FAIR datasets: *F*air *A*ugmented *I*ntegratable *R*econstruction Datasets

Anonymous CVPR submission

Paper ID *****

Abstract

While many large-scale datasets are used to train human body reconstruction (HBR) models, the demographics of these datasets are limited and thus do not wholly represent the intended end users. We conduct the first large-scale evaluation of biases in popular human image datasets used in HBR. To do so we generate a cost-efficient dataset to represent 324,000,000,000 images. *Our approach uses synthetic human image datasets generated using a* configurable human image generator*, which allows us to create large-scale data for diverse bodies under a multitude of environmental and camera settings. We provide an* automatic bias evaluator *for existing human body reconstruction methods that allows one to analyze the performance of HBR models across different bias categories such as skin tones, body sizes and shapes, lighting conditions, and camera poses. We present a method to neutralize these biases using our dataset, showing significant improvements in variance for all bias categories while maintaining performance in the original testing domains. Additionally we provide an efficient, low-cost framework to capture human data in the wild for future diverse real-world datasets.*

1. Introduction

Machine learning models learn by observing and mimicking training datasets, including biases that these datasets intentionally or unintentionally encoded. These biases include but are not limited to social prejudices, such as gender bias in natural language inferencing [**?**, **?**] and racial discrimination in facial recognition [**?**].

While some biases present in human body reconstruction (HBR) from images are partly due to the reconstruction frameworks, most are mainly a result of imbalanced representation in the training data. They can result in poorer reconstruction for some groups of people, camera poses, and environment conditions. Many HBR methods train on similar datasets, leading to trends in their biases although the variances still differ due to the nature of their algorithms and implementations.

Human body reconstruction is a promising and important topic with the potential for large-scale use in virtual try-on, online avatars, etc. While current large-scale datasets have enabled notable progress in HBR and allow for quantitative evaluation, many datasets have demographic and environment limitations that prevent us from evaluating how such a method may perform in the wild. Even the best reconstruction techniques may still fall short in some real-world scenarios. Therefore, addressing biases in reconstruction outcomes can further advance the success of this field as a whole.

In this work, we conduct the **first large-scale evaluation of biases in popular human image datasets** used in HBR. To do so we generate a **cost-efficient dataset to represent 324,000,000,000 images**. We observed biases toward specific body sizes, lighting conditions, backgrounds, and camera poses. We then studied HBR methods trained using these datasets. *We found that these reconstruction methods inherit many of these biases exemplified by the training data from the collected human body images.* We propose an **automated method to evaluate the biases in any HBR method** using our dataset and evaluation procedure, which will be made publicly available.

Ideally, when it comes to removing bias in machine learning, we aim to remove all existing biases from training data and re-train entire models using these new balanced datasets. However, in the case of human body reconstruction, such unbiased datasets do not exist, and repeating the training process can be computationally expensive. Instead, we propose **a fine-tuning approach to neutralize biases towards skin tones, body shapes and sizes, lighting conditions, and camera poses** using a small balanced dataset with additional data augmented to neutralize biases in specifically trained models. We found that our method reduces variance and biases across different groups and settings while maintaining performance in the original test domains.

With our goal of neutralizing biases, we present a **practical real-world human data collection pipeline for HBR from images**. This pipeline is highly accessible due to its ease of use and low cost. It allows one to collect both 2D and 3D groundtruth joint positions by interpolating multi-

view images. By making this pipeline publicly available, we hope to expedite future large-scale collection of diverse human data. Our code for these procedures, along with the dataset, will be published and made available publicly so that others working in HBR may evaluate and neutralize any biases.

2. Related Works

Many large-scale datasets have been used for human body reconstruction, but have limitations when it comes to representing all populations and environments. A summary of these limitations and findings can be found in Table 1. For human body reconstruction from images, the main components for the training data is composed of the image of a subject for reconstruction as well as a groundtruth format for evaluating the reconstruction results. This groundtruth format varies for different datasets, from the 2D pixel locations of the subject's joints on the images to overall true body shape and size parameters of the subject.

Additionally while work has been done to adversarially improve reconstruction results for subject data [**?**, **?**, 1, 5], not much prior work had been done to evaluate or reduce social or environmental biases from HBR. We discuss what potential biases have been considered and where we differ from existing datasets.

2.1. Datasets without 3D Groundtruth information

There are many datasets which contain human image data, as well as 2D groundtruth information. While some of these do not offer the most accurate training data for 3D reconstruction, they are still essential to training HBR methods, as they offer access to a large number of real-world annotated human images.

Some early datasets include multi-view silhouettes [**?**], MVIC [**?**], and MARCOnI [**?**, **?**]. Datasets commonly used in HBR include COCO [4], LSP and its extended version LSPET [**?**], and MPII [**?**]. Many of these datasets were created with certain focuses in mind, and are often used in combination with other HBR specific datasets in 3D groundtruth formats.

Take for example the Leeds Sports Dataset (LSP) [**?**] and its extension LSPET, the focus is on sportspersons, which does not necessarily represent the greater public. OCHuman [**?**] focuses on occluded humans using bounding-boxes and annotated human poses and instance masks.

