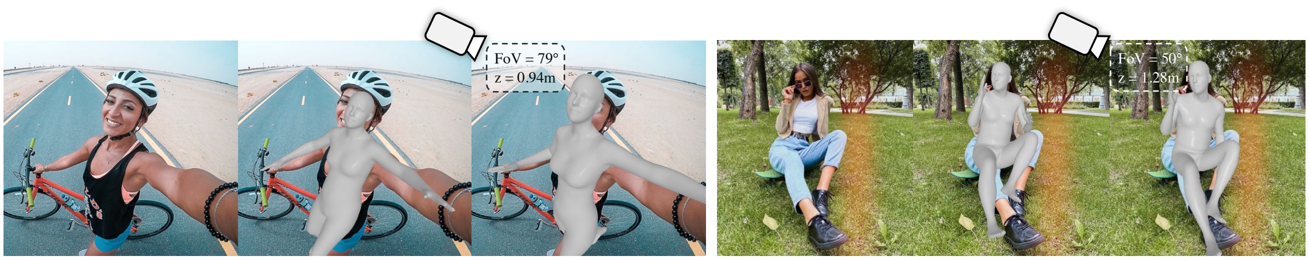
Zolly: Zoom Focal Length Correctly for Perspective-Distorted Human Mesh Reconstruction

Wenjia Wang[1,4] Yongtao Ge[2] Haiyi Mei[3] Zhongang Cai[3,4]
Qingping Sun[3] Yanjun Wang[3] Chunhua Shen[5] Lei Yang[3,4,†] Taku Komura[1]

[1] The University of Hong Kong [2] The University of Adelaide [3] SenseTime Research
[4] Shanghai AI Laboratory [5] Zhejiang University



(a) Input image (b) SPEC (c) Our results (d) Input image (e) SPEC (f) Our results

Figure 1: Close-up photography can make it difficult to discern 3D human pose in perspective-distorted images, while state-of-the-art methods often struggle with weak-perspective camera models or inaccurate focal length estimates. Our method overcomes these challenges, accurately recovering 3D human mesh from an approximate distance for fine-grained reconstruction. We could thus get focal length by our proposed approach.

Abstract

As it is hard to calibrate single-view RGB images in the wild, existing 3D human mesh reconstruction (3DHMR) methods either use a constant large focal length or estimate one based on the background environment context, which can not tackle the problem of the torso, limb, hand or face distortion caused by perspective camera projection when the camera is close to the human body. The naive focal length assumptions can harm this task with the incorrectly formulated projection matrices. To solve this, we propose Zolly, the first 3DHMR method focusing on perspective-distorted images. Our approach begins with analysing the reason for perspective distortion, which we find is mainly caused by the relative location of the human body to the camera center. We propose a new camera model and a novel 2D representation, termed distortion image, which describes the 2D dense distortion scale of the human body. We then estimate the distance from distortion scale features rather than environment context features. Afterwards, We integrate the distortion feature with image features to reconstruct the body mesh. To formulate the correct projection matrix and locate the human body position, we simultaneously use perspective and weak-perspective projection loss. Since existing datasets could not handle this task, we propose the first synthetic dataset PDHuman and extend two real-world datasets tailored for this task, all containing perspective-distorted human images. Extensive experiments show that Zolly outperforms existing state-of-the-art methods on both perspective-distorted datasets and the standard benchmark (3DPW). Code and dataset will be released at `https://wenjiawang0312.github.io/projects/zolly/`.

1. Introduction

Human pose and shape estimation from single-view RGB images is a long-standing research area of computer vision, as the reconstructed motion and mesh could empower various human-centered downstream applications like 3D animations, robotics, or AR/VR development. Previous works [19, 26, 23, 24, 41, 23, 27, 20, 55] formulate the problem under the assumption that the reconstructed people are far away from the camera, thus the torso and limb distortion caused by the perspective projection can be neglected.

*LY[†] is the corresponding author.

However, perspective distortion in close-up images is common in real-life scenarios, such as photographs of athletes/actors in sports events/films or selfies taken for social media. In such images, distortions are usually caused by aerial photography, overhead beat, or large depth variance among torsos and limbs, resulting in depth ambiguity in single-view RGB images, which makes it a big challenge to recover human pose and shape (See Fig. 1).

Previous methods typically assume a large fixed focal length [19, 26, 23, 24] or estimate a focal length [25] using pre-trained networks and calculate the translation from the estimated focal length. These settings are appropriate when people are far from the cameras, where the depth variance of the human body is negligible compared to the distance to the camera. However, these methods are inappropriate for handling scenarios in which human bodies are perspective distorted. Overestimating the focal length could lead to joint angle ambiguity or harm joint rotation learning. Several methodologies for pose estimation, as proposed by previous works [22, 29], assume a large field-of-view (FoV) angle. However, these methods may not show significant improvement when the focus is solely on non-distorted human images, as they often lack a conditioning for depth variance when the camera zooms in or out. Inaccuracies in estimating the depth variance with respect to translation can adversely impact re-projection loss, leading to erroneous results, as illustrated in Fig. 1. Actually, a correctly estimated distance and focal length also help with 2D alignment, which will be useful in downstream tasks.

To address the challenge of perspective distortion in close-up images, showing respect to Hitchcock's dolly zoom shot, we introduce Zolly (**Z**oom f**O**cal **L**ength correct**LY**) for perspective-distorted human mesh reconstruction. Our method utilizes 2D human distortion features to estimate the real-world distance to the camera center, enabling the reconstruction of the 3D human mesh in perspective-distorted images. The framework comprises of two parts: a translation estimation module for estimating the z-axis distance of the human body from the camera center, and a mesh reconstruction module for reconstructing 3D vertex coordinates in camera space. Additionally, we introduce a hybrid loss function that combines both perspective and weak perspective projection to boost performance.

Inspired by the iconic dolly-zoom shot [39] (also known as zolly shot), which creatively combines camera movement and zooming to create a distorted perspective and sense of unease, we propose a translation estimation module for the perspective-distorted 3DHMR task. This module highlights how the relative position of the human body to the camera affects the perspective distortion in images. Based on this insight, we introduce the distorted image as a new representation to capture the 2D shrinking or dilation scales of each pixel. Our translation network utilizes distortion and

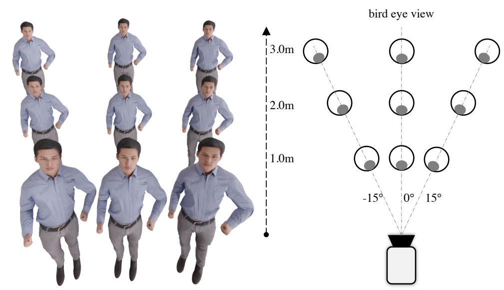


Figure 2: This figure showcases how the distortion scale of a person's limbs becomes more pronounced when closer to the camera center. Further, when two human bodies are at the same included angle with respect to the human-camera axis, they appear similar in facial direction. However, the horizontal translation may cause distinct distortion types on the left and right sides. We demonstrate that as the human body gets closer to the camera center, the distortion magnitude increases, leading to a more precise estimation of depth and rotation angles.

