BodyNet: Volumetric Inference of 3D Human Body Shapes

Gül Varol[1], Duygu Ceylan[2], Bryan Russell[2], Jimei Yang[2], Ersin Yumer[3], Ivan Laptev[1] and Cordelia Schmid[1]

[1]Inria [2]Adobe Research [3]Argo AI

ECCV 2018 European Conference on Computer Vision

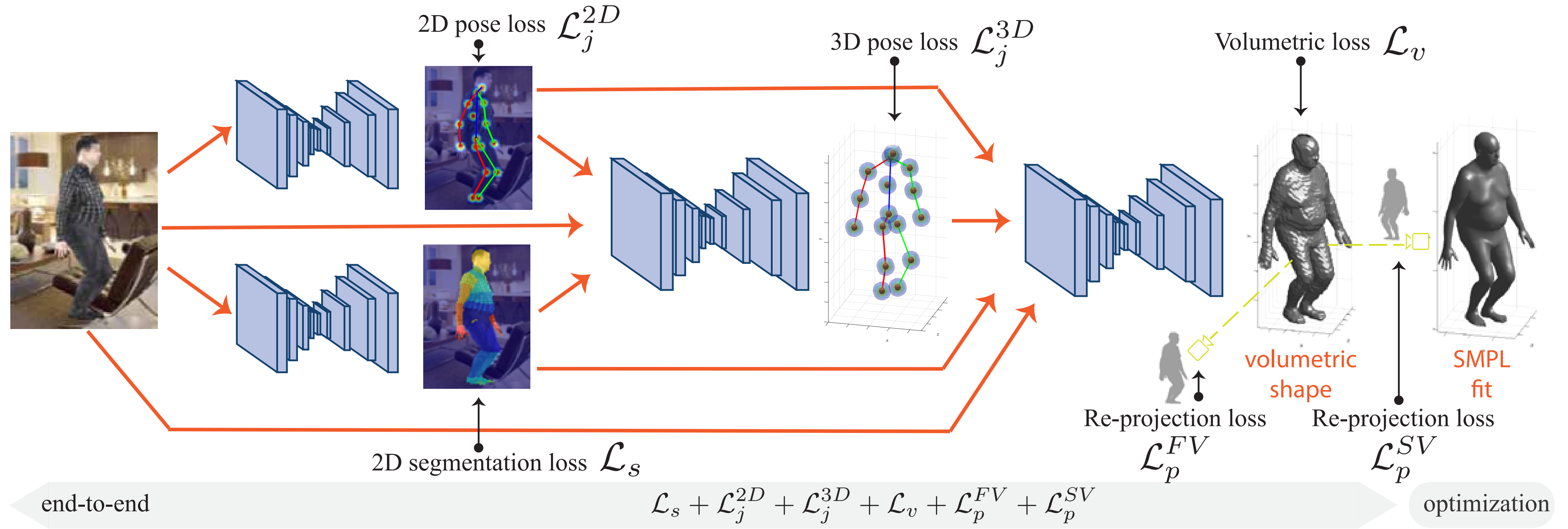
INTRODUCTION

- Goal & Contributions
  - Predicting 3D human body **pose and shape** given a **single RGB** image as input.
  - Demonstrating advantages of auxiliary body-related tasks in an end-to-end multi-task setting.
- Motivation
  - Volumetric representation of human bodies in the context of neural networks is not studied.
  - Volumetric representation is flexible, e.g. can capture clothing.

