Generative Proxemics: A Prior for 3D Social Interaction from Images

Lea Müller[1,4] Vickie Ye[1] Georgios Pavlakos[2,5] Michael Black[3] Angjoo Kanazawa[1]
[1]UC Berkeley [2]UT Austin [3]MPI for Intelligent Systems, Tübingen
{mueller,vye,kanazawa}@berkeley.edu, pavlakos@cs.utexas.edu, black@tuebingen.mpg.de

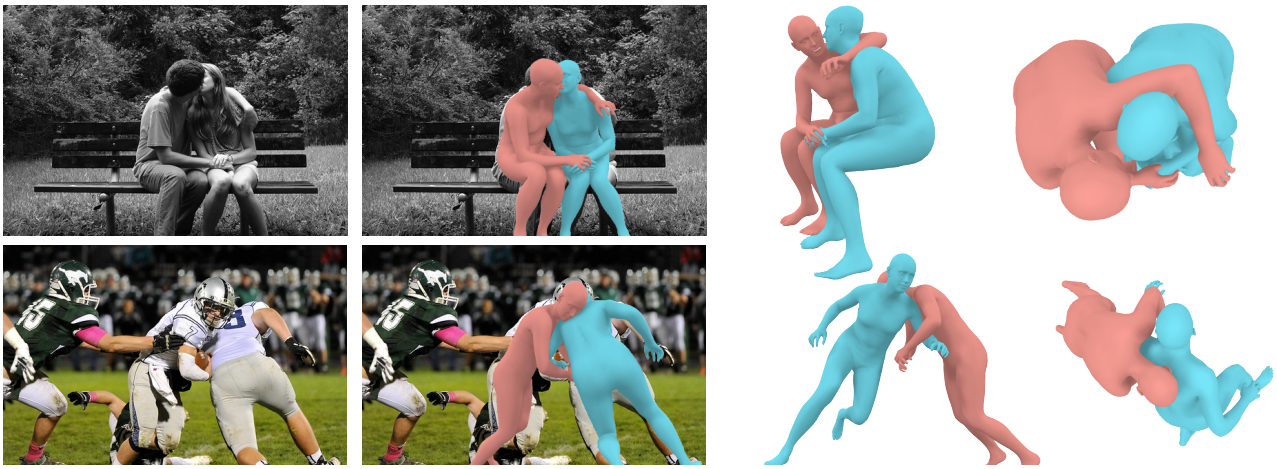

Figure 1. **Generative Proxemics.** We propose a diffusion model that learns a 3D generative model of two people in close social interaction. We show how the model can be used to generated samples or as a social prior in the downstream task of reconstructing two people in close proximity from images without any user annotation at test time. Shown here are input test images (left) and our predicted 3D bodies (right).

Abstract

Social interaction is a fundamental aspect of human behavior and communication. The way individuals position themselves in relation to others, also known as proxemics, conveys social cues and affects the dynamics of social interaction. Reconstructing such interaction from images presents challenges because of mutual occlusion and the limited availability of large training datasets. To address this, we present a novel approach that learns a prior over the 3D proxemics two people in close social interaction and demonstrate its use for single-view 3D reconstruction. We start by creating 3D training data of interacting people using image datasets with contact annotations. We then model the proxemics using a novel denoising diffusion model called BUDDI that learns the joint distribution over the poses of two people in close social interaction. Sampling from our generative proxemics *model produces realistic 3D human interactions, which we validate through a perceptual study. We use BUDDI in reconstructing two people in close proximity from an image without any contact annotation via an optimization approach that uses the diffusion model as a prior. Our approach recovers accurate 3D social interactions from noisy initial estimates, outperforming state-of-the-art methods. Our code, data, and model are available at:* `muelea.github.io/buddi`.

1. Introduction

Humans are social creatures, and physical interaction plays a crucial role in our daily lives, shaping our relationships. For example, research in behavioral science has shown that a slight touch between two people can cause a more friendly behaviour towards the touch-giver and lead to increased tips in restaurants [5]. Capturing and modeling scenarios of physical social interaction through computer vision will enable advancements in augmented and virtual reality and impact other fields like robotics and behavioral science. The recent progress in single-person mesh estimation is unfortunately not sufficient to model close interaction, as these methods are not trained to reason about the relative depth of people and the intricate interplay between two people's body poses, shapes, *and* proximity.

[4]Portions of this work were done while LM was at MPI-IS, Tübingen.
[5]This work was done while GP was at UC Berkeley.

In this work, we present the first approach that learns a generative model for 3D social proxemics and demonstrate its use as data-driven prior during an optimization routine. The diffusion model is trained using 3D human poses and shapes reconstructed from a large-scale image collection [8] using contact annotation, as well as using motion-capture (MoCap) data [8, 54]. The resulting model is able to generate the 3D pose and shape parameters of pairs of interacting people. When trained on bodies recovered from images, the model learns interactions depicted in photographs, such as people standing close together, playing sports, hugging, *etc*., see Figure 1. We further demonstrate the effectiveness of the learned prior by applying it to the challenging task of 3D human pose and shape reconstruction from a single image containing people engaged in social interaction.

Specifically, we propose BUDDI: a "BUDdies DIffusion Model". Diffusion models are established methods for image generation and are often used to model 3D human motion. In this work we use them to model 3D social proxemics. The majority of state-of-the-art diffusion-based methods for 3D human mesh generation operate on 3D joint locations [45]. This representation lacks information about the human body surface, which, intuitively, is important for reasoning about interpersonal contact. Our approach, in contrast, operates on the parameters of two parametric human body models, which represent the surfaces of two people closely interacting. After training, our model is able to generate samples of plausible pairs of 3D bodies in social interaction from pure noise. The model can also be conditioned in the output of a human pose and shape regressor. In this conditional case, the model takes the noisy output and generates similar poses but with realistic social interaction.

