

BLADE: Single-view Body Mesh Estimation through Accurate Depth Estimation

Supplemental Material

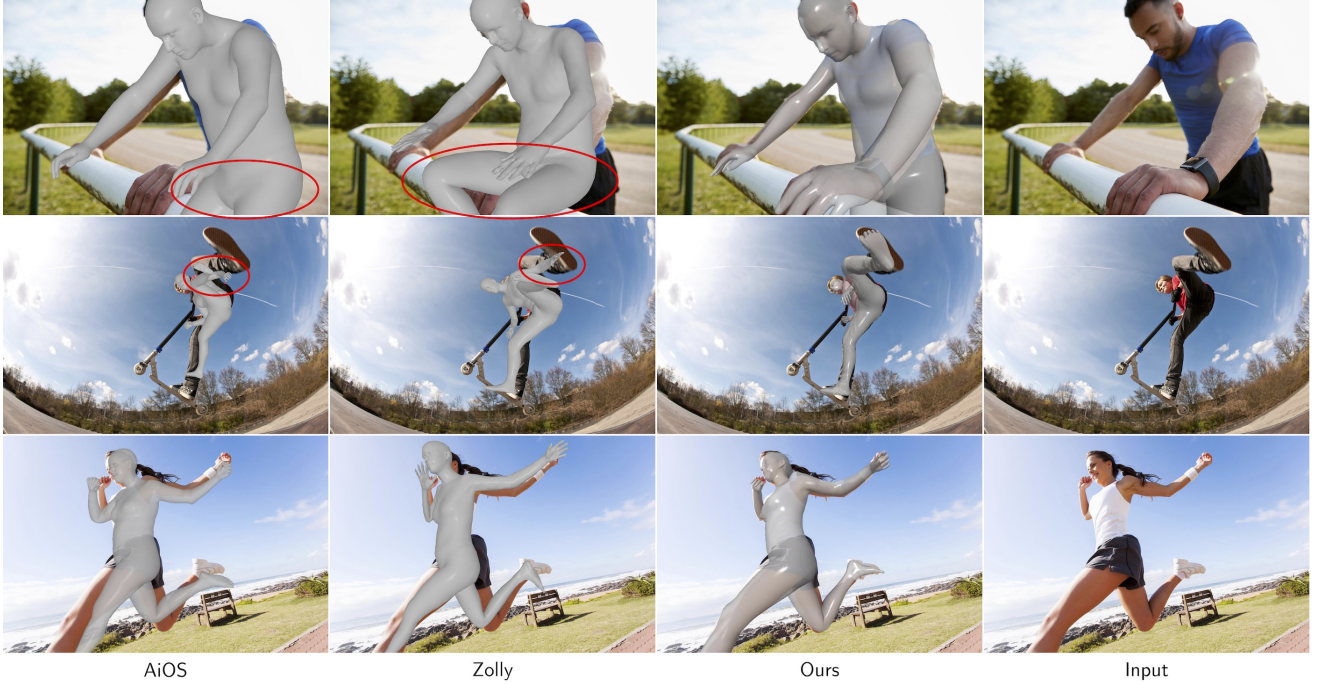


Figure A1. **More Qualitative Results.** BLADE not only achieves accurate 3D pose estimation, but also accurately recovers perspective projection parameters and thus achieves state-of-the-art alignment accuracy in image space.

A1. Overview

In this supplemental document, we (1) provide additional qualitative results on real-world images (Sec. A2); (2) examine the existing evaluation datasets and identify the need for a close-range evaluation dataset with accurate labels (Sec. A3); (3) report additional quantitative results of the various methods on more datasets and with additional metrics (Sec. A4); (4) elaborate on the ambiguity involved in single-image-based 3D human mesh recovery (Sec. A5); and (5) discuss the trade-off between achieving high depth estimation accuracy on close-range data versus far-range data (Sec. A6).

A2. Qualitative Results on Real-World Images

In Fig. A1, A4 and A5, we show more visual results with a comparison to recent state-of-the-art methods AiOS [21] and Zolly [23]. We achieve significant improvement in terms of alignment of the rendered 3D mesh to the input image, accuracy of perspective distortion, as well as

the estimated 3D pose. For example, in the first row of Fig. A1, only our method correctly estimates the camera’s close proximity to the person’s hand and that the person is standing, whereas AiOS and Zolly predict incorrect leg postures and distances to the person. In the second row of Fig. A1, both AiOS and Zolly wrongly estimate the person’s left hand behind their body, whereas BLADE recovers the correct position of the person’s hand and camera’s proximity to the person’s feet. A similar phenomenon can be observed in Fig. A4, A5, A6, and A7 as well.

Interestingly, Zolly [23] sometimes generates flattened meshes. For example, in the second image from top left in Fig. A4, Zolly predicts a mesh where the person’s head and arms are flattened. This is because, different from AiOS and our methods, Zolly directly predicts a mesh instead of parameters of the SMPL-X model. While this design gives Zolly more flexibility in generating difficult shapes, it can also lead to degenerate estimation at times.

Additionally, although BLADE leverages AiOS [21] as part of the pose estimator backbone, BLADE improves

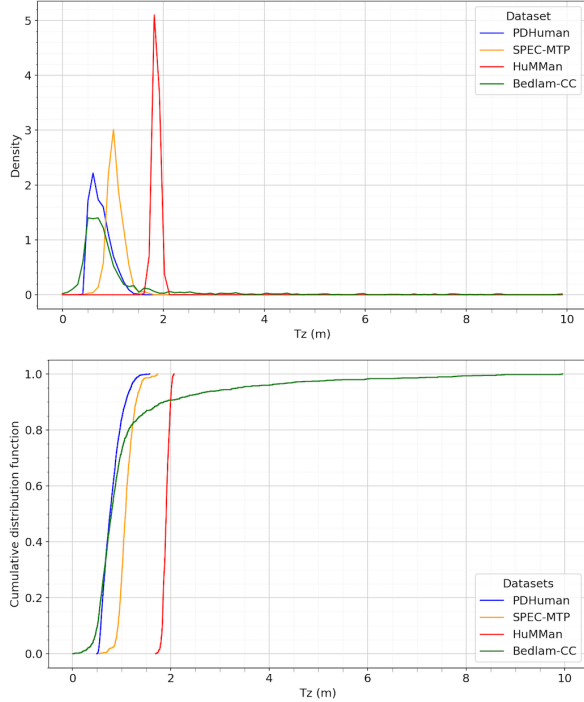


Figure A2. **Evaluation Dataset Distributions.** In the top diagram, we show the distribution of T_z values across different datasets. Notably, the majority of the HUMMAN dataset has T_z values concentrated in a small range around 1.9m. The HUMMAN dataset thus has much less perspective distortion compared to close-range datasets like the SPEC-MTP[13], PDHUMAN[23], and our BEDLAM-CC dataset. In the bottom, we show the cumulative distribution function of T_z values across datasets. Notably, our BEDLAM-CC dataset has a wider range of T_z values, and even smaller minimum T_z values than PDHUMAN. These traits make BEDLAM-CC a diverse evaluation dataset that is particularly well-suited for close-range HMR.

