GOAL: Generating 4D Whole-Body Motion for Hand-Object Grasping

Omid Taheri Vasileios Choutas Michael J. Black Dimitrios Tzionas
Max Planck Institute for Intelligent Systems, Tübingen, Germany
{otaheri, vchoutas, black, dtzionas}@tue.mpg.de

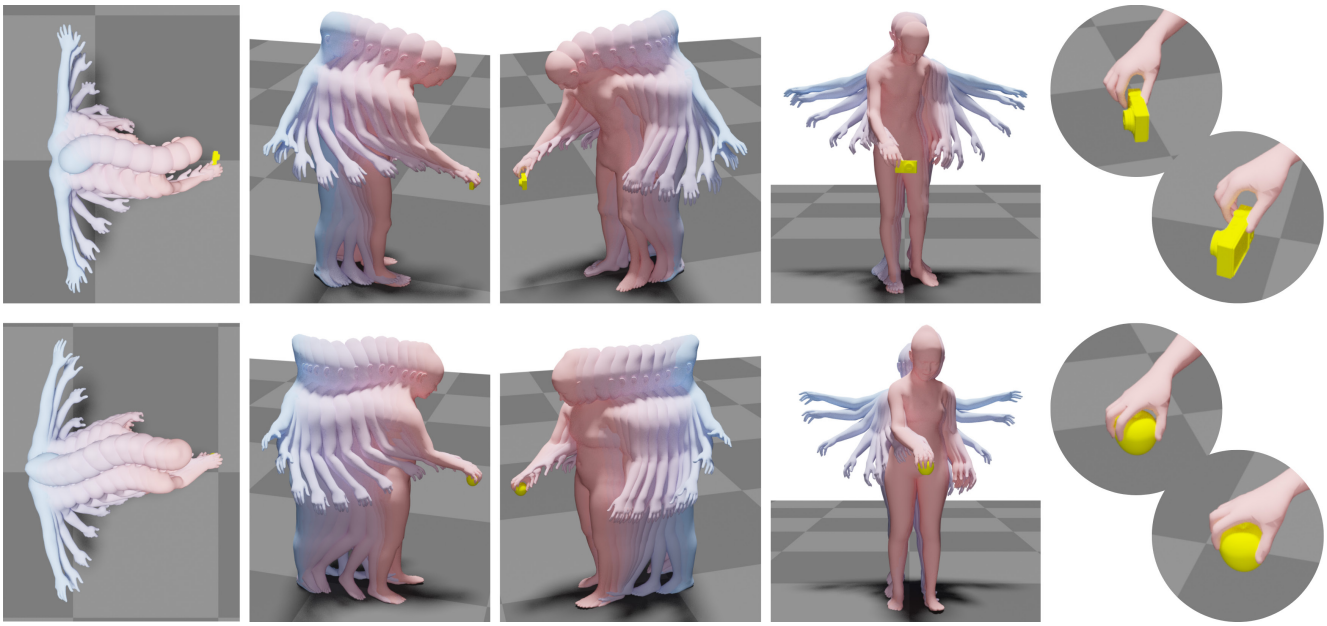

Figure 1. GOAL generates whole-body motions for approaching and grasping an unseen 3D object. The figure shows *generated motions* for 2 people (top, bottom), each grasping a different novel object. For each sequence we show 4 different views (left to right), as well as zoomed-in circular snapshots of the final grasp. GOAL is the first method to generate such a natural motion and grasp for the full body.

Abstract

Generating digital humans that move realistically has many applications and is widely studied, but existing methods focus on the major limbs of the body, ignoring the hands and head. Hands have been separately studied, but the focus has been on generating realistic static grasps of objects. To synthesize virtual characters that interact with the world, we need to generate full-body motions and realistic hand grasps simultaneously. Both sub-problems are challenging on their own and, together, the state space of poses is significantly larger, the scales of hand and body motions differ, and the whole-body posture and the hand grasp must agree, satisfy physical constraints, and be plausible. Additionally, the head is involved because the avatar must look at the object to interact with it. For the first time, we address the problem of generating full-body, hand and head motions of an avatar grasping an unknown object. As input, our method, called GOAL, takes a 3D object, its pose, and a starting 3D body pose and shape. GOAL outputs a sequence of whole-body poses using two novel networks. First, GNet generates a goal whole-body grasp with a realistic body, head, arm, and hand pose, as well as hand-object contact. Second, MNet generates the motion between the starting and goal pose. This is challenging, as it requires the avatar to walk towards the object with foot-ground contact, orient the head towards it, reach out, and grasp it with a realistic hand pose and hand-object contact. To achieve this the networks exploit a representation that combines SMPL-X body parameters and 3D vertex offsets. We train and evaluate GOAL, both qualitatively and quantitatively, on the GRAB dataset. Results show that GOAL generalizes well to unseen objects, outperforming baselines. A perceptual study shows that GOAL's generated motions approach the realism of GRAB's ground truth. GOAL takes a step towards generating realistic full-body object grasping motion. Our models and code are available at https://goal.is.tue.mpg.de.

1. Introduction

Virtual humans are important for movies, games, AR/VR and the metaverse. Not only do they need to look realistic, but also must move and *interact* realistically. Most work on human motion generation has focused only on bodies, without the head and hands. Often, these bodies are considered in "isolation", with no scene or object context. Other work focuses on bodies interacting with scenes, but ignores the hands. Similarly, work on generating hand grasps often ignores the body. We argue that these are all just parts of the problem. What we really need, instead, is to generate the motion of *whole-body* avatars *grasping* objects, by jointly considering the body, head, feet, hands, as well as objects. We address this here for the first time.

The problem is challenging and multifaceted. Think of how we grasp objects in real life (see Fig. 2); we walk towards the object with our feet contacting the ground, we orient our head to look at the object, lean our torso and extend our arms to reach it, and dexterously pose our hands to establish fine contact and grasp it. Humans are able to gracefully execute these steps, yet, these are challenging and involve motion planning, motor control, and spatial awareness. Some of these steps have been studied separately, but we cannot simply combine the partial solutions since the entire action must be *coordinated*. This is challenging because: (1) full bodies have a much higher-dimensional state space than bodies or hands alone; (2) the body and hands have very different sizes, motion scales and level of dexterity; (3) the body, head and hands must move in a coordinated fashion. Currently, there are no automatic tools to generate such coordinated full-body grasping motions.

We address this with *GOAL*, which stands for *Generating Object-interActing whoLe-body motions*. GOAL generates whole-body avatar motion for grasping an unknown object, by jointly considering the body, head, feet, hands and the object. GOAL takes three *inputs*: (1) a 3D object, (2) its position and orientation, and (3) a "starting" 3D body pose and shape, positioned near the object and roughly oriented towards it. As output, GOAL generates a sequence of 3D body poses from the starting pose through to an object grasp. To do so, GOAL uses two novel networks (for an overview see Fig. 3): (1) First, GNet generates a "goal" whole-body grasp, with a realistic body pose, head pose, arm pose, and hand pose, as well as realistic finger-object and foot-ground contact. GNet is formulated as a conditional variational auto-encoder (cVAE), thus, it learns a distribution over grasping poses, and can generate a variety of "goal" grasps. (2) Then, MNet inpaints the motion between the "starting" and "goal" poses, by generating a sequence of whole-body poses in an auto-regressive fashion. This is challenging because the avatar needs to (see Fig. 1) walk by taking a number of steps proportional to the distance to the object, while having natural foot-ground contact without "skating", and continuously orient the head to look at the object. Then, when it is near the object, it needs to slow down, stop walking, lean the torso, extend the arms to reach the object. It must also pose the hand to contact the object and grasp it. All body parts need to move gracefully and in full coordination, so that the motion looks natural.

