Synthetic Training for Accurate 3D Human Pose and Shape Estimation in the Wild

Akash Sengupta
as2562@cam.ac.uk
Ignas Budvytis
ib255@cam.ac.uk
Roberto Cipolla
rc10001@cam.ac.uk

Department of Engineering
University of Cambridge
Cambridge, UK

Abstract

This paper addresses the problem of monocular 3D human shape and pose estimation from an RGB image. Despite great progress in this field in terms of pose prediction accuracy, state-of-the-art methods often predict inaccurate body shapes. We suggest that this is primarily due to the scarcity of *in-the-wild* training data with *diverse and accurate* body shape labels. Thus, we propose STRAPS (Synthetic Training for Real Accurate Pose and Shape), a system that utilises proxy representations, such as silhouettes and 2D joints, as inputs to a shape and pose regression neural network, which is trained with synthetic training data (generated on-the-fly during training using the SMPL statistical body model) to overcome data scarcity. We bridge the gap between synthetic training inputs and noisy real inputs, which are predicted by keypoint detection and segmentation CNNs at test-time, by using data augmentation and corruption during training. In order to evaluate our approach, we curate and provide a challenging evaluation dataset for monocular human shape estimation, Sports Shape and Pose 3D (SSP-3D). It consists of RGB images of tightly-clothed sports-persons with a variety of body shapes and corresponding pseudo-ground-truth SMPL shape and pose parameters, obtained via multi-frame optimisation. We show that STRAPS outperforms other state-of-the-art methods on SSP-3D in terms of shape prediction accuracy, while remaining competitive with the state-of-the-art on pose-centric datasets and metrics.

1 Introduction

3D human shape and pose estimation from a single RGB image is a challenging computer vision problem, with widespread applications in computer animation and augmented reality. Recently, several deep-learning-based methods have been proposed [8, 15, 16, 22, 24, 27, 29, 32, 34]. Such methods provide impressive 3D pose reconstructions given single RGB images as inputs, by leveraging datasets of images of humans in a diverse range of labelled 3D poses [3, 7, 17, 21, 28, 30]. However, these approaches often predict inaccurate body shapes, as shown in Figure 1. We suggest that this is due to a lack of body shape diversity within the prevalent training datasets. Most learning-based models will struggle to generalise to unseen test data if the distribution of the test data is significantly different from the training

Figure 1: **STRAPS predicts body shapes with greater accuracy than other approaches** to monocular human 3D shape and pose estimation, such as SPIN [15], CMR [16] and HMR [8], without requiring training images annotated with 3D labels. The images shown in this figure are part of the dataset we provide, SSP-3D.

data distribution. Thus, increasingly inaccurate body shapes are predicted as the shape of the test subject is further removed from the training datasets' mean shape.

In this paper, we present **S**ynthetic **T**raining for **R**eal **A**ccurate **P**ose and **S**hape (STRAPS), a deep-learning-based framework that uses synthetic training data to overcome the lack of shape diversity in current datasets. Given an input image, inference occurs in two stages (see Figure 2a). First, we predict a proxy representation, which encodes the subject's silhouette and 2D joint locations, using off-the-shelf segmentation and 2D keypoint detection CNNs [4, 6, 31]. Then, we use a neural network regressor to predict the parameters of a statistical body model (SMPL [20]) from the proxy representation. The regressor is trained using synthetic input-target pairs generated on-the-fly during model training. This is done by sampling target SMPL shape and pose parameters from a training distribution and rendering the corresponding silhouettes and 2D joints, which act as inputs. Since we can choose the form of the training distribution, we have control over the diversity of human body shapes seen by the regressor during training. We utilise simple data augmentation and corruption techniques (see Figure 2a) to make our regressor robust to noisy inputs encountered at test-time.

Moreover, we curate and provide Sports Shape and Pose 3D (SSP-3D), a dataset which contains images of tightly-clothed sports-persons with a diverse range of body shapes in varied environments, obtained from the Sports-1M video dataset [9]. We use multi-frame optimisation, with forced shape consistency between frames, to obtain pseudo-ground-truth SMPL shape and pose parameters for the sports-person in each image. We evaluate our neural network regressor, along with several recent learning-based approaches, on SSP-3D and report shape prediction accuracy in terms of per-vertex Euclidean error in a neutral pose. Examples from SSP-3D are shown in Figure 3, along with statistics illustrating the greater body shape diversity in SSP-3D compared to widely-used 3D human datasets.

In summary, we have two main contributions: (i) a deep-learning framework which uses synthetic training data and simple data augmentation techniques to overcome the lack of body shape diversity within prevalent datasets and (ii) the SSP-3D evaluation dataset, which we use to show that our neural network regressor results in better shape prediction accuracy than other competing approaches. Our code and dataset are available for research purposes at https://github.com/akashsengupta1997/STRAPS-3DHumanShapePose.

2 Related Work

In this section, we discuss recent approaches to 3D human pose and shape estimation, as well as the training datasets typically used for this task.

