UCLA

Inferring Forces and Learning Human Utilities From Videos

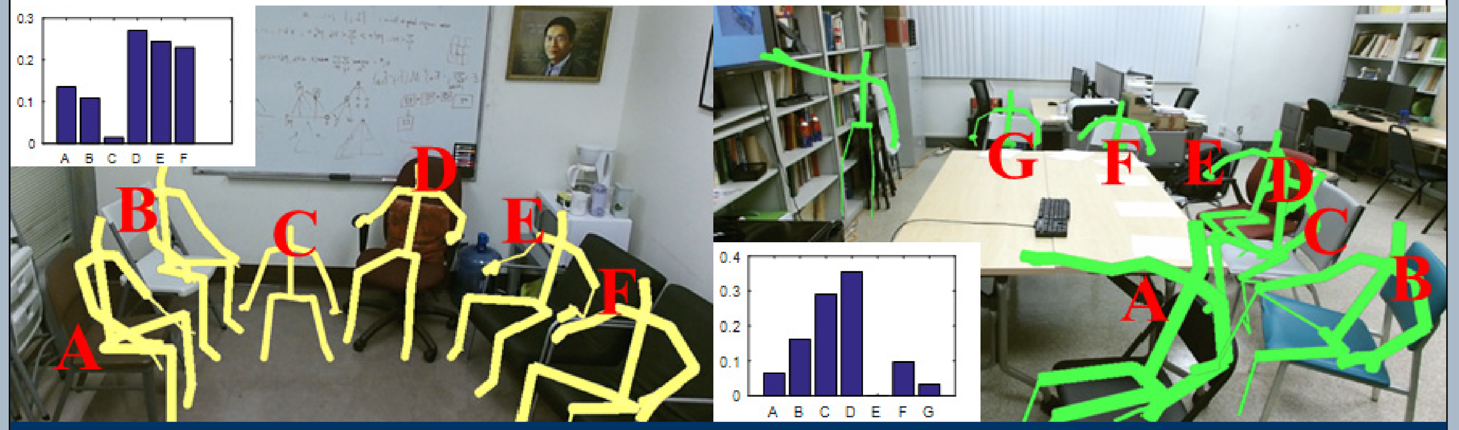
Yixin Zhu*♦ Chenfanfu Jiang*‡ Yibiao Zhao♦ Demetri Terzopoulos‡ Song-Chun Zhu♦ (* equal contribution)

♦ UCLA Center for Vision, Cognition, Learning, and Art ‡ UCLA Computer Graphics & Vision Laboratory

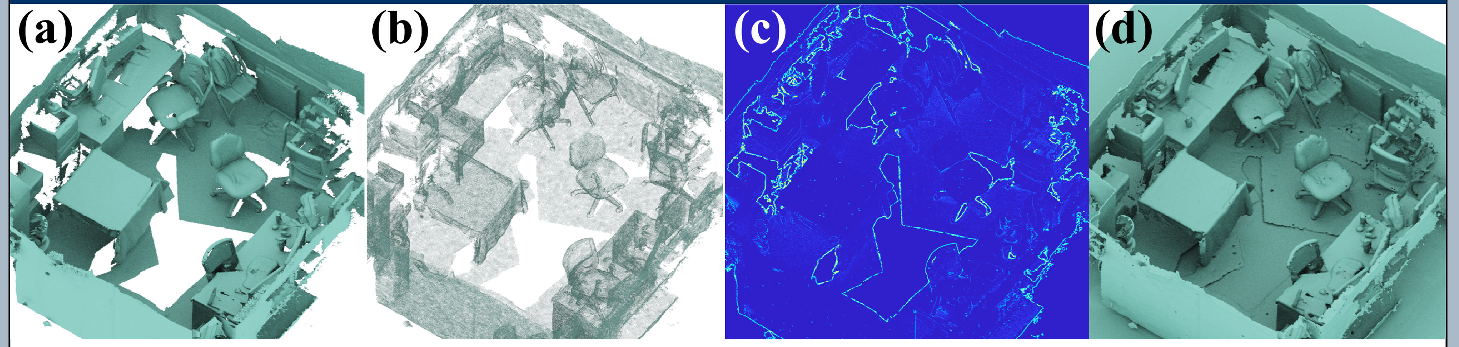
WELCOME TO Fabulous CVPR 2016 LAS VEGAS

Motivation

We propose a notion of affordance that takes into account **physical quantities** generated when the human body interacts with real-world objects, and introduce a learning framework that incorporates the concept of **human utilities**, which in our opinion provides a deeper and finer-grained account not only of object affordance but also of people's interaction with objects.