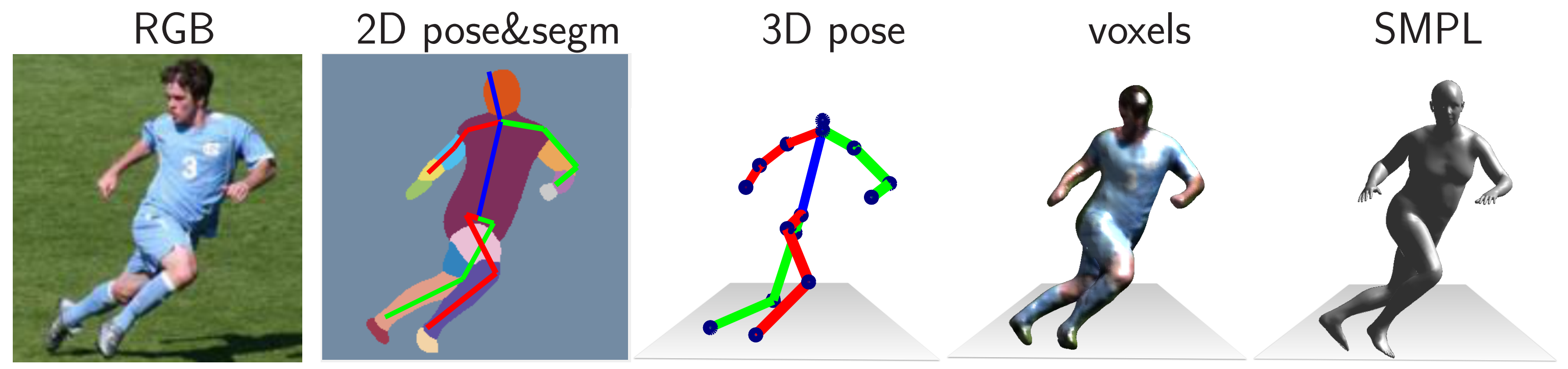
input output voxels output parts

BODYNET APPROACH

- The architecture benefits from the **multi-task** training of:
  - a **volumetric** 3D loss,
  - a **multi-view re-projection** loss,
  - intermediate supervision of **2D pose, 2D part segmentation, and 3D pose**.

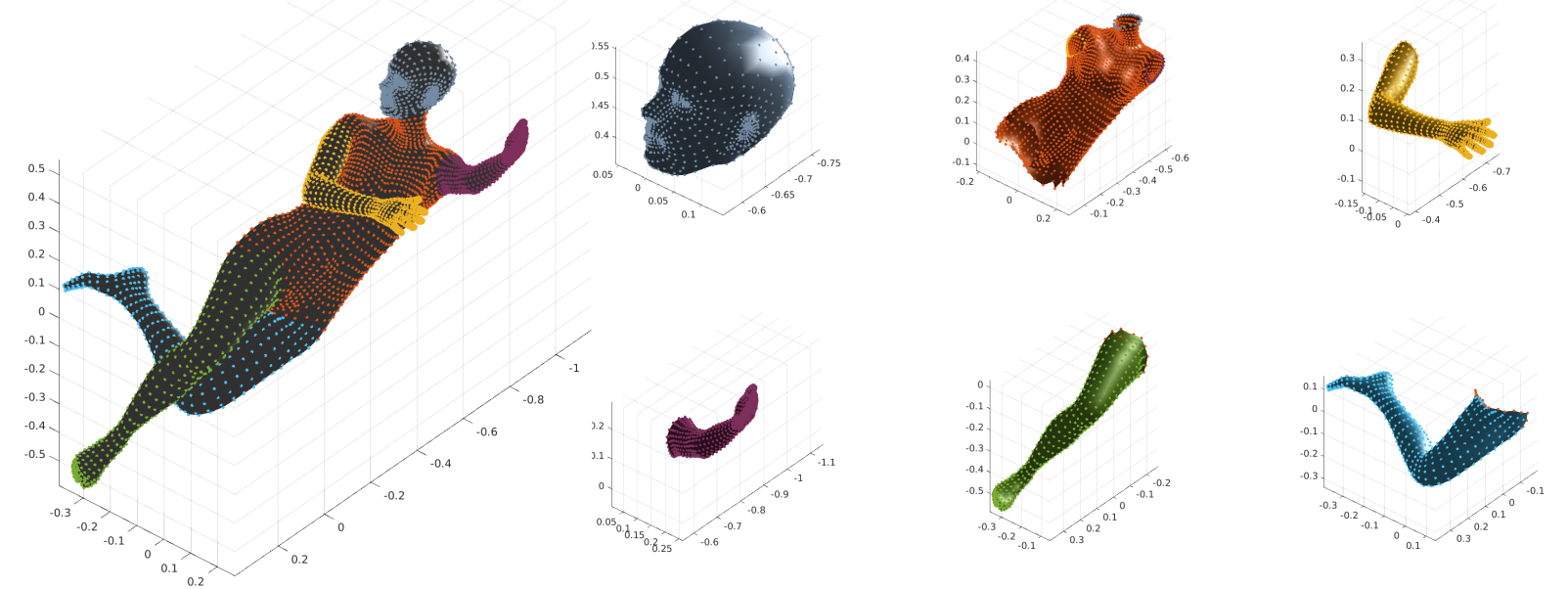

2D pose loss $\mathcal{L}_j^{2D}$
3D pose loss $\mathcal{L}_j^{3D}$
Volumetric loss $\mathcal{L}_v$
volumetric shape
SMPL fit
Re-projection loss $\mathcal{L}_p^{FV}$
Re-projection loss $\mathcal{L}_p^{SV}$
2D segmentation loss $\mathcal{L}_s$
end-to-end
$\mathcal{L}_s + \mathcal{L}_j^{2D} + \mathcal{L}_j^{3D}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

- We gradually increase the difficulty of the task to go from 2D to 3D:



