AGORA: Avatars in Geography Optimized for Regression Analysis

Priyanka Patel[1] Chun-Hao P. Huang[1] Joachim Tesch[1] David T. Hoffmann[2,3*]
Shashank Tripathi[1] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]University of Freiburg [3]Bosch Center for Artificial Intelligence
{ppatel, paul.huang, jtesch, dhoffmann, stripathi, black}@tuebingen.mpg.de

Figure 1: **AGORA** dataset examples. Top row: images with different scenes. Middle: SMPL-X ground-truth bodies rendered in the scene. Bottom: per-person segmentation masks including environmental occlusion (see bottom right image).

Abstract

While the accuracy of 3D human pose estimation from images has steadily improved on benchmark datasets, the best methods still fail in many real-world scenarios. This suggests that there is a domain gap between current datasets and common scenes containing people. To obtain ground-truth 3D pose, current datasets limit the complexity of clothing, environmental conditions, number of subjects, and occlusion. Moreover, current datasets evaluate sparse 3D joint locations corresponding to the major joints of the body, ignoring the hand pose and the face shape. To evaluate the current state-of-the-art methods on more challenging images, and to drive the field to address new problems, we introduce AGORA*, a synthetic dataset with high realism and highly accurate ground truth. Here we use 4240 commercially-available, high-quality, textured human scans in diverse poses and natural clothing; this includes 257 scans of children. We create reference 3D poses and body shapes by fitting the SMPL-X body model (with face and hands) to the 3D scans, taking into account clothing. We create around 14K training and 3K test images by rendering between 5 and 15 people per image using either image-based lighting or rendered 3D environments, taking care to make the images physically plausible and photoreal. In total, AGORA consists of 173K individual person crops. We evaluate existing state-of-the-art methods for 3D human pose estimation on this dataset. and find that most methods perform poorly on images of children. Hence, we extend the SMPL-X model to better capture the shape of children. Additionally, we fine-tune methods on AGORA and show improved performance on both AGORA and 3DPW, confirming the realism of the dataset. We provide all the registered 3D reference training data, rendered images, and a web-based evaluation site at* https://agora.is.tue.mpg.de/.

*This work was done while DTH was at MPI-IS.

1. Introduction

The field of 3D human pose and shape (3DHPS) estimation from images has advanced rapidly with steadily decreasing errors on standard benchmarks [17, 21, 26, 27, 36, 39, 47]. Large training datasets and benchmarks, with ground truth, enable progress and quantitative evaluation. These are difficult to obtain in the case of 3DHPS. Existing datasets have significant limitations and the rate of progress now suggests that these benchmarks are becoming saturated, making it difficult to evaluate how close the field is to fully robust and general solutions. These datasets often have limited clothing, focus on single subjects, have limited occlusion, are captured in laboratory environments, or have a limited range of ages and ethnicities. Additionally, accuracy is evaluated based on a small number of 3D joints, while the body is much more complex. To drive advances in the field, we propose a novel dataset that includes challenging scenarios neglected by earlier datasets and a more challenging evaluation protocol.

AGORA (Avatars in Geography Optimized for Regression Analysis) is a new publicly available dataset that includes high-resolution (4K) images with ground truth 3D bodies. AGORA goes beyond previous datasets in important ways. It includes accurate 3D body pose and shape of people in varied and complex clothing. People with varied poses, ages and ethnicities appear in complex natural scenes with natural lighting. Additionally, the dataset includes person-person occlusion, environmental occlusion, camera frame occlusion, crowds, children, face and hand pose, large field of view images and people appearing at a wide range of spatial scales. To the best of our knowledge, AGORA is the only dataset that provides all these features together with highly accurate 3D ground truth. Figure 1 shows a few representative examples from the dataset.

Since there is currently no technology to capture ground truth body shape and pose for real images of this complexity, we rely on synthetic data and a graphics rendering pipeline. Specifically, we purchased 4240 high-quality textured scans of people, which include 257 child scans from 3DPeople [1], AXYZ [2], Human Alloy [4] and Renderpeople [7]. These scans provide a rich variety of ethnicity, age, pose, and clothing variation with realistic textures. We also gathered a variety of scenes as HDRI panoramas and 3D environments. We randomly sample 3D people and place them in scenes at random distances and orientations. We then render them realistically using a game engine optimized for high-quality output [8].

For every scan, we fit the SMPL-X body model [37], taking great care to accurately capture the correct body shape, pose, hand shape, and facial shape; see Fig. 1 middle row. To generate the AGORA ground truth (or reference data), we take an optimization-based approach that fits SMPL-X to each scan. Specifically, we estimate both the pose and body shape under clothing, similar to [10, 56]. The estimated SMPL-X fits have an average error of 5mm, making them accurate enough to benchmark existing state-of-the-art (SOTA) methods. For backward compatibility with SMPL, we also provide ground truth in the gender-neutral SMPL format[1] [33] used by many current methods [25, 27, 47].

In addition to adults, the AGORA dataset contains images of children. It is probably the only dataset of children with reference 3D pose and shape. Existing 3DHPS methods focus on adult bodies and perform poorly on images of children. With AGORA, we evaluate this performance, but go further and extend the SMPL-X shape space to capture the variation in body shape beween infants and adults. Specifically, we introduce a shape dimension that interpolates between an adult SMPL-X body template and the infant SMIL template [18], which we convert to SMPL-X format. This results in an extra shape parameter that can be optimized like any other SMPL-X shape parameter.