We then demonstrate how exploit BUDDI's knowledge of human proxemics to guide 3D mesh reconstruction of people in a close social interaction from a single image. To this end, we introduce a novel optimization-based approach, which uses BUDDI as a data-driven prior. We initialize our optimization routine with samples from BUDDI, conditioned to the output of a state-of-the-art multi-person human mesh regressor [42]. We then optimize over SMPL-X pose, shape, and translation parameters to match detected 2D joint locations. We incorporate guidance from the diffusion model using a loss inspired by the Score-Distillation loss from the 3D object creation literature [35]: In each optimization step, BUDDI refines the current estimate towards a more plausible social interaction conditioned on the initial predictions. The refined pose, shape and translation serve as prior in the overall objective function.

Our contributions include (1) presenting the first generative model of a pair of 3D people in close social interaction and (2) a novel approach for reconstructing 3D human meshes from images without relying on ground-truth contact annotations. We perform extensive experiments with BUDDI to evaluate its performance on the FlickrCI3D Signatures dataset [8] as well as CHI3D, and the recent Hi4D dataset [54] and find that it outperforms the state of the art as well as strong baselines. We also evaluate the unconditional samples from the diffusion model in a perceptual study, where people find our samples more realistic 44.4% when compared over real samples, where 50% is the upperbound where they do not see any difference. Importantly, we find that our optimization approach significantly improves the results of [42] both quantitatively and qualitatively. This work opens up a new avenue of research on digital human synthesis, laying the foundation for a deeper understaning of human social behavior derived from image data. Our data, code, and model will be available for research.

2. Related Work

Generating 3D humans. There has been recent interest in generating 3D humans, in different contexts. Several methods automatically populate static 3D scenes with 3D humans [14, 60, 61], while more recent methods generate both body and hand poses to interact with 3D objects [43, 44, 49]. Other work generates human motions conditioned on different inputs such as audio [26, 47] or text [33, 34, 45]. Concurrent work proposes text-to-3D diffusion-based approaches to generate motion of two interacting humans [27, 40]. Neither method predicts the full body surface, but rather they synthesizes either 3D joint locations or SMPL pose parameters for the average body. These methods are not used as priors for reconstructing interacting people from images.

To model 3D human proxemics probabilistically, we employ diffusion models, which achieve impressive performance on image generation tasks [7, 15, 38, 39]. They have recently been adopted in 3D human motion generation scenarios: MDM [45] generates plausible motions conditioned on text input; PhysDiff [55] incorporates physical constraints in the diffusion process to generate physically plausible motions; and EDGE [47] uses a transformer-based diffusion model for dance generation. Related work [4, 6, 28] has investigated different modalities for the conditioning, *e.g*., audio, text, or action classes. EgoEgo [25] generates plausible full-body motions conditioned on the head motion. SceneDiffuser [16] focuses on the scene-conditioned setting. We also rely on techniques from the diffusion literature, but consider the unique setting where two people are in close interaction and leverage this for single-image 3D reconstruction.

Multi-person 3D human mesh estimation. An extensive line of work focuses on reconstructing the 3D human pose and shape of a single person from images using optimization [2, 11, 24, 32, 37, 46, 50] or regression approaches [1, 12, 19, 20, 22, 29, 31, 51, 58, 59]. Capitalizing

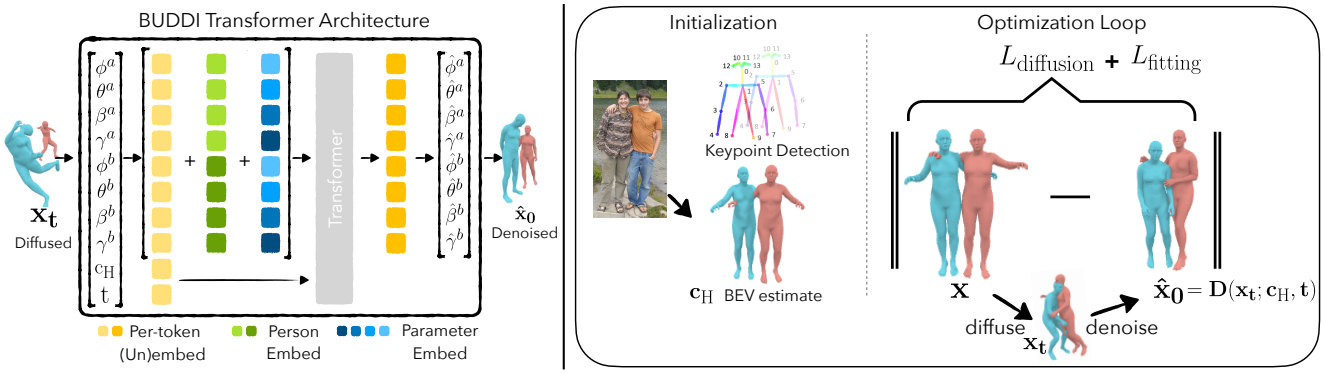

Figure 2. **BUDDI: BUddies DIffusion model.** On the left, we illustrate the architecture of BUDDI, our diffusion model for modeling 3D social proxemics between two people in close interaction. The diffusion process is applied directly on SMPL-X body parameters. To condition BUDDI on estimated body model parameters, c_{H}, we concatenate the parameters along the token dimension. On the right, we illustrate the optimization method with BUDDI as prior. Our optimization takes detected keypoints [3, 52] and an initial regressor estimate [42] as input. Given the regressor estimate, we sample from BUDDI to obtain $\tilde{x}$ which we use to initialize the optimization routine. In each optimization iteration, we take a single *diffuse-denoise* step on the current estimate using the learned denoiser model D. Our losses encourage the current estimate to be close to the refined meshes ($L_{\text{diffusion}}$) and to the initial estimate and detected keypoints (L_{fitting}).

on these techniques, recent approaches focus explicitly on reconstructing multiple people jointly from a single image. Zanfir *et al*. [56] propose an optimization solution, while Jiang *et al*. [18] and Sun *et al*. [41] rely on deep networks to regress the pose and shape for all people in the image. BEV [42] extends ROMP [41] to reason about the depth of people while taking age/height into account.