Monocular 3D human pose and shape estimation approaches can be classified into two paradigms: optimisation-based and learning-based. Optimisation-based approaches attempt to fit a parametric body model [1, 20, 25] to 2D observations, such as 2D joints [2, 25], body surface landmarks [17], silhouettes [17] or body part segmentations [33]. These approaches produce reliable results without requiring 3D-labelled datasets, which are expensive to obtain. However, they are slow at test-time, sensitive to initialisation and can get stuck in bad local minima, which motivates learning-based approaches.

Learning-based approaches can be further divided into two types: non-parametric 3D regression and body model parameter regression. Non-parametric 3D regression involves predicting a 3D human body representation from an image, such as a voxel occupancy grid [29] or vertex mesh [16]. However, each representation has associated drawbacks for body shape prediction: *e.g.* voxels are limited by the resolution of the voxel grid and direct mesh predictions can result in surface artifacts such as wrinkles and sharp protrusions. Body model parameter regression involves predicting the parameters of a statistical body model, such as SMPL [20], which provides a useful prior over body shape. Several approaches first predict a proxy representation from the input RGB image, such as surface keypoints [17, 24], silhouettes [24, 27], body part segmentations [22] or IUV maps [32, 34], and use this representation as the input to a regressor. Other approaches directly predict body model parameters from the input image [8, 15, 26]. Fundamentally, learning-based approaches are dependent on the label accuracy and sample diversity of the training datasets used. This results in a significant drawback when the training data distribution is not sufficiently diverse in terms of body shape, pose and image (*e.g.* background) conditions, as discussed below.

3D human pose and shape datasets. Learning-based approaches are trained using datasets of images paired with labels in the form of 3D joints or body model parameters. 3D labels may be obtained using motion capture (as for Human3.6M [7] and BML MoVi [3]), using inertial motion units (as for 3DPW [30]), or by optimisation (as for UP3D [17]). While current datasets contain varied and accurate 3D poses, they all suffer from limited body shape diversity, which greatly hampers the shape prediction accuracy of learning-based approaches. Additional drawbacks include: baggy clothing obscuring body shape and data captured in indoor MoCap environments being unrepresentative of in-the-wild images. We overcome these limitations of current training datasets by using synthetic training data. Furthermore, we create our own in-the-wild dataset, SSP-3D, to evaluate monocular 3D body shape predictions, which contains subjects with a greater variety of body shapes than current datasets.

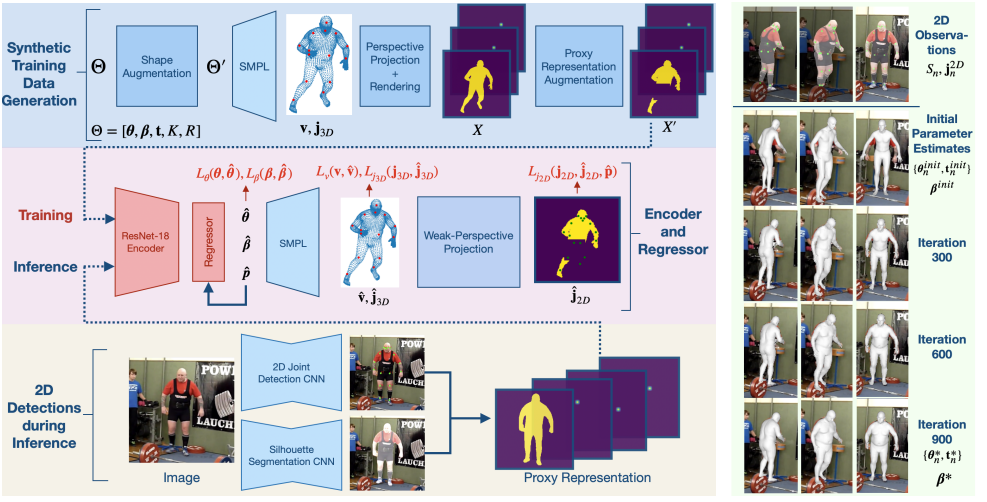
3 Method

In this section we describe STRAPS, our framework which utilises synthetic training data to overcome the lack of body shape diversity in real datasets. We also detail the multi-frame optimisation procedure used to create SSP-3D.

3.1 STRAPS

The proposed synthetic training process has two parts: synthetic data generation and neural encoder and regressor training, both of which use a parametric 3D body model.

