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

Shengze Wang[1*] Jiefeng Li[2] Tianye Li[2] Ye Yuan[2] Henry Fuchs[1]
Koki Nagano[2†] Shalini De Mello[2†] Michael Stengel[2†]
[1] UNC Chapel Hill [2] NVIDIA † Equal contribution
https://research.nvidia.com/labs/amri/projects/blade/

Figure 1. Our method enables accurate human mesh and camera parameter estimation for single-view in-the-wild images including close-ups with high levels of perspective distortion (pelvis depth T_z shown in meters).

Abstract

Single-image human mesh recovery is a challenging task due to the ill-posed nature of simultaneous body shape, pose, and camera estimation. Existing estimators work well on images taken from afar, but they break down as the person moves close to the camera. Moreover, current methods fail to achieve both accurate 3D pose and 2D alignment at the same time. Error is mainly introduced by inaccurate perspective projection heuristically derived from orthographic parameters. To resolve this long-standing challenge, we present our method BLADE which accurately recovers perspective parameters from a single image without heuristic assumptions. We start from the inverse relationship between perspective distortion and the person's Z-translation T_z, and we show that T_z can be reliably estimated from the image. We then discuss the important role of T_z for accurate human mesh recovery estimated from close-range images. Finally, we show that, once T_z and the 3D human mesh are estimated, one can accurately recover the focal length and full 3D translation. Extensive experiments on standard benchmarks and real-world close-range images show that our method accurately recovers projection parameters from a single image, and consequently attains state-of-the-art accuracy on both 3D pose estimation and 2D alignment for a wide range of images.

1. Introduction

Recent advances in 3D human mesh recovery (HMR) have started to democratize motion capture for media production, allowed computers to understand human gestures for human-computer interaction and enabled new applications in healthcare, fitness, and virtual try-on for E-commerce. Despite the many successes, current methods struggle in scenarios such as video conferencing and large-scale pose estimation on diverse images captured in the wild (Fig. 1).

Single-image human mesh recovery is challenging due to the under-constrained nature of estimating many parameters from a single view. Scale ambiguity and the unknown shape of the person contribute to the existence of potentially an infinite number of valid yet incorrect solutions [8]. Furthermore, intrinsic and extrinsic camera parameters are unknown for in-the-wild images and need to be estimated in addition to human shape and pose. It is thus exceptionally difficult to jointly estimate all of these variables at once.

Therefore, most existing methods reduce the number of unknowns by assuming near-orthographic projection, where the person is assumed to be far away and focal length is heuristically determined or calculated [10, 14–16, 19, 20, 31]. This leads to an unsatisfactory result, especially for close-ups that show a person with strong perspective distortion (Fig. 1). Recent work SPEC [15] targets this problem by directly estimating the camera focal length from images. ZOLLY [31] estimates both the depth of the person and a 2D affine transformation for an orthographic camera, which are then heuristically converted to a focal length and 3D translation with perspective projection. Both methods rely on inaccurate assumptions and fail to accurately recover the perspective parameters.

To simultaneously solve these manyfold challenges, we propose a new method for Body mesh Learning through Accurate Depth Estimation from a single image (BLADE).

* Shengze Wang was an intern at NVIDIA during the project.

Our key observation is that, mathematically, perspective distortion is driven by the distance between camera and person, but not affected by focal length (Fig. 3). The idea is that the Z-translation T_z of the person can be disentangled from other variables and be reliably estimated from the input image (Sec. 3.2). Once T_z is estimated, other variables become easier to solve. Motivated by this intuition as well as the success of recent one-shot metrical depth estimators [3, 27, 33], we train a T_z estimator to predict the depth of the person's pelvis with respect to the camera. We notice that that human pose estimators predict 3D human mesh from images that are affected by perspective distortion and that perspective distortion is determined by T_z. Therefore, we condition our pose estimator on T_z in order to improve accuracy of estimated human mesh. Lastly, the focal length and remaining translation parameters T_x and T_y can be obtained with knowledge of T_z and the 3D human mesh shape. Existing labeled datasets for HMR lack close-range images with strong perspective distortion. To augment them, we also contribute a new large-scale synthetic dataset with 2 million images tailored to this task. It helps our model learn accurate Z-translation of the human body and 3D pose across a wide range of depths.

On several benchmark datasets captured at diverse ranges, we outperform all existing SOTA methods at estimating subject depth, focal parameters, 3D pose, and 2D alignment. Our work contributes a new angle on accurate single-image 3D human pose estimation. It fully departs from the orthographic camera model and recover a fully perspective projection model without heuristics (Fig. 2), achieving high accuracy on 3D pose and 2D alignment on diverse depth ranges, including close-range images (Fig. 1 and 7).

In summary, we contribute:

1. A method for HMR that directly estimates perspective projection parameters given a single image without relying on heuristics. Our method achieves SOTA results on diverse depth ranges, including close-range images.
2. We identify that close-range pose estimation is heavily affected by Z-translation T_z, and we propose to condition the pose estimation on the estimated T_z to improve the accuracy of mesh recovery.
3. We correct the misconception that focal length affects image distortion, and we show the benefit of estimating focal length and XY-translation independently from T_z and mesh shape and pose.
4. We contribute a new large-scale synthetic dataset with a wide T_z variety.