2.2. Datasets with 3D Groundtruth information

Real-World human image datasets offer realistic inputs for training reconstruction methods, however, getting 3D groundtruth details for evaluation becomes difficult and time-consuming, limiting the number of subjects and environments from which data is collected. The size and shape of the subject has to be recorded physically, while capturing the 3D pose can be difficult as these joint positions are not easy to annotate. The alternative to annotation is motion capture, which forces data collection to occur in a lab like environment,further hindering the number of participants and scenarios. Examples of such datasets include HumanEva [**?**], Human3.6M [2], TotalCapture [**?**], MuPoTS-3D [**?**], 3DPW [**?**], MPI-INF-3DHP Test [**?**], HUMBI [**?**] and Panoptic Studio [**?**].

HumanEva [**?**], Human3.6M [2], and TotalCapture [**?**] utilize full body motion capture. While this allows for accurate 3D pose information to be collected, it requires a lab environment, and limits the clothing variety the participants were able to wear.

The remaining methods utilize marker-less motion capture where the capturing of 3D poses often relies on Inertial Measurement Unit (IMU) sensors. While this allows for more diversity in environment and clothing, the accuracy of the 3D groundtruth references for this data are less accurate than other dataset types we discuss.

Due to the the higher levels of effort required with capturing real-world datasets, aside from HUMBI [**?**] and Panoptic Studio [**?**], the other real-world datasets have under 20 participants captured and under 20 camera vantage points. Some datasets such as the 3DPW datasets, while using only a singular camera, allow for multiple vantage points of the subject(s) due to their mobile nature [**?**]. This leads to the 3DPW dataset being a popular method to validate the performance of HBR methods.

However these two datasets, HUMBI and Panoptic Studio, offer data only captured in a lab setting, removing diversity of environment.

Almost none of the real-world datasets offer in depth participant demographics information and even then the biases are clear with close to 70% of HUMBI's [**?**] 772 participants being of pale complexion and the total range of BMIs of Human3.6M [2] participants falling between the 17-29.

More recently several datasets have been created by fitting body models to existing real datasets to get new 3D "groundtruth" information for 2D human images. These datasets include EFT [**?**], STRAPS [**?**], SMPLy [**?**] and 3DOH50K [**?**].

Synthetic human image datasets using recent advances in computer graphics allows for the automatic generation of synthetic data, where the groundtruth is computer generated, therefore all details about the body sizes, shapes and poses are known. These human models can be parameterized to make their construction simpler. Some methods include the Skinned Multi-Person Linear Model (SMPL) [**?**], MakeHuman [**?**], Mixamo [**?**], and renderpeople [**?**]. The use of synthetic data in training HBR methods has proven to improve reconstruction results [**?**, 3].

Dataset	Sub. #	Body Demographics	Skintone Demographics	# Camera Poses	Multiple Environments	Groundtruth Format
EFT [?]	>1000	-	-	-	✓	SMPL
STRAPS [?]	62	-	-	-	✓	SMPL
3DOH50K [?]	-	-	-	6	1 (lab)	SMPL
SMPLy [?]	742	-	-	1*	✓	SMPL
HumanEva [?]	4	-	-	4/7	1 (lab)	3D joint positions
Human3.6M [2]	11	✓	-	14	1 (lab)	3D joint positions
TotalCapture [?]	5	-	-	8	1 (lab)	3D joint positions
MuPoTS-3D [?]	8	-	-	1*	✓	3D joint positions
3DPW [?]	7	✓	-	1*	✓	SMPL
MPI-INF-3DHP Test [?]	8	-	-	14	✓	3D joint positions
HUMBI [?]	772	-	✓	107	1 (lab)	meshes, SMPL
Panoptic Studio [?]	100	-	-	480	1 (lab)	3D joint positions
MPI-INF-3DHP Train [?]	8	-	-	14	✓	3D joint positions
AGORA [?]	>350	✓	✓	-	✓	SMPL-X, SMPL
MuCo-3DHP [?]	8	-	-	12	✓	3D joint positions
3DPeople [?]	80	✓	✓	4	✓	3D joint positions
SURREAL [7]	145	✓	-	1^t	✓	SMPL
Ours	22,500	✓	✓	2,500	✓	SMPL

Table 1. **Comparison of human image datasets used in HBR with 3D groundtruth annotations and characteristics that may cause social or environmental biases**. *: Single view video input from ambulant camera potentially offering multiple vantage points. ^{t}: camera perspective for each scene is random, but remains static throughout the scene.

Many synthetic datasets have been generated with the potential to be balanced in ways real-world data cannot be. These datasets include AGORA [?], MuCo-3DHP [?], 3DPeople [?], Human Optical Flow (MHOF) [?], SURREAL [7], renderpeople [?], and the MPI-INF-3DHP Train dataset [?]. While many of these datasets allow for images or videos to be generated of large-scale groups of persons, still very few offer clear demographics of the bodies being rendered.

Some datasets such as AGORA, 3DPeople, and SURREAL offer diverse bodies and rendering settings. However, they are limited to a smaller range of camera poses and are not easily configurable to create new perspectives. In the case of 3D people [?] or AGORA [?], the actual human models are not publicly available or are behind paywalls.

