BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion

Michael J. Black[1,*] Priyanka Patel[1,*] Joachim Tesch[1,*] Jinlong Yang[2,*,†]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]Google

Figure 1. **BEDLAM** is a large-scale synthetic video dataset designed to train and test algorithms on the task of 3D human pose and shape estimation (HPS). BEDLAM contains diverse body shapes, skin tones, and motions. Beyond previous datasets, BEDLAM has SMPL-X bodies with hair and realistic clothing animated using physics simulation. With BEDLAM's realism and scale, we find that synthetic data is sufficient to train regressors to achieve state-of-the-art HPS accuracy on real-image datasets without using any real training images.

Abstract

We show, for the first time, that neural networks trained only on synthetic data achieve state-of-the-art accuracy on the problem of 3D human pose and shape (HPS) estimation from real images. Previous synthetic datasets have been small, unrealistic, or lacked realistic clothing. Achieving sufficient realism is non-trivial and we show how to do this for full bodies in motion. Specifically, our BEDLAM dataset contains monocular RGB videos with ground-truth 3D bodies in SMPL-X format. It includes a diversity of body shapes, motions, skin tones, hair, and clothing. The clothing is realistically simulated on the moving bodies using commercial clothing physics simulation. We render varying numbers of people in realistic scenes with varied lighting and camera motions. We then train various HPS regressors using BEDLAM and achieve state-of-the-art accuracy on real-image benchmarks despite training with synthetic data. We use BEDLAM to gain insights into what model design choices are important for accuracy. With good synthetic training data, we find that a basic method like HMR approaches the accuracy of the current SOTA method (CLIFF). BEDLAM is useful for a variety of tasks and all images, ground truth bodies, 3D clothing, support code, and more are available for research purposes. Additionally, we provide detailed information about our synthetic data generation pipeline, enabling others to generate their own datasets. See the project page: https://bedlam.is.tue.mpg.de/.

1. Introduction

The estimation of 3D human pose and shape (HPS) from images has progressed rapidly since the introduction of HMR [32], which uses a neural network to regress SMPL [45] pose and shape parameters from an image. A steady stream of new methods have improved the accuracy of the estimated 3D bodies [21, 33, 35, 38, 41, 73, 93]. The progress, however, entangles two things: improvements to the architecture and improvements to the training data. This makes it difficult to know which matters most. To answer

*The authors contributed equally and are listed alphabetically.
†This work was performed when JY was at MPI-IS.

this, we need a dataset with real ground truth 3D bodies and not simply 2D joint locations or pseudo ground truth. To that end, we introduce a new, realistic, synthetic dataset called BEDLAM (Bodies Exhibiting Detailed Lifelike Animated Motion) and use it to analyze the current state of the art (SOTA). Fig. 1 shows example images from BEDLAM along with the ground-truth SMPL-X [57] bodies.

Theoretically, synthetic data has many benefits. The ground truth is "perfect" by construction, compared with existing image datasets. We can ensure diversity of the training data across skin tones, body shapes, ages, etc., so that HPS methods are inclusive. The data can also be easily repurposed to new cameras, scenes, and sensors. Consequently, there have been many attempts to create synthetic datasets to train HPS methods. While prior work has shown synthetic data is useful, it has not been sufficient so far. This is likely due to the lack of realism and diversity in existing synthetic datasets.

In contrast, BEDLAM provides the realism necessary to test whether "synthetic data is all you need". Using BEDLAM, we evaluate different network architectures, backbones, and training data and find that *training only using synthetic data* produces methods that generalize to real image benchmarks, obtaining SOTA accuracy on both 3D human pose and 3D body shape estimation. Surprisingly, we find that even basic methods like HMR [32] achieve SOTA performance on real images when trained on BEDLAM.

Dataset. BEDLAM contains monocular RGB videos together with ground truth 3D bodies in SMPL-X format. To create diverse data, we use 271 body shapes (109 men and 162 women), with 100 skin textures from Meshcapade [3] covering a wide range of skin tones. In contrast to previous work, we add 27 different types of hair (Reallusion [1]) to the head of SMPL-X. To dress the body, we hired a professional 3D clothing designer to make 111 outfits, which we drape and simulate on the body using CLO3D [2]. We also texture the clothing using 1691 artist-designed textures. The bodies are animated using 2311 motions sampled from AMASS [47]. Because AMASS does not include hand motions, we replace the static hands with hand motions sampled from the GRAB dataset [74]. We render single people as well as groups of people (varying from 3-10) moving in a variety of 3D scenes (8) and HDRI panoramas (95). We use a simple method to place multiple people in the scenes so that they do not collide and use simulated camera motions with various focal lengths. The synthetic image sequences are rendered using Unreal Engine 5 [5] at 30 fps with motion blur. In total, BEDLAM contains around 380K unique image frames with 1-10 people per image, for a total of 1M unique bounding boxes with people.

We divide BEDLAM into training, validation, and test sets with 75%, 20% and 5% of the total bounding boxes respectively. While we make all the image data available, we withhold the SMPL-X ground truth from the test set and provide an automated evaluation server. For the training and validation sets, we provide all the SMPL-X animations, the 3D clothing, skin textures, and all freely available assets. Where we have used commercial assets, we provide information about how to obtain the data and replicate our results. We also provide the details necessary for researchers to create their own data.