IUV images to accurately estimate z-axis translation, overcoming the limitations of traditional methods that rely on environmental information. IUV image could help eliminate the 2D shift and scale information in distorted images and represent 2D dense position information. For mesh reconstruction, we lift the 2D position feature to the 3D vertex position feature and sample the by-vertex distortion feature to regress 3D vertex coordinates. We use perspective projection to supervise correctly and weak-perspective projection to locate the 2D human body position in the image and help to calculate our focal length.

In summary, our contributions are as follows:

(1) We analyze the state-of-the-art (SOTA) 3DHMR methods and propose a novel approach tailored to the perspective-distorted 3DHMR task.

(2) We propose a novel learning-based method to tackle the perspective-distorted 3DHMR task without relying on extra camera information. The core of our method is a newly designed representation, termed distortion image, and a hybrid projection supervision that make use of both perspective and weak-perspective projection. (3) We build the first large-scale synthetic dataset PDHuman for the perspective-distorted 3DHMR task, with high-quality SMPL ground truth and camera parameters. To evaluate the performance on real images, we prepare two real-world benchmark datasets, SPEC-MTP [25] and HuMMan [6], which contain perspective-distorted images with well-fitted SMPL parameters and camera parameters.

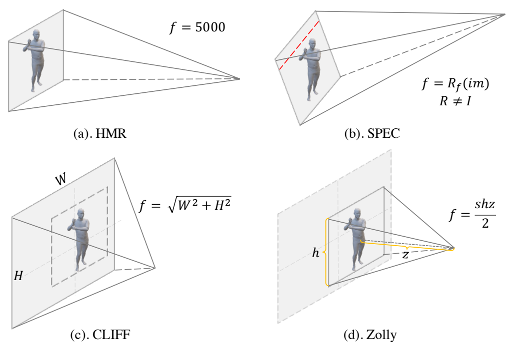


Figure 3: **Comparison of 4 different camera models in 3DHMR task**. For HMR [19], the focal length is fixed as 5000 pixels. Most methods follow this setting. For SPEC [25], the focal length is estimated by a network R_f pre-trained on other datasets. For CLIFF [29], during training, if without ground-truth focal length, will use the length of diagonal length. For Zolly, we use the estimated z-axis translation *z*, camera parameter *s*, and image height *h* to calculate focal length.

2. Related Work

Mainstream 3D human mesh reconstruction methods. 3D human pose estimation from a single RGB image is essentially an ill-posed problem. To obtain more realistic and manipulable human bodies, a parametric body model SMPL [33] was proposed, which uses 3D rotation representation to model human joint motions with defined LBS weights. To reconstruct human mesh from RGB images, there exist two mainstream pipelines: optimization-based methods and learning-based methods.

Optimization-based methods [5, 14] directly fit the body model parameters to 2D evidence via gradient back-propagation in an iterative manner. Learning-based approaches [19, 36, 52, 24] leverage a deep neural network to regress the human body model parameters or 3D coordinates of the human mesh, which can be further divided into model-based and model-free methods. Inspired by 3D mesh reconstruction tasks[10, 12, 54, 48, 47, 16], Model-based methods works [19, 25, 24, 28, 51, 50] utilize SMPL parameters to recover the human pose and shape. The milestone method HMR [19] takes it as a direct regression task.

Model-free works [27, 30, 8, 11] directly reconstructing 3D meshes from single view images.

Human mesh reconstruction with specific camera systems. In the previous trend led by HMR [19], the intrinsic camera model is formed as a weak-perspective camera, with a constant focal length of $5,000$ pixels. However, this assumption does not hold well when the person is close to the camera center, resulting in errors in the reconstructed 3D shape and pose. To address this issue, several recent works such as BeyondWeak [22], SPEC [25], and CLIFF [29] have proposed different camera system assumptions. SPEC predicts camera parameters (pitch, yaw, and FoV) from a single-view image, but its asymmetric Softargmax-$\mathcal{L}_2$ loss tends to overestimate focal length and translation, which is not suitable for distorted images. Moreover, SPEC regresses camera parameters through environmental information, which can sometimes be meaningless when the background lacks geometry information. CLIFF focuses on joint rotation variance caused by horizontal shift but has not conditioned the distance from the human body to the camera. CLIFF, following BeyondWeak[22], uses the diagonal length of the image as the focal length, which is not a close assumption for distorted problems since the focal length can be easily adjusted during image capture.

Compared to these methods, our framework estimates the z-axis translation from 2D human distortion features, and obtains a more accurate focal length from the estimated translation, leading to much better reconstruction accuracy on distorted images. See a comparison of the camera models in Fig. 3. In the Sup. Mat., we quantitatively demonstrate the bad re-projection influence caused by a wrongly formulated projection matrix.

3. Methodology

In this section, we first review the formulation of previous camera systems and then present our camera system customized for distorted images in Sec. 3.1. Sec. 3.2 presents our network architecture with two key components: (i) translation estimation module and (ii) mesh reconstruction module. Subsequently, we explain the proposed hybrid re-projection loss functions for distorted human mesh reconstruction in Sec. 3.3.

3.1. Preliminary

Camera system analysis. In weak-perspective projection, the inner depth variance is ignored in the human body, which means this projection model views the human body as a planar object without thickness. Thus the projection matrix should be as follows:

$$\begin{bmatrix} f & & \\ & f & \\ & & 1 \end{bmatrix} \begin{bmatrix} x+T_x \\ y+T_y \\ z+T_z \end{bmatrix} = \begin{bmatrix} f(x+T_x) \\ f(y+T_y) \\ T_z, \end{bmatrix}, z=0, \quad (1)$$

where f refers to the focal length in NDC (Normalized Device Coordinate) space, x, y, z refers to a vertex point on human body mesh and T_x, T_y, T_z refers to pelvis translation. The weak-perspective camera parameters (s, t_x, t_y), which represent 2D orthographic transformation, could be used to approximate the projection:

$$\begin{bmatrix} f(x+T_x)/T_z \\ f(y+T_y)/T_z \end{bmatrix} = \begin{bmatrix} s(x+t_x) \\ s(y+t_y) \end{bmatrix}. \quad (2)$$