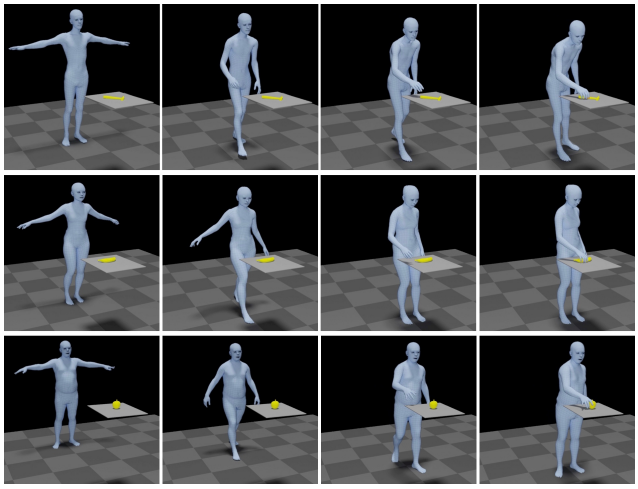

Figure 2. Grasping an object involves several motions. We walk towards the object with our feet contacting the floor, we orient our head to look at the object, we lean our torso, extend our arms, and pose our hand to contact and grasp the object. The depicted examples use motions captured in the GRAB dataset [57].

Achieving this level of realism requires technical novelties. GOAL goes beyond recent work [38, 65, 67] to *jointly* infer both SMPL-X [46] parameters and 3D offsets. GNet infers 3D hand-to-object vertex offsets to give spatial awareness and guide object grasping. MNet infers 3D SMPL-X vertex offsets to guide SMPL-X deformation from the previous to the current frame. These offsets lie in 3D Euclidean space, thus, they can be more accurately inferred than SMPL-X parameters, and are used in an offline optimization scheme to refine SMPL-X poses. We train GNet and MNet on the GRAB [57] dataset, which contains whole-body SMPL-X humans grasping objects.

We evaluate GOAL, both quantitatively and qualitatively, on withheld parts of the GRAB dataset. Specifically, we withhold 5 objects for testing. Results show that GOAL generalizes well and produces natural motions for full-body walking and object grasping; see Fig. 1. Quantitative evaluation shows that GOAL outperforms baselines, and ablation studies show a positive contribution of all major components. A perceptual study verifies the above, while showing that GOAL's generated motions achieve a level of realism comparable to GRAB's ground-truth motions.

GOAL takes a step towards generating whole-body grasp motion for realistic avatars. Models and code are available at https://goal.is.tue.mpg.de.

2. Related Work

Motion generation for bodies "in isolation": Research on human motion generation has a long history [2, 4, 61]. However, even recent methods [41, 49, 64, 67], mostly study the body "in isolation"; i.e., with no scene context. Most methods generate the motion of 3D skeletons [15, 24, 41–43, 64], while others [18, 49, 67] generate the motion of a human model like SMPL [39]. Typically, 1-2 seconds of motion synthesis is referred to as "long term". Early deep-learning methods employ RNNs [12, 15, 44], but they struggle with discontinuities between the observed and predicted poses, as well as with long-range spatio-temporal relations. Other methods tackle these with phase-functioned feed-forward neural networks [23, 56], i.e., by conditioning network weights on phase, but they focus on cyclic motions. More recent methods [36, 41, 49, 58] use attention [59].

Motion generation for bodies in 3D scenes: Most early methods extend MoCap databases with point annotations for foot and hand contact [14, 27, 33, 34]. Then, they fit motion to contacts with optimization and space-time constraints for 3D body motion re-targeting [14], and animating bodies that move over 3D terrain [27, 33, 34].

Some methods use deep reinforcement learning (RL) for body-scene [6, 47, 48] or hand-object [7, 8, 13] interactions. These methods show promising results for navigating terrain with varying height and gaps [47, 48], sitting on chairs [6, 56], hammering [13], opening a door [13], moving objects [8], and for in-hand object re-orientation [7]. Generalization to new bodies, object geometry, and interaction types remains a challenge.

Others follow a 3D geometric approach. Pirk et al. [50] place virtual sensors on objects to sense the flow of points sampled on an agent interacting with these, and build functional object descriptors. Al-Asqhar et al. [1] re-target body motion by encoding human joints w.r.t. fixed points sampled on a scene. Ho et al. [22] use body and object vertices to compute per-frame "interaction meshes", and minimize their Laplacian deformation to re-target body motion. These pure geometric methods are not robust to real-world noise.

In contrast, GOAL is in the category of data-driven methods. Corona et al. [9] generate the context-aware motion of a human skeleton interacting with objects, where "context" is encoded as a directed graph connecting person and object nodes. More relevant are methods for generating motion between a "start" and a "goal" pose in a 3D scene. Hassan et al. [19] estimate a "goal" position and interaction direction on an object, plan a 3D path from a "start" body pose to this, and finally generate a sequence of body poses with an auto-regressive cVAE for walking and interacting, e.g., sitting on a chair. Wang et al. [60] first estimate several "sub-goal" positions and bodies, divide these into short start/end pairs to synthesize short-term motions, and finally stitch these together in a long motion with an optimization process.

Motion generation for hands: ElKoura1 et al. [11] estimate physically plausible hand poses for music instruments, using a learned low-dimensional pose space. Pollard et al. [51] use MoCap to learn a controller for physically-based grasping. Kry et al. [32] capture hand MoCap and forces with instrumented objects, and build "interaction trajectories" for synthesizing or re-targeting motions with physics simulation. More related to us, Lie et al. [62] take as input MoCap data of body and object motion, and add the missing hand motion by searching for feasible contact-point trajectories and then generating smooth hand motion with space-time optimization that satisfies the estimated contacts.

Pose generation for bodies in 3D scenes: Early methods use contact annotations [37] or detections [29] on 3D objects, and fit human skeletons to these. Other methods use physics simulation to reason about contacts and sitting comfort [26, 35, 69]. Focusing on rooms, Grabner et al. [16] predict all areas on a 3D scene mesh where a 3D human mesh can sit, using proximity and intersection metrics. Recent methods generate static SMPL-X [46] humans interacting with a given scene; Zhang et al. [68] learn an implicit interaction representation given a depth image and semantic segmentation of the scene, Zhang et al. [66] use an explicit scene-centric representation of interaction, while Hassan et al. [20] use a human-centric one, called POSA, that they embed in the SMPL-X [46] statistical body model. In contrast, Yi et al. [63] reconstruct and refine object poses to better "support" a given human motion, using POSA-based contact, collision and relative-depth constraints.

Pose generation for hand-object grasps: Taheri et al. [57] infer MANO [55] grasps for a 3D object, by first predicting a rough grasp, and then refining it with distance and contact metrics. Grady et al. [17] first estimate contacts on both the hand and object, and then refine hand poses with optimization to satisfy contacts. Karunratanakul et al. [28] infer a "grasping" distance field and then fit MANO to it.