Parametric 3D body model. The SMPL [20] body model provides a fully-differentiable function $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta})$ that takes shape-space coefficients $\boldsymbol{\beta}$ and 3D joint rotations $\boldsymbol{\theta}$ as inputs and outputs a human vertex mesh $\mathbf{v} \in \mathbb{R}^{N\times 3}$. 3D joint locations are obtained as a linear



(a) STRAPS Training and Inference (b) SSP-3D Optimisation

Figure 2: **(a) Overview of the training and inference pipelines for STRAPS and (b) optimisation pipeline for SSP-3D.** Synthetic training data is generated by sampling SMPL [20] pose and shape parameters and decoding them into 3D vertices and joints, which are projected, rendered and corrupted to form an input proxy representation. The proxy representation is passed through an encoder and iterative regressor (both with trainable weights) that predicts pose, shape and camera parameters. Supervision signals are applied to SMPL parameters, 3D joints and vertices, and 2D joints. At test-time, off-the-shelf detection and segmentation CNNs are used to create the input proxy representation. Optimisation for SSP-3D involves fitting SMPL to multi-frame 2D joints and silhouettes of a subject, which yields optimised pose and camera parameters for each frame and shape parameters for the subject.

combination of the vertices, $\mathbf{j}^{3D} = J\mathbf{v}$, where $J \in \mathbb{R}^{L\times N}$ is a regression matrix for L joints of interest.

Synthetic data generation. SMPL is used to generate training data on-the-fly (top of Figure 2a). In each iteration of the training loop, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are sampled from any training dataset with SMPL parameters - paired images are not required. A camera translation vector $\mathbf{t}$ is sampled randomly, while camera intrinsics and rotation matrices, K and R, are fixed. To combat the insufficient body shape diversity in prevalent training datasets, we perform *body shape augmentation* by replacing $\boldsymbol{\beta}$ with a new random vector $\boldsymbol{\beta}'$, generated by sampling each shape parameter $\beta'_n \sim \mathcal{N}(\mu, \sigma_n^2)$, where σ_n is chosen (empirically) to provide greater body shape variance than current datasets. Then, the 3D vertices $\mathbf{v}$ and 3D joints $\mathbf{j}^{3D}$ corresponding to $\boldsymbol{\theta}$ and $\boldsymbol{\beta}'$ are perspective-projected and rendered [10] into a silhouette $S \in [0,1]^{H\times W}$ and 2D joint locations $\mathbf{j}^{2D} \in \mathbb{R}^{L\times 2}$. $\mathbf{j}^{2D}$ is transformed into 2D Gaussian joint heatmaps, $G \in \mathbb{R}^{H\times W\times L}$, where each channel corresponds to a separate joint location. We obtain our clean synthetic proxy representation (PR), $X \in \mathbb{R}^{H\times W\times(L+1)}$ by concatenating S and G along the channel dimension. Note that we opt for simple silhouettes and 2D joints as our PR, instead of more complex part segmentations or IUV maps [9], because the synthetic-to-real domain gap is smaller for a simple representation, and can be more easily bridged with *proxy representation augmentation* during training. This involves modelling the failure modes of

the off-the-shelf detection and segmentation CNNs used at test-time. In particular, noisy keypoint and silhouette predictions are modelled by adding uniform random noise to the 2D joint centres and silhouette edges in X. Occlusion is modelled by randomly removing body parts from and adding occluding boxes to the silhouette in X. The augmented PR X' serves as the training input to our neural encoder. The training labels consist of $\boldsymbol{\theta}, \boldsymbol{\beta}', \mathbf{v}, \mathbf{j}_{3D}$ and $\mathbf{j}_{2D}$.

Neural encoder and regressor. STRAPS is architecture-agnostic. For this paper, we use the same network architecture as [8, 15], which consists of a convolutional encoder for feature extraction and an iterative regressor that outputs predicted SMPL pose, shape and camera parameters ($\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{p}}$) given these features. Note that [8, 15] implement additional modules, namely an adversarial prior [8] or "in-the-loop" optimisation [15], necessitated by their use of training images with only 2D joint labels. However, 2D joints crucially fail to supervise 3D shape, unlike our strong 3D supervision, which also does not require such additional modules. Weak-perspective camera parameters are predicted during regression, represented by $\hat{\mathbf{p}} = [\hat{s}, \hat{\mathbf{t}}]$, where $s \in \mathbb{R}$ represents scale and $\hat{\mathbf{t}} \in \mathbb{R}^2$ represents x-y camera translation. We use a continuous 6-dimensional rotation representation for $\hat{\boldsymbol{\theta}}$, as proposed by [35], instead of the discontinuous Euler rotation vectors used by SMPL as default. From $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$, SMPL is used to obtain predicted 3D vertices and joints, $\hat{\mathbf{v}}$ and $\hat{\mathbf{j}}^{3D}$. Finally, projected 2D joint predictions are obtained by $\hat{\mathbf{j}}^{2D} = s\Pi(\hat{\mathbf{j}}^{3D} + \hat{\mathbf{t}})$, where Π represents an orthographic projection.

