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

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

This document provides supplementary material for the 3D human shape and pose prediction framework, STRAPS, and the 3D shape evaluation dataset, SSP-3D, presented in the main manuscript. Section 1 focuses on SSP-3D. We quantitatively compare pre- and post-optimisation SMPL [6] body model fits (where the latter forms the pseudo-ground-truth 3D labels in SSP-3D) and explain the role of human annotators in the dataset creation process. In Section 2, we formally define the PVE-T-SC metric, introduced in the main manuscript and used to measure 3D body shape prediction error. We also qualitatively compare 3D reconstructions obtained using ground-truth silhouettes versus silhouettes predicted by DensePose [10] as inputs to our SMPL regressor. Finally, relevant hyperparameters for both STRAPS and SSP-3D dataset optimisation are provided in in Table 2.

1 SSP-3D Dataset

In the main manuscript, we present the Sports Shape and Pose 3D (SSP-3D) dataset, which contains images of tightly-clothed sportspersons with a variety of body shapes. Pseudo-ground-truth SMPL [6] shape and pose parameters are obtained by fitting the SMPL body model to target 2D observations (joints and silhouettes) from multiple views, while forcing the shape parameters to be consistent across views. To prevent getting stuck in bad local minima, we initialised the pose, shape and camera parameters with estimates from VIBE [4], a method for SMPL prediction from video. Table 1 reports error metrics between target 2D observations and SMPL body model estimates, both before the optimisation process (i.e. using the initial VIBE estimates) and after the optimisation process.

1.1 Human annotation

A human annotator was involved in the creation of SSP-3D in two stages: pre-optimisation and post-optimisation.

Pre-optimisation annotation involved selecting frames where the silhouette segmentation (predicted by FPN [5] and PointRend [3]) and 2D joint detections, obtained using Keypoint-

Metric	Pre-optimisation	Post-optimisation
Silhouette IOU	0.64	0.85
Silhoeutte global accuracy	0.95	0.98
2D Joints Euclidean error	19.4	9.8

Table 1: **Optimisation metrics.** This table reports error metrics between SMPL body models and target 2D observations across all samples in SSP-3D. Reported metrics are silhouette intersection-over-union (IOU), silhouette global accuracy and 2D joints Euclidean error (in pixels). The post-optimisation SMPL body models have significantly lower error metrics.

RCNN [7], matched the subject in the image accurately. This was crucial since the predicted silhouettes and 2D joints act as target observations during optimisation. Selected frames also required good pose and shape initialisations, obtained using VIBE [9].

Post-optimisation annotation involved selecting good SMPL fits (i.e. optimised pose and shape parameters). The quality of an SMPL fit was judged by (i) the overlap between the SMPL silhouette and predicted silhouette, (ii) the distance between projected SMPL joints and predicted joints and (iii) the 3D plausibility of the SMPL body mesh, which was determined by rendering and viewing the mesh from 3 different camera angles.

2 STRAPS

In the main manuscript, we described STRAPS, a framework for 3D human shape and pose estimation. Here, we define the metrics used to evaluate STRAPS and give qualitative examples of predictions from STRAPS when ground truth silhouettes are used as inputs.

2.1 Evaluation metric definitions

For 3D pose accuracy evaluation, we report mean per joint position error (in mm) after rigid alignment between the ground truth and prediction using Procrustes analysis, abbreviated as MPJPE-PA. This is a common metric used to evaluate 3D pose estimation approaches. A formal definition is given in [7].

For 3D shape accuracy evaluation, we report two metrics: PVE-T-SC and mean intersection-over-union (mIOU). "PVE-T" represents mean per-vertex Euclidean error (in mm) in a neutral T-pose - i.e. the pose contribution to the per-vertex error is removed by first reposing the mesh to a neutral T-pose. Reposing is trivial using the SMPL body model, since the SMPL pose parameters (i.e. joint rotation vectors) can simply be set to zero. However, there is an inherent ambiguity between 3D subject size/scale and distance from camera. Since the true camera location relative to the 3D subject (and the focal length) is unknown, it is not possible to estimate the absolute size of the subject given an image. This is not accounted for by vanilla PVE-T. We want to measure shape accuracy *up to scale* by eliminating the contribution of this ambiguity to PVE-T. To do so, we carry out a scale-correction step (hence the "SC" in PVE-T-SC), which rescales the predicted vertex mesh such that the root mean square distance (RMSD) of the predicted vertices from their mean is the same as the RMSD of the target vertices from their mean. The RMSD of a vertex mesh with N vertices $\{\mathbf{v}_i\}_{i=1}^{N}$