Motion for full-body interactions: People use their body and hands together for interacting with the world. Hsiao et al. [25] build a database of whole-body grasps with a human operating an avatar, and perform imitation learning. Borras et al. [3] capture whole-body MoCap data [40] of people interacting with scene objects and handheld objects using a humanoid model, and define a pose taxonomy. Taheri et al. [57] capture whole-body SMPL-X [46] interactions with handheld objects, but learn a cVAE that generates only static grasping hands, due to the task complexity. Merel et al. [45] use deep RL and human MoCap demonstrations to learn a vision-guided neural controller for picking up and carrying boxes, or catching and throwing a ball.

Summary: The community has focused only on bodies or hands, using unrealistic models. We learn to generate full-body SMPL-X motion for walking towards an object up to grasping it, given a "start" object and human pose.

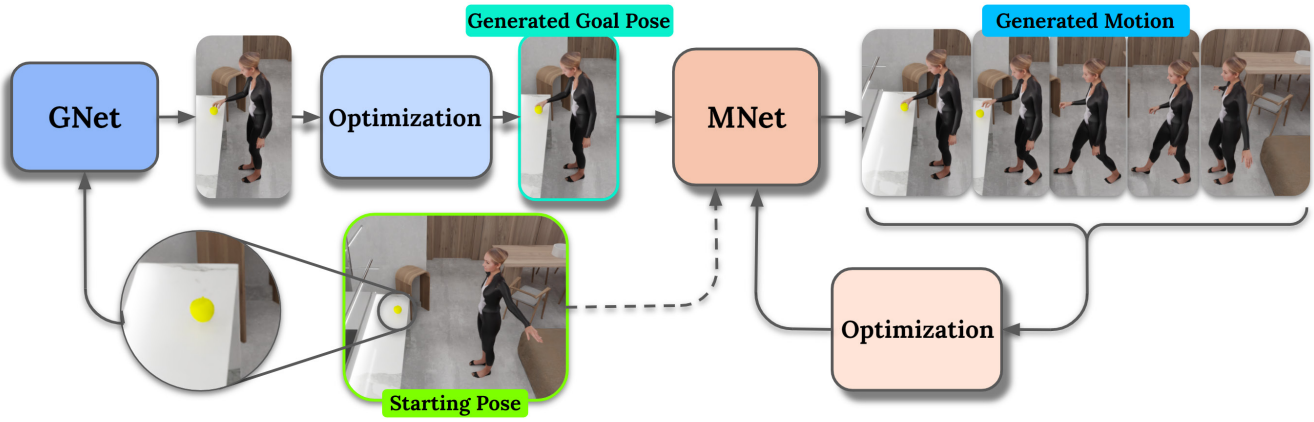

Figure 3. Overview of GOAL. There are two main stages: (1) GNet takes as input a 3D object, its position and orientation, and generates a "goal" whole-body grasping pose. GNet's output pose is refined with optimization post processing to look more realistic and physically plausible. (2) MNet takes as input a starting human pose and GNet's "goal" pose, and generates the motion in between as a sequence of poses in an auto-regressive fashion. MNet's output poses are refined with optimization post processing to better "reach" the "goal" pose.

3. Method

An overview of our method, GOAL, is shown in Fig. 3. GOAL takes three *inputs*: (1) a 3D object, (2) its position and orientation, and (3) a "starting" 3D body pose and shape, positioned near the object (roughly $0.5 - 1.5$ m) and oriented towards it (roughly $\pm 10°$). As *output* GOAL generates SMPL-X motion with two main networks: (1) GNet synthesizes a "goal" SMPL-X mesh that grasps the 3D object with a realistic body pose and hand-object contact; (2) MNet "inpaints" the motion from the "start" to the "goal" pose, by generating a sequence of "moving" SMPL-X bodies in an auto-regressive way. Without loss of generality, we model right-handed grasps; which can be transferred to the left hand easily through "mirroring" data and retraining. Extending these to two-handed grasps, with or without out hand coordination, is left for future work.

3.1. Human Model

We use the SMPL-X [46] statistical 3D whole-body model, which jointly represents the body, head, face and hands. SMPL-X is a differentiable function that takes as input shape, $\boldsymbol{\beta}$, pose, $\boldsymbol{\theta}$, and expression, $\boldsymbol{\psi}$, parameters and then outputs a 3D mesh, M, with $10,475$ vertices, V, and $20,908$ triangles, F. The shape vector $\boldsymbol{\beta} \in \mathbb{R}^{20}$ contains coefficients of a low-dimensional space, created via PCA on 3D meshes of roughly $4,000$ different people [54]. The vertices are posed with linear blend skinning with a learned rigged skeleton with joints $\mathcal{J} \in \mathbb{R}^{55\times 3}$. Let $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{\gamma}\}$ represent the articulated pose $\boldsymbol{\theta} \in \mathbb{R}^{55\times 6}$ [70] and translation $\boldsymbol{\gamma} \in \mathbb{R}^3$ of the body. In the following, instead of using all SMPL-X vertices, we sample N vertices on body areas that are important for interactions using GRAB's [57] contact heatmaps.

3.2. Interaction-Aware Attention

Two common representations for body-object interaction are: vertex-to-vertex distances and binary contact labels for mesh vertices, However, the former carries information that is irrelevant to the interaction (e.g., vertices far away from the object), while the latter is too compact and carries no information about 3D proximity before/after contact.

Here, we use vertex-to-vertex distances, but introduce a new "interaction-aware attention" (IAA) that focuses more on body vertices that are important for interaction (e.g., hands for grasping, feet for walking) and less on irrelevant vertices (e.g., knees are less relevant than hands for grasping). Our "interaction-aware" attention is formulated as:

$$I_w(\boldsymbol{d}) = e^{-w\boldsymbol{d}}, \tag{1}$$

where $w > 0$ is a scalar weight, $\boldsymbol{d} \in \mathbb{R}^N_+$ is a body-to-object distance, and N is the number of sampled vertices on SMPL-X; we sample $N_b = 400$ for the body and $N_h = 99$ for each hand. Our IAA gives exponentially more attention to vertices relevant for interaction. As visualized in Fig. 4, this focuses attention on body areas that are meaningful for interaction. We set $w = 5$, which empirically results in realistic grasps.

3.3. Goal Network (GNet)

GNet is a conditional variational auto-encoder (cVAE) [31] that generates a static whole-body grasp, conditioned on the given object and its pose. To do this, we first encode whole-body grasps into an embedding space.

Input: GNet's encoder takes as input:

$$\boldsymbol{X}^{\text{in}}_{\text{GNet}} = \left[\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{q}, \boldsymbol{\gamma}^{\circ}, \boldsymbol{b}^{\circ}, \boldsymbol{d}^{b\mapsto\circ}\right], \tag{2}$$

where $\boldsymbol{\Theta}$ and β are SMPL-X's pose and shape parameters, respectively, $\boldsymbol{v} \in \mathbb{R}^{N_b\times 3}$ are the 3D coordinates of the N_b

sampled SMPL-X vertices, $\boldsymbol{q} \in \mathbb{R}^3$ is a unit vector for head orientation, $\boldsymbol{\gamma}^\circ \in \mathbb{R}^3$ is the object translation, and $\boldsymbol{b}^\circ \in \mathbb{R}^{1024}$ is the Basis Point Set (BPS) [52] representation of the 3D object shape. Finally, $\boldsymbol{d}^{b\to\circ} \in \mathbb{R}^{N_b \times 3}$ denotes 3D offset vectors that encode the body-to-object proximity; for each of the N_b sampled body vertices, $\boldsymbol{v}$, it contains a 3D offset vector to the closest object vertex in $\boldsymbol{v}^\circ$.