We make the SMPL-X fits available for all the 14529 training and 1225 validation images, enabling training with AGORA. We withhold the ground truth bodies from the 3387 test images and instead provide an evaluation server. While we cannot provide the commercial scans, we provide a "shopping list" of the training and validation scans so that others can purchase them. Purchasing the scans extends the applications of AGORA to other problems such as 3D clothing modeling, neural avatars, and shape regression. We also provide the test scripts for researchers to test their methods on the validation set.

We use AGORA to evaluate SOTA 3DHPS methods with a novel protocol. In addition to the common 3D joint-based error measures, we provide a vertex-to-vertex error, and evaluate body, hands and face pose and shape. We also use 2D occlusion masks (Fig. 1 bottom row) to evaluate the performance of methods at varying levels of occlusion. Since our images contain multiple people, methods may detect too few as well as too many people. Consequently, we introduce an error measure that goes beyond the standard single-person measures and rewards methods for both 3DHPS and detection accuracy. We observe higher errors for SOTA methods on AGORA than on other datasets, suggesting that AGORA is more challenging. We also show that our training set can be used to improve recent 3D pose estimation methods [27] not only on AGORA but also on 3DPW [52]. This validates that the synthetic data is sufficiently real to be useful.

In summary, we contribute a new, varied and challenging dataset to evaluate and improve the SOTA in 3D human pose and shape estimation and to push the field in new

[1] Given a SMPL-X mesh, we convert it to gender-neutral SMPL format by fitting the gender-neutral SMPL template to it. In this work, SMPL fits are always generated through this process unless otherwise stated.

directions. The dataset is synthetic but diverse and realistic. We use new evaluation metrics and provide detailed analysis of limitations of current methods. We also introduce a new child model to generate better ground truth shape for children. We provide the training and validation set images with SMPL-X and SMPL ground truth and 2D masks. We also provide test images along with evaluation code and will maintain a web evaluation server: https://agora.is.tue.mpg.de/.

2. Related Work

Many datasets have been proposed for 3DHPS estimation, but each has limitations as summarized in Table 1. While there are many 2D datasets, we focus on those with 3D ground truth of one form or another.

Datasets with real images. Unlike 2D annotation, 3D body poses are difficult for humans to annotate since the task is ambiguous and requires metric accuracy. Consequently, existing benchmarks rely on multiple synchronized cameras. For example, HumanEva [46] , Human3.6M [20], and TotalCapture [48] synchronize video cameras with motion capture (mocap) systems that provide ground truth through optical markers. While providing accurate 3D pose, the image complexity is limited: lack of background variation in lab scenarios, only one subject in each image, no scene occlusions, and little clothing variety due to the attachment of markers, which, unfortunately are also visible in the images. These methods typically evaluate accuracy based on 3D joint locations. Note that, while the 3D joints are commonly treated as "ground truth", they are not directly observed, but rather are inferred by the mocap system based on an approximate skeletal body structure.

Alternatively, several methods use marker-less motion capture, e.g. MuPoTS-3D [35], PanopticStudio [24], MPI-INF-3DHP-Test [34], and HUMBI [54]. Such methods are typically less accurate than marker-based systems, but they avoid intrusive markers, allow more varied clothing, and sometimes are used in more realistic scenes e.g. outdoors. IMU sensors provide another way to measure 3D poses, which is less intrusive than mocap markers but also less accurate due to yaw drift. Von Marcard et al. [52] explicitly account for this by combining IMU data with monocular video, enabling in-the-wild capture. We consider these datasets as reference data rather than "ground truth" because the accuracy of the method is evaluated in a separate process (e.g. using mocap data) and not on the image data in the benchmark. In contrast, for AGORA we report how close the SMPL-X meshes are to these reference scans, directly indicating the fidelity of our pseudo ground truth.

All the above are limited in the complexity of the clothing, occlusions, scene variety, ethnicity, etc. Of the above only PanopticStudio [24] and HUMBI [54] consider the face and hands together with bodies.

Synthetic datasets. Computer graphics has the potential to synthesize large-scale image datasets, where ground truth is generated by animating parametric 3D human models such as SMPL [33], MakeHuman [5], or Mixamo [6]. The main challenge for such methods lies in creating data that is sufficiently realistic in terms of body shape, ethnicity, motion, cloth deformation, texture, and interaction with environments. In several datasets, images are created by compositing 3D people on image backgrounds. MHOF [40], LTSH [19], 3DPeople [38], and SURREAL [50] render 3D people on the background image, while MPI-INF-3DHP-Train [34] and MuCo-3DHP [35] paste a segmented real human foreground on top of the background. Such composition does not faithfully reflect the local statistics of pixel intensity in real images and does not support methods that learn how humans interact with scenes. Most similar to us is SimPose [60], which poses 17 rigged commercial scans [7] and SURREAL data rendered in a 3D scene. The 3D scenes are simplistic, the scans lack diversity, there is no evaluation site, and the dataset is not public.

A recent promising direction synthesizes realistic looking people in images [55, 59]. Zanfir et al. [55] use a learned human synthesis method to insert generated people in images such that they make sense relative to the scene geometry and lighting. While they can condition the generated person on pose and shape, the resulting images contain artifacts that are common to generative models, making the results unsuitable as ground truth.

Other human-related datasets. There are many other datasets of real humans in images that do not contain 3D ground truth. For example, OCHuman [57] focuses on occlusion in real single-view images and provides 2D joint landmarks and human segmentation masks. Early multiview sequences e.g. Adobe data [51], MVIC [32], and MARCOnI [14] also consider 2D landmarks and silhouettes as evaluation measures. The hunger for large training corpora for deep learning motivates self-supervised strategies [25, 26, 49] that leverage 2D landmark annotations in LSP-Extended [22], COCO [31], and MPII [9]. Several recent datasets, e.g. EFT [23], STRAPS [44] and 3DOH50K [58], are generated by fitting a body model to the images, while others fit to videos of complex scenes with "frozen" people [30] using structure from motion [28] or multi-view matching [45]. Methods like EFT and SMPLy [28] provide image variety, which is good for robustness, but with unknown accuracy in body shape and pose.

