SHARE: *S*ingle-view *H*uman *A*dversarial *RE*construction

Anonymous submission

Abstract

We propose a novel adversarial training and fine-tuning method to improve the robustness of human body reconstruction (HBR) against ubiquitous perturbations from the camera poses of input images, which can considerably influence the reconstruction of human body meshes. To analyze how camera location and orientation impact HBR, we generate large augmented datasets from a variety of camera perspectives and introduce an adversarial training method using such datasets on pre-existing human body reconstruction methods. Our experiments show that this approach significantly reduces the mean joint errors on single-view images from varying camera positions, while not affecting the baseline performance of existing models, and in several challenging cases improving overall performance. With our findings we propose an application for user-assisted camera adjustment to optimize camera poses for improved body reconstruction results. Additionally, we create a real-world image dataset automatically acquired by a precision-controlled robotic arm for further experimentation on the robustness of image-based 3D HBR.

1 Introduction

The ability to reconstruct a human body with accurate pose and shape has been an active research area with interests from various industries, including fashion, healthcare, special effects, computer animation, virtual and augmented reality (Liang and Lin 2019; Hu et al. 2018). Due to the simplicity and practicality of using a single uncalibrated image, robust and accurate single-view or monocular 3D human body reconstruction has been studied with renewed interest.

Having multi-view images or videos of the human body can offer improved reconstruction results (Liang and Lin 2019; Smith et al. 2019a; Sengupta, Budvytis, and Cipolla 2021; Kocabas, Athanasiou, and Black 2020; Kanazawa et al. 2019). However, to understand and characterize the impact of input image variations on the 3D human body reconstruction (HBR) systematically, the objective of this work is to improve reconstruction results against ubiquitous perturbations by first focusing on single-view images to establish the foundation of such experimental analyses. Furthermore, these types of perturbations are commonly found in a standard setting expected from user input images requiring minimal efforts using commodity mobile devices.

For single-view HBR, it is more critical to address the robustness issues on 3D body reconstruction subject to ubiquitous perturbations. Many single-view HBR models use feature extraction on input images. Relative camera poses (e.g. orientation and position) can have a considerable effect on captured images and estimated depth due to occlusions on the human form (Puscas et al. 2019; Kocabas et al. 2021b). Some angles may conceal or exaggerate different features of the object or person that the camera is trying to capture (Puscas et al. 2019; Zins et al. 2021; Alcorn et al. 2019). For HBR, many techniques rely on these 2D captures to guide the process, it is critical to understand what effects different camera poses may have on the reconstruction results. Moreover in the large-scale consumer settings of human body reconstruction such as virtual try-on or healthcare, such effects can have considerable impact on the user experience.

In this paper, we study the influence of different camera poses on human body reconstruction from images and pro-

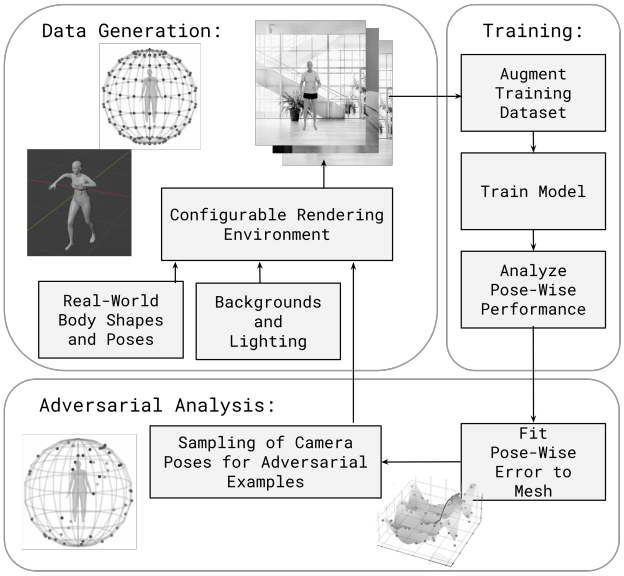

Figure 1: **The overall pipeline of SHARE**, consisting of three major components: Synthetic human data generation (Sec. 4.1) from various camera poses, the training of any existing model using the augmented image dataset, and the selection of camera poses for adversarial training (Sec. 5).

pose a novel methods to minimize disparities due to such image variations. Our adversarial training method can be applied to numerous different HBR techniques and improve robustness against unorthodox camera poses.

The key contributions of this paper include:

1. An automated creation of large-scale image datasets for given bodies, cameras and scene settings (Sec. 4.1);
2. A systematic study and analysis on *impact of camera poses* and lighting variations on the quality of *body reconstruction* using real and simulated images (Sec. 4.2);
3. A real-world image dataset automatically acquired by a precision-controlled robot arm to further research on robustness of image-based body reconstruction (Sec 4.3);
4. An *adversarial data augmentation* technique for a more robust learning-based human body reconstruction algorithm against image variation due to camera poses using *differentiable sampling techniques* (Sec. 5);
5. A *user-assisted camera adjustment* application to efficiently optimize camera poses for image capturing with improved body reconstruction results (Sec. 6).

2 Related Works

In this section, we review recent works on human body reconstruction, existing datasets on human bodies, and adversarial techniques against ubiquitous perturbations.

2.1 Human Body Reconstruction Methods

Methods for human body reconstruction (HBR) can be categorized by their inputs. They can range from single vantage point images (single-view) (Kanazawa et al. 2018; Bogo et al. 2016; Omran et al. 2018; Zheng et al. 2019; Pavlakos et al. 2019; Zanfir et al. 2021; Choutas et al. 2020), multiple vantage point images (multi-view), video input (Liang and Lin 2019; Sengupta, Budvytis, and Cipolla 2021; Zins et al. 2021; Hu, Ho, and Munteanu 2021; Lu et al. 2018), or a combination of visual inputs and other sensory inputs (Cheng et al. 2016; Xiang, Joo, and Sheikh 2019). The focus of this work is on single-view body reconstruction.

Many have noted gaps in single-view HBR capabilities, leading to new frameworks on using multi-view images, or additional video input (Liang and Lin 2019; Sengupta, Budvytis, and Cipolla 2021; Zins et al. 2021; Biggs et al. 2020). These methods that utilize additional data do improve accuracy (Liang and Lin 2019; Sengupta, Budvytis, and Cipolla 2021; Zins et al. 2021). The analysis, insights, and the experimental and analytical foundations from this work can also generalize to these techniques.

2.2 Synthetic and Real-World Human Datasets

Existing data-driven methods for 3D human body reconstruction have the ability to produce impressively accurate representations of human beings, and the use of large synthetic datasets aids in training these models (Chen et al. 2016). As a result, various synthetic datasets have been generated for HBR, but most have limitations on camera or human poses. (Bogo et al. 2016; Omran et al. 2018; Kanazawa et al. 2018; Tan, Budvytis, and Cipolla 2017; Alldieck et al. 2019). More recently some have generated synthetic datasets with more variance in these variables. For example datasets created by Liang et. al and Varol et. al generated realistic synthetic data from multiple view-points and distances, wherein the synthetic humans had a large variety of clothing, shapes and pose, however only a small number of camera poses were used, which is not enough to study the impact of camera pose on reconstruction results (Liang and Lin 2019; Varol et al.; Smith et al. 2019b).

