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

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.

Scene Population
- \( f_d \) and \( f_s \) are the observed distance and semantic features.
- Penetration penalty \( \mathcal{L}_{pen} = \lambda_{pen} \sum (f_d)^2 \).
- Regularization \( \mathcal{L}_{reg} = \lambda_{reg} \| \theta - \theta_{init} \|_2^2 \).
- \( \theta_{init} \): Initial pose

Training Data
- PROX: Pseudo GT
- SMPL-X meshes in 3D scenes
- PROX-E: Semantic labels

Random Sampling
\[ P(f_{\text{Gen}} | z, V_b) \]

Random Sampling
- Generate a feature map \( P(f_{\text{Gen}} | z, V_b) \).
- 1. Optimize
- 2. Replace hand-crafted contact features of PROX with learned POSA feature maps.

Scene Population
- \( \mathcal{L}_{afford} \)
- \( \mathcal{L}_{afford} = \lambda_{afford} \sum || f_{Gen} - f_d ||_2^2 + \lambda_2 \sum_{i} \text{CCE} (f_{Gen}, f_i) \)

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.

References
1. Hassan et al. Resolving 3D Human Pose Ambiguities with 3D Scene Constraints.

Pose Estimation
- Replace hand-crafted contact features of PROX with learned POSA feature maps.
- \( E(f, \theta, \psi, \tau, M_s) = E_{\text{smplifyX}} + \mathcal{L}_{afford} + \mathcal{L}_{pen} \)

References
1. Hassan et al. Resolving 3D Human Pose Ambiguities with 3D Scene Constraints.