The above methods do not address contact between people. To do so, Fieraru *et al*. [8] introduce the first datasets with ground-truth labels for the body regions in contact between humans. Labels are collected using MoCap (CHI3D) or human annotators (FlickrCI3D Signatures). They propose an optimization approach that requires the ground-truth contact map to reconstruct people in close proximity at test time. More recently, REMIPS [9] introduces a transformer-based method that regresses the 3D pose of multiple people. REMIPS is trained using the above datasets while taking into account contact and interpenetration. In this work, we take a very different approach by learning and exploiting a 3D generative proxemics prior. We use the ground-truth contact maps to generate pseudo-ground truth 3D human fits from which we learn the diffusion model; once this is learned, we show that it can be used as a prior to recover plausible bodies in close proximity from images without explicit knowledge of contact maps.

Data-driven priors in optimization. Optimization-based methods for 3D human pose and shape estimation, like SMPLify [2], are versatile and allow different data-driven prior terms to be incorporated in the objective function. Different methods have been used to learn pose priors including GMMs [2], VAEs [32], neural distance fields [46], and normalizing flows [57]. ProHMR [23] learns a pose prior conditioned on image pixels. HuMoR [37] incorporates a data-driven motion prior in the iterative optimization. POSA [14] learns a prior for human-scene interaction from PROX data [13] and uses it in their optimization. In contrast to these methods, we use a diffusion model to capture the joint distribution over SMPL-X parameters for two people interacting and show that we can both sample from the model and use during optimization to improve the pose estimates of interacting people.

3. Method

We introduce BUDDI, a generative model of two people in close social interaction. Because of the complexity and multimodality of the data, we turn to denoising diffusion probabilistic models [15] to address this task. In Sec. 3.1, we describe the basics of diffusion, and the parameterization we employ to model people in contact. In addition to sampling new body meshes from our model, our generative model can serve as a prior for reconstructing 3D humans from images. In Sec. 3.2, we describe an optimization procedure that incorporates BUDDI as a prior to recover two SMPL-X meshes from observed 2D keypoints.

For all of the following, we use the SMPL-X [32] body model to represent the human bodies. SMPL-X is a differentiable function that maps pose, $\boldsymbol{\theta} \in \mathbb{R}^{21\times3}$, shape, $\boldsymbol{\beta} \in \mathbb{R}^{10}$, and expression, $\boldsymbol{\psi} \in \mathbb{R}^{10}$ parameters to a mesh consisting of $N_v = 10,475$ vertices $V \in \mathbb{R}^{N_v\times3}$. We place the generated meshes in the world by rotating and translating them by $\boldsymbol{\phi} \in \mathbb{R}^3$ and$\boldsymbol{\gamma} \in \mathbb{R}^3$. We denote person a's parameters as $X^a = [\boldsymbol{\phi}^a, \boldsymbol{\theta}^a, \boldsymbol{\beta}^a, \boldsymbol{\gamma}^a]$ and $X^b = [\boldsymbol{\phi}^b, \boldsymbol{\theta}^b, \boldsymbol{\beta}^b, \boldsymbol{\gamma}^b]$. For simplicity, we refer to both people when no index is specified, *e.g*., X refers X^a and X^b

3.1. Diffusion Model for 3D Proxemics

Denoising diffusion models are latent variable generative models that learn to transform random noise into the desired data distribution p_{data} through a forward and reverse process. The forward inference process is a Markov chain over T steps given by transitions $q(\mathbf{x}_{t+1}|\mathbf{x}_t)$, which gradually adds Gaussian noise to clean samples $\mathbf{x}_0$ from the data distribution according to a fixed variance schedule σ_t.

The reverse process $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ then gradually denoises noisy samples back into the data distribution. The reverse process transitions follow a Gaussian distribution when conditioned on x_0, but must be inferred during the generative process. Following Ramesh et al. [36], we train a neural network D that predicts a sample $\hat{\mathbf{x}}_0 = D(\mathbf{x}_t; t)$ from a noisy sample $\mathbf{x}_t$ given the noise level t.For the task of reconstructing humans from images, when we have initial estimates of the SMPL-X parameters of two humans, we condition the denoising network D on $\mathbf{c}_H$, the predicted SMPL-X parameters of the two humans by a regressor.

We refer to the process of adding noise as *diffusion* and the process of removing the noise via D as *denoising*. Specifically, we diffuse a ground-truth sample $\mathbf{x}_0$ by uniformly sampling a noise level t with $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$ to obtain the noisy sample $\mathbf{x}_t = \sqrt{\sigma'_t}\mathbf{x}_0 + \sqrt{1-\sigma'_t}\epsilon_t$ with $\sigma'_t = \prod_{i=1}^{t}(1-\sigma_t)$.

We then train D to minimize

$$\mathbb{E}_{\mathbf{x}_0 \sim p_{\text{data}}} \mathbb{E}_{t \sim \mathcal{U}\{0,T\}, \mathbf{x}_t \sim q(\cdot|\mathbf{x}_0)} ||D(\mathbf{x}_t; t, \mathbf{c}_H) - \mathbf{x}_0||, \quad (1)$$

where we set $\mathbf{c}_H = \varnothing$ for 20% of conditional model training, and all of unconditional model training.