AiOS’ pose and shape accuracy. For example, in the top left of Fig. A6, BLADE predicts the person’s body shape more accurately than AiOS. In the second and bottom row in Fig. A6, predictions of the person’s legs from AiOS and Zolly are both wrong whereas BLADE shows robustness in both situation. In the top row of Fig. A7, BLADE correctly recovers both the orientation and the leg posture of the person, whereas AiOS does not. In the second row of Fig. A7, BLADE correctly recovers the position and angle of the person’s ankles, whereas predictions from AiOS are inaccurate.

A3. Examining the Evaluation Datasets

In this section, we examine the strengths and shortcomings of various standard benchmark datasets used to evaluate the task of single-image-based human mesh recovery (HMR). We find that there is a lack of close-range test data with

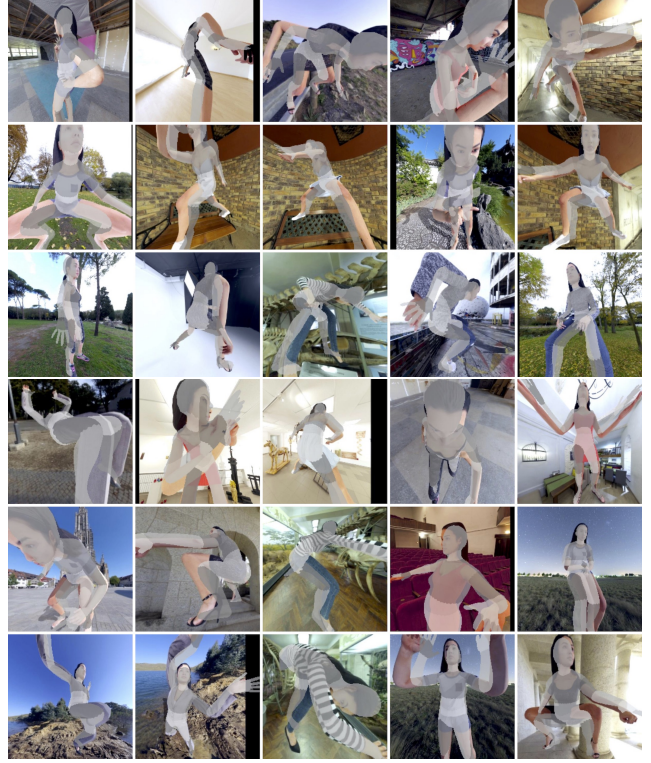


Figure A3. **Inaccurate Pose Labels in PDHuman[23].** We find that a high percentage of pose labels in PDHuman do not align with the corresponding images. In the above examples, we visualize the SMPL labels superimposed on top of the corresponding images. The SMPL renderings (gray overlays) are generated using the authors’ original code base used for IoU calculations.

accurate ground truth annotations, and we thus introduce BEDLAM-CC to fill this void.

In Fig. A2, we show the distribution of T_z , *i.e.* the depth of the pelvis of a person, across different datasets. As mentioned in Fig. 3 (main paper), T_z has significant impact on the level of perspective distortion observed in an image and becomes more impactful to 3D HMR, the closer the person gets to the camera. An ideal evaluation dataset for HMR of strongly perspective images should thus contain a large number of samples with persons within close-range to the camera, which we loosely define to be less than 1.5 meter.

HuMMan [4]: This dataset is captured in a studio environment. A person stands in the middle of a circle of cameras and performs different actions. This dataset is useful for performing 3D reconstruction on human subjects due to its multi-view camera setup. However, it is very limited in terms of visual diversity due to it being captured in the same studio environment. More importantly, as shown in Fig. A2 (red distribution), this dataset contains very limited variation in terms of T_z , distributed closely around 1.9m, farther from the close range of <1.5 m distance. Therefore, due to