In summary, no single dataset can address all needs of the community. AGORA provides realistic textures, complex body shapes and clothing, complex varied scenes and lighting, high-resolution (4K) imagery, varied occlusion, all with high-quality 3D ground truth. This new benchmark reveals limitations of current approaches while providing novel, high-quality training data for multiple applications.

Dataset	Sub. #	Image	Complexity	Clothing	Body anno.	Ground truth format
HumanEva [46]	4	lab	1 subject, no occlusion	limited	B	3D joint locations
Human3.6M [20]	11	lab	1 subject, minor occlusion	limited	B	3D joint locations
TotalCapture [48]	5	lab	1 subject, no occlusion	mocap suit	B	3D joint locations
PanopticStudio [24]	∼100	lab	multiple subjects & furniture	varied	BFH	3D joint locations
HUMBI [54]	772	lab	1 subject, no occlusion	rich	BFH	meshes, SMPL
3DPW [52]	18	natural	multiple subjects in the wild	varied	B	SMPL
MuPoTS-3D [35]	8	natural	multiple subjects in the wild	varied	B	3D joint locations
MPI-INF-3DHP-Train [34]	14	both	1 subject, minor occlusion	varied	B	3D joint locations
3DOH50K [58]	n/a	lab	1 subject, object occlusion	limited	B	SMPL
EFT [23]	> 1000	natural	multiple subjects, in the wild	varied	B	SMPL
STRAPS [44]	62	natural	1 subject, in the wild	limited	B	SMPL
SMPLy [28]	742	natural	multiple subjects, in the wild, frequent occlusion	rich	B	SMPL
MuCo-3DHP [35]	8	composite†	multiple subjects in the lab	limited	B	3D joint locations
MPI-INF-3DHP-Test [34]	14	composite†	1 subject, minor occlusion	varied	B	3D joint locations
SURREAL [50]	145	composite†	1 subject, no occlusion	texture¶	B	SMPL
3DPeople [38]	80	composite†	1 subject, no occlusion	synthetic‡	B	3D joint locations
AGORA (ours)	>350	realistic††	multiple subjects in the wild, frequent occlusion	rich	BFH	SMPL-X, SMPL, masks

Table 1: Comparison of datasets that provide images and 3D human pose annotations. Body annotation type B, F, and H correspond to body, face, and hands respectively. †: 2D foreground layers pasted on background images. ††: 3D models positioned in 3D with panoramic background or full 3D scenes. ¶: unclothed human body with clothing texture. ‡: clothed human body with texture.

3. Method: Obtaining reference data

To construct AGORA, we purchased high-quality textured 3D scans from 3DPeople [1], AXYZ [2], Human Alloy [4] and Renderpeople [7]. We selected 4240 scans for inclusion in the dataset spanning more than 350 unique subjects. A scan $\mathcal{S}$ comprises a set of 3D points $S \subset \mathbb{R}^3$ and their connectivity $F_{\mathcal{S}}$, $\mathcal{S} = \{S, F_{\mathcal{S}}\}$. To each scan $\mathcal{S}$ we fit a parametric SMPL-X body model $\mathcal{M} = \{M, F_{\mathcal{M}}\}$, whose vertex locations $M(\theta, \beta, \psi) \subset \mathbb{R}^3$ are controlled by parameters for pose θ, shape β, and facial expression ψ [37]. θ consists of body pose θ_b and hand pose θ_h. Hand pose θ_h is a function $\theta_h(Z_h)$ of a PCA latent vector $Z_h \in \mathbb{R}^6$.

Fitting a SMPL-X mesh $\mathcal{M}$ to a scan $\mathcal{S}$ amounts to solving for the optimal parameters (θ, β, ψ) such that $\mathcal{M}$ resembles $\mathcal{S}$. The fitting process takes into account that SMPL-X explains the body in minimal clothing while the scans are typically clothed. In this process we exploit the fact that a person may appear in multiple scans and their shape parameter, β, should be the same across scans.

We first initialize the parameters by an approach that extends the single-view SMPLify-X fitting [37] to multi-view images rendered using C pre-defined virtual cameras. The initial mesh, M, obtained by multi-view SMPLify-X fitting is only approximately aligned with S. While sparse 2D landmarks constrain the 3D pose, they provide little information about body shape. To refine the shape and pose, we fit SMPL-X to the 3D scan surface. However, this is challenging because SMPL-X cannot model things like hair and clothing that are present in the scans. To address this, we use the idea of fitting body shape under clothing [10, 56].

Similar to [56], we define energy terms E_{skin} and E_{cloth} for skin and clothing, respectively. Both aim to bring the model surface close to the scan, whereas E_{cloth} additionally penalizes body vertices being outside the clothing. In other words, our objective function tries to move the model as close as possible to the scan near the visible skin while discouraging the clothing vertices from penetrating the model. We label skin and cloth vertices on scan using Graphonomy [15]. We keep the 2D landmark data term E_J^c that penalizes differences between projected and observed keypoints from multi-view fitting, as they provide information complementary to E_{skin} and E_{cloth}. See Sup. Mat. for more details.