Improved datasets such as AGORA and RenderPeople are available but generally do not allow for the generation of new data based on specified parameters. Parameters such as camera angle, lighting or background,as well as skin tone, clothing or height and weight (Patel et al. 2021; renderpeople 2018).

Almost all methods of HBR utilize large real-world datasets such as UP-3D, and the Leeds Sports Pose dataset (Zhang et al. 2019; Li et al. 2021), both of which tend to have single views of a particular subject (Johnson and Everingham 2010; Lassner et al. 2017). Many also used the Human3.6M dataset, however this dataset used 11 different actors, so only 11 different bodies, from 4 camera positions(Ionescu et al. 2013). Other commonly used datasets such as MPII , MPI-INF-3DHP, or 3DPW only contain a set of fixed camera positions, and do not include many of the camera positions which may be taken in the real world (Andriluka et al. 2014; Mehta et al. 2017; von Marcard et al. 2018).

With the existing datasets, when it comes to training HBR models, the conditions for accurate reconstruction are restricted in different ways. This propagates biases from the training data to the human body reconstruction application. These biases include those towards certain camera poses, which one can see in Fig. 3,improved performance under specific lighting conditions as well as poorer performance for different skin-tones, and body types.

Our synthetic data generator aims to fill the gaps and improve the biases caused by previous synthetic datasets. Our real-world dataset precisely collected using a robotic arm allows us to collate data from specific angles and positions repeatedly, offering an accurate way to emulate both virtual and real world settings.

2.3 Robustness & Adversarial Techniques:

Many methods of HBR, discuss failures due to challenging depth ambiguities (Kanazawa et al. 2018; Bogo et al. 2016; Omran et al. 2018). Different ubiquitous perturbations create various lighting and camera pose conditions, these conditions have the most prominent effect on depth perception (Todd et al. 2007). Recent works have proposed adversarial techniques for 3D HBR, however mostly the focus has been solely on human pose and do not address camera pose variation specifically (Wandt and Rosenhahn 2019; Chou, Chien, and Chen 2018; Ke et al. 2011; Kanazawa et al. 2018).

Liu et. al, Sun et. al. and Sardari et. al among others have also worked extensively on creating more camera pose invariant methods of HBR, but also solely focus on human pose estimation (Sun et al. 2019; Liu et al. 2021; Sardari, Ommer, and Mirmehdi 2021). Some have looked into re-

gression networks to combat occlusions created by human or camera pose and improve overall reconstruction (Jackson, Manafas, and Tzimiropoulos 2018; Chen et al. 2022; Guler and Kokkinos 2019; Kocabas et al. 2021a).

While pose may be an key part of reconstruction; shape, size and other features are also integral to effective HBR and in some applications are more important. Our approach is more holistic, focusing on all details of HBR, not pose alone. SHARE takes a novel approach to adversarial learning, focusing on making HBR models robust against camera pose variation.

With reference to adversarial techniques Shen et. al. described an adversarial technique to improve the robustness of autonomous steering through the use of adversarial training. In their implementation, they adversarially feed their model images with varying types and levels of degradation (Shen et al. 2021).

3 Overview

Our overall aim is to improve robustness in human body reconstruction (HBR) against ubiquitous perturbations caused by camera pose variation. As shown in Fig. 1, the overall pipeline of our framework, SHARE, begins with the large-scale generation of synthetic data (Sec. 4.1) to analyze the biases in pre-existing HBR methods. Our synthetic data generator aims to emulate all of the possible camera poses that may occur in the real world as well as edge cases and has the capabilities to emulate a vast variety of skin tones, lighting settings, body types and clothing.

Once this data is generated we can study the impact of different camera poses on human body reconstruction results from single-view images (Sec. 4.2). An analysis of these results will indicate which camera poses perform poorly as well as which camera poses perform better. Through this systematic sensitivity analysis, we can quantify to which degree camera pose affects the reconstruction. From these quantified results, we further implement two remedy methods. An adversarial framework to rectify the disparities caused by camera pose variation (Sec. 5) and a technique to optimize camera pose in the real world (Sec. 6). Additionally we created robot-arm controlled image capturing to emulate our data generator in the real world (Sec. 4.3)

4 Data Generation and Collection

4.1 Synthetic Datasets

To begin our experiments we require a large synthetic dataset of diverse human bodies from various camera poses in different environments. Such a comprehensive dataset does not already exist.

Many Human Body Reconstruction (HBR) methods utilize synthetic datasets to allow for validation of their reconstruction results (Liang and Lin 2019; Sengupta, Budvytis, and Cipolla 2020). However these datasets are generally created with reference to a few body shapes and poses, and have only a few vantage points, limited ranges for skin tones and lighting environments. Other datasets generated from video or animations do not emulate the real-world image generation needed for a systematic study.

Figure 2: Examples of images from our augmented image dataset, showing different views of the body, body types/shapes of wide distributions, poses, skin tones, backgrounds, clothing, and lighting environments. Please refer to the supplemental materials for additional images.

We propose to create a fully automated synthetic dataset generator. For this image generator, we need a rendering engine, a human body model, and a way to efficiently configure both. To create and configure human bodies in our synthetic dataset generator, we utilized the Skinned Multi-Person Linear Model (SMPL). The SMPL model allows us to create a large range of body shapes, sizes and poses, as controlled and parameterized by 82 SMPL parameters (Loper et al. 2015). We required a large number of poses, and to ensure they were realistic, we utilized real-world poses from large-scale datasets, such as those provided by (Kolotouros et al. 2019; Mehta et al. 2017; Bogo et al. 2016) for body poses.

To vary the lighting settings we included parameters to control the type, positions, colors and intensity of the light. The lighting of the body is also affected by the light reflected off of the background and setting of the body. For these variegated background settings we created a dataset of common backgrounds, ranging from wooded trails to office spaces and bedrooms.

For our automatic generator, a python script was written to accept specified features and settings to be rendered. These features and settings range from the body proportions, clothing, and skin tone of the human model, as well as the lighting, background and camera poses of the renderer. Given these requirements, the script dynamically creates a configuration file which is supported by multiple rendering environments. We have tested our image generator on both Unity (Juliani et al. 2018) and Blender (Community 2018), creating scripts to accept different configuration files and return the rendered images along with corresponding details.

Using the rendered images and the respective camera position from which they are taken, we are able to project the 3-D human joint locations onto their 2-D pixel locations in the image. For this we use the image and the combination of the camera intrinsic and extrinsic matrix from the rendered environment.

Using our image generator, we synthesized images from multiple view points around the human body using the polar coordinate system. Fig.(2) shows a few example images from our synthetic dataset while additional examples are included in the supplemental material.