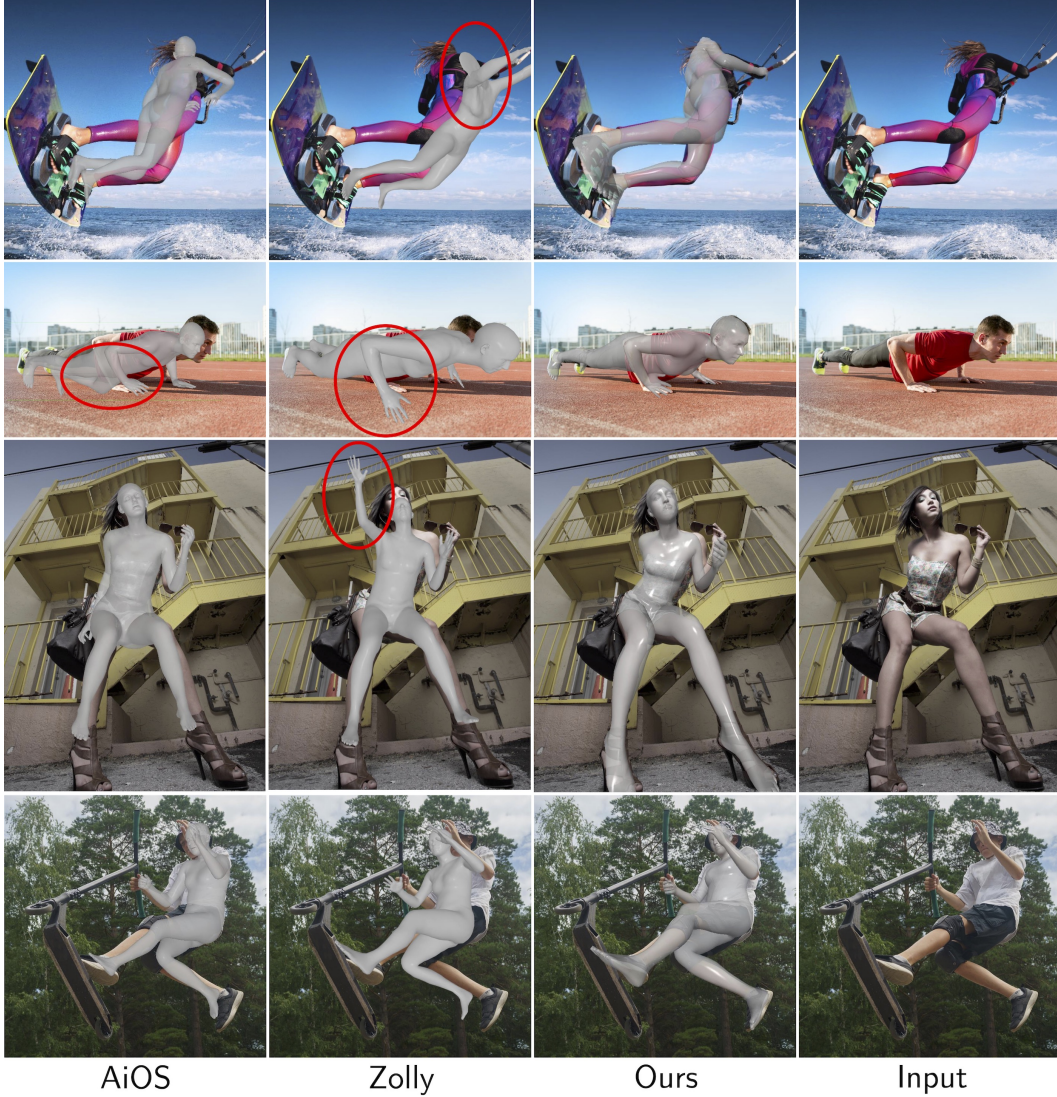


Figure A4. **More Qualitative Results.** In addition to achieving accurate pose estimation, our method BLADE recovers precise perspective projection parameters, ensuring the predicted 3D human mesh is well-aligned with the input image.

its limited visual diversity, T_z variation, and the absence of close-range data with $T_z < 1.5\text{m}$, this dataset is not ideal for evaluating close-range HMR methods intended to operate on images in-the-wild. Performance on it, thus, is not reflective of performance on highly unconstrained images in the real world.

SPEC-MTP [13]: This dataset is captured using smartphones in the real world with diverse identities, lighting conditions, and poses. It is captured by having one person move the camera around another person as they pose for the camera. 3D pose labels are then generated from the video frames. As shown in Fig. A2 (yellow distribution), SPEC-MTP’s T_z values fall within the desired 1.5m threshold and center around 1m. This T_z distribution and

the appearance diversity from real-world capture settings makes SPEC-MTP[13] a good dataset for evaluating close-range HMR methods. We find the provided labels to be mostly accurate, while inevitable errors in calibration and video-based reconstruction lead to inaccurate pose labels in a small portion of the test samples.

PDHuman [23]: This is a synthetic dataset generated using 630 photogrammetry-scanned human models from Renderpeople [2] and animated using Mixamo [1]. 3D labels are converted to SMPL by optimizing for a set of pose and shape parameters that best fit the 3D human models used to generate the rendered data. As shown in Fig. A2 (blue distribution), PDHuman’s T_z values are mostly within 1m, leading to high levels of perspective distortion in this

dataset. However, we find that a high percentage of its pose labels are inaccurate with respect to the input images. In Fig. A3, we visualize the SMPL labels overlaid on top of the corresponding images. The SMPL renderings (gray overlays) are generated by using the scripts provided for IoU calculations in the authors’ original code base. We postulate that this inaccuracy may have been the result of inaccurate conversion from the animated RenderPeople models to SMPL.

Considering that quantitative results on PDHuman may not also correctly reflect actual performance, we conclude that there is a lack of accurate and diverse data to quantitatively benchmark performance of close-range HMR for images taken at a T_z depth closer than 1m. Therefore, we curate a new dataset with accurate labels to facilitate evaluation of close-range HMR.

A3.1. BEDLAM-CC: A Close-Range Synthetic Dataset with Accurate 3D Labels

We create a new close-range evaluation dataset utilizing assets provided with the BEDLAM dataset [3] and name our dataset BEDLAM-CC. As discussed in the main paper, perspective distortion is non-linear w.r.t. the distance between the camera and the subject [18]. In particular, it changes rapidly when the distance gets closer (0.3m to 1.2m), because of its inverse relationship to distance. The perspective projection gradually approximates orthographic projection at distances of 5m and higher. Therefore, to concentrate our evaluation on close-range HMR, we enforce that 80% of our dataset locates T_z within the range of $0.5\text{m} \leq T_z \leq 1.2\text{m}$ and the remaining samples are in the range of $1.2\text{m} < T_z \leq 10\text{m}$. From the 2 million generated images there are a total of 1314 images in the evaluation split.

We carefully curate the camera poses in our dataset to generate images with diverse viewpoints relative to the person. With a T_z value being sampled as described above, the camera is positioned on a sphere with the radius given by T_z and randomly sampled spherical coordinates $\theta \in [0, 2\pi]$ and $\phi \in [0.1\pi, 0.7\pi]$, where θ is the azimuth angle and ϕ represents the elevation. The camera rotation is evaluated by a LookAt() function towards a randomized target bone along the SMPL-X spine given by a randomized bone index $i \in [0, 3, 6, 9, 12, 15]$ and an added random noise vector $v \in \mathbb{R}^3$. To keep the person at a reasonable size relative to the frame we set the focal length using a dolly zoom with a default value f_d of 15mm at 1m distance with a camera sensor size of 36x36mm. We then uniformly randomize the focal length $f_{GT} \in [0.7, 1.3] \cdot f_d$. In addition, we randomize the lighting setup including skylight (background image and intensity), and directional sun light (position, color, intensity). We show example images of our BEDLAM-CC dataset in Figure A11. Since our dataset is generated through SMPL-X and Unreal Engine, we do not

need to convert the data to SMPL-X format and thus avoid conversion errors.