2. Related Work

Human mesh recovery (HMR) from images and video is a long-standing problem and has received broad attention in research. Tian et al. [30] provides a comprehensive review

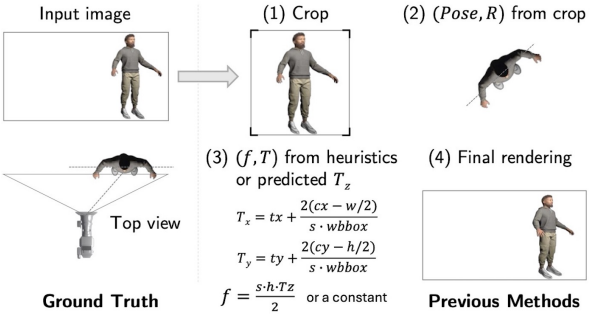

Figure 2. **Pose error introduced by camera heuristics.** (1,2) Previous methods estimate the pose of the person from image crops, leading to pose inaccuracy compared to the ground truth (left). (3) Focal length and 3D translation (f, T) are heuristically converted from a 2D affine transformation (s, t_x, t_y), which is only suitable from afar but not for close-range images. (4) Due to the incorrect pose and perspective parameters, the final estimation is inaccurate.

of the SOTA in HMR from monocular images. Additional surveys include recovery from multi-view images, videos, and body-worn sensors [7, 23, 38, 39]. In the following, we focus on methods for single-view single-person 3D HMR. This is an important distinction as we target general pose labeling of in-the-wild and internet-scale image datasets for which usually no data beside the images is available. To obtain realistic and manipulable human bodies, the parametric body model SMPL [24] and its successor SMPL-X [28] have been proposed. These models use linear blend skinning for the person's shape along with 3D joint positions and rotations for the pose.

Various methods estimate the body mesh directly using different neural network architectures such as a graph neural network [17], transformer [6], and a hybrid of the two [22]. Other methods regress on the SMPL(-X) body model parameters [8, 10, 14, 16, 20, 32, 35] using a multi-stage process that includes cropping of the body parts using detected bounding boxes followed by utilizing distinct models for individual reconstruction of those parts. In contrast, SMPLer-X [5], OSX [21], and AiOS [29] regress the body model as a whole, which reduces artifacts stemming from individual part reconstruction. Additionally, AiOS [29] utilizes a one-stage framework that directly recovers the human mesh from the entire image, omitting body cropping.

Due to the lack of camera information for in-the-wild images, all mentioned methods use orthographic camera models assuming that the person is sufficiently far from the camera. This is not always true in practice. As shown in Fig. 2, the weak-perspective assumption often involves estimating a 2D affine transform and heuristically converting the 2D scale and image space translations to focal length and 3D translations.

Different from these, few prior works do consider perspective distortion [13, 15, 20, 31]. Nagano *et al*. eval-

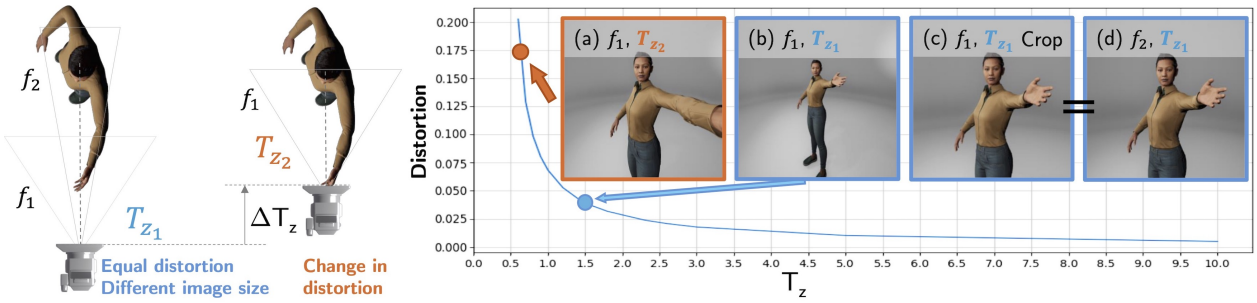

Figure 3. **Influence of T_z on perspective distortion.** A person is captured with different focal length and Z-translation T_z from the camera. (b&d) Changing the focal length from a short lens f_1 to a long lens f_2 changes the zoom factor but does not change the perspective distortion, as shown by the equivalence between (c) and (d). (a) Changing the Z-translation by a ΔT_z changes the level of perspective distortion in the image. This effect is particularly pronounced for close-range imagery (blue curve). See Sec. 3.1 for detailed discussion.

uate the distortion of faces for perspective projection and propose a generative adversarial network to normalize face images with distortion into near-orthographic ones [26]. Zhao *et al*. propose an approach to learning perspective undistortion for face portraits [37]. BeyondWeak [13] and CLIFF [20] show for HMR that a correction of camera translation from the box crop around the person to the full image improves performance. BeyondWeak [13] and W-HMR [34] propose to derive the focal length from image resolution as an approximation. SPEC [15] predicts camera parameters by learning field of view, camera pitch, and roll. However, the mentioned methods tend to overestimate focal length and translation and are therefore not reliable for close-up images.

TokenHMR specifically studies the influence of near-orthographic assumptions on the HMR quality [8]. The method reveals that current focal length estimators are inaccurate and unreliable and as a result, improving alignment to the 2D image deteriorates the accuracy of the 3D pose. TokenHMR proposes a Threshold-Adaptive Loss Scaling function to achieve both high 2D and 3D accuracy but only for a distant camera. Our approach is different from TokenHMR as we do not generate perspective projection parameters from an orthographic camera model. Instead, we directly solve for camera intrinsic and extrinsic parameters.