4.2 Sensitivity Analysis on Impact of Camera Poses

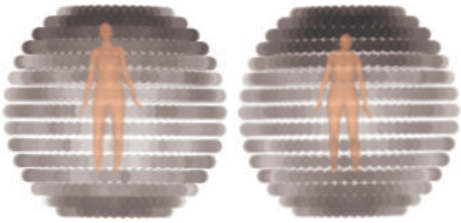

Figure 3: Error heat-map of camera poses and their respective reconstruction results. Darker black poses show poor reconstruction results, while lighter white poses show good reconstruction results for images taken from those positions. Here we show the front (left) and back (right) views of the body. Images taken from below the waist have demonstrated improved reconstruction results. In the center is a placeholder to represent where the various bodies are placed in reference to the camera poses. To see what our synthetic data/bodies look like, please refer to Sec. 4.1.

Our goal after generation was to identify which camera poses provided images for the best reconstruction results and those resulting in poor body reconstruction. As a baseline, we first tested these various camera positions on the Human Mesh Recovery framework (HMR) in the End-to-end Recovery of Human Shape and Pose (Kanazawa et al. 2018).

Using the true joint locations and the reconstructed ones, we were able to calculate the average mean errors when reconstructing the human body through images from our dataset on HMR. We use the standard mean per joint position error (MPJPE) metric to evaluate reconstructed human bodies, as well as a variation that includes Procrustes alignment (PA-MPJPE) (Ionescu et al. 2013). The average PA-MPJPE from all camera poses was 154.04 mm, which is significantly greater than the reported average PA-MPJPE of 67.53 mm on their test dataset. This indicates that the larger variety of camera poses results in a great impact on reconstruction results. Thus, rectifying any losses due to camera pose variations can improve reconstruction accuracy.

Along with these average scores, we also are able to see which camera poses performed better or worse than others. Fig. 3 shows a heat-map of the performance of images taken from different camera poses. Black represents a larger error or the worse camera poses, while White gives us the lowest error, and the best camera poses. One sees that the best images for reconstruction are taken from the front towards the center of the body but at a lower angle than we had initially anticipated, the next best views were taken from the corners of the body. This result confirms our hypotheses, as these views offer the best depth perception with minimal occlusion of the body. Interestingly images taken from above hip-level performed far worse than other images, with the worst images being from the top-backside. This is most likely due to a lack of depth perception with the chest and head occluding the rest of the body.

Images taken in the real world are not curated by experts for the purpose of training human body reconstruction methods may not fit the same guidelines that the images used in training or testing provide. If user input is to be reliably used for human-body reconstruction, the quality of reconstruction must be robust to occlusion and lack of depth perception due to camera position and angle.

The large variance in results from the different camera poses prove that there is a need to ameliorate human body reconstruction methods against camera pose variation. Therefore, we propose an adversarial data augmentation technique to optimize camera poses for better reconstruction results.

4.3 Robot-Assisted Data Collection

For applications such as virtual try-on, the images for HBR will come from the user. The input images for reconstruction methods will thus be subject to different ubiquitous perturbations. Emulating these perturbations in a virtual environment is easily possible, however virtual environments fail to fully capture the effects of real-world images.

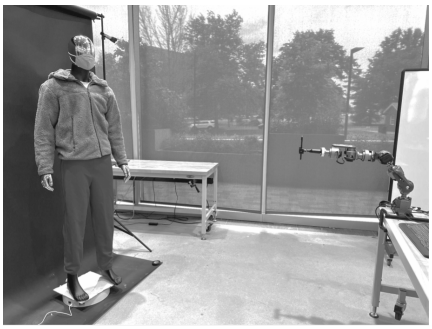

Figure 4: Set up of our mannequin on motorized turn table as well as the Amber B1 robot. The three images show different robot arm positions as well as 3 different angles from the turn table. Results on these images can be seen in table 2.

Current real-world datasets are obtained from a few view points, and human subjects cannot stand entirely still or recover precise positions for iterations of data collection. For this reason, we propose a robot-assisted data capturing of human body data. We utilize robot arms to mimic how a human may a take a picture. For our set up, we utilized an Amber B1 robot arm, a Google Pixel 4a, and a motorized turn-table with the ability to move a single degree at a time.

Fig. 4 shows the real-world data collection configuration, with further implementation details and specifications included in the supplementary document. The robot arm provides the ability to capture images from precise positions and from those positions we are able to modulate the panning and tilting of the phone camera. The motorized turntable offers us a 360° view of the mannequin, allowing us to generate a comprehensive real-world dataset. Results from these images can be seen in the supplementary materials, along with more comprehensive information regarding the robot-assisted data collection.

This setup allows us to take images as a human would but to repeat the action with precision and accuracy to carry out controlled experiments that use real-world collected data.

With this set up we are able to emulate our configurable virtual environment in the real world. Although limited to one mannequin in our current experiments, this data collection method can be expanded to create large-scale real-world datasets, in a manner similar to that in the virtual world.

5 Adversarial Data Augmentation Technique

Our method, Single-view Human Adversarial Reconstruction, or SHARE, may be applied in two ways, either as an adversarial training method, or as an adversarial fine-tuning method for pre-trained HBR models.

Our algorithm works by adversarially augmenting and modifying the synthetic training data of a single-view HBR model at specific intervals. One interval consists of a number of epochs, in our setup specifically, 5 epochs to one interval. After each interval of training we augment the training data set with images generated from sampled camera poses. These camera poses are sampled using a differentiable sampling method. Fig. 1 gives an overview of the SHARE pipeline.

We first generate our synthetic training data and evaluation data using our synthetic data generator, described in Sec. 4.1. Our evaluation dataset consists of images from all camera poses along with their corresponding expected outputs. Specific regarding our experimental setup are described in the experimental setup description (Sec. 7). For the very first interval of training, the synthetic training dataset consists of images from all camera poses as the initial round of sampling has yet to occur.

After an interval of epochs,, we extract an example-wise or camera-pose-wise error. This is done so that we may analyze the error on each camera pose, rather than a single global average error. Based on these camera-pose-wise errors, we utilize a differentiable sampling method to select a set of camera poses. This set of poses is then used to generate a new synthetic training dataset. The training dataset of the model is then augmented with the new synthetic training dataset, and the next interval of training begins.

During the new synthetic dataset generation, to improve human body and pose diversity in the data, we randomly choose SMPL (Bogo et al. 2016) parameters from the MPI-INF-3DHP (Mehta et al. 2017) dataset, which are based on the real-world bodies and poses, and use them to generate new joints and meshes. We additionally randomly sample various backgrounds and clothing for our synthetic humans. This allows for a more diverse set of bodies, poses, backgrounds, and lighting situations.

Utilization of Existing Models. Training for the selected HBR method runs typically within each interval, allowing SHARE to be a framework that can be applied to a host of HBR methods. The only requirement is that the models have similar input-output formats, which many of the state-of-art HBR methods employ, thereby making SHARE, our single-view human adversarial reconstruction technique, *shareable* in nature.