At training time, GNet's encoder maps the inputs X to the parameters of a normal distribution, $\{\boldsymbol{\mu}, \boldsymbol{\sigma}\} \in \mathbb{R}^{16}$. At inference time, we "skip" the encoder, and sample a latent whole-body grasp code, $\boldsymbol{z}_g \in \mathbb{R}^{16}$, from this distribution.

Output: GNet's decoder takes the grasp code, $\boldsymbol{z}_g$, and the input conditions for the object, $\boldsymbol{C} = [\boldsymbol{b}^\circ, \boldsymbol{\gamma}^\circ]$, and infers SMPL-X pose parameters $\hat{\boldsymbol{\Theta}}$, the head direction vector $\hat{\boldsymbol{q}}$, and 3D offset vectors $\hat{\boldsymbol{d}}^{h\to\circ}$ from the N_h sampled hand vertices, $\boldsymbol{v}_h \subset \boldsymbol{v}$, to the closest object vertex.

Output space: We make two empirical observations: (1) Networks struggle to predict accurate SMPL-X parameters, possibly due to their non-Euclidean space. (2) Networks predict interaction features in a Euclidean space much more precisely. These observations are in line with recent work [38, 65, 67]. However, we go beyond prior work by inferring 3D *offsets* together with SMPL-X parameters, instead of regressing vertex positions and fitting SMPL-X to these.

Training: GNet is trained with the loss:

$$\begin{aligned}\mathcal{L}_{\text{GNet}} = \; & \lambda_{\boldsymbol{v}}\mathcal{L}_{\boldsymbol{v}} + \lambda^h_{\boldsymbol{v}}\mathcal{L}^h_{\boldsymbol{v}} + \lambda_{\boldsymbol{\Theta}}\mathcal{L}_{\boldsymbol{\Theta}} + \\ & \lambda_{\text{q}}\mathcal{L}_{\text{q}} + \lambda^{h\to\circ}_d\mathcal{L}^{h\to\circ}_d + \lambda_{KL}\mathcal{L}_{KL},\end{aligned} \tag{3}$$

where $\mathcal{L}_{\boldsymbol{v}} = \|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_1$, $\mathcal{L}^h_{\boldsymbol{v}} = \|\boldsymbol{v}^h - \hat{\boldsymbol{v}}^h\|_1$, $\mathcal{L}_{\boldsymbol{\Theta}} = \|\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}\|_2$, $\mathcal{L}_{\text{q}} = \|\boldsymbol{q} - \hat{\boldsymbol{q}}\|_2$, $\mathcal{L}^{h\to\circ}_d = \|\boldsymbol{d}^{h\to\circ} - \hat{\boldsymbol{d}}^{h\to\circ}\|_1$, $\mathcal{L}_{KL}$ is the Kullback-Leibler divergence, and λ are weights. Hat variables are inferred; non-hat ones are ground truth. GNet's encoder and decoder use fully-connected layers with skip connections. For architecture details, see Sup. Mat.

Optimization: We use the predicted offsets to refine our SMPL-X predictions with optimization post processing. Specifically, we *optimize* over SMPL-X parameters, $\boldsymbol{\Theta}$, initialized with GNet's predictions. Instead of hand-crafted contact constraints [10, 21, 60] during optimization, we use data-driven constraints *generated* from GNet, namely: (1) hand-to-object vertex offsets, (2) the head-orientation vector, (3) pose coupling to the initial value, and (4) foot-ground penetration. In technical terms, to refine the hand to realistically grasp the object, we define a term that penalizes differences between GNet's inferred offsets $\hat{\boldsymbol{d}}^{h\to\circ}$, and offsets $\boldsymbol{d}^{h\to\circ}$, computed online during the optimization, from SMPL-X's hand vertices to the closest object vertices:

$$\boldsymbol{E}^{h\to\circ}_{\boldsymbol{d}}(\boldsymbol{\theta}, \boldsymbol{\gamma}; \hat{\boldsymbol{d}}^{h\to\circ}) = \|\boldsymbol{d}^{h\to\circ} - \hat{\boldsymbol{d}}^{h\to\circ}\|_1. \tag{4}$$

Coupling for pose, $\boldsymbol{\theta}$, and translation, $\boldsymbol{\gamma}$, parameters discourages deviation from GNet's inferred values, $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$:

$$\boldsymbol{E}_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}) = \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2, \quad \boldsymbol{E}_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}; \hat{\boldsymbol{\gamma}}) = \|\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\|_1. \tag{5}$$

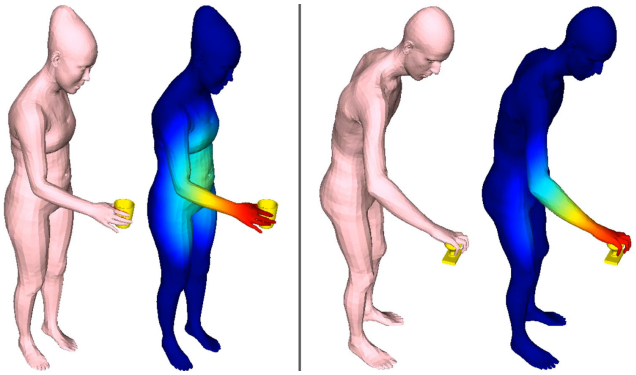

Figure 4. Visualization of the "interaction-aware attention" (IAA) for the body-to-object vertex distances, $I_w(\boldsymbol{d})$, of Sec. 3.2. For each pair: **(Left)** Input 3D meshes for the human (pink) and the object (yellow). **(Right)** The color-coded body mesh visualizes the interaction-aware attention; blue denotes body vertices that are far from the object (i.e., irrelevant for the specific interaction), and red denotes vertices that are near the object (i.e., very relevant).

Similarly, head-orientation coupling is formulated as:

$$\boldsymbol{E}_{\text{q}}(\boldsymbol{\theta}, \boldsymbol{\gamma}; \hat{\boldsymbol{q}}) = \|\boldsymbol{q}(\boldsymbol{\theta}, \boldsymbol{\gamma}) - \hat{\boldsymbol{q}}\|_1. \tag{6}$$

To encourage ground contact and discourage penetration, given the current SMPL-X parameters in every optimization step, we find, online, the lowest vertex of the body along the "y" vertical axis and encourage its y-coordinate to be zero:

$$\boldsymbol{E}_{\text{f}} = |\boldsymbol{v}_y(k)|, \quad k = \arg\min_i \boldsymbol{v}_y(i), \tag{7}$$

where i and k are vertex indices in the body mesh. Our final energy is a combination of the above terms:

$$\boldsymbol{E}_{\text{GNet}} = \lambda^{h\to\circ}_d \boldsymbol{E}^{h\to\circ}_d + \lambda_{\boldsymbol{\theta}}\boldsymbol{E}_{\boldsymbol{\theta}} + \lambda_{\boldsymbol{\gamma}}\boldsymbol{E}_{\boldsymbol{\gamma}} + \lambda_{\text{q}}\boldsymbol{E}_{\text{q}} + \lambda_f \boldsymbol{E}_f. \tag{8}$$

3.4. Motion Network (MNet)

MNet generates the motion from the "start" to the "goal" frame; the latter is generated by GNet, described above in Sec. 3.3. The length of a sequence depends on several factors, such as the object location w.r.t. the body and motion speed. Therefore, to generate motion of arbitrary duration, we use an auto-regressive network architecture [19, 56].