ZOLLY [31] is a perspective-aware SOTA method which allows HMR from close-range images. The method predicts SMPL body parameters inside a bounding box containing the person and estimates the orthographic projection, which is an affine transformation containing a scaling factor s. ZOLLY follows existing heuristics to estimate the focal length as $f = s \cdot h \cdot T_z/2$ and 3D translation as a function of 2D translation and bounding box properties (Fig. 2). Here, h is the image height, and T_z is the estimated depth of the SMPL pelvis. However, these heuristics are inaccurate approximations that lead to incorrect projections. In this work, we also estimate T_z as part of our method, but we avoid relying on heuristics for estimation. Instead, we disentangle the parameters to achieve better HMR performance and a more accurate recovery of camera parameters. There exists no method that can estimate the accurate 3D translation $[T_x, T_y, T_z]$ or correct focal length from a single image. The problem is inherently ill-posed because there are not enough constraints from a single image to solve for all variables. On the other hand, significant advancement has been made in solving two major sub-problems, *i.e*. depth estimation [1, 3, 11, 27, 33] and 3D pose estimation [5, 8, 10, 20, 21, 31]. Therefore, we leverage these efforts to solve for the remaining variables, namely $[f, T_x, T_y]$.

3. Method

Given a single image, our goal is to estimate an accurate 3D mesh of the person as SMPL-X parameters [28] while simultaneously achieving good 2D alignment. Although it is unreliable to directly estimate camera focal length and extrinsics from a single image, we show that they are essentially scaling and alignment parameters, which can be determined once the person's Z-translation T_z is estimated. Building upon this insight, we introduce a 3-step HMR pipeline (Fig. 4) that solves for all essential parameters in perspective projection: (1) Z-translation T_z of the person with respect to the camera (Sec. 3.2), (2) the 3D human pose and shape (β, θ) (Sec. 3.3), and finally (3) the person's XY-translations (T_x, T_y) and focal length f (Sec. 3.4).

3.1. Perspective Projection and its Implication

SMPL-X provides a differentiable function $M(\beta, \theta)$ that takes the pose parameters θ and the shape parameters β and outputs a body mesh $M \in \mathbb{R}^{N\times 3}$ with $N = 10475$ vertices and joint location $J \in \mathbb{R}^{K\times 3}$ with $K = 54$ joints. [1] The shape parameters $\beta \in \mathbb{R}^{10}$ are the first 10 PCA coefficients to model body shape variations. The pose parameters $\theta \in \mathbb{R}^{3K}$ model the joint rotation including the body orientation. One can obtain camera space coordinates of

[1] We omit facial expressions and hand gestures due to the lack of such labels in the existing close-range datasets.

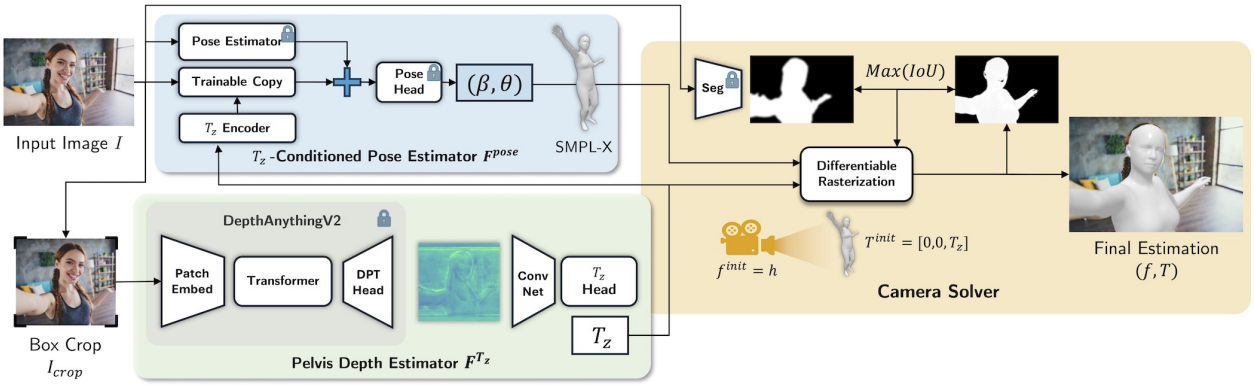

Figure 4. **Overview.** Starting with a bounding box image crop I_{crop} of the person, the *Pelvis Depth Estimator* F^{T_z} (green box) estimates the Z-translation of the person's pelvis, T_z. Then, the *Pose Estimator* F^{pose} (blue box) estimates SMPL-X shape and pose (β, θ) from the full input image while considering the image distortion induced by T_z. Finally, through differentiable rasterization, the *Camera Solver* (brown box) recovers the optimal focal length and 3D translations that best aligns the rasterized SMPL-X mesh with the segmented mask of the person. We are thus able to solve for the full perspective projection model without heuristic assumptions.