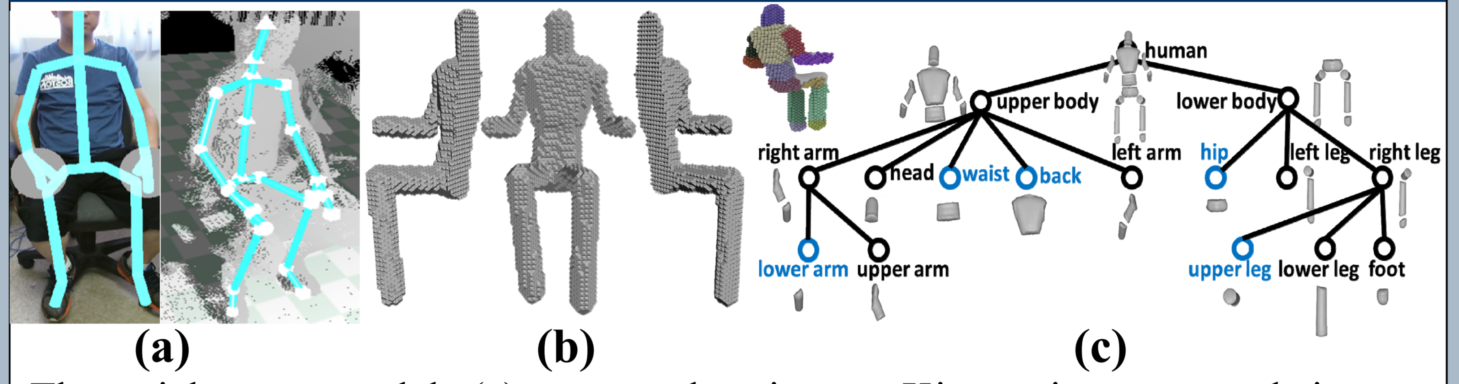


Reconstructing Watertight Scenes

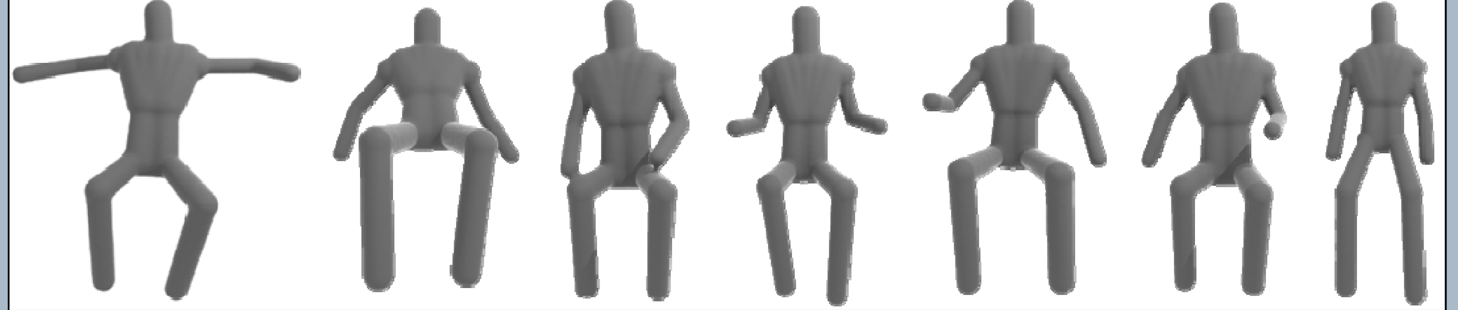


From a reconstructed 3D indoor scene (a), we uniformly sample vertices in the input mesh with Poisson disk sampling (b), then convert them into a watertight mesh (d) with well-defined interior and exterior regions. Differences (c) between the input mesh and the converted watertight mesh.