Architecture. Because we aim to model close contact between people, we choose a model state space that can express the full surface of the human body. Specifically, in contrast to prior work in human motion diffusion that operate only on joint angles and locations [45, 55], we directly operate on the full SMPL-X parameters of the two people. A sample $\mathbf{x}$ thus corresponds to the concatenation of two bodies:

$$\mathbf{x} = [X^a, X^b] = [\boldsymbol{\phi}^a, \boldsymbol{\theta}^a, \boldsymbol{\beta}^a, \boldsymbol{\gamma}^a, \boldsymbol{\phi}^b, \boldsymbol{\theta}^b, \boldsymbol{\beta}^b, \boldsymbol{\gamma}^b].$$

We denoise a sample $\mathbf{x}_t$ with a transformer encoder block on tokenized parameters. Specifically, each parameter of each person is tokenized into 152-dimensional latent vectors with per-parameter and per-person embedding layers. We tokenize the noise level t similarly with a noise embedding. When conditioning is available, i.e. SMPL-X estimates for person a and b, we similarly tokenize the parameters to be used as additional tokens. We pass the available tokens into the transformer encoder, and similarly decode the output tokens with per-token embeddings. We illustrate the denoiser architecture in Figure 2.

Losses. We employ standard human pose and shape regularization losses. We write our training objective as

$$L_D = L_\theta + L_\beta + L_{\boldsymbol{\gamma}} + L_{v2v}, \quad (2)$$

where L_θ, L_β, $L_{\boldsymbol{\gamma}}$ denote squared L2-losses on respective body model parameters, and L_{v2v} denotes a squared L2 loss on model vertices. We use 6D rotation representations [62] for global orientation and pose, and model the relative translation between a and b. We show generated samples from our unconditional model in Fig. 3.

3.2. Optimization with the Proxemics Prior

Reconstructing 3D human meshes from a single image is an extremely under-constrained problem, and priors over human pose and shape are crucial in an optimization based framework for recovering plausible meshes [2, 32, 56]. Our problem involves people in close contact, which requires correctly placing the meshes in context with each other, which has only been done when given ground truth contact annotations at test time [8]. We remove the need for ground-truth contact maps by using our generative model as a prior during reconstruction with a score distillation approach [35, 48].

During inference, we observe detected 2D keypoints $\tilde{J}_{2D}$ and initial body model parameter estimates $\mathbf{c}_H$ from a regressor [42]. We then optimize the body parameters of two people to minimize

$$L_{\text{Optimization w. BUDDI}} = L_{\text{fitting}} + L_{\text{diffusion}}. \quad (3)$$

L_{fitting} ensures that the solution stays close to the image evidence, while $L_{\text{diffusion}}$ is a data-driven prior using our conditional diffusion model. We treat this prior as similar to those used for 3D pose in previous works such as GMM [2] and V-Poser [32], but for 3D proxemics. We illustrate the optimization procedure in Fig. 2 right.

We initialize our optimization by generating a sample $\tilde{\mathbf{x}}$ from the conditional model. We sample with DDIM sampling with 100 evenly spaced steps. We then use the data fitting loss:

$$L_{\text{fitting}} = \lambda_J L_J + \lambda_{\tilde{\theta}} L_{\tilde{\theta}} + \lambda_P L_P, \quad (4)$$

where L_J denotes 2D re-projection error between the reprojected 3D joints of the current estimate and the detected 2D keypoints, and $L_{\tilde{\theta}}$ is a prior for the solution to be close to the denoised initialization.

L_P denotes an interpenetration loss between two people that pushes inside vertices to the surface, which we use winding numbers [17] to find intersecting vertices between low-resolution SMPL-X meshes of the current estimates:

$$L_P = \sum_{v \in V_I^a} \min_{u \in V^b} \|v - u\|^2 + \sum_{v \in V_I^b} \min_{u \in V^a} \|v - u\|^2, \quad (5)$$

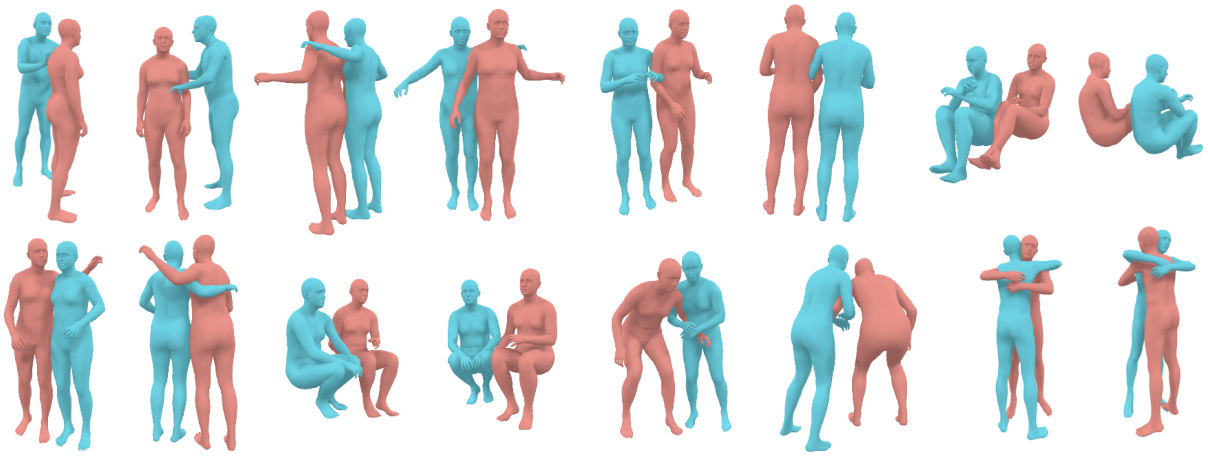

Figure 3. **Generative Proxemics: Samples from BUDDI.** All samples are unconditionally generated from pure noise using the trained diffusion model. We select several representative examples and show two views per sample. These samples reveal that BUDDI has learned the distribution of people in close contact including embracing each other, playing sports, sitting side by side, and taking photographs.

where V_I^a denotes vertices of M^a intersecting the low-resolution mesh of M^b; and vice versa for V_I^b.