A4. Additional Quantitative Results

In this section, we report additional quantitative results for various evaluation datasets using more metrics. Specifically, we test the various methods on the SPEC-MTP [13], PDHUMAN [23], BEDLAM-CC, and HUMMAN [4] datasets. We use the commonly used metrics, including, Mean Per-Joint Position Error (MPJPE), Procrustes Analysis Mean Per-Joint Position Error (PAMJPPE), Per-Vertex Error (PVE), mean Intersection over Union (mIoU), and Body Part mean Intersection over Union (P-mIoU). As discussed in the main paper, we introduce new metrics to evaluate the accuracy of recovered perspective projection parameters. Specifically, we measure the accuracy of the recovered focal length as its percentage error relative to the ground truth focal length:

$$E_f = |f_{pred} - f_{GT}| / f_{GT}. \quad (1)$$

Given that T_z has an inverse relationship with respect to the amount of distortion in the image (Fig. 3, main paper), whereas (T_x, T_y) do not, we separately evaluate T_z and (T_x, T_y) errors as E_{T_z} and $E_{T_{xy}}$ in meters. Additionally, since T_z ’s accuracy is less important at far distances, we also calculate an inverse T_z error E_{1/T_z} , reflecting this property:

$$E_{T_{xy}} = \|T_{xy}^{pred} - T_{xy}^{GT}\|_2, \quad (2)$$

$$E_{T_z} = |T_z^{pred} - T_z^{GT}|, \quad (3)$$

$$E_{1/T_z} = |1/T_z^{pred} - 1/T_z^{GT}|. \quad (4)$$

In Table. A1, we show that BLADE achieves state-of-the-art accuracy for a majority of the metrics across the four datasets: SPEC-MTP[13], PDHUMAN[23], BEDLAM-CC, and HUMMAN[4]. Among these SPEC-MTP[13], PDHUMAN[23], and BEDLAM-CC are perspective distorted datasets with many persons with $T_z < 1.5\text{m}$. On perspective distorted datasets, BLADE is state of the art in terms of recovering accurate perspective projection parameters (measured by E_{T_z} , E_{1/T_z} , $E_{T_{xy}}$, and E_f) and accurate 3D mesh recovery (measured by PVE). Additionally, BLADE achieves joint accuracies (measured by PAMJPPE and MPJPE) better than or comparable to state-of-the-art methods. The accurate recovery of projection parameters and 3D geometry results in state-of-the-art alignment from the rendered mesh to the input image. This is shown by BLADE’s significantly higher mIoU and P-mIoU performances. For example, on SPEC-MTP[13], BLADE’s mIoU is 69.9%, whereas the second best method PARE[12] achieves 55.8%. Similarly, on PDHUMAN [23] and BEDLAM-CC, BLADE achieves mIoU values of 67.3%



Figure A5. **More Qualitative Results.** Beyond accurate pose estimation, our approach BLADE effectively reconstructs perspective projection parameters, allowing the predicted 3D human mesh to align closely with the input image.

and 72.8%, respectively, whereas the second best methods achieve 53.0% and 54.6%. Moreover, BLADE consistently achieves high IoU values of around 70%, whereas prior methods show significant degradation on the three perspective distorted datasets. On the less distorted HUMAN[4] dataset, we achieve state-of-the-art accuracy on T_z estimation (E_{T_z} , E_{1/T_z}) and focal length estimation (E_f). BLADE achieves significantly better joint precisions (PA-MPJPE, MPJPE) and 3D mesh reconstruction than the recent state-of-the-art methods (AiOS[21], SMPLer-X[5], and TokenHMR[7]) and is comparable to Zolly.

In addition to the close-range benchmarks, we evaluate the methods on 3DPW [22] and HUMAN3.6M [9], which are the captured farther away and thus less perspectively

distorted. These two datasets have average pelvis depths at around 5m and thus exhibit much less perspective distortion of the person. We tested all models on 3DPW without training with 3DPW. Additionally, as 3DPW images are often crowded with people and the test subjects can be very far away from the camera and thus tiny in the image, we retrain a version of BLADE that processes cropped images instead of the original full images. Although our model is optimized for close-range pose estimation, our method also outperforms recent state-of-the-art methods (AiOS, SMPLer-X, Zolly) on the farther-away 3DPW dataset, demonstrating its robustness to different use cases. On HUMAN3.6M, we perform similar to the SOTA models, but we notice that the HUMAN3.6M test set only contains 2 subjects and