We fit each model $M_i(\beta_i, \theta_i, \psi_i)$ to the corresponding scan S_i in parallel, for scans, i, of the same identity. We optimize jointly for θ_i, ψ_i and β_i while minimizing the shape (inter-beta) distance E_{ib} between scans of the same identity. The objective function is:

$$E(\beta_1, \dots, \beta_N, \theta_1, \dots, \theta_N, \psi_1, \dots, \psi_N) =$$
$$\sum_{i=1}^{N} \left(\lambda_J \sum_{c=1}^{C} E_J^{c,i} + \lambda_{\text{s}} E_{\text{skin}}^i + \lambda_{\text{c}} E_{\text{cloth}}^i + E_{\text{reg}}^i \right) + \lambda_{\text{ib}} E_{\text{ib}},$$
$$E_{\text{ib}} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \|\beta_i - \beta_j\|_2^2,$$
$$E_{\text{reg}} = \lambda_{\theta_b} E_{\theta_b}(\theta_b) + \lambda_{\theta_h} E_{\theta_h}(\theta_h) + \lambda_\beta E_\beta(\beta) + \lambda_\mathcal{E} E_\mathcal{E}(\psi),$$

where E_{reg} contains L_2 priors used to constrain the body shape, pose and expression, as defined in [37]. Different weights denoted by λ are used for each term.

This approach exploits semantic information (2D landmarks, skin/clothing segmentation) as well as geometry (3D shapes) to obtain accurate fits as demonstrated in Sec. 4.1.

3.1. Fitting child scans

AGORA also contains 257 child scans. Fitting SMPL-X directly to these scans results in distorted fits as shown in

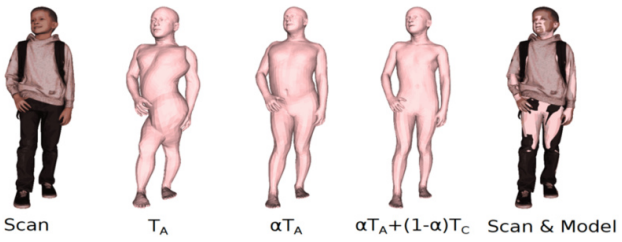

Figure 2: Fitting child scans using an adult template (T_A), a scaled adult template (αT_A), and our proposed approach, which interpolates between adult and infant templates ($\alpha T_A + (1-\alpha)T_C$).

the 2nd column of Fig. 2, because SMPL-X cannot represent children. Naively scaling the adult SMPL-X template, i.e. αT_A, by optimizing for a global scale parameter α is better but still unnatural since children have different proportions than adults. To solve this problem, we take the mean infant body template from SMIL [18] and convert it to SMPL-X topology, T_C. We find that interpolating between the adult SMPL-X template T_A and the SMIL infant template T_C approximately captures the shape of children. See Sup. Mat. Note that, while not perfect, we use the adult shape space for children and only vary the template shape. Incorporating children then involves only a minor change to the fitting process. In addition to optimizing the shape parameters, β, we also optimize for a weight, $\alpha \in [0, 1]$, the linearly interpolates the templates. This produces more accurate body shapes for children, as shown in Fig. 2.

4. AGORA Dataset

AGORA consists of 4240 scans spanning more than 350 unique subjects, all paired with SMPL-X fits (we also supply SMPL fits for backward compatibility). While generally robust, the fitting approach fails sometimes for hands and faces. This typically happens when the hands are grasping objects. Since we want high-quality ground truth, we manually curate the results of the automatic process and create two different sets: (1) those with well aligned body, face and hands (3161, BFH); and (2) those only with well aligned bodies (1079, B). When evaluating algorithms for 3DHPS estimation, only body joints and vertices are considered for B scans, while body, hand and face joints and vertices are evaluated for BFH scans.

We sample 1051 scans to create the test set; for these, the SMPL-X fits are withheld from public release. We make sure there are no overlapping scans between these 1051 scans and the rest, and no selected scans have been included to train previous work [43]. Our test set spans 105 subjects. We create approximately 3387 images from the test scans, including many challenging scenarios, with the goal of making them photorealistic. From the remaining scans, we sample 2930 scans as a training set and 259 scans as a validation set and create 14529 training and 1225 validation images, whose ground-truth SMPL-X parameters and 2D segmentation masks are included in the released dataset. Since the ground truth 3D scans are commercially available, it is possible for people to cheat by using test scans for training. We have also built in several countermeasures to detect cheating that we do not describe in this paper.

AGORA images are rendered using perspective cameras with focal lengths 18mm, 28mm and 50mm. All images are rendered using Unreal Engine [8] on a single Windows 10 PC with NVIDIA RTX 2080 graphics hardware. Renderings are generated either with image-based lighting using freely available HDR backgrounds [3] or with free and commercial 3D environments obtained from the Unreal Marketplace. The image-based lighting scenes used hardware-accelerated ray-tracing for accurate ground plane shadows. For scans in seated poses, we insert random chairs at the appropriate height so that the scans appear naturally supported. See Sup. Mat. for details of the rendering process.

4.1. Fitting Accuracy

To evaluate the accuracy of results on AGORA, and to know whether improvements are significant, we first need to know the accuracy of our ground truth[2]. We define accuracy relative to the high-quality 3D scans in the following ways:

1. Skin error. For the visible skin vertices on the scan, we compute the Euclidean distance of the nearest point on the triangle of the reconstructed model M. An accurate model fit should fit closely to the skin. We report the weighted mean distance as our final error value where the weight is the probability of the scan vertex belonging to skin calculated using Graphonomy [15].

2. Penetrating clothing error. The SMPL-X fits are supposed to be fully inside the clothing. We report two values for clothing vertices on the scans: (1) the percentage of them that penetrate the body model. (2) for those penetrating vertices, we calculate their distance to the closest point on the model surface and compute the weighted avg. error.