To use the prior on human interaction into account, we use the learned denoising model D from BUDDI and perform a single *diffuse-denoise* step, with a noise level at $t = 10$, on the current estimate. The denoised estimate, $\hat{\mathbf{x}}_0 = D(\mathbf{x}_t; t, \mathbf{c}_H)$, regularizes the current estimate via

$$L_{\text{diffusion}} = ||D(\mathbf{x}_t; t, \mathbf{c}_H) - \mathbf{x}||, \tag{6}$$

where $\mathbf{x}_t = \sqrt{\sigma_t'}\mathbf{x}_{\texttt{no-grad}} + \sqrt{1-\sigma_t'}\epsilon_t$ denotes the diffused body model parameters of the current estimate, and $\mathbf{x}_{\texttt{no-grad}}$ denotes the current estimate with detached gradients. $\hat{\mathbf{x}}_0$, and encourages $\mathbf{x}$ to be close to $\hat{\mathbf{x}}_0$. In practice, we penalize the decoded parameters of $\mathbf{x}$ and $\hat{\mathbf{x}}_0$ directly as

$$\begin{aligned} L_{\text{diffusion}} =& \lambda_{\hat{\boldsymbol{\phi}}}||\hat{\boldsymbol{\phi}}_0 - \boldsymbol{\phi}|| + \lambda_{\hat{\boldsymbol{\theta}}}||\hat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}|| \\ &+ \lambda_{\hat{\boldsymbol{\beta}}}||\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}|| + \lambda_{\hat{\boldsymbol{\gamma}}}||\hat{\boldsymbol{\gamma}}_0 - \boldsymbol{\gamma}||. \end{aligned} \tag{7}$$

Intuitively, this loss uses the learned denoiser D to take a step from the current estimate towards the data distribution of two people in close proximity, conditioned on the regressor prediction.

4. Implementation Details

Training Data. There are few datasets containing 3D ground truth of humans in close social interaction [8, 54]. Such datasets are usually captured in lab environments, consequently they are small and do not contain the variety of interactions between humans "in the wild," *e.g.* when playing sports or taking social pictures. To address this lack of data, we create **Flickr Fits**, *i.e.* SMPL-X fits for Flickr images portraying humans in contact scenarios. For this, we use FlickrCI3D Signatures [8], a dataset of images showing interacting humans collected from Flickr with discrete human-human contact annotations. Specifically, the SMPL-X body surface is divided into $R = 75$ regions such that each region, r, roughly covers a similar area. For a given photo, the human annotators assign a binary label indicating contact between a region on one person and a region on the other. For two meshes, M^a and M^b, the annotation can be represented as a binary contact map $\mathcal{C}^D \in \{0,1\}^{R\times R}$, where

$$\mathcal{C}_{ij}^D = \begin{cases} 1, & \text{if } r_i \text{ of } M^a \text{ is in contact with } r_j \text{ of } M^b \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

We use these ground-truth contact maps in an optimization routine for fitting two people to detected keypoints, similar to Sec. 3.2 but replace the diffusion model prior with standard image fitting priors. The dataset contains 10,631/1,139 train/test images, with one image containing multiple contact annotations. Note that we use pairs of people in contact as labeled in the FlickrCI3D Signatures dataset [8] and only use images containing matching regressor estimates, 2D keypoints, and contact labels. See Fig. 4 for qualitative examples and the Sup. Mat. for more details about the selection process and an evaluation of the optimization method with ground-truth contact labels.

We also augment our training data with available MoCap data, which is considerably smaller than those obtained from image fits: **CHI3D** [8] contains 3/2 pairs of training/test subjects performing 127 sequences of two-person interactions like hugs or kicks with ground-truth SMPL-X bodies. One frame per sequence has contact map annota-

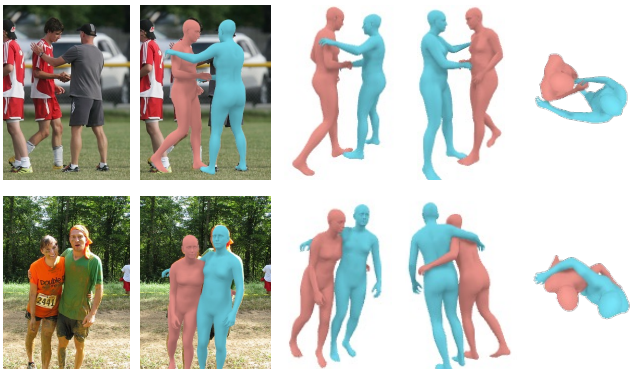

Figure 4. **Flickr Fits.** We visualize the output of the optimization process that reconstructs two people in close proximity using ground-truth contact maps, shown from four different views. We use these 3D fits as training data for BUDDI.

tions. We use only the contact frame of the sequences from two subject pairs, resulting in 247 mesh pairs for training, and the third pair for evaluation. **Hi4D** [54] contains sequences of 20 pairs of people interacting with each other. The interactions include actions like hugging, dancing, and fighting. We randomly split the data into 14/3/3 pairs for train/val/test and use every fifth frame of the subsequence involving contact as labeled in Hi4D, resulting in about 1K mesh pairs for training. The body representation format in Hi4D is SMPL, which we transfer to SMPL-X using the SMPL-X code repository [32]. Please see the Sup. Mat. for more details of the datasets. Note that while we use SMPL-X model, BUDDI is not trained on hands because none of these datasets contain hand poses.

BUDDI Training. BUDDI is trained with meshes from FlickrCI3D Signatures Fits, CHI3D, and Hi4D. We use 60% Flickr, 20% CHI3D, and 20% Hi4D data distribution per batch with batch size 512. The transformer backbone has six layers and eight heads; we use 10% dropout and randomly shuffle the order of people during training. To train BUDDI, we randomly sample noise levels t up to 1000 using a cosine noise schedule [30]. We use the Adam optimizer [21] with learning rate 10^{-4}. We train two separate networks: an unconditional model for generation and the conditional version for reconstruction. For the conditional model, we use all camera views of the MoCap datasets, *i.e.* 4/8 cameras for CHI3D/Hi4D. The unconditional model is trained on 3D MoCap fits in the world coordinate system. To sample new poses, we use DDIM sampling starting at noise levels $t = 1000$ in steps of 10.