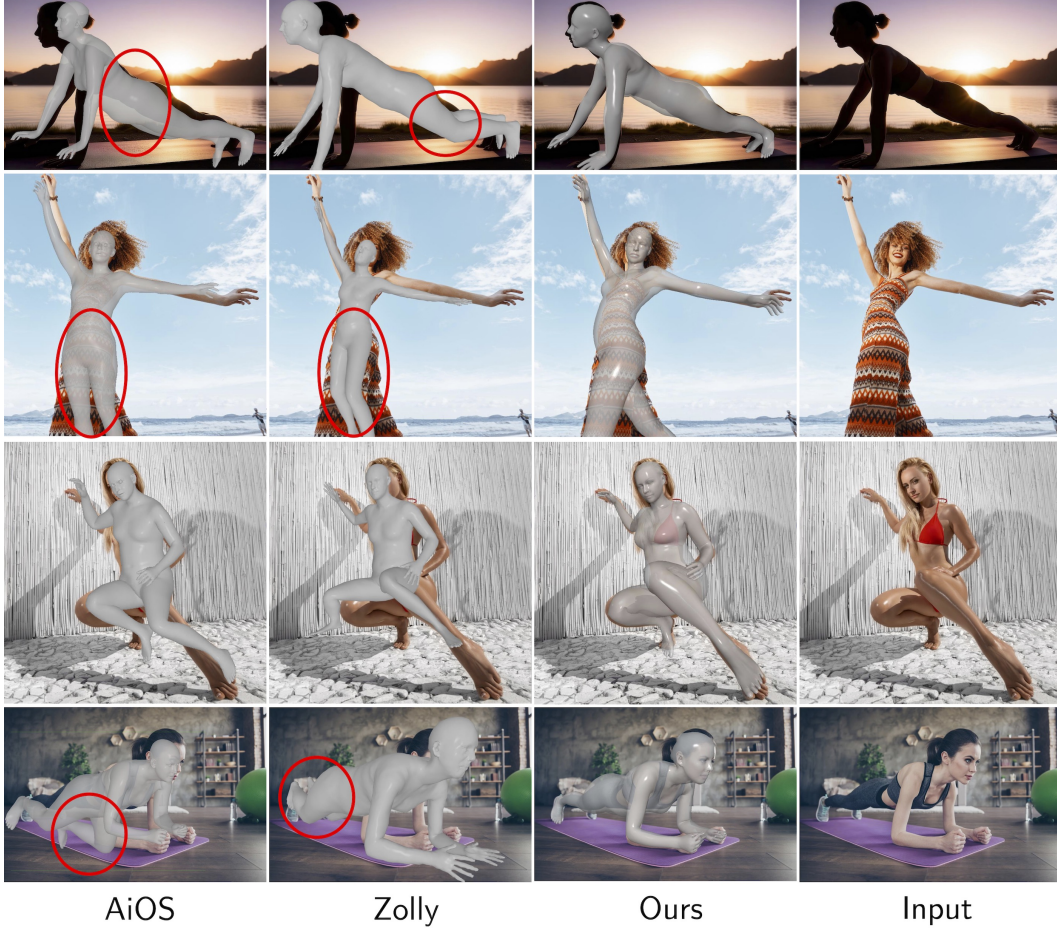


Figure A6. **More Qualitative Results.** Our approach BLADE not only estimates 3D shape and pose precisely but also accurately retrieves perspective projection parameters, enabling the predicted 3D human mesh to align seamlessly with the input image.

might not be a representative benchmark.

A5. Single-Image Ambiguity in 3D Human Mesh Recovery (3D HMR)

In Fig. A9 and A10, we visually illustrate the ambiguity in single-image human mesh recovery. To achieve both accurate 3D mesh recovery and 2D alignment, one needs to solve for both the 3D mesh of the person as well as the camera intrinsic and extrinsic parameters. However, given that none of the aforementioned parameters is known, and that they are heavily entangled, this problem is well known to be ill-posed and has potentially infinite solutions. For example, as shown in Fig. A9, it is difficult for a model to correctly predict the two poses from the input images because it has no information about the shape of the person’s legs and shoes. Moreover, due to the nature of projected geometry, the reconstructions are always up to scale unless additional knowledge of scale is provided, *e.g.* the camera’s movement is measured in physical units. For example, as shown in Fig. A10, images of people of different sizes can

result in very similar images. Therefore, the reverse problem of reconstructing the person from the images can also result in 3D meshes of different sizes.

Human Height Bias While the aforementioned ambiguities are inherent to the problem, much prior work [17, 19] have leveraged the regularity of the human body to arrive at reasonable solutions for this ill-posed problem. For example, one such regularity [20] is that 95% of men have a height between 163.2cm and 193.6cm and 95% of women have a height between 150.6cm and 178.84cm. However, it is possible that the model learns a very narrow range of human height to make the problem trivial to solve. Therefore, we visualize the height distribution of BEDLAM-CC in Fig. A8 (left) along with the human population distribution. The height distribution in our dataset BEDLAM-CC, which uses neutral SMPL-X, is quite similar to the combination of adult male and female human population (in meters: BEDLAM-CC: mean 1.714, $\sigma=0.095$, real-world male population: 1.777, $\sigma=0.078$, female population: 1.662, $\sigma=0.067$ [20]). Furthermore, both our ground truth test data

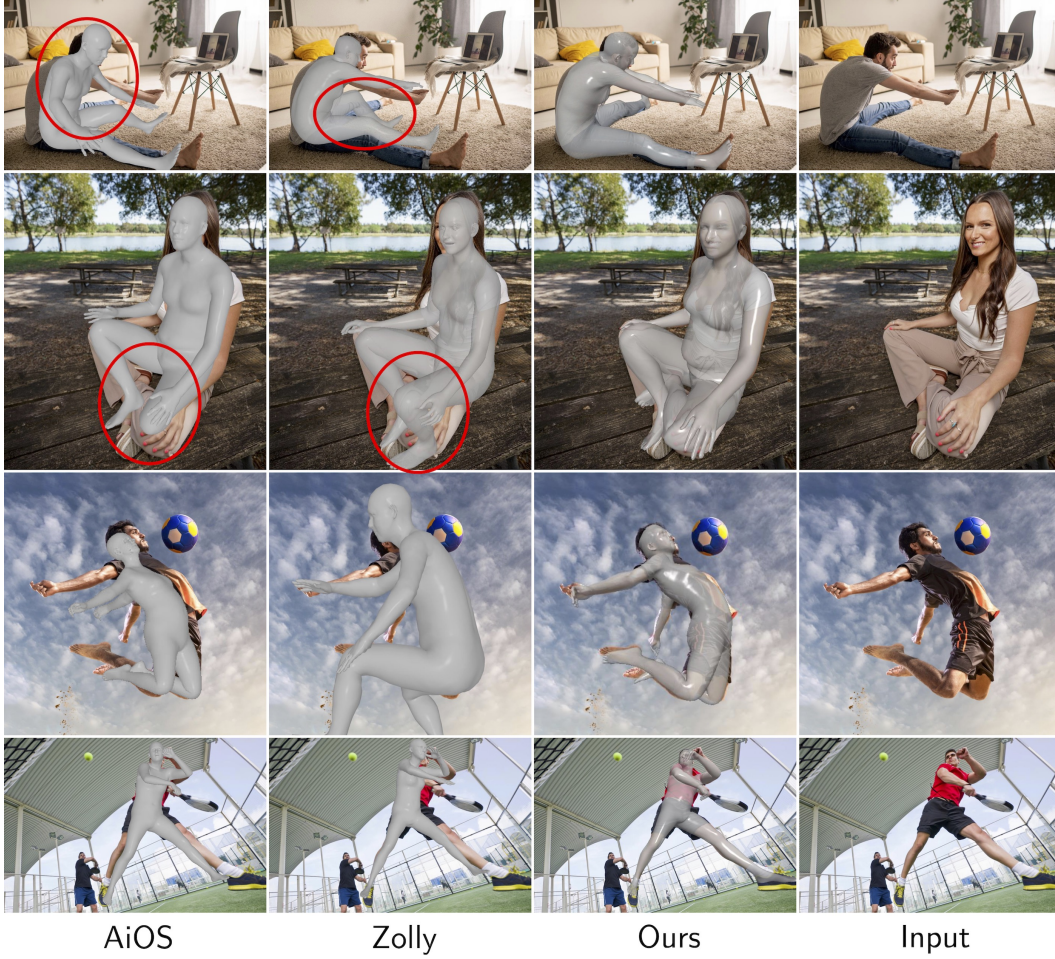


Figure A7. **More Qualitative Results.** BLADE not only achieves accurate pose estimation, but also recovers accurate perspective projection parameters and thus can align the predicted 3D human mesh to the input image well.

and BLADE’s predictions (Fig. A8, right) have significant diversity. However, we recognize that more samples should be added for heights $< 1.5\text{m}$ and $> 1.85\text{m}$ and that the distribution of heights should be better normalized without gaps.