Sampling for Selecting Adversarial Examples. To sample camera poses we fit the camera-pose-wise errors to a mesh to create a loss landscape. Since the camera positions are in spherical coordinates, and our radius is fixed for the evaluation dataset, we can use the θ and ϕ of the camera position as the x and y coordinates of each camera pose in our mesh. We then train a multi-layer perceptron to predict the error given a θ and ϕ location. This predicted $error$ is used along the z axis for each camera pose. We then scale the $error$, θ and ϕ into the range [-1, 1]. One such mesh is shown in Fig. 5.

Different sampling methods obtain different results. The most effective sampling method proved to be the Voxel Sampling. We perform an ablation study to find the optimal sampling strategy and refer the readers to the supplementary document for more details.

The Voxel sampling method works on the idea that we can assign a local average to a small area or voxel on our loss landscape. Suppose that we have N samples on our mesh. We calculate the first $G1_n$ and second $G2_n$ derivatives at every point. Our samples now have the specifications $\{(X_n, Y_n, W_n)\}_{n=1}^{N}$, where X_n, Y_n are coordinates of the sample on our fitted mesh, and W_n is the sum of the error, E, and the first and second derivatives of the sample. Thus, $W_n = |E| + |G1_n| + |G2_n|$. We then set a variable P that represents the numbers of partitions to be created for our mesh. For example, if $P = 3$ we create $3^3 = 27$ voxels or sub-areas within our mesh.

We calculate a threshold $\tau = Mean(W_n)$ of all samples and remove all voxels that do not contain any samples passing the threshold τ. Next, we calculate the $Mean(W_n)$ of all samples within each of the remaining voxels, this set of average voxel scores is denoted by $\{AW_v\}$. If the average voxel score is greater than the threshold τ, that is $AW_v > \tau$, we randomly select two samples from the voxels. If the average voxel score is less than the threshold, that is, $AW_v \leq \tau$, we select only one sample.

These selected samples are the results of voxel sampling. This form of sampling chooses samples in such a way to give importance to regions of camera poses that perform poorer than others. The camera poses of the selected samples are the ones utilized to generate new synthetic adversarial examples for the next iteration of SHARE.

6 User-Assisted Camera Pose Optimization

During the sampling process we train a multi-layer perceptron (MLP) to predict the loss given the θ and ϕ of the camera (camera pose). This means we are able to predict the loss for any camera pose in the polar coordinate space, which allows us to estimate how well or how poorly an image captured from a camera pose will perform when used for human body reconstruction. This is useful as there is no other way to check the accuracy of real-world scenarios, as we do with synthetic ones.

If we know our current camera pose is in a position likely to result in higher errors, we can then guide the user to a nearby pose for improved results. Even better, if possible, we can guide the user to the optimal camera pose to take the best possible image for human body reconstruction results. This path to the minimum error is found via an iterative gradient descent technique: given an initial camera pose, we find its point on the mesh and its gradient, we find another

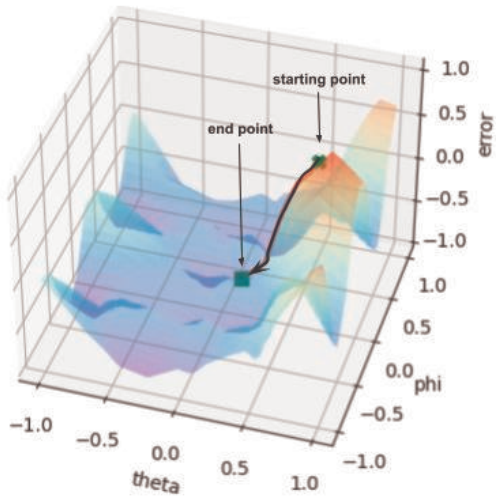

Figure 5: This plot represents the loss landscape of the pre-trained model tested on our synthetic validation set, and the gradient-descent path for optimization drawn shown with an arrow. The X and Y axis are used to plot the phi and theta location of each camera pose while the Z axis represents the average Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE). This is the basis for our camera pose optimization process.

point amongst its surroundings that allows us to move to an area of lower errors. This point then becomes our new starting point, and the descent continues. While this search runs, we keep track of all of the points we have visited. We continue descending until the error converges. Finally, the list of visited points in order become the path to the minimum error within a polar coordinate space.

Fig. 5 shows the loss landscape created by fitting the ϕ, θ and the error of a camera pose to a mesh. This loss landscape has the path for optimization marked with an arrow. Following the path from the initial result, we are led to a different perspective from which we can capture an image for improved reconstruction results.

7 Results

Experimental Setup: Our setup for our primary experiments utilized HMR (Kanazawa et al. 2018) and SPIN (Kolotouros et al. 2019), which we configure the same way as described in the OpenMMLab 3D Human Parametric Model Toolbox and Benchmarks, under the Apache License 2.0 (Contributors 2021). For our adversarial training method, HMR + SHARE we reduce the ratio of of different datasets to include our synthetic data. For this we reduced Human3.6M (Ionescu et al. 2013) from 35% to 30%, Common Objects in Context (Lin et al. 2014) from 20% to 10%, and then set our synthetic dataset to be 15%.

We sample 50,000 from our synthetic training dataset and 10,000 from our synthetic testing dataset. One training interval of SHARE consisted of 5 epochs. We trained our model on multiple servers: a dual NVIDIA 3090 machine takes around 1 day for 40 epochs, and a dual NVIDIA 2080 Ti machine takes 1.5 day for 20 epochs.

7.1 Overall Performance

Method	Augmented Data (MPJPE/PA-MPJPE in mm)	3DPW
HMR	357.62/154.04	112.34/67.53
HMR + SHARE	**113.66/107.85**	117.26/68.16
SPIN	364.87/145.98	96.06/59.06
SPIN + SHARE	**138.15/115.73**	99.34/60.27

Table 1: **Performance of SHARE on SPIN and HMR Compared with the Performance of the Baseline Model**. We observe error reductions (*with* and *without* SHARE for both HMR and SPIN) due to the adversarial training. SHARE offers the best reconstruction results on the augmented dataset with a large body shape distribution, while maintaining comparable performance on full 3DPW dataset.

We demonstrate the capabilities of SHARE on different methods, to show improvements in performance as well as the *shareable* quality of such a framework. The two human body reconstruction techniques we applied SHARE on are the Human Mesh Recovery (HMR) framework (Kanazawa et al. 2018) as well as SMPL Optimization in the loop (SPIN) framework (Kolotouros et al. 2019).

The importance of testing on synthetic data is due to the fact that it is impractical and error-prone (esp. with soft, deformable, and articulated human bodies) to validate the exact shapes and poses using real-world data, especially on large-scale real-world datasets. For this reason, we utilize our synthetic dataset (Sec. 4.1) on a hard-surface mannequin with test and validation images taken by a precision-controlled robot arm (with less uncontrollable variability).