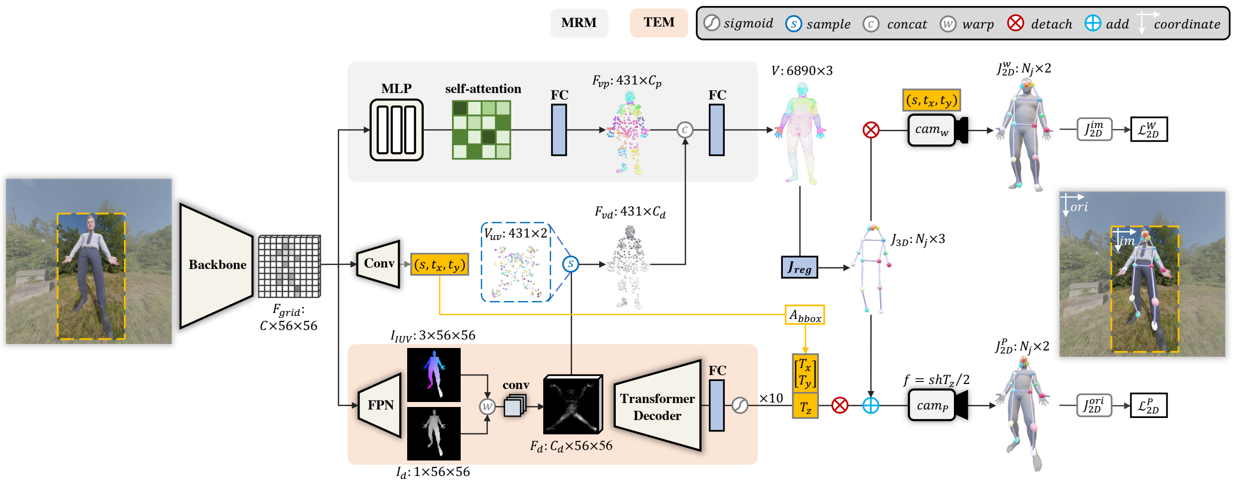


Figure 4: **Zolly pipeline overview**. The whole pipeline mainly consists of two modules and a hybrid re-projection supervision. MRM denotes the mesh reconstruction module. TEM indicates the translation estimation module. F_{grid} is the spatial feature from the backbone. F_{vp} and F_{vd} represents per-vertex position and distortion feature . (s, t_x, t_y) are the weak-perspective parameters. J_{2D}^{im} denotes 2D joints in the cropped image coordinate system. J_{2D}^{ori} denotes 2D joints in original image coordinate system before cropped. h denotes image height.

Finally, we can get:

$$s \times T_z = f, T_x = t_x, T_y = t_y. \tag{3}$$

However, the perspective projection actually is:

$$\begin{bmatrix} x_{2D} \\ y_{2D} \end{bmatrix} = \begin{bmatrix} f(x + T_x)/(z + T_z) \\ f(y + T_y)/(z + T_z) \end{bmatrix} = \begin{bmatrix} s(x + t_x) \\ s(y + t_y) \end{bmatrix}. \tag{4}$$

From Eq. (4), if z gets smaller, the projected x_{2D}, y_{2D} will be larger. This phenomenon causes the closer points on the close-up photographed image to dilate and the farther points to shrink. *Thus, the 3D translation results in pixel-level distortion on the limbs, torso, or faces in the 2D projected image.* Usually, when the human body is farther than 5 m, the distortion is subtle. Under these circumstances, a weak-perspective projection could be used.

Following the weak-perspective assumption, we take the $f = s \times T_z$ (f is in NDC space) as an approximation. T_z is the z-axis translation of the pelvis, which could be viewed as a mean translation of the whole human body. The difference compared to previous methods [19, 25, 29] is that we first estimate the body translation and then calculate the focal length. So we still need to estimate weak-perspective camera parameters (s, t_x, t_y) to compute the focal length and obtain the 2D location in the image. Following SPEC [25], we get the T_x, T_y in the full image from the t_x, t_y by affine transformation using the bounding box.

Distortion image. As described in Sec. 3.1, our approach projects the 3D translations of human body points into 2D images where the limbs dilate or shrink. We adopt the x-y plane of the pelvis as the reference plane representing a 'scale equals 1' plane. When the human is distant from the camera, it can be approximated as a zero-thickness plane, where all distortion scales are 1. And as shown in Eq. (4), distortion scales are inversely proportional to the z-distance from the camera when the body is closer. To quantify limb distortion caused by perspective projection, we introduce a distortion image I_d, where $I_d = T_z / I_{Depth}$, where I_{Depth} represents a depth image. The distorted image and its pixel value enable a visual and numerical representation of the limb dilation or shrinkage caused by the perspective camera. For instance, when the pelvis is fixed, a finger can appear twice as dilated when z of the finger reaches from $1\,\mathrm{m}$ to $0.5\,\mathrm{m}$. See Fig. 5 for the demonstration of distortion image.

3.2. Network Structure

Given a monocular image, Zolly applies an off-the-shelf Convolution Neural Network, *e.g.* [15, 40], as an image encoder; the output multi-level features can be used as an input for the translation module and the mesh reconstruction module that we describe next.

Translation module. As shown in Fig. 4, we estimate the distortion image I_d and IUV image I_{IUV} with an FPN [31] structure. There are two main advantages for this setting. Firstly, we can distillate the geometry information on the human body without background context. Another advantage is that this dense correspondence predicting task is easy to train. As noted in Sec. 3.1, the distortion type corresponds to one certain translation, and the distortion is determined when the image was captured, whether or not cropped or rotated afterwards. So we further warp the distorted image into the continuous UV space [52] to eliminate the 2D scale, shift, and rotation. We treat T_z as a learnable embedding. A 1×1 convolution is first applied to up-sample the channels of the warped distortion image. Then cross-attention [44] is performed between the warped distortion feature and the z-axis embedding, with a fully connected layer to output T_z. Note that we use sigmoid then $\times 10$ to restrict T_z to be between 0m and 10m. Following SPEC [25], we get T_x and T_y by applying affine transform on the estimated t_x, t_y with ground-truth bounding boxes (See Sup. Mat. for more details). The loss function of the translation module is formulated as follows:

$$\mathcal{L}_{Transl} = \lambda_{IUV}\mathcal{L}^2_{IUV} + \lambda_d\mathcal{L}^2_d + \lambda_z\mathcal{L}^1_z, \tag{5}$$

where $\mathcal{L}^2_{IUV}$ is the $\mathcal{L}2$ loss of the IUV image, $\mathcal{L}^2_d$ is the $\mathcal{L}2$ loss of the distortion image, and $\mathcal{L}^1_z$ is the $\mathcal{L}1$ loss of z-axis translation.