Many of the mentioned synthetic and real-world datasets aid in improving diversity in HBR training data, however, with the large variance in data between one another, it is clear there is a need for some formal mechanism to evaluate biases in human body reconstruction.

2.3. Evaluating and Mitigating Bias

While the biases in HBR have not yet been examined systematically in the past, the discussion of biases within reconstruction results has already begun [?, ?, 1, 5]. Additionally techniques which have been used to mitigate bias in other fields have the potential to be ported for human body reconstruction as well.

Recent works in HBR have proposed adversarial techniques to improve reconstruction results, however the main focus has been on human pose recovery [?, ?, ?, ?]. Some have looked to improve robustness against camera poses and body pose occlusion, but only tackle a single feature [?, ?, ?, ?, ?, ?, 6].

While not much work has been done to mitigate other biases such as skin tone, or body size in HBR, many works have explored mitigating other biases and improving robustness in datasets and machine learning (ML) models. Some techniques such apply different corruptions (or biases) on to existing datasets [?]. These types of datasets can be used to benchmark robustness of neural networks. Additionally, these sorts of datasets can be used to perform adversarial machine learning. Adversarial machine learning is a learning method by which one can improve robustness in machine learning models. One such method is via adversarial data augmentation. For example Ghosh et al. analyzes the performance of convolutional neural networks on quality degradations and introduces a method to improve the learning outcome [?]. Cubuk et al. propose a method to automatically search for improved data augmentation policies [?]. AugMix [?], Cutout [?], MixUp [?], and CutMix [?] are frameworks in which model robustness is improved using data augmentation.

More recently work like that of Shen. et. al., decouple different perturbations (or biases) and use adversarial training for improving robustness in autonomous driving [?]. Such a technique could prove effective within HBR.

While we explore methods to successfully reduce biases

in models, we want to provide a framework to prevent biases and improve diversity in future datasets.

2.4. Diverse Real-World Data Collection

Many real-world datasets are captured with constraints either environmental such as consistent lighting, fixed camera positions, and motion capturing devices, or subject, such as a few paid actors [**?**, **?**, 2].

Many of the initial real-world datasets only offered 2D joint positions, this is because 2D Pose Estimation is more straightforward and accurate than 3D, existing models like ViTPose [**?**], TransPose [**?**], HRNet [**?**], and OpenPose [**?**] can be used to detect 2D joint positions. For 3D joint positions, relative camera pose and other camera parameters are required. Well known techniques for 3D joint position calculation include COLMAP [**?**, **?**] and OpenCV [**?**]. More recently, traditional methods like COLMAP [**?**, **?**] and machine learning-based methods like NeRF [**?**] can recover 3D meshes of the intended scene.

With a combination of these techniques, it is possible to perform cost-effective and portable real-world data collection, allowing the expedited capture of large-scale diverse human data.

3. Methodology

We begin by describing the process by which we generate our diverse datasets to be used for evaluating current HBR models. We describe the bias parameters we are analysing and how we are able to represent a large set of features using a much smaller cost-efficient dataset. We analyze the biases and trends present in current methods, then present a technique to neutralize such biases. We follow our bias neutralization with a novel approach for cheap large scale human data collection so that in the future we are able to collect more diverse and balanced data.

3.1. Data Generation

We create our dataset using a large-scale synthetic human data generator composed of 3 components: a rendering framework, a human body model with a realistic pose, and the environment details, including backgrounds, lighting details, and camera poses.

For the human body model, we chose to parameterize the body using a Skinned Multi-Person Linear model (SMPL) [**?**]. By controlling and parameterizing the 82 parameters present in the SMPL model, we can generate an extensive range of body shapes, sizes, and poses [**?**].

For realistic poses, we utilized real-world human poses from large-scale datasets, such as those provided by Kolotouros et al., Mehta et al., Bogo et al., etc. [**?**, **?**, 1].

To make our dataset generation pipeline easy to configure, we created python scripts to specify the details of the images one may wish to render. Using these scripts we pass on environment details to the rendering framework.

This contains details such as skin tones, clothing, and lighting settings.

To vary the lighting settings, we included parameters to control the light source's type, position, color, and intensity. Since light reflected off the background objects affects lighting on the body's clothing and skin, we include parameters to specify backgrounds.

3.2. Evaluation Set for Body Reconstruction

For our evaluation dataset, our goal is to create a set of parameters not often described or checked for in existing datasets. By studying these prior works and common biases in trained HBR models, we ultimately chose our parameters to be **body size/shape, skin tone, lighting, and camera poses**.

From Table. 1, one can see that very few datasets offer demographic information about the human subjects included in their datasets. While this information is difficult to verify and request with large-scale real-world data, it is much simpler to parameterize in a synthetic setting.

We want to represent a large number of body sizes, from those with very little body fat or low adiposity to those with a large amount of body fat or high adiposity. We also want each body in our dataset to be unique. To analyze reconstruction results on different body sizes, we group body sizes by using the waist-to-height ratio. This ratio is a well-established measure for body fat distribution. It is considered a good global indicator as it requires no age, sex, or ethnic-specific information [**?**].