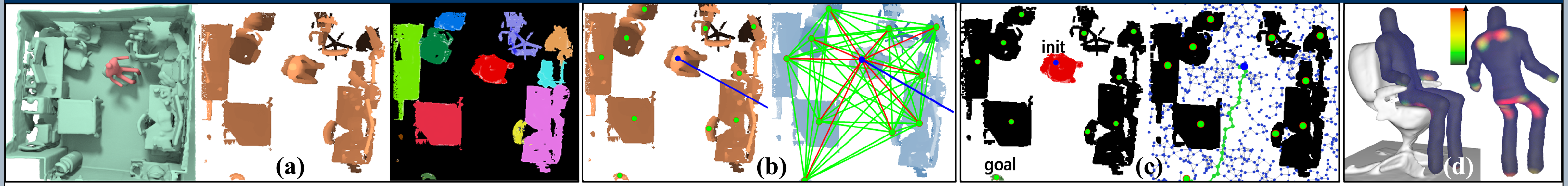
Modeling Volumetric Human Pose



The stick-man model (a) captured using a Kinect is converted into a tetrahedralized human model (b) and then segmented into 14 body parts (c).



Data Pre-processing and Feature Extractions



(a) Data pre-processing. Given a reconstructed 3D scene, we project it down onto a planar map, and segment 3D objects from the scene.
(b) Spatial Entities and Relations in 3D Spaces. Visualization of 3D object positions (green dots), human head position (blue dot), and orientation (blue line). Spatial features $\phi_s(G)$ are defined as human-object (red lines) and object-object (green lines) relative distances and orientations.
(c) Human Utilities in Time. Temporal features $\phi_t(G)$ are defined as the plan cost from a given initial position to a goal position.
(d) Physical Quantities of Human Utilities. Using FEM simulation, the physical quantities $\phi_p(G)$ are estimated at each vertex of the FEM mesh.

Simulating Human Interactions with Scenes

We used **Finite Element Method** to simulate human tissue dynamics.
Input: reconstructed watertight scenes and volumetric human poses.
Outputs: forces and pressures.

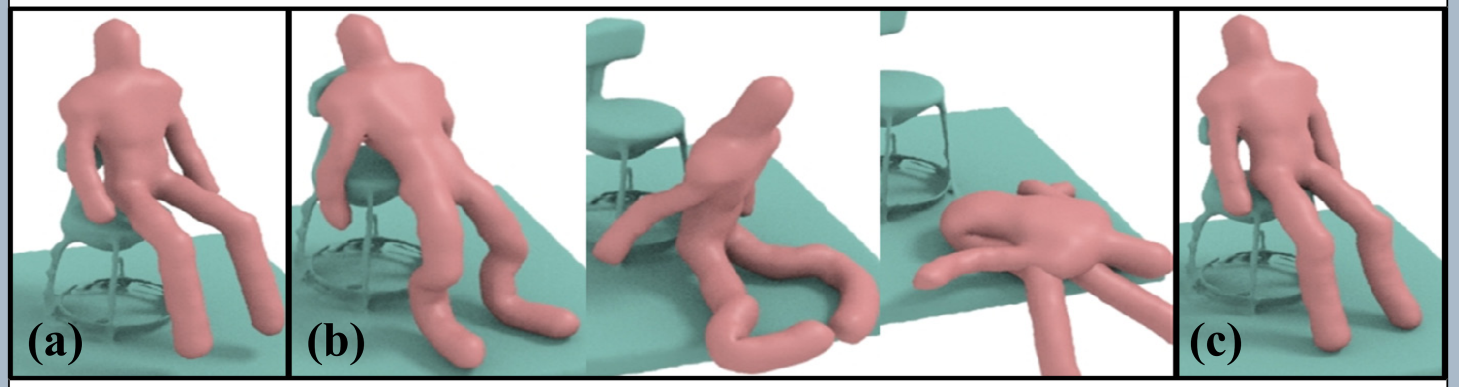
Timestep: $1 \times 10^{-3} s$	Density: $1000 kg/m^3$	Young's modulus: $0.15 kPa$	Poisson's ratio: 0.3
Collision stiffness: $1 \times 10^4 kg/s^2$	Friction coeff: 1×10^{-3}	Damping coeff: $50 kg/s$	Gravity: $9.81 m/s^2$

Elasticity: The human body is modeled as an elastic material. The total elastic potential energy is defined as: $\Phi^E(\mathbf{x}) = \int_\Omega \Psi^E(\mathbf{x}) d\mathbf{x} \approx \sum_e V_e^0 \Psi^E(\mathbf{F}(\mathbf{x}))$.