is defined as

$$\text{RMSD}\big(\{\mathbf{v}_i\}_{i=1}^N\big) = \sqrt{\frac{\sum_{i=1}^N ||\mathbf{v}_i - \mathbf{v}_{\text{mean}}||_2^2}{N}}, \tag{1}$$

where $\mathbf{v}_{\text{mean}} \in R^3$ is the mean vertex location. Given predicted vertices $\{\hat{\mathbf{v}}_i\}_{i=1}^N$ and target vertices $\{\mathbf{v}_i\}_{i=1}^N$, scale-corrected predicted vertices are obtained by

$$\hat{\mathbf{v}}_{\mathbf{i}}^{\text{SC}} = \frac{\text{RMSD}\big(\{\mathbf{v}_i\}_{i=1}^N\big)}{\text{RMSD}\big(\{\hat{\mathbf{v}}_i\}_{i=1}^N\big)}(\hat{\mathbf{v}}_{\mathbf{i}} - \hat{\mathbf{v}}_{\text{mean}}) + \mathbf{v}_{\text{mean}}. \tag{2}$$

After scale correction, the PVE-T-SC over an evaluation dataset with M samples is given by

$$\text{PVE-T-SC} = \sum_{m=1}^{M}\sum_{i=1}^{N} \frac{||\mathbf{v}_i^m - \hat{\mathbf{v}}_i^{\text{SC}m}||_2}{NM}. \tag{3}$$

We use PVE-T-SC (in mm) as a shape accuracy metric since it is invariant to pose and scale.

2.2 Qualitative examples using ground truth silhouettes as inputs

In the main manuscript, we use DensePose [1] to obtain silhouettes at test-time. However, DensePose may predict erroneous silhouettes, particularly for subjects with outlier body shapes. In Figure 1, we compare predictions made using DensePose silhouettes as inputs versus using ground truth silhouettes from SSP-3D. The ground-truth silhouettes were obtained using FPN [5] with PointRend [3] and curated by human annotators when creating SSP-3D, as described in Section 1. We show that having access to high-quality silhouettes allows our method to predict body shape accurately, even for challenging examples with extreme outlier subjects.

Hyperparameter	**Symbol** (used in paper)	**Value**
Synthetic data generation:		
Shape augmentation sampling variance	σ_n^2	2.25 for all $n \in \{1, ..., 10\}$
Uniform noise range for 2D joints	-	[-8, 8] pixels
Uniform noise range for vertices (for silhouette edge corruption)	-	[-10, 10] mm
Limb removal probability	-	0.1
Occlusion box probability	-	0.5
Occlusion box size	-	48 pixels
Camera translation sampling mean	Mean of $\mathbf{t}$	(0, 0.2, 42.0)
Camera translation sampling variance	Variance of $\mathbf{t}$	(0.05, 0.05, 8.3)
Camera rotation	R	I
Camera focal length	Part of K	5000.0
Proxy representation width and height	W and H	256×256 pixels
Optimisation:		
Reprojection error weight	λ_j	1
Silhouette error weight	λ_S	50
Angle prior weight	λ_a	1
GMM pose prior weight	λ_θ	1
Shape regulariser weight	λ_β	2
Pose regulariser weight	$\lambda_{\theta^{\text{init}}}$	1000
Silhouette width and height	W and H	512×512 pixels

Table 2: **Hyperparameter values not provided in the main manuscript.**

Figure 1: **Qualitative comparison between 3D predictions using DensePose [11] versus ground-truth silhouettes (from SSP-3D) as inputs.** This figure illustrates that high-quality silhouettes that accurately follow the subject's body shape can improve 3D shape predictions using our approach.

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Kostantinos G Derpanis, and Kostas Daniilidis. Monocap: Monocular human motion capture using a CNN coupled with a geometric prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2018.