3DV 2018

Neural Body Fitting: Unifying Deep Learning and Model-Based Human Pose and Shape Estimation

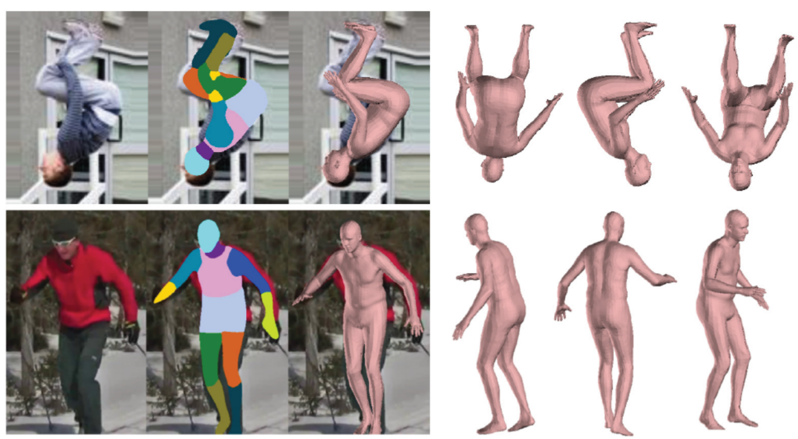

M. Omran


C. Lassner


G. Pons-Moll


P.V. Gehler


B. Schiele



max planck institut informatik

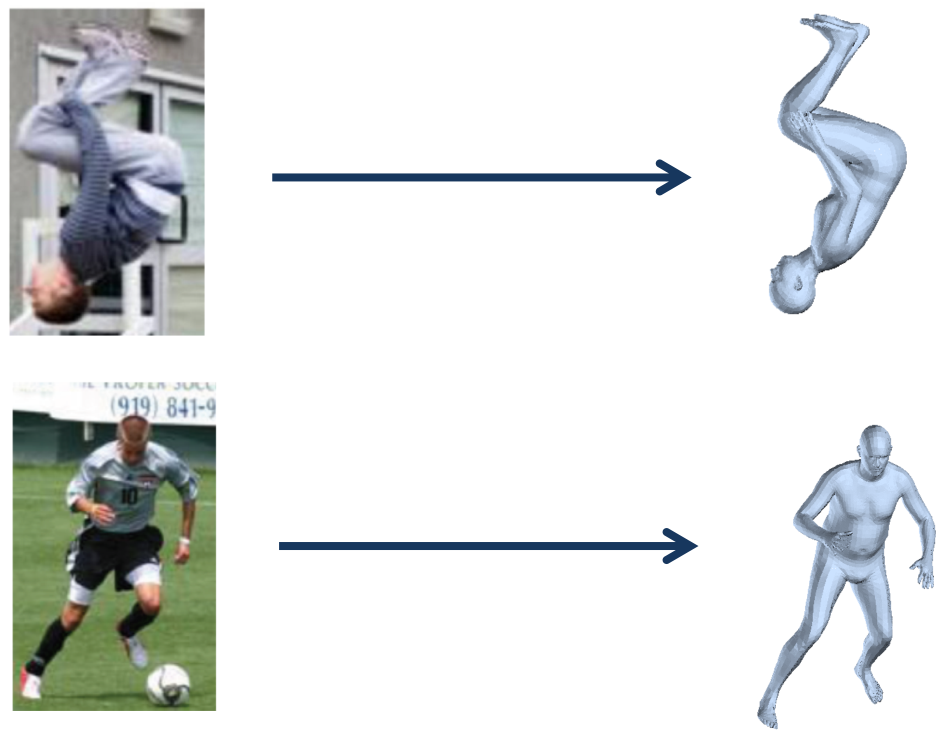
SIC Saarland Informatics Campus

Max Planck Institute for Intelligent Systems

amazon

Goal

predict full 3D human body mesh from a single 2D image



Input (2D) Output (3D)

Model-Based Approaches

starting point: parametrized body model (e.g. SMPL)

$M(\boldsymbol{\theta}, \boldsymbol{\beta})$ mesh parametrized by pose θ and shape β

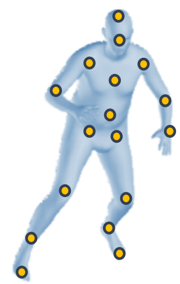
Loper et al., 2015

Model-Based Approaches

$$\arg\min_{\boldsymbol{\theta},\boldsymbol{\beta}} \operatorname{dist}\left(\hat{\mathbf{z}}\left(M(\boldsymbol{\theta},\boldsymbol{\beta})\right),\mathbf{z}\right)$$

