DenseRaC: Joint 3D Pose and Shape Estimation by Dense Render-and-Compare

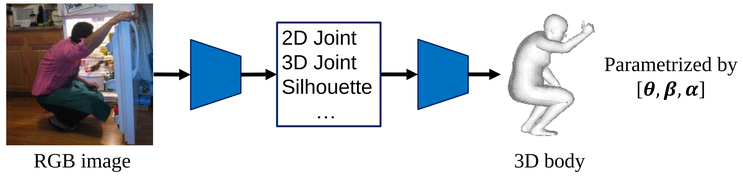
Yuanlu Xu[1,2], Song-Chun Zhu[2], Tony Tung[1] [1]Facebook Reality Labs, Sausalito [2]University of California, Los Angeles (UCLA) merayxu@gmail.com, sczhu@stat.ucla.edu, tony.tung@fb.com

Motivation

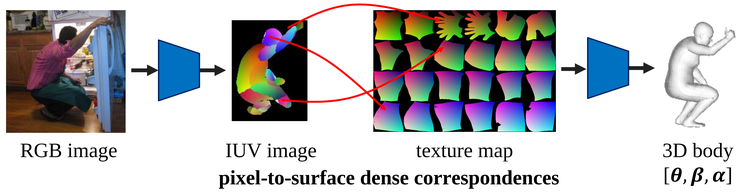
In this paper, we propose an end-to-end framework for jointly estimating 3D human pose and body shape from a monocular RGB image.

- **A Two-Step Framework**

Existing frameworks: direct regression, little information about **body details**

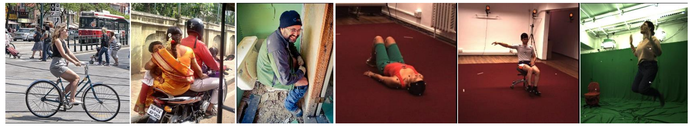


Our framework: **IUV map** as proxy, richer weak supervisions, less nuisances (light, shading texture)



- Augmented Training from **Synthetic Data**

Existing framework: lacking paired training data (w./o. **ground truth mesh**)



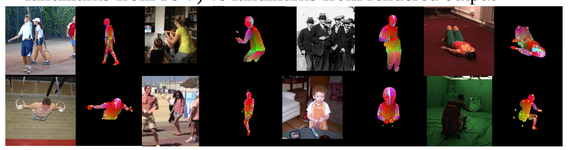
Our framework: augmented training from **fully paired** synthetic data

MOCA Dataset: large-scale synthetic dataset (2M+ images) from covering various camera views, human actions and body shapes, with fully paired ground truth.

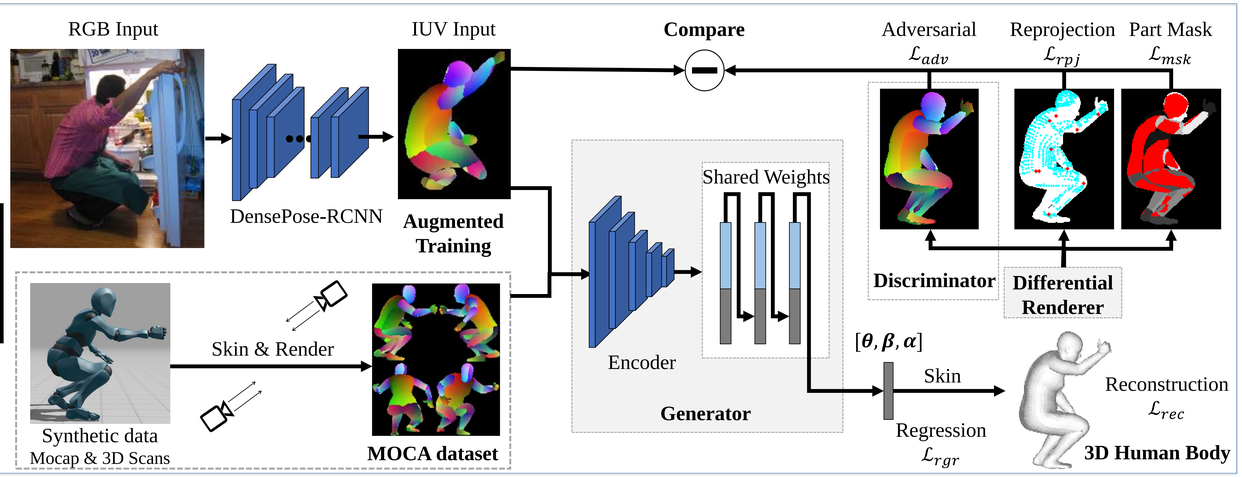
- Poses from web-crawled Mocap sequences and 3D animations (260K frames)
- Body shapes 3D human scans (CAESAR dataset, 2,781 samples)
- Random camera view, cropping and scaling.

A Render-and-Compare Framework

- Minimizing differences between input and rendered output
- **Regression Loss*** L_{rgr}: GT parameters vs. regressed parameters
- **Reconstruction Loss*** L_{rec}: GT mesh/skeleton vs. reconstructed mesh/skeleton
- **Reprojection Loss** L_{rpj}: GT 2d landmarks (joints, dense landmarks from IUV) vs landmarks from rendered output



- **Part Mask Loss** L_{msk}: GT part masks vs part masks from rendered output
- **Adversarial Loss** L_{adv}: real/fake 3D body configurations
- **Training Set**: MOCA, LSP, MPII, COCO, H3.6M, MPI-INF-3DHP.