Input: MNet takes as input (auto-regressive fashion):

$$\boldsymbol{X}^{in}_{\text{MNet}} = [\boldsymbol{\Theta}_{t\text{-}5:t}, \boldsymbol{\beta}, \boldsymbol{v}_t, \dot{\boldsymbol{v}}_t, d^h_{t\to g}, \boldsymbol{b}^h_g], \tag{9}$$

where t is the current frame, $\boldsymbol{\Theta}_{t\text{-}5:t}$ are SMPL-X parameters of the last 5 frames, $\boldsymbol{\beta}$ is the subject's shape, $\boldsymbol{v}_t$ and $\dot{\boldsymbol{v}}_t$ are the locations and velocities of the N_b sampled body vertices in the current frame, and $\boldsymbol{d}^h_{t\to g}$ are the hand vertex offsets from the current frame, t, to the "goal" frame, g. Finally, $\boldsymbol{b}^h_g$ is the BPS representation [52] of the hand in the "goal"

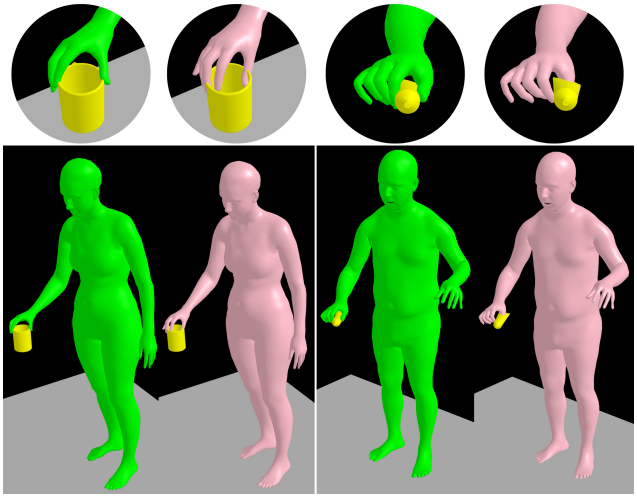

Figure 5. GNet's generated SMPL-X grasp poses (Sec. 3.3) before (pink) and after optimization (green). Results show that optimization-based post-processing effectively refines the initial prediction towards a more realistic and physically plausible grasp.

grasping frame using the same BPS basis points as for the object; using the same basis points for the BPS representation encodes the spatial relationship between the hand and the object in the "goal" frame, and is empirically important for "guiding" the motion towards a realistic grasp.

We empirically find that using as input the pose, $\boldsymbol{\Theta}$, of more than 1 past frame leads to a smoother motion prediction, in agreement with Starke et al. [56]; using more than 5 frames does not lead to noticeable improvement.

Output: MNet produces as output:

$$\boldsymbol{X}^{\text{out}}_{\text{MNet}} = [\Delta\boldsymbol{\Theta}_{t:t+10}, \Delta\boldsymbol{v}_{t:t+10}, \Delta\boldsymbol{d}^h_{t:t+10}] \tag{10}$$

where $t:t$+10 is the future 10 frames, $\Delta\boldsymbol{\Theta}_{t:t+10}$ is the change of SMPL-X parameters, $\Delta\boldsymbol{v}_{t:t+10}$ is the change of SMPL-X vertex positions, and $\Delta\boldsymbol{d}^h_{t:t+10}$ is the change of hand vertex offsets. All changes, Δ, are relative to the current frame.

Output space: MNet focuses both on SMPL-X parameters and Euclidean-space interaction features, similarly to GNet. This empirically helps inference; the generated motion is smoother and better "reaches" the "goal" grasp.

Auto-regression: MNet estimates SMPL-X parameters for future poses, and then these are fed back to MNet as inputs for the next iteration, along with other inputs in Eq. (9). For architecture details, see Sup. Mat. Unlike HuMoR [53] where in each iteration only 1 future frame is generated, MNet's generated motion improves when generating more number of future frames; note that the improvement saturates for 10 future frames, see Tab. 2-right in Sec. 4.2.

Training: MNet is trained with the loss: $\mathcal{L}_{\text{MNet}} =$

$$\lambda_{\boldsymbol{v}}\mathcal{L}_{\boldsymbol{v}} + \lambda^h_{\boldsymbol{v}}\mathcal{L}^h_{\boldsymbol{v}} + \lambda_{\boldsymbol{\Theta}}\mathcal{L}_{\boldsymbol{\Theta}} + \lambda^{h\to\circ}_{\boldsymbol{d}}\mathcal{L}^{h\to\circ}_{\boldsymbol{d}} + \lambda^{\mathbf{f}}_{\boldsymbol{v}}\mathcal{L}^{\mathbf{f}}_{\boldsymbol{v}}, \tag{11}$$

where the losses on body vertices, $\mathcal{L}_{\boldsymbol{v}}$, hand vertices, $\mathcal{L}^h_{\boldsymbol{v}}$, SMPL-X parameters, $\mathcal{L}_{\boldsymbol{\Theta}}$, and hand-to-object offsets, $\mathcal{L}^{h\to\circ}_{\boldsymbol{d}}$ are borrowed from Eq. (3). Finally, $\mathcal{L}^{\mathbf{f}}_{\boldsymbol{v}} = \|\boldsymbol{v}^{\mathbf{f}} - \hat{\boldsymbol{v}}^{\mathbf{f}}\|_1$ is a new loss on foot vertices that are close to the ground. This loss and the input velocities of Eq. (9) help foot-ground contact and reduce sliding; see the video on our website.

Optimization: We refine MNet's generated motion with a post-processing optimization such that the final hand grasp gets closer to the "goal" grasp generated by GNet. Since we need precision only when the hand is close to the object, we conduct optimization only when MNet's estimated hand vertices get closer than 10 cm to the "goal" hand vertex positions. Following GNet's scheme, we use MNet's predictions from Eq. (10) as constraints, instead of hand-crafted ones. Specifically, we first compute the average per-vertex velocity of MNet's predicted hands, $\dot{\boldsymbol{v}}^h_t$. Then, we linearly interpolate hand vertices for the next frame, v^h_{t+1}, between the current, v^h_t, and "goal" ones, v^h_g:

$$v^h_{t+1} = v^h_t + \|\dot{v}^h_t\| l, \quad \text{where} \quad l = \frac{v^h_g - v^h_t}{\|v^h_g - v^h_t\|} \tag{12}$$

where $\|\dot{v}^h_t\|$ is the average-velocity magnitude, and l is a unit vector pointing from current to "goal" hand vertices. In practice, we "force" hands to move towards the "goal" grasp in a locally linear trajectory; this is simple and intuitive, but might result in some hand-object penetrations. Since our focus here is the hand grasp, for the rest of the body we keep MNet's predicted pose and velocity intact.

The energy function for optimization is:

$$\boldsymbol{E}_{\text{MNet}} = \lambda_{\boldsymbol{\Theta}}\boldsymbol{E}_{\boldsymbol{\Theta}} + \lambda^h_{\boldsymbol{v}}\boldsymbol{E}^h_{\boldsymbol{v}}, \tag{13}$$

where the term $\boldsymbol{E}_{\boldsymbol{\Theta}}$ is on SMPL-X parameters, and $\boldsymbol{E}^h_{\boldsymbol{v}}$ is on hand vertices; their definition is borrowed from Eq. (3).