3D world

2D keypoints $\mathbf{z}$



$P(\cdot)$

$$\hat{\mathbf{z}}\left(M(\boldsymbol{\theta},\boldsymbol{\beta})\right)$$

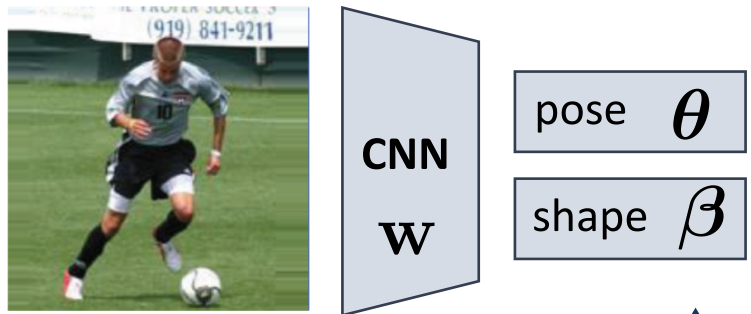
Bogo et al. '16
Lassner et al. '17

Optimization can be **slow and complicated**

Optimization requires **careful initialization**

Learning-Based Approaches

2D Input 3D Output

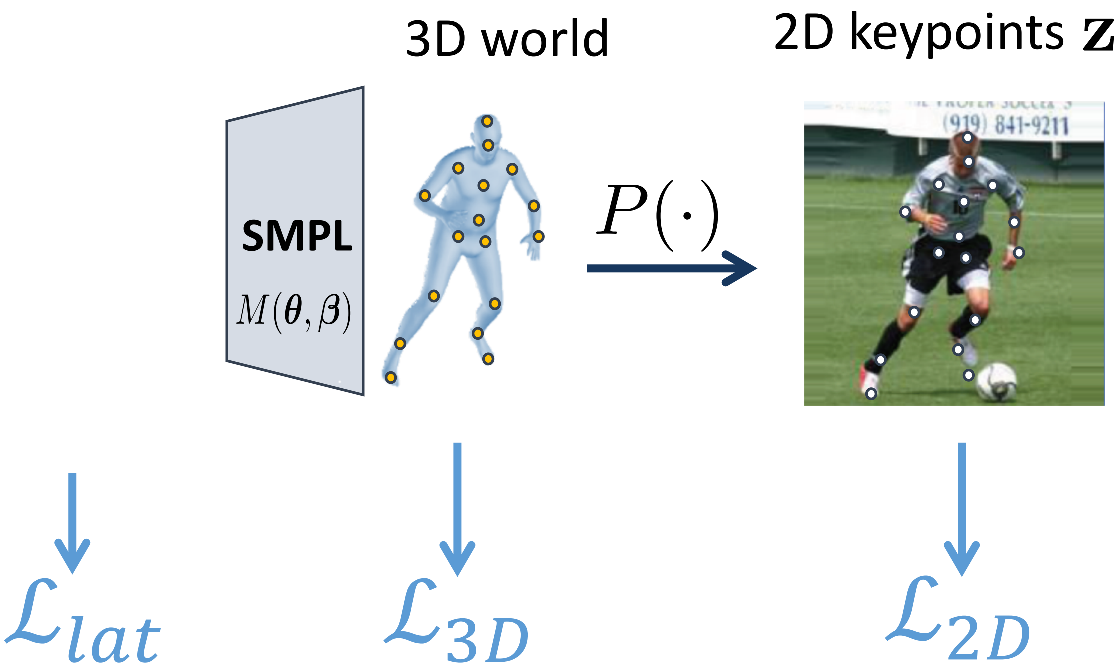


Training data hard to obtain!

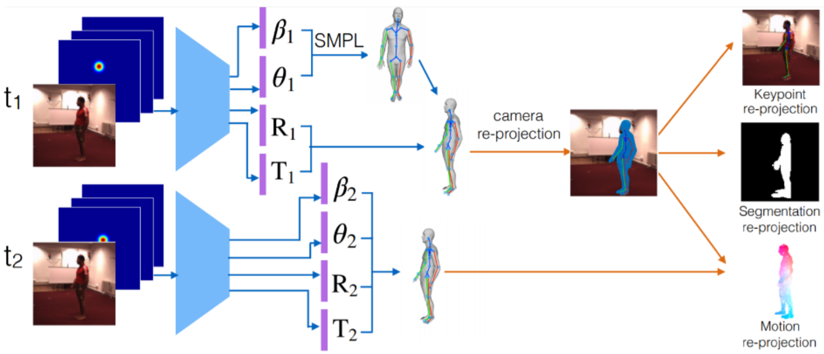
Also: no feedback between estimates and observations

Our Hybrid Approach

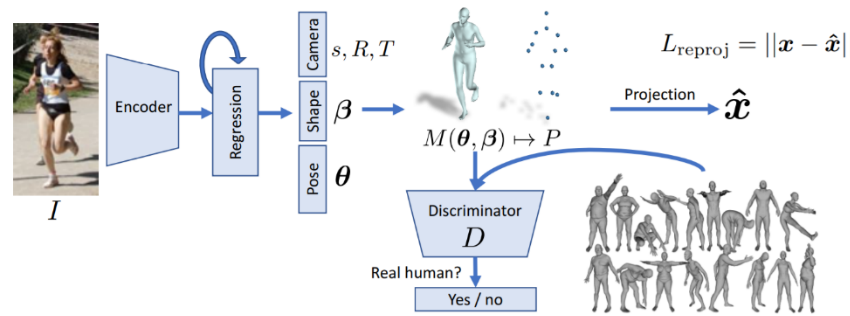
combines aspects of model- and learning-based approaches



