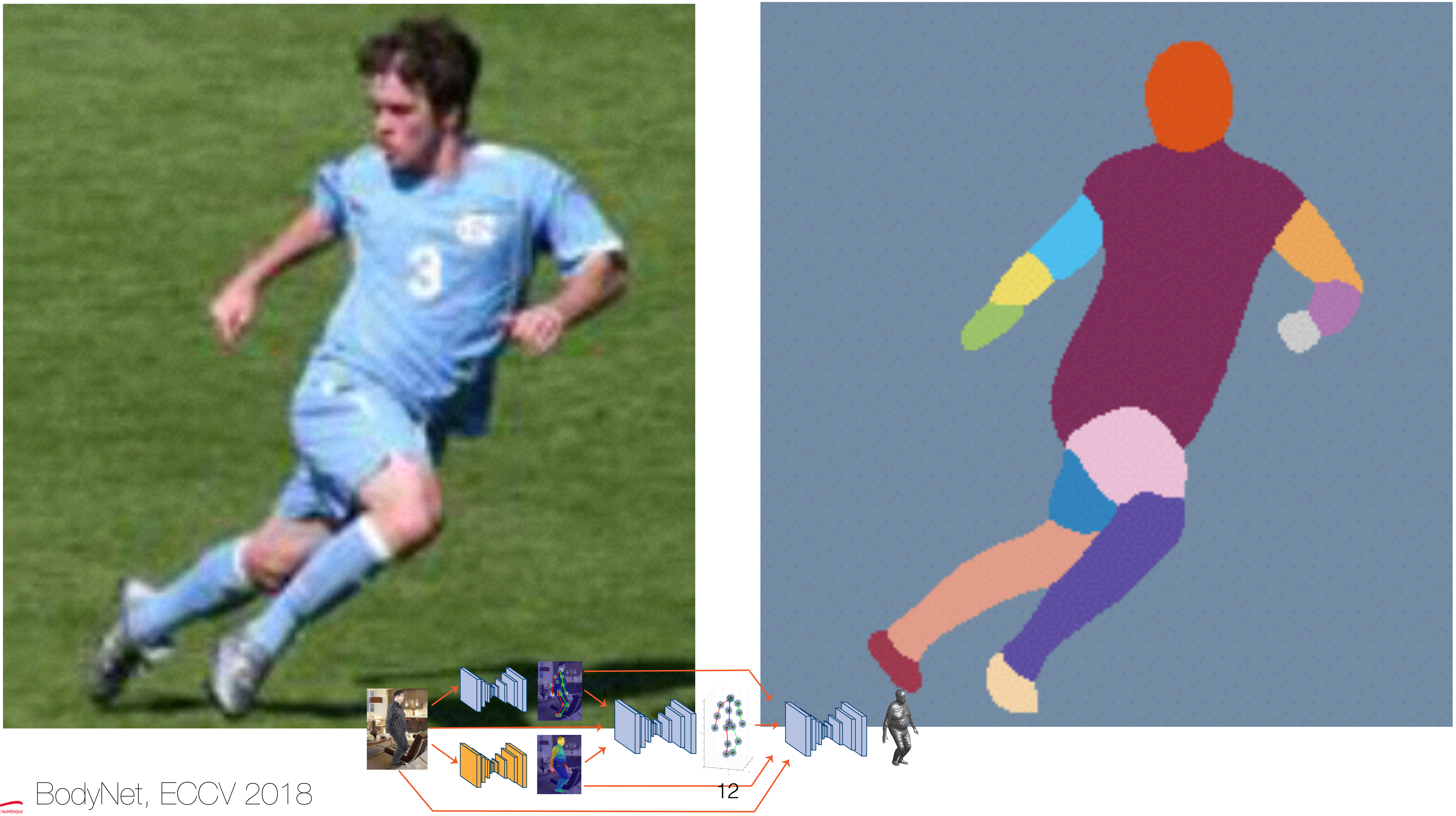
optimization



Subnetwork 1:

RGB input

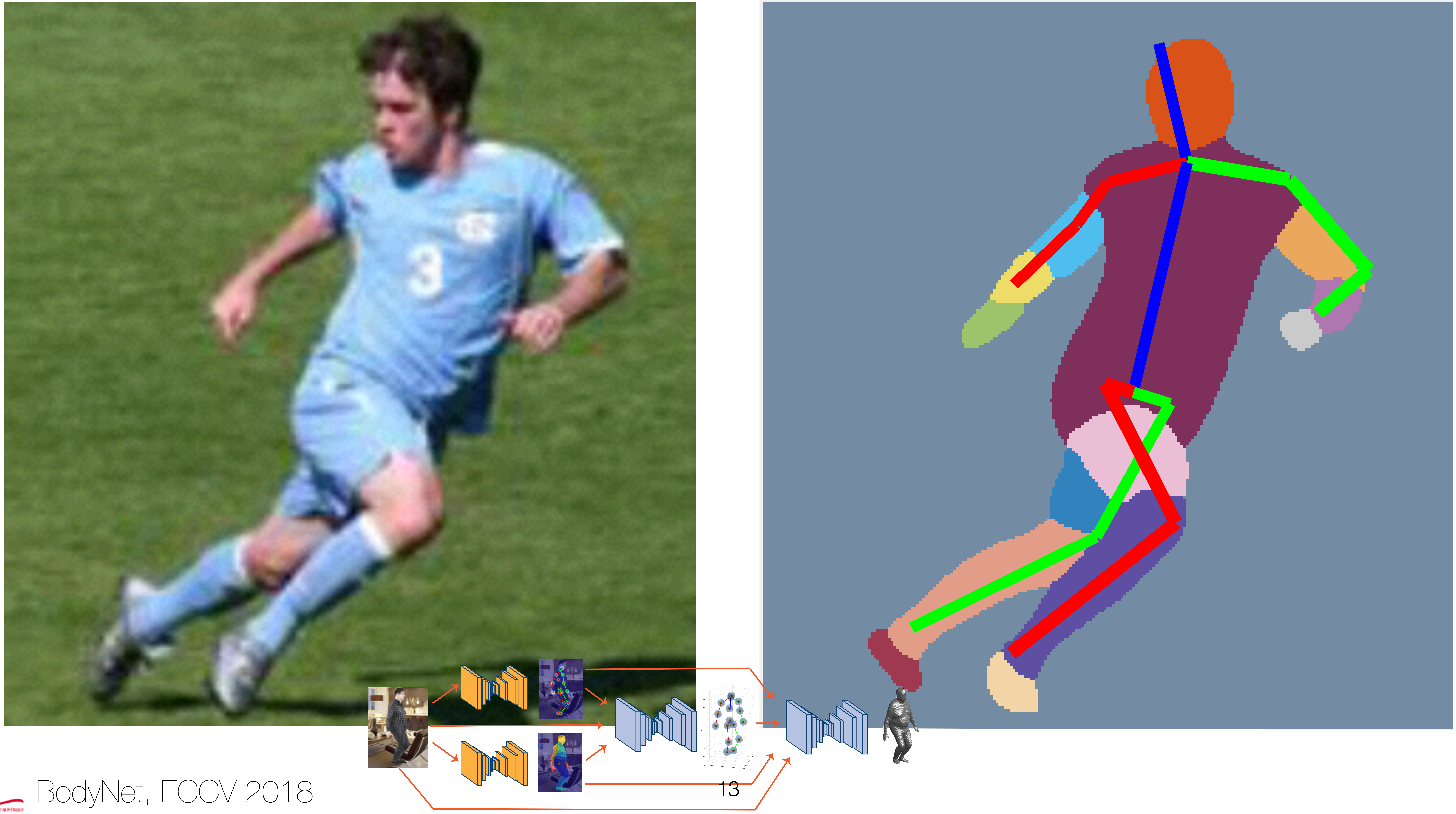
2D segmentation



Subnetwork 1&2:

2D segmentation & 2D pose predictions

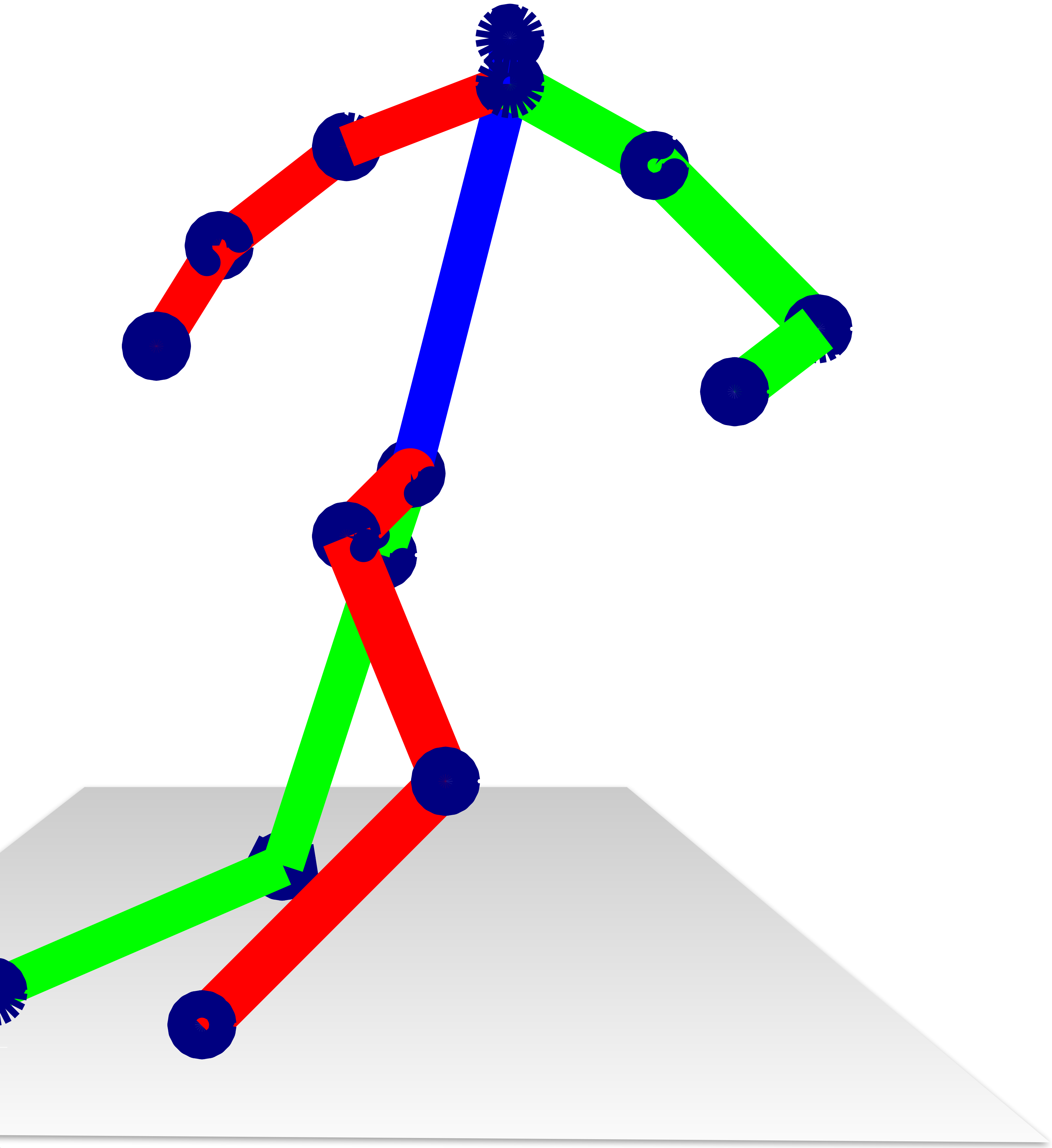
RGB input

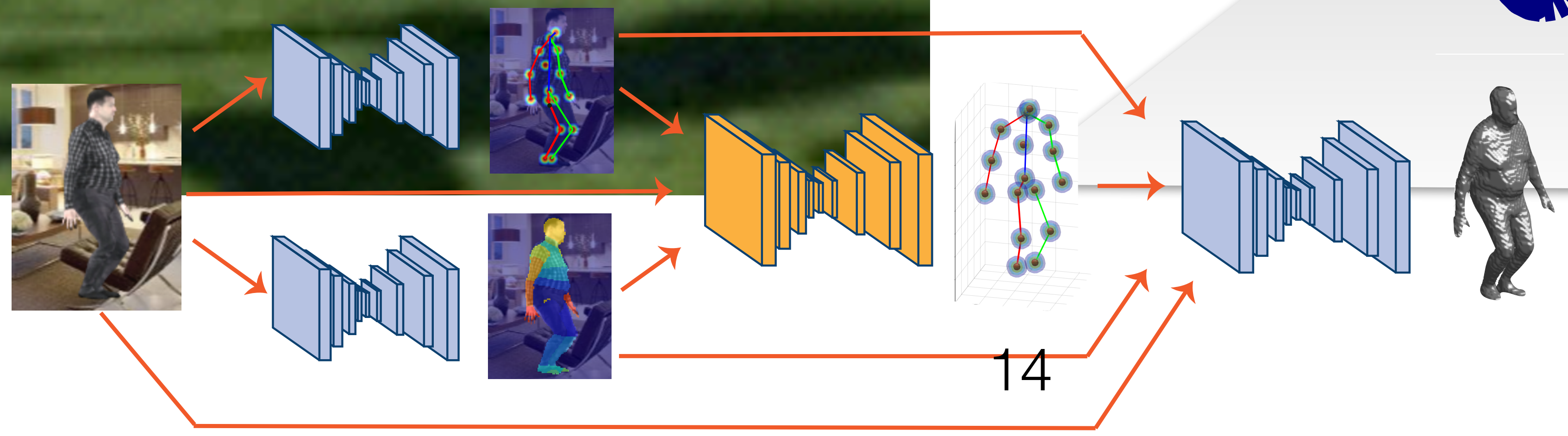


RGB input

Subnetwork 3:

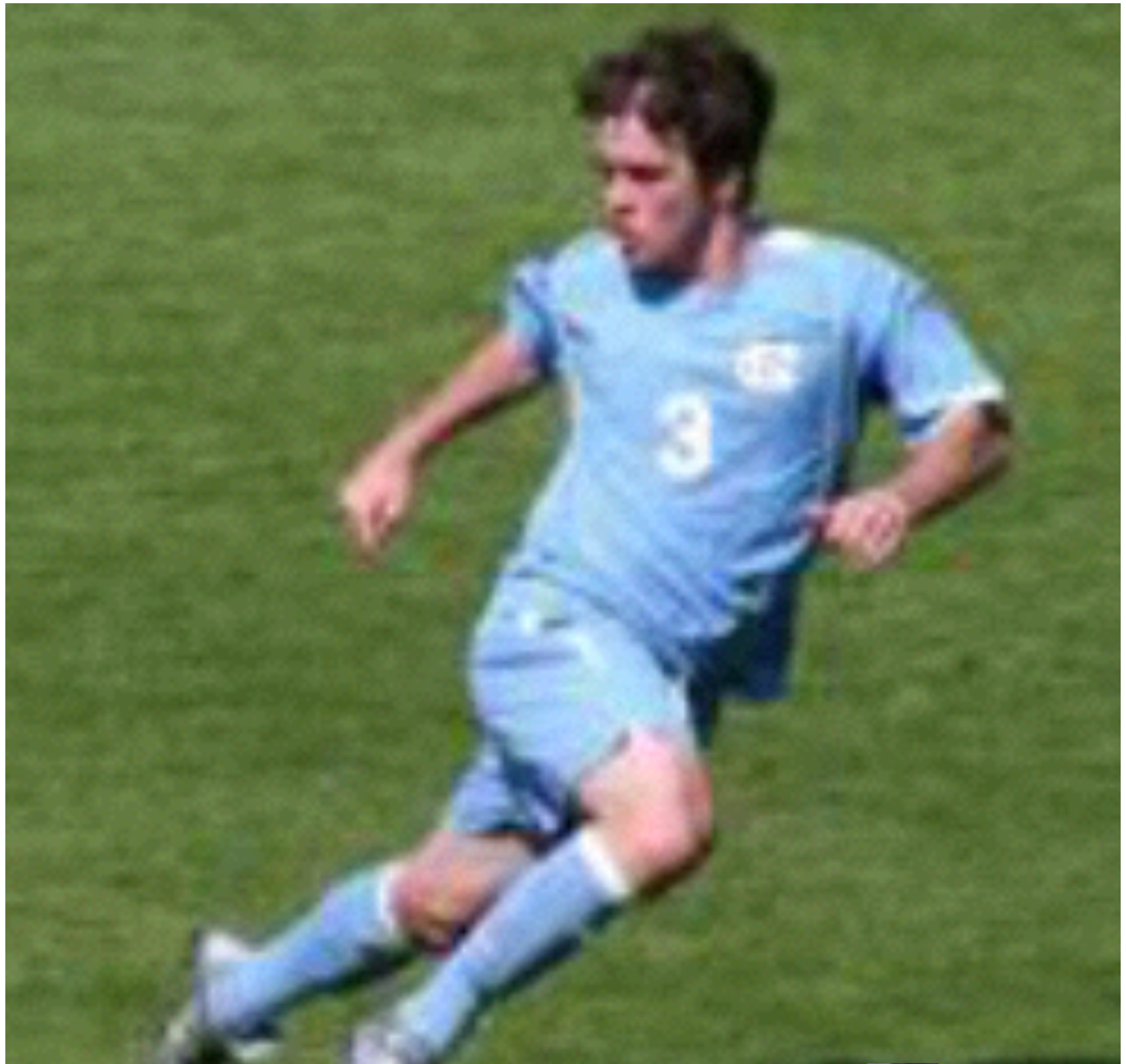
3D pose prediction



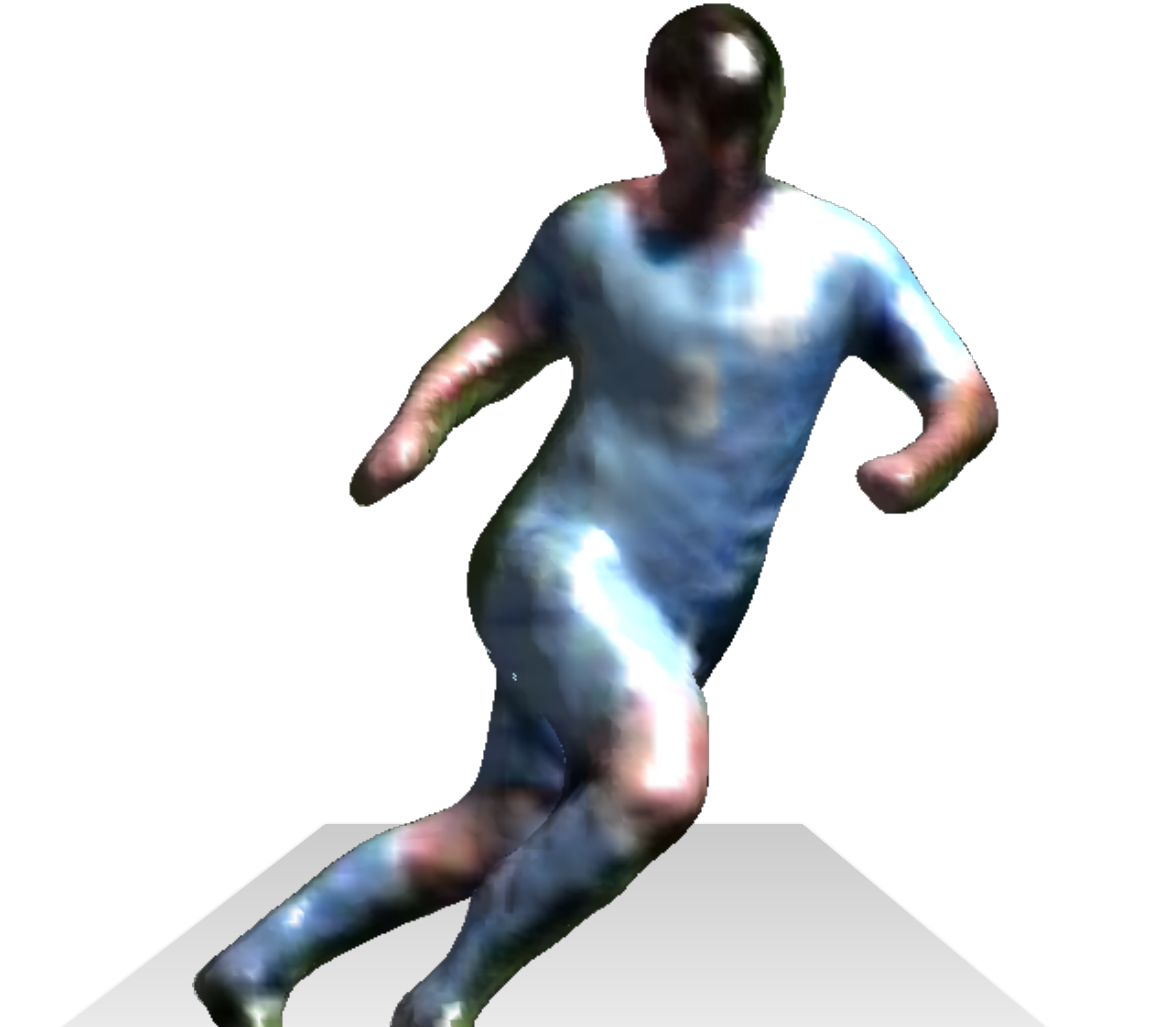


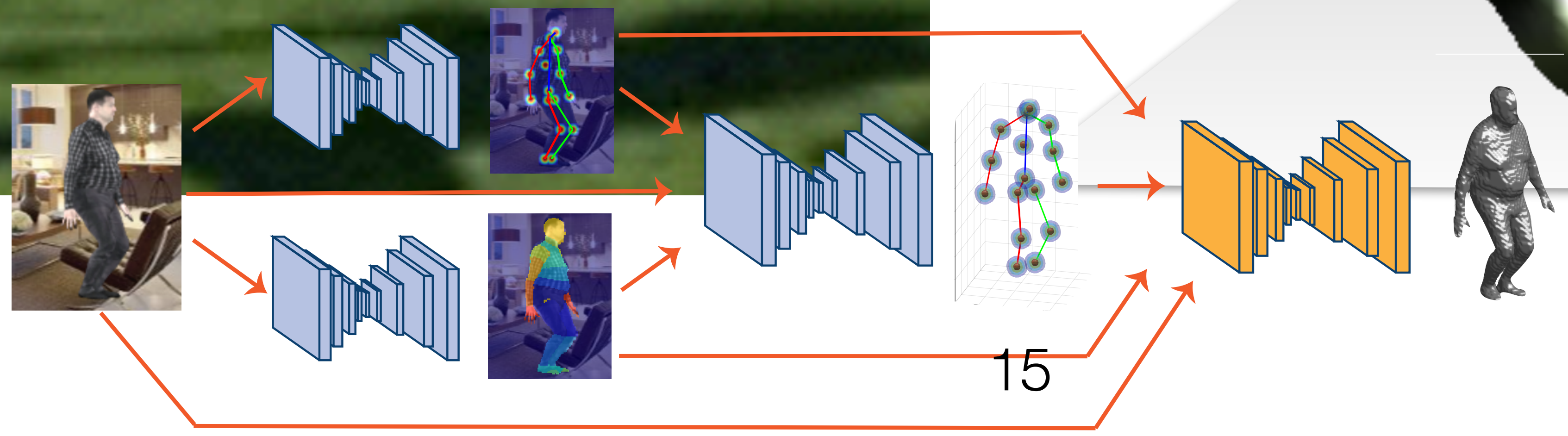
Subnetwork 4:

RGB input



Volumetric shape prediction

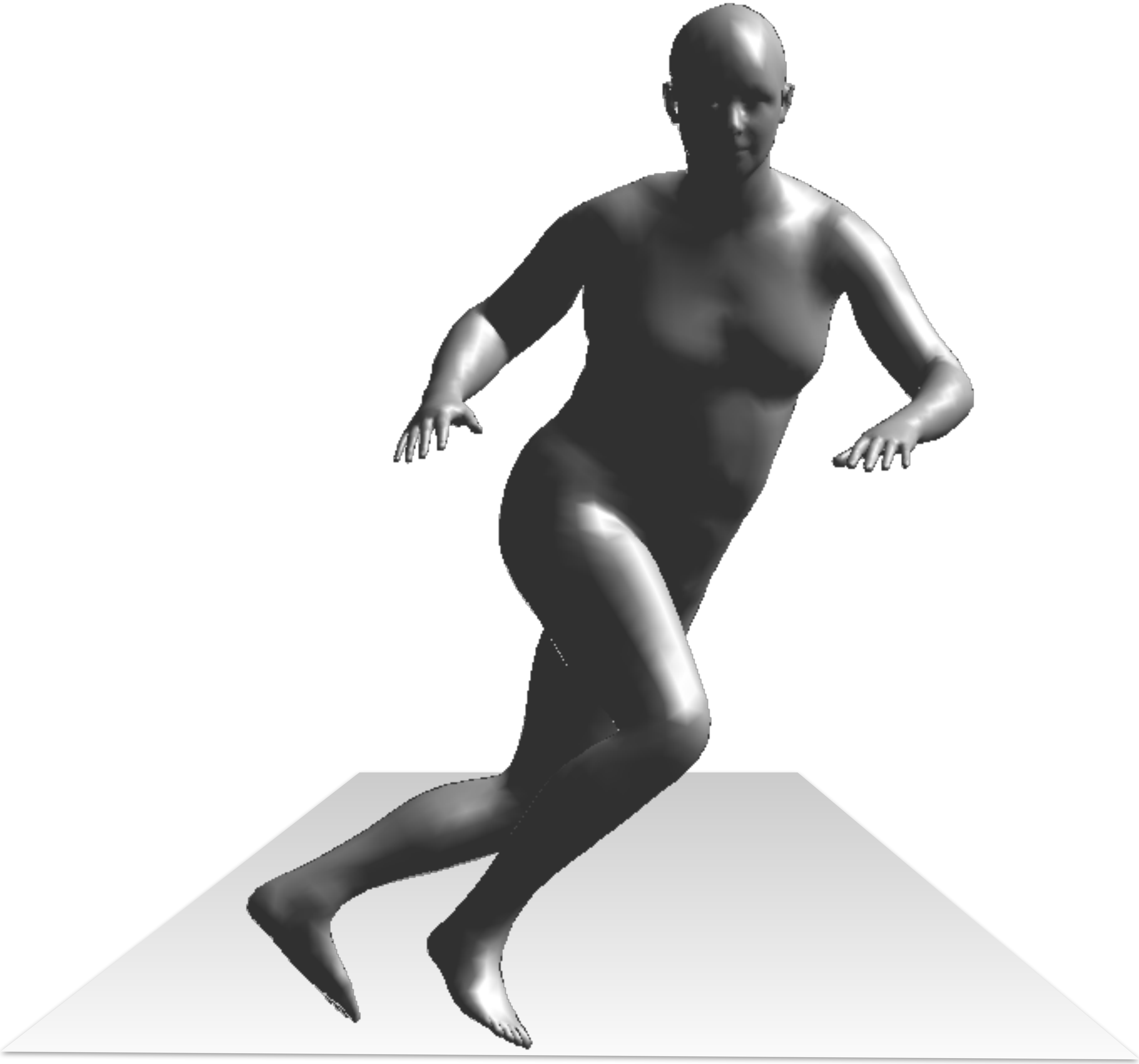




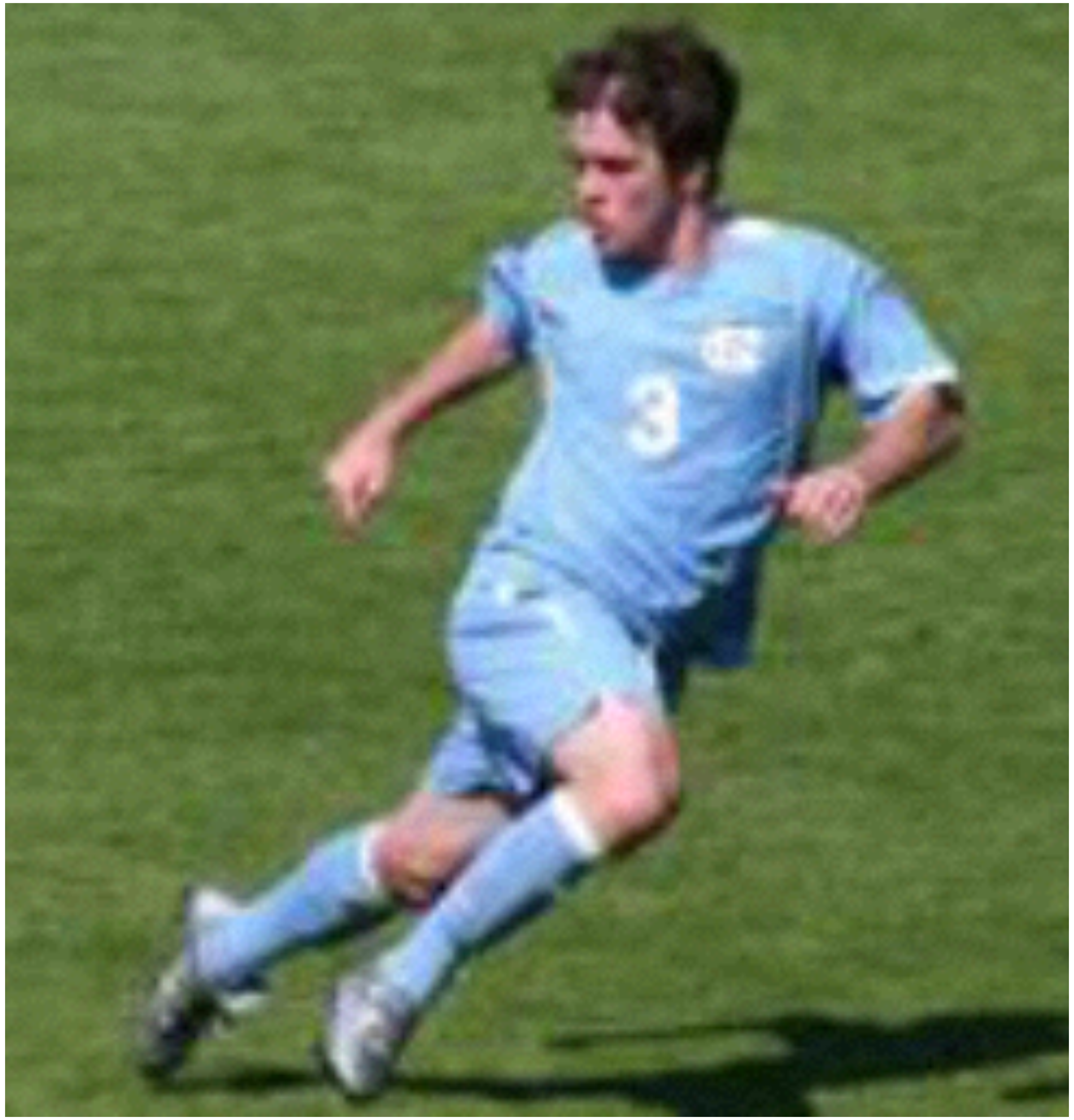
(Optional) Fitting:

RGB input

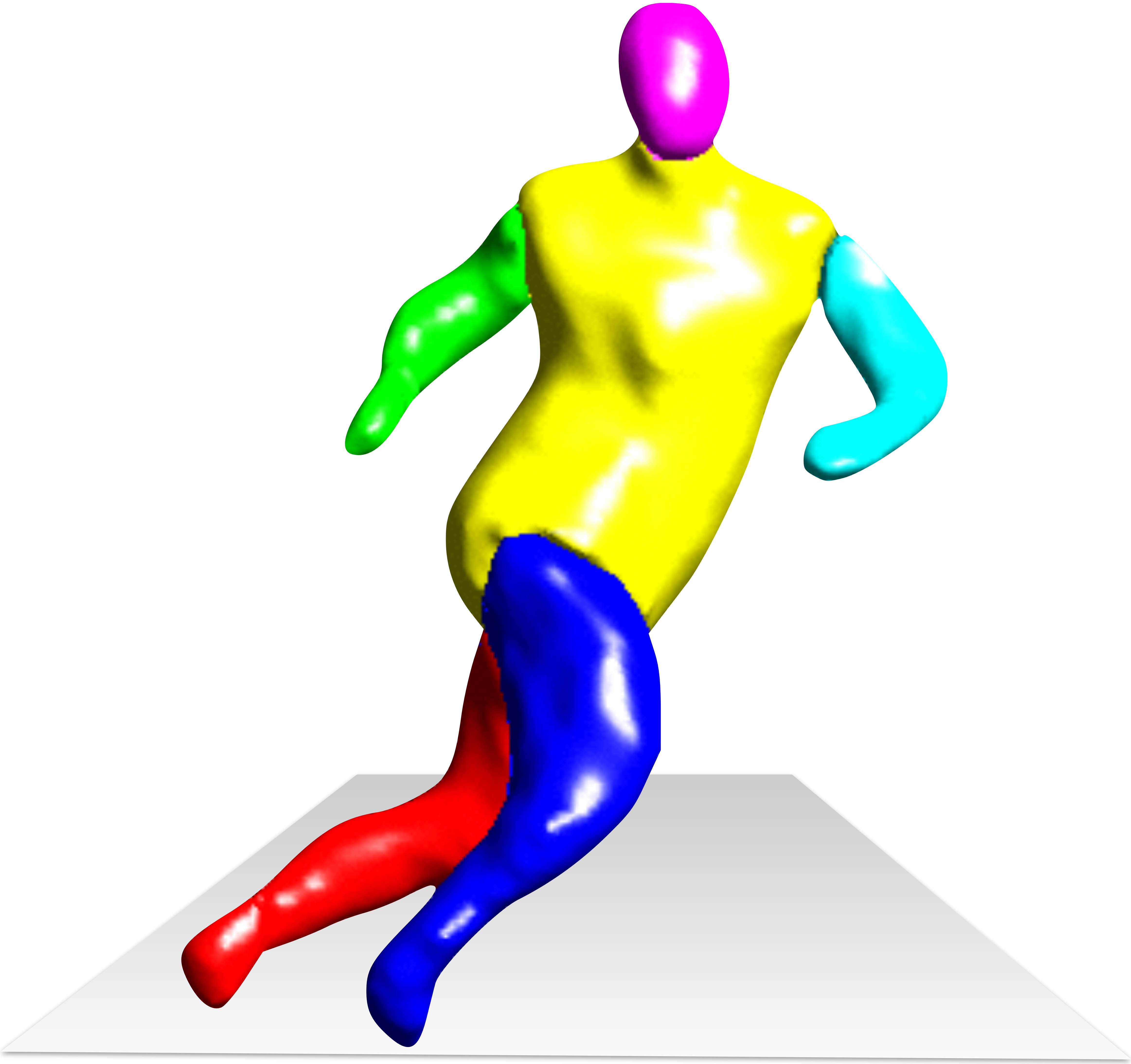
SMPL



RGB input

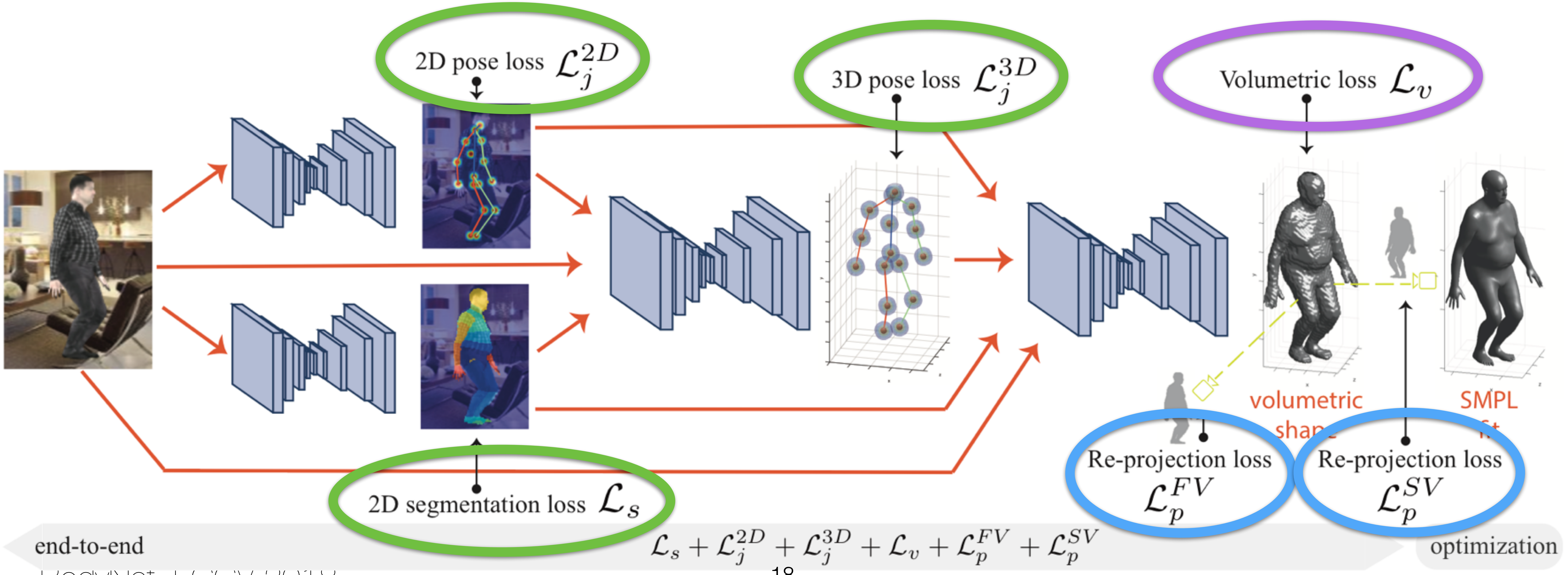


Output 3D body parts prediction



BodyNet is an end-to-end trainable network that benefits from:

- **a volumetric 3D loss,**
- **a multi-view re-projection loss,**
- **intermediate supervision of 2D pose, 2D body part segmentation, and 3D pose.**


2D pose loss $\mathcal{L}_j^{2D}$
3D pose loss $\mathcal{L}_j^{3D}$
Volumetric loss $\mathcal{L}_v$
volumetric shape
SMPL fit
Re-projection loss $\mathcal{L}_p^{FV}$
Re-projection loss $\mathcal{L}_p^{SV}$
2D segmentation loss $\mathcal{L}_s$
end-to-end
$\mathcal{L}_s + \mathcal{L}_j^{2D} + \mathcal{L}_j^{3D} + \mathcal{L}_v + \mathcal{L}_p^{FV} + \mathcal{L}_p^{SV}$
optimization

- **volumetric 3D loss**

$$\mathcal{L}_v = \sum_{x=1}^{W}\sum_{y=1}^{H}\sum_{z=1}^{D} V_{xyz}\log \hat{V}_{xyz} + (1 - V_{xyz})\log(1 - \hat{V}_{xyz})$$