3.5. Implementation Details

Optimization: We refine GNet's and MNet's inferred SMPL-X bodies with gradient descent using Adam [30].

Data: We use GRAB [57], a dataset of whole-body 3D SMPL-X humans grasping objects; it has a separate training and testing set. For data preparation, see Sup. Mat.

4. Experiments

4.1. Qualitative Evaluation

GNet: Figure 5 shows representative generated static grasps before and after optimization. Before optimization, body and head poses are plausible, but hand grasps can be improved (pink). The optimization refines hands for more realistic and physically plausible grasps (green). Figure 7 shows GNet 's generalization ability with grasps generated for 2 unseen and complex objects from the YCB dataset [5].

MNet: Figure 6 shows motions generated for several object shapes and locations, body shapes and "start" poses.

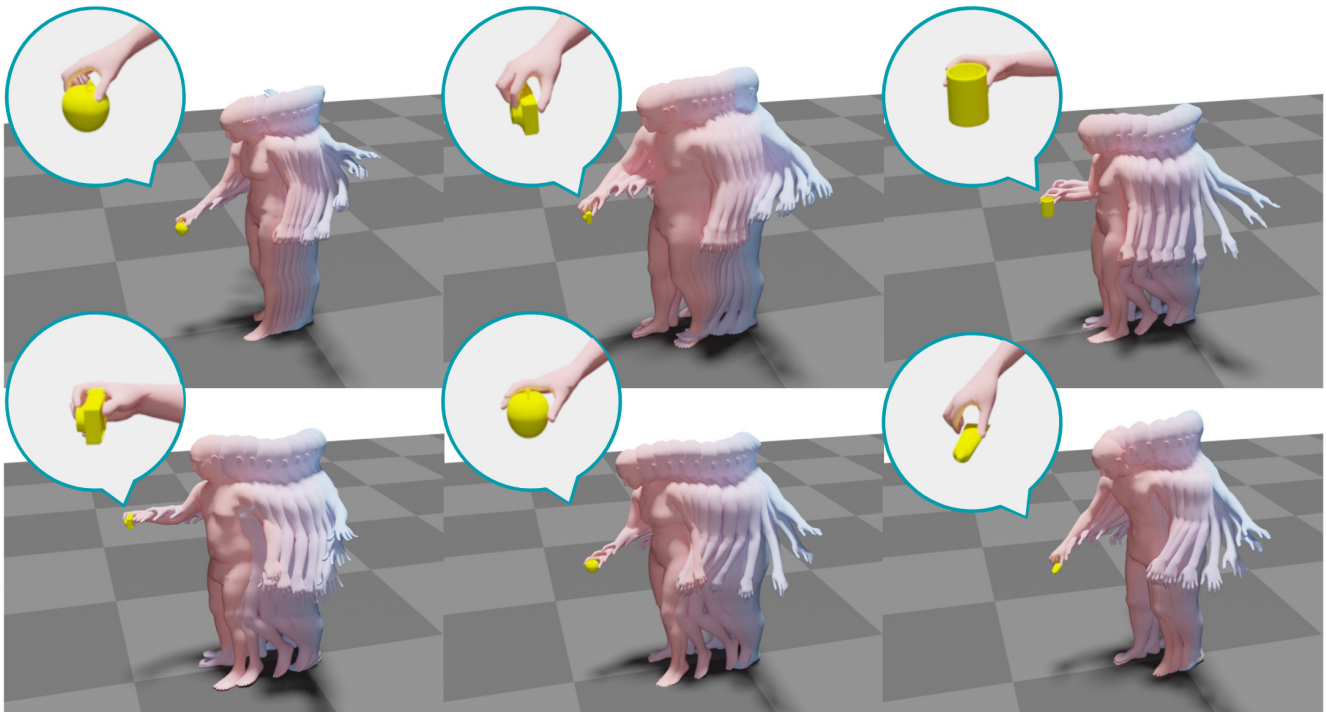

Figure 6. Representative motions generated by GOAL (GNet and MNet) for several object shapes, locations, body shapes and "start" poses.

Grasp Synthesis	Penetration Vol. (cm^3) ↓	Contact Ratio [68]
1. GrabNet [57]	2.65	**1.00**
2. –"– SMPL-X	7.33	0.87
3. GNet-w/o-opt	5.32	0.87
4. GNet (ours)	**2.22**	**1.00**
GRAB (GT)	1.95	1.00

Table 1. Penetration and contact-ratio evaluation for GNet. We compare GNet with-/out optimization and GrabNet variants.

Motion Network	V2V (mm) ↓ Body	Hand	Feet
1. MNet-Pose	22.0	30.2	10.4
2. MNet-Marker [67]	21.1	30.1	**9.8**
3. MNet-w/o-IAA	21.0	29.1	10.5
4. MNet (ours)	**19.7**	**28.0**	9.9

# Output Frames	V2V (mm) ↓ Body	Hand	Feet
1	26.7	35.7	17.9
2	21.5	29.6	13.2
3	20.3	25.5	12.2
5	19.7	28.1	10.5
10	**19.7**	**28.0**	**9.9**

Table 2. **(Left)** Effect of MNet outputs and use of "Interaction Aware Attention" (IAA) on the V2V error. "Pose" refers to SMPL-X pose parameters, and "Markers" to a MOJO-like [67] setup for the whole body. **(Right)** As the number of MNet's output frames increase, results improve but saturate around 10 frames.

4.2. Quantitative Evaluation

GNet: Table 1 reports the penetration volume (cm^3) and contact ratio [68] of four models: (1) "GrabNet" [57], which generates MANO grasps, (2) "GrabNet-SMPL-X", a variant that uses SMPL-X, (3) GNet without optimization, and (4) "GNet" with optimization. We see that generating whole-body grasps (row 2) is harder than hand-only grasps (row 1), yet "GNet" (row 4) outperforms baselines. Thus, post-processing optimization helps improve contact and reduce penetrations; small penetrations are inevitable as SMPL-X does not model soft tissue deformation; see [17].

MNet – IAA & output features: Table 2-left compares MNet with similar models that infer: (1) only SMPL-X pose parameters, "MNet-Pose", (2) only markers akin to MOJO [67], "MNet-Marker", and (3) MNet outputs without Interaction-Aware Attention, "MNet-w/o-IAA". We report vertex-to-vertex (V2V) errors for the full body, hands and feet on GRAB's held-out test set. Errors drop with: (1) using IAA features and (2) jointly inferring SMPL-X pose and marker offsets as output. We empirically observe that our combination of inputs and outputs, inspired from work on character control [56], leads to more realistic results.

MNet – Output frames number: We train 5 networks with outputs ranging from 1 to 10 frames, and report in Tab. 2-right the vertex-to-vertex (V2V) error for the body, feet, and hands between the generated and ground-truth meshes. Results show that generating more frames in each iteration of our auto-regressive scheme helps generate better results. We empirically observe that, when inferring a small number of future frames, sometimes the motion does not converge to a grasp and hands gradually deviate away from the object, instead of contacting it.