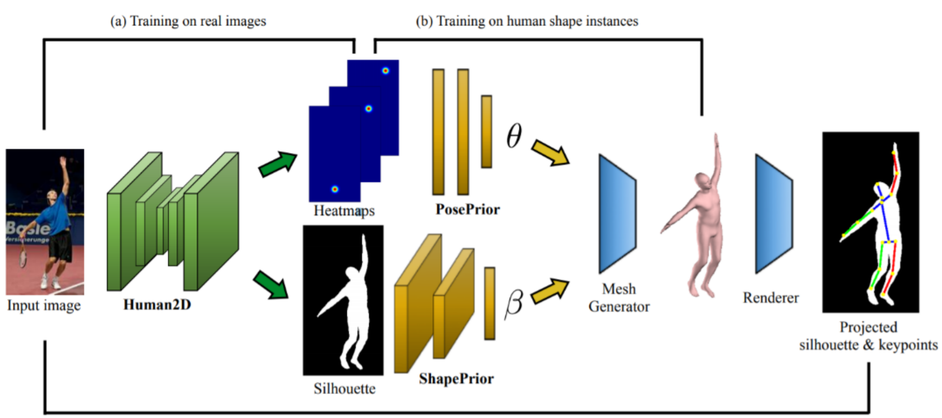
Other Hybrid Approaches



(Tung et al. 2017)

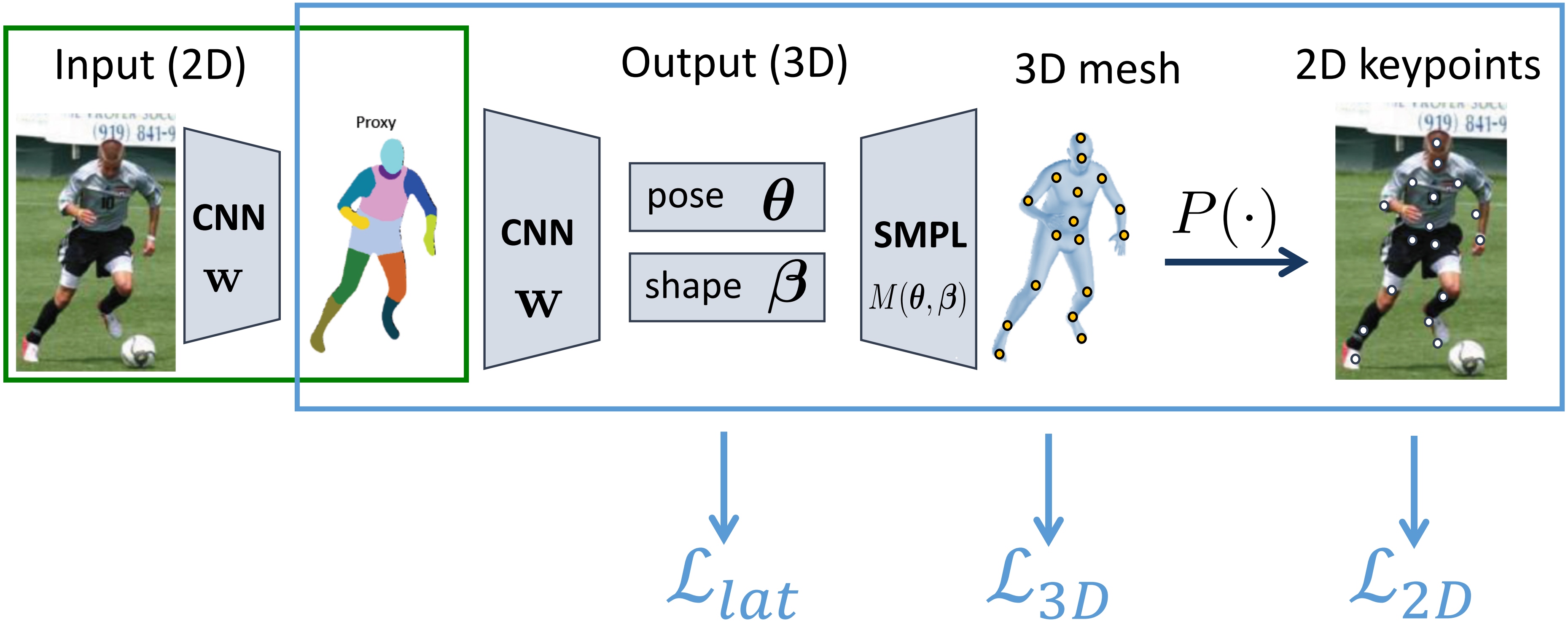


(Kanazawa et al. 2018)



(Pavlakos et al. 2018)

Our Hybrid Approach



1) Use intermediate 2D representation?
2) Amount of 2D vs 3D supervision?