Since we are using the SMPL model to represent our synthetic humans, there are 10 shape parameters to control body size and shape. The first 3 shape parameters control the overall height and adiposity of the body. If we wanted two or more unique bodies of a similar size, we can particularize these first 3 parameters, and randomly sample the remaining 7. Using this technique we can generate multiple distinct bodies of a similar size.

Finally, we specify 9 body groups with waist-to-height ratio from 0.4 (low adiposity) to 0.75(high adiposity). Each body group contains 2,500 distinct bodies of a similar waist-to-height ratio, demonstrating diversity at every body size.

For our skin tones, we wanted to ensure that all skin tones are represented from the deepest to the palest. We looked to shade ranges from popular cosmetic companies touted for inclusivity [**?**]. We aimed to use their color palettes as our own. Although we found a large number of skin tones [**?**, **?**], we downsize them to 8 skin tones since we do not need to represent specific undertones as cosmetic companies do(cool-olive, warm-pink, neutral-gold).

Many datasets offered data from multiple lighting and environmental conditions. However, they lacked statisti-

cal details regarding these conditions within the data and a straightforward way to distinguish them. Additionally, many real-world datasets only include images collected in lab environments.

For the lighting conditions, we first picked 4 potential different lighting positions, 1 global illumination or infinite area light, labeled as *Sun*, and 3 spotlights, positions below the body, to the side of the body, and above the body, labeled *Bottom, Side,* and *Top* respectively. Then, we chose 3 light hues: *warm-tone*, *cool-tone*, and *neutral*, allowing us to simulate the most common real-world lighting settings. Finally, we selected 3 different intensities for these lights, including *low*, *mid*, and *high* intensity, forming dim to extremely bright lighting conditions.

For the environments, we included 20 different backgrounds. The colors and objects within the background are reflected onto the synthetic human at the time of rendering.

Most of the existing human image datasets have a limited number camera positions. We chose to place cameras in polar coordinates around the body to capture images from variegated viewpoints. By uniformly sampling in this system, we obtained 2,500 camera positions and use them for our dataset.

Given our extensive set of parameters, consisting of **9** different size ranges, with **2,500** distinct bodies in each range, **8** different skin tones, **4** lighting positions, **3** light hues, **3** light intensities, **20** background environments, and finally, **2,500** camera poses. These variations give us a total of ***324,000,000,000*** sample combinations for our image dataset, which will take decades for a cluster to generate. To overcome this problem, we study the impact of different features on reconstruction results and down-sample to create a balanced and reduced-size dataset, representing the critical features for each category of parameters.

3.3. Data Size Reduction

Generating a prohibitively large dataset representative of true social diversity is infeasible. Instead, we down-sample to create a much smaller and more manageable dataset of 152,500 samples, representing critical features that maximally affect reconstruction results. We plan to make this dataset publicly available so all researchers can evaluate biases in their HBR methods.

To find these critical features, we generated smaller subsets of data for each feature while altering other parameters. We keep a single parameter invariant for all other parameters while generating each subset of data. For example, with skin tones, we generate a subset of data for each of our 8 skin tones while randomly sampling from each camera pose, body size range, lighting condition, and background environment with equivalent probability. Then, we can identify the changes in performance caused by different skin tones since other features are identical for all subsets.

This approach makes our overall dataset composed of 4 subsets, one for each parameter: skin tones, lighting conditions, body sizes, and camera poses. Each parameter dataset is then composed of subsets for each distinct feature; for example, the body sizes parameter dataset comprises 9 subsets, one for each waist-to-height ratio group. With this method of downsizing, we can represent all of the crucial features using **152,500** images, which is much less than 0.00005% of the original **324,000,000,000** images. We then sample 20% of each subset for testing, as described in Sec. 3.4, while the remaining 80% for training, as described in Sec. 3.5. The exact and detailed composition of our dataset along with its statistical information in included in the supplemental materials.

Figure 1. **Examples of images from our dataset**, showing widely distributed body types/shapes, poses, skin tones, backgrounds, clothing, lighting environments, and camera poses.

Fig. 1 displays images from our synthetic dataset, demonstrating the large variety of bodies, lighting conditions, environments, and camera poses our dataset provides, and it will be publicly available.

3.4. Automatic Bias Evaluation

We use our testing dataset to evaluate the biases of existing datasets and HBR models by performing the following steps: (1) Train an HBR model using the datasets one wishes to evaluate. (2) Test each subset of image data within the broader bias parameters to find which categories perform the worst.

To evaluate the biases in existing datasets and trained models, we select and train baseline models using their orig-

inal dataset selection. We chose Human Mesh Recovery (HMR) introduced by Kanazawa et al. [?], SMPL with optimization IN the loop (SPIN) introduced by Kolotouros et al. [?] and PARE introduced by Kocabas et al. [?] for evaluation comparison.

Specifically, for HMR and SPIN, the training dataset was composed of 35% of the Moshed Human3.6M dataset, [2], 20% COCO [4], 15% MPI-INF-3DHP [?], 10% LSP [?], 10% LSPET [?], and 10% MPII [?]. For PARE the training dataset was composed of 50% Moshed Human3.6M dataset [2], 23.3% EFT-COCO [?, 4], 20% MPI-INF-3DHP [?], 4.6% EFT-LSPET [?,?], and 2.1% EFT-MPII [?, ?].