Mesh reconstruction module. Different from previous methods that use graph convolution [27] or transformers [8, 30] for building long-range dependence among different vertices, we adopt a light-weight MLP-Mixer [42] structure to model the attention among different vertices, followed by a fully connected layer that lifts per-vertex position features F_{vp} from the spatial feature F_{grid} which was used to predict I_{IUV}.

As illustrated in Fig. 4, since the distortion feature has already been warped into UV space, we could easily sample the per-vertex distortion feature F_{vd} from the warped distortion feature F_d by pre-defined Vertex UV coordinates V_{uv} [52]. We concatenate F_{vd} with F_{vp} and use fully connected layers to predict the coordinates of a coarse mesh of the body that is composed of 431 vertices. The coarse mesh is up-sampled using two fully connected layers, resulting in an intermediate mesh with $1,723$ vertices and a full mesh with $6,890$ vertices. 3D joint coordinates are obtained using a joint regression matrix provided by the SMPL [33] body model. The total loss for the mesh reconstruction module is:

$$\begin{aligned}\mathcal{L}_{Mesh} = \lambda_{J_{3D}}\mathcal{L}^1_{J_{3D}} + \lambda_{J^P_{2D}}\mathcal{L}^1_{J^P_{2D}} + \lambda_{J^W_{2D}}\mathcal{L}^1_{J^W_{2D}} \\ + \lambda_V(\mathcal{L}^1_{V''} + \mathcal{L}^1_{V'} + \mathcal{L}^1_V),\end{aligned} \tag{6}$$

where $\mathcal{L}^1_{J_{3D}}$ is $\mathcal{L}1$ loss of 3D joints, $\mathcal{L}^1_{V''}$, $\mathcal{L}^1_{V'}$ and $\mathcal{L}^1_V$ is $\mathcal{L}1$ loss of coarse, intermediate vertices, and full vertices respectively. $\mathcal{L}^1_{J^P_{2D}}$ and $\mathcal{L}^1_{J^W_{2D}}$ represents loss of perspective and weak-perspective re-projected 2D joints, and will be further illustrated in Sec. 3.3.

3.3. Hybrid Re-projection Supervision

Most existing methods [19, 29, 8] usually use a pre-defined focal length f. SPEC [25] train a CamCalib network to estimate the focal length. Then, z-axis translation T_z can be calculated by $T_z = 2f/hs$. On the contrary, as illustrated in Eq. (3), we aim to get the focal length f by directly predicting the orthographic scale s and z-axis translation T_z. Following HMR [19], we still use the weak-perspective projection besides perspective projection.

Weak-perspective re-projection. For weak-perspective projection loss, we follow HMR [19], use focal length f_W as $5,000$ pixels, and thus formulate the weak-perspective intrinsic matrix and translation separately as:

$$K_W = \begin{bmatrix} f_W & & h/2 \\ & f_W & h/2 \\ & & 1 \end{bmatrix}, T_W = \begin{bmatrix} t_x \\ t_y \\ 2f_W/sh \end{bmatrix}. \tag{7}$$

Then we project the 3D joints $\hat{J}_{3D}$ and measure the difference with 2D keypoints in image coordinates as:

$$\hat{J}^W_{2D} = K_W(\hat{J}^{\otimes}_{3D} + T_W), \tag{8}$$

$$\mathcal{L}^W_{2D} = \sum_{i=1}^{N_j} \frac{1}{d_{J[i]}} \|\hat{J}^W_{2D}[i] - J^{im}_{2D}[i]\|^1_F, \tag{9}$$

where $\hat{J}^{\otimes}_{3D}$ *means we detach the gradient from the body model joints in weak-perspective projection.* This means we only update the weak-perspective camera (s, t_x, t_y) and do not want this wrong projection to harm the body pose gradient flow. (s, t_x, t_y) are mainly used to locate the human body's position in image coordinates and compute the focal length f_P. For better position alignment, we divide a distortion weight $d_{J[i]}$, which is sampled from distortion image I_d by $J^{im}_{2D}[i]$ for every joint. This forces the dilated limbs to get a smaller weight while the shrunk limbs get a bigger weight.

Perspective re-projection. Perspective re-projection is mainly used to supervise pose or mesh reconstruction with the correct projection matrix. Firstly, we have 3D joints by $\hat{J}_{3D} = \mathcal{J}_{reg}V$. We use ground-truth focal length f_P to stabilize the training. For samples without ground-truth focal length, we will use a focal length of $1,000$ pixels for 224×224 images. This will make the translation range approximately from 5 to 10 meters. During inference, according to Eq. (4), we compute the focal length in screen space for perspective projection by $f_P = shT_z/2$ pixels, where h represents cropped image height, equals 224 pixels in our setting. Thus we can formulate the perspective intrinsic matrix K_P and projected 2D joints $\hat{J}^P_{2D}$ as:

$$K_P = \begin{bmatrix} f_P & & H/2 \\ & f_P & H/2 \\ & & 1 \end{bmatrix}, \hat{J}^P_{2D} = K_P(\hat{J}_{3D} + T^{\otimes}_P), \tag{10}$$

where $T^{\otimes}_P$ is the translation estimated by translation head in Sec. 3.2. We detach it as well to avoid the alignment conflicting of two re-projection. We project the 3D joints $\hat{J}_{3D}$

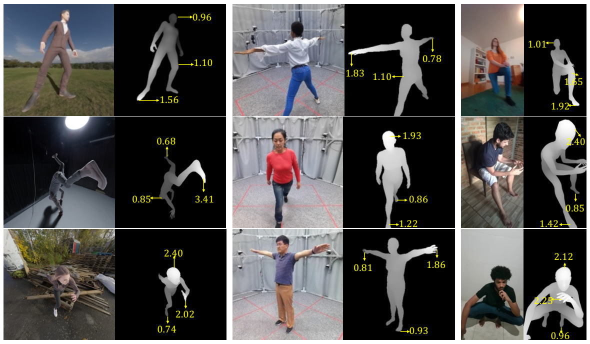


Figure 5: **Demonstration of distortion image.** The three columns from left to right are our PDHuman dataset, HuMMan dataset, and SPEC-MTP dataset, respectively. The value from the arrow (yellow) indicates the distortion scale of the pixel.

and measure the difference with the original 2D keypoints in the image coordinates before cropped.

4. Experiments

4.1. Datasets

PDHuman. Despite perspective distortion being a common problem, no existing public dataset is specifically designed for this task. Inspired by recent synthetic datasets [49, 7, 3, 35], we introduce a synthetic dataset named PDHuman. The dataset contains $126,198$ images in the training split and $27,448$ images in the testing split, with annotations including camera intrinsic matrix, 2D/3D keypoints, SMPL parameters ($\boldsymbol{\theta}$, $\boldsymbol{\beta}$), and translation for each image. The testing split is further divided into 5 protocols by the max distortion scale of each image sample. We define the max distortion scale for each sample as τ; this value will be used in splitting protocols.