Comparisons to State-of-the-Art

- Dataset: Human3.6M
- Error metric: mean per-joint error (in mm) after Procrustes Alignment

Class	Method	Mean	Median
Learning-Based	Akhter & Black [1]	181.1	158.1
	Ramakrishna et al. [45]	157.3	136.8
	Zhou et al. [68]	106.7	90.0
Model-Based	SMPLify [6]	82.3	69.3
	SMPLify (dense) [24]	80.7	70.0
Hybrid	SelfSup [64]	98.4	-
	Pavlakos et al. [38]	75.9	-
	HMR (H36M-trained) [22]	77.6	72.1
	HMR [22]	**56.8**	-
	Ours (H36M-trained)	59.9	52.3

Experimental Analysis

Datasets

Unite the People (UP)
(Lassner et al., 2017)

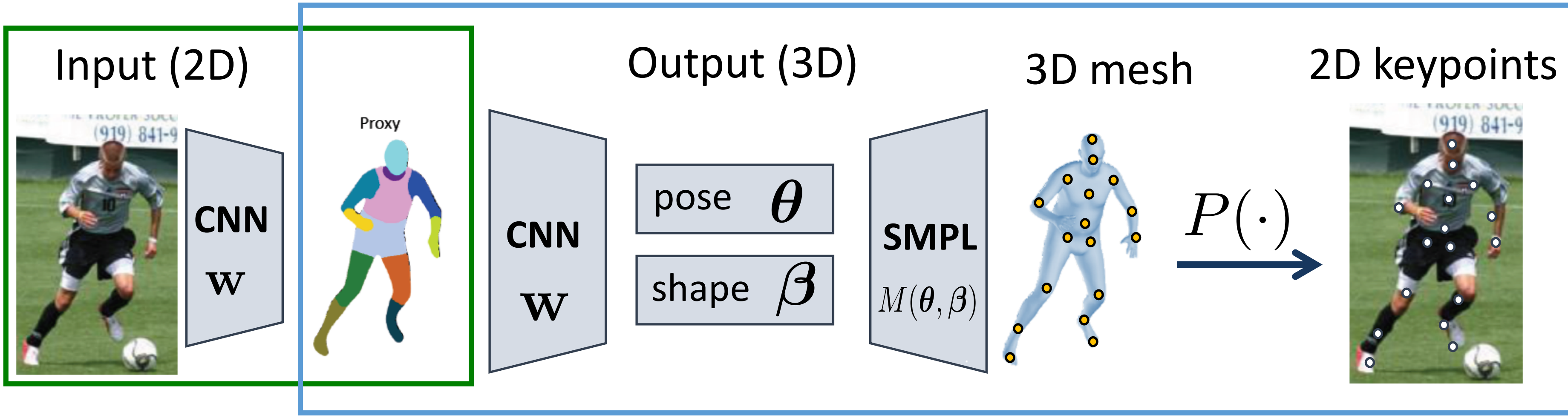
- “in-the-wild”
- 8126 images
- SMPL fits provided

(image from Hossain, 2017)

Human3.6M
(Ionescu et al., 2014)

- controlled environment
- 210 video sequences
  7 subjects / 2*15 actions
- MoCap data provided, SMPL fits
  via MoSH (Loper et al., 2015)

Our Hybrid Approach

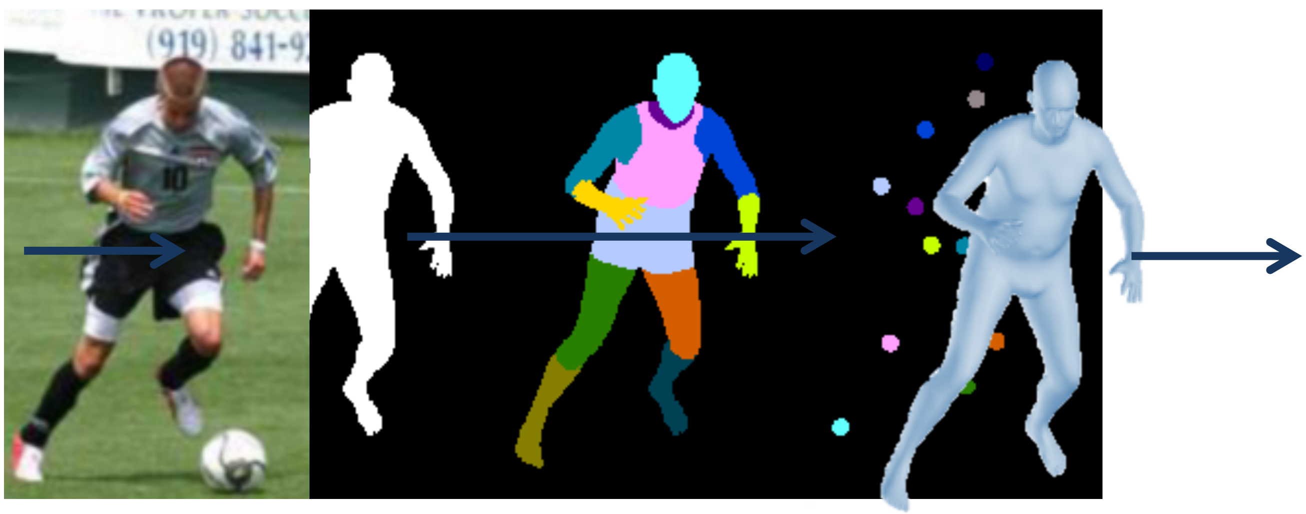


- training:
  - separate training for both components
  - up to 12 hours + 6 hours (Volta V100)
  - scale information constrained to shape parameters (camera parameters and distance to observer assumed fixed)
- test time: 0.2s (segmentation) + 0.05s (fitting)