Evaluation. With sufficient high-quality training data, fairly simple neural-network architectures often produce SOTA results on many vision tasks. Is this true for HPS regression? To tackle this question, we train two different baseline methods (HMR [32] and CLIFF [38]) on varying amounts of data and with different backbones; HMR represents the most basic method and CLIFF the recent SOTA. Since BEDLAM provides paired images with SMPL-X parameters, we train methods to directly regress these parameters; this simplifies the training compared with methods that use 2D training data. We evaluate on natural-image datasets including 3DPW [79] and RICH [26], a laboratory dataset (Human3.6M [27]), as well as two datasets that evaluate body shape accuracy (SSP-3D [66] and HBW [16]).

Surprisingly, despite its age, we find that training HMR on synthetic data produces results on 3DPW that are better than many recently published results and are close to CLIFF. We find that the backbone has a large impact on accuracy, and pre-training on COCO is significantly better than pre-training on ImageNet or from scratch. We perform a large number of experiments in which we train with just synthetic data, just real data, or synthetic data followed by fine tuning on real data. We find that there is a significant benefit to training on synthetic data over real data and that fine tuning with real data offers only a small benefit.

A key property of BEDLAM is that it contains realistically dressed people with ground truth body shape. Consequently, we compare the performance of methods trained on BEDLAM with two SOTA methods for body shape regression: SHAPY [16] and Sengupta et al. [67] using both the HBW and SSP-3D datasets. CLIFF trained with BEDLAM does well on both datasets, achieving the best overall of all methods tested. This illustrates how methods trained on BEDLAM generalize across tasks and datasets.

Summary. We propose a large synthetic dataset of realistic moving 3D humans. We show that training on synthetic dataset alone, even with a basic network architecture, produces accurate 3D human pose and shape estimates on real data. BEDLAM enables us to perform an extensive meta-ablation study that illuminates which design decisions are most important. While we focus on HPS, the dataset has many other uses in learning 3D clothing models and action recognition. BEDLAM is available for research purposes together with an evaluation server and the assets needed to generate new datasets.

2. Related work

There are four main types of data used to train HPS regressors: (1) Real images from constrained scenarios with high-quality ground truth (lab environments with motion capture). (2) Real images in-the-wild with 2D ground truth (2D keypoints, silhouettes, etc.). (3) Real images in-the-wild with 3D pseudo ground truth (estimated from 2D or using additional sensors). (4) Synthetic images with perfect ground truth. Each of these has played an important role in advancing the field to its current state. The ideal training data would have perfect ground truth 3D human shape and pose information together with fully realistic and highly diverse imagery. None of the above fully satisfy this goal. We briefly review 1-3 while focusing our analysis on 4.

Real Images. Real images are diverse, complex, and plentiful. Most methods that use them for training rely on 2D keypoints, which are easy to manually label at scale [7, 28, 42, 48]. Such data relies on human annotators who may not be consistent, and only provides 2D constraints on human pose with no information about 3D body shape. In controlled environments, multiple cameras and motion capture equipment provide accurate ground truth [10,13,14,24,26,27,31,37,53,69,77,79,88,94]. In general, the cost and complexity of such captures limits the number of subjects, the variety of clothing, the types of motion, and the number of scenes.

Several methods fit 3D body models to images to get pseudo ground truth SMPL parameters [30, 35, 51]. Networks trained on such data inherit any biases of the methods used to compute the ground truth; e.g. a tendency to estimate bent knees, resulting from a biased pose prior. Synthetic data does not suffer such biases.

Most image datasets are designed for 3D pose estimation and only a few have addressed body shape. SSP-3D [66] contains 311 in-the-wild images of 62 people wearing tight sports clothing with pseudo ground truth body shape. Human Bodies in the Wild (HBW) [16] uses 3D body scans of 35 subjects who are also photographed in the wild with varied clothing. HBW includes 2543 photos with "perfect" ground truth shape. Neither dataset is sufficiently large to train a general body shape regressor.

In summary, real data for training HPS involves a fundamental trade off. One can either have diverse and natural images with low-quality ground truth or limited variability with high-quality ground truth.

Synthetic. Synthetic data promises to address the limitations of real imagery and there have been many previous attempts. While prior work has shown synthetic data to be useful (e.g. for pre-training), no prior work has shown it to be sufficient without additional real training data. We hypothesize that this is due to the fact that prior datasets have either been too small or not sufficiently realistic. To date, no state-of-the-art method is trained from synthetic data alone.

Recently, Microsoft has shown that a synthetic dataset of faces is sufficiently accurate to train high-quality 2D feature detection [81]. While promising, human bodies are more complex. AGORA [56] provides realistic images of clothed bodies from *static* commercial scans with SMPL-X ground truth. SPEC [34] extends AGORA to more varied camera views. These datasets have limited avatar variation (e.g. few obese bodies) and lack motion.

Synthetic from real. Since creating realistic people using graphics is challenging, several methods *capture* real people and then render them synthetically in new scenes [22,49,50]. For example, MPI-INF-3DHP [49] captures 3D people, augments their body shape, and swaps out clothing before compositing the people on images. Like real data, these capture approaches are limited in size and variety. Another direction takes real images of people plus information about body pose and, using machine learning methods, synthesizes new images that look natural [63, 90]. This is a promising direction but, to date, no work has shown that this is sufficient train HPS regressors.

Synthetic data without clothing. Synthesizing images of 3D humans on image backgrounds has a long history [70]. We focus on more recent datasets for training HPS regressors for parametric 3D human body models like SCAPE [8] (e.g. Deep3DPose [15]) and SMPL [45] (e.g. SURREAL [78]). Both apply crude textures to the naked body and then render the bodies against random image backgrounds. In [15, 25], the authors use domain adaptation methods to reduce the domain gap between synthetic and real images. In [78] the authors use synthetic data largely for pre-training, requiring fine tuning on real images.

