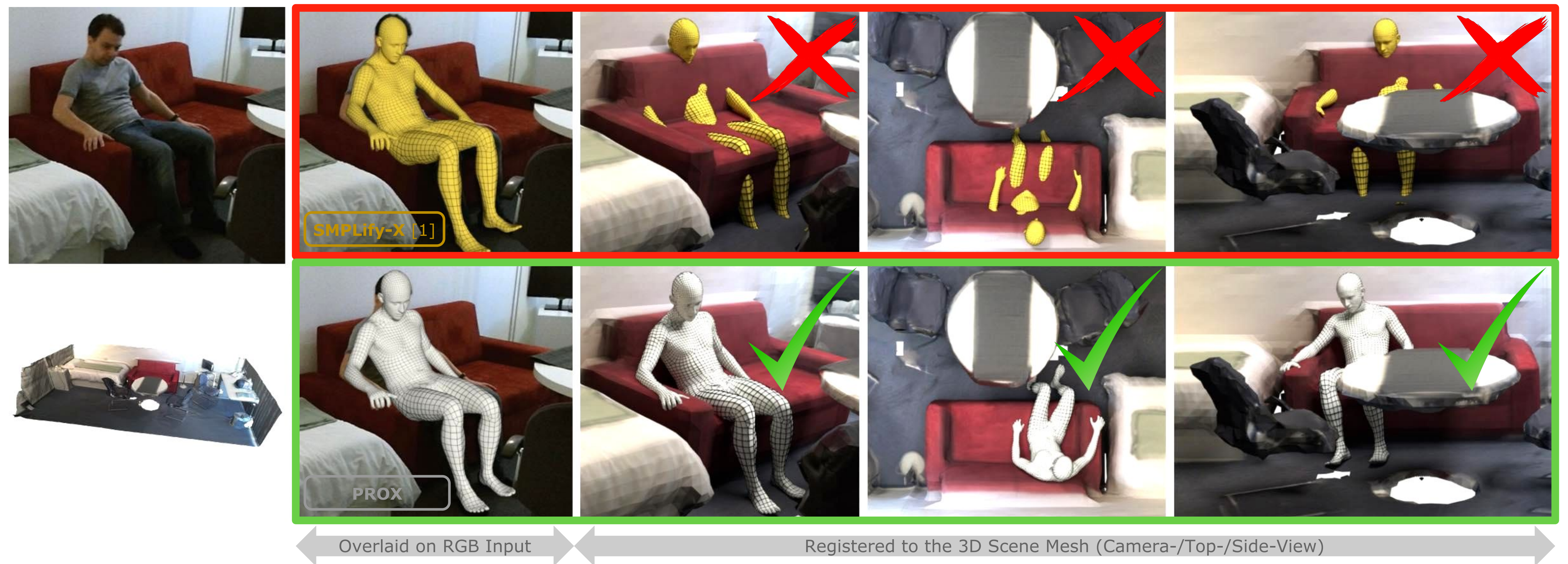
Resolving 3D Human Pose Ambiguities with 3D Scene Constraints

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas and Michael J. Black
Max Planck Institute for Intelligent Systems
prox@tuebingen.mpg.de

ICCV 2019 Seoul, Korea

prox.is.tuebingen.mpg.de
Data & Code

MAX-PLANCK-GESELLSCHAFT



Goal

- Capture humans moving in, and ***interacting*** with, the world.

- Most pose estimation methods, like **SMPLify-X** [1] do ***not*** take into account the scene.
- Therefore, estimated bodies are usually ***not consistent*** with the 3D scene.
- However, the world ***constrains*** the body, and vice-versa.

- We use ***3D scene*** knowledge to ***improve*** human pose estimation from single RGB images.
- Our method enforces *Proximal Relationships with Object eXclusion* & is called ***PROX***.
- We formulate two types of ***scene constraints***:
  - ***Penetration*** constraint.
  - ***Contact*** constraint.