SMPL-X vertices $[x_m, y_m, z_m]$ as:

$$[x, y, z] = [x_m, y_m, z_m] + [T_x, T_y, T_z], \quad (1)$$

where $T = [T_x, T_y, T_z]$ is the position of the person's pelvis in the camera coordinate. With perspective projection, the projected coordinate is:

$$\begin{bmatrix} u \\ v \end{bmatrix} = f \cdot \begin{bmatrix} x/z \\ y/z \end{bmatrix} = f \cdot \begin{bmatrix} (x_m + T_x)/(z_m + T_z) \\ (y_m + T_y)/(z_m + T_z) \end{bmatrix}. \quad (2)$$

According to Eq. 2, the projected image coordinate is globally linear with respect to the focal length f, indicating that *focal length only acts as a uniform scaling and does not affect perspective distortion.* In contrast, the distance T_z and 3D geometry, which influence the position z_m, have a nonlinear impact on the projected image. In Fig. 3, we show how perspective distortion, defined as the difference between perspective and orthographic projection, decreases as T_z increases, whereas perspective distortion quickly increases as T_z decreases in the close range. This phenomenon presents two key insights: (1) The amount of perspective distortion observed in an image is strongly correlated to the subject's Z-distance T_z to the camera and hence can be exploited to reliably estimate T_z directly from the image (Sec. 3.2). (2) The same person and pose can result in significantly different projections in the image depending on T_z. Thus, when estimating the 3D mesh of the person, the model needs to consider the influence of T_z (Sec. 3.3).

3.2. Predicting Z-Translation T_z

The amount of perspective distortion of a person in an image I is determined by T_z, *i.e.*, their distance to the camera (Fig. 3). Thus, we build a pelvis depth estimator F^{T_z} that directly estimates the depth of their pelvis from their appearance in a cropped image I_{crop} around them, $T_z = F^{T_z}(I_{crop})$. For F^{T_z} we employ a state-of-the-art pre-trained monocular depth prediction network DAv2 [33] as a pre-trained backbone to extract appearance features from I_{crop}, but we do not use their depth maps directly because they are highly inaccurate. Instead, we feed the appearance features into a learnable ConvNet followed by a transformer module to estimate the pelvis depth T_z. We find DAv2 [33] to be the best-performing among recent alternatives [12, 27, 33] (Tab. 2). Note that, as depth can increase to infinity, it is impractical to accurately predict depth for the entire unbounded range due to the model's limited learning capacity. We show in the supplemental material that current backbones struggle to simultaneously achieve high accuracy for both near ranges (SPEC-MTP [15]) and farther ranges (HuMMan [4]). Hence, it is more important for the model to learn accurate depth prediction for $<$1.2m, where perspective distortion manifests more strongly, versus farther ranges. To encourage this, while training F^{T_z} we weigh the T_z error inversely in proportion to the ground truth depth T_z^{GT} resulting in the weighted L_1 depth loss:

$$L_{depth} = 1/T_z^{GT} \cdot \left\| T_z - T_z^{GT} \right\|_1. \quad (3)$$

To avoid unstable division by small T_z values, we remove samples with $T_z \leq$0.5m from training.

3.3. T_z-aware Pose Estimation

As discussed in Sec. 3.1 and Fig. 2, T_z affects the appearance of the human body in the image and thus the accuracy of pose estimation. Therefore, we design a T_z-aware pose estimation block F^{pose} (Fig. 4) that takes the input image I and T_z translation to predict the human mesh as SMPL-X parameters, *i.e.* (β, θ). Specifically, BLADE employs the HMR algorithm AiOS [29], which directly predicts human meshes from the original uncropped image I. The method extracts features from a pre-trained backbone and contains a transformer-based encoder and non-autoregressive decoder for set prediction of the poses of all persons in an image. It

is trained on large amounts of real-world and synthetic images making it highly generalizable. However, its training data mostly contains distant persons, making it not accustomed to close-range people with strong perspective distortion. We find that naively fine-tuning AiOS with smaller close-range datasets employed in [31] results in over-fitting and undermines its generalizability (Table 3).

To achieve both generalizability and T_z-awareness, our pose estimator F_{pose} retains the existing knowledge of the pretrained AiOS while injecting additional depth information $T_z = F^{T_z}(I)$ through a ControlNet [36] style architecture (Fig. 4, pose estimator block). Specifically, we freeze AiOS and create a trainable copy of its backbone. The trainable copy is initialized with the pretrained weights, and its output is passed through a zero-initialized MLP before summing with the original output from the frozen backbone. Before training starts, the zero-MLP creates a zero residual and thus guarantees the same performance as the original AiOS. Once training starts, the zero-MLP becomes non-zero and allows the trainable backbone to improve upon the original AiOS. To condition the pose backbone on T_z, we use two MLPs to encode T_z into deep features, and we inject the T_z features into the trainable backbone by summing them with the backbone's encoder features. This way, the existing knowledge is retained in the frozen backbone while the trainable backbone acquires new knowledge about how the T_z distance affects the appearance of the human body in close-range images.

We input the predicted shape and pose parameters (β, θ) to the SMPL-X function M to obtain the vertices V and joints J with the pelvis joint at the origin:

$$(\beta, \theta) = F^{pose}(I|T_z), \qquad (V, J) = M(\beta, \theta). \quad (4)$$