Once we had trained our baseline models using their base datasets, we could evaluate all of our testing scenarios. To do so, we developed a framework to automatically test each dataset within a category and return category-specific results, learning which subcategories of bias parameters performed worst for each model.

In Figures 2, 3 and 4, we illustrated the performances of our baseline models. Fig. 2 shows that *all methods perform worse on the lightest skin tones*. One possible explanation is that finding 2D details for Human Body Reconstruction heavily depends on edge detection. We also find biases towards smaller or shorter bodies, as seen in Fig. 3, and biases towards global or top-down light sources, as shown in Fig. 4. These biases seem consistent across all methods, leaving us to believe that, to some degree, these biases come from the training data.

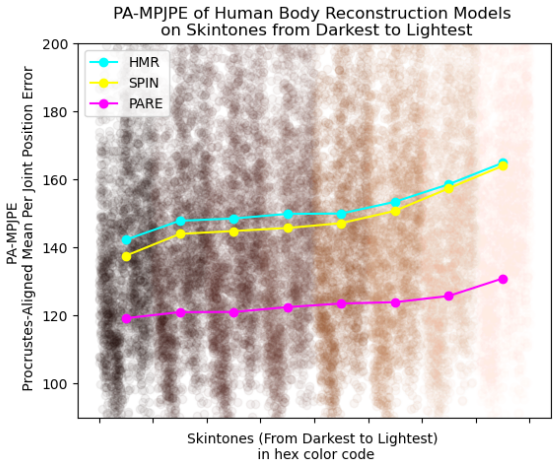

Figure 2. **Performance of HBR methods on images of synthetic persons of varying skin tone**. The skin tones are ordered from darkest to lightest, with the colors of the points representing the skin tones themselves. A lower PA-MPJPE is better. All methods, including the SOTA methods like PARE, perform worse on lighter skin tones.

Our fairness experiments observe which biases are ingrained in existing HBR techniques. Next, we present a novel framework to neutralize these biases.

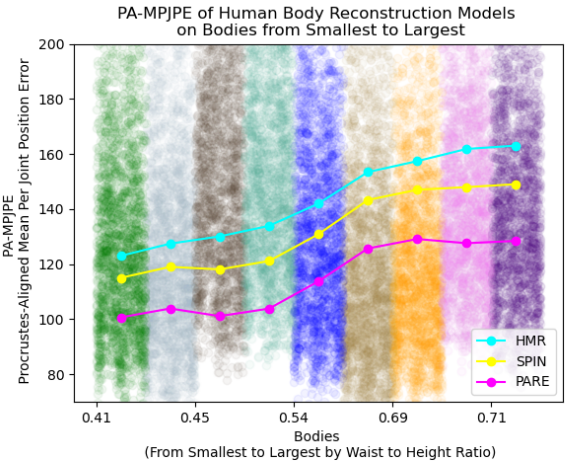

Figure 3. **Performance of HBR methods on images of synthetic persons of varying bodies**. The bodies are ordered from the smallest to the largest by waist-to-height ratio. A lower PA-MPJPE is better. All methods, including the SOTA methods like PARE, perform worse on larger bodies.

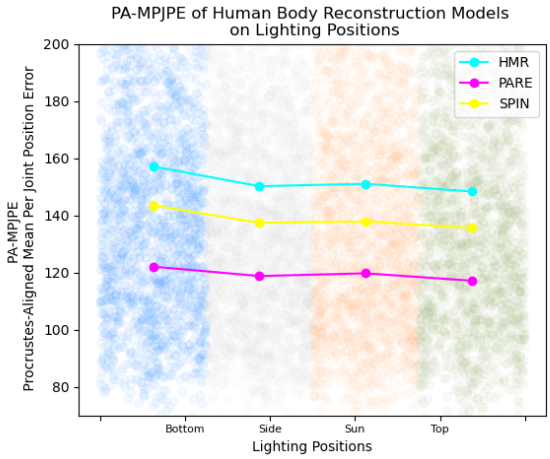

Figure 4. **Performance of HBR methods on images of synthetic persons lit from varying light positions**. The lighting positions vary from the global illumination (sun) to the location of the spotlight, including below (bottom), above (top), and to the side of the body. Different light hues and intensities are sampled for each lighting scenario at equal probabilities to allow for the same testing conditions with only the location changing. Here we see poorer reconstruction results occurring when the light source locates below or to the subject's side.

3.5. Bias Neutralizing Fine-Tuning

Our bias neutralization technique follows an iterative min-max training approach: an adversarial fine-tuning cy-

cle inter-weaved with our automatic bias evaluation used to choose new adversarial examples.