PROX Dataset

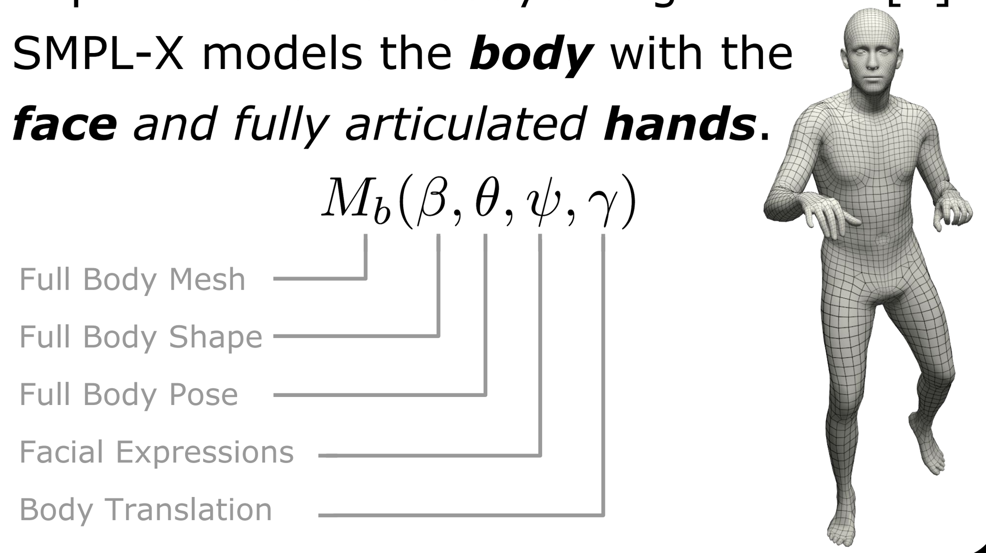
- 12 reconstructed *scene meshes* M_s.
- With Occipital Structure & Skanect.
- 20 subjects (16m/4f).
- Kinect-One camera.
- 100K RGB-D frames.
- Synced & aligned RGB & Depth cameras.
- RGB-D camera aligned to 3D scene.
- Capture humans interacting naturally with rigid scenes.

Formulation

SMPL-X

- **Interaction** = 3D *surfaces* in **contact**.
- *Skeletons* can ***not*** fully capture this [2,3].
- Represent human body using SMPL-X [1].
- SMPL-X models the ***body*** with the ***face*** and fully articulated ***hands***.

$$M_b(\beta, \theta, \psi, \gamma)$$



Monocular Fitting

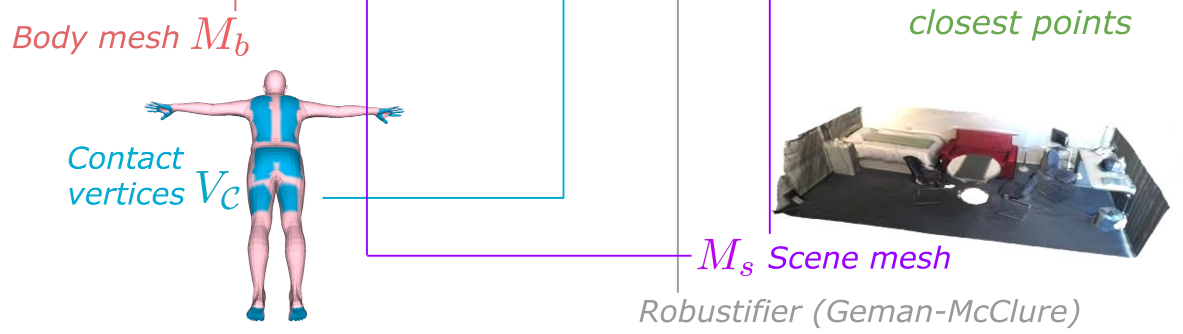
- Extend SMPLify-X [1] to include scene constraints.
- Minimize the objective function:

$$\begin{aligned}E(\beta,\theta,\psi,\gamma,M_s) = & E_J & \text{2D Joint Reprojection [2]}\\ & +\lambda_D E_D & \textit{(optional)}\text{ Depth Data}\\ & +\lambda_{\theta_b}E_{\theta_b} + \lambda_{\theta_f}E_{\theta_f} + \lambda_{\theta_h}E_{\theta_h} + \lambda_\alpha E_\alpha & \text{Priors}\\ & +\lambda_\beta E_\beta + \lambda_\mathcal{E} E_\mathcal{E} & \text{Priors}\\ & +\lambda_\mathcal{P} E_\mathcal{P} & \textit{Penetration Constraint}\\ & +\lambda_\mathcal{C} E_\mathcal{C} & \textit{Contact Constraint}\end{aligned}$$

Scene Constraints

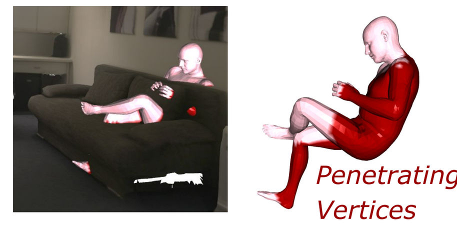
Contact $\mathcal{C}$: Encourage *proximity* between:
- likely *contact vertices* $V_\mathcal{C}$ of *body* M_b.
- the 3D *scene mesh* M_s.

$$E_\mathcal{C}(\beta,\theta,\gamma,M_s) = \sum_{v_\mathcal{C} \in V_\mathcal{C}} \rho_\mathcal{C}\Big(\min_{v_s \in V_s} \|v_\mathcal{C} - v_s\|\Big)$$



Penetration $\mathcal{P}$: Penalize body vertices that are *penetrating* the scene:

- Compute a voxel grid for each scene.
- Voxel p_i stores the distance d_i to the closest scene point $p_{s,i} \in \mathbb{R}^3$ of M_s with normal $n_{s,i}$.
- Signed distance: $d_i > 0 \rightarrow$ free space. $d_i = 0 \rightarrow$ lying on M_s. $\boxed{d_i < 0} \rightarrow$ penetrating M_s.