We use 630 human models from RenderPeople [2] and $1,710$ body pose sequences from Mixamo [1], with 500 HDRi images with various lighting conditions as backgrounds. We use the dolly-zoom effect to generate random camera extrinsic and intrinsic matrices with random rotations, translations, and focal lengths. The distance from the human body to the cameras is set from 0.5m to 10m, so our dataset contains severely distorted, slightly distorted, and nearly non-distorted images. Then we use Blender [4] to render the RGB images. See Fig. 5 for brief demonstration. For detailed rendering procedures and more image demonstrations, please refer to Sup. Mat.

SPEC-MTP and HuMMan Datasets. SPEC-MTP dataset [25] is proposed to test human pose reconstruction in world coordinates. It includes many close-up shots, mostly taken from below or overhead views, leading to images with distorted human bodies. HuMMan dataset [6] is captured by multi-view RGBD cameras and has accurate ground truth because the SMPL parameters are fitted based on 3D keypoints and point clouds. HuMMan also contains images with distorted human bodies since the actors were close to the cameras, all less than 3 meters away. In our paper, we extend both datasets into real-world perspective-distorted datasets. SPEC-MTP is used only for testing. For HuMMan, we split it into training and testing parts. When testing, we divide these two datasets into three protocols based on their maximum distortion scale τ. See Fig. 5 for brief demonstration.

Non-distorted Datasets. For non-distorted datasets, we use Human3.6M [17], COCO [32], MPI-INF-3DHP [34] and LSPET [18] as our training data. Following [30, 8], we also report the results fine-tuned on 3DPW [45] training data.

4.2. Evaluation Metrics

To measure the accuracy of reconstructed human mesh, we follow the previous works [19, 25, 29] by adopting MPJPE (Mean Per Joint Position Error), PA-MPJPE (Procrustes Analysis Mean Per Joint Position Error) and PVE (Per Vertex Error) as our 3D evaluation metrics. They all measure the Euclidean distances of 3D points or vertices between the predictions and ground truth in millimeters (mm).

To measure the re-projection results in perspective distorted datasets such as PDHuman, SPEC-MTP and HuMMan, we leverage metrics widely used in segmentation tasks, MeanIoU [13] as our 2D metric. We both report foreground and background MeanIoU marked as mIoU and body part MeanIoU marked as P-mIoU. We use the 24-part vertex split provided by official SMPL [33] for body part segmentation. During the evaluation, for weak-perspective methods like HMR, we will render the predicted segmentation masks with a focal length of $5,000$ pixels. And we use the corresponding focal length on methods with specific camera models, such as SPEC [25], CLIFF [29], and proposed Zolly.

4.3. Implementation Details

Unless specified, we use ResNet-50 [15] and HRNet-w48 [40] backbones for model-free Zolly. We also design a model-based variant, Zolly$^{\mathcal{P}}$ ($\mathcal{P}$ stands for parametric), by changing the mesh reconstruction module to a model-based pose and shape estimation module. The details of Zolly$^{\mathcal{P}}$ can be found in the Sup. Mat. All backbones are initialized by COCO [32] key-point dataset pre-trained models. We use Adam [21] optimizer with a fixed learning rate of $2e^{-4}$. All experiments of Zolly are conducted on 8 A100 GPUs for around 160 epochs, 14∼18 hours. Our training pipeline was built based on MMHuman3D [9] code base. For samples with ground-truth focal length and translations, we render IUV and distortion images online during the training by PyTorch3D [38].

Methods	PDHuman ($\tau = 3.0$)					SPEC-MTP ($\tau = 1.8$)					HuMMan ($\tau = 1.8$)				
	PA-MPJPE↓	MPJPE↓	PVE↓	mIoU↑	P-mIoU↑	PA-MPJPE↓	MPJPE↓	PVE↓	mIoU↑	P-mIoU↑	PA-MPJPE↓	MPJPE↓	PVE↓	mIoU↑	P-mIoU↑
HMR (R50) [19]	62.5	91.5	106.7	48.9	21.7	73.9	121.4	145.6	48.8	16.0	30.2	43.6	52.6	65.1	39.5
HMR-f (R50) [19]	61.6	90.2	105.5	45.2	20.4	72.7	123.2	145.1	52.3	20.1	29.9	43.6	53.4	62.7	34.9
SPEC (R50) [25]	65.8	94.9	109.6	43.4	19.6	76.0	125.5	144.6	49.9	18.8	31.4	44.0	54.2	51.4	25.6
CLIFF (R50) [29]	66.2	99.2	115.2	51.4	24.8	74.3	115.0	132.4	53.6	23.7	28.6	42.4	50.2	68.8	44.7
PARE (H48) [24]	66.3	95.9	116.7	48.2	20.9	74.2	121.6	143.6	55.8	23.2	32.6	53.2	65.5	66.5	38.3
GraphCMR (R50)	62.0	85.8	98.4	47.9	21.5	76.1	121.4	141.6	53.5	22.0	29.5	40.6	48.4	61.6	37.5
FastMETRO (H48) [8]	58.6	83.6	95.4	50.1	22.5	75.0	123.1	137.0	53.5	20.5	26.3	38.8	45.5	68.3	45.2
Zolly$^{\mathcal{P}}$ (R50)	54.3	80.9	93.9	**54.5**	**27.4**	72.9	117.7	138.2	54.7	22.4	24.4	36.7	45.9	70.4	**45.4**
Zolly (R50)	54.3	76.4	87.6	51.4	24.0	74.0	122.1	135.5	58.9	24.9	25.5	36.7	43.4	67.0	38.4
Zolly (H48)	**49.9**	**70.7**	**82.0**	53.0	26.5	**67.4**	**114.6**	**126.7**	**62.3**	**30.4**	**22.3**	**32.6**	**40.0**	**71.2**	45.1

Table 1: Results of SOTA methods on PDHuman, SPEC-MTP [25] and HuMMan [6] datasets. Here we report the largest distortion protocol. R50 terms ResNet-50 [15], and H48 terms HRNet-w48 [40] here.

For comparison on PDHuman, SPEC-MTP, and HuMMan, we follow the official codes of HMR [19], SPEC [25], PARE [24], GraphCMR[27], FastMETRO[8]. We re-implemented CLIFF [29] since the authors have not released the training codes. All SOTA methods are trained on 8 A100 GPUs until convergence, following the officially released hyper-parameters. All methods are trained on the same datasets with the same proportion, *e.g.*, Human3.6M [17] (40%), PDHuman (20%), HuMMan [6] (10%), MPI-INF-3DHP [34] (10%), COCO [32] (10%), LSPET [18] (5%).