Input Representation

Mapping directly from 2D image to 3D shape and pose is challenging

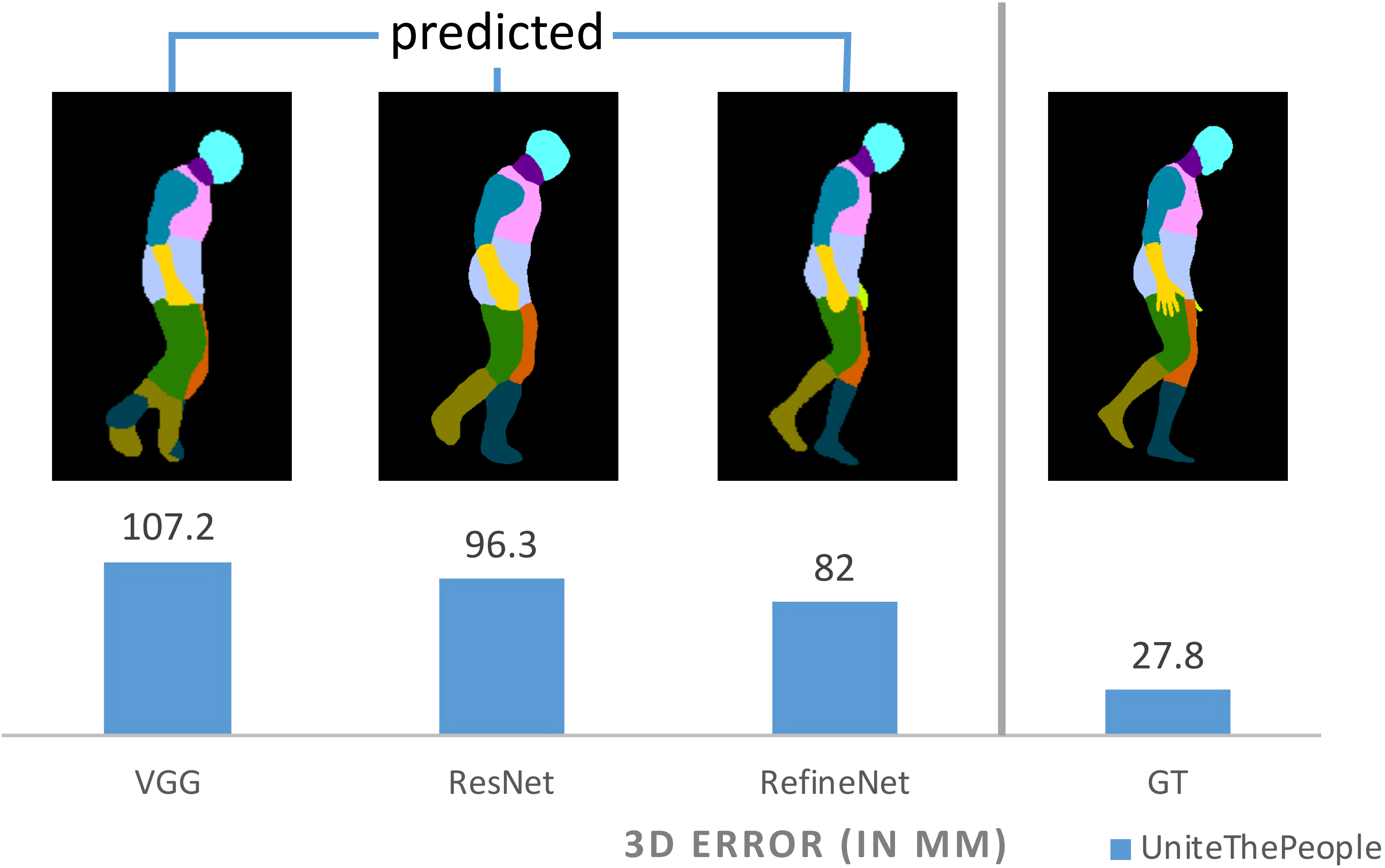
Input (2D) Proxy representation Output (3D)



Would an intermediate representation help?
If yes, which?