To supervise the estimation of human shape, we calculate a shape loss L_{shape} as the L_1 distance between the ground truth shape vector β_{GT} and predicted shape vector β:

$$L_{shape} = L_1(\beta, \beta_{GT}). \quad (5)$$

To supervise the pose parameters, we use an angular error between the predicted joint rotations θ and ground truth joint rotations θ_{GT} (including the root joint orientation):

$$L_{pose} = E_{ang}(\theta, \theta_{GT}). \quad (6)$$

We also supervise the position of the estimated SMPL-X joints using a joint location loss L_{joint} as the L_1 distance between the predicted joint locations J and ground truth joint locations J_{GT}:

$$L_{joint} = L_1(J, J_{GT}). \quad (7)$$

Finally, we supervise the prediction of the mesh vertices by calculating the vertex loss L_{vert} as the distance between ground truth vertices V_{GT} and predicted vertices V:

$$L_{vert} = L_1(V, V_{GT}). \quad (8)$$

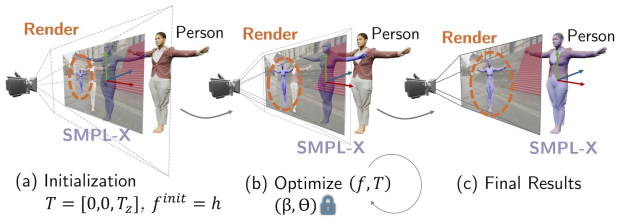

Figure 5. **Solving for $(\mathbf{f}, \mathbf{T_x}, \mathbf{T_y})$** : (a) With initial $(f, T_x, T_y) = [h, 0, 0]$, the estimated T_z and human mesh parameters (β, θ), the optimal (f, T_x, T_y, T_z) is derived (b) by optimizing the image space alignment through differentiable rasterization [18]. (c) The optimized parameters correctly align the projected 3D human mesh to the person in the image.

In summary, the total loss of our pose network is:

$$L = w_{shape} \cdot L_{shape} + w_{pose} \cdot L_{pose} + w_{joint} \cdot L_{joint} + w_{vert} \cdot L_{vert}, \quad (9)$$

where we use $w_{shape} = 1$, $w_{pose} = 1$, $w_{joint} = 5$, $w_{vert} = 5$ to balance the magnitudes of the different losses.

3.4. Solving for Focal Length and 3D Translation

The foundation of our method is the observation that, once T_z is determined, $[f, T_x, T_y]$ can be solved as alignment parameters. This is because when T_z is fixed, $[T_x, T_y]$ controls movements in the $z = T_z$ plane and f controls the scale of the image. Therefore, we reformat the problem as alignment and solve it through differentiable rasterization (Fig. 4, brown box). We render the predicted SMPL-X mesh with an initial translation $T = [0, 0, T_z]$ and the initial focal length equals to the image height $f^{init} = h$. Specifically, we rasterize the SMPL-X model as a binary mask, where pixels are 1 for the projected mesh surface and 0 otherwise. Then, through differentiable rasterization [18], we optimize for a tensor (f, T_x, T_y) that maximizes the intersection-over-union between the rasterized SMPL-X mask and the mask of the person, which is generated using an off-the-shelf segmentation method [25]. To ensure smooth gradient flow over the entire image, we apply Gaussian smoothing to both the rasterized and segmented masks. The process is visualized in Fig. 5 where (1) the purple SMPL-X model shifts to the right such that its projection aligns with the person in the image, and (2) the camera adjusts its focal length to align the sizes of the rasterized and segmented masks. Additionally, we find that optimizing for T_z, and potentially pose and global orientation, often further improves the quality of estimated pose and camera parameters.

3.5. Synthetic Dataset

While perspective distortion is more severe for the depth range smaller than 1.2m (Sec. 3.1), existing datasets [4, 9] for HMR do not contain enough data for this range. An evaluation of T_z distribution for various datasets is included in the supplemental material. Therefore, we cre-

Figure 6. **Examples of our synthetic BEDLAM-CC dataset.** High variation in lighting and camera angles as well as strong close-up distortion are intentionally part of the data. Images are rendered with very wide FoVs to enable arbitrary crop augmentation without re-rendering.

ate a new large synthetic dataset we name BEDLAM-CC ("close camera") utilizing assets provided with the BEDLAM dataset [2]. It contains 2 million synthetically rendered images enhancing current data for depth estimation. We show example images of our dataset in Fig. 6. Focused on challenging close-range images, we uniformly sample the inverse depth $1/T_z$ approximating the perspective distortion curve (Fig. 3) to generate this data. We enforce that 80% of the samples are within the range of 0.3m $\leq T_z \leq$ 1.2m and the remaining samples in the range of 1.2m $< T_z \leq$ 10m. BEDLAM-CC is used alongside other datasets to train our Pelvis Depth Estimator F^{T_z}. For fair comparisons during pose estimation, we do not use BEDLAM-CC during pose learning. We also create a separate test set from it for evaluation to provide more accurate ground truth data with a higher depth range. Please refer to the supplemental material for more details on the BEDLAM-CC dataset generation.

4. Experiments

We evaluate our method using existing benchmarks and also present extensive results on real-world images. Our approach recovers both camera parameters and the human mesh, achieving high 3D accuracy as well as precise 2D alignment, whereas prior methods typically excel at only one or the other [8].

4.1. Datasets

We train our model using a subset of 3D datasets employed in ZOLLY [31], *i.e.* H36M [9], PDHUMAN [31], and HUMMAN [4]. These datasets provide labeled camera and SMPL parameters, which we convert to the state-of-the-art SMPL-X model using the method from Choutas *et al*. [28]. Following ZOLLY [31], we evaluate our method on datasets with strong perspective distortions including SPEC-MTP, HUMMAN, PDHUMAN, and our dataset BEDLAM-CC. SPEC-MTP [15] is a real-world dataset with distances ranging from 0.5m to 2m. PDHUMAN [31] is a synthetic dataset with distances ranging from 0.5m to 1.8m, where many samples are around 0.6m. We identified some inconsistencies in the ground truth labels of PDHUMAN, which we visualize in the supplementary material. HUMMAN [4] is a multi-view dataset captured in a studio, exhibiting limited visual diversity and a narrow distance range of 1.75m to 2.2m. To address the above shortcomings, we perform additional evaluations on our BEDLAM-CC which provides accurate ground truth labels and diverse depth ranging from 0.3m to 10m (Sec. 3.5), with 80% of the samples within 1.2m. We report performance on HUMMAN, 3DPW, and H36M in the supplementary material, alongside visualization of inconsistencies in PDHUMAN, distributions of depth and body height, and runtime.