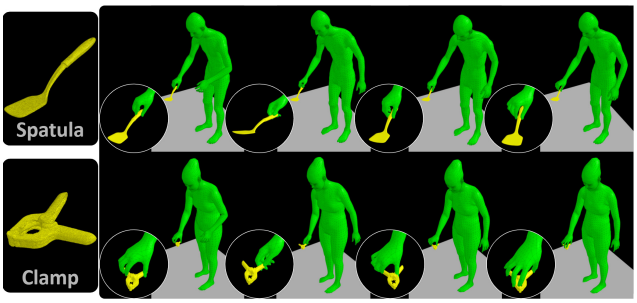

Figure 7. We show how GOAL generalizes by sampling 4 grasps from GNet's latent space for 2 complex, unseen YCB [5] objects.

Foot sliding: We evaluate foot-ground contact with a "foot sliding" metric. For each frame, we find the closest body vertex to the ground and compute its velocity. For contact, velocity should ideally be zero; if it is higher than 1 cm per frame, we consider the foot to be "sliding". GOAL generates sequences with 13.7% "foot-sliding" frames; GRAB's ground truth has 6.7%. Although there is room for improvement, GNet's feet "slide" less than existing work [19, 67].

4.3. Perceptual Evaluation

We evaluate GNet and MNet by generating grasping poses and motions, respectively, on GRAB's test set, and running a perceptual study via Amazon Mechanical Turk.

GNet: For each test-set object, we generate 2 "goal" whole-body grasps and render a "turntable animation" of these, before and after optimization, and the corresponding ground-truth grasps. Participants rate the quality of 4 features: (1) grasping pose, (2) foot-ground contact, (3) hand-object grasp, and (4) head orientation. They rate the realism of each feature on a Likert scale of scores between 1 (unrealistic) to 5 (very realistic). Each grasp is evaluated by at least 10 Participants. To remove invalid ratings, e.g., those who do not understand the task, we use catch trials similar to [57]. Results are shown in Tab. 3. The optimization step improves the realism of whole-body grasps. Moreover, it improves head orientation compared to the ground truth; this is because some GRAB subjects look away from the object while grasping it, while GNet produces heads oriented towards the object due to the explicit head orientation, $\boldsymbol{q}$, in Eq. (2). The same applies also for foot-ground contact, due to the explicit foot term, $\mathcal{L}_v^{\mathrm{f}}$, in Eq. (11). Overall, the quality of the generated grasps is close to the ground truth.

MNet: We show generated and ground-truth sequences to participants, and ask to rate the quality of: (1) overall body motion, (2) foot-ground contact, (3) hand-object grasp at the end of motion, (4) head orientation. Table 4 shows the results; GOAL generates grasping motions that approach the realism of ground truth. Note that MNet has a harder task than GNet as it generates a full motion. Also, Tabs. 3 and 4 show that ground truth is rated higher for motions than for static poses; this is harder for MNet to match.

Metric	GNet	GNet + Opt	Ground-truth [57]
Overall Grasping Pose ↑	3.89 ± 0.93	**3.98** ± 0.94	3.78 ± 1.06
Foot-Ground Contact ↑	3.98 ± 1.06	**4.10** ± 0.93	3.82 ± 1.11
Hand-Object Grasp ↑	2.70 ± 1.37	3.63 ± 1.16	**3.98** ± 1.04
Head Orientation ↑	3.83 ± 1.01	**4.01** ± 0.97	3.84 ± 1.07
Average ↑	3.60 ± 1.22	**3.93** ± 1.02	3.86 ± 1.07

Table 3. Perceptual study for GNet with-/out optimization. Subjects rate the realism of the grasp from 1 (unrealistic) to 5 (very realistic). We report the average rating value $\pm$ the standard deviation, computed across all valid participants. The optimization step ("GNet + Opt") improves all of the four studied features.

Metric	GOAL	Ground-truth [57]
Overall Body Motion ↑	3.74 ± 0.97	4.20 ± 0.90
Foot-Ground Contact ↑	3.88 ± 1.14	4.18 ± 1.05
Final Hand-Object Grasp ↑	3.66 ± 1.05	4.32 ± 0.91
Head Orientation ↑	3.86 ± 1.03	4.18 ± 1.00
Average ↑	3.79 ± 1.05	4.22 ± 0.97

Table 4. Evaluation of MNet motions. Participants rate the generated and ground-truth motions on a Likert scale of 1 (unrealistic) to 5 (very realistic) for 4 factors:overall body motion realism, feet-ground contact, final hand-object grasp, and head orientation.

5. Conclusion

We introduce GOAL, the first model to generate realistic human motions to grasp previously unseen 3D objects. We use two novel networks (GNet and MNet) to first generate a static "goal" grasp and then inpaint the motion between the frames. We exploit the ability of both networks to infer interaction features in Euclidean space and introduce an optimization step after each network to improve the quality of the grasps and motion based on the regressed features. The evaluation shows that our framework is able to synthesize natural and physically plausible grasping motions.

Future work: GOAL opens up many possibilities for future studies on grasping motion generation. Even though GOAL generates realistic grasping motions, it is constrained to be close to the object and can not generate motions when the body is far away. We plan to extend this to synthesize longer walking motions, prior to interaction with objects. In addition, here we focus on human-object interaction; we plan to combine GOAL with human-scene interaction models.

***Acknowledgements*:** This work was supported by the International Max Planck Research School for Intelligent Systems and the Max Planck ETH Center for Learning Systems. We thank Tsvetelina Alexiadis, Taylor McConnell, Joachim Tesch, and Benjamin Pellkofer, for helping with experiments, renderings, and the website.
Disclosure: https://files.is.tue.mpg.de/black/CoI_CVPR_2022.txt

References

[1] Rami Ali Al-Asqhar, Taku Komura, and Myung Geol Choi. Relationship descriptors for interactive motion adaptation. In *Symposium on Computer Animation (SCA)*, pages 45–53, 2013.

[2] Norman I. Badler, Cary B. Phillips, and Bonnie Lynn Webber. *Simulating Humans: Computer Graphics Animation and Control*. Oxford University Press, Inc., USA, 1993.

[3] Júlia Borras and Tamim Asfour. A whole-body pose taxonomy for loco-manipulation tasks. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 1578–1585, 2015.

[4] Matthew Brand and Aaron Hertzmann. Style machines. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 183–192, 2000.

[5] Berk Çalli, Arjun Singh, Aaron Walsman, Siddhartha S. Srinivasa, Pieter Abbeel, and Aaron M. Dollar. The YCB object and model set: Towards common benchmarks for manipulation research. In *International Conference on Advanced Robotics (ICAR)*, pages 510–517, 2015.

[6] Yu-Wei Chao, Jimei Yang, Weifeng Chen, and Jia Deng. Learning to sit: Synthesizing human-chair interactions via hierarchical control. In *Conference on Artificial Intelligence (AAAI)*, pages 5887–5895, 2021.

[7] Tao Chen, Jie Xu, and Pulkit Agrawal. A system for general in-hand object re-orientation. *Conference on Robot Learning (CoRL)*, 2021.

[8] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-Grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, 2022.

[9] Enric Corona, Albert Pumarola, Guillem Alenyà, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6990–6999, 2020.

[10] Enric Corona, Albert Pumarola, Guillem Alenyà, Francesc Moreno-Noguer, and Gregory Rogez. GanHand: Predicting human grasp affordances in multi-object scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5030–5040, 2020.

[11] George ElKoura and Karan Singh. Handrix: Animating the human hand. In *Symposium on Computer Animation (SCA)*, pages 110–119, 2003.

[12] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *International Conference on Computer Vision (ICCV)*, pages 4346–4354, 2015.

[13] Guillermo Garcia-Hernando, Edward Johns, and Tae-Kyun Kim. Physics-based dexterous manipulations with estimated hand poses and residual reinforcement learning. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 9561–9568, 2020.