We consider only the scans without any large objects for the error calculation and report an average skin error of approximately 4.73mm. Only 16% of cloth vertices are inside the body with an average distance of 4.63mm. An error of approximately 5mm is significantly below any industry standards for the measurement of live humans and is less than the soft-tissue motion of mocap markers on the body [11]. Thus, we believe that the SMPL-X fits provide valid pseudo ground truth.

4.2. Evaluation metrics

A common practice in evaluating 3DHPS methods is applying Procrustes alignment [16] before computing the er-

[2]Note that with traditional mocap ground truth, only the accuracy of the markers is known – the accuracy of the 3D joints is actually unknown.

	Method	MPJPE ↓				MVE ↓				NMJE ↓		NMVE ↓		F1 score↑
		B	LH/RH	F	FB	B	LH/RH	F	FB	B	FB	B	FB	
SMPL	HMR [25]	180.5	N/A	N/A	N/A	173.6	N/A	N/A	N/A	226.0	N/A	217.0	N/A	0.80
	CenterHMR [47]	168.1	N/A	N/A	N/A	161.4	N/A	N/A	N/A	242.3	N/A	233.9	N/A	0.69
	EFT [23]	165.4	N/A	N/A	N/A	159.0	N/A	N/A	N/A	203.6	N/A	196.3	N/A	**0.81**
	SPIN [27]	175.1	N/A	N/A	N/A	168.7	N/A	N/A	N/A	223.1	N/A	216.3	N/A	0.78
	SPIN-ft (ours)	**153.4**	N/A	N/A	N/A	**148.9**	N/A	N/A	N/A	**199.2**	N/A	**193.4**	N/A	0.77
SMPL-X	SMPLify-X [37]	182.1	**46.5/49.6**	**52.9**	231.8	187.0	**48.3/51.4**	**48.9**	236.5	256.5	326.5	263.3	333.1	0.71
	ExPose [13]	**150.4**	72.5/68.8	55.2	**215.9**	**151.5**	74.9/71.3	51.1	**217.3**	**183.4**	**263.3**	**184.8**	**265.0**	**0.82**
	Frankmocap [42]	165.2	52.3/53.1	N/A	N/A	168.3	54.7/55.7	N/A	N/A	204.0	N/A	207.8	N/A	0.81

Table 2: Comparison of SOTA 3DHPS methods on the AGORA testset. SPIN-ft is SPIN after finetuning on the AGORA training set described in Sec. 5.2. SMPL-based methods are evaluated on B+BFH and SMPL-X-based methods are evaluated on the BFH subset of AGORA. Error metrics are described in Sec. 4.2. All numbers are in millimeters.

ror. Doing so eliminates discrepancies in scale, translation and rotation, measuring only the error in poses (PA-MPJPE) and shapes (PA-MVE/V2V). This convention is largely due to the fact that existing HPS datasets, e.g. [23, 52], contain only pose and shape annotations, and HPS methods estimate the body relative to the camera. In contrast, AGORA provides *complete* 3D pseudo ground truth: body parameters of each person and their spatial arrangement in the 3D scene, enabling a more comprehensive error measure.

Consequently, we do not apply Procrustes alignment but only align at the pelvis, i.e. MPJPE and MVE/V2V, because estimating absolute depth is ambiguous. Furthermore, since AGORA has 5-15 people per image, methods may not detect every person leading to misses, i.e. false negatives. Due to occlusions, methods may also detect bodies where there are actually no people, i.e. false positives. Accuracy on AGORA means high detection performance and low error for every correct detection; consequently, we must penalize false negatives and false positives. Thus, we normalize the MPJPE and MVE/V2V error by the standard detection metric, F1 score (the harmonic mean of recall and precision), and refer to this as *Normalized Mean Joint Error (NMJE)* and *Normalized Mean Vertex Error (NMVE)*. F1 score punishes both misses and false alarms so NMJE/NMVE increase the reported error for methods that make either type of mistake in detection. As a result, to reduce the overall NMJE/NMVE, the method needs to miss no one, detect no spurious bodies, and estimate accurate poses and shapes for each correct detection, making NMJE/NMVE more challenging and comprehensive than other metrics.

We evaluate 3DHPS methods along different dimensions and also provide a combined score. For SMPL-based methods, we just evaluate on body joints and vertices using both B and BFH scans. For SMPL-X-based methods, we evaluate separately on the body, hands and face and also provide a weighted sum of the three as a full body (FB) error. SMPL-X-based methods are evaluated only for BFH scans.

B-MPJPE is evaluated on 24 body joints of SMPL and 22 body joints of SMPL-X after aligning the pelvis. **LH-MPJPE, RH-MPJPE** are evaluated on 15 hands joints on the left and right hands, respectively, after aligning the wrist joint. **F-MPJPE** is evaluated on 51 facial landmarks after aligning the neck joint. **FB-MPJPE** is a weighted sum of the above 4 errors. Since the number of hand joints and face landmarks outweigh the number of body joints, we define the FB error as $\text{FB} = \text{B} + (\text{LH+RH+F})/3$.

While 3D joint error evaluates pose, it does not provide evaluation of shape, for which we have ground truth. To encourage research on body shape estimation, we also evaluate the methods on vertices. We segment the body, left hand, right hand and face vertices using SMPL [33], MANO [41] and FLAME [29] vertex indices of the SMPL-X template and calculate **B-MVE**, **LH-MVE**, **RH-MVE** and **F-MVE** respectively. **FB-MVE** uses the same weighted combination as joint error, FB-MPJPE. We also calculate **B-NMJE**, **B-NMVE**, **FB-NMJE**, **FB-NMVE** and penalize the methods for missed detections and false positives.