Optimization Details. During optimization, we experiment with different noise levels, between 10 and 100, and find that $t = 10$ does not disturb the inputs too much, but enough for D to generate new configurations. We use detected 2D keypoints from OpenPose [3] and ViTPose [52] and BEV [42] estimates as conditioning. Unlike single-person mesh regressors, BEV is designed to predict multiple people including their relative depth. Please see Sup. Mat. for more details.

5. Experiments

Baselines. We compare our reconstruction method with BEV [42], which is also used as an input to our conditional model. Since there is no other available work that reasons about people in close social interaction, we experiment with simple but effective baselines. We train the transformer model of BUDDI to directly predict SMPL-X parameters of people in contact from BEV input, essentially a deterministic, single-step ablation of BUDDI. We also evaluate the direct conditional denoised output of BEV by BUDDI without any optimization. As another baseline, we propose an optimization routine that replaces $L_{\text{diffusion}}$ with a simple heuristic that takes the minimal distances between two meshes predicted by BEV and minimizes their distance during optimization along with the other energy terms. Finally, to compare the generation ability we train a VAE which we also use during the optimization routine in a similar manner to VPoser [32] but for two people by optimizing the VAE latent space instead of SMPL-X parameters. We refer to these models as *Transformer*, *BUDDI (gen.)*, *Contact Heuristic*, and *VAE*, respectively. All baselines are trained on the same datasets as BUDDI with the same sampling strategies. Details about our baselines are provided in the Sup. Mat..

Metrics. We use standard evaluation metrics from the human pose and shape estimation literature. We also report the joint PA-MPJPE computed by performing Procrustes alignment of both people together. In addition to per-person metrics, this captures the relative orientation and translation of the two people. Since our method directly estimates 3D humans we propose a new metric similar to PCK [53] from the 2D pose literature called **PCC**, the percentage of correct contact points with respect to a radius r. Specifically, given two meshes, M^a/M^b and a contact map $\mathcal{C}^D$ we compute the pairwise vertex-to-vertex Euclidean distances $d_{\text{eucl}}(\mathcal{C}^D)$ between annotated contact regions and consider the pair to be correct when $\min(d_{\text{eucl}}(\mathcal{C}^D)) < r$.

5.1. Unconditional Generation

We qualitatively evaluate BUDDI by showing samples generated from the model in Fig. 1 and 3. Our approach is able to generate people in close proximity including embraces, handshakes, having a conversation, sitting side by side, and, in general, plausibly interacting with each other. Since it is trained on Internet image collections, it also learns to generate people posing for photographs or playing sports.

We further run a perceptual study to evaluate the realism of the generated social interactions against other methods. In a forced choice study, we compare our generated samples with samples from the real data distribution according to the 60/20/20 per-batch ratio for Flickr/CHI3D/Hi4D