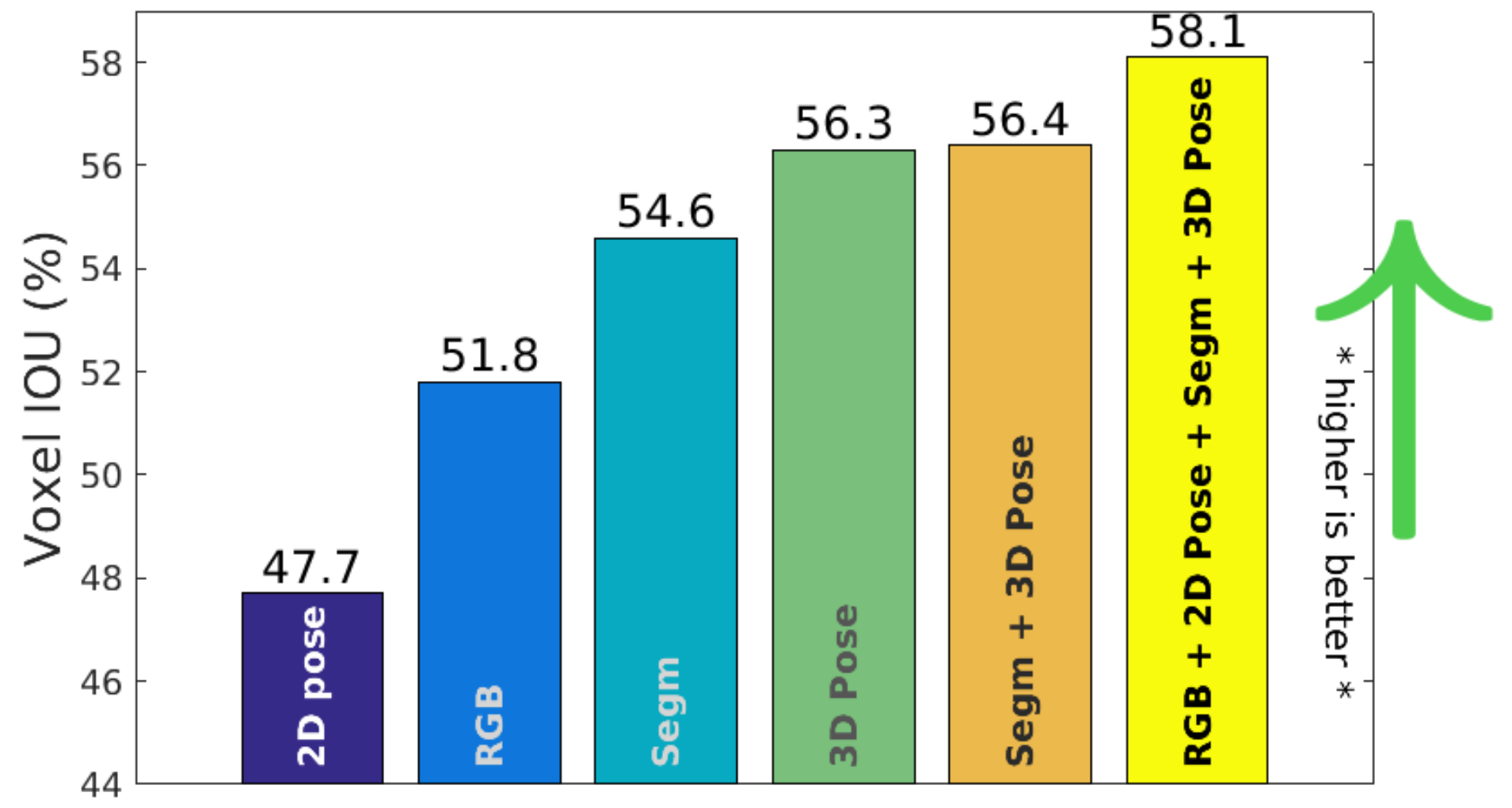
EXTENDING TO 3D BODY PART SEGMENTATION

- Last layer weights are duplicated as many times as the number of parts to initialize training for **part voxels**.