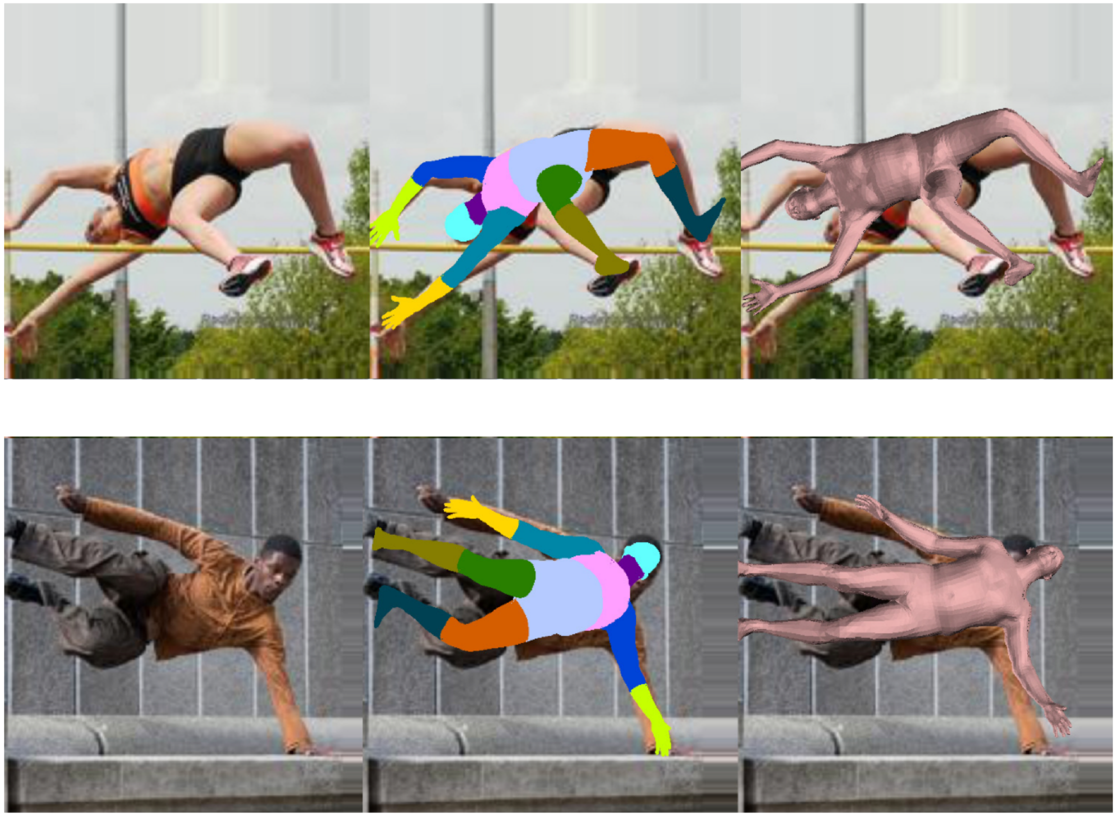
Input Representation

Segmentation Quality



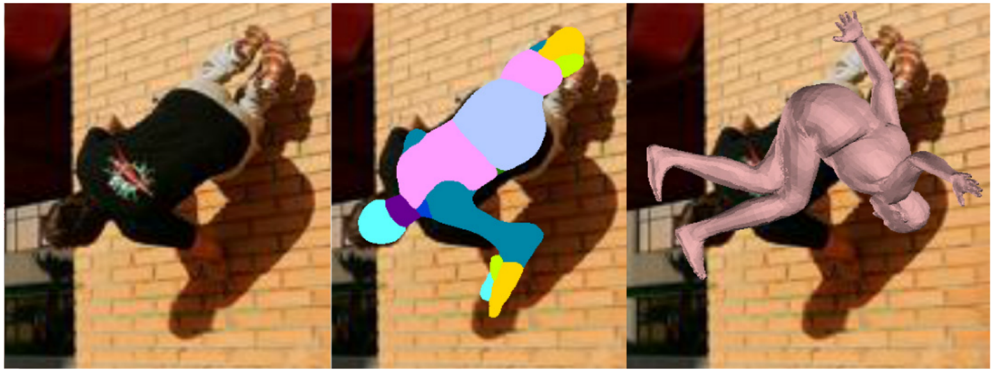
Segmentation Quality Matters

Worst fits when using ground truth segmentations:

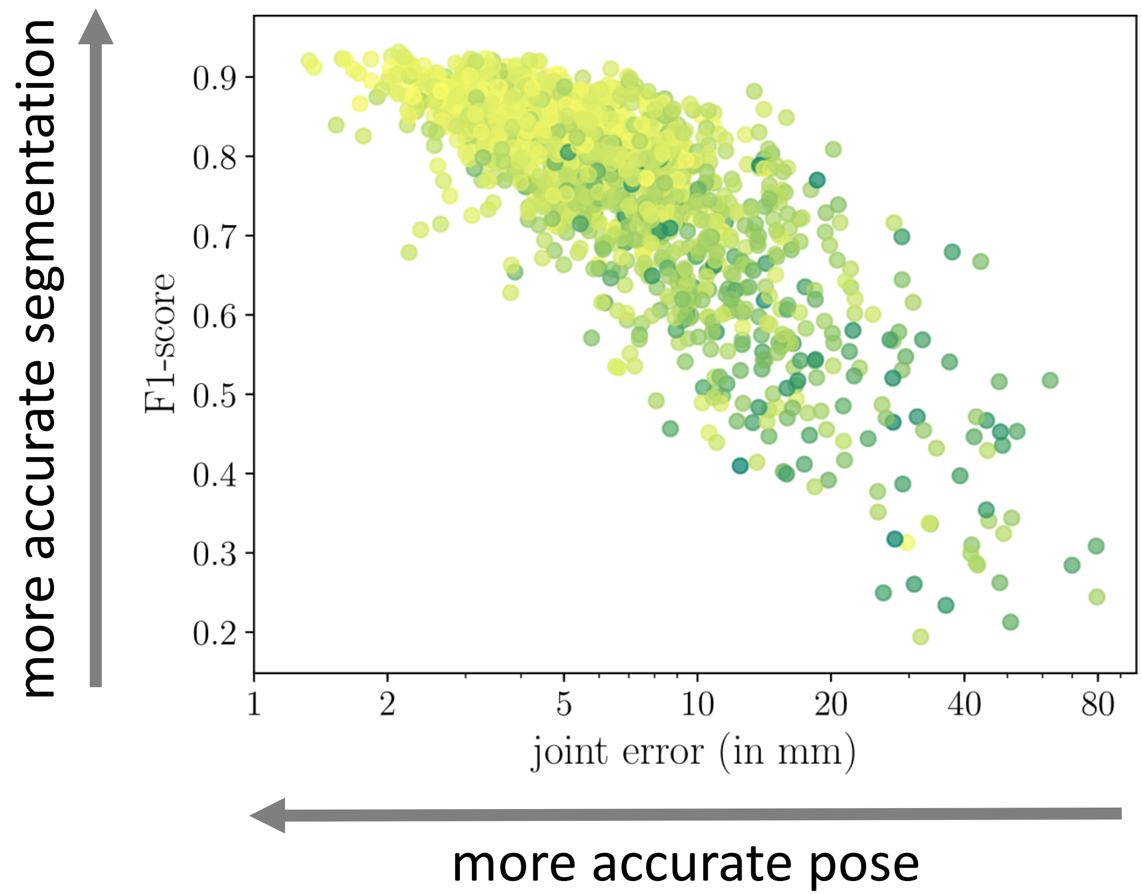


Segmentation Quality Matters

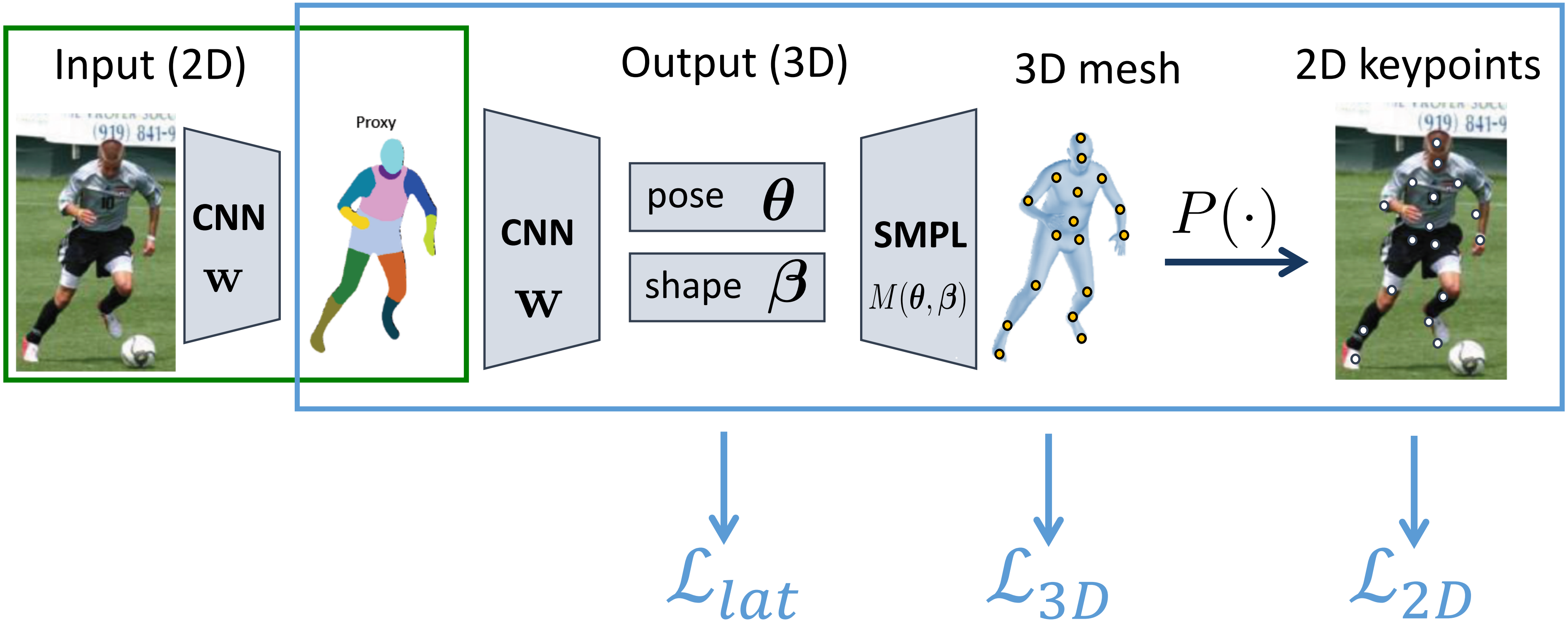
Worst fits when using predicted segmentations:



Pose vs. Segmentation Accuracy



Our Hybrid Approach



1) Use intermediate 2D representation?
2) Amount of 2D vs 3D supervision?