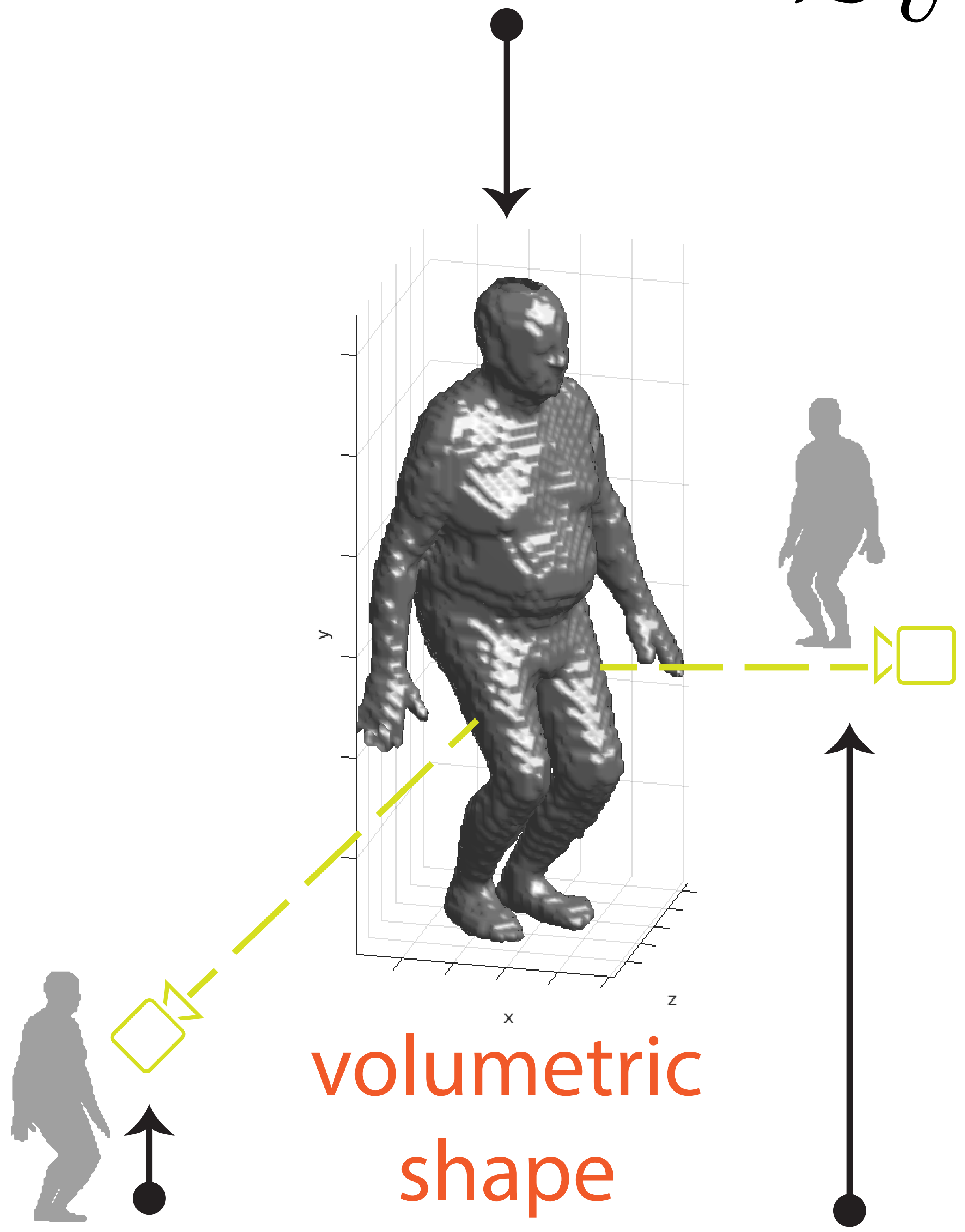
- **multi-view re-projection loss**

$$\hat{S}^{FV}(x,y) = \max_z \hat{V}_{xyz} \quad \text{and} \quad \hat{S}^{SV}(y,z) = \max_x \hat{V}_{xyz}.$$

$$\mathcal{L}_p^{FV} = \sum_{x=1}^{W}\sum_{y=1}^{H} S(x,y)\log \hat{S}^{FV}(x,y) + (1 - S(x,y))\log(1 - \hat{S}^{FV}(x,y)),$$

$$\mathcal{L}_p^{SV} = \sum_{y=1}^{H}\sum_{z=1}^{D} S(y,z)\log \hat{S}^{SV}(y,z) + (1 - S(y,z))\log(1 - \hat{S}^{SV}(y,z)).$$

Volumetric loss $\mathcal{L}_v$



volumetric shape

Re-projection loss $\mathcal{L}_p^{FV}$

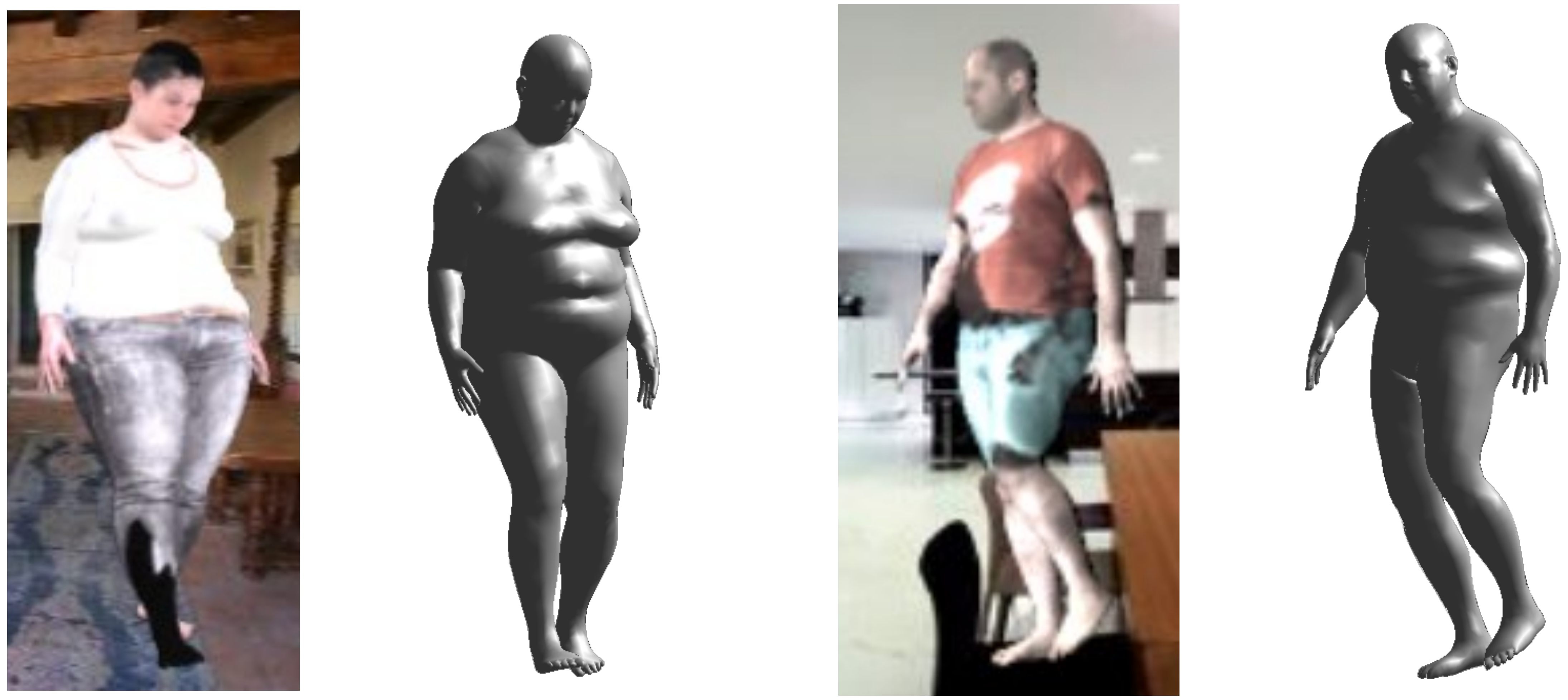
Re-projection loss $\mathcal{L}_p^{SV}$

SMPL fitting

$$\{\theta^\star, \beta^\star\} = \underset{\{\theta,\beta\}}{\operatorname{argmin}} \sum_{\mathbf{p}^n \in \mathbf{V}^n} \min_{\mathbf{p}^s \in \mathbf{V}^s(\theta,\beta)} w^n \|\mathbf{p}^n - \mathbf{p}^s\|_2^2 +$$

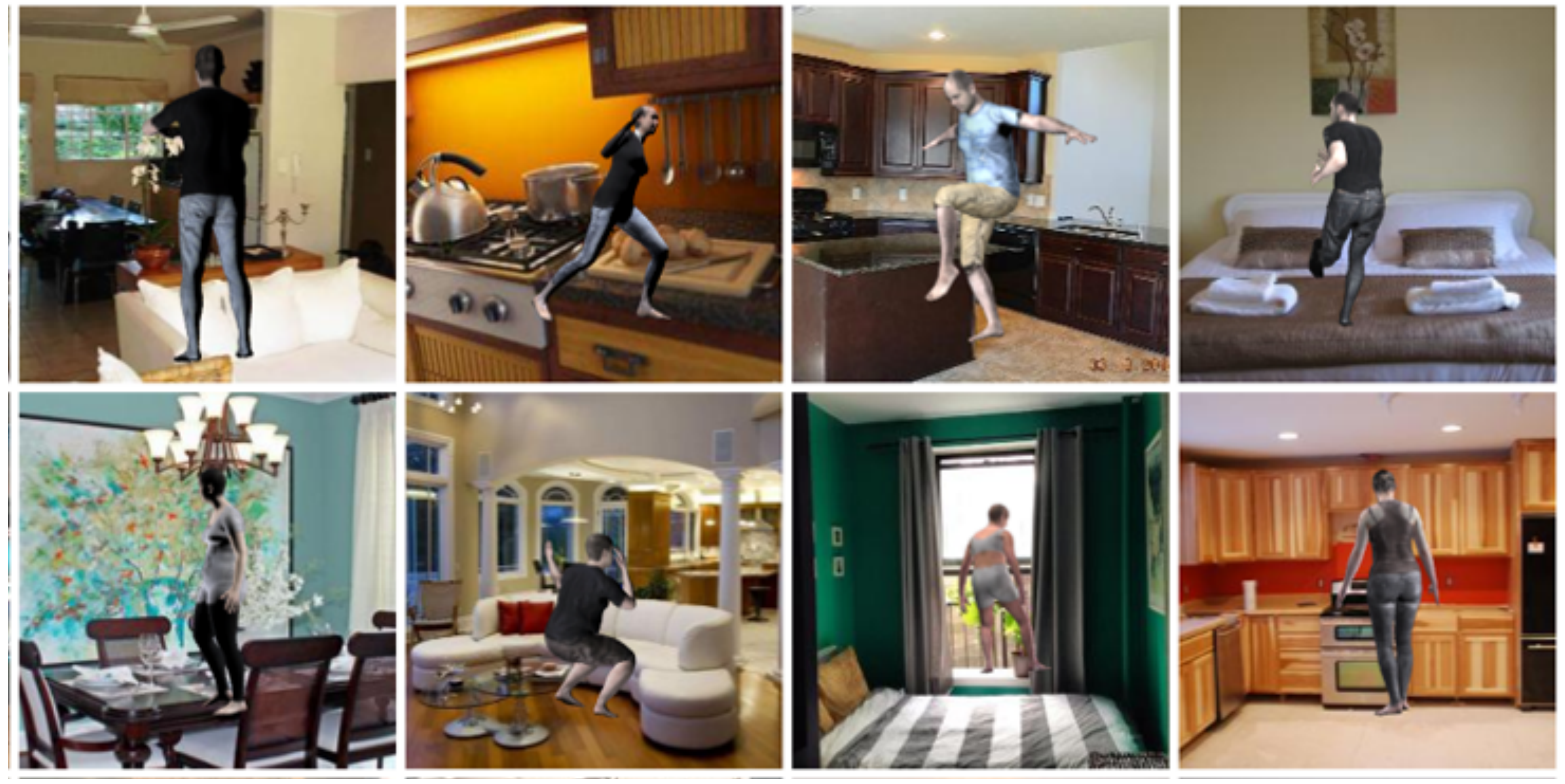
$$\sum_{\mathbf{p}^s \in \mathbf{V}^s(\theta,\beta)} \min_{\mathbf{p}^n \in \mathbf{V}^n} w^n \|\mathbf{p}^n - \mathbf{p}^s\|_2^2 + \lambda \sum_{i=1}^{J} \|\mathbf{j}_i^n - \mathbf{j}_i^s(\theta, \beta)\|_2^2.$$

θ	Pose parameter of SMPL
β	Shape parameter of SMPL
$\mathbf{p}^n$	Vertex coordinate predicted by the network
$\mathbf{p}^s$	Closest vertex coordinate of fitted SMPL
$\mathbf{j}_i^n$	3D joint coordinate predicted by the network
$\mathbf{j}_i^s$	3D joint coordinate of fitted SMPL

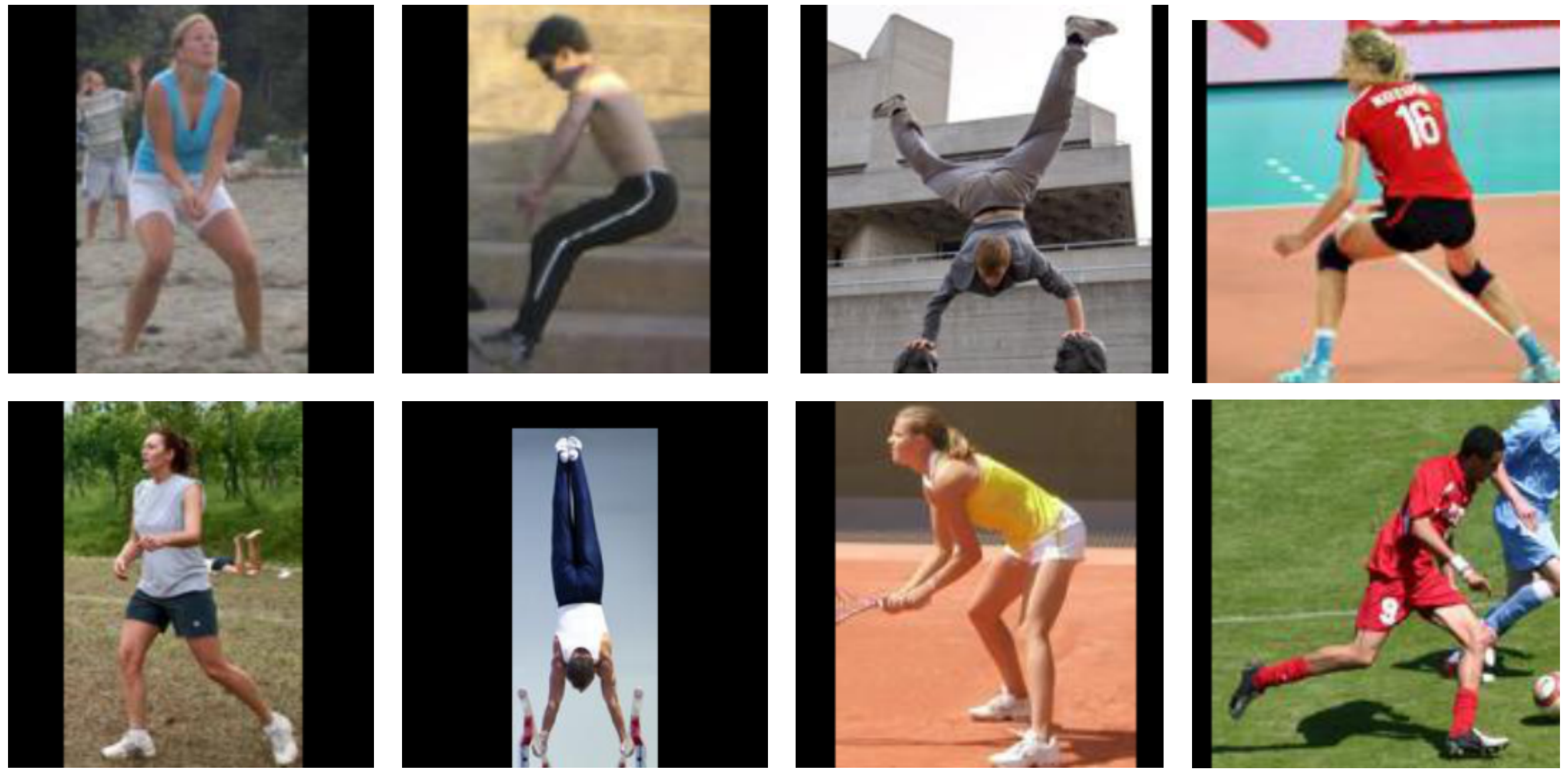


Experiments - Datasets

- SURREAL (Varol et al. CVPR 2017)
  - synthetically rendered images
  - perfect ground truth

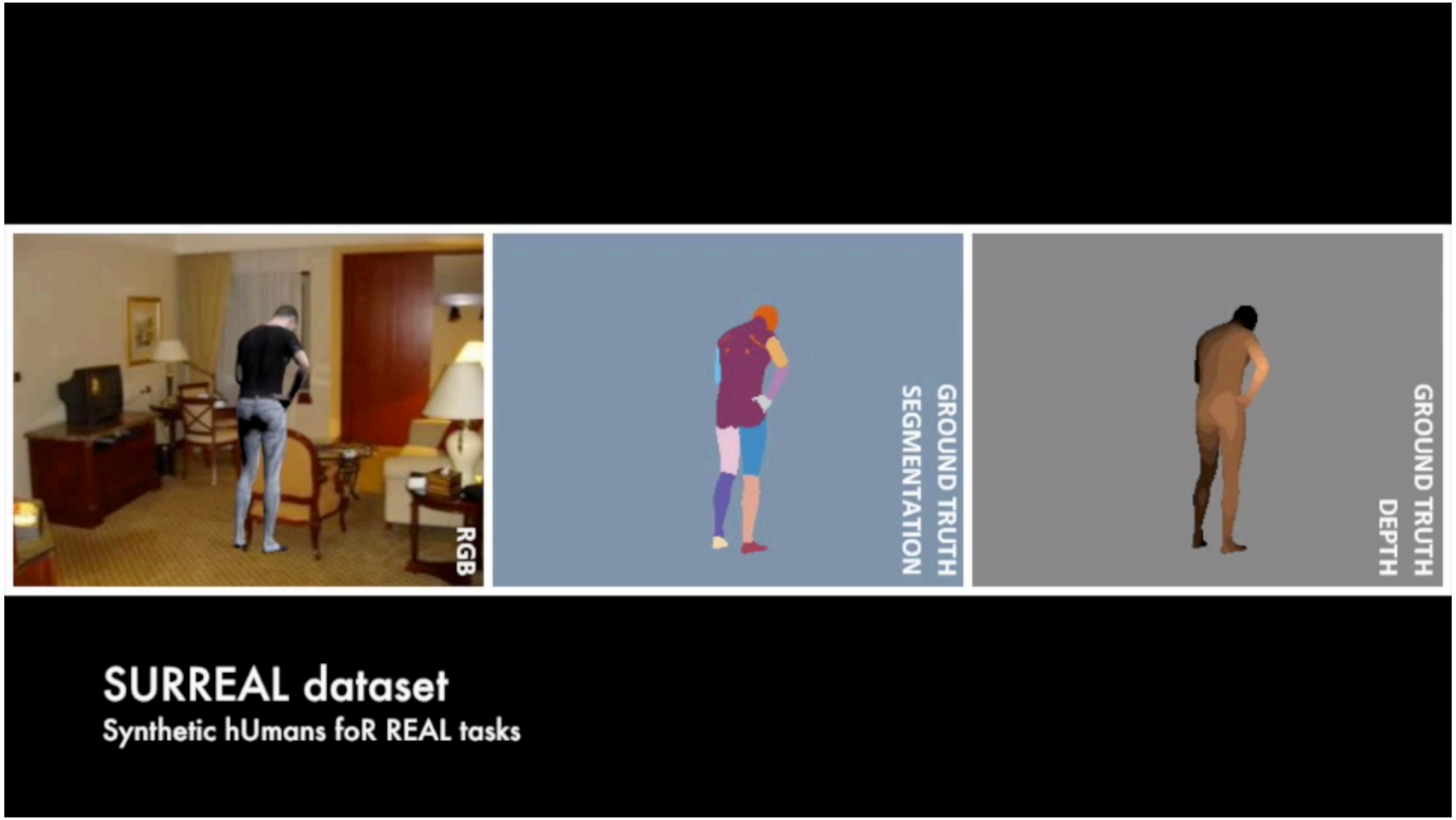


- Unite the People (Lassner et al. CVPR 2017)
  - real images
  - noisy 3D ground truth
  - manual 2D segmentation annotation