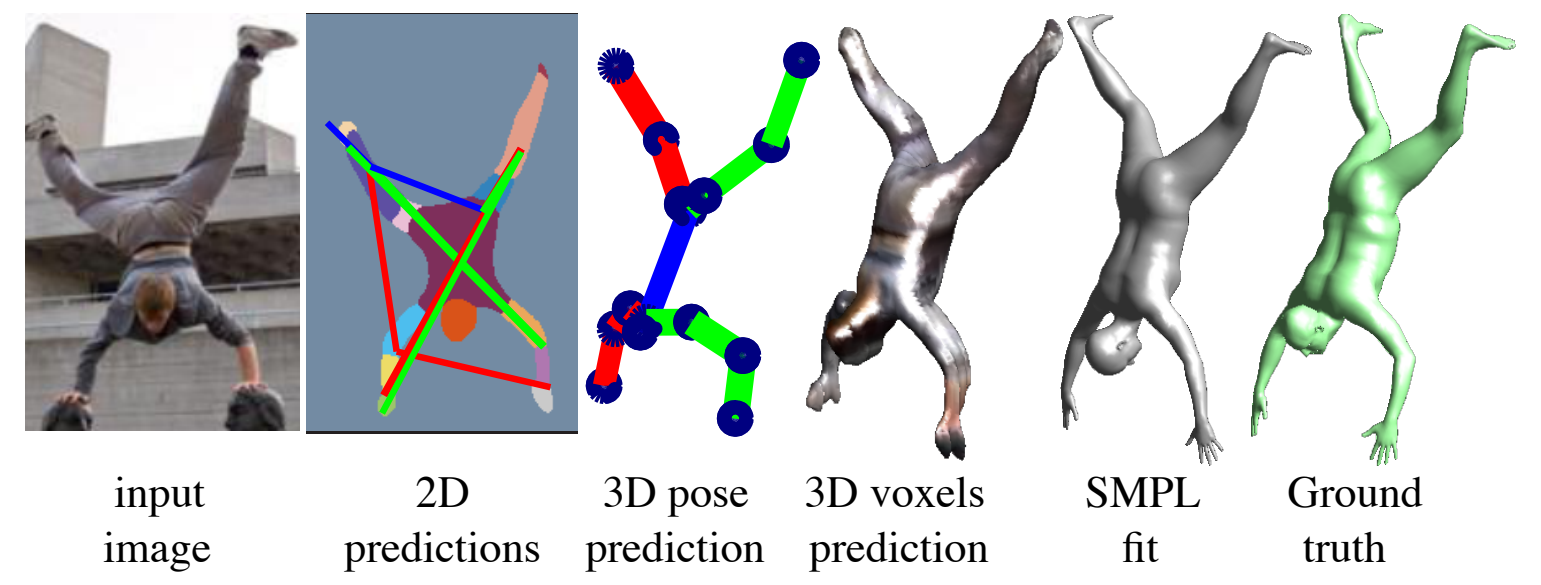
ARCHITECTURE STUDY

- Effect of **additional inputs**
  - 3D shape estimation on SURREAL



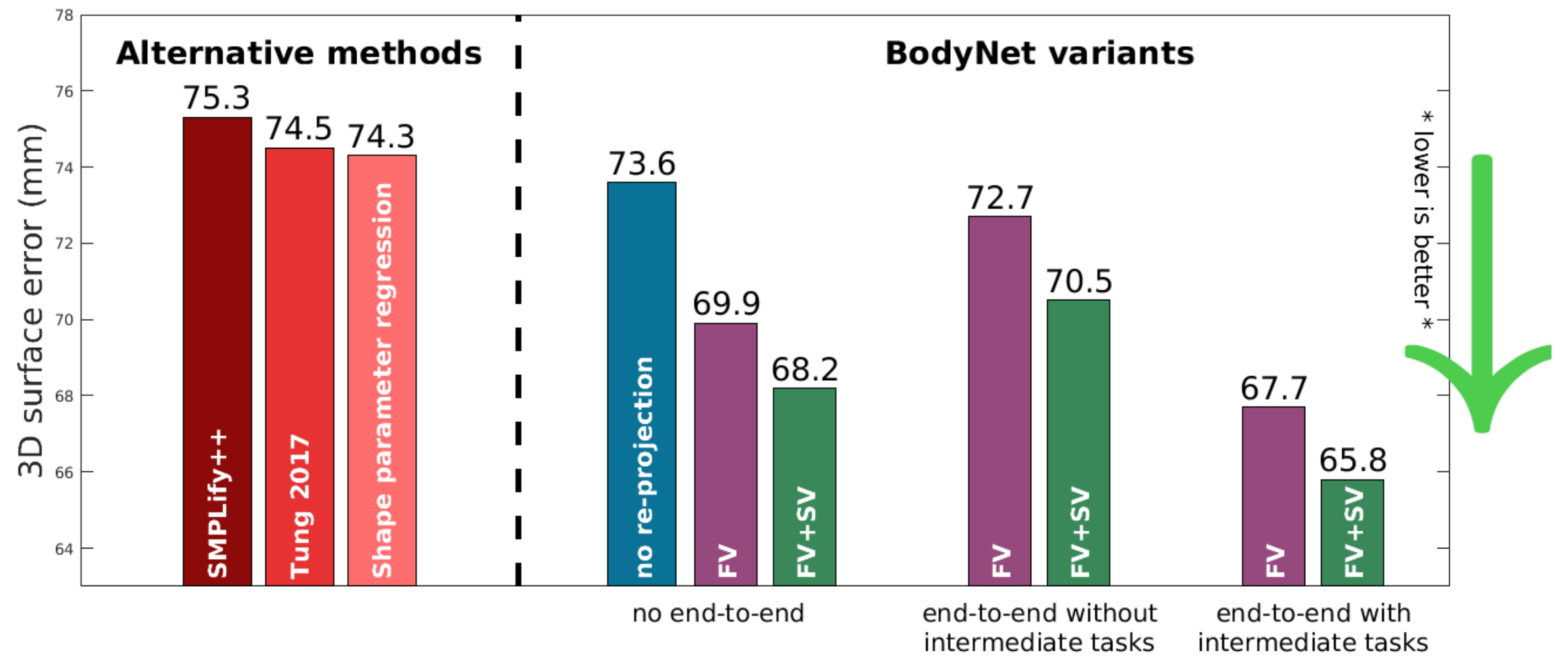
  - 3D pose estimation (mm)