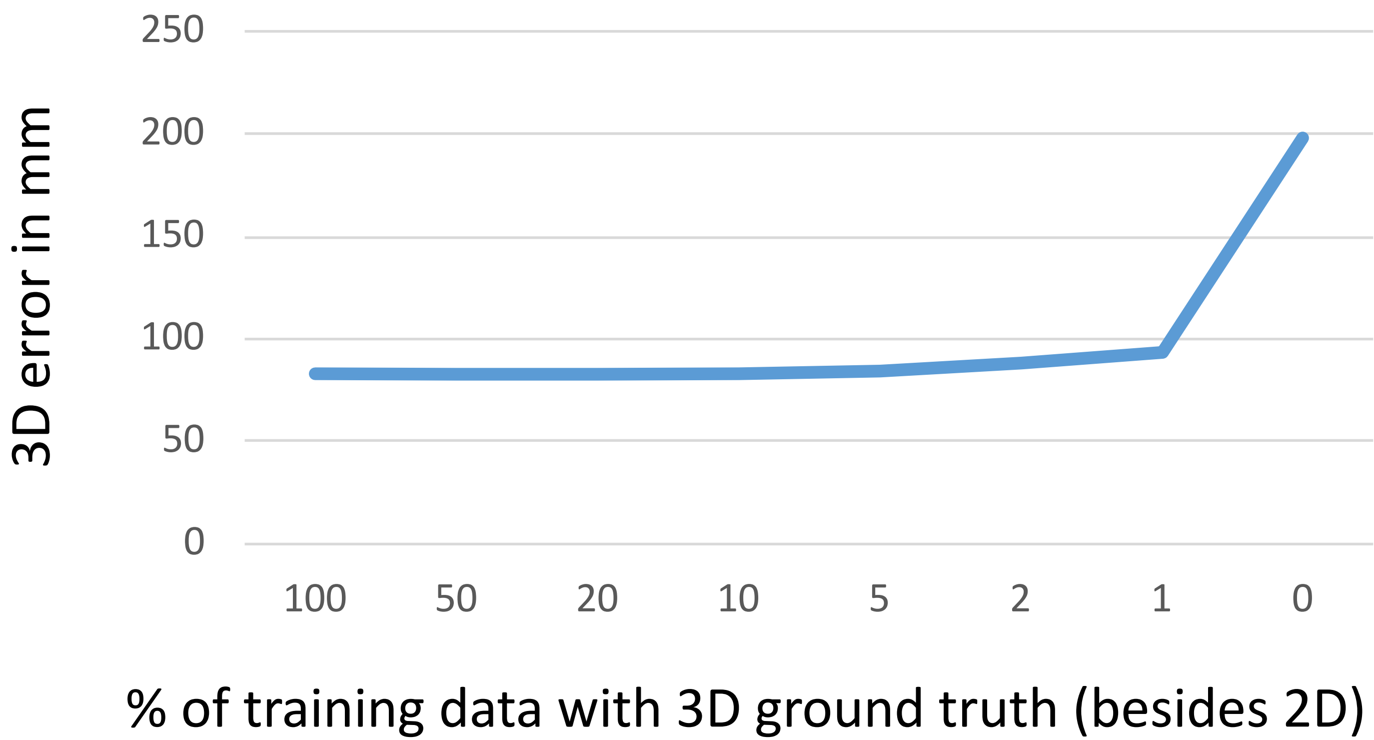
Which Type of Supervision

	Errors		
Loss	**3D joints (in mm)**	**2D joints (PCKh)**	**joint rotation (in quaternions)**
$\mathcal{L}_{2D}$	198.0	94.0	1.971
$\mathcal{L}_{3D}$	83.7	93.5	1.962
$\mathcal{L}_{lat}$	83.7	93.1	0.278
$\mathcal{L}_{lat} + \mathcal{L}_{3D} + \mathcal{L}_{2D}$	82.0	93.5	0.279

- supervising with SMPL parameters:
  -> better joint localization (in 2D and 3D) + joint rotations

How Much 3D Supervision?

Experiment: given training data with 2D ground truth (keypoints)
vary size of subset that also has 3D ground truth (shape/pose)



Qualitative Results

Conclusions

- Our hybrid method combines aspects of model-based and learning-based approaches to address some shortcomings of both

- Using intermediate part-based representation provides a helpful abstraction for predicting shape and pose.

- A small amount of 3D annotations are already useful when used in conjunction with 2D annotations

Future Work

- introducing test-time refinement to fully leverage the incorporated model and improve over the strong initial estimate we provide
- closer integration of the localization, segmentation and fitting components
- addressing alignment / estimation of absolute scale
- considering multiple (possibly occluded) people




Thank you for your Attention!

Code available here soon:

www.github.com/mohomran/neural_body_fitting