Since realistic clothes and textures are hard to generate, several methods render SMPL silhouettes or part segments and then learn to regress HPS from these [58,64,85]. While one can generate an infinite amount of such data, these methods rely on a separate process to compute silhouettes from images, which can be error prone. For example, STRAPS [66] uses synthetic data to regress body shape from silhouettes.

Synthetic data with rigged clothing. Another approach renders commercial, rigged, body models for which the clothing deformations are not realistic. For example PSP-HDRI+ [19], 3DPeople [59], and JTA [20] use rigged characters but provide only 3D skeletons so they cannot be used for body shape estimation. The Human3.6M dataset [27] includes mixed-reality data with rigged characters inserted into real videos. There are only 5 sequences, 7.5K frames, and a limited number of rigged models, making it too small for training. Multi-Garment Net (MGN) [12] constructs a wardrobe from rigged 3D scans but renders them on images with no background. Synthetic data has also been used to estimate ego-motion from head-mounted cameras

Figure 2. **Dataset construction.** Illustration of each step in the process, shown for a single character. Left to right: (a) sampled body shape. (b) skin texture. (c) clothing simulation. (d) cloth texture. (e) hair. (f) pose. (g) scene and illumination. (h) motion blur.

Figure 3. Skin tone diversity. Example body textures from 50 male and 50 female textures, covering a wide range of skin tones.

[6, 76, 84]. HSPACE [9] uses 100 rigged people with 100 motions and 100 3D scenes. To get more variety, they fit GHUM [83] to the scans and reshape them. They train an HPS method [91] on the data and note that "models trained on synthetic data alone do not perform the best, not even when tested on synthetic data." This statement is consistent with the findings of other methods and points to the need for increased diversity to achieve generalization.

Simulated clothing with images. Physics-based cloth simulation provides greater realism than rigged clothing and allows us to dress a wide range of bodies in varied clothing with full control. The problem, however, is that physics simulation is challenging and this limits the size and complexity of previous datasets. Liang and Lin [39] and Liu et al. [44] simulate 3D clothing draped on SMPL bodies. They render the people on image backgrounds with limited visual realism. BCNet [29] uses both physics simulation and rigged avatars but the dataset is aimed at 3D clothing modeling more than HPS regression. Other methods use a very limited number of garments or body shapes [17, 80].

Simulated clothing without images. Several methods drape clothing on the 3D body to create datasets for learning 3D clothing deformations [11, 23, 55, 65, 75]. These datasets are limited in size and do not contain rendered images.

Summary. The prior work is limited in one or more of these properties: body shapes, textures, poses, motions, backgrounds, clothing types, physical realism, cameras, etc. As a result, these datasets are not sufficient for training HPS methods that work on real images.

3. Dataset

Each step in the process of creating BEDLAM is explained below and illustrated in Fig. 2. Rendering is performed using Unreal Engine 5 (UE5) [5]. Additionally, the Sup. Mat. provides details about the process and all the 3D assets. The Supplemental Video shows example sequences.

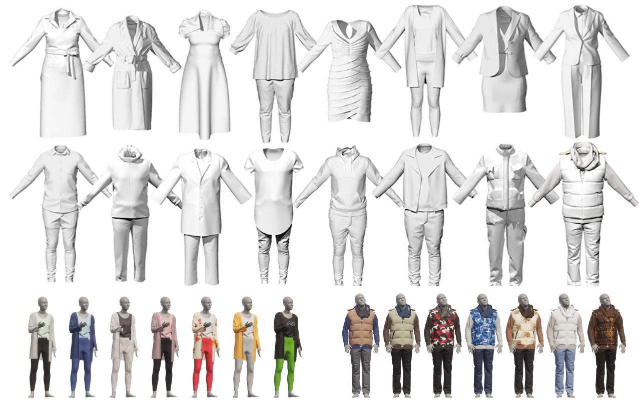

Figure 4. Diversity of clothing and texture. Top: samples from BEDLAM's 111 outfits with real-world complexity. Bottom: each outfit has several clothing textures. Total: 1691.

3.1. Dataset Creation

Body shapes. We want a diversity of body shapes, from slim to obese. We get 111 adult bodies in SMPL-X format from AGORA dataset. These bodies mostly correspond to models with low BMI. To increase diversity, we sample an additional 80 male and 80 female bodies with $\mathrm{BMI} > 30$ from the CAESAR dataset [62]. Thus we sample body shapes from a diverse pool of 271 body shapes in total. The ground truth body shapes are represented with 11 shape components in the SMPL-X gender-neutral shape space. See Sup. Mat. for more details about the body shapes.

Skin tone diversity. HPS estimation will be used in a wide range of applications, thus it is important that HPS solutions be inclusive. Existing HPS datasets have not been designed to ensure diversity and this is a key advantage of synthetic data. Specifically, we use 50 female and 50 male commercial skin albedo textures from Meshcapade [3] with minimal clothing and a resolution of 4096x4096. These artist-created textures represent a total of seven ethnic groups (African, Asian, Hispanic, Indian, Mideast, South East Asian and White) with multiple variations within each. A few examples are shown in Fig. 3.

3D Clothing and textures. A key limitation of previous synthetic datasets is the lack of diverse and complex 3D clothing with realistic physics simulation of the clothing in motion. To address this, we hired a 3D clothing designer to create 111 unique real-world outfits, including but not

Figure 5. Clothing as texture maps for high-BMI bodies. Left: example simulated clothing. Right: clothing texture mapped on bodies with BMIs of 30, 40, and 50.