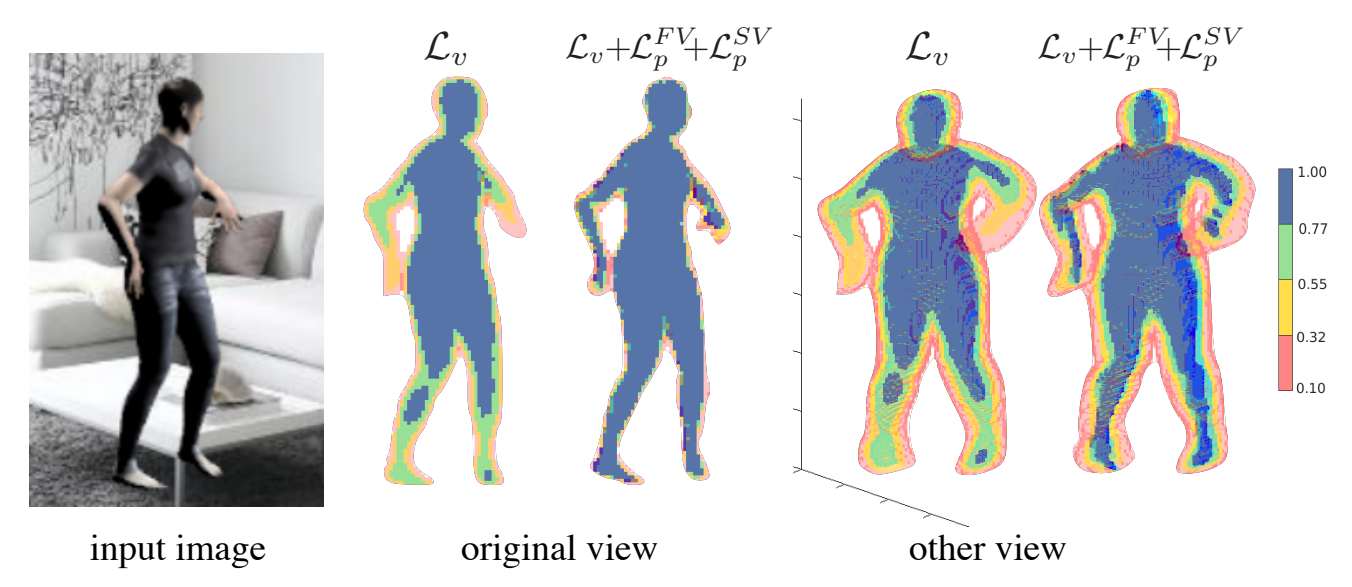
Input	SURREAL	Human3.6M
RGB	49.1	51.6
2D pose	55.9	57.0
Segm	48.1	58.9
2D pose + Segm	47.7	56.3
RGB + 2D pose + Segm	**46.1**	**49.0**