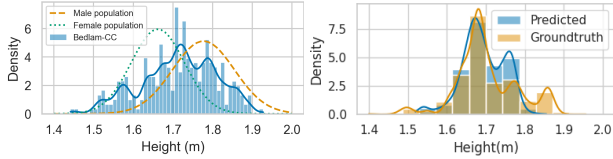


Figure A8. Left: Height distribution of BEDLAM-CC against world adult population [Roser et al., 2021] Right: Body height distribution of ground truth and prediction from BLADE on SPEC-MTP, PDHuman, and HuMMan.

A6. Trade-Off between Close and Far Range T_z Estimation

For T_z estimators trained without our BEDLAM-CC dataset, we observe that it is difficult for them to achieve accurate

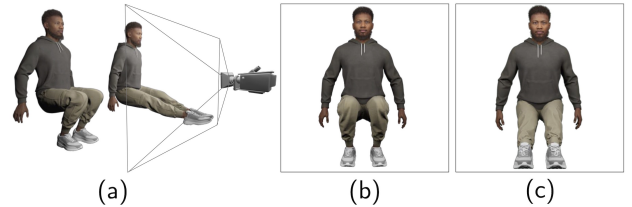


Figure A9. **The Ambiguity of Single Image 3D Human Pose Estimation.** Although being significantly different in pose and distance to the camera (a) both presented configurations result in similar camera views (b, c). Therefore, due to the ill-posed nature of single-image 3D pose estimation, different combinations of pose and camera distance can result in valid but incorrect reconstructions.

T_z estimation for both close and far range images. The various T_z estimators with different backbones oscillate between achieving high accuracy on close-range or on far-

| Methods | SPEC-MTP [13] (real-world capture) | | | | | | | | | | PDHUMAN [23] (synthetic) | | | | | | | | | |
|--------------------|------------------------------------|-----------------------|------------------------|-----------------|-------------|-------------|-------------|-------------|-------------|---------------------|--------------------------|------------------------|-----------------|-------------|-------------|-------------|-------------|-------------|--|--|
| | $E_{T_z}\downarrow$ | $E_{1/T_z}\downarrow$ | $E_{T_{xy}}\downarrow$ | $E_f\downarrow$ | PA-MPJPE↓ | MPJPE↓ | PVE↓ | mIoU↑ | P-mIoU↑ | $E_{T_z}\downarrow$ | $E_{1/T_z}\downarrow$ | $E_{T_{xy}}\downarrow$ | $E_f\downarrow$ | PA-MPJPE↓ | MPJPE↓ | PVE↓ | mIoU↑ | P-mIoU↑ | | |
| HMR [10] | - | - | - | - | 73.9 | 121.4 | 145.6 | 48.8 | 16.0 | - | - | - | - | 62.5 | 91.5 | 106.7 | 48.9 | 21.7 | | |
| HMR- <i>f</i> [10] | - | - | - | - | 72.7 | 123.2 | 145.1 | 52.3 | 20.1 | - | - | - | - | 61.6 | 90.2 | 105.5 | 45.2 | 20.4 | | |
| SPEC [13] | - | - | - | - | 76.0 | 125.5 | 144.6 | 49.9 | 18.8 | - | - | - | - | 65.8 | 94.9 | 109.6 | 43.4 | 19.6 | | |
| CLIFF [16] | - | - | - | - | 74.3 | 115.0 | 132.4 | 53.6 | 23.7 | - | - | - | - | 66.2 | 99.2 | 115.2 | 51.4 | 24.8 | | |
| PARE [12] | - | - | - | - | 74.2 | 121.6 | 143.6 | 55.8 | 23.2 | - | - | - | - | 66.3 | 95.9 | 116.7 | 48.2 | 20.9 | | |
| GraphCMR [14] | - | - | - | - | 76.1 | 121.4 | 141.6 | 53.5 | 22.0 | - | - | - | - | 62.0 | 85.8 | 98.4 | 47.9 | 21.5 | | |
| FastMETRO [6] | - | - | - | - | 75.0 | 123.1 | 137.0 | 53.5 | 20.5 | - | - | - | - | 58.6 | 83.6 | 95.4 | 50.1 | 22.5 | | |
| Zolly [23] | 0.899 | 0.394 | 0.906 | 106.3 | 67.4 | 114.6 | 126.7 | 62.3 | 30.4 | 0.255 | 0.355 | 0.051 | 27.3 | 49.9 | 70.7 | 82.0 | 53.0 | 26.5 | | |
| SMPLer-X* | 0.980 | 0.450 | 0.109 | 112.1 | 55.5 | 90.9 | 102.6 | 53.0 | 15.9 | 2.223 | 1.030 | 0.126 | 55.0 | 96.8 | 148.2 | 161.2 | 47.6 | 17.1 | | |
| TokenHMR* | 0.909 | 0.436 | 0.095 | 112.1 | 64.2 | 107.1 | 124.3 | 49.8 | 19.0 | 2.280 | 1.034 | 0.068 | 55.0 | 92.1 | 141.5 | 156.7 | 53.0 | 27.8 | | |
| AiOS* | 1.035 | 0.464 | 0.121 | 112.1 | 62.8 | 101.6 | 110.9 | 48.7 | 11.3 | 2.312 | 1.024 | 0.149 | 55.0 | 106.6 | 170.6 | 183.4 | 49.5 | 16.0 | | |
| Ours | 0.129 | 0.114 | 0.056 | 16.3 | 61.0 | 105.3 | 111.9 | 68.6 | 39.8 | 0.106 | 0.176 | 0.043 | 21.6 | 49.6 | 69.7 | 80.5 | 67.3 | 44.6 | | |
| Ours (real-world) | 0.127 | 0.112 | 0.044 | 15.9 | 56.7 | 94.1 | 99.6 | 69.9 | 41.5 | 0.107 | 0.178 | 0.049 | 22.3 | 61.4 | 90.1 | 102.6 | 65.2 | 41.4 | | |