For distorted datasets, we only report the results on the protocols with the largest distortion scales. The full results of all protocols are shown in Sup. Mat.

4.4. Main Results

Results on PDHuman, SPEC-MTP, and HuMMan. We report PA-MPJPE, MPJPE, PVE, mIoU, and P-mIoU on these three datasets. For model-based methods, we compare with HMR [19], SPEC [25], CLIFF [29], PARE [24]. We compare model-free methods with GraphCMR [27] and FastMETRO [8]. From Tab. 1, we can see that SPEC [25] performs poorly on these distorted datasets. This is mainly due to their wrong focal length assumption, which has negative rather than positive effects on their supervision. (Note that our re-implemented SPEC has higher performance than the official code, see in Sup. Mat.) CLIFF performs well on SPEC-MTP, while badly on PDHuman. Because their focal length assumption is about 53° for 16:9 images, close to SPEC-MTP images. Although HMR-f is trained with the same focal length as Zolly, it improves little compared to HMR since they have not encoded the distortion or distance feature into their network. Zolly-H48 outperforms SOTA methods on most metrics, especially the 3D ones. Some 2D re-projection metrics, *e.g.* mIoU and P-mIoU, of Zolly-H48 are lower than Zolly$^{\mathcal{P}}$-R50 version. We conjecture that model-based methods have better reconstructed shapes. Please refer to Fig. 6 for qualitative results. More qualitative results and failure cases can be found in Sup. Mat.

Methods	Backbone	w. 3DPW	Metrics		
			PA-MPJPE↓	MPJPE↓	PVE↓
HybrIK[28]	ResNet-34	×	48.8	80.0	94.5
HybrIK[28]	ResNet-34	✓	45.0	74.1	86.5
GraphCMR [27]	ResNet-50	×	70.2	-	-
HMR [19]	ResNet-50	×	72.6	116.5	-
SPIN [26]	ResNet-50	×	59.2	96.9	116.4
PyMAF [53]	ResNet-50	×	58.9	92.8	110.1
SPEC [25]	ResNet-50	✓	52.7	96.4	-
PARE [24]	ResNet-50	×	52.3	82.9	99.7
FastMETRO [8]	ResNet-50	✓	48.3	77.9	90.6
CLIFF [29]	ResNet-50	✓	45.7	72.0	85.3
PARE [24]	HRNet-w32	✓	46.5	74.5	88.6
CLIFF [29]	HRNet-w48	✓	43.0	69.0	81.2
Graphormer [30]	HRNet-w64	✓	45.6	74.7	87.7
FastMETRO [8]	HRNet-w64	✓	44.6	73.5	84.1
Zolly$^{\mathcal{P}}$	ResNet-50	×	48.9	80.0	92.3
Zolly	ResNet-50	×	49.2	79.6	92.7
Zolly	ResNet-50	✓	44.1	72.5	84.3
Zolly	HRNet-w48	×	47.9	76.2	89.8
Zolly	HRNet-w48	✓	**39.8**	**65.0**	**76.3**

Table 2: Results of SOTA methods on 3DPW. Zolly$^{\mathcal{P}}$ terms our parametric-based variant. 'w/o PD' means trained without the proposed distorted dataset. 'w/ f' means trained with ground-truth focal length if provided.

Results on 3DPW. This study compares our proposed method, Zolly, with SOTA methods [28, 27, 19, 26, 53, 25, 24, 8, 29, 30], including both model-based and model-free approaches. As shown in Tab. 2, Zolly-R50 achieves comparable results to the SOTA method FastMETRO-R50 even without being fine-tuned on the 3DPW training set. Moreover, after fine-tuning, Zolly with both backbone structures show a significant improvement in performance. Furthermore, when using HRNet-w48, our approach outperforms all SOTA methods in all three metrics, surpassing model-based SOTA method CLIFF [29] and model-free SOTA method FastMETRO [8]. This superiority can be attributed to two main factors: on one hand, in training, we use ground-truth focal length and translation from 3DPW raw data to supervise the rendering of IUV images and distortion images; on the other hand, given that 96% of the 3DPW images were captured within a distance of 1.2m to 10m, with more than half of them captured within 4m, there exist