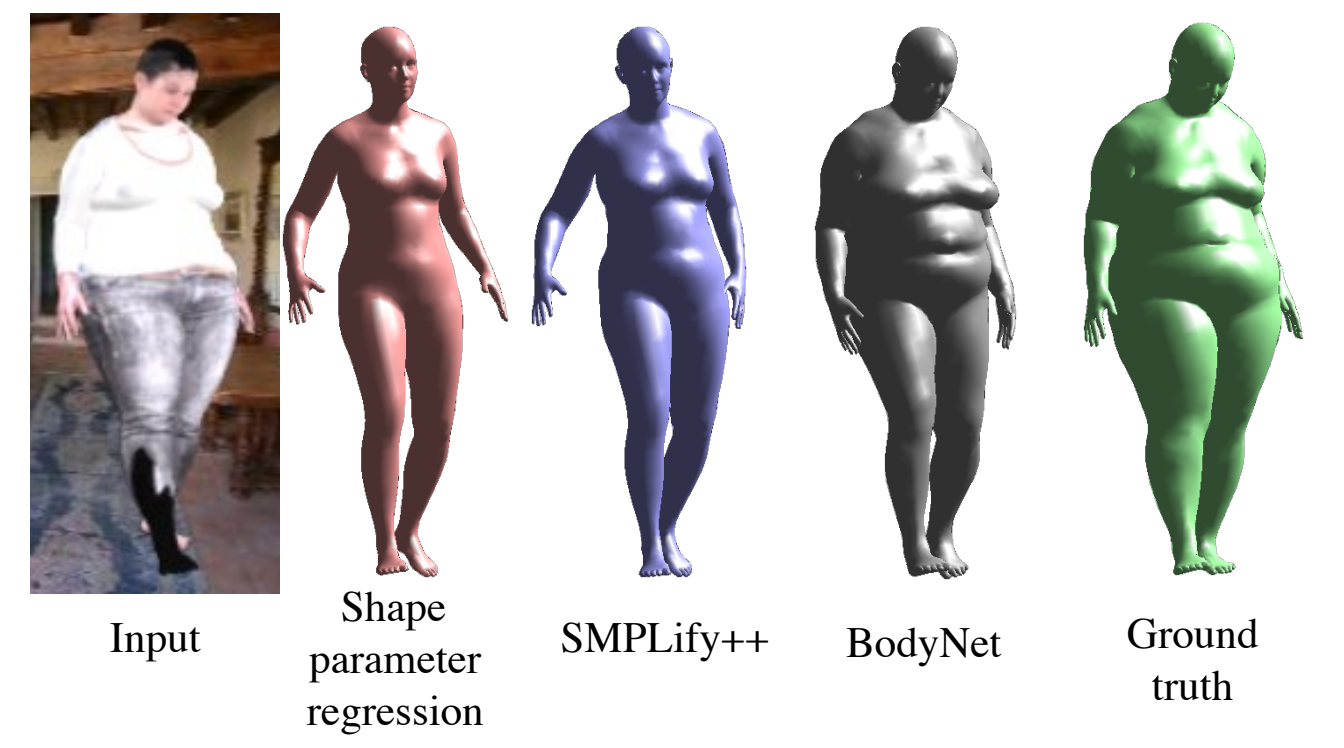


RESULTS: SURREAL dataset [Varol et al. CVPR 2017]

- Effect of multi-view **re-projection loss**
- Effect of **end-to-end** training
- Comparison with **alternative methods**