| | BEDLAM-CC (synthetic) | | | | | | | | | | HUMMAN [4] (studio capture) | | | | | | | | | |
|--------------------|-----------------------|-----------------------|------------------------|-----------------|-------------|-------------|--------------|-------------|-------------|---------------------|-----------------------------|------------------------|-----------------|-------------|-------------|-------------|-------------|-------------|--|--|
| | $E_{T_z}\downarrow$ | $E_{1/T_z}\downarrow$ | $E_{T_{xy}}\downarrow$ | $E_f\downarrow$ | PA-MPJPE↓ | MPJPE↓ | PVE↓ | mIoU↑ | P-mIoU↑ | $E_{T_z}\downarrow$ | $E_{1/T_z}\downarrow$ | $E_{T_{xy}}\downarrow$ | $E_f\downarrow$ | PA-MPJPE↓ | MPJPE↓ | PVE↓ | mIoU↑ | P-mIoU↑ | | |
| HMR [10] | - | - | - | - | - | - | - | - | - | - | - | - | - | 30.2 | 43.6 | 52.6 | 65.1 | 39.5 | | |
| HMR- <i>f</i> [10] | - | - | - | - | - | - | - | - | - | - | - | - | - | 29.9 | 43.6 | 53.4 | 62.7 | 34.9 | | |
| SPEC [13] | - | - | - | - | - | - | - | - | - | - | - | - | - | 31.4 | 44.0 | 54.2 | 51.4 | 25.6 | | |
| CLIFF [16] | - | - | - | - | - | - | - | - | - | - | - | - | - | 28.6 | 42.4 | 50.2 | 68.8 | 44.7 | | |
| PARE [12] | - | - | - | - | - | - | - | - | - | - | - | - | - | 32.6 | 53.2 | 65.5 | 66.5 | 38.3 | | |
| GraphCMR [14] | - | - | - | - | - | - | - | - | - | - | - | - | - | 29.5 | 40.6 | 48.4 | 61.6 | 37.5 | | |
| FastMETRO [6] | - | - | - | - | - | - | - | - | - | - | - | - | - | 26.3 | 38.8 | 45.5 | 68.3 | 45.2 | | |
| Zolly [23] | 0.539 | 0.634 | 0.081 | 46.1 | 68.8 | 107.8 | 131.8 | 51.8 | 21.2 | 0.228 | 0.072 | 0.034 | 9.4 | 22.3 | 32.6 | 40.0 | 71.2 | 45.1 | | |
| SMPLer-X* | 2.057 | 1.172 | 0.087 | 134.9 | 69.5 | 120.3 | 140.0 | 53.0 | 21.3 | 2.461 | 0.300 | 0.125 | 41.6 | 38.7 | 56.4 | 65.8 | 51.8 | 11.1 | | |
| TokenHMR* | 2.378 | 1.200 | 0.096 | 134.9 | 59.9 | 114.3 | 136.4 | 54.1 | 22.3 | 2.599 | 0.307 | 0.044 | 41.6 | 46.4 | 72.2 | 82.0 | 60.9 | 31.1 | | |
| AiOS* | 2.340 | 1.197 | 0.111 | 134.9 | 71.6 | 125.7 | 143.0 | 54.6 | 19.9 | 2.311 | 0.292 | 0.033 | 41.6 | 66.1 | 91.8 | 99.4 | 72.0 | 44.3 | | |
| Ours | 0.326 | 0.306 | 0.066 | 26.2 | 59.4 | 90.5 | 111.6 | 72.7 | 44.5 | 0.188 | 0.058 | 0.055 | 8.5 | 24.9 | 44.4 | 56.3 | 69.8 | 37.9 | | |
| Ours (real-world) | 0.325 | 0.305 | 0.065 | 25.7 | 57.8 | 85.8 | 106.8 | 72.8 | 44.5 | 0.187 | 0.058 | 0.056 | 8.3 | 23.8 | 41.1 | 52.3 | 70.6 | 38.2 | | |

Table A1. Results of SOTA methods on the SPEC-MTP [13], PDHUMAN [23], BEDLAM-CC, and HUMMAN [4] datasets. For baselines at the top of the tables, we use the results reported by Zolly [23] and omit the ones not available. Additionally, we re-evaluate newer state-of-the-art methods AiOS [21], SMPLer-X [5], and TokenHMR [7]. These models are noted using “*”.

| 3DPW (not in training, avg. $T_z=4.6m$) | | | | PA-MPJPE \downarrow | MPJPE \downarrow | PVE \downarrow |
|--|--|--|--|-----------------------|--------------------|------------------|
| AiOS | | | | 50.5 | 80.8 | 95.1 |
| SMPLer-X | | | | 49.5 | 88.2 | 92.3 |
| Zolly | | | | 47.9 | 76.2 | 89.8 |
| Ours (BLADE) | | | | 45.7 | 75.1 | 89.5 |

| H3.6M (in training, avg. $T_z=5.1m$) | | | | PA-MPJPE \downarrow | MPJPE \downarrow | PVE \downarrow |
|---------------------------------------|--|--|--|-----------------------|--------------------|------------------|
| AiOS | | | | 46.3 | 68.9 | no |
| SMPLer-X | | | | 38.9 | 75.3 | ground |
| Zolly | | | | 32.3 | 49.4 | truth |
| Ours (BLADE) | | | | 40.5 | 55.1 | |

Table A2. Results of SOTA methods on less perspectively distorted datasets (3DPW [22] and HUMAN3.6M [9]). All methods are evaluated without training on 3DPW.

range images, exemplified by their accuracies on the close range dataset SPEC-MTP [13] and the farther range dataset HUMMAN [4]. For example, when using Sapiens[11] as the backbone for our T_z estimator, its best T_z error on SPEC-MTP[13] is 21cm, but it scores a high T_z error on HUMMAN. On the other hand, using a model checkpoint with a low T_z error of 60cm on HUMMAN results in an 85cm error on SPEC-MTP. Similarly, when using DepthAnythingV2 [24] as the backbone, our T_z estimator can achieve a low T_z error of 15.4cm on SPEC-MTP [13], but at the same time suffers from a high T_z error of 23cm on HUMMAN [4]. When using a checkpoint that can achieve 3.1cm T_z error on HUMMAN, the model in turn suffers from a high T_z error of 67.6 on SPEC-MTP.