4.3. Evaluation protocol

When a method estimates a body, the matching ground truth body in AGORA is not known. Therefore, to match the predicted person with the ground truth, we project the estimated 3D keypoints to the image plane and find the closest ground-truth subject in terms of 2D joint error. If there is no match found for a particular ground truth body, we count it as a miss (see Sup. Mat. for details). Similarly, if a detection does not match any ground truth, we count it as false positive. For the correctly matched predictions, we calculate all the errors as described in Sec. 4.2.

5. Experiments

We evaluate existing methods on AGORA to determine whether the dataset provides new insights about the current SOTA. We also evaluate whether the AGORA training set can help improve the accuracy of SOTA methods by using it to fine-tune SPIN [27].

5.1. Baseline Evaluation.

The evaluation protocol for existing methods is shown in Fig. 3. Most current methods assume that the input im-

	Method	MPJPE (mm) ↓ B	MVE (mm) ↓ B
SMPL	HMR [25]	219.4	209.3
	CenterHMR [47]	207.4	198.5
	EFT [23]	202.7	193.5
	SPIN [27]	203.7	193.2
	SPIN-ft	**191.7**	**186.7**
SMPL-X	SMPLify-X [37]	208.3	213.3
	ExPose [13]	**176.6**	**174.0**
	Frankmocap [42]	203.7	204.2

Table 3: Performance of SOTA methods on "AGORA kids."

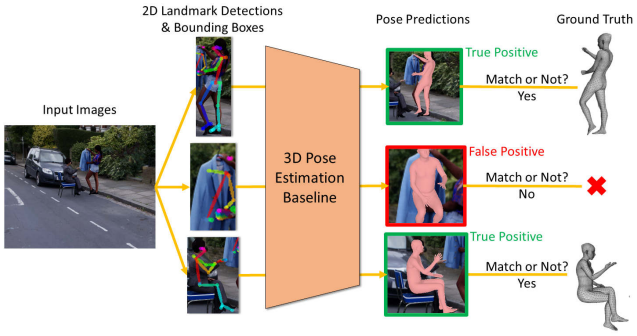

Figure 3: Baseline Evaluation. Given a test image, we detect keypoints with [12] to obtain bounding boxes centered at each detected person, followed by network inference to reconstruct a human mesh for each cropped image. We identify true positives (to compute pose error), false negatives (misses) and false positives by matching predictions and ground truth. See Sec. 4.3 for details.

age is tightly cropped around the person [13, 23, 25, 27, 42] or require 2D keypoint detections [37]. Therefore, to fairly test prior methods on AGORA with the same input, we use OpenPose [12] to detect people and their respective keypoints and construct tight bounding boxes based on these detections. For single-stage approaches e.g. [47] we directly use the entire image as input without any cropping.

Table 2 reports results for multiple baselines on the AGORA testset using the evaluation metrics described in Sec. 4.2. To compare our new metrics with metrics used in earlier work, we also report MPJPE and MVE without penalizing for missed detections and false positives. See Fig. 7 in Sup. Mat. for qualitative results. While SPIN fine-tuned on the AGORA training set (Sec. 5.2) outperforms other SMPL-based SOTA methods by a large margin in terms of MPJPE and MVE error, its error increases under our new NMJE and NMVE metrics because of misses and false positives. This shows that MPJPE alone is not enough to evaluate performance on multi-person images. We hope AGORA will drive research on multi-person pose estimation.

We notice that among SMPL-X based methods, ExPose [13] performs best for the body while the optimization-based method SMPLify-X [37] beats regression based

Models	3DPW (14)		3DPW (24)		AGORA (24)
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE
SPIN-pt [27]	96.9	59.3	95.5	65.5	175.1
SPIN-ft-EFT [23]	97.4	59.7	95.3	66.1	173.7
SPIN-ft (ours)	**85.7**	**55.3**	**83.7**	**61.8**	**153.4**

Table 4: Pretrained SPIN vs. SPIN finetuned with AGORA and EFT([MPII+LSPet+COCO]). Parens.: (#joints).

methods in hand and face estimation. These errors are further analysed w.r.t. different parameters like occlusion, child shape, distance to the center of the image and orientation (Sup. Mat. for orientation).

Occlusion. Using the ground-truth segmentation masks, Fig. 4 plots the error of SOTA methods vs. the percentage of occlusion. Since this is analyzed on ground-truth bodies in which false positives are not included, we normalize the MPJPE by recall (correctly detected and matched bodies divided by total number of bodies), denoted as recall-NMJE and we also plot it for different ranges of occlusion. As expected, the MPJPE for correct detections increases with increasing occlusion and the percentage of misses also increases as shown in the left and middle plots in Fig. 4. We observe that CenterHMR performs well for high occlusion but suffers from many misses, particularly with small people. See Fig. 7 in Sup. Mat. This shows that bottom-up methods that work on the full image are good in dealing with images of multiple people but need to improve their detection accuracy. FrankMocap and SPIN are highly sensitive to occlusion, leading to large errors as occlusion increases. We also notice that fine-tuning with AGORA improves the performance of SPIN for high occlusion cases.

Distance from center. Most methods rely on a weak perspective camera assumption, which breaks when people occur off-center in images, as also pointed out by [53]. The large field-of-view images in AGORA facilitate the analysis of this error. We plot the B-MPJPE error of the selected SOTA methods vs. horizontal distance from the center of the image in Fig. 5. We find that error consistently increases for all methods as the distance from the center increases. This effect is less significant for CenterHMR [47], the only method that works on full images instead of crops.

Child Shape. AGORA contains child scans with pseudo ground truth shape generated as described in Sec. 3.1. We calculate body joint and vertex error for the SOTA on a "kids" subset of AGORA. We find that performance is significantly worse for predicting child shape compared to adult shape as shown in Table 3. We hope that this, together with the child shape representation we provide, will encourage work in this direction.