[14] Michael Gleicher. Retargetting motion to new characters. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 33–42, 1998.

[15] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G. Ororbia. A neural temporal model for human motion prediction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12116–12125, 2019.

[16] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? In *Computer Vision and Pattern Recognition (CVPR)*, pages 1529–1536, 2011.

[17] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1471–1481, 2021.

[18] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2Motion: Conditioned generation of 3D human motions. In *International Conference on Multimedia (MM)*, pages 2021–2029, 2020.

[19] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J. Black. Stochastic scene-aware motion prediction. In *International Conference on Computer Vision (ICCV)*, pages 11374–11384, 2021.

[20] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 14708–14718, 2021.

[21] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, 2019.

[22] Edmond S. L. Ho, Taku Komura, and Chiew-Lan Tai. Spatial relationship preserving character motion adaptation. *Transactions on Graphics (TOG)*, 29(4):33:1–33:8, 2010.

[23] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *Transactions on Graphics (TOG)*, 36(4):1–13, 2017.

[24] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *Transactions on Graphics (TOG)*, 35(4):138:1–138:11, 2016.

[25] Kaijen Hsiao and Tomas Lozano-Perez. Imitation learning of whole-body grasps. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 5657–5662, 2006.

[26] Changgu Kang and Sung-Hee Lee. Environment-adaptive contact poses for virtual characters. *Computer Graphics Forum (CGF)*, 33(7):1–10, 2014.

[27] Mubbasir Kapadia, Xu Xianghao, Maurizio Nitti, Marcelo Kallmann, Stelian Coros, Robert W. Sumner, and Markus Gross. Precision: Precomputing environment semantics for contact-rich character animation. In *Symposium on Interactive 3D Graphics (SI3D)*, 2016.

[28] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. Grasping Field: Learning implicit representations for human grasps. In *International Conference on 3D Vision (3DV)*, pages 333–344, 2020.

[29] Vladimir G. Kim, Siddhartha Chaudhuri, Leonidas Guibas, and Thomas Funkhouser. Shape2pose: Human-centric shape analysis. *Transactions on Graphics (TOG)*, 33(4):120:1–120:12, 2014.

[30] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[31] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[32] Paul G. Kry and Dinesh K. Pai. Interaction capture and synthesis. *Transactions on Graphics (TOG)*, 25(3):872–880, 2006.

[33] Jehee Lee, Jinxiang Chai, Paul S. A. Reitsma, Jessica K. Hodgins, and Nancy S. Pollard. Interactive control of avatars animated with human motion data. *Transactions on Graphics (TOG)*, 21(3):491–500, 2002.

[34] Kang Hoon Lee, Myung Geol Choi, and Jehee Lee. Motion patches: Building blocks for virtual environments annotated with motion data. *Transactions on Graphics (TOG)*, 25(3):898–906, 2006.

[35] Kurt Leimer, Andreas Winkler, Stefan Ohrhallinger, and Przemyslaw Musialski. Pose to seat: Automated design of body-supporting surfaces. *Computer Aided Geometric Design (CAGD)*, 79, 2020.

[36] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI Choreographer: Music conditioned 3D dance generation with AIST++. In *International Conference on Computer Vision (ICCV)*, pages 13401–13412, 2021.

[37] Juncong Lin, Takeo Igarashi, Jun Mitani, Minghong Liao, and Ying He. A sketching interface for sitting pose design in the virtual environment. *Transactions on Visualization and Computer Graphics (TVCG)*, 18(11):1979–1991, 2012.

[38] Matthew Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *Transactions on Graphics (TOG)*, 33(6):220:1–220:13, 2014.

[39] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A Skinned Multi-Person Linear Model. *Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015.

[40] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The KIT whole-body human motion database. In *International Conference on Advanced Robotics (ICAR)*, pages 329–336, 2015.

[41] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision (ECCV)*, volume 12359, pages 474–489, 2020.

[42] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *International Conference on Computer Vision (ICCV)*, pages 9488–9496, 2019.

[43] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4674–4683, 2017.

[44] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3D human pose estimation. In *International Conference on Computer Vision (ICCV)*, pages 2659–2668, 2017.

[45] Josh Merel, Saran Tunyasuvunakool, Arun Ahuja, Yuval Tassa, Leonard Hasenclever, Vu Pham, Tom Erez, Greg Wayne, and Nicolas Heess. Catch & Carry: Reusable neural controllers for vision-guided whole-body tasks. *Transactions on Graphics (TOG)*, 39(4):39, 2020.

[46] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[47] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *Transactions on Graphics (TOG)*, 37(4):143:1–143:14, 2018.

[48] Xue Bin Peng, Glen Berseth, and Michiel Van de Panne. Terrain-adaptive locomotion skills using deep reinforcement learning. *Transactions on Graphics (TOG)*, 35(4):81:1–81:12, 2016.

[49] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *International Conference on Computer Vision (ICCV)*, pages 10985–10995, 2021.

[50] Sören Pirk, Vojtech Krs, Kaimo Hu, Suren Deepak Rajasekaran, Hao Kang, Yusuke Yoshiyasu, Bedrich Benes, and Leonidas J. Guibas. Understanding and exploiting object interaction landscapes. *Transactions on Graphics (TOG)*, 36(3):31:1–31:14, 2017.

[51] Nancy S. Pollard and Victor Brian Zordan. Physically based grasping control from example. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 311–318, 2005.

[52] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4331–4340, 2019.

[53] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. HuMoR: 3D human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, pages 11488–11499, 2021.

[54] Kathleen M. Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, Scott Fleming, Tina Brill, David Hoeferlin, and Dennis Burnsides. Civilian American and European Surface Anthropometry Resource (CAESAR) final report. Technical Report AFRL-HE-WP-TR-2002-0169, US Air Force Research Laboratory, 2002.

[55] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG)*, 36(6):245:1–245:17, 2017.

[56] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *Transactions on Graphics (TOG)*, 38(6):209:1–209:14, 2019.

[57] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, volume 12349, pages 581–600, 2020.

[58] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamics. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 935–941, 2018.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[60] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3D human motion and interaction in 3D scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9401–9411, 2021.

[61] Jack M. Wang, David J. Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(2):283–298, 2008.

[62] Yuting Ye and C. Karen Liu. Synthesis of detailed hand manipulations using contact sampling. *Transactions on Graphics (TOG)*, 31(4):41:1–41:10, 2012.

[63] Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J. Black. Human-aware object placement for visual environment reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2022.

[64] Ye Yuan and Kris Kitani. DLow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision (ECCV)*, volume 12354, pages 346–364, 2020.

[65] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. THUNDR: Transformer-based 3D human reconstruction with markers. In *International Conference on Computer Vision (ICCV)*, pages 12971–12980, 2021.

[66] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *International Conference on 3D Vision (3DV)*, pages 642–651, 2020.

[67] Yan Zhang, Michael J. Black, and Siyu Tang. We are more than our joints: Predicting how 3D bodies move. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3372–3382, 2021.

[68] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. Generating 3D people in scenes without people. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6193–6203, 2020.

[69] Youyi Zheng, Han Liu, Julie Dorsey, and Niloy J. Mitra. Ergonomics-inspired reshaping and exploration of collections of models. *Transactions on Visualization and Computer Graphics (TVCG)*, 22(6):1732–1744, 2015.

[70] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019.