4.2. Training

Our framework contains two modules that require training, namely the pelvis depth estimator F^{T_z} and the pose estimator F^{pose}. We train them in two stages. During the first stage, we train the pelvis depth estimator F^{T_z} with a total batch size of 128 on 8 NVIDIA A100 GPUs for 4 epochs. In the second stage, we freeze F^{T_z}, feed its prediction of T_z to the pose estimator F^{pose}, and train F^{pose}. The second stage of training uses a batch size of 336 on 48 NVIDIA A100 GPUs for 4 epochs. The optimization of focal length, and translation vector $T = [T_x, T_y, T_z]$ requires no training.

4.3. Evaluation Metrics and Baselines

We evaluate the quantitative performance of all methods using standard metrics and introduce new metrics to evaluate the recovered camera parameters. We use mean Intersection-over-Union (mIoU) percentage to measure the accuracy of 2D alignment between the rendered mesh and the ground truth mask in the image. We use the Per-Vertex Error (PVE) in millimeters to measure the accuracy of the 3D mesh as the mean Euclidean distance between the 3D vertices of predicted and ground truth meshes. We also notice that existing metrics ignore the accuracy of the estimated camera parameters, which is crucial to achieving consistent 3D pose estimation and 2D pose alignment. Therefore, we introduce new metrics to evaluate the accuracy of the recovered camera parameters. The common camera model includes focal length and the translation and rotation of the subject in camera space. We measure the accuracy of the recovered focal length as the percentage error

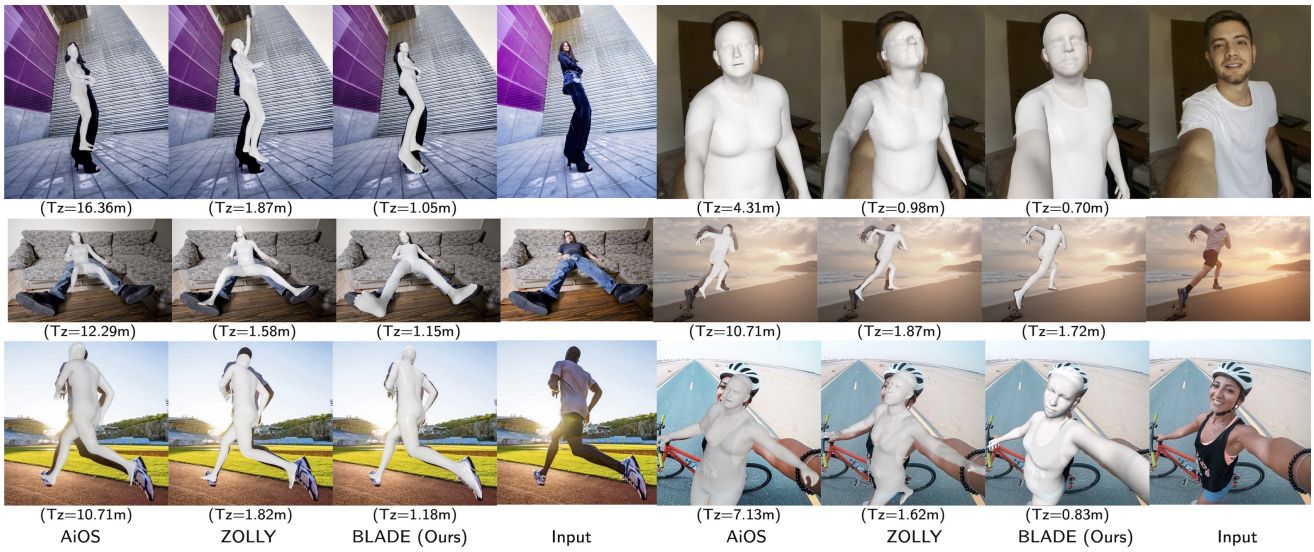

Figure 7. **Qualitative SOTA comparison.** We compare with SOTA methods for single-view human mesh recovery including AiOS [29], and ZOLLY [31]. Our method BLADE is consistently more accurate in terms of estimated pelvis depth T_z of the person (metrical distances given in parenthesis), focal length, and 2D alignment. Notice the improvements for areas with strong perspective effects close to the camera. Image sources are given in the supplemental material.

with respect to the ground truth focal length:

$$E_f = |f_{pred} - f_{GT}|/f_{GT}. \tag{10}$$

Given that T_z has a direct inverse relationship with the amount of distortion in the image (Fig. 3), 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:

$$E_{T_{xy}} = \|T_{xy}^{pred} - T_{xy}^{GT}\|_2, \tag{11}$$

$$E_{T_z} = |T_z^{pred} - T_z^{GT}|, \tag{12}$$

$$E_{1/T_z} = |1/T_z^{pred} - 1/T_z^{GT}|. \tag{13}$$

We omit a dedicated 3D rotation error given that 3D rotation is already evaluated as a part of MPJPE.