SURREAL Dataset

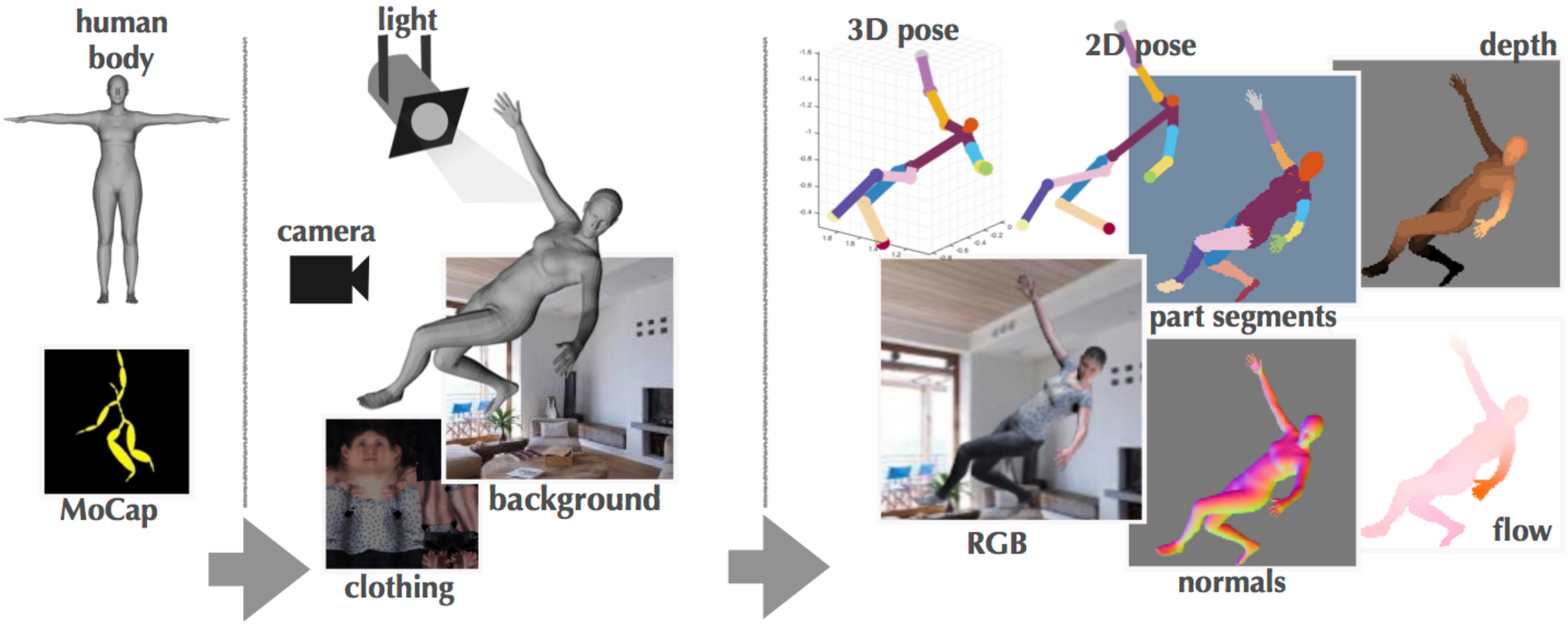
Synthetic hUmans foR REAL tasks



Varol et al. Learning from Synthetic Humans, CVPR 2017

SURREAL Dataset

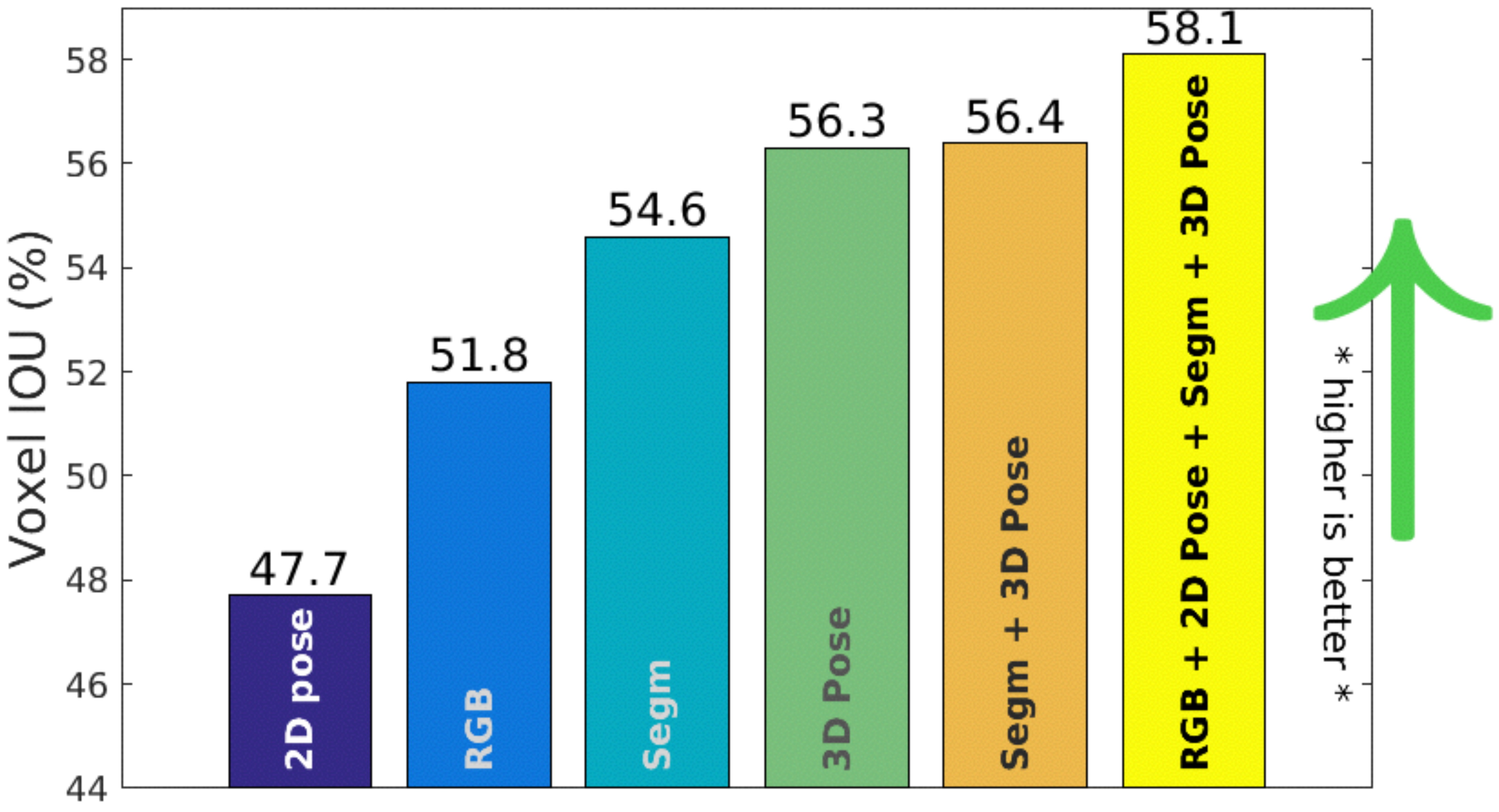
Synthetic hUmans foR REAL tasks

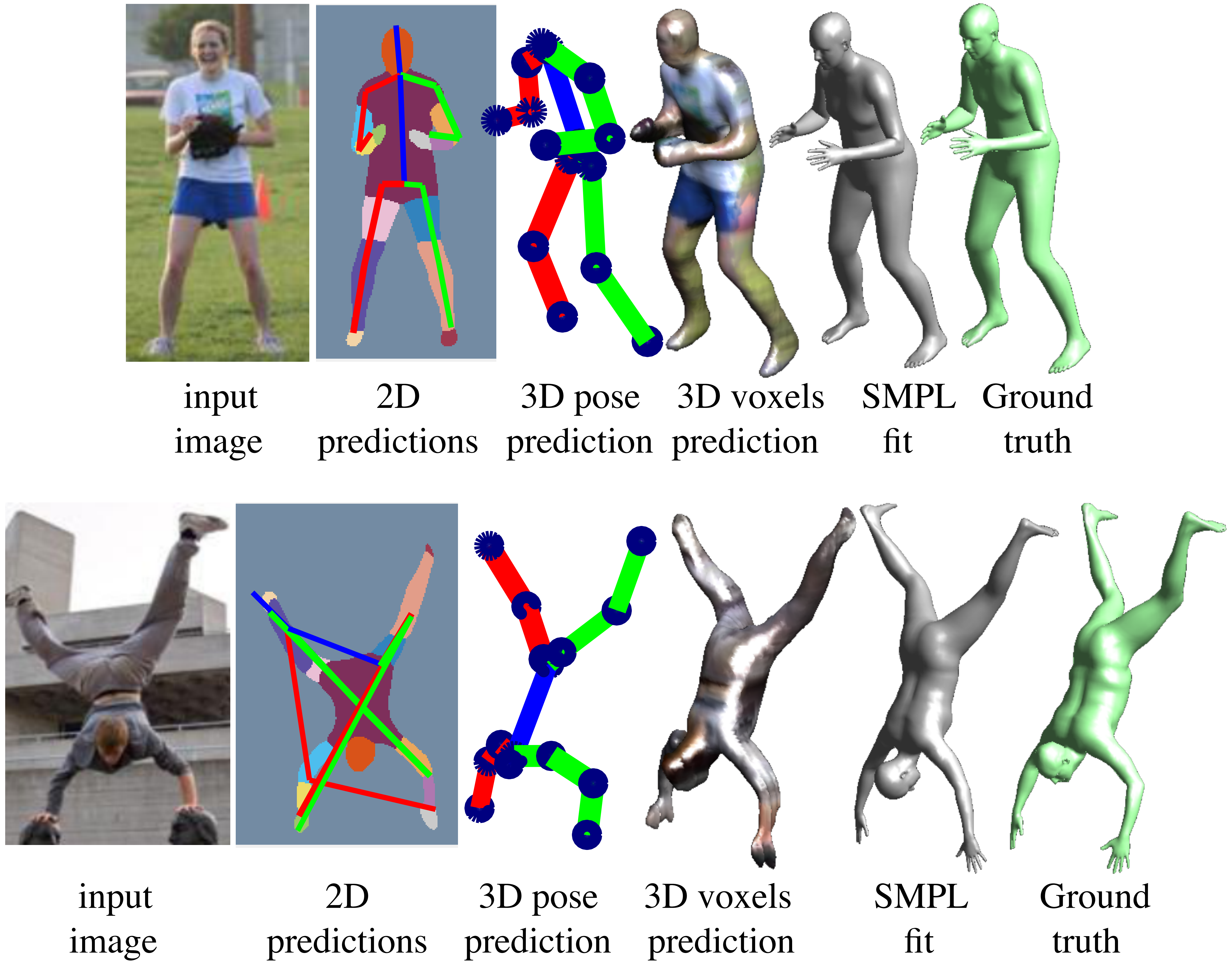


Varol et al. Learning from Synthetic Humans, CVPR 2017

Experiments

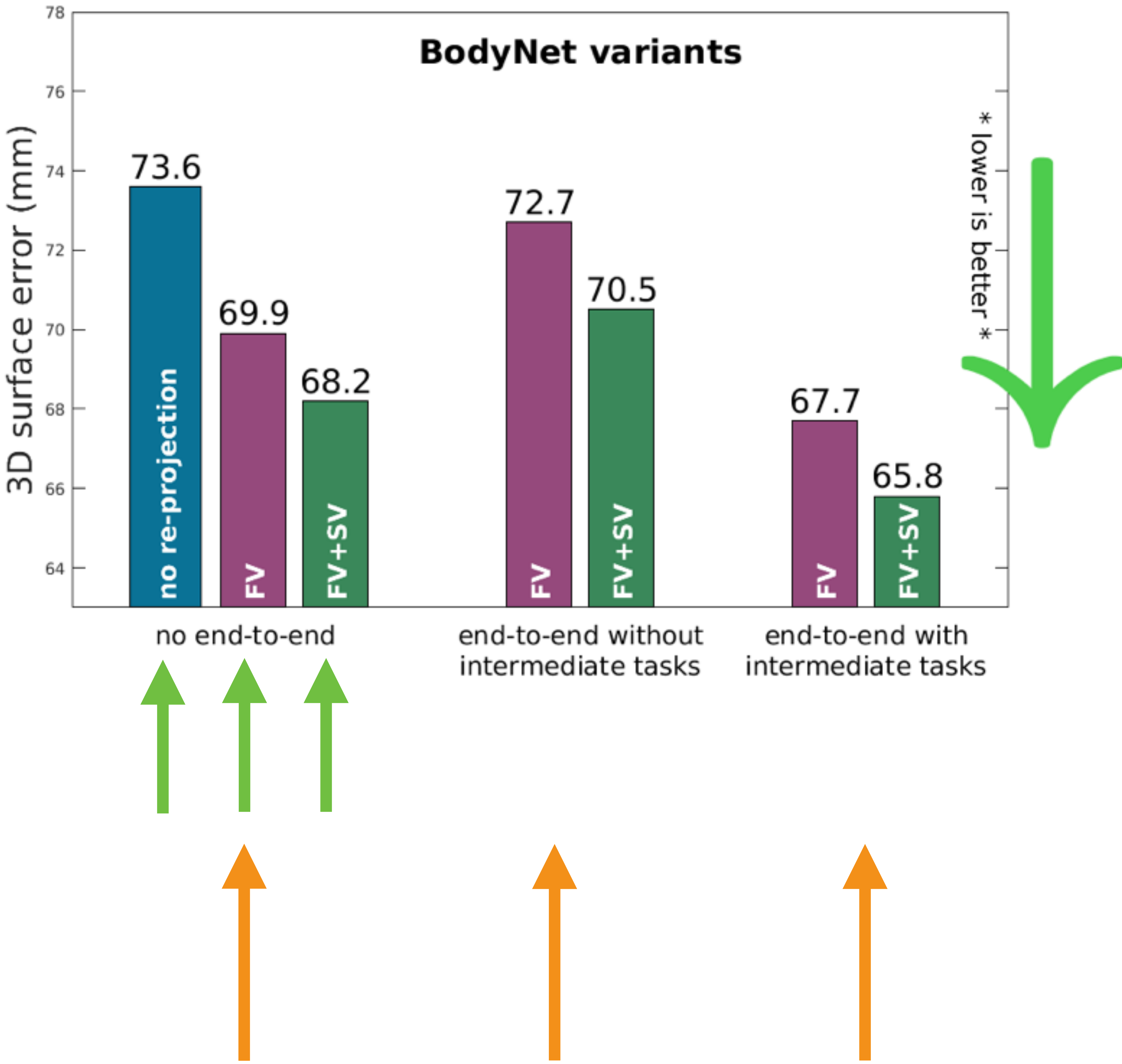
- Effect of additional inputs





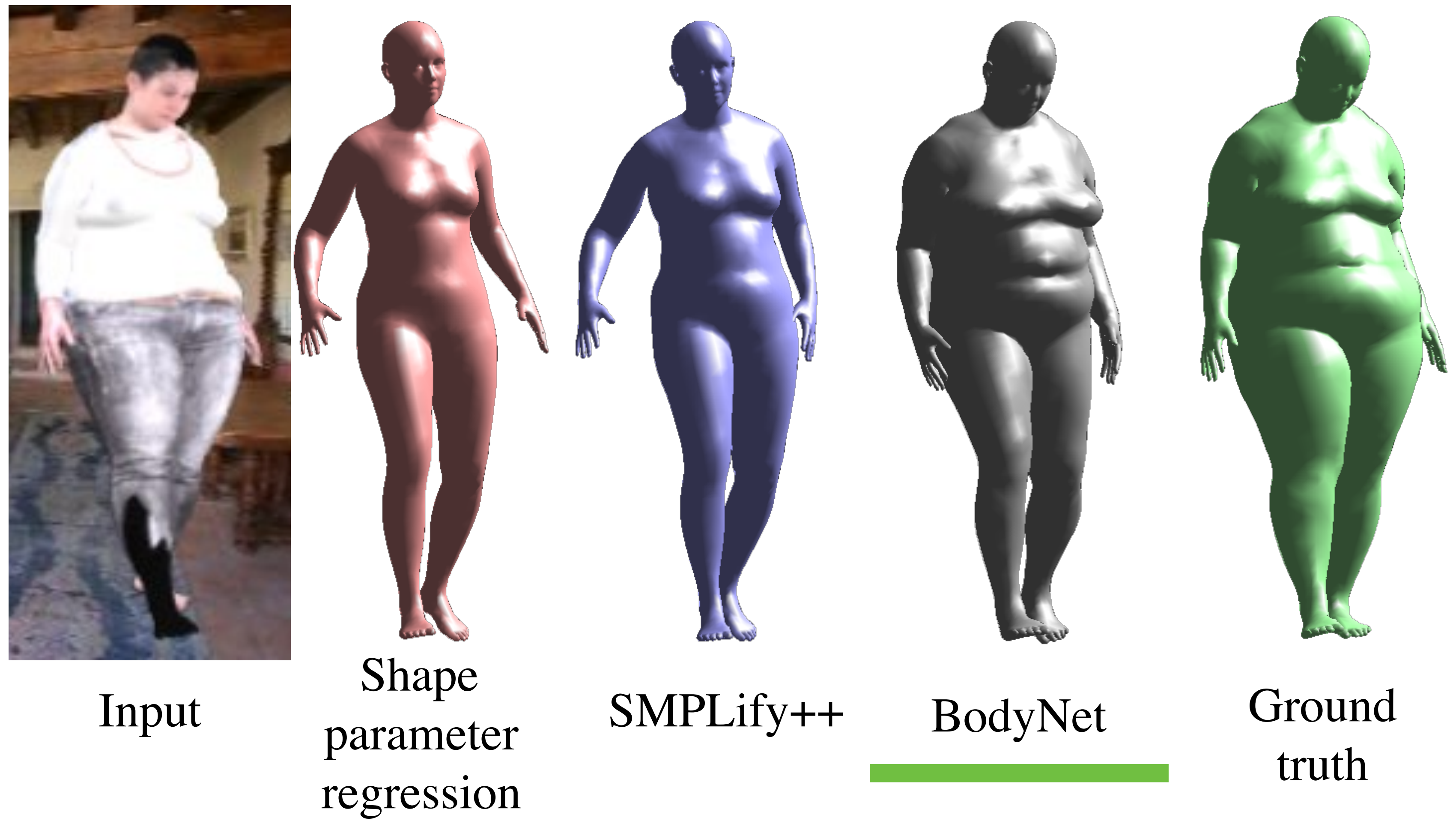
Experiments

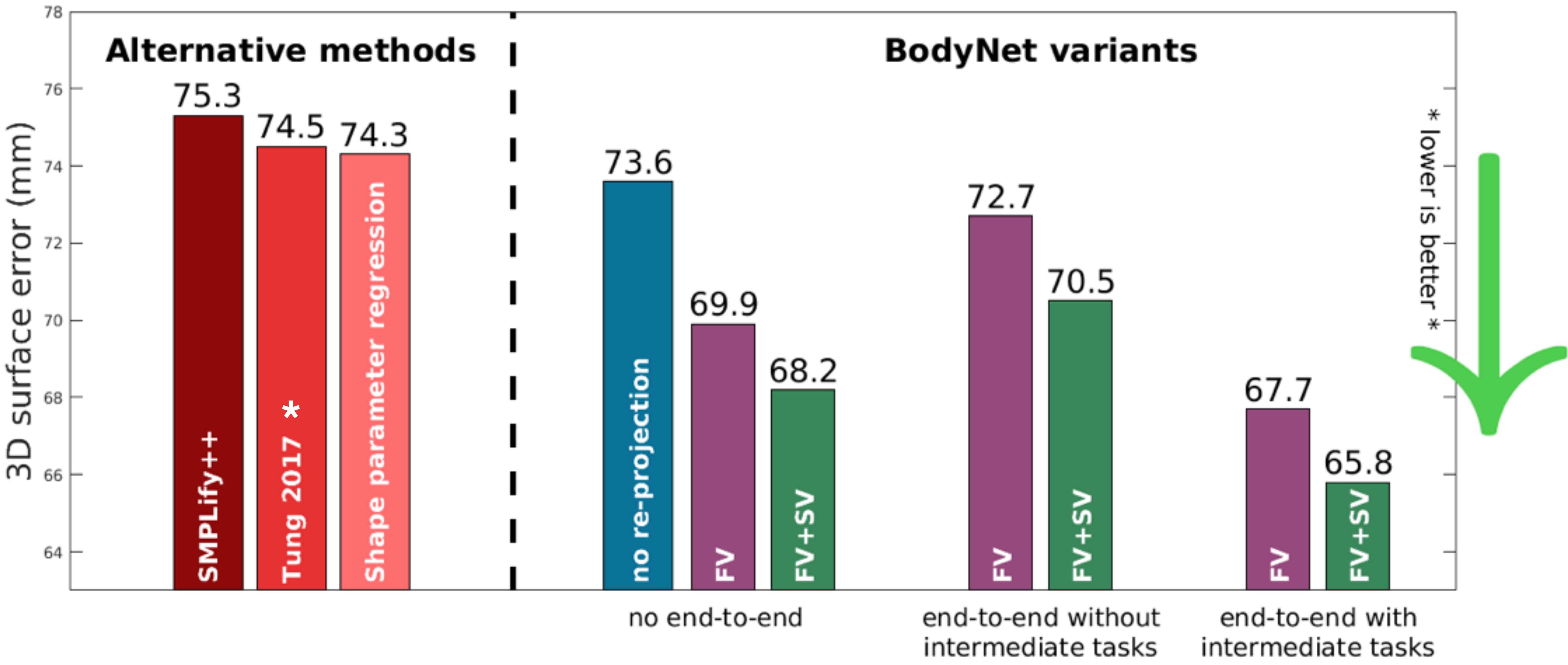
- **Effect of multi-view re-projection**
- **Effect of multi-task training**