Initially we augment the original training datasets with a balanced set of data from our training dataset. This step is followed by training for some set number of fine-tuning epochs. The loss function for training is the PA-MPJPE or Procrustes-Aligned Mean Per Joint Position Error:

$$L = \frac{1}{J}\sum_{j=1}^{J}\|x_{3d,j} - x_{3d,j}^{gt}\|_2 \quad (1)$$

where $x_{3d,j}$ and $x^g t_{3d,j}$ are the estimated and groundtruth coordinates of the 3D joints of a body. At the end of training for each iteration we use mean loss or ($\bar{L}$) to evaluate all subsets of our bias parameters(skin tones, lighting conditions, body sizes, camera poses). We then choose 2 subset datasets for each parameter dataset to maximize our loss.

These subsets of data are then used to perform additional adversarial data augmentation for the next iteration of adversarial fine-tuning. Such a method introduces the maximally disruptive data so that the model may learn to counter and neutralize these biases.

Our method emulates traditional adversarial training in that we improve model robustness by training to minimize losses using the base training dataset with the addition of the bias datasets with the maximum losses. The loss function for bias neutralization is the following:

$$\min_{\theta}\max_{\mathbf{p}}\bar{L}(\theta, U_{\mathbf{p}}), \quad (2)$$

Here $\mathbf{p}$ represents the union of all subsets of bias parameters; θ denotes the model parameters; and $U_{\mathbf{p}}$ is the training dataset.

Generally such a task of adversarial training would require a large amount of data. Due to our ability to represent a large number of scenarios with a much smaller critical dataset, we offer *adversarial fine-tuning for cheap*. We demonstrate the results of fine-tuning bias neutralization using our dataset in Sec. 4, including the overall variance for all parameters and showing that the mean error drops significantly after bias neutralization.

Debiasing while maintaining general reconstruction performance Fine-tuning a converged neural network on data from a distinct domain typically leads to catastrophic forgetting of the original domain [**?**]. To ensure this does not happen, we only perform adversarial augmentation on 15% of the total training dataset. Due to this, the new ratios of the training data became the following.

For HMR and SPIN, their training data was composed of 30% of the Moshed Human3.6M dataset, [2], 15% COCO [4], 10% MPI-INF-3DHP [**?**], 10% LSP [**?**], 10% LSPET [**?**], 10% MPII [**?**] and 15% of our debiasing synthetic data.

Algorithm 1 Neutralize Biases in HBR from Images

Input: Existing trained model parameterized by θ
Output: Bias neutralized model
Initialization:
Initialize $t = 0$
T, the total # of iterations,
k, # training epochs in each iteration of bias neutralization,
$\theta^{(0)}$, the model parameters at $t = 0$.
B, the original base training data.

Let I represent the set of bias parameters $\{lighting, bodies, skintones, camera\,poses\}$.

Sample $D_{i,j}$ from each bias parameter dataset where $i \in I$ and j is the subset of the ith parameter.

Process:

while $t \leq T$ **do** {

Train the model for k epochs and update
$\theta^{(t+1)}$ = train($\theta^{(t)}$, $U_{\mathbf{p}}$, I) to minimize $\bar{L}(\theta^{(t+1)}, U_{\mathbf{p}})$

For each bias parameter, select D_{i,p_i} to maximize $\bar{L}$, s.t. $p_i = \arg\max_j \bar{L}(\theta^{(t)}, D_{i,j})$

Merge all selected datasets; augment with base training data
$U_{\mathbf{p}} = (\bigcup_{i=1}^{n} D_{i,p_i}) \bigcup B$
t = t+1
}

For PARE, the training data was composed of 30% Moshed Human3.6M dataset [2], 25% EFT-COCO [**?**, 4], 10% MPI-INF-3DHP [**?**], 10% EFT-LSPET [**?**, **?**], 10% EFT-MPII [**?**, **?**] and 15% of debiasing synthetic data.

3.6. Efficient Real-World Human Data Collection

We designed a novel data collection pipeline to collect data efficiently in arbitrary environments and without expensive equipment.

First, we use any camera, including mobile phones, to collect 10 seconds of videos of any steady pose. Since anyone can use this method at any time and anywhere, we can collect data on a larger scale in the future.

Then, we use either COLMAP [**?**, **?**] with SIFT feature extractor or OpenCV [**?**] to calculate both cameras' intrinsic and extrinsic parameters. By obtaining the frame-to-frame camera rotation and translation matrices, we can visualize the track of camera movement.

After that, we use TransPose [**?**] to estimate the human pose for each frame automatically. We also use the insightface python library [**?**] to detect facial landmarks and apply

Gaussian Blurry to these faces to reduce privacy concerns. However, since using a pose estimation model introduces bias and facial blurry does not have a 100% success rate, we check all images to adjust joint positions, remove invisible joints, and ensure all faces are blurred.

Finally, we craft a simple model that projects 3D points to 2D with the above camera parameters and minimizes the L_1 distance between projected and estimated points from TransPose [**?**]. Compared to finding the least-square intersection of rays directly, this method provides us with more accurate points.