We train our network using a combination of 5 loss functions in a highly-multi-task framework. We use homoscedastic uncertainty [11] to adaptively learn the loss weights during training, which results in an objective function of the form

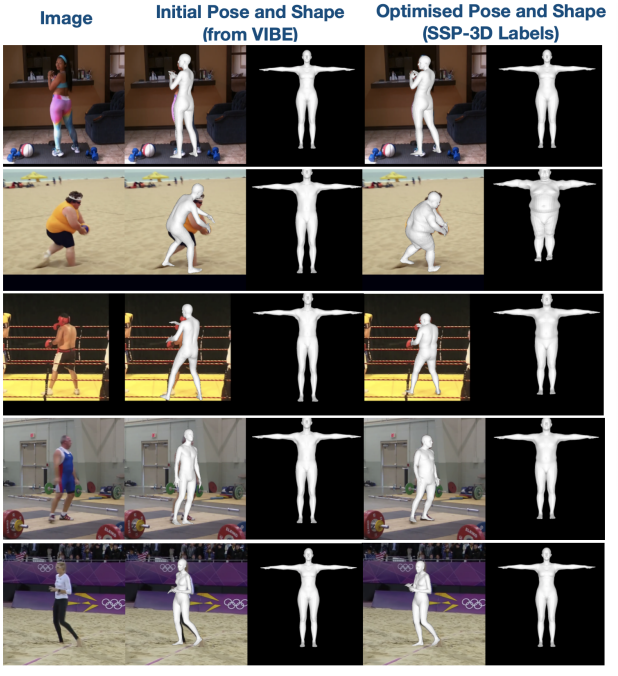
$$L = \frac{1}{\sigma_\beta^2} L_\beta + \frac{1}{\sigma_\theta^2} L_\theta + \frac{1}{\sigma_v^2} L_v + \frac{1}{\sigma_{j_{3D}}^2} L_{j_{3D}} + \frac{1}{\sigma_{j_{2D}}^2} L_{j_{2D}} + \log(\sigma_\beta \sigma_\theta \sigma_v \sigma_{j_{3D}} \sigma_{j_{2D}}), \quad (1)$$

where the σ^2 terms represent task uncertainties and the L terms represent mean squared error losses. Empirically, we found that redundancy in the multi-task objective - *e.g.* applying losses on SMPL parameters as well as on 3D vertices, despite $\mathbf{v}$ being fully determined by $(\boldsymbol{\theta}, \boldsymbol{\beta})$ - improved both network convergence and final performance. We hypothesise that this is because each supervision signal has a different granularity. For instance, a loss on vertices provides a finer-scale supervision signal than a loss on 3D joints. Thus, we apply mean squared error losses on 3D joints ($L_{j_{3D}}$), 3D vertices (L_v), SMPL pose parameters in the 6D rotation representation of [35] (L_θ), and SMPL shape coefficients (L_β). We employ an additional loss on projected 2D joints ($L_{j_{2D}}$) to enforce image-model alignment.

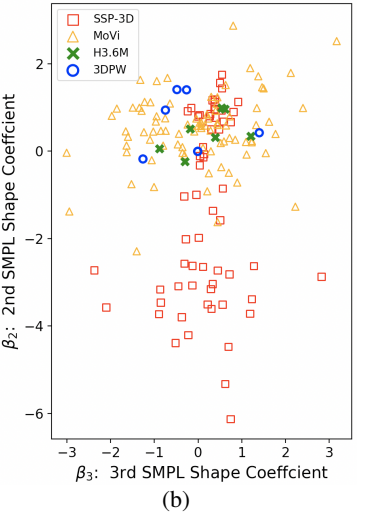
3.2 SSP-3D

SSP-3D contains 311 in-the-wild images of 62 tightly-clothed sports-persons (selected from the Sports-1M video dataset [9]) with a diverse range of body shapes, along with corresponding pseudo-ground-truth SMPL shape and pose labels. Figure 3 illustrates the greater body shape diversity in SSP-3D compared to Human3.6M [7], 3DPW [30] and MoVi [3]. Note that Human3.6M and MoVi are not in-the-wild and have homogeneous backgrounds.

SMPL shape and pose labels were acquired via optimisation, using an extended version of SMPLify [2] in a similar manner to the UP-3D dataset [17]. Unlike [2] and [17], we used multiple frames of the same subject in parallel, obtaining different optimised poses $\boldsymbol{\theta}_n^*$ and camera translations $\mathbf{t}_n^*$ for each frame n, but *forcing the optimised shape* $\boldsymbol{\beta}^*$ *to be the same across all frames* to exploit multi-view information. We used 2D joints, acquired using Keypoint-RCNN [6], and pixel-accurate silhouettes, acquired using PointRend [13] on top of FPN [18], as the target 2D observations into which we fit the SMPL model. The use of multi-frame silhouettes ensures that SSP-3D has significantly more accurate body shape labels than



(a)



(b)

Dataset	No. of Subjects	β_2 Var	β_3 Var
SSP-3D	62	4.44	0.52
H3.6M [21]	7	0.17	0.41
3DPW [50]	7	0.39	0.49
MoVi [6]	86	0.68	1.40

(c)

Figure 3: **SSP-3D samples and statistics.** (a) shows RGB images and corresponding optimised SMPL body shape and pose labels provided in the SSP-3D dataset. It also illustrates the improvement in shape and pose parameter estimates after optimisation, compared to initial estimates obtained with VIBE [24]. (b) and (c) demonstrate the greater body shape diversity in SSP-3D compared to widely-used datasets, by considering the distribution of the 2nd and 3rd shape coefficient labels β_2 and β_3. (b) plots β_2 and β_3 for each sample in each dataset. (c) gives the number of subjects and the variance of β_2 and β_3 labels in each dataset. Note that β_2 is strongly correlated with variation in body fat content. β_1 is not used because it is strongly correlated with overall size, which is ambiguous in monocular predictions.

