Populating 3D Scenes by Learning Human-Scene Interaction

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, Michael J. Black
Max Planck Institute for Intelligent Systems
MAX-PLANCK-GESELLSCHAFT

CVPR VIRTUAL JUNE 19-25

Nutshell

Goal

- Learn how humans interact with the scene.
- Enable virtual characters to do the same.

What is POSA?

- A novel body-centric human scene interaction model.

Applications

- Place 3D people in 3D scenes
- Improve monocular pose estimation.

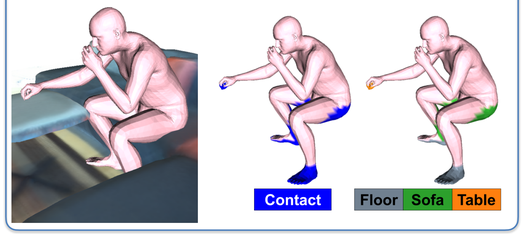
Representation

From GT extract:

$$f : (V_b, M_s) \rightarrow [f_c, f_s]$$

V_b : Body vertices f_c : Contact label

M_s :Scene mesh f_s : Semantic label



Learning

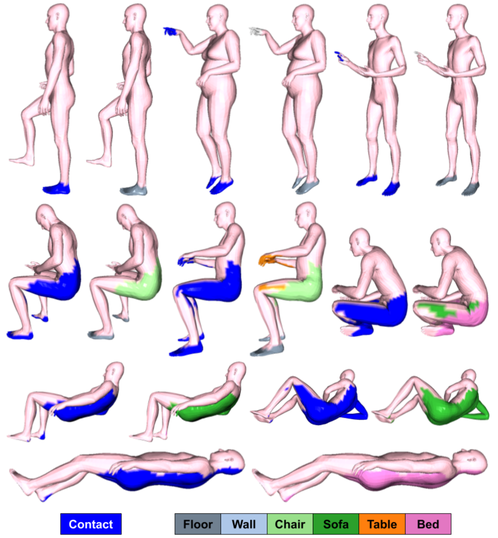
- Learn the mapping f.
- Train a CVAE

$$\mathcal{L}_{\text{total}} = \alpha * \mathcal{L}_{KL} + \mathcal{L}_{rec}$$

- Architecture is based on Spiral convolution and fully connected layers.

Random Sampling

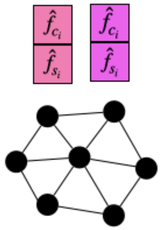
$$P(f_{Gen} \mid z, V_b)$$



Scene Population

1. Generate a feature map $P(f_{Gen} \mid z, V_b)$
2. Optimize

$$E(\tau, \theta_0, \theta) = \mathcal{L}_{afford} + \mathcal{L}_{pen} + \mathcal{L}_{reg}$$



τ: body translation

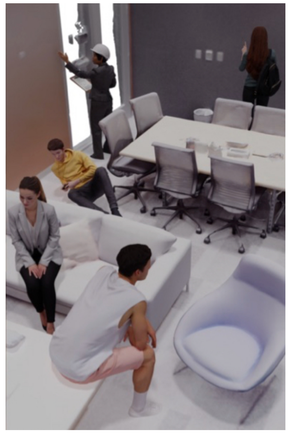
θ_0: global body orientation

θ: body pose

$$\mathcal{L}_{afford} = \lambda_1 * ||f_{Gen_c} \cdot f_d||_2^2 + \lambda_2 * \sum_i \text{CCE}\left(f_{Gen_s}^i, f_s^i\right)$$

Scene Population

- f_d and f_s are the observed distance and semantic features
- Penetration penalty $\mathcal{L}_{pen} = \lambda_{pen} * \sum_{f_d^i<0} \left(f_d^i\right)^2$
- Regularization $\mathcal{L}_{reg} = \lambda_{reg} * ||\theta - \theta_{\text{init}}||_2^2$
- θ_{init} : Initial pose



Eval

	POSA-variant ↑	PLACE ↓
POSA (contact only)	60.7%	39.3%
POSA (contact + semantics)	**61.0%**	39.0%

Pose Estimation

Replace hand-crafted contact features of PROX with learned POSA feature maps.

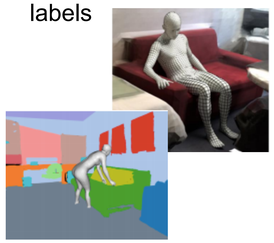
$$E(\beta, \theta, \psi, \tau, M_s) = E_{\text{smplifyX}} + ||f_{Gen} \cdot f_d|| + \mathcal{L}_{pen}$$

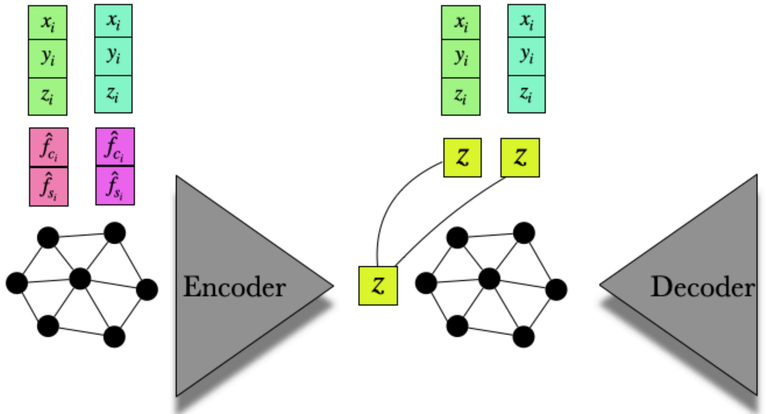
(mm)	PJE ↓	V2V ↓
RGB	220.27	218.06
PROX	167.08	166.51
POSA	**154.33**	**154.84**

Training Data

PROX: Pseudo GT SMPL-X meshes in 3D scenes

PROX-E: Semantic labels





References

1. Hassan et al. Resolving 3D Human Pose Ambiguities with 3D Scene Constraints.
2. Zhang et al. Generating 3D People in Scenes without People.
3. Pavlakos et al. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image.
4. Patel et al. AGORA: Avatars in Geography Optimized for Regression Analysis.
5. Bouritsas et al. Spiral convolutional networks for 3D shape representation learning and generation.
6. Zhang et al. PLACE: Proximity learning of articulation and contact in 3D environments.