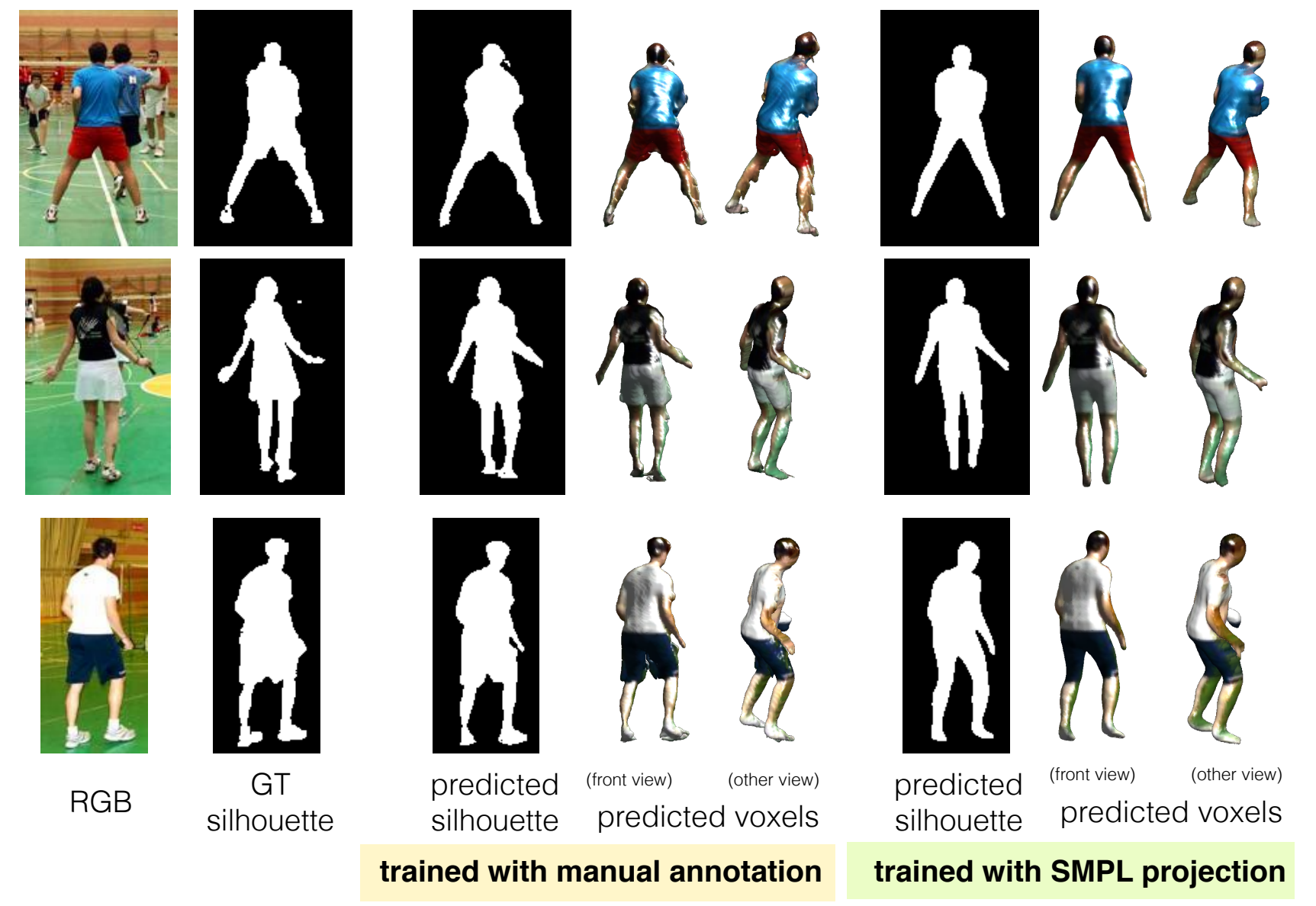

$\mathcal{L}_v$
$\mathcal{L}_v + \mathcal{L}_p^{FV} + \mathcal{L}_p^{SV}$
$\mathcal{L}_v$
$\mathcal{L}_v + \mathcal{L}_p^{FV} + \mathcal{L}_p^{SV}$
1.00
0.77
0.55
0.32
0.10
input image
original view
other view



RESULTS: Unite the People dataset [Lassner et al. CVPR 2017]