UP-3D. To prevent getting stuck in bad local minima, we obtained an initialisation for per-frame pose and camera parameters, $\boldsymbol{\theta}_n^{\text{init}}$ and $\mathbf{t}_n^{\text{init}}$, and shape coefficients, $\boldsymbol{\beta}^{\text{init}}$, by using VIBE [24], a method for SMPL prediction from video (see Figure 3a). Suitable frames for optimisation, with good SMPL initialisations and accurate target silhouettes and 2D joints, were hand-picked by human annotators. Our objective function is the sum of 6 error terms:

$$E(\boldsymbol{\beta}, \{\boldsymbol{\theta}_n, \mathbf{t}_n\}_{n=1}^N) = \lambda_j E_j + \lambda_S E_S + \lambda_a E_a + \lambda_\theta E_\theta + \lambda_\beta E_\beta + \lambda_{\theta^{\text{init}}} E_{\theta^{\text{init}}}, \tag{2}$$

where N is the number of frames and the λ terms represent weights. The silhouette error term, E_S, penalises the L_1 difference between target and SMPL silhouettes. It is defined as

$$E_S(\boldsymbol{\beta}, \{\boldsymbol{\theta}_n, \mathbf{t}_n\}_{n=1}^N; \{S_n\}_{n=1}^N) = \frac{1}{N}\sum_{n=1}^{N} \frac{\|\hat{S}(\Pi_K(\hat{\mathbf{v}}_n(\boldsymbol{\beta}, \boldsymbol{\theta}_n) + \mathbf{t}_n)) - S_n\|_1}{WH}, \tag{3}$$

where W, H are the width and height of the target silhouettes S_n and $\hat{\mathbf{v}}_n(\boldsymbol{\beta}, \boldsymbol{\theta}_n)$ represents the SMPL vertices for the n-th frame. $\Pi_K()$ is a perspective-projection with intrinsic camera parameters K. Neural Mesh Renderer [10] is used to differentiably render SMPL silhouettes $\hat{S}_n$ from projected vertices.

$E_{\theta^{\text{init}}}$ is a pose regularisation term, which penalises the L_2 distance between the current and initial estimates (from VIBE [14]) of the SMPL pose parameters in rotation matrix form. We observed that the optimiser would use perspective effects to fit the SMPL model to target silhouettes of large persons, instead of updating the shape parameters. For example, the global rotation parameters would be updated to make the SMPL body lean towards the camera, enlarging the rendered silhouette. $E_{\theta^{\text{init}}}$ was incorporated to prevent such effects. It is defined as

$$E_{\theta^{\text{init}}}(\{\boldsymbol{\theta}_n\}_{n=1}^N;\{\boldsymbol{\theta}_n^{\text{init}}\}_{n=1}^N) = \frac{1}{N}\sum_{n=1}^{N}\frac{\|\mathbf{r}(\boldsymbol{\theta}_n)-\mathbf{r}(\boldsymbol{\theta}_n^{\text{init}})\|_2^2}{|\mathbf{r}(\boldsymbol{\theta}_n)|}, \tag{4}$$

where $\mathbf{r}(\boldsymbol{\theta}_n)\in\mathbb{R}^{216}$ represents the vector of flattened and concatenated rotation matrices (for each of the 24 SMPL joints) corresponding to $\boldsymbol{\theta}_n$.

E_j, E_a, E_θ and E_β are derived from SMPLify and full definitions can be found in [2]. In short, E_j is a weighted 2D joint reprojection error, E_a is an angle prior term which penalises unnatural bending of the elbow and knees, E_θ is the negative log-likelihood of a Gaussian mixture model pose prior and E_β is a L_2 regularisation penalty upon shape parameters.

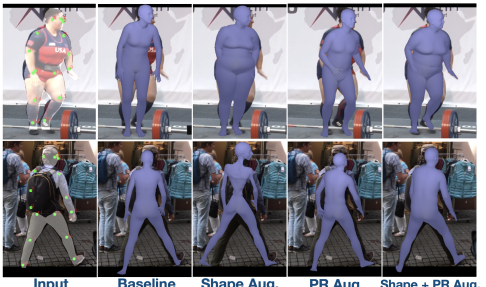
We optimise our objective function using the Adam [12] optimiser with a learning rate of 0.01. After convergence, a human annotator selects good SMPL fits. Details on the human annotation, as well as all hyperparameter values, are available in the supplementary material.