We also test on the 3DPW dataset (von Marcard et al. 2018). The 3DPW dataset covers a wide variety of activities, these activities offer many different angles; however, the angles from these activities are not ones common to the use cases for single-view reconstruction, such as virtual try-on, metaverse, virtual avatars, etc. Additionally, the 3DPW dataset, though widely used to compare human body reconstruction methods, contains only **5** different human actors, making it not representative of real-world body distribution for testing robustness. Further information regarding the demographics of the 3DPW dataset and our augmented test datasets is included in the supplemental materials.

As seen from Table 1, SHARE improves the performance of HMR by around **30%** and SPIN by around **25%** when tested on the augmented dataset with diverse camera poses with bodies having much larger variations in BMI. In contrast, the 3DPW dataset contains only 5 human bodies conducting many activities, with limited shape variation, and does not have similar data distribution to our augmented dataset, there is no significant drop in performance – the difference between the baseline models and the augmented models using SHARE is less than 1.5%.

Furthermore, our overall goal is to ensure that not only the accuracy of joint positions is improved, but also the overall dimensions of the *human body measurements* are also pre-

Source	Height	Neck	Chest	Waist	Hips	Arms	Legs
Groundtruth	185.65±1.3	37.985±0.5	94.91±2.5	80.18±1.0	100.21±1.1	30.8±0.4	56.28±0.5
HMR	175.52	**37.12**	108.82	100.00	110.45	**38.05**	45.51
HMR + SHARE	**186.29**	34.96	**95.45**	**82.90**	**102.21**	**38.05**	**51.10**
SPIN	168.73	33.46	97.79	87.50	104.12	36.35	**44.70**
SPIN + SHARE	**189.80**	**34.93**	**96.30**	**85.5**	**102.93**	**35.70**	43.75

Table 2: **Body Measurements of SHARE vs. Baseline Models, Compared with the Ground-truth Measurements of the Mannequin**. We observe that the adversarial training provided by SHARE offers the most accurate body measurements.

served. Table 2 demonstrates how *SHARE improves preservation of body shape and dimensions*. We take measurements from the mannequin as 'ground truth' and compare the mean measurements taken from the reconstructed bodies using the different models. We observe that *for most body measurements*, such as height, chest, waist, and hips, **SHARE provides the best results**.

These studies illustrate that SHARE improves robustness of models against variations in camera poses and enables better preservation of reconstructed body poses and shapes.

7.2 Ablation Study on Synthetic Dataset and Adversarial Training

To show that SHARE gains from adversarial learning, we need to analyze the impact of synthetic datasets in training. To do so, we train the HMR model using our pipeline with $50,000$ synthetic training images from all camera views, and disabled the adversarial augmentation.

Method	Augmented Data (MPJPE/PA-MPJPE in mm)	3DPW
HMR	357.62/154.04	112.34/67.53
HMR + Synthetic	131.44/118.48	132.63/73.70
HMR + SHARE	**113.66/107.85**	117.26/68.16
SPIN	364.87/145.98	96.06/59.06
SPIN + Synthetic	255.45/143.73	106.83/64.24
SPIN + SHARE	**138.15/115.73**	99.34/60.27

Table 3: Performance of HMR and SPIN with the inclusion of synthetic data compared with HMR and SPIN with SHARE on our augmented dataset and 3DPW data. As seen, adversarial training by SHARE provides the best results.

From Table 3 we can compare the performance of HMR and SPIN with the sole inclusion of the synthetic dataset as well as HMR and SPIN with SHARE. The simple inclusion of the synthetic dataset improves reconstruction results for HMR, but the true improvement, we can see, comes from the adversarial data augmentation of SHARE offers.

Additional Studies and Results We have performed additional studies such as an Ablation study on different sampling strategies with SHARE. These can be found in the supplementary materials.

8 Conclusion

In this work, we have analyzed current human body reconstruction methods and demonstrated that they are not robust against ubiquitous perturbations due to variations in camera poses, largely due to a lack of diversity in camera poses in the training data and occlusions that occur during image capturing. To address these problems, we provide an automatic synthetic data generator to create large datasets with sampled camera views as well as a technique to utilize a robot arm for real-world data collection.

We further present two techniques to improve the reconstruction results: (1) A *differentiable* user-assisted camera adjustment to optimize camera pose for image capturing and improving body reconstruction results, and (2) An *adversarial data augmentation* technique for a more robust learning-base human body reconstruction algorithm against pose-induced image variations using differentiable sampling techniques, known as Single-View Human Adversarial REconstruction (SHARE). We demonstrate that SHARE effectively minimizes the errors cause by camera variation on existing methods. We also construct an automated system to collect real-world image data using a precision-controlled robot arm, with the collected dataset to be released along with this publication to further research in this area.

Future Work. It is important to note that other perturbations such as extreme light variation may still limit reconstruction results. However, SHARE can be generalized to address these types of image degradation as well, similar to (Shen et al. 2021). Based on SHARE, we can also provide a more comprehensive adversarial training for handling all types of image degradation in human body reconstruction. For the automatic optimization of camera poses, we hope to deploy it fully as a mobile application for automatic correction and/or improvement of reconstruction results based on camera poses.

References

Alcorn, M. A.; Li, Q.; Gong, Z.; Wang, C.; Mai, L.; Ku, W.-S.; and Nguyen, A. 2019. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4845–4854.

Alldieck, T.; Pons-Moll, G.; Theobalt, C.; and Magnor, M. 2019. Tex2Shape: Detailed Full Human Body Geometry From a Single Image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Biggs, B.; Novotny, D.; Ehrhardt, S.; Joo, H.; Graham, B.; and Vedaldi, A. 2020. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. *Advances in Neural Information Processing Systems*, 33: 20496–20507.

Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; and Black, M. J. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. 561–578.

Chen, D.; Song, Y.; Liang, F.; Ma, T.; Zhu, X.; and Jia, T. 2022. 3D human body reconstruction based on SMPL model. *The Visual Computer*, 1–14.

Chen, W.; Wang, H.; Li, Y.; Su, H.; Wang, Z.; Tu, C.; Lischinski, D.; Cohen-Or, D.; and Chen, B. 2016. Synthesizing training images for boosting human 3d pose estimation. In *2016 Fourth International Conference on 3D Vision (3DV)*, 479–488. IEEE.

Cheng, K.-L.; Tong, R.-F.; Tang, M.; Qian, J.-Y.; and Sarkis, M. 2016. Parametric Human Body Reconstruction Based on Sparse Key Points. *IEEE Transactions on Visualization and Computer Graphics*, 22(11): 2467–2479.

Chou, C.-J.; Chien, J.-T.; and Chen, H.-T. 2018. Self adversarial training for human pose estimation. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 17–30. IEEE.

Choutas, V.; Pavlakos, G.; Bolkart, T.; Tzionas, D.; and Black, M. J. 2020. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, 20–40. Springer.

Community, B. O. 2018. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam.