Inspired by recent works in monocular depth estima-

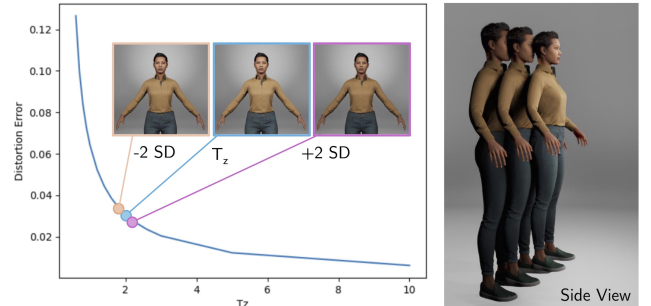


Figure A10. **Ambiguous Human Size from a Single Image.** The problem of metric-scale mesh estimation problem is inherently ill-posed, and capturing people of different sizes from different distances can result in similar images. The side view reveals the actual sizes of the subjects and their distances T_z to the camera. When the image of a taller person captured farther away can be similar to the image of a shorter person captured at a closer distance. The corresponding T_z values are also shown on the left. However, given that the heights of 95 % of all human [8] (± 2 standard deviations) lie within a small range, the size variation thus correspond to a narrow T_z variation as shown on the left curve. The mean size is the blue inset and the range of ± 2 standard deviations are shown as yellow and violet insets.

tion [24, 25], we focus on providing the networks with more high-quality close-range training samples by curat-



Figure A11. **Examples of our synthetic BEDLAM-CC dataset.** The strong variation in lighting and camera angles as well as occasional extreme close-up distortion are intentionally part of the data.

ing our own BEDLAM-CC dataset (Sec. A3). With more high-quality close-range training samples, our final T_z estimator achieves a low error of 12.7cm on the close-range dataset SPEC-MTP [13] while maintaining a reasonable T_z error of 18.7cm on the farther-range HUMMAN dataset (Table. A1).

A7. Speed

Currently, our implementation targets quality over speed although not being unreasonably slow. Processing 1 image on an RTX3090 GPU takes 0.009s for depth estimation and 0.33s for pose estimation. For quantitative evaluations, we ran the camera solver for 50 iterations (0.963s) to ensure good accuracy but found that 10 iterations (0.210s) give good accuracy already. The optimization time can be

further improved by using Nvdiffraft [15] instead of PyTorch3d.

Dataset license information. The assets of the BEDLAM dataset [3] have been published by Max Planck Institute for Intelligent Systems under a *No distribution* license¹.

With the publication of our work, we will publish

- our code changes with respect to the BEDLAM dataset to render the BEDLAM-CC dataset, and
- instructions to render the BEDLAM-CC dataset.

For recreation of the BEDLAM-CC dataset, the render pipeline needs to be set up according to the guidelines of the BEDLAM dataset. We will publish our data under license terms to allow usage for research purposes.

Image Sources

- Main Paper Figure 1: Adobe Stock image ids: 16532441, 688449553, 868801378.²
- Main Paper Figure 4: Adobe Stock Image id: 789510049.
- Main Paper Table 1: Row 1-2 Adobe Stock image ids: 415527042, 344928073, 71230339, 605587274. Last row: Images from Zolly [23].
- Figure A1: Adobe Stock image ids: 184701266, 21677394, 60240732.
- Figure A4: Adobe Stock image ids: 859644245, 81892568, 21197764, 902825438.
- Figure A5: Adobe Stock image ids: 892029686, 71230339, 688449514, 615119495.
- Figure A6: Adobe Stock image ids: 1061297360, 765162341, 547882981, 355426702.
- Figure A7: Adobe Stock image ids: 348174880, 583910785, 219801712, 63038620.

References

- [1] Mixamo, 2022. <https://www.mixamo.com/>. 3
- [2] Render People, 2020. <https://hdrihaven.com/>. 3
- [3] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 4, 10
- [4] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. HuMMan: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, pages 557–577. Springer, 2022. 2, 4, 5, 8
- [5] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems*, 2023. 5, 8
- [6] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *European Conference on Computer Vision*, pages 342–359. Springer, 2022. 8
- [7] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 5, 8
- [8] Cheryl D Fryar, Qiuping Gu, and Cynthia L Ogden. Anthropometric reference data for children and adults; United States, 2007-2010. 2012. 8
- [9] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2014. 5, 8
- [10] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 8
- [11] Rawal Khrodgar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2025. 8
- [12] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11127–11137, 2021. 4, 8
- [13] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J Black. SPEC: Seeing people in the wild with an estimated camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11035–11045, 2021. 2, 3, 4, 8, 9
- [14] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4501–4510, 2019. 8
- [15] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 10
- [16] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 8
- [17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 34(6):248:1–248:16, 2015. 6
- [18] Koki Nagano, Huiwen Luo, Zejian Wang, Jaewoo Seo, Jun Xing, Liwen Hu, Lingyu Wei, and Hao Li. Deep face nor-

¹<https://bedlam.is.tuebingen.mpg.de/license.html>

²<https://stock.adobe.com/>

malization. *ACM Transactions on Graphics (TOG)*, 38(6): 1–16, 2019. [4](#)

- [19] 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 *CVPR*, 2019. [6](#)
- [20] Max Roser, Cameron Appel, and Hannah Ritchie. Human height. *Our World in Data*, 2021. <https://ourworldindata.org/human-height>. [6](#)
- [21] Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, et al. AiOS: All-in-One-Stage Expressive Human Pose and Shape Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1834–1843, 2024. [1](#), [5](#), [8](#)
- [22] 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 *ECCV*, 2018. [5](#), [8](#)
- [23] Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, Yanjun Wang, Chunhua Shen, Lei Yang, and Taku Komura. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. *ICCV*, 2023. [1](#), [2](#), [3](#), [4](#), [8](#), [10](#)
- [24] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. *arXiv:2406.09414*, 2024. [8](#)
- [25] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. 2023. [8](#)