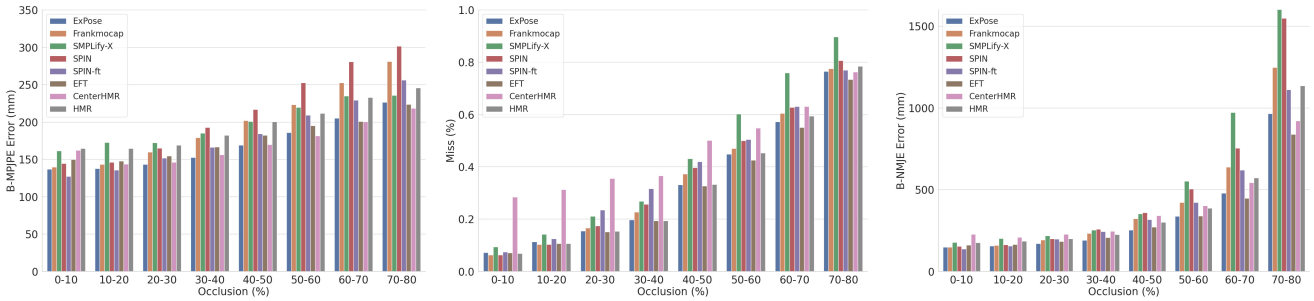

Figure 4: SOTA evaluation: B-MPJPE for the correct predictions (left), percentage of misses (center) and B-recall-NMJE for the correct predictions (right). Evaluated on BFH subset of AGORA for 22 SMPL-X and 24 SMPL joints.

5.2. Baseline Improvement.

To evaluate the efficacy of the AGORA training set, we fine-tune a pretrained SPIN model (SPIN-pt) using only crops from AGORA training images and refer to the fine-tuned model as SPIN-ft. We chose SPIN for the fine-tuning experiment as it uses HMR as a backbone, which is a base for many 3DHPS methods, e.g. EFT [23], VIBE [26] and FrankMocap [41]. For fair comparison, we use the same hyperparameters and loss functions as SPIN. However, unlike SPIN, we do not use SMPLify in loop, replacing that supervision with AGORA ground truth.

While AGORA images are rendered using perspective cameras, SPIN assumes a weak perspective camera, which is unable to capture perspective warping, especially when people occur off-center in images (see Sec. 5.1). This makes the global orientation of the ground-truth 3D joints and vertices in AGORA inconsistent with SPIN predictions. During SPIN fine-tuning, we therefore set global orientation to zero before calculating all 3D losses, such that information about global orientation comes only from the 2D keypoint loss. We note that this is required only due to the weak perspective camera assumption in SPIN and we hope AGORA will encourage research with more realistic perspective camera models.

Since SPIN reports Procrustes aligned MPJPE (PA-MPJPE) on 3DPW, we compare SPIN-pt and SPIN-ft on both MPJPE and PA-MPJPE. We also compare the AGORA training set with the EFT dataset, $[\text{MPII+LSPet+COCO}]_{\text{EFT}}$, We call fine-tuning on EFT, SPIN-ft-EFT. We evaluate SPIN-pt, SPIN-ft-EFT and SPIN-ft on the 3DPW testset with known association and on AGORA testset without known association. Training with AGORA leads to significant improvement in performance on both datasets, with MPJPE improving by $\sim$12% for 3DPW and $\sim$13% for AGORA; see Table 4. Higher MPJPE on AGORA compared to 3DPW also shows that AGORA is more challenging than 3DPW. Note that we calculate the error using 14 joints to compare with original SPIN-pt results.

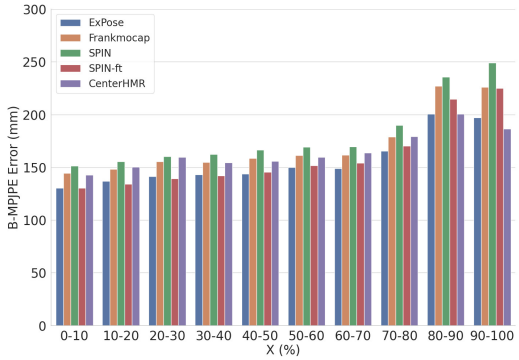

Figure 5: Horizontal distance from center of the image: B-MPJPE for selected SOTA baselines. Evaluated on the BFH subset of AGORA for 22 SMPL-X and 24 SMPL joints.

6. Conclusions and Future Work

We have presented AGORA, a new dataset that goes beyond current datasets to include challenging cases of environmental occlusion, person-person occlusion, scale variation, children, crowds, etc. AGORA is challenging and reveals limitations of existing methods. Despite being synthetic, fine-tuning on AGORA improves performance of a SOTA method on the natural 3DPW dataset. We introduce a new metric to include misses and false positives and facilitate analysis of the SOTA methods on images with multiple people. We also introduce a simple child body model and provide better 3D ground truth for images with children. Future work should include adding images of varied camera height, indoor scenes, multi-view images, larger crowds, animals, and movement.

Acknowledgements. We thank Galina Henz, Taylor McConnell and Tsvetelina Alexiadis for the help in data labeling. **Disclosure.** MJB has received research gift funds from Adobe, Intel, Nvidia, Facebook, and Amazon. While MJB is a part-time employee of Amazon, his research was performed solely at, and funded solely by, Max Planck. MJB has financial interests in Amazon, Datagen Technologies, and Meshcapade GmbH.