Contact forces: To model contact forces, we need to penalize penetrations of the human body mesh into the scene mesh. If a penetration is detected for vertex i, a collision energy that penalizes the penetration distance in the normal direction is assigned to the corresponding vertex: $\Phi^C(\mathbf{x}_i) = \frac{1}{2} k_c (\mathbf{x}_i - \mathcal{P}(\mathbf{x}_i))^2$.

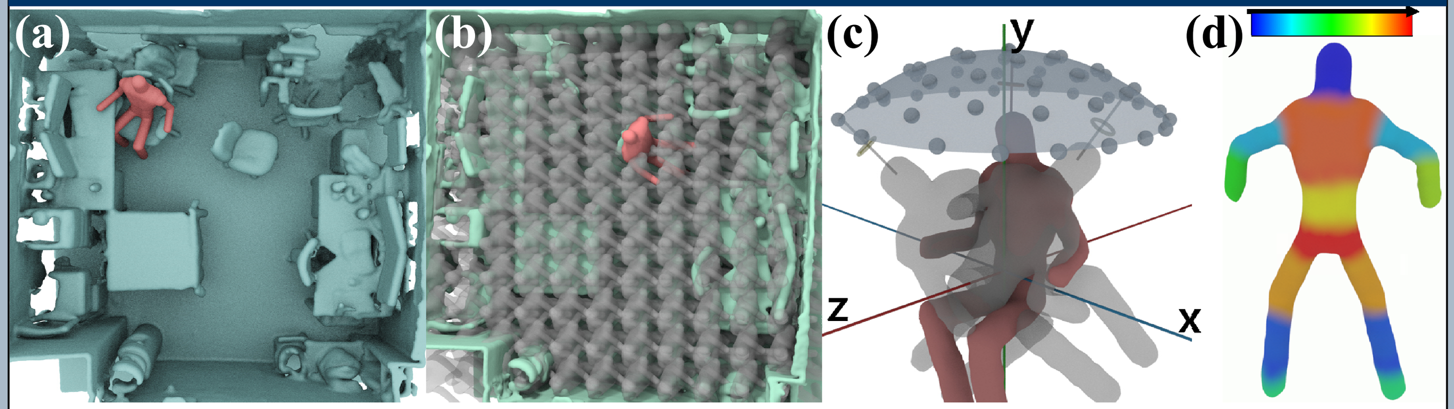
Dynamics integration: Backward Euler time integration is used to solve the momentum equation. From time n to n + 1, the nonlinear system to solve is:

$$\mathrm{M}\frac{\mathbf{v}^{n+1} - \mathbf{v}^n}{\Delta t} = \mathbf{f}(\mathbf{x}^{n+1}, \mathbf{v}^{n+1}) + \mathrm{M}g,$$
$$\mathbf{f}(\mathbf{x}^{n+1}, \mathbf{v}^{n+1}) = \mathbf{f}^E(\mathbf{x}^{n+1}) + \mathbf{f}^C(\mathbf{x}^{n+1}) + \mathbf{f}^D(\mathbf{v}^{n+1}),$$
$$\mathbf{x}^{n+1} - \mathbf{x}^n = \mathbf{v}^{n+1}\Delta t.$$