Experiments

- **Comparison with alternative methods**

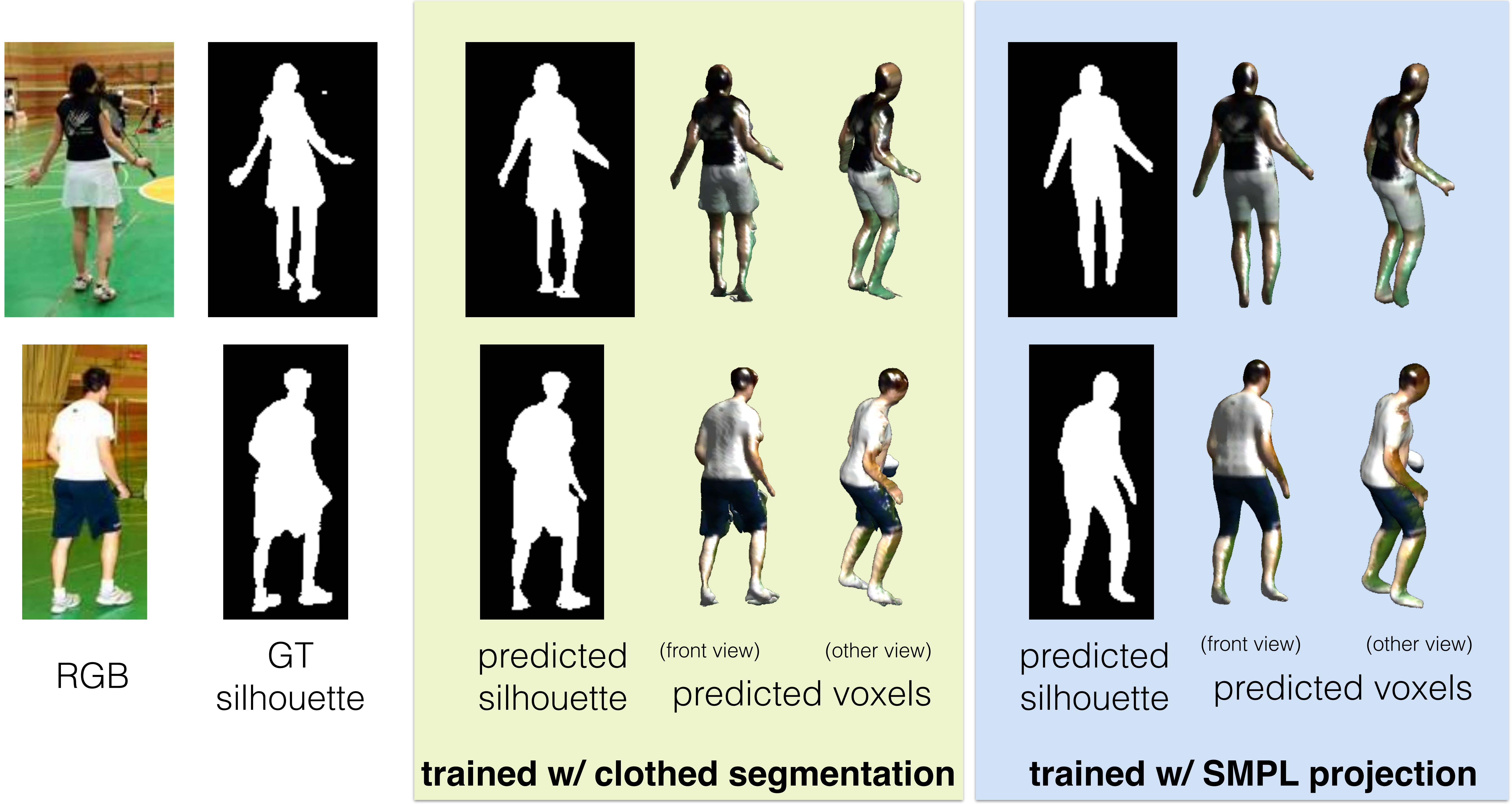




* Tung et al. Self-supervised learning of motion capture, NIPS 2017

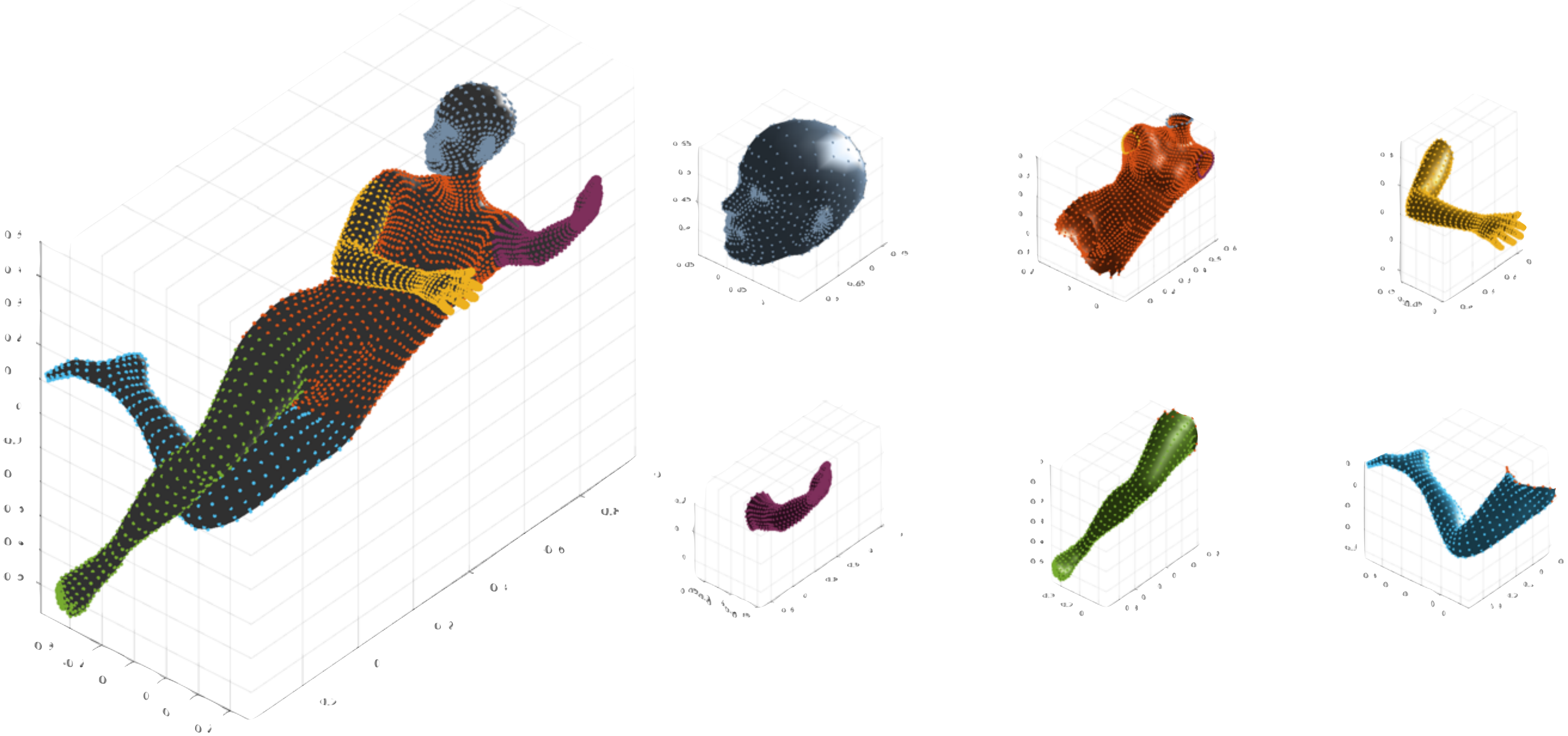
Potential to capture clothing

If the re-projection loss is supervised with 2D clothed segmentation, volumetric output captures clothing.

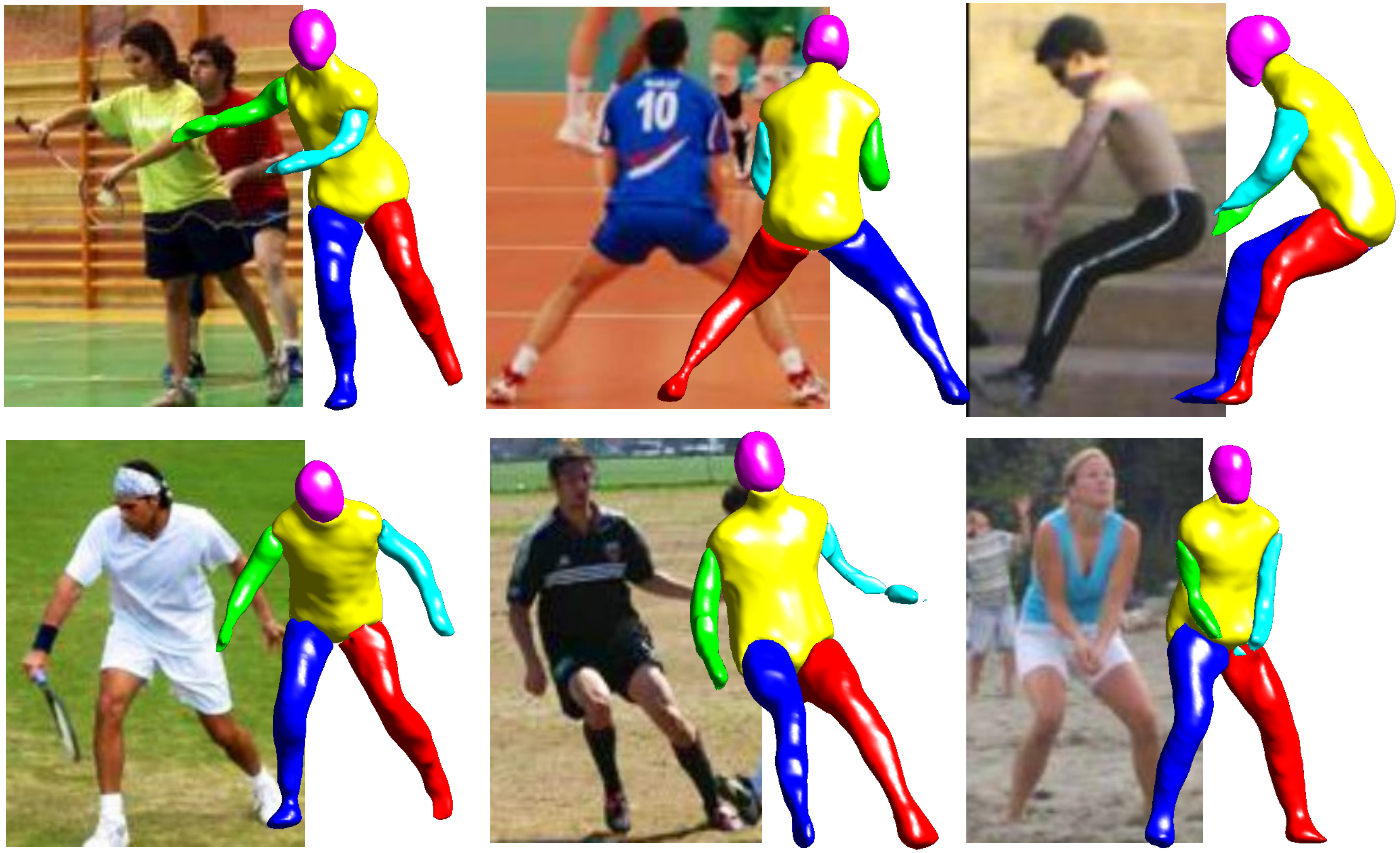


3D Body Part Segmentation

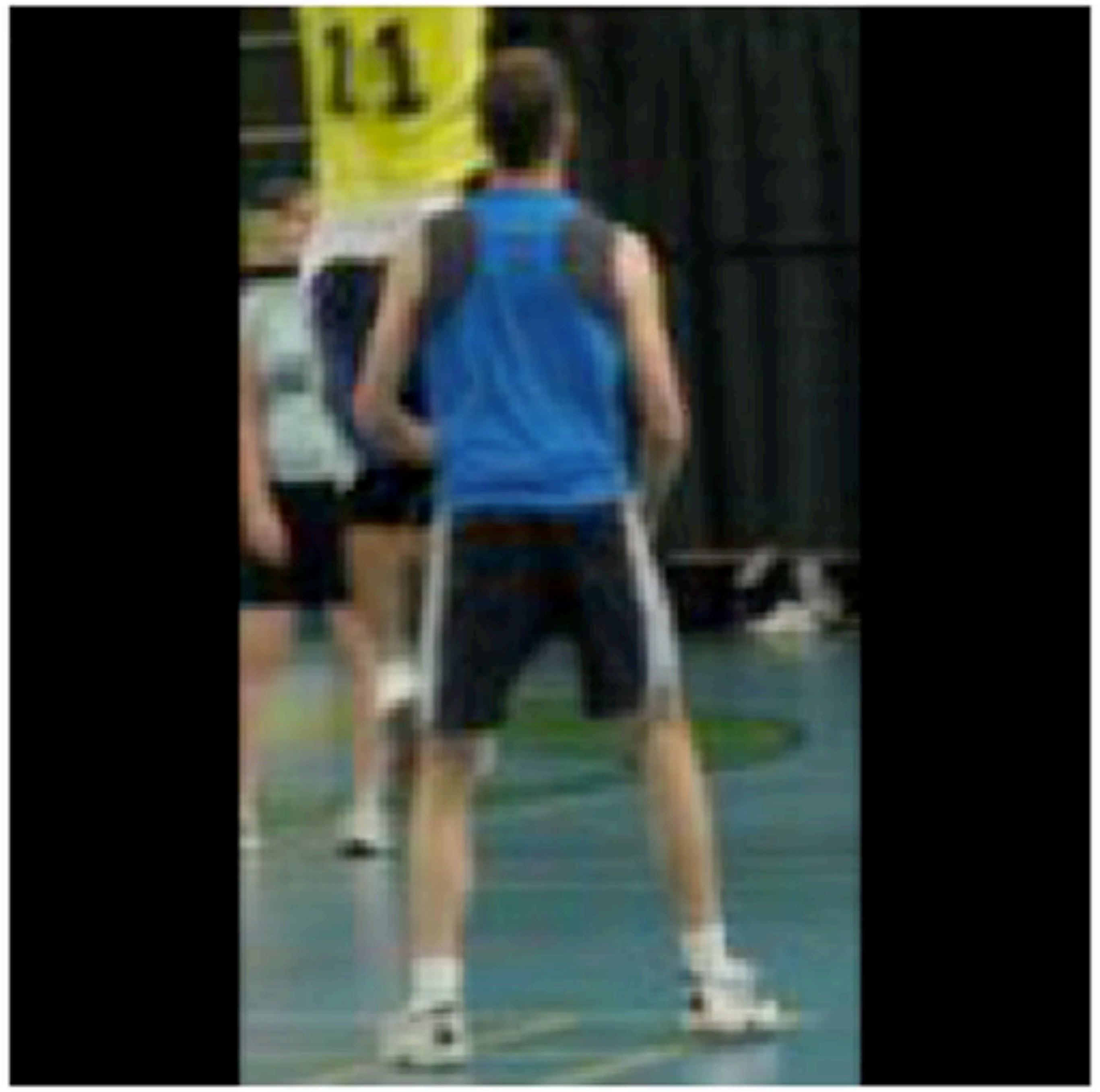
- Initialize part segmentation network with the foreground segmentation network weights by copying last layer as many times as the number of parts.