Figure 6. 10 examples of BEDLAM's 27 hairstyles.

limited to T-shirts, shirts, jeans, tank tops, sweaters, coats, duvet jackets, suits, gowns, bathrobes, vests, shorts, pants, and skirts. Unlike existing synthetic clothing datasets, our clothing designs have complex and realistic structure and details such as pleats, pockets, and buttons. Example outfits are shown in Fig. 4. We use commercial simulation software from CLO3D [2] to obtain realistic clothing deformations with various body motions for the bodies from the AGORA dataset (see Supplemental Video). This 3D dataset is a unique resource that we will make available to support a wide range of research on learning models of 3D clothing.

Diversity of clothing appearance is also important. For each outfit we design 5 to 27 clothing textures with different colors and patterns. In total we have 1691 unique clothing textures. Examples are shown in Fig. 4.

For high-BMI bodies, physics simulation of clothing fails frequently due to the difficulty of garment auto-resizing and interpenetration between body parts. For such situations, we use clothing texture maps that look like clothing "painted" on the body. Specifically, we auto-transfer the textures of 1738 simulated garments onto the body UV-map using Blender. We then render high-BMI body shapes using these textures (see Fig. 5).

Hair. We use the Character Creator (CC) software from Reallusion [1] and purchased hairstyles to generate 27 hairstyles (Fig. 6). We auto-align our SMPL-X female and male template mesh to the CC template mesh and then transfer the SMPL-X deformations to it. We then apply the hairstyles in the CC software to match our custom head-shapes. We export the data to Blender to automatically process the hair mesh vertices so that their world vertex positions are relative to the head node positioned at the origin. Note that vendor-provided plugins take care of the extensive shader setup needed for proper rendering of these hair-card-based meshes. Finally the "virtual toupees" are imported into Unreal Engine where they are attached to the head nodes of the target SMPL-X animation sequences. The world-pose of each toupee is then automatically driven by the Unreal Engine animation system.

Human motions. We sample human motions from the AMASS dataset [47]. Due to the long-tail distribution of motions in the dataset, a naive random sampling leads to a strong bias towards a small number of frequent motions, resulting in low motion diversity. To avoid this, we make use of the motion labels provided by BABEL [60]. Specifically, we sample different numbers of motion sequences for each motion category according to their motion diversity (see Sup. Mat. for details). This leads to 2311 unique motions. Each motion sequence lasts from 4 to 8 seconds. Naively transferring these motions to new body shapes in the format of joint angle sequences may lead to self-interpenetration, especially for high-BMI bodies. To avoid this, we follow the approach in TUCH [52] to resolve collisions among body parts for all the high-BMI bodies. While the released dataset is rendered at 30fps, we only use every 5^{th} frame for training and evaluation to reduce pose redundancy. The full sequences will be useful for research on 3D human tracking, e.g. [61, 72, 86, 89].

Unfortunately, most motion sequences in AMASS contain no hand motion. To increase realism, diversity, and enable research on hand pose estimation, we add hand motions sampled from the GRAB [74] dataset. While these hand motions do not semantically "match" the body motion, the rendered sequences still look realistic, and are sufficient for training full-body and hand regressors.

Scenes and lighting. We represent the environment either through 95 panoramic HDRI images [4] or through 8 3D scenes. We manually select HDRI panoramas that enable the plausible placement of animated bodies on a flat ground plane up to a distance of 10m. We randomize the viewpoint into the scenes and use the HDRI images for image-based lighting. For the 3D scenes we focus on indoor environments since the HDRI images already cover outdoor environments well. To light the 3D scenes, we either use Lightmass precalculated global illumination or the new Lumen real-time global illumination system introduced in UE5 [5].

Multiple people in the scene. For each sequence we randomly select between 1 and 10 subjects. For each subject a random animation sequence is selected. We leverage binary ground occupancy maps and randomly place the moving people into the scene such that they do not collide with each other or scene objects. See Sup. Mat. for details.

Cameras. For BEDLAM, we focus on cameras that one naturally encounters in common computer vision datasets. For most sequences we use a static camera with randomized camera extrinsics. The extrinsics correspond to typical ground-level hand-held cameras in portrait and landscape mode. Some sequences use additional extrinsics augmentation by simulating a cinematic orbit camera shot. Camera intrinsics are either fixed at HFOV of 52 and 65 or zoom in from 65 to 25 HFOV.

Rendering. We render the image sequences using the UE5 game engine rasterizer with the cinematic camera model simulating a 16:9 DSLR camera with a 36x20.25mm sensor size. The built-in movie render subsystem (Movie Render Queue) is used for deterministic and high-quality image sequence generation. We simulate motion blur caused by the default camera shutter speed by generating 7 temporal image samples for each final output image. A single Windows 11 PC using one NVIDIA RTX3090 GPU was used to render all color images and store them as 1280x720 lossless compressed PNG files with motion blur at an average rate of more than 5 images/s.

Depth maps and segmentation. While our focus is on HPS regression, BEDLAM can support other uses. Since the data is synthetic, we also render out depth maps and segmentation masks with semantic labels (hair, clothing, skin). These are all available as part of the dataset release. See Sup. Mat. for details.

3.2. Dataset Statistics

In summary, BEDLAM is generated from a combination of 271 bodies, 27 hairstyles, 111 types of clothing, with 1691 clothing textures, 2311 human motions, in 95 HDRI scenes and 8 3D scenes, with on average 1-10 person per scene, and a variety of camera poses. See Sup. Mat. for detailed statistics. This results in 10K motion clips, from which we use 380K RGB frames in total. We compute the size of the dataset in terms of the number of unique bounding boxes containing individual people. BEDLAM contains 1M such bounding boxes, which we divide into sets of about 750K, 200K, and 50K examples for training, validation, and test, respectively. See Sup. Mat. for a detailed comparison of BEDLAM's size and diversity relative to existing real and synthetic datasets.