Contributors, M. 2021. OpenMMLab 3D Human Parametric Model Toolbox and Benchmark. https://github.com/open-mmlab/mmhuman3d.

Guler, R. A.; and Kokkinos, I. 2019. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10884–10894.

Hu, P.; Ho, E. S.-L.; and Munteanu, A. 2021. 3DBodyNet: fast reconstruction of 3D animatable human body shape from a single commodity depth camera. *IEEE Transactions on Multimedia*, 24: 2139–2149.

Hu, P.; Li, D.; Wu, G.; Komura, T.; Zhang, D.; and Zhong, Y. 2018. Personalized 3D mannequin reconstruction based on 3D scanning. *International Journal of Clothing Science and Technology*.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. 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.

Jackson, A. S.; Manafas, C.; and Tzimiropoulos, G. 2018. 3d human body reconstruction from a single image via volumetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 0–0.

Johnson, S.; and Everingham, M. 2010. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *Proceedings of the British Machine Vision Conference*.

Juliani, A.; Berges, V.-P.; Teng, E.; Cohen, A.; Harper, J.; Elion, C.; Goy, C.; Gao, Y.; Henry, H.; Mattar, M.; et al. 2018. Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627*.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end Recovery of Human Shape and Pose.

Kanazawa, A.; Zhang, J. Y.; Felsen, P.; and Malik, J. 2019. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5614–5623.

Ke, S.-R.; Hwang, J.-N.; Lan, K.-M.; and Wang, S.-Z. 2011. View-invariant 3D human body pose reconstruction using a monocular video camera. In *2011 Fifth ACM/IEEE International Conference on Distributed Smart Cameras*, 1–6. IEEE.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 2020. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5253–5263.

Kocabas, M.; Huang, C.-H. P.; Hilliges, O.; and Black, M. J. 2021a. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11127–11137.

Kocabas, M.; Huang, C.-H. P.; Tesch, J.; Müller, L.; Hilliges, O.; and Black, M. J. 2021b. SPEC: Seeing people in the wild with an estimated camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11035–11045.

Kolotouros, N.; Pavlakos, G.; Black, M. J.; and Daniilidis, K. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2252–2261.

Lassner, C.; Romero, J.; Kiefel, M.; Bogo, F.; Black, M. J.; and Gehler, P. V. 2017. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6050–6059.

Li, J.; Xu, C.; Chen, Z.; Bian, S.; Yang, L.; and Lu, C. 2021. HybrIK: A Hybrid Analytical-Neural Inverse Kinematics Solution for 3D Human Pose and Shape Estimation.

Liang, J.; and Lin, M. C. 2019. Shape-aware human pose and shape reconstruction using multi-view images. 4352–4362.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, T.; Sun, J. J.; Zhao, L.; Zhao, J.; Yuan, L.; Wang, Y.; Chen, L.-C.; Schroff, F.; and Adam, H. 2021. View-invariant, occlusion-robust probabilistic embedding for human pose. *International Journal of Computer Vision*, 1–25.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6): 248:1–248:16.

Lu, Y.; Zhao, S.; Younes, N.; and Hahn, J. K. 2018. Accurate nonrigid 3d human body surface reconstruction using commodity depth sensors. *Computer animation and virtual worlds*, 29(5): e1807.

Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; and Theobalt, C. 2017. Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE.

Omran, M.; Lassner, C.; Pons-Moll, G.; Gehler, P.; and Schiele, B. 2018. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, 484–494. IEEE.

Patel, P.; Huang, C.-H. P.; Tesch, J.; Hoffmann, D. T.; Tripathi, S.; and Black, M. J. 2021. AGORA: Avatars in Geography Optimized for Regression Analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10975–10985.

Puscas, M. M.; Xu, D.; Pilzer, A.; and Sebe, N. 2019. Structured coupled generative adversarial networks for unsupervised monocular depth estimation. In *2019 International Conference on 3D Vision (3DV)*, 18–26. IEEE.

renderpeople. 2018. Renderpeople. Https://renderpeople.com/3d-people/.

Robotics, A. 2022. AMBER B1 Wiki How. https://github.com/AmberInside/Amber_AI_ROS2/wiki/AMBER-B1-Wiki-&-How.

Sardari, F.; Ommer, B.; and Mirmehdi, M. 2021. Unsupervised View-Invariant Human Posture Representation. *arXiv preprint arXiv:2109.08730*.

Sengupta, A.; Budvytis, I.; and Cipolla, R. 2020. Synthetic training for accurate 3d human pose and shape estimation in the wild. *arXiv preprint arXiv:2009.10013*.

Sengupta, A.; Budvytis, I.; and Cipolla, R. 2021. Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16094–16104.

Shen, Y.; Zheng, L.; Shu, M.; Li, W.; Goldstein, T.; and Lin, M. C. 2021. Improving Robustness of Learning-based Autonomous Steering Using Adversarial Images. *arXiv preprint arXiv:2102.13262*.

Smith, B. M.; Chari, V.; Agrawal, A.; Rehg, J. M.; and Sever, R. 2019a. Towards accurate 3D human body reconstruction from silhouettes. In *2019 International Conference on 3D Vision (3DV)*, 279–288. IEEE.

Smith, B. M.; Chari, V.; Agrawal, A.; Rehg, J. M.; and Sever, R. 2019b. Towards Accurate 3D Human Body Reconstruction from Silhouettes. In *2019 International Conference on 3D Vision (3DV)*, 279–288.

Sun, Y.; Ye, Y.; Liu, W.; Gao, W.; Fu, Y.; and Mei, T. 2019. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5349–5358.

Tan, J. K. V.; Budvytis, I.; and Cipolla, R. 2017. Indirect deep structured learning for 3d human body shape and pose prediction. *British Machine Vision Conference*.

Todd, J. T.; Thaler, L.; Dijkstra, T. M.; Koenderink, J. J.; and Kappers, A. M. 2007. The effects of viewing angle, camera angle, and sign of surface curvature on the perception of three-dimensional shape from texture. *Journal of vision*, 7(12): 9–9.

Varol, G.; Ceylan, D.; Russell, B.; Yang, J.; Yumer, E.; Laptev, I.; and Schmid, C. ???? Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 20–36.

von Marcard, T.; Henschel, R.; Black, M.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera.

Wandt, B.; and Rosenhahn, B. 2019. RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7774–7783.

Xiang, D.; Joo, H.; and Sheikh, Y. 2019. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10965–10974.

Zanfir, A.; Bazavan, E. G.; Zanfir, M.; Freeman, W. T.; Sukthankar, R.; and Sminchisescu, C. 2021. Neural descent for visual 3d human pose and shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14484–14493.

Zhang, H.; Cao, J.; Lu, G.; Ouyang, W.; and Sun, Z. 2019. Danet: Decompose-and-aggregate network for 3d human shape and pose estimation. In *Proceedings of the 27th ACM International Conference on Multimedia*, 935–944.

Zheng, Z.; Yu, T.; Wei, Y.; Dai, Q.; and Liu, Y. 2019. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7739–7749.