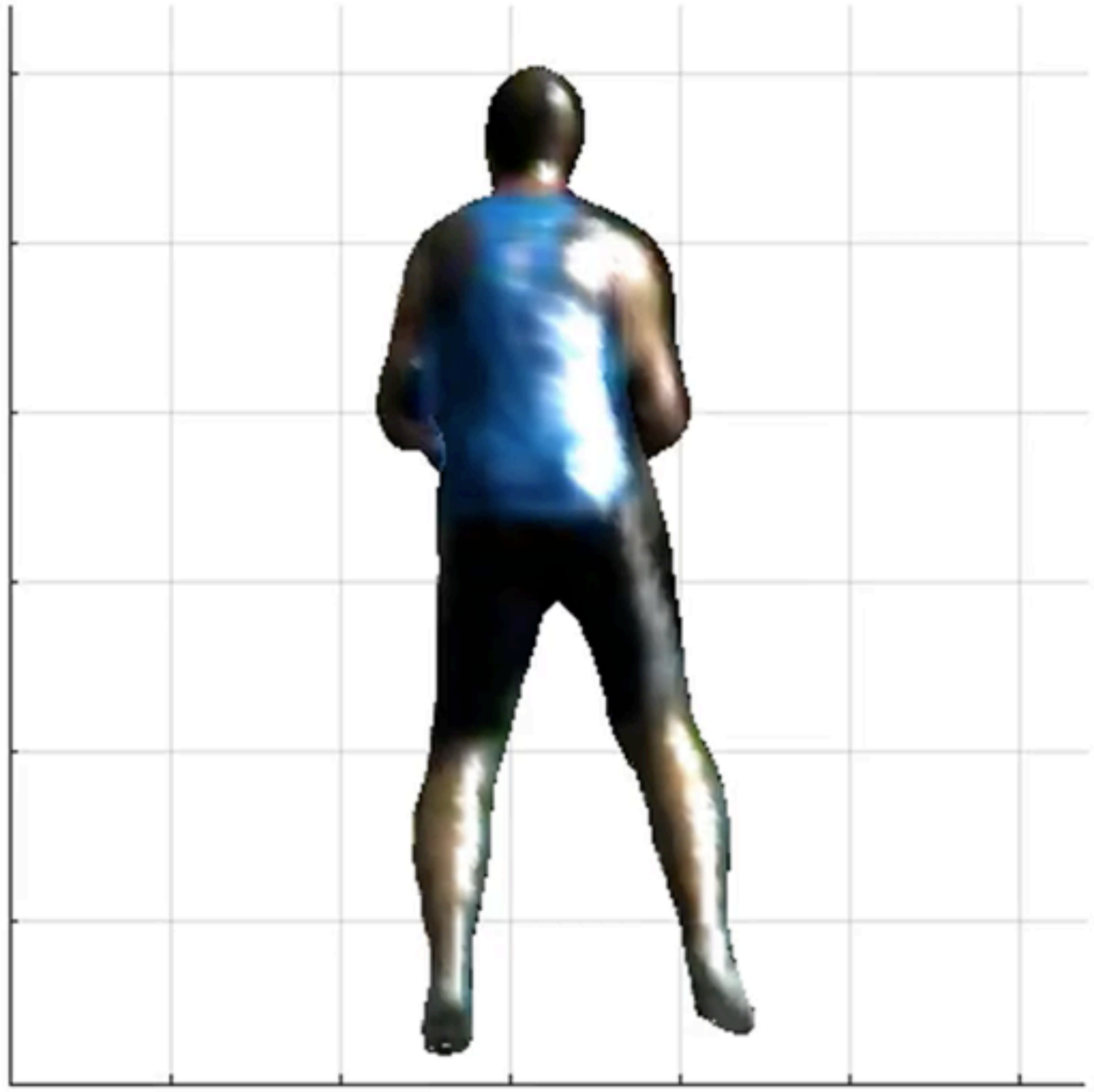


3D Body Part Segmentation

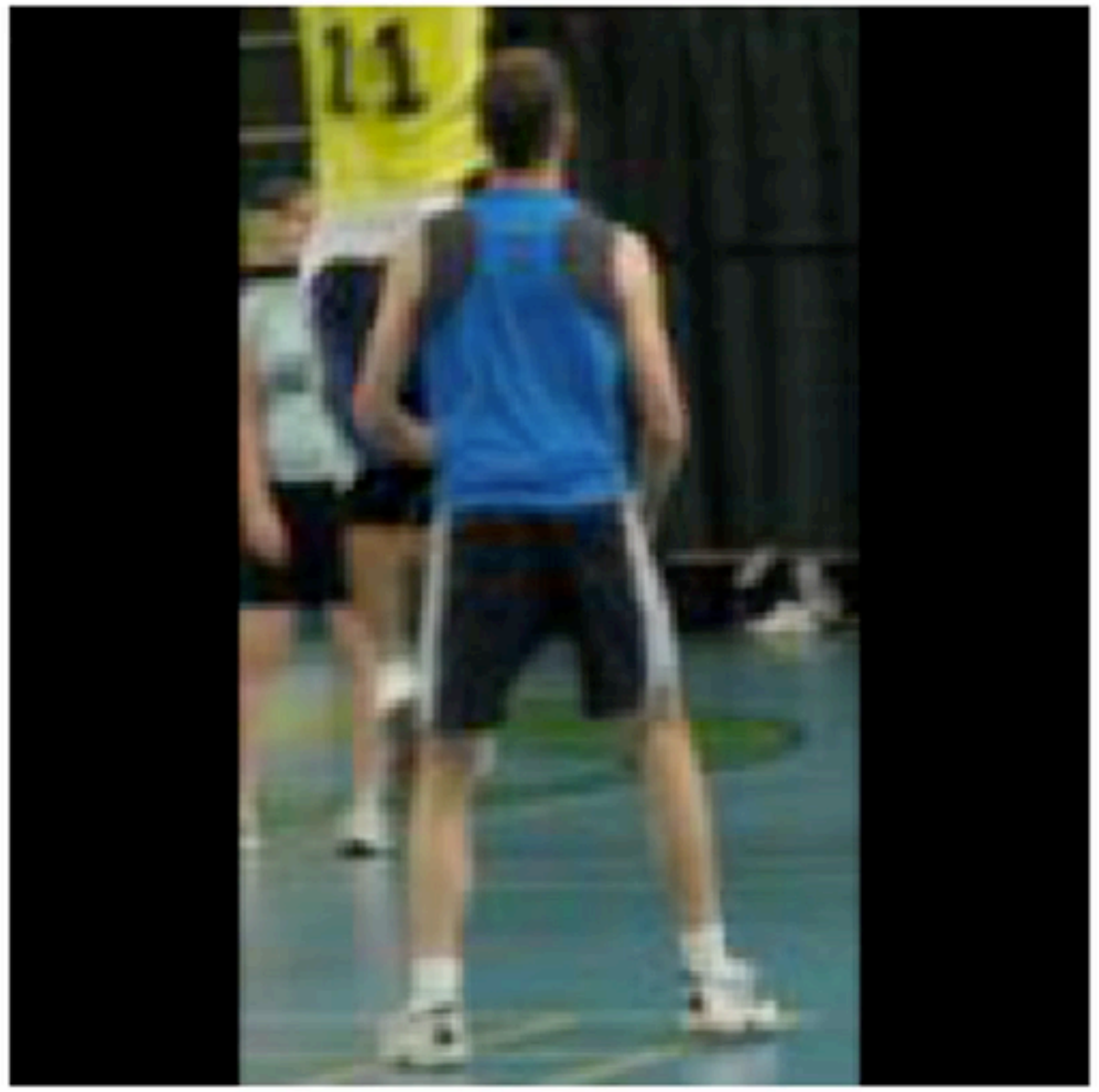


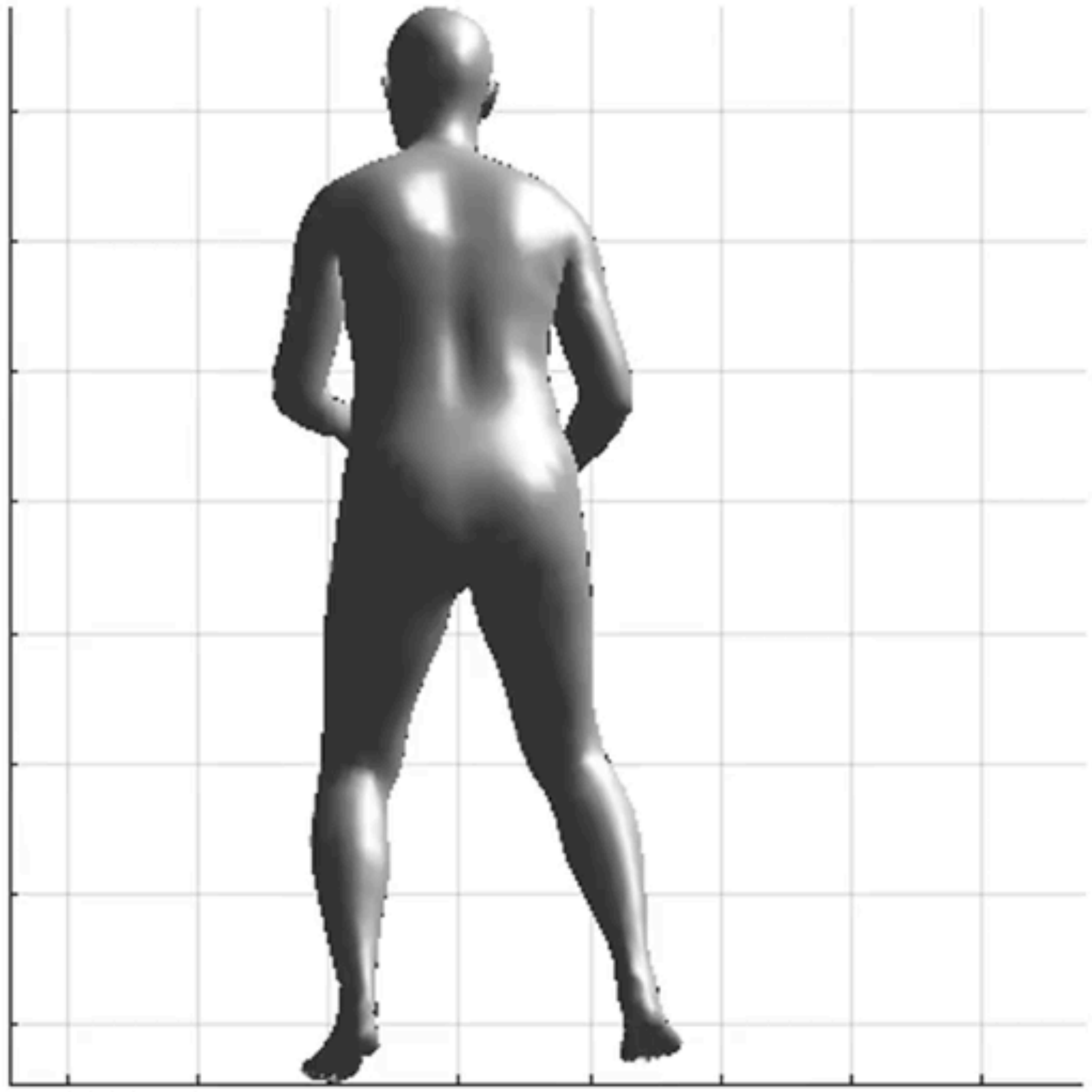
Results - Volumetric Shape



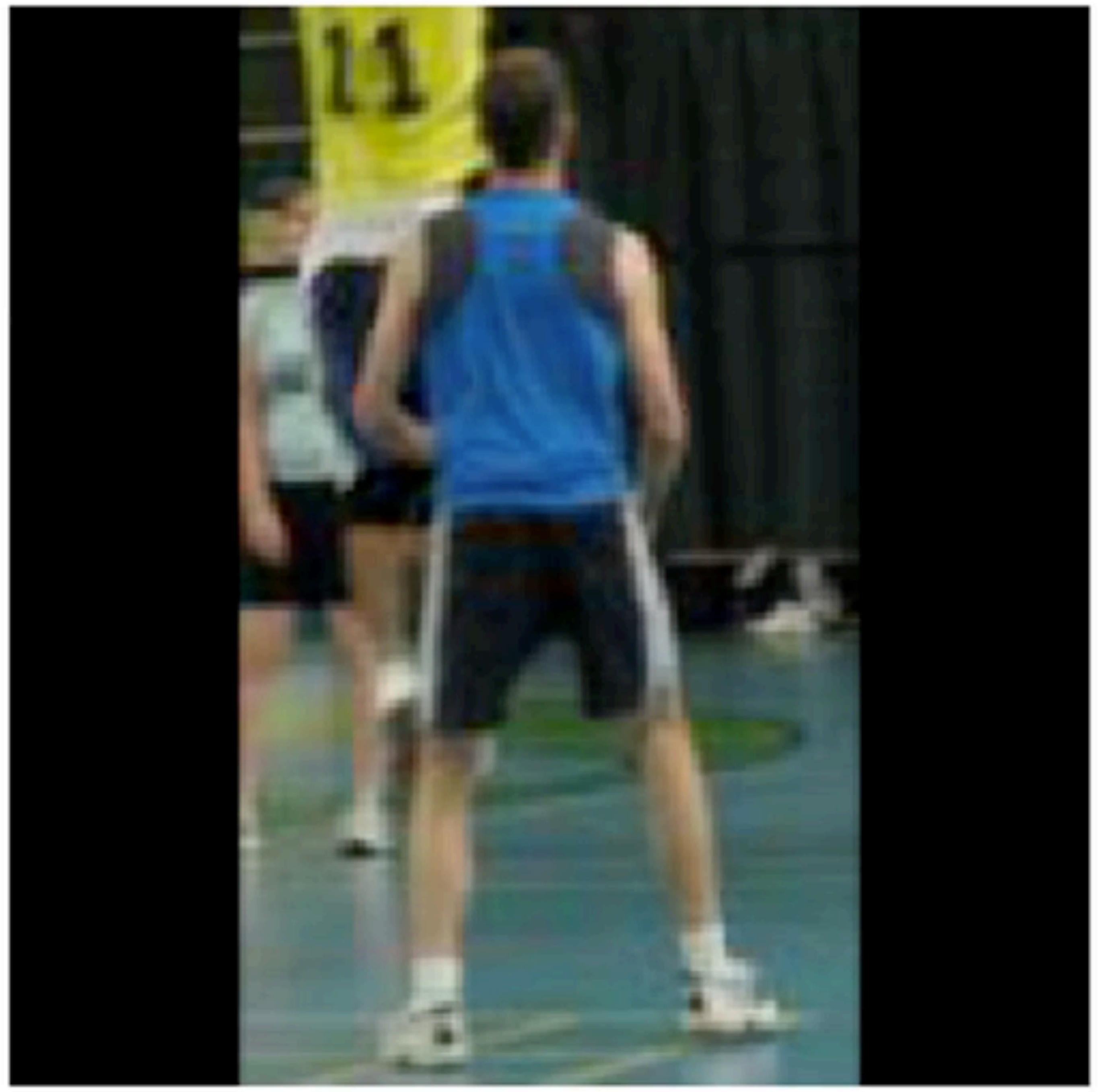


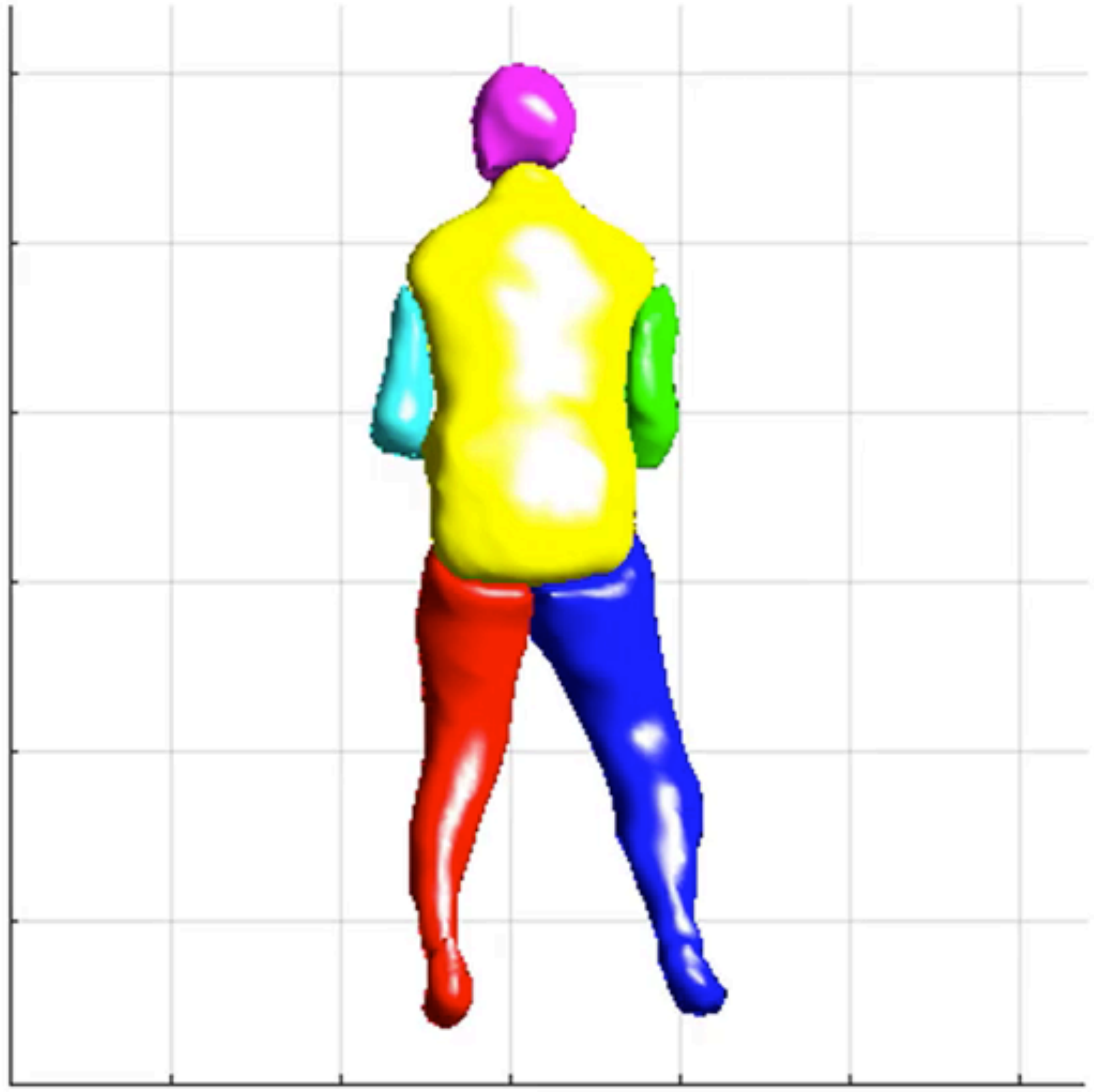
Results - SMPL fit



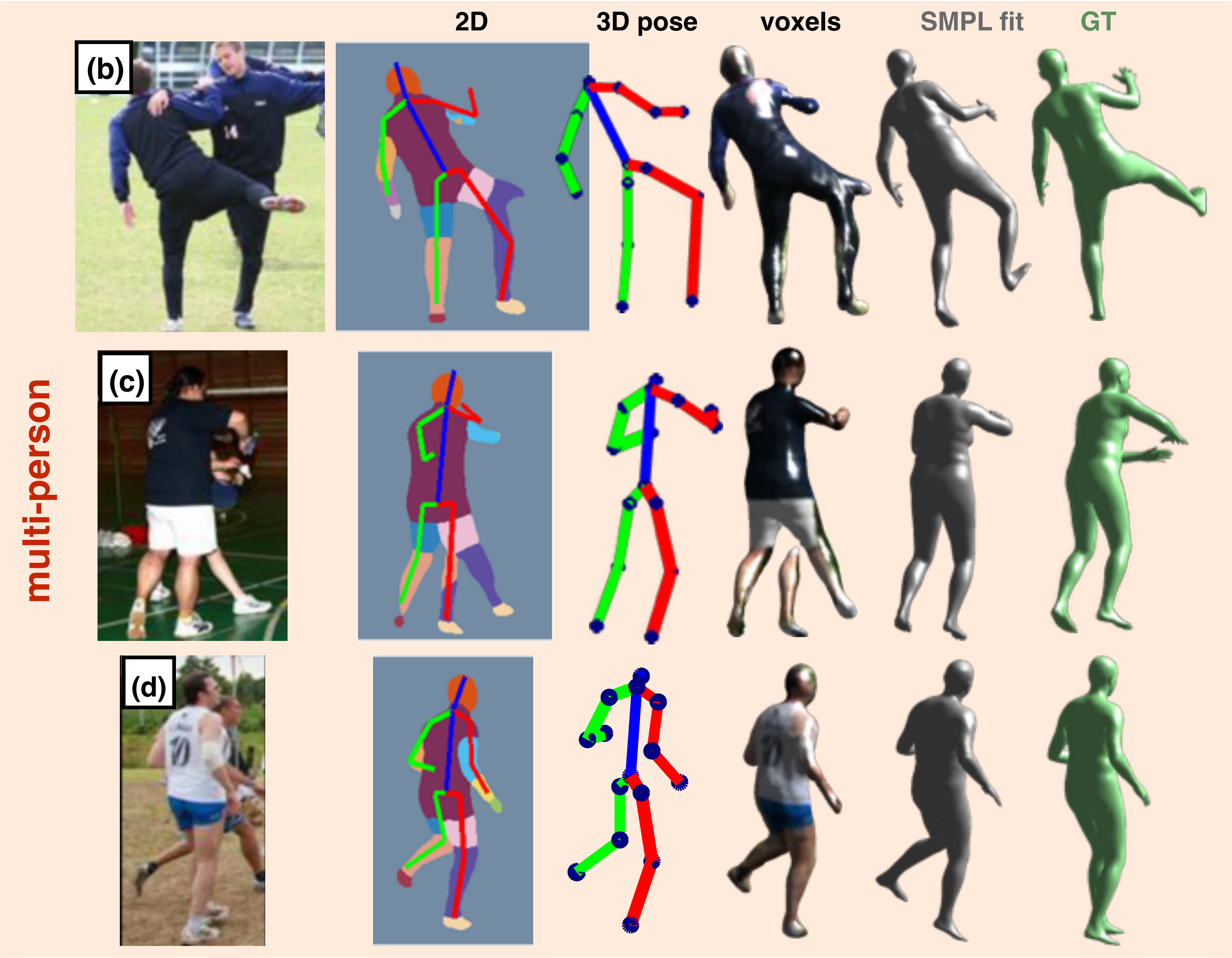


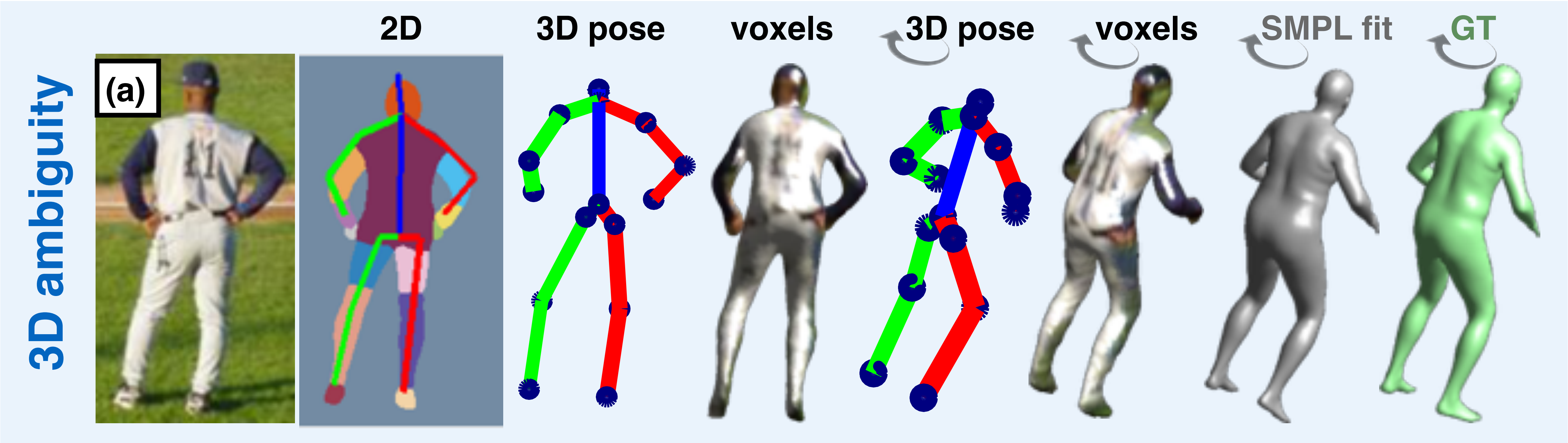
Results - 3D Body Parts





Results - Failure Cases

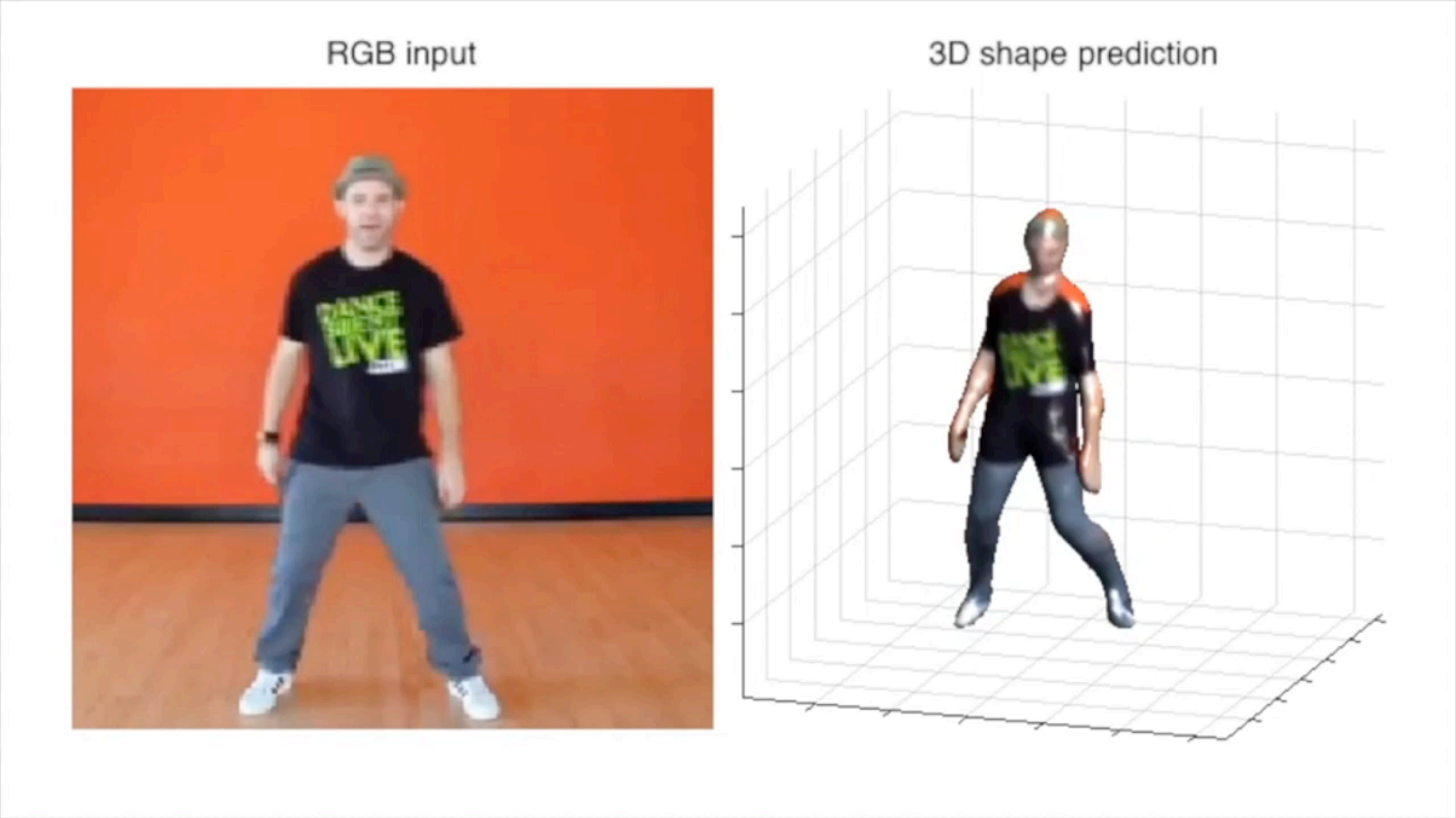




Conclusions

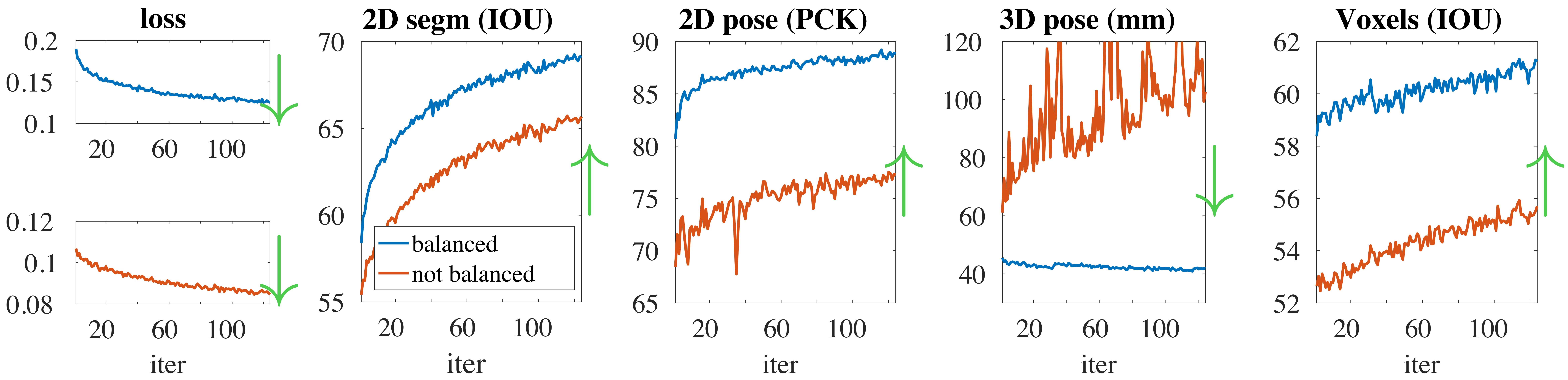
- **Volumetric** human body shape representation is a **flexible** and **effective** alternative to deformable model parameters.
- **Re-projection** loss is critical to obtain confident body surface.
- **Multi-task** training of relevant tasks such as 2D/3D pose and 2D segmentation helps 3D shape estimation.

Thanks



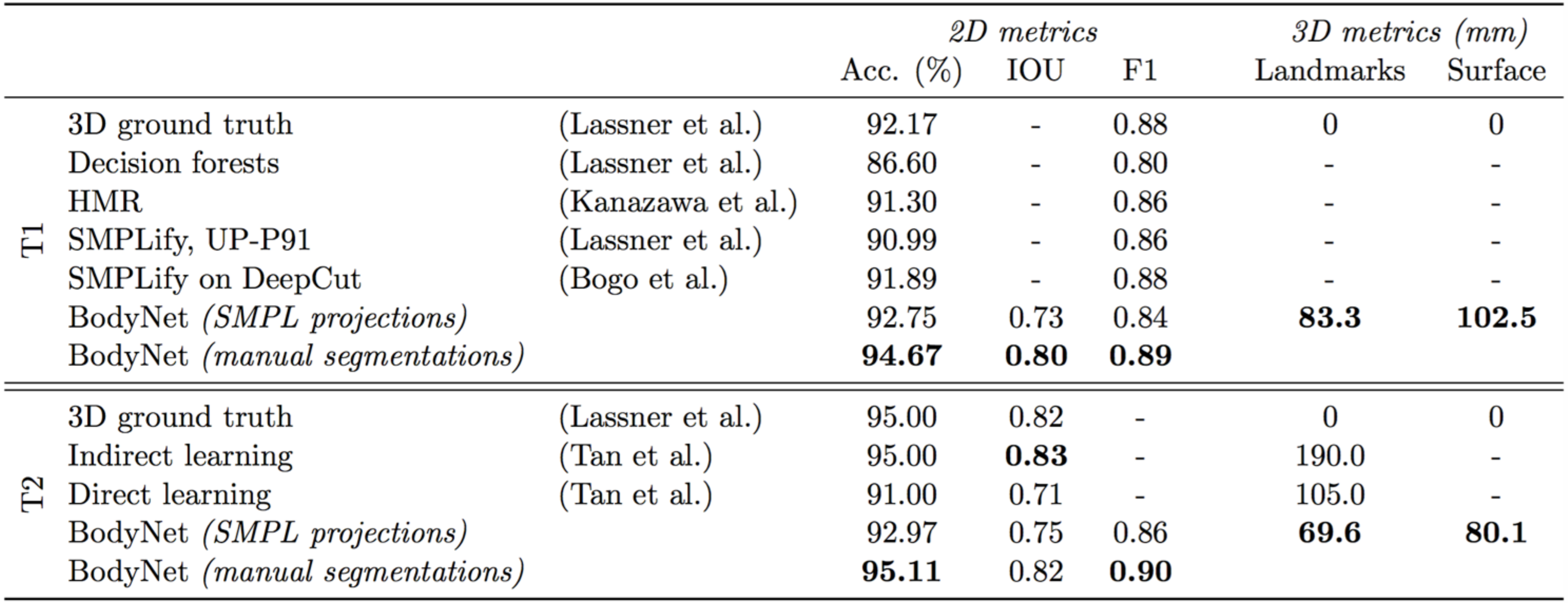
Code is available at:
www.di.ens.fr/willow/research/bodynet/

Balancing multi-task losses