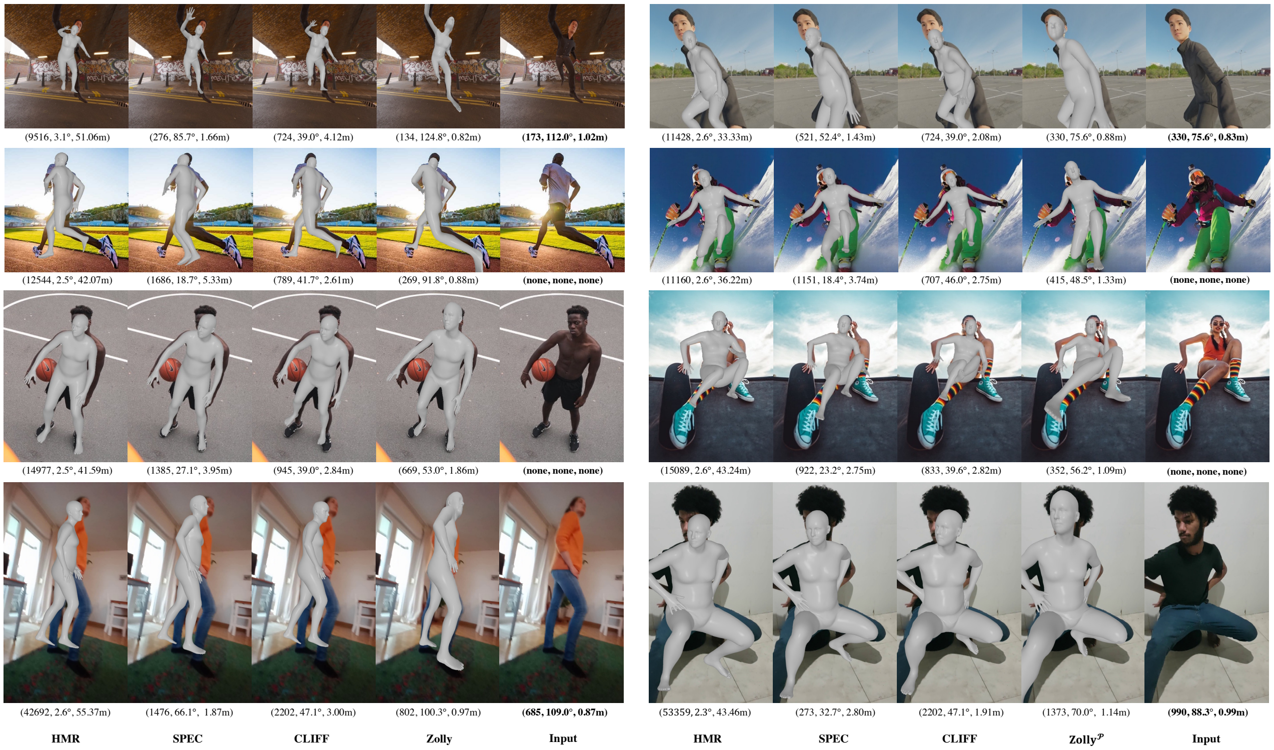


Figure 6: **Qualitative results of SOTA methods.** Besides Zolly, we visualize the results of three methods with specific camera models: HMR [19], SPEC [25], CLIFF [29]. Zolly$^{\mathcal{P}}$ terms our model-based variance. We show results come from different data sources. Row 1: PDHuman test. Row 2, 3: web images. Row 4: SPEC-MTP. The number under each image represents predicted/ground-truth f, FoV angle, and T_z. The ground-truth f and T_z for SPEC-MTP are pseudo labels. The focal lengths here are all transformed to pixels in full image.

many perspective-distorted images. We provide a detailed analysis of the results for samples captured from different distances in the 3DPW dataset in the Sup. Mat.

Results on Human3.6M: During training, we get the ground-truth focal length and translation from Human3.6M [17] training set for our supervision. When evaluating Human3.6M, we follow HybrIK [28] by using SMPL joints as the ground truth for evaluation. As shown in Tab. 3, our method performs well on Human3.6M through it is not a perspective-distorted dataset. Zolly-H48 achieves the best result on the PA-MPJPE metric and achieves comparable results on the MPJPE metric. CLIFF achieves the best results on MPJPE while they also need ground-truth bounding boxes during testing.

Methods	Backbone	Metrics	
		PA-MPJPE↓	MPJPE↓
HybrIK[28]	ResNet-34	34.5	54.4
HMR [19]	ResNet-50	56.8	88.0
GraphCMR [27]	ResNet-50	50.1	-
SPIN [26]	ResNet-50	41.1	62.5
PyMAF [53]	ResNet-50	40.5	57.7
FastMETRO [8]	ResNet-50	37.3	53.9
CLIFF [29]	ResNet-50	35.1	50.5
CLIFF [29]	HRNet-w48	32.7	**47.1**
Graphormer [30]	HRNet-w64	34.5	51.2
FastMETRO [8]	HRNet-w64	33.7	52.2
Zolly$^{\mathcal{P}}$	ResNet-50	34.7	54.0
Zolly	ResNet-50	34.2	52.7
Zolly	HRNet-w48	**32.3**	49.4

Table 3: Results of SOTA methods on Human3.6M [17].

4.5. Ablation study

Ablation on training settings on the standard benchmark 3DPW. In Table 4, we present the results of our ablation study on the standard 3DPW benchmark [45], where we investigate the impact of different training settings on the performance of our method. By controlling two different variables, we show that introducing perspective-distorted datasets and fine-tuning with ground-truth focal length both lead to a slight improvement in performance. Notably, our method Zolly-H48 still outperforms the current state-of-the-art methods even without using perspective-distorted data or ground-truth focal length.

This study evaluates the effectiveness of the distortion feature and the hybrid re-projection loss function. The evaluation is conducted on the PDHuman ($\tau = 3.0$), as this ex-

Figure 7: Qualitative results for 3DPW. Zolly achieves good alignment with the characters in the original image, but other SOTA methods have difficulty aligning images that suffer from distortion caused by overhead shots, which causes upper body dilation and lower body shrinkage. The number under each image represents predicted/ground-truth f, FoV angle, and T_z.

w/ PD	w/ 3DPW	w/ gt f	Metrics		
			PA-MPJPE	MPJPE	PVE
×	×	-	48.3	78.0	92.0
×	✓	×	41.3	67.4	78.9
×	✓	✓	40.9	67.2	78.4
✓	×	-	47.9	76.2	89.8
✓	✓	×	40.9	66.4	78.3
✓	✓	✓	**39.8**	**65.0**	**76.3**

Table 4: Ablation study of Zolly-H48 of different training settings on 3DPW dataset. w/ PD means whether trained on perspective-distorted datasets (PDHuman, HuMMan). w/ 3DPW means whether fine-tuned on 3DPW [45] dataset. w/ gt f means using ground-truth focal length when fine-tuned on 3DPW.

hibits the highest degree of distortion. More experimental results are provided in the Sup. Mat.

Effect of distortion feature. In Tab. 5, w/o $w(I_d)$ terms without warp distortion image into UV space, and w/o $c(F_d)$ terms without concatenating distortion feature to per-vertex feature. We can see that, while the mIoU and P-mIoU change a little, the 3D metrics increase significantly with the correct distortion feature. This study validates our intuition that distortion information helps the network predict more accurate vertex coordinates.

Effect of the hybrid re-projection loss function. We experimented with different re-projection loss configurations and found that relying solely on weak-perspective loss significantly decreases 2D alignment. Incorporating perspective loss improved 3D metrics slightly but increased the 2D segmentation error significantly. Moreover, using per-joint distortion weight to supervise the weak-perspective camera improved the alignment of the human mesh and resulted in more accurate 3D supervision without increasing the 2D segmentation error.

We conclude that utilizing dense distortion features and an accurate camera model greatly improves the performance of our proposed method.

Architecture	Loss	Metrics				
		PA-MPJPE	MPJPE	PVE	mIoU	P-mIoU
Zolly w/o $w(I_d), c(F_d)$	$\Sigma d_J L_W^J + L_P$	60.2	86.8	99.0	52.0	24.9
Zolly w/o $c(F_d)$	$\Sigma d_J L_W^J + L_P$	57.0	83.3	95.0	51.2	23.6
Zolly	L_W	56.4	80.0	92.2	47.3	21.2
Zolly	$L_W + L_P$	56.1	79.1	91.0	52.5	25.5
Zolly	$\Sigma d_J L_W^J + L_P$	54.3	76.4	87.6	51.4	24.0

Table 5: Ablation study of Zolly-H48 structure on PDHuman ($\tau = 3.0$). L_W indicates weak-perspective re-projection loss. L_P indicates perspective re-projection loss. $\sum d_J L_W^J$ terms dividing the per-joint distortion weight in our weak-perspective loss.

5. Conclusion

We present Zolly, the first 3DHMR method that focuses on human reconstruction from perspective distorted images. Our proposed camera model and focal length solution accurately reconstruct the human body, especially in close-view photographs. We introduce a new dataset, PDHuman, and extend two datasets containing perspective-distorted images. Our results show significant value for human mesh reconstruction in perspective-distorted images and can empower many downstream tasks, such as monocular clothed human reconstruction [43, 37, 46] and human motion reconstruction in live shows, vlogs, and selfie videos. Further improvements and broader applications could be explored in the future.