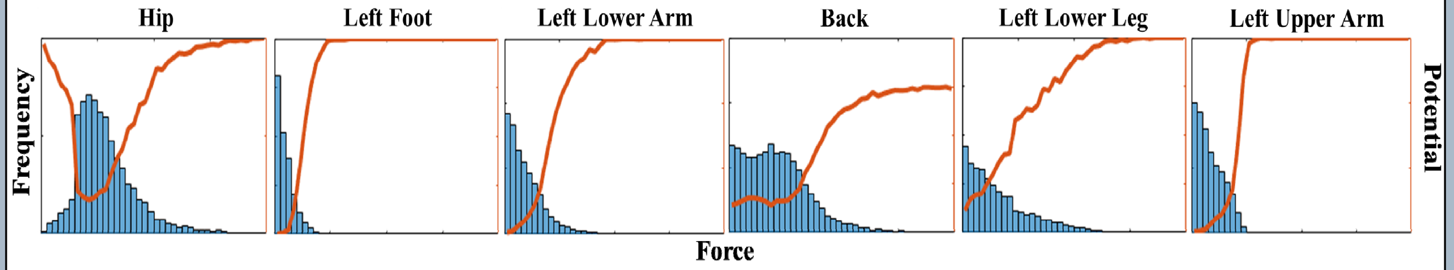


Given an initial human pose in a 3D scene subject to gravity (a), without adequate damping (b) the human body is too energetic and produces unnaturally bouncy motion. With proper damping, the simulation converges to a physically stable rest pose (c) in a small number of timesteps.

Learning Human Utilities



(a) We assume that the observed demonstration is near-optimal, and therefore regard it a positive example. The learning algorithm then imagines different configurations by initializing with different human poses P_a, (b) translations T_b, and (c) orientations O_c. The imagined randomly generated configurations are regarded negative examples. (d) The average forces of each body part remapped to a T pose.



Learning the ranking function is equivalent to finding the coefficient vector such that the maximum number of the inequalities are satisfied:

$$\langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}^\star) \rangle > \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}_i) \rangle, \ \ \forall i \in \{1, 2, \cdots, n\}$$

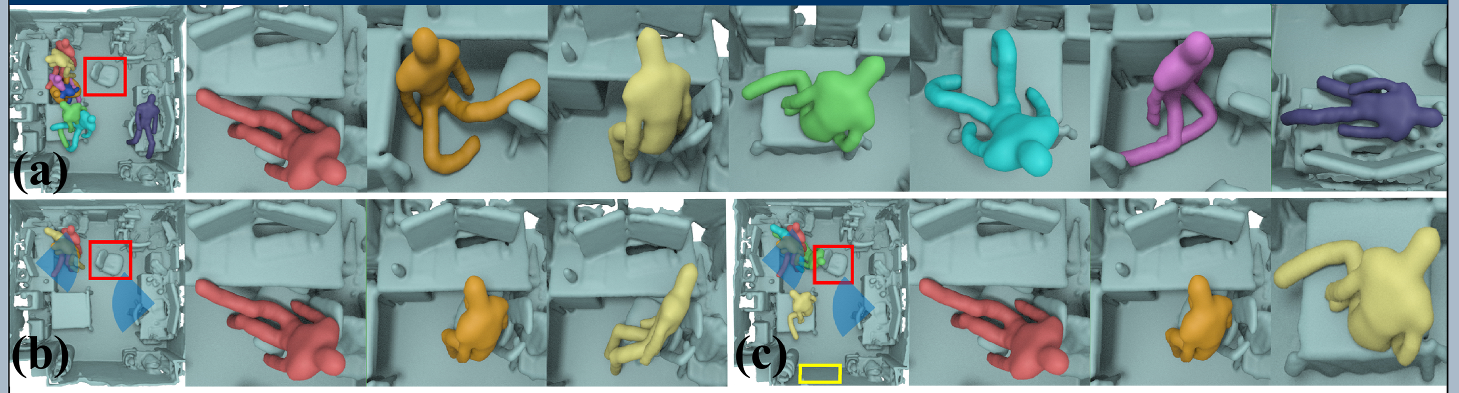
To approximate the solution, we introduce non-negative slack variables

$$\min \frac{1}{2}\langle \boldsymbol{\omega}, \boldsymbol{\omega} \rangle + \lambda \sum_i^n \xi_i^2, \ \ \forall i \in \{1, \cdots, n\}$$
$$\text{s.t. } \xi_i \geq 0, \ \ \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}^\star) \rangle - \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}_i) \rangle > 1 - \xi_i^2$$