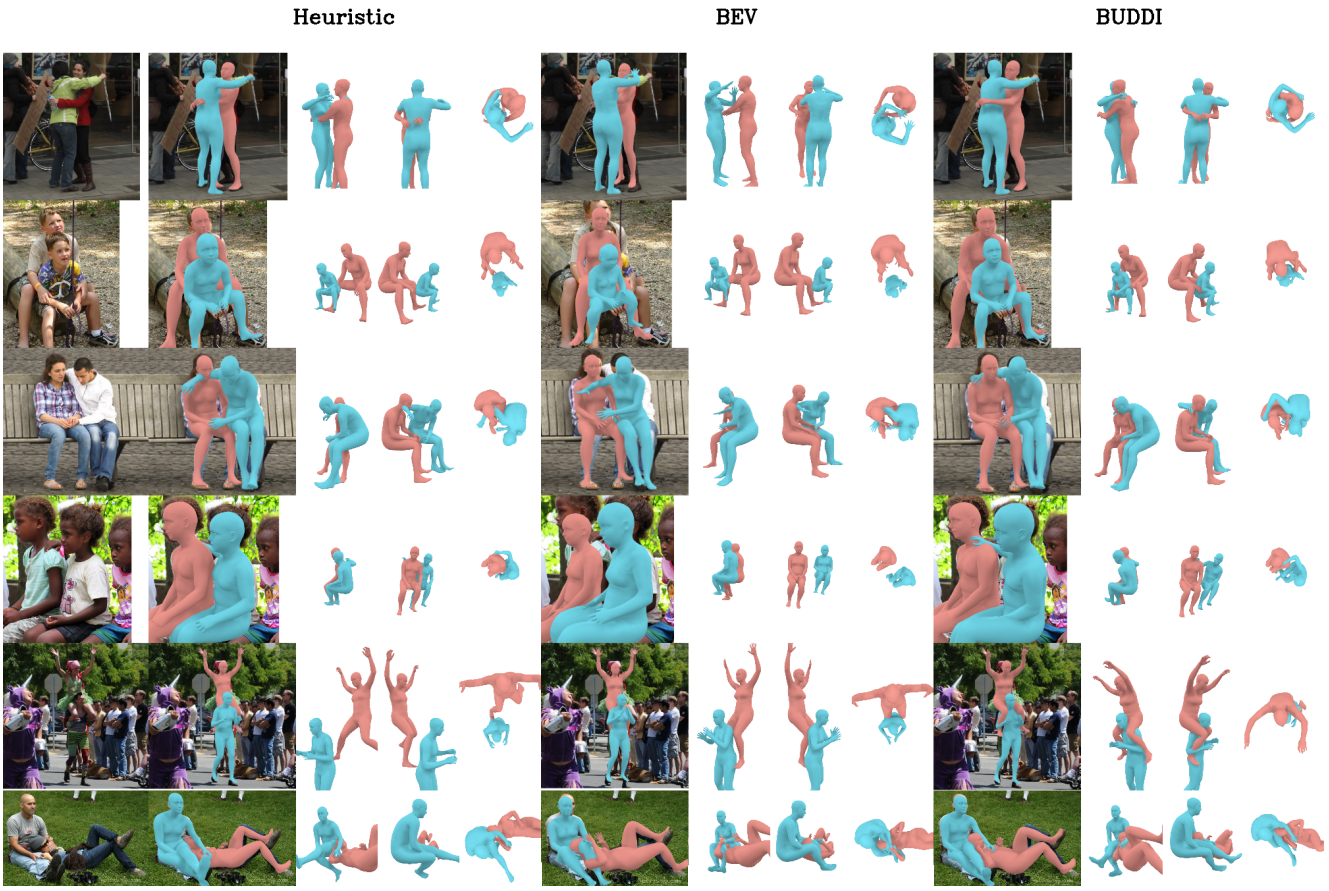

Figure 5. **Automatic reconstruction of people in close social interaction.** We show qualitative results from a) BEV, b) contact heuristics, which takes the BEV output and encourages the closest parts to be in contact, and c) our method, which optimizes the BEV estimates against the image evidence with the BUDDI prior. Our approach recovers a plausible reconstruction with subtle details.

	JOINT ↓	PCC at radius ↑				
	PA-MPJPE	5	10	15	20	25
BEV	106	-	-	-	-	-
Transformer	86	14	40	60	**73**	**82**
BUDDI (gen.)	92	15	39	58	71	80
Heuristic	68	14	34	49	61	70
VAE	101	11	28	42	55	65
BUDDI	**66**	**19**	**44**	**62**	**73**	81

Table 1. **3D Pose Evaluation on FlickrCI3D Signatures.** We evaluate methods against the Flickr fits using their joint (two-person) PA-MPJPE expressed in mm. We also evaluate the percentage of correct contact points (PCC) for radius r mm.

used during training. We also compare BUDDI against generations from the VAE and a non-parametric random baseline that samples meshes from the pseudo-ground truth after centering the two people. We do a forced choice comparison between BUDDI and these there other methods, asking workers on Amazon Mechanical Turk to choose the sample that shows a more realistic close social interaction. We use 256 samples per method. We collect ratings for 768 pairwise comparisons. In this study, BUDDI was chosen over random in 71.23% of the comparisons, over the VAE in 60.17%, and over the training data in 44.4%. Note that 50% is the upper bound for such forced choice comparisons, in which participants cannot tell the difference between real and generated samples.

For a quantitative evaluation, we compute the FID score between samples from BUDDI and samples from the *VAE* on concatenated SMPL-X parameters. We sample 8K examples per method and from our training data following the dataset ratio per batch. BUDDI has a lower FID score (1.6) compared to the *VAE* (3.3).

5.2. Fitting with BUDDI

We show qualitative results in Fig. 5 comparing BUDDI against BEV and the Contact Heuristic. Our approach is able to generate various types of human interactions with plausible contact and depth placement. It is also able to capture close interaction between a child and a parent. Although the Contact Heuristic (center) is able to move two

	PER PERSON ↓ PA-MPJPE	JOINT ↓ PA-MPJPE	JOINT PA-MPJPE ↓ backhug	basketball	cheers	dance	fight	highfive	hug	kiss	pose	sidehug	talk
BEV	78 / 84	136	200	126	109	135	121	106	163	139	142	131	118
Heuristic	**67 / 71**	121	168	**83**	94	131	**94**	**68**	159	159	118	113	109
BUDDI (F, C)	70 / 77	115	200	94	92	128	108	100	133	114	104	107	91
Transformer	79 / 85	120	161	141	103	138	123	128	117	106	120	105	100
BUDDI (gen.)	82 / 90	117	152	139	120	137	130	96	101	97	115	102	101
VAE	80 / 82	138	175	133	114	141	119	87	176	162	135	140	113
BUDDI	70 / 76	**98**	**127**	95	**92**	**113**	109	72	**105**	**85**	**88**	**96**	**81**

Table 2. **Evaluation of BUDDI on Hi4D.** We compare the output of BUDDI to the proposed baseline methods on the Hi4D challenge. The first block shows methods that do not use Hi4D data during training or are optimization based without access to priors trained on Hi4D. BUDDI (F,C) in particular, is our model BUDDI trained on Flickr and CHI3D data only. All errors are reported in mm for 3D Joints.

	PER PERSON ↓ PA-MPJPE	JOINT ↓ PA-MPJPE
BEV	50 52	96
Transformer	54 56	105
BUDDI (gen.)	53 53	80
Heuristic	49 **46**	105
VAE	54 54	103
BUDDI	**48** 47	**68**

Table 3. **Quantitative Evaluation on CHI3D.** We compare the output of our model to the baselines on CHI3D (pair s03). All errors reported in mm for 3D Joints.

people closer together, which helps with image alignment, upon close observation it is not able to capture the subtle interaction between people that happens during intimate interaction. BUDDI's estimates are more realistic and better capture the subtle details of interaction. We provide additional qualitative examples in the Sup. Mat.

We further report the percentage of correct contact (PCC) with respect to the ground truth contact map on the FlickrCI3D Signatures test set in Table 1. The table also shows the pose reconstruction accuracy against our Flickr Fits. All metrics show improvement over BEV, in particular the joint PA-MPJPE. Non-optimization methods, *i.e*. *Transformer* and *BUDDI (gen.)*, are able to predict plausible contacts, with similar PCC accuracy to BUDDI, but struggle to reconstruct the data with a worse joint PA-MPJPE. The *Heuristic*, in contrast, achieves a lower reconstruction error, but worse PCC. Our approach which leverages the learned prior during optimization can recover both the relative positions and contacts between the two people. To provide insights into the performance of single-person mesh regressors when evaluated on the two-person reconstruction task, we run 4D Humans [10] on Flickr Fits. The joint PA-MPJPE is 344 mm which is high, as expected, since these methods are not trained to reason about proximity.