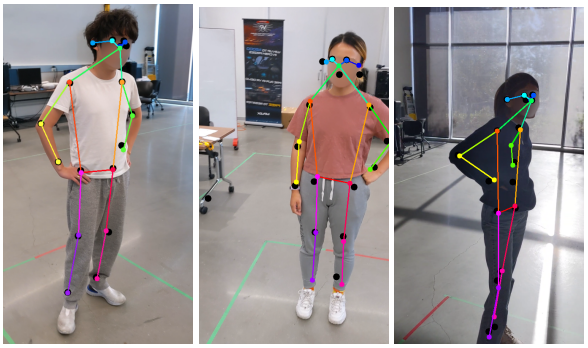

Figure 5. The visualization of joints in our dataset, while colorful points are points esitimated by our methods, and black ones are predicted by TransPose [**?**]

4. Results

Our goal with this challenge dataset was to explore the biases generated during training HBR methods using the currently available training data. Table. 2 demonstrates the ability to utilize our challenge dataset for bias neutralization, resulting in far less variance across bias categories and an overall lower average error. In Figure 2, 3 and 4 in Sec. 3.4, we visualized the biases present in the baseline models. Similarly, we present the bias-neutralized models' visualization in the supplemental materials. Our experiments show that PARE and HMR have benefited more than SPIN. We find improvements for mean PA-MPJPE to be as large as 90% for HMR, 70% for PARE, and 60% for SPIN. Biases toward skin tones obtained the maximum benefit from our bias neutralization technique. We also see extreme improvements in variance and mean PA-MPJPE.

Our dataset proves to be a good indicator of biases in HBR, and our bias-neutralizing adversarial fine-tuning method has proven to reduce biases.

5. Conclusion

With a large number of applications and goals for Human Body Reconstruction, this work evaluates the biases introduced by the training data due to various factors, including skin tones, body shapes, poses, lighting conditions, and camera poses, that affect the performance of human body reconstruction algorithms. We plan to publicly release our dataset and the code for automatic bias evaluation and neutralization. We aim to improve awareness of such partialities in HBR and reduce the embedded biases due to the training data.

Method	Challenge Set σ	Challenge Set $\bar{x}$	3DPW $\bar{x}$
Diverse Bodies			
HMR [**?**]	50.23	143.57	67.53
HMR+ b.n.	**26.74**	**106.99**	70.12
SPIN [**?**]	48.88	132.38	59.06
SPIN + b.n.	**34.64**	**120.31**	60.78
PARE [**?**]	39.54	114.89	50.78
PARE + b.n.	**22.29**	**94.48**	50.06
Diverse Skintones			
HMR	39.63	151.95	67.53
HMR + b.n.	**2.58**	**78.19**	70.12
SPIN	41.63	148.94	59.06
SPIN + b.n.	**10.49**	**92.77**	60.78
PARE	33.92	123.46	50.78
PARE + b.n.	**1.45**	**72.35**	50.06
Diverse Lighting Conditions			
HMR	56.16	151.69	67.53
HMR + b.n.	**31.78**	**112.09**	70.12
SPIN	54.30	138.60	59.06
SPIN + b.n.	**40.29**	**124.73**	60.78
PARE	44.00	119.43	50.78
PARE + b.n.	**24.28**	**98.79**	50.06
Diverse Camera Poses			
HMR	58.98	159.16	67.53
HMR + b.n.	**32.29**	**112.19**	70.12
SPIN	57.71	145.98	59.06
SPIN + b.n.	**40.84**	**127.19**	60.78
PARE	47.43	124.28	50.78
PARE + b.n.	**26.81**	**100.17**	50.06

Table 2. **Performance of baseline models before and after bias neutralization**. One can see the average error (PA-MPJPE) and variance improves significantly (by up to **95.7**%) on our challenge dataset after bias neutralization. Additionally overall performance on common test datasets, such as 3DPW [**?**], does not suffer.

References

[1] 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. pages 561–578, 2016. 2, 3, 4

[2] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2, 3, 4, 6, 7

[3] Junbang Liang and Ming C Lin. Shape-aware human pose and shape reconstruction using multi-view images. pages 4352–4362, 2019. 2

[4] 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*, pages 740–755. Springer, 2014. 2, 6, 7

[5] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018. 2, 3

[6] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5349–5358, 2019. 3

[7] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017. 3

A. Dataset Composition

Our dataset can be structured into 4 smaller datasets: bodies, skin tones, lighting and camera poses. The bodies dataset is composed of 9 subsets of data, one for each waist-to-height distribution. The average waist-to-height ratio of each subset from smallest to largest is the following: 0.41, 0.44, 0.45, 0.48, 0.54, 0.63, 0.69, 0.71, and 0.73.

The skin tones dataset is composed of 8 subsets of data, one for each skin tone. The hexadecimal color values from darkest to lightest are: #24100d, #401c17, #52241e, #714137, #a15c33, #c9865b, #edbc9d, and #ffece3.

The lighting datasets is composed of 4 subsets, categorized by the location and type of light used in the dataset, these are a 'below the body' spotlight labeled bottom, a 'to the side of the body' spotlight labeled side, an 'above the body' spotlight labeled top and an infinite area light which offers global illumination labeled 'sun'. For each subset the lights have sampled hues (3 possible), and intensities (3 possible), which are each sampled with equal probabilities. This approach was chosen as we saw no great changes in reconstruction results across light hues, however coupled with intensity this was an important attribute to vary.

For the bodies, skintones and lighting datasets, each camera pose from our possible 2500 is represented.