4.4. Comparison to State-of-the-Art Methods

Quantitative Results: In Table 1, we compare our method BLADE with state-of-the-art single image HMR methods. BLADE surpasses the current SOTA for close-range HMR, ZOLLY [31], on all datasets and achieves the best overall 2D alignment, 3D localization, and pose estimation. Notably, BLADE obtains a relative improvement of **85.9**% E_{T_z} and **21.4**% PVE on the SPEC-MTP [15] dataset and **44.8**% mIoU on the BEDLAM-CC dataset. We also evaluate recent SOTA methods AiOS [29], TokenHMR [8] and SMPLer-X [5], using their respective publicly released models. These methods don't explicitly estimate focal length and instead use a constant focal length of 5000. They estimate accurate 3D meshes with low PVE values but are inaccurate in terms of 2D alignment, focal length and 3D translation. The common tradeoff between 2D and 3D accuracy is discussed in detail in TokenHMR [8].

Additionally, we find that good performance on the synthetic PDHUMAN dataset [31] does not reflect good performance in real-world usage. As shown in Table 1, recent SOTA methods [5, 8, 29] perform well on the real-world dataset SPEC-MTP but substantially worse on PDHUMAN in terms of PVE. Whereas ZOLLY [31] performs well on PDHUMAN but less so on SPEC-MTP [15]. We suspect that this potential gap is due to: (1) the extreme distortion in PDHUMAN which is not present in real-world data, and (2) inconsistencies in its ground truth labels (detailed in the supplementary). We thus show two versions of BLADE: (i) "Ours" trained with a balanced distribution across the 3 training datasets; and (ii) "Ours (real-world)" trained with increased sampling from HUMAN3.6M and decreased sampling from PDHUMAN. "Ours" performs well on each dataset compared to other methods and performs best on PDHUMAN. "Ours (real-world)" performs the best on SPEC-MTP, BEDLAM-CC, and in real-world usage. Please refer to the supplementary for an expanded version of Table 1 with all metrics and additional results.

Qualitative Results: In Fig. 1 and Fig. 7, we show results of SOTA methods AiOS [29] and Zolly [31], and our method on real-world images. BLADE performs significantly better than compared methods in terms of 2D alignment of the mesh to the image, 3D body mesh, and the accuracy of perspective distortion. The alignment of body parts close to the camera is specifically improved by our method. More visual results are included in the supplementary.

Methods	SPEC-MTP [15] (real-world capture)						PDHUMAN [31] (synthetic)						BEDLAM-CC (synthetic)					
	$E_{T_z}\downarrow$	$E_{1/T_z}\downarrow$	$E_{T_{xy}}\downarrow$	$E_f\downarrow$	PVE↓	mIoU↑	$E_{T_z}\downarrow$	$E_{1/T_z}\downarrow$	$E_{T_{xy}}\downarrow$	$E_f\downarrow$	PVE↓	mIoU↑	$E_{T_z}\downarrow$	$E_{1/T_z}\downarrow$	$E_{T_{xy}}\downarrow$	$E_f\downarrow$	PVE↓	mIoU↑
ZOLLY [31]	0.899	0.394	0.906	1.063	126.7	62.3	0.255	0.355	0.267	0.273	82.0	53.0	0.539	0.634	0.564	0.461	131.8	51.8
SMPLer-X*[29]	0.980	0.450	0.109	1.121	102.6	53.0	2.223	1.030	0.126	0.550	161.2	47.6	2.057	1.172	0.087	1.349	139.9	53.0
TokenHMR*[8]	0.909	0.436	0.095	1.121	124.3	49.7	2.280	1.034	0.068	0.550	156.7	53.0	2.378	1.200	0.096	1.349	136.4	54.2
AiOS*[29]	1.035	0.464	0.121	1.121	110.9	48.7	2.312	1.024	0.149	0.550	183.4	49.5	2.340	1.197	0.111	1.349	143.0	54.6
Ours	0.129	0.114	0.056	0.163	111.9	68.7	**0.106**	**0.176**	**0.043**	**0.216**	**80.5**	**67.3**	0.326	0.305	0.079	0.257	111.6	74.6
Ours (real-world)	**0.127**	**0.112**	**0.044**	**0.159**	**99.6**	**69.5**	0.107	0.178	0.049	0.223	102.6	65.2	**0.325**	**0.305**	**0.076**	**0.212**	**106.8**	**75.0**

Table 1. **Quantitative comparison to SOTA methods.** Evaluation on SPEC-MTP [15], PDHUMAN [31], and BEDLAM-CC [2] datasets. Our method achieves SOTA results. Best results indicated by bold numbers. For additional metrics and test datasets please refer to the supplemental material. * symbol indicates pre-trained public models. Model version "Ours" is trained using 3D datasets used in ZOLLY [31] whereas "Ours (real-world)" is trained with increased sampling frequency for real-world data HUMAN3.6M [9].

	DiNOv2 [27]	Sapiens [12]	DAv2 [33]	Ours
$E_{T_z}\downarrow$	0.300	0.210	0.154	**0.127**

Table 2. **Ablation study for depth backbone.** Test on SPEC-MTP [15]. "Ours" is using DAv2 as the depth backbone [33] and fine-tuned using different augmentations.

	PA-MPJPE↓	MPJPE↓	PVE↓
raw AiOS	62.816	101.577	110.851
ft. AiOS	64.932	113.173	120.582
Ours (T_z cond.)	**56.666**	**94.050**	**99.635**

Table 3. **Ablation study for conditioning.** Test on SPEC-MTP [15]. Architecture: DAv2 [33] used in pelvis depth estimator. First row: AiOS [29] used as pose estimator. Second and third row "Ours": AiOS [29] with ControlNet [36] used as pose estimator with and without conditioning on T_z.