4 Implementation Details

Training datasets. To generate synthetic training data, we sample SMPL pose parameters from the training sets of UP-3D [17] and 3DPW [30], and from Human3.6M [7] subjects S1, S5, S6, S7 and S8 (after applying MoSh [19] to obtain SMPL poses from 3D joint labels). For our baseline experiments without shape augmentation, we also use the SMPL shape parameters from these datasets. Synthetic silhouettes are cropped and resized to 256×256.
Evaluation datasets. We report metrics on Human3.6M (Protocol 2 [8] subjects S9, S11), 3DPW (test set), BML-MoVi [3] (F-PG1 videos) and SSP-3D. For Human3.6M and 3DPW, we report mean per joint position error after rigid alignment with Procrustes analysis (MPJPE-PA [21]). For BML-MoVi and SSP-3D, we report scale-corrected per-vertex Euclidean error in a neutral pose (or T-pose), i.e. PVE-T-SC. A description of the scale-correction technique used to combat scale ambiguity is given in the supplementary material. We also report silhouette mean intersection-over-union on SSP-3D.
Architecture. We use a ResNet-18 [5] encoder, the output of which is average pooled, producing a feature vector $\boldsymbol{\phi}\in\mathbb{R}^{512}$. The iterative regression network consists of two fully connected layers with 512 neurons each, followed by an output layer with 157 neurons. We use the Adam [12] optimiser to train our encoder and regressor, with a learning rate of 0.0001 and a batch size of 140. We train for 240 epochs, which takes 5 days on a single 2080Ti GPU. During inference, 2D joint predictions are obtained using Keypoint-RCNN [6] and silhouette predictions are obtained using DensePose [9]. All implementations are in PyTorch [23]. Inference runs at $\sim$4fps, 90% of which is silhouette and joint prediction.

5 Empirical Evaluation

In this section, we present results from our ablative study, which investigates the effects of shape and proxy representation augmentation during synthetic training. We also compare our method to other approaches in terms of shape and pose accuracy.

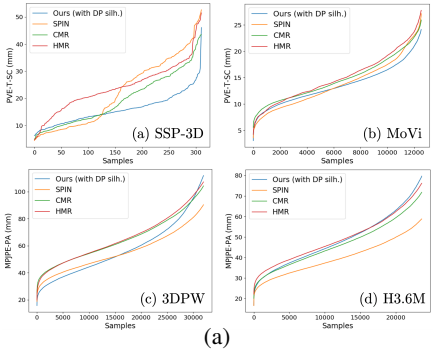


(a)

Input	Augmentation	SSP-3D PVE-T-SC	H3.6M MPJPE-PA	3DPW MPJPE-PA
GT Synthetic	Baseline	14.4	40.4	39.0
	Shape aug.	10.1	34.1	**34.9**
	PR aug.	16.1	42.2	44.7
	Shape+PR aug.	**10.0**	**33.1**	37.6
DP + KPRCNN	Baseline	20.1	70.5	71.3
	Shape aug.	24.1	75.5	88.6
	PR aug.	18.9	61.0	69.9
	Shape+PR aug.	**15.9**	**55.4**	**66.8**

(b)

Figure 4: **Ablation study.** (a) illustrates that applying shape and proxy representation (PR) augmentation improves predictions of non-typical body shapes and develops robustness against noisy inputs. (b) reports pose (MPJPE-PA in mm) and shape (PVE-T-SC in mm) metrics when using synthetic proxy representations (rendered from ground-truth SMPL labels) versus predicted proxy representations (from DensePose and Keypoint-RCNN) as inputs.



(a)

Method	SSP-3D PVE-T-SC	SSP-3D mIOU	MoVi PVE-T-SC	H3.6M MPJPE-PA	3DPW MPJPE-PA
Ours (baseline)	20.1	0.62	**13.2**	70.5	71.3
Pavlakos *et al.* [24]	-	-	-	75.9	-
HMR (unpaired) [8]	20.8*	0.61*	14.2*	66.5	-
SPIN (unpaired) [20]	-	-	-	62.0	-
Ours	**15.9**	**0.80**	**14.0**	55.4	66.8
HMR [8]	22.9*	0.69*	15.5*	56.8	71.5*
NBF [23]	20.9*	-	14.4*	59.9	90.7*
CMR [14]	19.5*	0.68*	15.2*	50.1	70.3*
SPIN [20]	22.2*	0.70*	14.3*	**41.1**	**59.2**

(b)

Figure 5: **Quantitative comparison with the SOTA** on SSP-3D, MoVi, Human3.6M (Protocol 2) and 3DPW. We report PVE-T-SC (mm) and mIOU on shape-centric datasets SSP-3D and MoVi and MPJPE-PA (mm) on pose-centric datasets 3DPW and Human3.6M. (a) plots the (sorted) distributions of metrics per evaluation sample. (b) lists mean metrics over all samples. Methods in the top part of (b) do not require training data comprised of images paired with 3D ground truth, while methods in the bottom part do. Numbers marked with * were evaluated for this paper, all other numbers are reported by the respective papers.