For the camera poses dataset, we have a dataset of 100,000 images taken from 2500 camera poses, 40 images per camera pose, with 40 images from each camera pose. The body, lighting and background are randomly sampled for each image.

B. Results After Bias Neutralization

We discuss our results and include our quantative results in Table 2, sec. 4.

While we see significant improvements in HMR [**?**] and PARE [**?**], SPIN [**?**] did not see the same results. In some cases we see HMR + bias neutralization, outperforming SPIN after bias neutralization. In some specific cases we also see the variance of SPIN's performance degrades a little after bias neutralization, while on average we still see improvements in mean PA-MPJPE.

In the case of bias neutralization against different bodies, we see that while mean PA-MPJPE drops, there is still some large variance across models. This may be due to incomplete training for this bias parameter. A future experiment may be to train adversarially against each parameter one by one, to ensure the model converges for each bias parameter.

We see the best improvements in neutralizing biases against variation in lighting and skin tone. We plot the results of the models after bias neutralization (+ BN) against the original distributions of the baselines in Fig. 6, 8, and 7. In addition to quantitative results, we show qualitative examples in Fig. 9 to demonstrate scenarios where biases

have been neutralized.

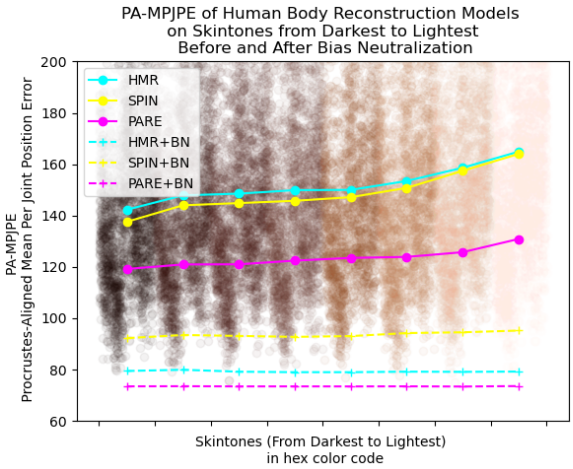

Figure 6. **Performance of HBR methods on images of synthetic persons of varying skin tone**. Here we see significant improvements in variance of results and mean PA-MPJPE with bias neutralization (+ BN).

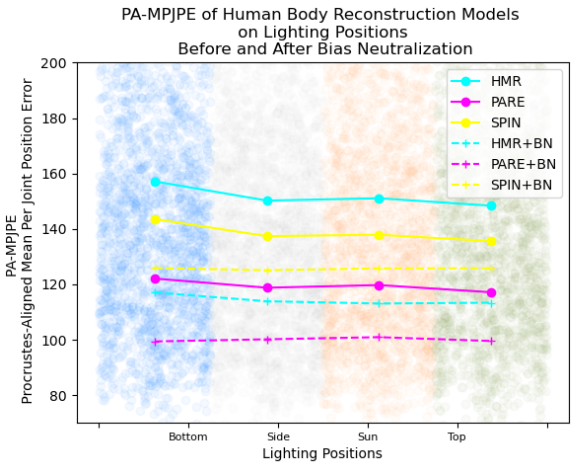

Figure 7. **Performance of HBR methods on images of synthetic persons lit from varying light positions**. Here we see improvements in variance and mean PA-MPJPE across all models after bias neutralization (+ BN).

C. Social Impact Assessment

This basis for this paper is to analyze and rectify biases in Human Body Reconstruction, and its related datasets. Our dataset allows HBR researchers to evaluate biases in their reconstruction methods. We offer a technique to then rectify and neutralize those biases. We also provide an efficient, effective and accessible way to collect large scale real-world human data so that future human image datasets

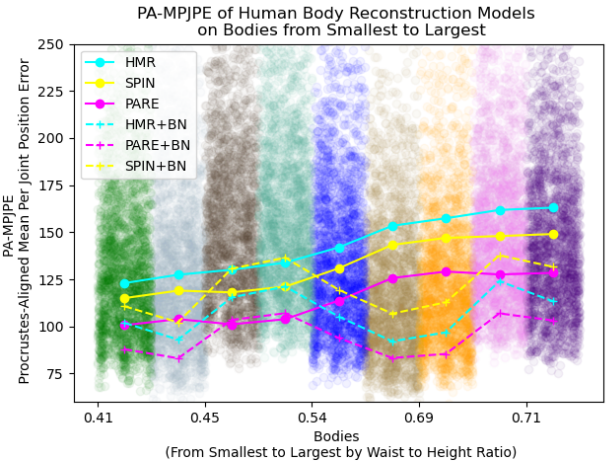

Figure 8. **Performance of HBR methods on images of synthetic persons of varying bodies**. Here we see significant improvements on mean PA-MPJPE for PARE and HMR with the bias neutralization (+ BN), while effects on variance are not as significant.

Figure 9. **Performance of HBR methods on real-world images before and after bias neutralization**. Here we see significant improvements in quality of body shape and size in reconstruction.

may contain more diverse populations, environmental conditions and camera perspectives. Thus, we see our work in this paper having the potential to greatly improve social interactions, due to more faithful and fair human body reconstruction, subjecting to less biases due to the training data.