4. Experiments

4.1. Implementation Details

We train both HMR and CLIFF on the synthetic data (BEDLAM+AGORA) using an HRNet-W48 [71] backbone and refer to these as BEDLAM-HMR and BEDLAM-CLIFF respectively. We conduct different experiments with the weights of the backbone initialized from scratch, using ImageNet [18], or using a pose estimation network trained on COCO [82]. We represent all ground truth bodies in a gender neutral shape space to supervise training; we do not use gender labels. We remove the adversary from HMR and set the ground truth hand poses to neutral when training BEDLAM-HMR and BEDLAM-CLIFF. We apply a variety of data augmentations during training. We experiment with a variety of losses; the final loss is a combination of MSE loss on model parameters, projected keypoints, 3D joints, and an L1 loss on 3D vertices.

We re-implement CLIFF (called CLIFF†) and train it on only real image data using the same settings as BEDLAM-CLIFF. Following [38], we train CLIFF† using Human3.6M [27], MPI-INF-3DHP [49], and 2D datasets COCO [43] and MPII [7] with pseudo-GT provided by the CLIFF annotator. Table 1 shows that, when trained on real images, and fine-tuned on 3DPW training data, CLIFF† matches the accuracy reported in [38] on 3DPW and is even more accurate on RICH. Thus our implementation can be used as a reference.

We also train a full body network, BEDLAM-CLIFF-X, to regress body and hand poses. To train the hand network, we create a dataset of hand crops from BEDLAM training images using the ground truth hand keypoints. Since hands are occluded by the body in many images, MediaPipe [46] is used to detect the hand in the crop. Only the crops where the hand is detected with a confidence greater than 0.8 are used in the training. For details see Sup. Mat.

4.2. Datasets and Evaluation Metrics

Datasets. For training we use around 750K crops from BEDLAM and 85K crops from AGORA [56]. We also finetune BEDLAM-CLIFF and BEDLAM-HMR on 3DPW training data; these are called BEDLAM-CLIFF* and BEDLAM-HMR*. To do so, we convert the 3DPW [79] GT labels in SMPL-X format. We use 3DPW for evaluation but, since it has limited camera variation, we also use RICH [26] which has more varied camera angles. Both 3DPW and RICH have limited body shape variation, hence to evaluate body shape we use SSP-3D [66] and HBW [16]. In Sup. Mat. we also evaluate on Human3.6M [27] and observe that, without fine-tuning on the dataset, training on BEDLAM produces more accurate results than training using real images; that is, BEDLAM generalizes better to the lab data. To evaluate the output from BEDLAM-CLIFF-X, we use the AGORA and BEDLAM test sets.

Evaluation metrics. We use standard metrics to evaluate body pose and shape accuracy. PVE and MPJPE represent the average error in vertices and joints positions, respectively, after aligning the pelvis. PA-MPJPE further aligns the rotation and scale before computing distance. PVE-T-SC is per-vertex error in a neutral pose (T-pose) after scale-correction [66]. $\text{P2P}_{20\text{k}}$ is per-vertex error in a neutral pose, computed by evenly sampling 20K points on SMPL-X's surface [16]. All errors are in mm.

For evaluation on 3DPW and SSP-3D, we convert our predicted SMPL-X meshes to SMPL format by using a vertex mapping $D \in \mathbb{R}^{10475 \times 6890}$ [57]. The RICH dataset has ground truth in SMPL-X format but hand poses are less reliable than body pose due to noise in multi-view fitting.

Figure 7. Example BEDLAM-CLIFF results from all test datasets. Left to right: SSP-3D $\times$ 2, HBW $\times$ 3, RICH, 3DPW.

Method	3DPW (14)			RICH (24)		
	PA-MPJPE	MPJPE	PVE	PA-MPJPE	MPJPE	PVE
PARE* [33]	46.5	74.5	88.6	60.7	109.2	123.5
METRO* [40]	47.9	77.1	88.2	64.8	114.3	128.9
CLIFF* [38]	43.0	69.0	81.2	56.6	102.6	115.0
CLIFF†*	43.6	68.8	82.1	55.7	91.6	104.4
BEDLAM-HMR*	43.3	71.8	83.6	50.9	88.2	101.8
BEDLAM-CLIFF*	**43.0**	**66.9**	**78.5**	**50.2**	**84.4**	**95.6**
HMR [32]	76.7	130	N/A	90.0	158.3	186.0
SPIN [35]	59.2	96.9	116.4	69.7	122.9	144.2
SPEC [34]	53.2	96.5	118.5	72.5	127.5	146.5
PARE [33]	50.9	82.0	97.9	64.9	104.0	119.7
HybrIK [36]	48.8	80	94.5	56.4	96.8	110.4
Pang et. al. [54]	47.3	81.9	96.5	63.7	117.6	136.5
CLIFF†	**46.4**	73.9	87.6	55.7	90.0	102.0
BEDLAM-HMR	47.6	79.0	93.1	53.2	91.4	106.0
BEDLAM-CLIFF	46.6	**72.0**	**85.0**	**51.2**	**84.5**	**96.6**

Table 1. Reconstruction error on 3DPW and RICH. *Trained with 3DPW training set. †Trained on real images with same setting as BEDLAM-CLIFF. Parenthesis: (#joints).

Hence, we use it only for evaluating body pose and shape. We convert the ground truth SMPL-X vertices to SMPL format using D after setting the hand and face pose to neutral. To compute joint errors, we use 24 joints computed from these vertices using the SMPL joint regressor. For evaluation on AGORA-test and BEDLAM-test, we use a similar evaluation protocol as described in [56].

