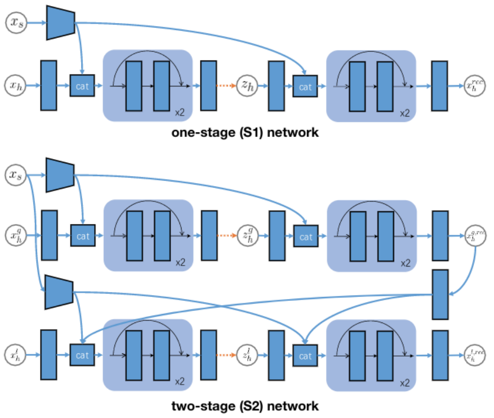
Generating 3D People in Scenes without People

Zhang, Hassan, Neumannm, Black, Tang (CVPR 2020) 4/7/2021 Soma Nonaka



What: Generate human mesh models with natural poses in a given 3D scene.

Why: Past methods generate humans as stick figures, but this method generates meshes.

Generating humans in 3D scenes is useful for applications such as XR, video games, and creating training data for downstream tasks. Realistic human-scene interactions such as sitting, grasping require a detailed human body model to represent the contact between humans and objects. Past methods cannot represent these contacts because they generate humans as stick figures. The authors propose a method to generate human SMPL-X meshes from scene contexts, enabling more detailed modeling of human-scene interaction.

How: Learn a human mesh generator conditioned on scene depth and semantics, then fit the generated human model to scenes so that the human-scene interaction is encouraged but collision and interpenetration are discouraged.

The human mesh generator takes initial body parameters (*e.g.* initial SMPL-X parameters, global positions, and rotations) and scene features of a given 3D scene (*e.g.* depth and semantics) as inputs. The generator outputs the modified SMPL-X parameter that represents natural poses guided by the 3D scene.

As shown in Fig. 2, the authors built CVAE to model the probability of modified body features (x_h^{rec}) conditioned on input body (x_h) and scene (x_s) features. They made two variations of models. The one-stage network (S1) is a simple CVAE that outputs x_h^{rec} from x_h and x_s. The two-stage network (S2) firstly processes the global body features (*i.e.* global rotation and position), and then processes the local features (*i.e.* SMPL-X parameters) given the global features.

They trained the network with five loss functions. 1: Reconstruction loss of body features (Eq. 3). 2: KL-divergence loss that encourages the latent space to become Gaussian (Eq. 4). 3: VPoser loss. VPoser is a VAE trained on a large-scale human pose dataset, and it encodes a natural prior of human poses. VPoser loss is the norm of the generated body pose, therefore it penalizes unnatural body poses (Eq. 5). 4: Collision loss, which is the overlapping area between the body mesh and scene mesh. This loss provides penalties when the body is interpenetrating scene objects (Eq. 6). 5: Contact loss, which is the minimum distance between the body mesh and scene mesh. It encourages contact between humans and objects (Eq. 7).

When training, all of the 5 loss functions are used to train the CVAE. Then, the generated poses are further refined by minimizing VPoser loss, collision loss, and contact loss to encourage human-scene interaction.

Results: Outperforms previous methods in reconstruction error, diversity of generated poses, physical realism, perceptual plausibility rated by crowd workers. The generated poses are useful for training pose estimation models.

The proposed model was trained on PROX-E (3d body pose dataset with 3D scene geometries and semantics), and evaluated on PROX-E and MP3d-R (3d scene dataset without people). The training images were augmented by placing virtual cameras in the scene, and 70K images were created in total.

As a baseline for comparison, the authors used an extension of Li *et al.*'s method. Because Li *et al.*'s method outputs stick figures of people, the authors modified the method to output SMPL-X models. Also, they used the same input data (scene depth + semantics) and applied VPoser prior for a fair comparison.

The authors firstly evaluated reconstruction error between generated and true body poses, and the proposed model outperforms Li *et al.*'s method. Secondly, they evaluated the diversity of generated poses by computing the entropy of the poses, and the proposed method also performs better. They also investigated the number of collisions and contacts between the body and scenes. Results show that the proposed method successfully avoids collision and encourages human-scene contacts. Rating by AMT workers indicates the proposed method and Li *et al.*'s method generate equally plausible 3D human bodies.

The authors also used their method as a prior for the 3D human pose estimation. They generate human bodies with their method, and then fit a SMPL-X model to a person in the input image while using one of the generated poses as the initial value. This method improves accuracy over past 3D pose estimation methods.

Thoughts: Scene depth and semantics are effective to learn human-scene interaction. The proposed method can be further extended.

These scene features as well as CVAE framework may be useful for gaze estimation using scene contexts. The proposed method itself can be extended for human motion generation, motion prediction, or generating multiple human interactions