Societal Impacts. Misuse by gaming or animation companies for motion capture can lead to copyright infringement and discourage original content. Fair use advocacy and negotiation with creators can promote sustainable creativity.

Acknowledgement Taku Komura and Wenjia Wang are partly supported by Technology Commission (Ref:ITS/319/21FP) and Research Grant Council (Ref: 17210222), Hong Kong.

References

[1] *Mixamo*, 2022. https://www.mixamo.com. 6

[2] *Renderpeople*, 2022. https://renderpeople.com. 6

[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. 6

[4] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam, 2022. 6

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Eur. Conf. Comput. Vis.*, pages 561–578. Springer, 2016. 3

[6] 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 *Eur. Conf. Comput. Vis.*, pages 557–577. Springer, 2022. 2, 6, 7

[7] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3d human recovery. *arXiv preprint arXiv:2110.07588*, 2021. 6

[8] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *Eur. Conf. Comput. Vis.*, pages 342–359. Springer, 2022. 3, 5, 6, 7, 8

[9] MMHuman3D Contributors. Openmmlab 3d human parametric model toolbox and benchmark. https://github.com/open-mmlab/mmhuman3d, 2021. 6

[10] Zhiyang Dou, Cheng Lin, Rui Xu, Lei Yang, Shiqing Xin, Taku Komura, and Wenping Wang. Coverage axis: Inner point selection for 3d shape skeletonization. In *Computer Graphics Forum*, volume 41, pages 419–432. Wiley Online Library, 2022. 3

[11] Zhiyang Dou, Qingxuan Wu, Cheng Lin, Zeyu Cao, Qiangqiang Wu, Weilin Wan, Taku Komura, and Wenping Wang. Tore: Token reduction for efficient human mesh recovery with transformer. *arXiv preprint arXiv:2211.10705*, 2022. 3

[12] Zhiyang Dou, Shiqing Xin, Rui Xu, Jian Xu, Yuanfeng Zhou, Shuangmin Chen, Wenping Wang, Xiuyang Zhao, and Changhe Tu. Top-down shape abstraction based on greedy pole selection. *IEEE Transactions on Visualization and Computer Graphics*, 27(10):3982–3993, 2020. 3

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 6

[14] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3d human pose by watching humans in the mirror. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 4, 6, 7

[16] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022. 3

[17] 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 Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2013. 6, 7, 8

[18] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1465–1472. IEEE, 2011. 6, 7

[19] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7122–7131, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[20] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1715–1725, 2022. 1

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv: Comp. Res. Repository*, 2014. 6

[22] Imry Kissos, Lior Fritz, Matan Goldman, Omer Meir, Eduard Oks, and Mark Kliger. Beyond weak perspective for monocular 3d human pose estimation. In *Eur. Conf. Comput. Vis.*, pages 541–554. Springer, 2020. 2, 3

[23] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5253–5263, 2020. 1, 2

[24] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Int. Conf. Comput. Vis.*, pages 11127–11137, 2021. 1, 2, 3, 7

[25] 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 *Int. Conf. Comput. Vis.*, pages 11035–11045, 2021. 2, 3, 4, 5, 6, 7, 8

[26] 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 *Int. Conf. Comput. Vis.*, pages 2252–2261, 2019. 1, 2, 7, 8

[27] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4501–4510, 2019. 1, 3, 5, 7, 8

[28] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3383–3393, 2021. 3, 7, 8

[29] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in

full frames into human pose and shape estimation. *arXiv: Comp. Res. Repository*, 2022. 2, 3, 4, 5, 6, 7, 8

[30] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Int. Conf. Comput. Vis.*, pages 12939–12948, 2021. 3, 5, 6, 7, 8

[31] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2117–2125, 2017. 4

[32] 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 *Eur. Conf. Comput. Vis.*, 2014. 6, 7

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):1–16, 2015. 3, 5, 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 *Int. Conf. 3D. Vis.*, 2017. 6, 7

[35] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2021. 6

[36] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 459–468, 2018. 3

[37] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9054–9063, 2021. 9

[38] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv: Comp. Res. Repository*, 2020. 6

[39] Steven Spielberg. Jaws, June 1975. 2

[40] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5693–5703, 2019. 4, 6, 7

[41] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Int. Conf. Comput. Vis.*, pages 11179–11188, 2021. 1

[42] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Adv. Neural Inform. Process. Syst.*, 34:24261–24272, 2021. 5

[43] Denis Tome, Thiemo Alldieck, Patrick Peluse, Gerard Pons-Moll, Lourdes Agapito, Hernan Badino, and Fernando De la Torre. Selfpose: 3d egocentric pose estimation from a headset mounted camera. *arXiv: Comp. Res. Repository*, 2020. 9

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 30, 2017. 5

[45] 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 *Eur. Conf. Comput. Vis.*, 2018. 6, 8, 9

[46] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13296–13306, June 2022. 9

[47] Rui Xu, Zhiyang Dou, Ningna Wang, Shiqing Xin, Shuangmin Chen, Mingyan Jiang, Xiaohu Guo, Wenping Wang, and Changhe Tu. Globally consistent normal orientation for point clouds by regularizing the winding-number field. *arXiv preprint arXiv:2304.11605*, 2023. 3

[48] Rui Xu, Zixiong Wang, Zhiyang Dou, Chen Zong, Shiqing Xin, Mingyan Jiang, Tao Ju, and Changhe Tu. Rfeps: Reconstructing feature-line equipped polygonal surface. *ACM Transactions on Graphics (TOG)*, 41(6):1–15, 2022. 3

[49] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, et al. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. *arXiv preprint arXiv:2303.17368*, 2023. 6

[50] Ailing Zeng, Xuan Ju, Lei Yang, Ruiyuan Gao, Xizhou Zhu, Bo Dai, and Qiang Xu. Deciwatch: A simple baseline for $10\times$ efficient 2d and 3d pose estimation. In *European Conference on Computer Vision*, pages 607–624. Springer, 2022. 3

[51] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision*, pages 625–642. Springer, 2022. 3

[52] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3d human mesh regression with dense correspondence. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7054–7063, 2020. 3, 5

[53] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Int. Conf. Comput. Vis.*, 2021. 7, 8

[54] Nan Zhang, Li Liu, Zhiyang Dou, Xiyue Liu, Xueze Yang, Doudou Miao, Yong Guo, Silan Gu, Yuguo Li, Hua Qian, et al. Close contact behaviors of university and school students in 10 indoor environments. *Journal of Hazardous Materials*, 458:132069, 2023. 3

[55] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7376–7385, 2020. 1