4.3. Comparison with the State-of-the-Art

Table 1 summarizes the key results. (1) Pre-training on BEDLAM and fine-tuning with a mix of 3DPW and BEDLAM training data gives the most accurate results on 3DPW and RICH (i.e. BEDLAM-CLIFF* is more accurate than CLIFF†* or [38]). (2) Using the same training, makes HMR (i.e. BEDLAM-HMR*) nearly as accurate on 3DPW and more accurate than CLIFF†* on RICH. This suggests that even simple methods can do well if trained on good data. (3) BEDLAM-CLIFF, with no 3DPW fine-tuning, does nearly as well as the fine-tuned version and generalizes better to RICH than CLIFF with, or without, 3DPW fine-tuning. (4) Both CLIFF and HMR trained only on synthetic data outperform the recent methods in the field. This suggests that more effort should be put into obtaining high-quality data. See Sup. Mat. for SMPL-X results.

Table 2 shows that BEDLAM-CLIFF has learned to estimate body body shape under clothing. While SHAPY [92] performs best on HBW and Sengputa et al. [67] performs best on SSP-3D, both of them perform poorly on the other dataset. Despite not seeing either of the training datasets, BEDLAM-CLIFF ranks 2nd on SSP-3D and HBW. BEDLAM-CLIFF has the best rank averaged across the datasets, showing its generalization ability.

Qualitative results on all these benchmarks are shown in Fig. 7. Note that, although we do not assign gender labels to any of the training data, we find that, on test data, methods trained on BEDLAM predict appropriately gendered body shapes. That is, they have automatically learned the association between image features and gendered body shape.

4.4. Ablation Studies

Table 3 shows the effect of varying datasets, backbone weights and percentage of data; see Sup. Mat. for the full table with results for HMR. We train with synthetic data only and measure the performance on 3DPW. Note that the backbones are pre-trained on image data, which is standard practice. Training them from scratch on BEDLAM gives worse results. It is sufficient to train using simple 2D task for which there is plentiful data. Similar to [54], we find that training the backbone on a 2D pose estimation task (COCO) is important. We also vary the percentage of BEDLAM crops used in training. Interestingly, we find that uniformly sampling just 5% of the crops from BEDLAM produces reasonable performance on 3DPW. Performance monotonically improves as we add more training data. Note that 5% of BEDLAM, i.e. 38K crops, produces better results than 85K crops from AGORA, suggesting that BEDLAM is more diverse. Still, these synthetic datasets are complementary, with our best results coming from a combination of the two. We also found that realistic clothing simulation leads to significantly better results than training with textured bodies. This effect is more pronounced when using a backbone pre-trained on ImageNet rather than COCO. See Sup. Mat. for details.

Method	Model	SSP-3D		HBW		Average
		PVE-T-SC	Rank	$P2P_{20k}$	Rank	Rank
HMR [32]	SMPL	22.9	8	-	-	-
SPIN [35]	SMPL	22.2	7	29	4	5.5
SHAPY [16]	SMPL-X	19.2	6	21	1	3.5
STRAPS [66]	SMPL	15.9	4	47	6	5
Sengupta et al. [68]	SMPL	15.2	3	-	-	-
Sengupta et al. [67]	SMPL	13.6	1	32	5	3
CLIFF†	SMPL	18.4	5	27	3	4
BEDLAM-CLIFF	SMPL-X	14.2	2	22	2	2

Table 2. Per-vertex 3D body shape error on the SSP-3D and HBW test set in T-pose (T). SC refers to scale correction.

Method	Dataset	Backbone	Crops %	PA-MPJPE	MPJPE	PVE
CLIFF	B+A	scratch	100	61.8	97.8	115.9
CLIFF	B+A	ImageNet	100	51.8	82.1	96.9
CLIFF	B+A	COCO	100	47.4	73.0	86.6
CLIFF	B	COCO	5	54.0	80.8	96.8
CLIFF	B	COCO	10	53.8	79.9	95.7
CLIFF	B	COCO	25	52.2	77.7	93.6
CLIFF	B	COCO	50	51.0	76.3	91.1
CLIFF	A	COCO	100	54.0	88.0	101.8
CLIFF	B	COCO	100	50.5	76.1	90.6

Table 3. Ablation experiments on 3DPW. B denotes BEDLAM and A denotes AGORA. Crop %'s only apply to BEDLAM.

5. Limitations and Future Work

Our work demonstrates that synthetic human data can stand in for real image data. By providing tools to enable researchers to create their own data, we hope the community will create new and better synthetic datasets. To support that effort, below we provide a rather lengthy discussion of limitations and steps for improvement; more in Sup. Mat.

Open source assets. There are many high-quality commercial assets that we did not use in this project because their licences restrict their use in neural network training. This is a significant impediment to research progress. More open-source assets are needed.

Motion and scenes. The human motions we use are randomly sampled from AMASS. In real life, clothing and motions are correlated, as are scenes and motions. Additionally, people interact with each other and with objects in the world. Methods are needed to automatically synthesize such interactions realistically [87]. Also, the current dataset has relatively few sitting, lying, and complex sports poses, which are problematic for cloth simulation.

Hair. BEDLAM lacks hair physics, long hairstyles, and hair color diversity. Our solution, based on hair cards, is not fully realistic and suffers from artifacts under certain lighting conditions. A strand-based hair groom solution would allow long flowing hair with hair-body interaction and proper rendering with diverse lighting.