References

[1] *3DPeople*, 2020. https://3dpeople.com. 2, 4

[2] *AXYZ*, 2020. https://secure.axyz-design.com. 2, 4

[3] *HDRIHaven*, 2020. https://hdrihaven.com. 5

[4] *HumanAlloy*, 2020. https://humanalloy.com. 2, 4

[5] *MakeHuman*, 2020. http://www.makehumancommunity.org. 3

[6] *Mixamo*, 2020. https://www.mixamo.com. 3

[7] *Renderpeople*, 2020. https://renderpeople.com. 2, 3, 4

[8] *Unreal*, 2020. https://www.unrealengine.com. 2, 5

[9] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. Human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686 – 3693, 2014. 3

[10] A. Balan and M. J. Black. The naked truth: Estimating body shape under clothing,. In *European Conference on Computer Vision*, volume 5304, pages 15–29, 2008. 2, 4

[11] V. Camomilla, T. Bonci, and A. Cappozzo. Soft tissue displacement over pelvic anatomical landmarks during 3-D hip movements. *Journal of Biomechanics*, 62:14 – 20, 2017. 5

[12] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021. 7

[13] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, pages 20–40, 2020. 6, 7

[14] Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchulin, Mykhaylo Andriluka, Christoph Bregler, Bernt Schiele, and Christian Theobalt. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3810–3818, 2015. 3

[15] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7450–7459, 2019. 4, 5

[16] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 5

[17] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3D human reconstruction in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019. 2

[18] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J Black, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger, et al. Learning an infant body model from RGB-D data for accurate full body motion analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 792–800, 2018. 2, 5

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

[20] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 3, 4

[21] Karim Iskakov, Egor Burkov, Victor S. Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *International Conference on Computer Vision*, pages 7717–7726, 2019. 2

[22] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1465–1472. IEEE, 2011. 3

[23] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. *arXiv preprint arXiv:2004.03686*, 2020. 3, 4, 6, 7, 8

[24] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):190–204, 2019. 3, 4

[25] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 2, 3, 6, 7

[26] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: video inference for human body pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5252–5262, 2020. 2, 3, 8

[27] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision*, pages 2252–2261, 2019. 2, 6, 7

[28] V. Leroy, Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, and Grégory Rogez. SMPLy benchmarking 3d human pose estimation in the wild. *International Conference on 3DVision*, pages 301–310, 2020. 3, 4

[29] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 2D scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 6

[30] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4521–4530, 2019. 3

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

[32] Yebin Liu, Juergen Gall, Carsten Stoll, Qionghai Dai, H Seidel, and Christian Theobalt. Markerless motion capture of multiple characters using multi-view image segmentation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, pages 2720–2735, 2013. 3

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2, 3, 6

[34] 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 *International Conference on 3DVision*, pages 506–516, 2017. 3, 4

[35] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *International Conference on 3DVision*, pages 120–130, 2018. 3, 4

[36] Gyeongsik Moon, Juyong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. In *International Conference on Computer Vision*, pages 10132–10141, 2019. 2

[37] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2, 4, 6, 7

[38] Albert Pumarola, Jordi Sanchez-Riera, Gary P. T. Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3Dpeople: Modeling the geometry of dressed humans. In *International Conference on Computer Vision*, pages 2242–2251, 2019. 3, 4

[39] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3D human pose estimation. In *International Conference on Computer Vision*, pages 4341–4350, 2019. 2

[40] Anurag Ranjan, David T Hoffmann, Dimitrios Tzionas, Siyu Tang, Javier Romero, and Michael J Black. Learning multi-human optical flow. *International Journal of Computer Vision*, pages 873–890, 2020. 3

[41] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6):1–17, 2017. 6, 8

[42] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3D hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020. 6, 7

[43] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision*, pages 2304–2314, 2019. 5

[44] Akash Sengupta, Roberto Cipolla, and Ignas Budvytis. Synthetic training for accurate 3D human pose and shape estimation in the wild. In *British Machine Vision Conference*, 2020. 3, 4

[45] Yeji Shen and C-C Jay Kuo. Multi-view matching (MVM): Facilitating multi-person 3D pose estimation learning with action-frozen people video. In *arXiv preprint*, 2020. 3

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

[47] Yu Sun, Qian Bao, Wu Liu, Yili Fu, and Tao Mei. CenterHMR: a bottom-up single-shot method for multi-person 3D mesh recovery from a single image. *arXiv preprint arXiv:2008.12272*, 2020. 2, 6, 7

[48] Matt Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3D human pose estimation fusing video and inertial sensors. In *2017 British Machine Vision Conference (BMVC)*, 2017. 3, 4

[49] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing*, pages 5236–5246. Curran Associates, Inc., 2017. 3

[50] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4627–4635, 2017. 3, 4

[51] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *ACM SIGGRAPH 2008 papers*, pages 1–9. 2008. 3

[52] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision*, pages 614–631, 2018. 2, 3, 4, 6

[53] Frank Yu, Mathieu Salzmann, Pascal Fua, and Helge Rhodin. PCLs: Geometry-aware neural reconstruction of 3D pose with perspective crop layers. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 7

[54] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. HUMBI: A large multiview dataset of human body expressions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2987–2997, 2020. 3, 4

[55] Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human synthesis and scene compositing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12749–12756, 2020. 3

[56] Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4191–4200, 2017. 2, 4

[57] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul L Rosin, Zixi Cai, Han Xi, Dingcheng Yang, Hao-Zhi Huang, and Shi-Min Hu. Pose2Seg: Detection free human instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 889–898, 2019. 3

[58] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7376–7385, 2020. 3, 4

[59] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. Generating 3D people in scenes without people. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6193–6203, 2020. 3

[60] Tyler Zhu, Per Karlsson, and Christoph Bregler. SimPose: Effectively learning densepose and surface normals of people from simulated data. In *European Conference on Computer Vision*, pages 225–242, 2020. 3