Zins, P.; Xu, Y.; Boyer, E.; Wuhrer, S.; and Tung, T. 2021. Data-Driven 3D Reconstruction of Dressed Humans From Sparse Views. In *2021 International Conference on 3D Vision (3DV)*, 494–504. IEEE.

A Ablation Study on Sampling Strategies

Analyzing which camera positions to augment images from is an integral component of SHARE. We proposed two different camera pose sampling algorithms: Greedy Sampling and Voxel Sampling. Thus, we train different models with the exact same parameters using different strategies to compare their performances. The description for **voxel sampling** is included in Sec 5.

Greedy Sampling

The greedy method works by sampling the worst performing camera poses from the loss landscape. This is different from voxel sampling as we do not look into regions of poor performance but rather specific points.

Suppose that we have N samples on our mesh. Our samples have the specifications $\{(x_n, y_n, e_n)\}_{n=1}^{N}$, where x_n, y_n are coordinates of the sample on our fitted mesh,and e_n is the error. We then set a variable K to be the number of camera poses we want to sample.

We then find the distance d_i between each sample with the specifications $\{(x_i, y_i, e_i)\}$ and the rest of the samples to create the set $\{d_i\}$.

This distance, $d_i = \sqrt{(x_n - x_i)^2 + (y_n - y_i)^2} + \alpha * |e_n - e_i|$ where α is some small constant, in our case $\alpha = 3$.

With the set of distances $\{d_i\}$, we create a variable $d_{max} = 0$. Then for K iterations we find $d_{max} = max(d_{max}, Min(\{d_i\}))$, removing the d_{max} sample from the $\{d_i\}$ set each iteration and adding it to our list of camera poses to be sampled.

Using this min-max method we can approximate the worst performing camera poses in $O(n^2)$.

Method	Augmented Data (MPJPE/PA-MPJPE in mm)	3DPW
HMR	357.62/154.04	112.34/67.53
HMR + SHARE (greedy)	115.32/**99.70**	116.63/69.37
HMR + SHARE (voxel)	**113.66**/107.85	117.26/68.16

Table 4: **Greedy vs Voxel Sampling for SHARE**. While both sampling methods greatly improve reconstruction results, voxel sampling offers the most accurate results for MPJPE, while greedy sample works best with Procrustes Alignment. Both sampling strategies do not drop performance significantly when evaluated with the real-world dataset.

From Table 4 depending on the users preferred error metric, the various sampling methods perform differently, while both still significantly improve performance. If one is using MPJPE, the Voxel sampling is better, whereas if one is using PA-MPJPE, Greedy sampling is more appropriate.

B Types of Ubiquitous Variations in Data and Their Effects

The end users of HBR methods is a diverse population and the environmental settings in which images are collected will vary just as much as the bodies do.

Here we study the effects of lighting, skin-tone and body size and shape on reconstruction results, while keeping all other variables constant. The findings show contrast between the body and and the background lead to improved reconstruction results, and smaller bodies tend to frame better. Along with the results for camera pose variation from Sec 4.2 it is important to ensure that HBR methods are robust so that more users feel confident in their reconstruction capabilities. To test the effects of these variations we utilized our configurable synthetic data generator, described in Sec 4.1.

Lighting We tested 36 different lighting settings, with 4 locations for the light source (global ambient or Sun, top, bottom and side lighting), 3 types of light intensity (dim, medium, bright) and 3 different light hues (cool, warm, neutral). For each setting, we generated images from 2,500 camera poses.

Figure 6: Images from the worst and best lighting scenarios, the left is the worst lighting, Bottom-lit, Bright, Cool-tone lighting and the right is the best lighting condition, Top-lit Bright Neutral-tone lighting.

The results from Table 5 are sorted from best to worst by PA-MPJPE. They show that under all tested light settings, lighting from the top down improved reconstruction results the most and lighting from the bottom impaired reconstruction results. Warm-tone light settings improved reconstruction result for most of lighting positions, while cool light performing the worst. There were no general trends for light intensity, the best light intensity setting differed across light colors and positions. For example, the dimmest light setting worked best for bottom lit warm-tone light, whereas the brightest light setting worked best for the top lit coo-tone light and medium intensity worked best for top lit neutral tone lighting.

Fig. 6 contains images taken in the worst and best lighting conditions. We can observe how lighting plays a significant role in how differently bodies may appear.

Skin-Tone It is a common goal to ensure there is no biases towards skin-tones in applications. Here we decided to vary the skin tones of our virtual human to see if there are differences in reconstruction results. We tested with 8 different skin tones, in 3 different background settings, a darker background, a lighter background and with random backgrounds from our background dataset. For each setting, we generated images from 2,500 camera poses.

Lighting Setting	Average Error (MPJPE/PA-MPJPE in mm)
Top Bright Neutral	305.79/136.95
Top Bright Warm	305.55/137.51
Top Mid Neutral	304.61/137.82
Top Dim Neutral	306.97/138.13
Global Dim Warm	307.23/138.50
Top Mid Warm	311.02/139.01
Global Mid Warm	314.79/139.54
Top Bright Cool	309.78/139.83
Top Dim Warm	316.23/140.47
Top Mid Cool	314.10/140.98
Top Dim Cool	315.43/141.25
Global Dim Neutral	321.95/141.50
Global Dim Cool	319.86/142.16
Global Bright Warm	327.21/142.32
Global Mid Neutral	325.27/143.76
Global Mid Cool	325.07/144.47
Global Bright Neutral	335.01/146.07
Side Dim Warm	323.37/146.17
Global Bright Cool	333.76/146.84
Side Mid Warm	326.01/147.00
Side Dim Neutral	327.99/147.74
Side Bright Warm	327.99/147.74
Side Dim Cool	336.14/148.98
Side Mid Neutral	331.53/149.07
Side Bright Neutral	335.38/149.30
Side Mid Cool	338.87/150.21
Side Bright Cool	342.32/151.29
Bottom Dim Warm	348.33/153.82
Bottom Mid Warm	353.07/155.84
Bottom Bright Warm	352.60/157.99
Bottom Dim Neutral	360.12/158.85
Bottom Mid Neutral	355.76/160.16
Bottom Dim Cool	365.59/162.11
Bottom Mid Cool	364.65/163.09
Bottom Bright Neutral	358.28/163.16
Bottom Bright Cool	367.62/166.22

Table 5: Reconstruction results for various lighting conditions sorted from best to worst by PA-MPJPE. The beest being Top, Bright, Neutral lighting while the worst being Bottom,Bright,Cool lighting.

Tone:	PA-MPJPE for Background 1: (lighter)	PA-MPJPE for Background 2: (darker)	PA-MPJPE for Random Backgrounds:
(36,16,13)	144.27	149.41	148.78
(64,28,23)	143.97	146.19	147.87
(82,36,30)	144.76	144.88	148.54
(113,65,55)	145.69	144.56	149.90
(161,92,51)	145.85	145.35	149.99
(201,134,91)	146.68	146.03	153.45
(237,188,157)	150.09	147.69	158.61
(255,236,227)	159.00	150.68	164.89

Figure 7: This table displays the reconstruction results of skin-toned virtual humans against 3 different background settings. On the left column is the tone and the RGB values of the tone.