Body shape diversity. Our distribution of body shapes is not uniform (see Sup. Mat.). Future work should use a more even distribution and add children and people with diverse body types (scoliosis, amputees, etc.). Note that draping high-BMI models in clothing is challenging because the mesh self-intersects, causing failures of the cloth simulation. Retargeting AMASS motions to high-BMI subjects is also problematic. We describe solutions in Sup. Mat.

More realistic body textures. Our skin textures are diverse but lack details and realistic reflectance properties. Finding high-quality textures with appropriate licences, however, is difficult.

Shoes. BEDLAM bodies are barefoot. Adding basic shoes is fairly straightforward but the general problem is actually complex because shoes, such as high heels, change body posture and gait. Dealing with high heels requires retargeting, inverse kinematics, or new motion capture.

Hands and Faces. There is very little mocap data with the full body and hands and even less with hands interacting with objects. Here we ignored facial motion; there are currently no datasets that evaluate full body and facial motion.

6. Discussion and Conclusions

Based on our experiments we can now try to answer the question "Is synthetic data all you need?" Our results suggest that BEDLAM is sufficiently realistic that methods trained on it generalize to real scenes that vary significantly (SSP-3D, HBW, 3DPW, and RICH). If BEDLAM does not well represent a particular real-image domain (e.g. surveillance-camera footage), then one can re-purpose the data by changing camera views, imaging model, motions, etc. Synthetic data will only get more realistic, closing the domain gap further. Then, does architecture matter? The fact that BEDLAM-HMR outperforms many recent, more sophisticated, methods argues that it may be less important than commonly thought.

There is one caveat to the above, however. We find that HPS accuracy depends on backbone pre-training. Pretraining the backbone for 2D pose estimation on COCO exposes it to all the variability of real images and seems to help it generalize. We expect that pre-training will eventually be unnecessary as synthetic data improves in realism.

We believe that there is much more research that BEDLAM can support. None of the methods tested here estimate humans in *world coordinates* [72, 86]. The best methods also do not exploit temporal information or action semantics. BEDLAM can support new methods that push these directions. BEDLAM can also be used to model 3D clothing and learn 3D avatars using implicit shape methods.

Acknowledgments. We thank STUDIO LUPAS GbR for creating the 3D clothing, Meshcapade GmbH for the skin textures, Lea Müller for help removing self-intersections in high-BMI bodies and Timo Bolkart for aligning SMPL-X to the CC template mesh. We thank T. Alexiadis, L. Sánchez, C. Mendoza, M. Ekinci and Y. Fincan for help with clothing texture generation.

Disclosure: https://files.is.tue.mpg.de/black/CoI_CVPR_2023.txt

References

[1] Character Creator (CC), Reallusion. https://www.reallusion.com/character-creator, 2022. 2, 5

[2] CLO. https://www.clo3d.com, 2022. 2, 5

[3] Meshcapade GmbH, Tübingen, Germany. https://meshcapade.com, 2022. 2, 4

[4] Poly Haven. https://polyhaven.com/hdris, 2022. 5

[5] Unreal Engine 5. https://www.unrealengine.com, 2022. 2, 4, 5

[6] Hiroyasu Akada, Jian Wang, Soshi Shimada, Masaki Takahashi, Christian Theobalt, and Vladislav Golyanik. UnrealEgo: A new dataset for robust egocentric 3D human motion capture. In *European Conference on Computer Vision (ECCV)*, 2022. 4

[7] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014. 3, 6

[8] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape completion and animation of people. *Transactions on Graphics (TOG)*, 24(3):408–416, 2005. 3

[9] Eduard Gabriel Bazavan, Andrei Zanfir, Mihai Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. HSPACE: Synthetic parametric humans animated in complex environments. *arXiv*, 2112.12867, 2021. 4

[10] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The IKEA ASM dataset: Understanding people assembling furniture through actions, objects and pose. In *Winter Conference on Applications of Computer Vision (WACV)*, 2021. 3

[11] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. CLOTH3D: Clothed 3D humans. In *European Conf. on Computer Vision (ECCV)*, pages 344–359. Springer International Publishing, 2020. 4

[12] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-Garment Net: Learning to dress 3D people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 3

[13] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 3

[14] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yangmin Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-modal 4D human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, 2022. 3

[15] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3D pose estimation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 479–488. IEEE, 2016. 3

[16] Vasileios Choutas, Lea Müller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. Accurate 3D body shape regression using metric and semantic attributes. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2718–2728, June 2022. 2, 3, 6, 8

[17] R. Daněček, E. Dibra, C. Öztireli, R. Ziegler, and M. Gross. DeepGarment: 3D garment shape estimation from a single image. *Comput. Graph. Forum*, 36(2):269–280, may 2017. 4

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009. 6

[19] Salehe Erfanian Ebadi, Saurav Dhakad, Sanjay Vishwakarma, Chunpu Wang, You-Cyuan Jhang, Maciek Chociej, Adam Crespi, Alex Thaman, and Sujoy Ganguly. PSP-HDRI+: A synthetic dataset generator for pre-training of human-centric computer vision models. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. 3

[20] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[21] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, pages 792–804, 2021. 1

[22] Valentin Gabeur, Jean-Sebastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3D human shape estimation from single images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2232–2241, 2019. 3

[23] Peng Guan, Loretta Reiss, David Hirshberg, Alex Weiss, and Michael J. Black. DRAPE: DRessing Any PErson. *ACM Trans. on Graphics (Proc. SIGGRAPH)*, 31(4):35:1–35:10, July 2012. 4

[24] 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, Oct. 2019. 3

[25] David T. Hoffmann, Dimitrios Tzionas, Michael J. Black, and Siyu Tang. Learning to train with synthetic humans. In *German Conference on Pattern Recognition (GCPR)*, pages 609–623, 2019. 3