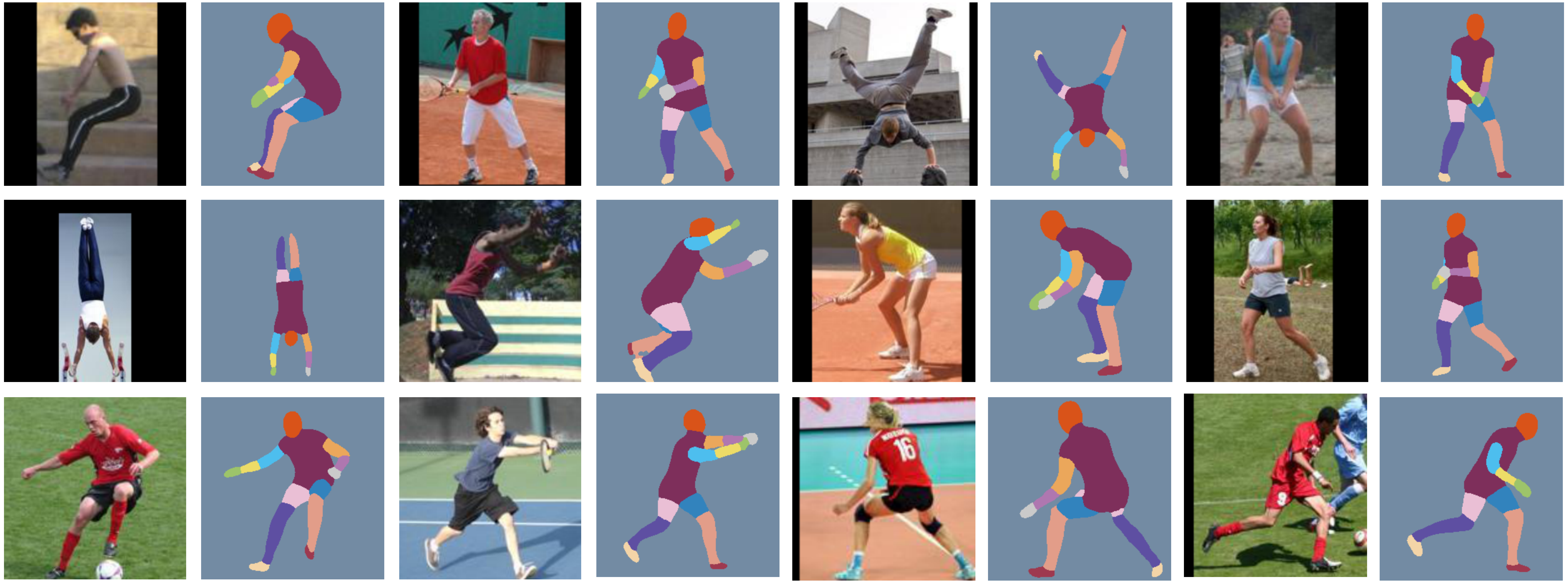
Experiments - Unite the People

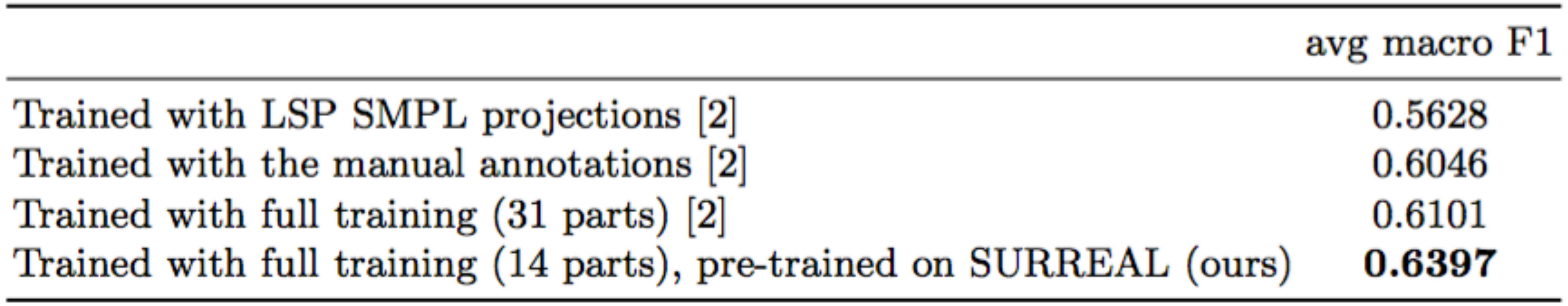
- Comparison to state-of-the-art
- Re-projection supervision



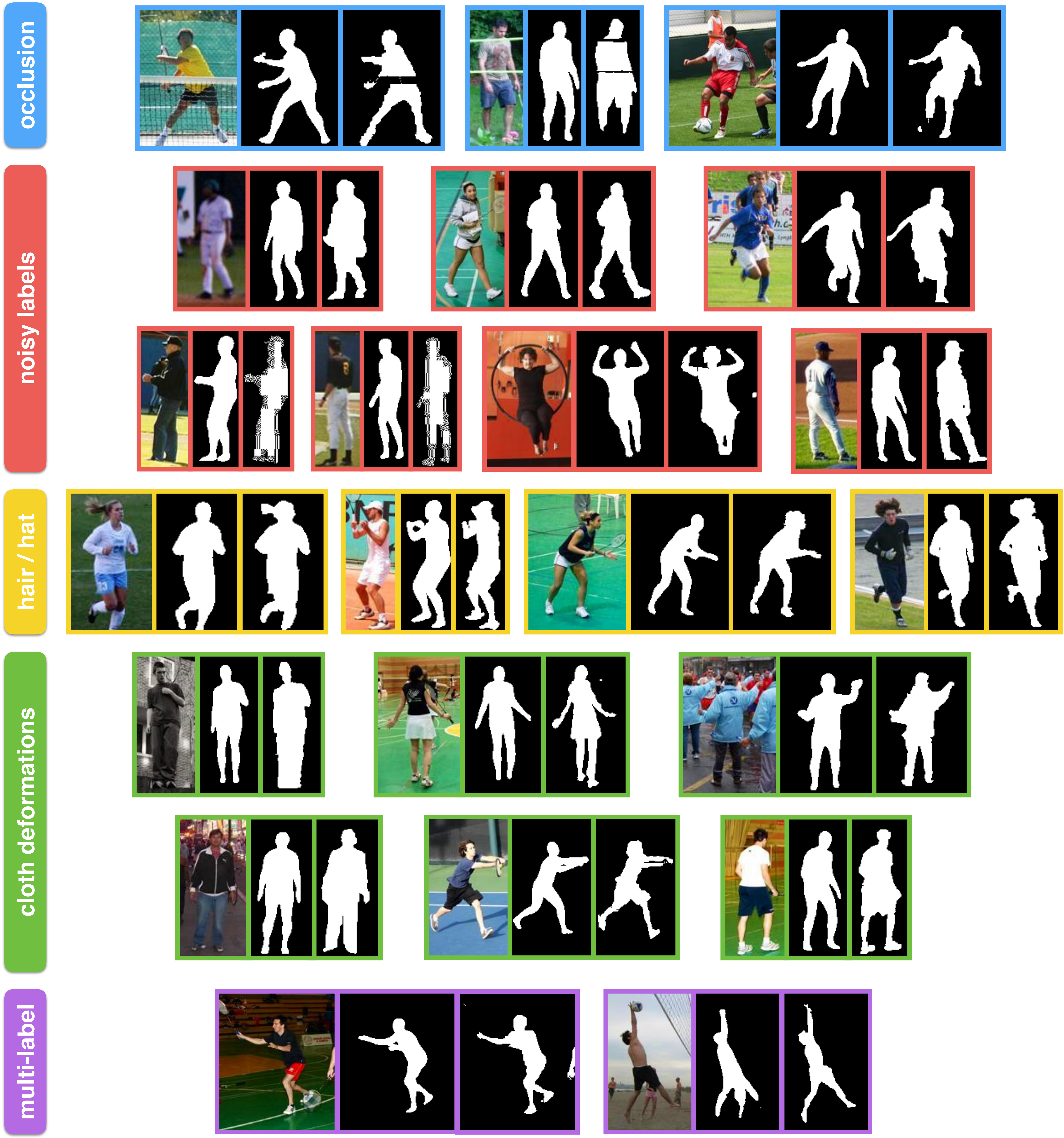
			2D metrics			*3D metrics (mm)*	
			Acc. (%)	IOU	F1	Landmarks	Surface
T1	3D ground truth	(Lassner et al.)	92.17	-	0.88	0	0
	Decision forests	(Lassner et al.)	86.60	-	0.80	-	-
	HMR	(Kanazawa et al.)	91.30	-	0.86	-	-
	SMPLify, UP-P91	(Lassner et al.)	90.99	-	0.86	-	-
	SMPLify on DeepCut	(Bogo et al.)	91.89	-	0.88	-	-
	BodyNet *(SMPL projections)*		92.75	0.73	0.84	**83.3**	**102.5**
	BodyNet *(manual segmentations)*		**94.67**	**0.80**	**0.89**		
T2	3D ground truth	(Lassner et al.)	95.00	0.82	-	0	0
	Indirect learning	(Tan et al.)	95.00	**0.83**	-	190.0	-
	Direct learning	(Tan et al.)	91.00	0.71	-	105.0	-
	BodyNet *(SMPL projections)*		92.97	0.75	0.86	**69.6**	**80.1**
	BodyNet *(manual segmentations)*		**95.11**	0.82	**0.90**		