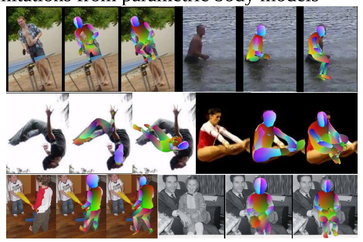


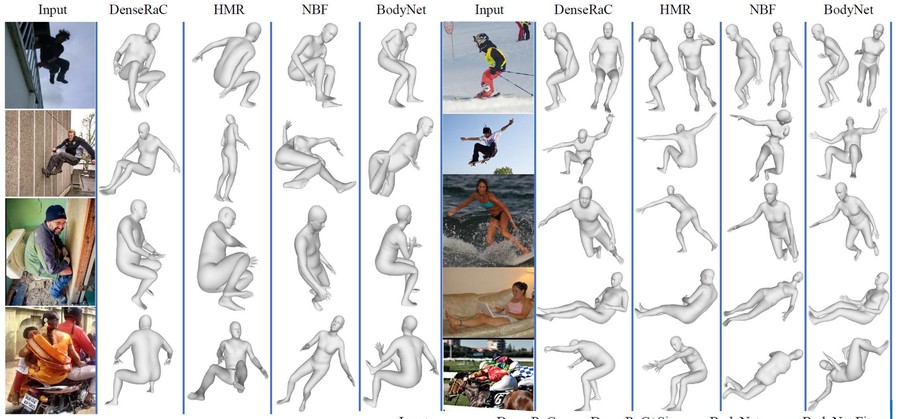
Experiments

- DenseRaC obtains superior performance on human-related tasks: **3D pose estimation** (Human3.6M, MPI-INF-3DHP), **semantic segmentation** (UP-3D, MOCA).
- DenseRaC obtains more accurate poses and body shapes (e.g., slim, overweight), more natural hands (opened, clenched, etc.)

Failure cases:

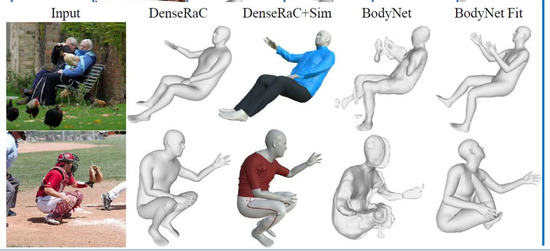
Heavy occlusions, errors from DensePose, limitations from parametric body models





Application: Virtual Dressing

- Under-clothes body reconstruction + clothes simulation
- Volumetric reconstruction + under-clothes body mesh fitting
- **Fidelity** (visually plausible vs. details)
- **Robustness** (crowd, occlusions)
- **Flexibility** (beyond reconstruction)



3D Pose Estimation

H3.6M	Protocol #1	Protocol #2	Protocol #3
	MPJPE	MPJPE	MPJPE
Lassner *et al.* (CVPR'17) [26]	80.7	-	-
HMR (CVPR'18) [22]	88.0	56.8	77.3
Pavlakos *et al.* (CVPR'18) [42]	-	75.9	-
NBF (3DV'18) [39]	-	59.9	-
DenseRaC baseline	82.4	53.9	77.0
+ render-and-compare	79.5	51.4	75.9
+ synthetic data	76.8	48.0	74.1

MPI-INF-3DHP	Protocol #1			Protocol #2		
	PCK	AUC	MPJPE	PCK	AUC	MPJPE
Mehta *et al.* (3DV'17) [36]	75.7	39.3	117.6	-	-	-
Mehta *et al.* (TOG'17) [37]	76.6	40.4	124.7	83.9	47.3	98.0
HMR (CVPR'18) [22]	72.9	36.5	124.2	86.3	47.8	89.8
DenseRaC baseline	73.1	36.7	123.1	86.8	47.8	88.7
+ render-and-compare	74.7	38.6	124.9	87.5	48.3	86.7
+ synthetic data	**76.9**	**41.1**	**114.2**	**89.0**	**49.1**	**83.5**

Semantic Segmentation

UP-3D	Body Part		Fg/Bg	
	Accuracy	F1	Accuracy	F1
SMPL on DpCut (ECCV'16) [4]	87.7	0.64	91.9	**0.88**
SMPL, UP-P91 (ICCV'17) [26]	87.3	0.61	91.0	0.86
HMR (CVPR'18) [22]	87.1	0.60	91.7	0.87
BodyNet (ECCV'18) [55]	-	-	**92.8**	0.84
DenseRaC	**87.9**	**0.64**	92.4	**0.88**

MOCA	Body Part		Fg/Bg	
	Accuracy	F1	Accuracy	F1
HMR (CVPR'18) [22]	86.6	0.19	92.1	0.60
DenseRaC	**89.3**	**0.27**	**96.4**	**0.68**

Reference:

[1] DensePose – Guler et al, "DensePose: Dense human pose estimation in the wild", CVPR18.
[2] HMR – Kanazawa et al, "End-to-end recovery of human shape and pose", CVPR18.
[3] NBF – Omran et al, "Neural body fitting: Unifying deep learning and model-based human pose and shape estimation", 3DV18.
[4] BodyNet –Varol et al, "Bodynet: Volumetric inference of 3d human body shapes", ECCV18.