[26] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 3, 6

[27] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2013. 2, 3, 6

[28] Umar Iqbal, Anton Milan, and Juergen Gall. PoseTrack: Joint multi-person pose estimation and tracking. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4654–4663, 2017. 3

[29] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. BCNet: Learning body and cloth shape from a single image. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XX*, pages 18–35, 2020. 4

[30] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *International Conference on 3D Vision (3DV)*, pages 42–52, 2020. 3

[31] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic Studio: A massively multiview system for social interaction capture. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(1):190–204, 2019. 3

[32] 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. 1, 2, 7, 8

[33] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 1, 7

[34] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *Proceedings International Conference on Computer Vision (ICCV)*, pages 11035–11045. IEEE, Oct. 2021. 3, 7

[35] 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. 1, 3, 7, 8

[36] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3383–3393, 2021. 7

[37] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. AI choreographer: Music conditioned 3D dance generation with AIST++. In *Int. Conf. Comput. Vis.*, 2021. 3

[38] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, 2022. 1, 2, 6, 7

[39] Junbang Liang and Ming C Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4352–4362, 2019. 4

[40] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1954–1963. Computer Vision Foundation / IEEE, 2021. 7

[41] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12939–12948, 2021. 1

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, volume 8693, pages 740–755, 2014. 3

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Eur. Conf. Comput. Vis.*, 2014. 6

[44] Jian Liu, Naveed Akhtar, and Ajmal Mian. Temporally coherent full 3D mesh human pose recovery from monocular video. *arXiv preprint arXiv:1906.00161*, 2019. 4

[45] 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. 1, 3

[46] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172, 2019. 6

[47] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, pages 5442–5451, 2019. 2, 5

[48] Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. JRDB: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021. Early access. 3

[49] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 3, 6

[50] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *3DV*, 2018. 3

[51] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2307, 2022. 3

[52] 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. 5

[53] Aiden Nibali, Joshua Millward, Zhen He, and Stuart Morgan. ASPset: An outdoor sports pose video dataset with 3D keypoint annotations. *Image and Vision Computing*, 111:104196, 2021. 3

[54] Hui En Pang, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Benchmarking and analyzing 3d human pose and shape estimation beyond algorithms. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 7

[55] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 4

[56] Priyanka Patel, Chun-Hao Paul Huang, Joachim Tesch, David Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, 2021. 3, 6, 7

[57] 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. 2, 6

[58] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018. 3

[59] Albert Pumarola, Jordi Sanchez, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3DPeople: Modeling the Geometry of Dressed Humans. In *International Conference in Computer Vision (ICCV)*, 2019. 3

[60] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, June 2021. 5

[61] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3D appearance, location & pose. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 5

[62] 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. 4

[63] Grégory Rogez and Cordelia Schmid. MoCap-guided data augmentation for 3D pose estimation in the wild. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3116–3124, Red Hook, NY, USA, 2016. Curran Associates Inc. 3

[64] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3d human recovery in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5340–5348, 2019. 3

[65] Igor Santesteban, Miguel A Otaduy, and Dan Casas. SNUG: Self-Supervised Neural Dynamic Garments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[66] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3D human pose and shape estimation in the wild. In *British Machine Vision Conference (BMVC)*, 2020. 2, 3, 6, 8

[67] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3D human shape and pose estimation from images in the wild. In *International Conference on Computer Vision (ICCV)*, pages 11219–11229, 2021. 2, 7

[68] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 16094–16104, 2021. 8

[69] Leonid Sigal, Alexandru Balan, and Michael J Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, 87(1):4–27, 2010. 3

[70] Cristian Sminchisescu, Amit Kanaujia, and Dimitris Metaxas. Learning joint top-down and bottom-up processes for 3D visual inference. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1743 – 1752, 02 2006. 3

[71] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 6

[72] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. TRACE: 5D temporal regression of avatars with dynamic cameras in 3D environments. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 5, 8

[73] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13243–13252, 2022. 1

[74] 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)*, 2020. 2, 5

[75] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. SIZER: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing. In *European Conference on Computer Vision (ECCV)*. Springer, August 2020. 4

[76] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xR-EgoPose: Egocentric 3D human pose from an HMD camera. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7728–7738, 2019. 4

[77] Matt Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3D human pose estimation fusing video and inertial sensors. In *Brit. Mach. Vis. Conf.*, 2017. 3

[78] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid.

Learning from synthetic humans. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635, 2017. 3

[79] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, volume 11214, pages 614–631, 2018. 2, 3, 6

[80] Tuanfeng Y. Wang, Duygu Ceylan, Jovan Popović, and Niloy J. Mitra. Learning a shared shape space for multimodal garment design. *ACM Trans. Graph.*, 37(6), dec 2018. 4

[81] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljevic, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljkovic, Tom Cashman, and Julien Valentin. 3D face reconstruction with dense landmarks. In *European Conf. on Computer Vision (ECCV)*, 2022. 3

[82] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*, 2018. 6

[83] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 4

[84] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. $\mathrm{Mo}^2\mathrm{Cap}^2$: Real-time mobile 3D motion capture with a cap-mounted fisheye camera. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):2093–2101, 2019. 4

[85] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7760–7770, 2019. 3

[86] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 5, 8

[87] Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3D scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 8

[88] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. HUMBI: A large multiview dataset of human body expressions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 3

[89] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[90] Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human synthesis and scene compositing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12749–12756, Apr. 2020. 3

[91] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. THUNDR: Transformer-based 3D human reconstruction with markers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 4

[92] Chao Zhang, Sergi Pujades, Michael Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5484–5493, 2017. 7

[93] 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. 1

[94] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 3