Ablation studies. Our ablative study investigates the effects of shape and proxy representation (PR) augmentation applied during synthetic training. We compare four networks, trained with: (i) no augmentation (baseline), (ii) only shape augmentation, (iii) only PR augmentation and (iv) shape + PR augmentation. Evaluations are carried out with two types of input proxy representations: synthetic silhouettes and 2D joints generated from GT SMPL labels and "real" silhouettes and 2D joints predicted from test RGB images using DensePose [4] and Keypoint-RCNN [6] respectively. SSP-3D evaluates 3D shape prediction across a diverse range of body shapes, while Human 3.6M and 3DPW evaluate 3D pose prediction.

The quantitative performance of the baseline network on GT synthetic inputs (see Figure 4b first row) motivates the use of synthetic training data, since, in this ideal case, it achieves greater than SOTA accuracy (compare with Figure 5b). However, in the practically-applicable situation using "real" inputs, the baseline network has two key failure modes:

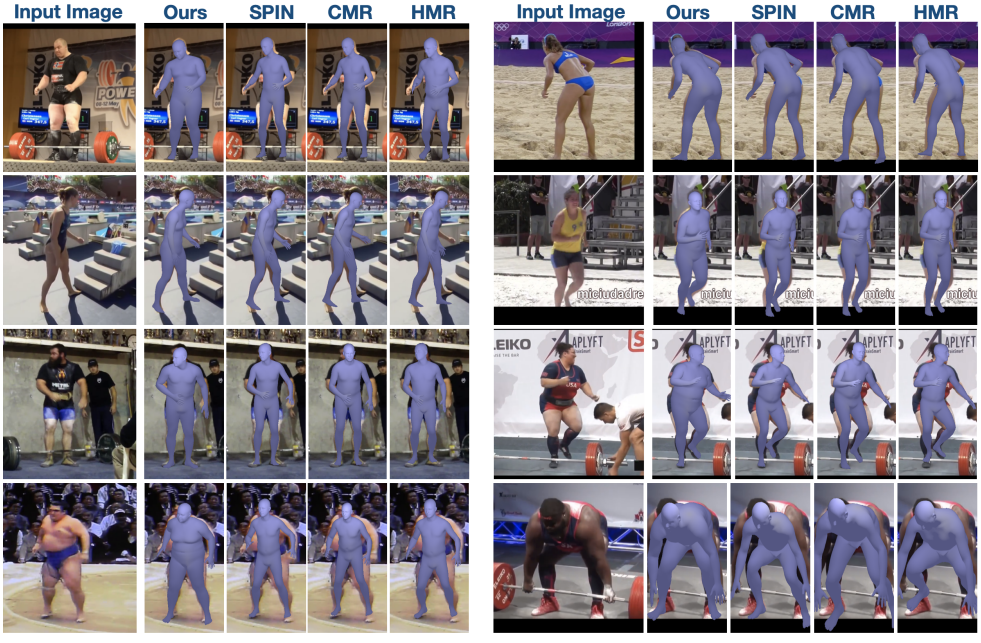


Figure 6: **Qualitative comparison on SSP-3D.** Each row shows examples from different PVE-T-SC quartiles (for our method), top to bottom: 0-25%, 25-50%, 50-75%, 75-100%. Results from SPIN [15], CMR [16] and HMR [8] are shown for comparison. Our method is able to accurately predict a diverse range of body shapes, whereas other approaches are biased towards an average body shape prediction.

firstly, the predicted body shape is inaccurate, particularly for non-typical subjects (see Figure 4a, top row) and secondly, the network becomes reliant on the perfectly-rendered synthetic inputs and is unable to deal with noisy real inputs. Incorporating shape augmentation alleviates the first problem, since the network sees a greater variety of shapes during training. However, the second problem is greatly exacerbated, particularly in cases with occluded silhouettes (see Figure 4a, bottom row). Hence, the network trained with shape augmentation results in better shape (and pose) metrics on synthetic inputs compared to the baseline, as shown in Figure 4b, while the metrics on real inputs are poor. Incorporating PR augmentation shrinks the performance deterioration when using real versus synthetic inputs by explicitly modelling input noise and occlusion during the synthetic training process. By combining PR and shape augmentation, we are able to predict a diverse range of body shapes, improve our pose accuracy significantly over the baseline and produce semantically-plausible outputs on all datasets, even when the input is heavily corrupted (see Figure 4a).

Comparison with the state-of-the-art. Our method, with shape and PR augmentation, surpasses the state-of-the-art in terms of PVE-T-SC and mIOU on SSP-3D and MoVi. The distribution of errors per SSP-3D sample, shown in Figure 5a, suggests that our method is able to maintain shape prediction accuracy for challenging evaluation samples while the performance of competing approaches degrades for samples featuring non-average body shapes, qualitative examples of which are given in Figure 6. Our method may give erroneous reconstructions for outlier body shapes, in which case DensePose fails to predict an accurate silhouette, or due to poses with substantial self-occlusion, as shown in Figure 6 bottom row.