4.5. Ablation Study

Ablation of pelvis depth estimator. Accurate depth estimation is the core to solving for other variables. In Table 2 we evaluate various foundation models including DiNOv2 [27], Sapiens [12], and DAv2 [33] as the backbone to our pelvis depth estimator F^{T_z}. The models are trained using HUMMAN [4], PDHUMAN [31], and HUMAN3.6M [9]. On the most challenging real-world SPEC-MTP [15] dataset, DAv2 achieves the best accuracy with E_{T_z} = 15.4cm. Finally, "Ours" is a version of the DAv2-based F^{T_z} trained with improved augmentation and additional data from our BEDLAM-CC dataset (Sec. 3.5), which provides many close-range images (<1m), and thus further reduces the T_z error from 15.4cm to 12.7cm.

Conditioning the pose estimator. In Table 3, we evaluate various architectures of pose estimator on the task of 3D pose estimation and mesh recovery on the challenging close-range real-world SPEC-MTP dataset [15]. The publicly available "raw AiOS" performs well. However, after fine-tuning ("ft. AiOS") with the HUMMAN, PDHUMAN, H36M datasets, which mostly contain faraway subjects and synthetic images, its performance degrades on the close-range real-world SPEC-MTP dataset [15], by losing its good generalization to real-world data. On the other hand, conditioning raw AiOS [29] in T_z through a ControlNet-style architecture [36] that we proposed in BLADE (Fig. 4), leads to significant improvements in pose estimation performance. It enables the pose backbone to retain its previous knowledge while learning the correct relationship between T_z and the image to enhance 3D pose estimation.

Limitations. We currently only consider single-person images. For the future, we plan to extend our method to process videos where more information can be leveraged for better accuracy. We also do not consider lens distortion or camera types other than the standard pin-hole camera such as fish eye lenses. Lastly, the estimation of (f, T_x, T_y) can fail when the segmentation mask is very inaccurate. A promising direction is learnable optimization to substitute differentiable rasterization for better robustness.

5. Conclusion

In this work, we propose BLADE – a method for human mesh recovery and perspective camera estimation from single images. This is a long-standing challenging and open problem. Different from previous work, we provide a solution to estimating perspective projection parameters without conversion from an orthographic camera model. We underscore the significance of accurate and disentangled pelvis depth estimation, followed by depth-conditioned human pose estimation, and finally optimization of camera focal length and XY-translation. We also introduce a large-scale synthetic single-person dataset, BEDLAM-CC, containing a large number of close-range images with ground truth labels for the perspective camera and SMPL-X body parameters. Our framework BLADE achieves state-of-the-art accuracy on a variety of benchmarks and across a wide range of depths. Among other use cases, the method can be applied for accurate pose labeling of in-the-wild image datasets to train robust human-centric models.

References

[1] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth, 2023. 3

[2] 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, 8

[3] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 2, 3

[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. 4, 5, 6, 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. 2, 3, 7

[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. 2

[7] Shradha Dubey and Manish Dixit. A comprehensive survey on human pose estimation approaches. *Multimedia Systems*, 29(1):167–195, 2023. 2

[8] 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. 1, 2, 3, 6, 7, 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, 6, 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. 1, 2, 3

[11] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[12] Rawal Khirodkar, 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. 4, 8

[13] Imry Kissos, Lior Fritz, Matan Goldman, Omer Meir, Eduard Oks, and Mark Kliger. Beyond weak perspective for monocular 3d human pose estimation. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 541–554. Springer, 2020. 2, 3

[14] 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. 1, 2

[15] 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. 1, 2, 3, 4, 6, 7, 8

[16] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2252–2261, 2019. 1, 2

[17] 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. 2

[18] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (ToG)*, 39(6):1–14, 2020. 5

[19] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021. 1

[20] 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. 1, 2, 3

[21] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. *CVPR*, 2023. 2, 3

[22] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12939–12948, 2021. 2

[23] Yang Liu, Changzhen Qiu, and Zhiyong Zhang. Deep learning for 3d human pose estimation and mesh recovery: A survey. *Neurocomputing*, page 128049, 2024. 2

[24] 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. 2

[25] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 5

[26] Koki Nagano, Huiwen Luo, Zejian Wang, Jaewoo Seo, Jun Xing, Liwen Hu, Lingyu Wei, and Hao Li. Deep face normalization. *ACM Transactions on Graphics (TOG)*, 38(6): 1–16, 2019. 3

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3, 4, 8

[28] 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. 2, 3, 6

[29] 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. 2, 4, 7, 8

[30] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2023. 2

[31] 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, 5, 6, 7, 8

[32] Yufu Wang and Kostas Daniilidis. Refit: Recurrent fitting network for 3d human recovery. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[33] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. *arXiv:2406.09414*, 2024. 2, 3, 4, 8

[34] Wei Yao, Hongwen Zhang, Yunlian Sun, Yebin Liu, and Jinhui Tang. W-hmr: Monocular human mesh recovery in world space with weak-supervised calibration. *arXiv preprint arXiv:2311.17460*, 2023. 3

[35] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11446–11456, 2021. 2

[36] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 5, 8

[37] Yajie Zhao, Zeng Huang, Tianye Li, Weikai Chen, Chloe LeGendre, Xinglei Ren, Ari Shapiro, and Hao Li. Learning perspective undistortion of portraits. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7849–7859, 2019. 3

[38] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37, 2023. 2

[39] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2