$$E_\mathcal{P} = \sum_{\boxed{d_i < 0}} \|d_i n_{s,i}\|^2$$

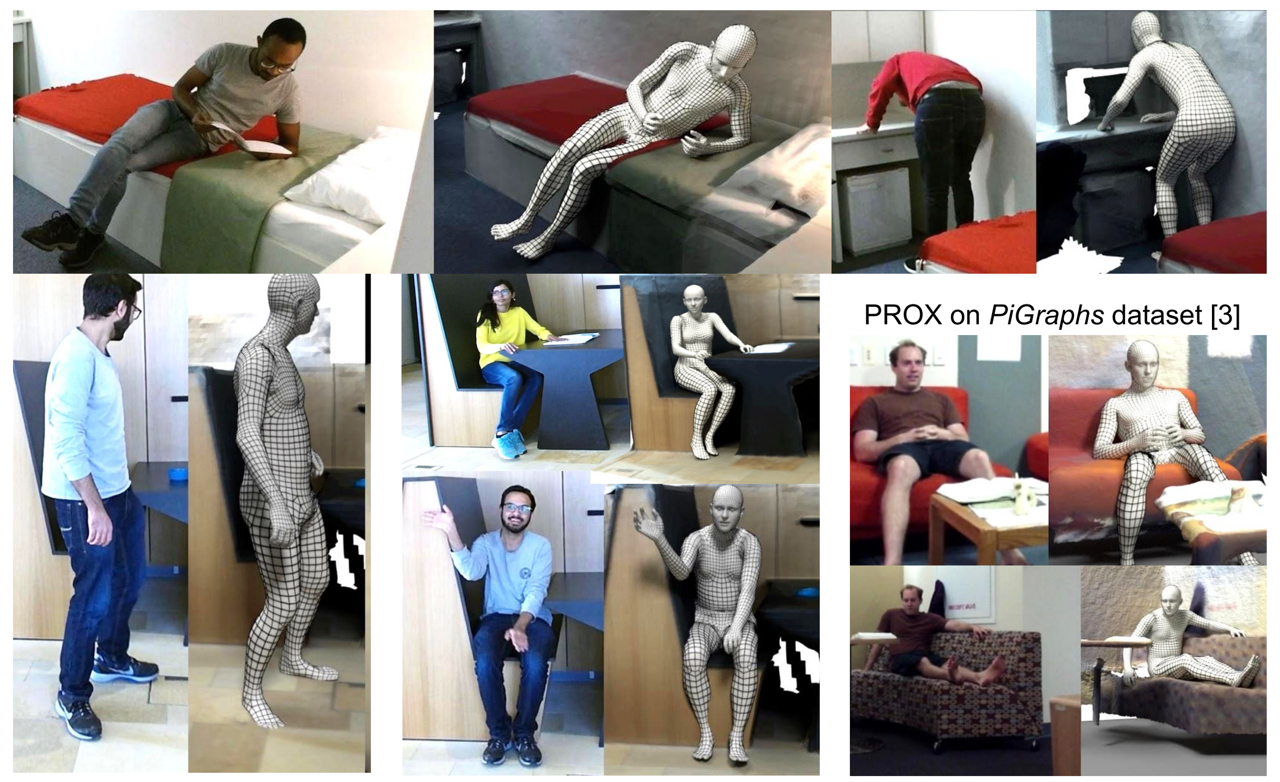
Optional Depth Term

Depth used *only* for generating pseudo ground-truth (GT) SMPL-X fits.

$$E_D = \sum_{p \in P} \rho_D\Big(\min_{v \in V_b^v} \|v - p\|\Big)$$

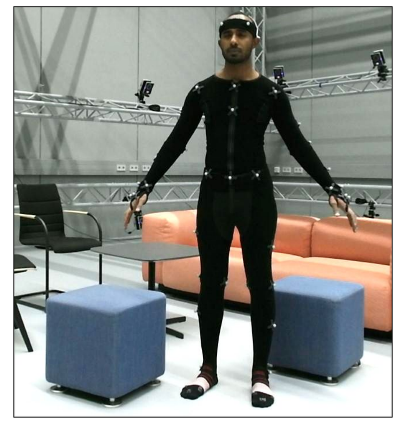
Body-to-PCL closest points
Visible body vertices
Robustifier (Geman-McClure)
Depth point cloud (PCL)

Qualitative Evaluation



Quantitative Evaluation

- Vicon MoCap system.
- 54 high-res cameras.
- Living room in the capture space.
- 180 RGB-D frames.
- Pseudo ground-truth SMPL-X meshes with MoSh++ [4].



Objective fn terms				Error (mm)				
E_J	$E_\mathcal{C}$	$E_\mathcal{P}$	E_D	PJE	V2V	p.PJE	p.V2V	
✓	✗	✗	✗	220.27	218.06	73.24	60.80	
✓	✓	✗	✗	208.03	208.57	72.76	60.95	
✓	✗	✓	✗	190.07	190.38	73.73	62.38	
✓	✓	✓	✗	167.08	166.51	71.97	61.14	→ PROX
✓	✗	✗	✓	72.91	69.89	55.53	48.86	
✓	✓	✓	✓	68.48	60.83	52.78	47.11	→ PROX-D

References

[1] Pavlakos et al. *Expressive Body Capture: 3D Hands, Face, and Body from a Single Image* CVPR 2019
[2] Cao et al. *OpenPose: Realtime Multi-person 2D Pose Estimation using Part Affinity Fields* TPAMI 2019
[3] Savva et al. *PiGraphs: Learning Interaction Snapshots from Observations* SIGG 2016
[4] Mahmood et al. *AMASS: Archive of Motion Capture as Surface Shapes* ICCV 2019