2D Segmentation on UP

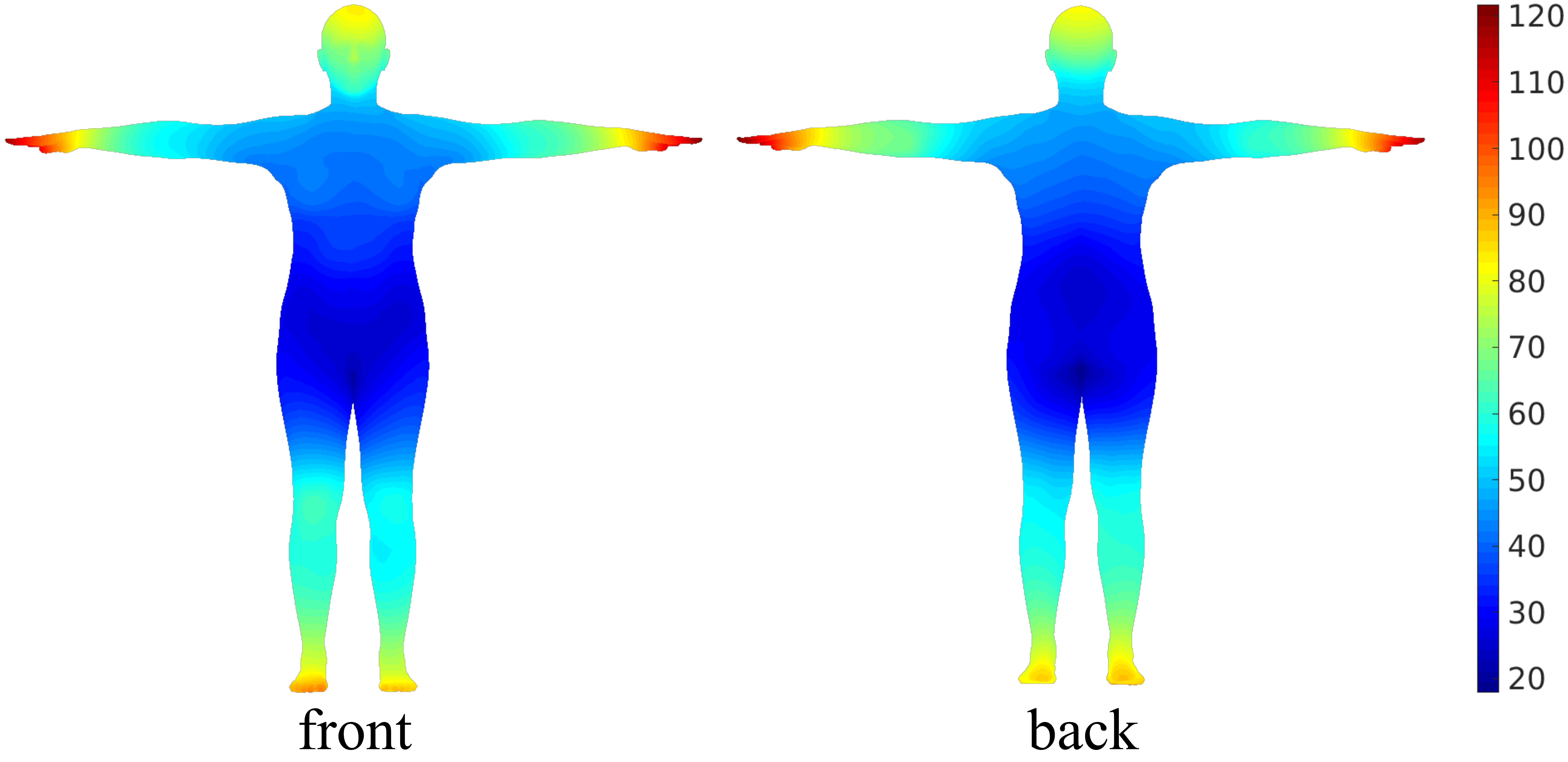


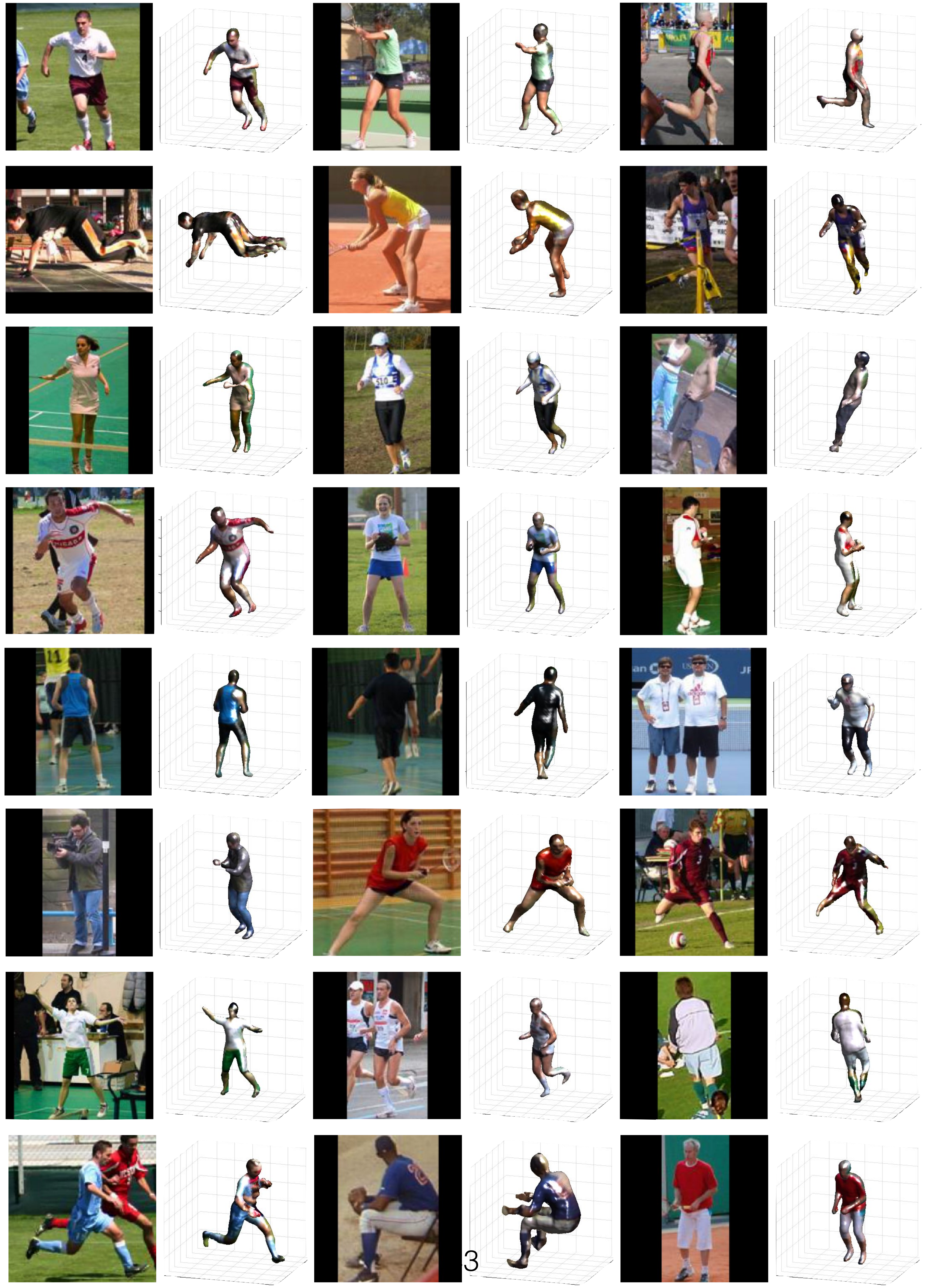


	avg macro F1
Trained with LSP SMPL projections [2]	0.5628
Trained with the manual annotations [2]	0.6046
Trained with full training (31 parts) [2]	0.6101
Trained with full training (14 parts), pre-trained on SURREAL (ours)	**0.6397**

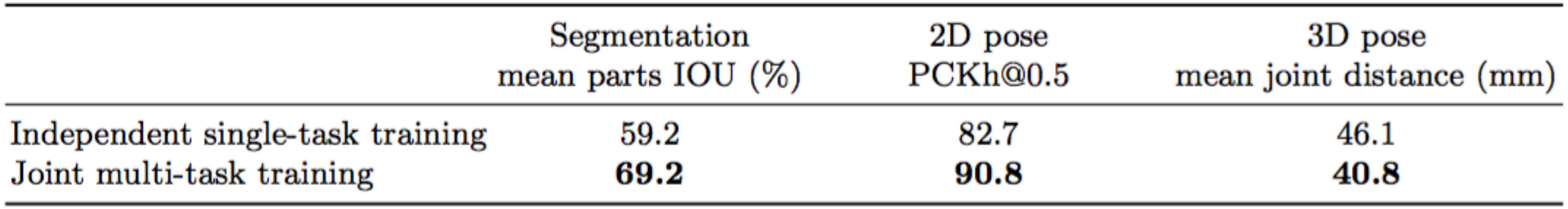

occlusion
noisy labels
hair / hat
cloth deformations
multi-label

SMPL Surface Error



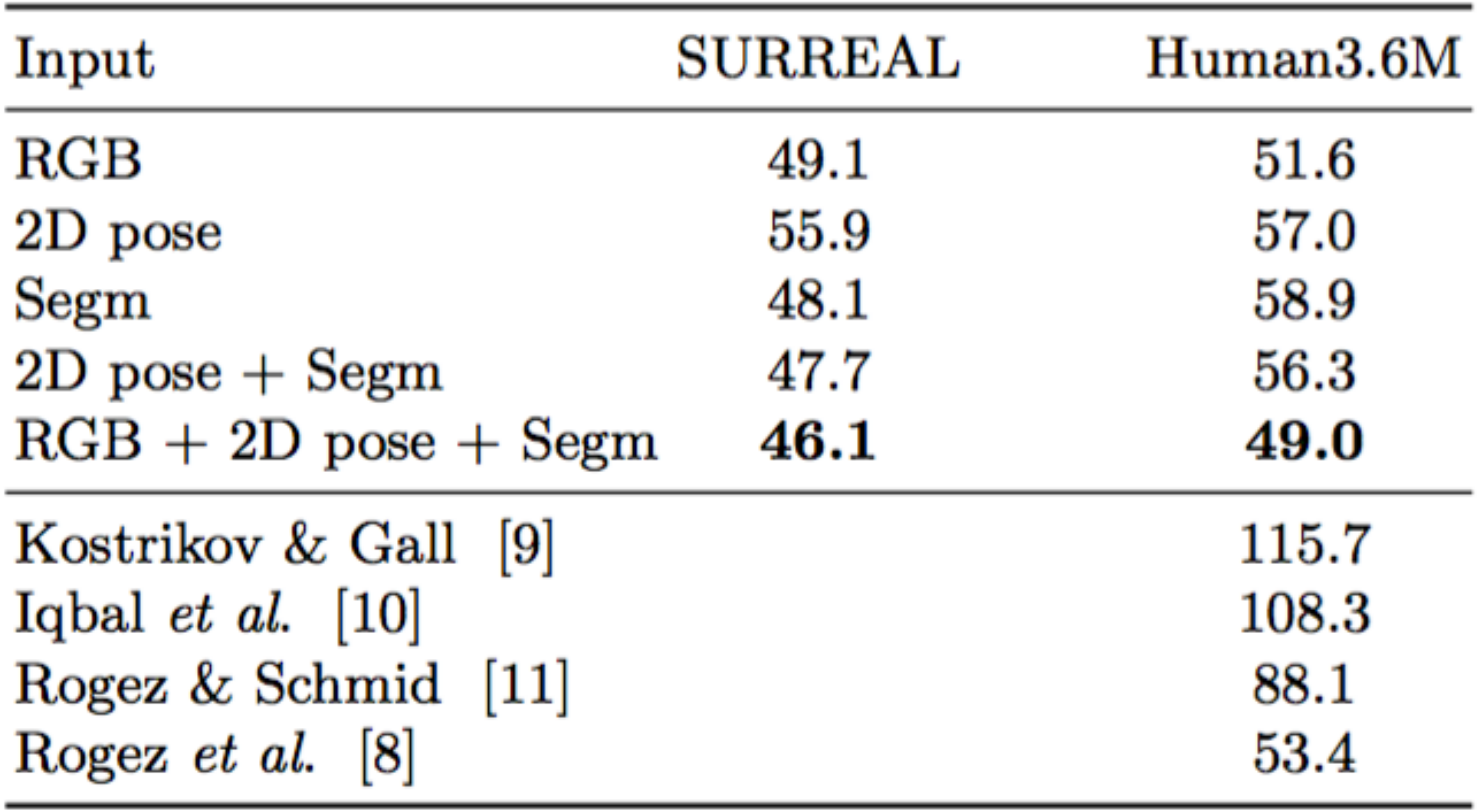


Intermediate Tasks after Multi-task Fine-tuning



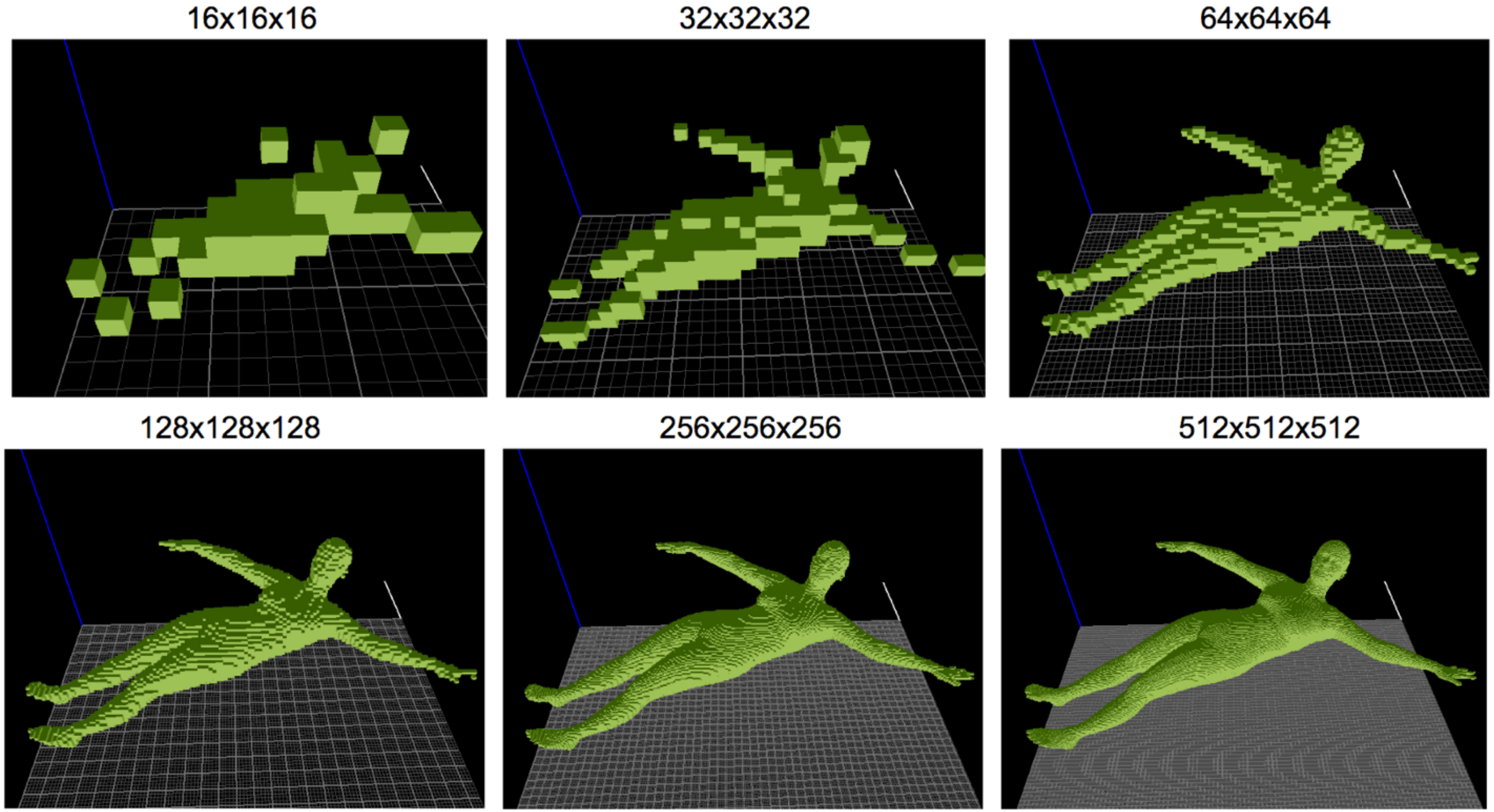
	Segmentation mean parts IOU (%)	2D pose PCKh@0.5	3D pose mean joint distance (mm)
Independent single-task training	59.2	82.7	46.1
Joint multi-task training	**69.2**	**90.8**	**40.8**

3D Pose Experiments



Input	SURREAL	Human3.6M
RGB	49.1	51.6
2D pose	55.9	57.0
Segm	48.1	58.9
2D pose + Segm	47.7	56.3
RGB + 2D pose + Segm	**46.1**	**49.0**
Kostrikov & Gall [9]		115.7
Iqbal *et al.* [10]		108.3
Rogez & Schmid [11]		88.1
Rogez *et al.* [8]		53.4

Voxel grid resolution



16x16x16

32x32x32

64x64x64

128x128x128

256x256x256

512x512x512