We further evaluate our model against ground truth MoCap data in Tables 2 and 3. Optimization with BUDDI consistently improves the two-person reconstruction error over BEV and other baselines. When evaluated per action, the strongest improvements over BEV come from complex close social interactions like hugging or kissing, at 58mm and 54mm absolute improvement over BUDDI respectively. The *Heuristic* baseline achieves a low PA-MPJPE reconstruction error on all three datasets in particular for poses with a few physical contact points, such as a handshake, whereas more complex contact, such as a hug, requires data-driven priors like BUDDI; we provide an analysis in the Sup. Mat. to quantify this hypothesis. *Transformer* and *BUDDI (gen.)* have lower joint PA-MPJPE errors than BEV and the *Heuristic*, but worse per-person reconstruction errors. The *VAE* results suggest that directly operating in the latent space of a generative model is challenging and not sufficient to accurately recover close social interactions. BUDDI, in contrast, is able to model a wide variety of poses, as supported by the numerical results.

6. Conclusion

We propose BUDDI, a diffusion model for close human-human interaction. We train BUDDI from 3D fits obtained from a large-scale dataset of images with ground truth contact annotations as well as a small set of available mocap data. BUDDI enables unconditional sampling of people in close social interaction. More importantly, we also demonstrate how BUDDI can be used as effective prior for single-view 3D reconstruction of two people in close proximity.

Our core contribution of a generative proxemics prior provides the foundation for future work on modeling and capturing human interaction. For example, future work could iteratively apply our method to reconstruct more than two people in close proximity, explore other conditioning modalities like pixel features, text, action labels, or, by taking the outputs of recent single-person mesh regressors into account, address reconstructing fine-grained interaction including finger pose and facial expressions.

Acknowledgements: We thank A. Holynski, E. Weber, F. Warburg, B. Peebles, J. Rajasegaran, K. Mangalam and N. Athanasiou for insightful discussions, and A. Cseke, T. McConnell and T. Alexiadis for running the user study.

Disclosure: `https://files.is.tue.mpg.de/black/CoI_CVPR_2024.txt`

References

[1] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3D human pose estimation in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, pages 561–578. Springer International Publishing, 2016.

[3] Zhe Cao, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43 (1):172–186, 2019.

[4] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023.

[5] April H Crusco and Christopher G Wetzel. The midas touch: The effects of interpersonal touch on restaurant tipping. *Personality and Social Psychology Bulletin*, 10(4): 512–517, 1984.

[6] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, 2023.

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Conference on Neural Information Processing Systems (NeurIPS)*, 34:8780–8794, 2021.

[8] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7214–7223, 2020.

[9] Mihai Fieraru, Mihai Zanfir, Teodor Szente, Eduard Bazavan, Vlad Olaru, and Cristian Sminchisescu. REMIPS: Physically consistent 3D reconstruction of multiple interacting people under weak supervision. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 19385–19397, 2021.

[10] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023.

[11] Peng Guan, Alexander Weiss, Alexandru Balan, and Michael J. Black. Estimating human shape and pose from a single image. In *International Conference on Computer Vision (ICCV)*, pages 1381–1388, 2009.

[12] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[13] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision (ICCV)*, pages 2282–2292, 2019.

[14] 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.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Conference on Neural Information Processing Systems (NeurIPS)*, pages 6840–6851, 2020.

[16] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[17] Alec Jacobson, Ladislav Kavan, and Olga Sorkine-Hornung. Robust inside-outside segmentation using generalized winding numbers. *ACM Transactions on Graphics (TOG)*, 32(4): 1–12, 2013.

[18] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5578–5587, 2020.

[19] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human model fitting towards in-the-wild 3D human pose estimation. In *International Conference on 3D Vision (3DV)*, 2021.

[20] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018.

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[22] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019.

[23] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *International Conference on Computer Vision (ICCV)*, 2021.

[24] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6050–6059, 2017.

[25] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023.

[26] 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.

[27] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion

generation under complex interactions. *arXiv preprint arXiv:2304.05684*, 2023.

[28] Jianxin Ma, Shuai Bai, and Chang Zhou. Pretrained diffusion models for unified human motion synthesis. *arXiv preprint arXiv:2212.02837*, 2022.

[29] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9990–9999, 2021.

[30] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, pages 8162–8171, 2021.

[31] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *International Conference on 3D Vision (3DV)*, pages 484–494, 2018.

[32] 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.

[33] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10985–10995, 2021.

[34] Mathis Petrovich, Michael J Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, pages 480–497, 2022.

[35] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022.

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[37] 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)*, 2021.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[40] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023.

[41] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11179–11188, 2021.

[42] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13243–13252, 2022.

[43] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13263–13273, 2022.

[44] Purva Tendulkar, Dídac Surís, and Carl Vondrick. FLEX: Full-body grasping without full-body grasps. In *Computer Vision and Pattern Recognition (CVPR)*, 2023.

[45] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[46] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*, 2022.

[47] Jonathan Tseng, Rodrigo Castellon, and C Karen Liu. EDGE: Editable dance generation from music. *Computer Vision and Pattern Recognition (CVPR)*, 2023.

[48] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023.

[49] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision (ECCV)*, pages 257–274, 2022.

[50] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6183–6192, 2020.

[51] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare. In *International Conference on Computer Vision (ICCV)*, 2019.

[52] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022.

[53] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012.

[54] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Computer Vision and Pattern Recognition (CVPR)*, 2023.

[55] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16010–16021, 2023.

[56] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene

constraints. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2148–2157, 2018.

[57] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3D human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision (ECCV)*, pages 465–481, 2020.

[58] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3D human pose and shape. In *Computer Vision and Pattern Recognition (CVPR)*, pages 14484–14493, 2021.

[59] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *International Conference on Computer Vision (ICCV)*, pages 11446–11456, 2021.

[60] 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.

[61] 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 6194–6204, 2020.

[62] 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.