- Effect of the **re-projection type**

			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**		

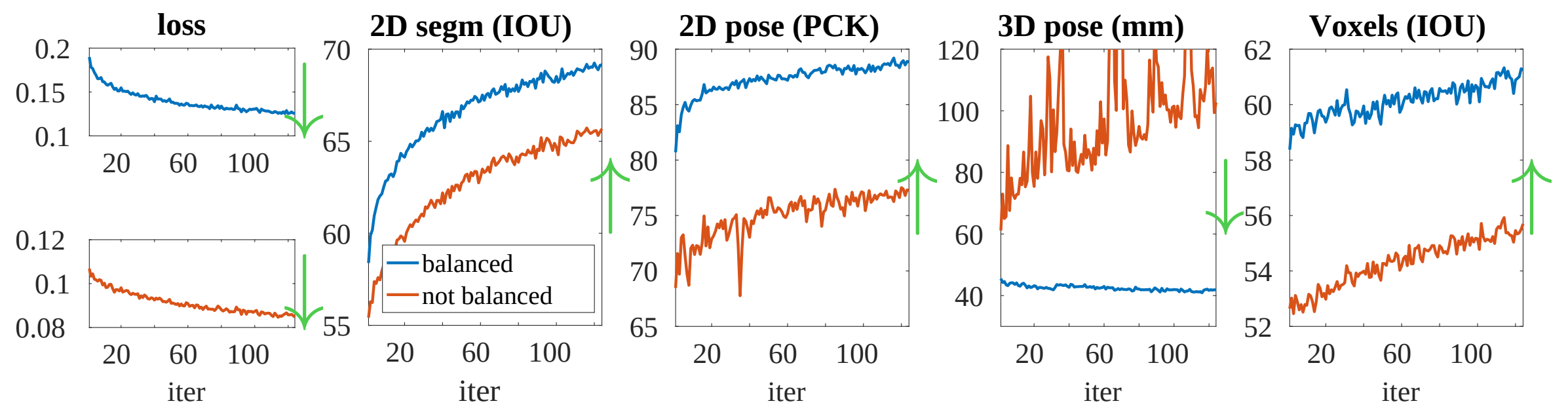


INTERMEDIATE TASKS

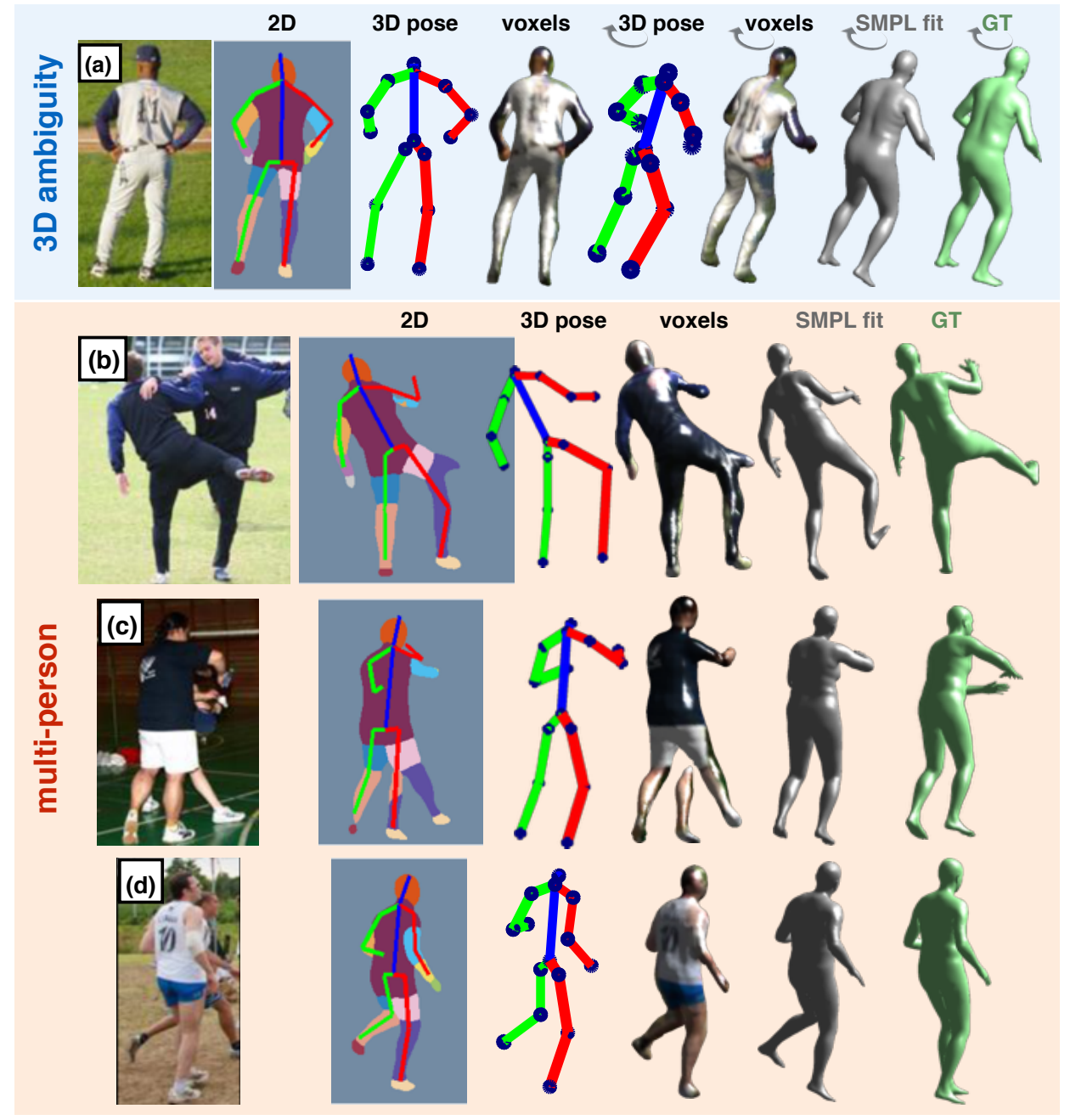
- **All tasks improve** with end-to-end training.

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

- Weight **balancing** is important.



LIMITATIONS



CONCLUSIONS

- Volumetric representation is **flexible** and effective.
- Re-projection loss is critical to obtain **confident body surface**.
- Multi-task training of **relevant tasks** helps.

REFERENCES

- *SMPLify:* Bogo et al. ECCV 2016
- *Unite the People:* Lassner et al. CVPR 2017
- *Indirect learning:* Tan et al. BMVC 2017
- *Self-supervised learning:* Tung et al. NIPS 2017
- *HMR:* Kanazawa et al. CVPR 2018

Code is available!

http://www.di.ens.fr/willow/research/bodynet

gul.varol@inria.fr