From Table 7 one can see that the best reconstruction results came from the images with the highest contrast between skin-tone and background. With most indoor backgrounds having white or paler walls, we see improvements in reconstruction results with the darker or deeper skin tones when on lighter or random backgrounds. Darker backgrounds only benefit redder or more vibrant skin-tones as the contrast is greater there.

Body Size and Shape To test the changes in reconstruction results for varying body shapes and sizes, we modified the first 3 shape parameters of the SMPL model (Loper et al. 2015). These parameters control the size, weight and visceral fat of the human model respectively. We used 10 different bodies with images generated from 2500 camera positions.

Figure 8: This figure displays the worst and best performing body shape/size. On the left we have the worst performing body with an average PA-MPJPE of 169.48 mm while on the right we have the best performing body with an average PA-MPJPE of 130.06 mm. We found that changes in height had much more significant effects on reconstruction results over width/weight.

The main factor for change in performance was found to be height. Those bodies with the shortest height led to the best reconstruction results, while those bodies with the tallest height had the poorest reconstruction results. The difference in weight or width of the body only altered reconstruction results minimally, with a maximum variation of around 5%, while the difference in height led to a maximum variation of around 30%. Fig. 8 displays the worst and best performing bodies.

C Synthetic and Real-World Datasets

In order to discuss the performance of Human Body Reconstruction methods, we must ensure that there is accurate performances across all human bodies. To achieve this datasets for testing reconstruction methods should emulate the real world.

A commonly used dataset for testing human body reconstruction methods is via the 3D Poses in the Wild Dataset (3DPW) (von Marcard et al. 2018). The 3DPW test dataset covers many activities but does not offer a large enough variety of human bodies, with the test dataset only has 5 different actors and range of skin-tones is far more limited. Fig. 9 contains images from both the 3DPW dataset, as well as our

synthetic dataset; one can see there is much more diversity in shapes, sizes and skin tones of the virtual humans in our dataset.

While testing on real world data is imperative, synthetic datasets allow us to generate a wide variety of human data, that may have yet to be collected in a real world setting along with the specific 3D dimensions of the body. Thereby, synthetic data is a valuable addition into 3D human body reconstruction from images. This has also been demonstrated by several prior studies (Liang and Lin 2019; Smith et al. 2019b; Alldieck et al. 2019).

Additionally to speed up the real-world data collection process, we have developed a system with a robot arm to reproduce the method in which we collect data in the virtual scenes. This is described in Sec. 4.3 with additional information in Sec. D.

Figure 9: The top 2 rows of this figure displays images from the 3DPW test dataset, where there are only a few different human bodies throughout the whole dataset. The second 2 rows contains images from our synthetic dataset with a wider variety of human bodies in terms of skin tones, size and shape.

C.1 Qualitative Results on Real-World Data

Figure 10: This figure demonstrates the qualitative results of the baseline and those with the addition of SHARE on real world images taken of people outside the general BMI range. On the left is the images, in the middle is the original model and on the right is the same model with SHARE. One can see in these images SHARE improves shape and pose.

Figure 11: This figure demonstrates the qualitative results of the baseline and SHARE on images of a human in a non-canonical pose. From the rendered results one can see that SHARE creates a much more accurate representation of the real-world data.

Aside from the 3DPW dataset, most real world datasets also contain bodies of shapes and sizes within a smaller BMI range, this can neglect body types that lie outside of this range. Wider and Slimmer bodies should not be biased against in the learning-based algorithms. Through our synthetic datasets, SHARE improves reconstruction on these bodies as seen in Fig. 10 and poses as seen in Fig. 11.

D Robot-Arm Data Collection

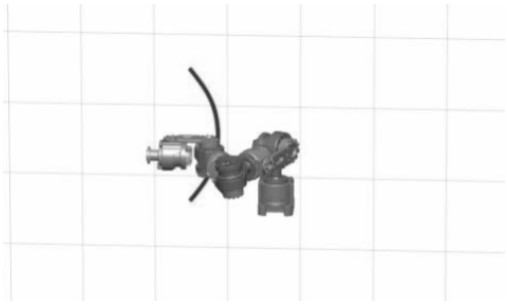

Figure 12: **The trajectory of the AMBER B1 robot arm.** Shown here is the configuration of the arm at one of the positions in Matlab environment, with the specifications of the real-world robot in Figure 4.

Figure 13: **A sample of the different images in our robot-arm collected real-world data**. Images of our set-up can be seen in Figure 4 We demonstrate the wide variety of camera orientations, as well as positions with respect to the body.

For our real world data collection, we utilized a robot arm for precise-control of the camera position and orientation. Our research is focused on improving robustness against ubiquitous perturbations, especially those that may be created, when the everyday user is the one taking the images for body reconstruction.

To emulate this process, we utilize a Pixel 4a as our "camera" since many people use cameras from their mobile devices. As mentioned in Sec. 4.3, we use the Amber B1 robot arm from Amber Robotics, which has 7 axis of movement and a 581 mm reach (Robotics 2022). Using the specifications of the robot and its respective Unified Robotics Description Format (URDF) files, we are able to map the joint positions of the robot using inverse kinematics.

Using Matlab to perform the inverse kinematics, we plotted a trajectory for the robot arm. This trajectory and one of the robot arm positions can be seen in Fig. 12.

For our current real-world dataset, the robot arm stops at 8 positions, equidistant from one another along this trajectory. Equipped with a phone holding attachment, the robot arm can then act as a real-world stand in for the human-user's camera positions.

The Amber b1 robot takes in arrays of positions along with time for the time to move and stay in that position. We created an android application for the Pixel 4a to accept the time stamps for the robot movements and to collect images from each position the robot moves to. This setup, combined with the 360-degree rotation capabilities of the turn-table, allows us to curate a real world data set of human body images from a variety of camera poses in a precise manner. For the current dataset, we collected images from all 8 positions for every 22.5° or $2\pi/16$ interval.

Fig. 13 shows several images taken using our setup from different camera positions and orientations with reference to the body. We can see how different camera orientations affect the appearance of the limbs, while the different body-to-camera positions occlude different parts of the body.

E Social Impact Assessment

Human body reconstruction can be used for many applications and can greatly improve engagement among users, due to a life-like avatar of themselves. However, we hypothesize that the further away the reconstruction results are from the true body specifications, the more likely it is to met with a dissatisfied user response. SHARE can provide more accurate and robust results towards ubiquitous perturbations, created by user data collection and can potentially minimize the issue of dissatisfaction. For applications such as virtual try-on or virtual health evaluations, the most accurate and robust portrayal of human body shape and pose is need, and SHARE helps to offer improved results.