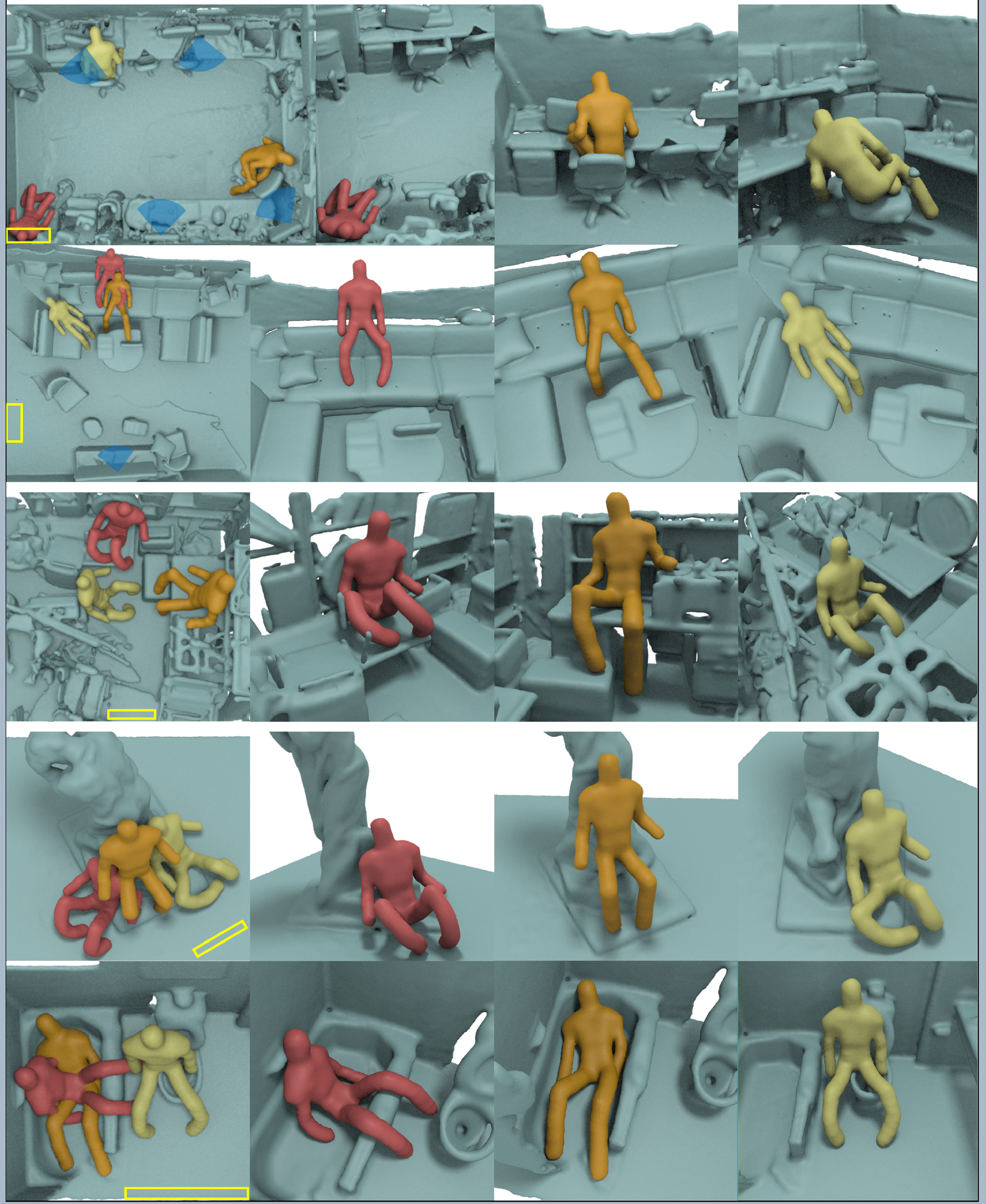
In the **inference** phase, the goal is to find, among all the imagined configurations in the solution space, the best configuration that receives the highest score:

$$\mathcal{G}^\star = \arg\max_{\mathcal{G}_i} \langle \boldsymbol{\omega}, \boldsymbol{\phi}(\mathcal{G}_i) \rangle$$

Experiments



(a) The top 7 human poses using physical quantities. The algorithm seeks physically comfortable sitting poses, resulting in casual sitting styles; e.g., lying on the desk. (b) Improved results after adding spatial features to restrict the human-object relative orientations and distances. Further including temporal features yields the most natural poses (c).



The work reported herein was supported by DARPA SIMPLEX grant N66001-15-C-4035, ONR MURI grant N00014-16-1-2007, and DoD CDMRP grant W81XWH-15-1-0147