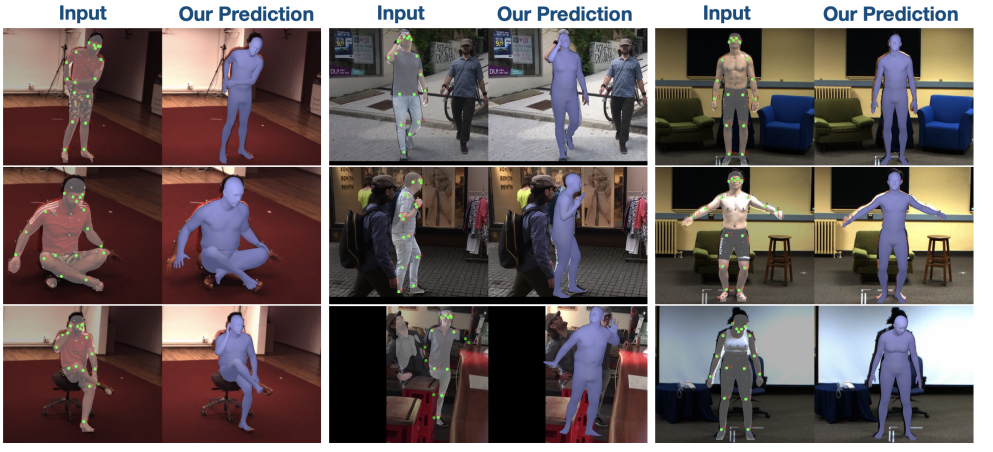


Figure 7: **Qualitative results on Human3.6M (left), 3DPW (middle) and MoVi (right).** Each row shows examples from different error metric quantiles (MPJPE-PA for H3.6M and 3DPW, PVE-T-SC for MoVi). Top to bottom: 0-33%, 33-66%, 66-100%. Input silhouettes and input 2D keypoints are visualised over each image.

Although we focus on shape prediction, our method is competitive with the SOTA on H3.6M and 3DPW in terms of MPJPE-PA and outperforms other methods that do not require training data comprised of images paired with expensive-to-obtain 3D labels. Qualitative examples are given in Figure 7. We observe that we perform relatively better on 3DPW than H3.6M, as compared to other methods, particularly up to the median error (Figure 5a). This is because methods trained on images captured in an indoor MoCap environment (like H3.6M) do not maintain the same pose prediction accuracy for test images with unconstrained background and lighting conditions (like in 3DPW). We create our input proxy representation using 2D segmentation and detection CNNs, which are more easily trained to be invariant to such variables. Thus, we match or surpass the SOTA on 3DPW for samples up to the median MPJPE-PA. However, 3DPW contains samples with severe occlusion (beyond what is modelled by PR augmentation) and overlapping persons, which cause DensePose to predict erroneous silhouettes and results in worse MPJPE-PA in the 75-100% quartile.

6 Conclusion

In this paper, we addressed the problem of monocular 3D human shape and pose estimation. In particular, we observed that current approaches often predict inaccurate body shapes, particularly for non-typical subjects, due to a lack of body shape diversity in prevalent 3D human datasets. Thus, we proposed STRAPS, a learning framework that overcomes the lack of diversity by generating synthetic training data with diverse body shapes on-the-fly, such that the regressor sees a new body shape at every training iteration. To evaluate our approach, we created a challenging evaluation dataset for monocular human shape estimation, SSP-3D, which consists of RGB images of tightly-clothed sports-persons with a variety of body shapes and corresponding pseudo-ground-truth SMPL shape and pose parameters. We showed that STRAPS outperforms other approaches on SSP-3D in terms of shape prediction accuracy, while remaining competitive with the state-of-the-art on pose-centric datasets.

References

[1] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape completion and animation of people. In *ACM Transactions on Graphics (TOG) - Proceedings of SIGGRAPH*, volume 24, pages 408–416, 2005.

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

[3] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F. Troje. MoVi: A Large Multipurpose Motion and Video Dataset, 2020. URL https://doi.org/10.5683/SP2/JRHDRN.

[4] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[7] 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 (PAMI)*, 36 (7):1325–1339, July 2014.

[8] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[9] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[10] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3D mesh renderer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[11] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[13] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. PointRend: Image segmentation as rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[14] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

[16] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[17] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the People: Closing the loop between 3D and 2D human representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] Matthew Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. In *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH Asia*, volume 33, pages 220:1–220:13. ACM, 2014.

[20] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH Asia*, volume 34, pages 248:1–248:16. ACM, 2015.

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

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

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035. Curran Associates, Inc., 2019.

[24] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[25] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[26] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3D human recovery in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[27] Vince J. K. Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3D human shape and pose prediction. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[28] Gül 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 (CVPR)*, 2017.

[29] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[30] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[31] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

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

[33] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[34] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. Danet: Decompose-and-aggregate network for 3D human shape and pose estimation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 935–944, 2019.

[35] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.