CameraHMR: Aligning People with Perspective

Priyanka Patel[1]
Meshcapade, Germany
priyanka@meshcapade.com

Michael J. Black
Max Planck Institute for Intelligent Systems
black@tuebingen.mpg.de

Figure 1. **Putting people in perspective.** In contrast to common methods like HMR2.0, **CameraHMR** estimates 3D human shape and pose using a *perspective camera* by leveraging a learned regressor, **HumanFoV**, to estimate the appropriate camera intrinsics. Note how this improves the estimated pose when there is strong *foreshortening*. Our approach exploits new pseudo ground-truth data and a new dense surface keypoint detector that improve body shape estimation; this is particularly visible for the heavier people in the images. CameraHMR defines the new state-of-the-art for 3D human pose and shape accuracy from a single image.

Abstract

We address the challenge of accurate 3D human pose and shape estimation from monocular images. The key to accuracy and robustness lies in high-quality training data. Existing training datasets containing real images with pseudo ground truth (pGT) use SMPLify to fit SMPL to sparse 2D joint locations, assuming a simplified camera with default intrinsics. We make two contributions that improve pGT accuracy. First, to estimate camera intrinsics, we develop a field-of-view prediction model (HumanFoV) trained on a dataset of images containing people. We use the estimated intrinsics to enhance the 4D-Humans dataset by incorporating a full perspective camera model during SMPLify fitting. Second, 2D joints provide limited constraints on 3D body shape, resulting in average-looking bodies. To address this, we use the BEDLAM dataset to train a dense surface keypoint detector. We apply this detector to the 4D-Humans dataset and modify SMPLify to fit the detected keypoints, resulting in significantly more realistic body shapes. Finally, we upgrade the HMR2.0 architecture to include the estimated camera parameters. We iterate model training and SMPLify fitting initialized with the previously trained model. This leads to more accurate pGT and a new model, CameraHMR, with state-of-the-art accuracy. Code and pGT is available for research purposes.

[1]This work was done when PP was at MPI-IS.

1. Introduction

The field of monocular 3D human pose and shape (HPS) estimation has advanced rapidly. Updated architectures, stronger backbones, and more extensive training data have all led to improvements in robustness and accuracy. We argue that a key remaining source of error lies in the fact that many HPS methods use a simplified weak-perspective camera model. We describe how the wrong camera model introduces error and we propose a solution. Specifically, we collect a dataset of images of people with varied field of view (FoV) and train a network to directly predict FoV from pixels. We then leverage this predicted FoV in training and show how this leads to state-of-the-art accuracy.

Recent HPS methods, such as HMR2.0 [13], achieve notable 2D alignment by leveraging large-scale real image datasets for training. However, this success in 2D alignment comes at the cost of reduced 3D accuracy as described in [10]. The core issue lies in the fact that these large-scale real image datasets frequently lack camera intrinsic parameters. Training involves first creating 3D pseudo ground truth (pGT) data by fitting a parametric 3D body model like SMPL [33] to 2D features such as keypoints. This fitting process typically uses a weak perspective camera model or default camera intrinsics. When the camera model is wrong, fitting 2D keypoints accurately forces the 3D pose to be wrong. Consequently, methods trained on these pGT datasets learn to replicate the 3D errors.

To achieve both accurate 2D alignment and 3D poses, it is crucial to use the correct camera intrinsics in creating the pGT. Unfortunately, estimating intrinsics from a single image is challenging. While there are many state-of-the-art approaches for camera calibration from monocular images [18, 29, 48], they are trained on datasets such as Google Street View [4] and SUN360 [44]. Such datasets focus on outdoor or indoor scenes rather than people. Methods trained on datasets containing panoramic images of streets, natural landscapes, urban scenes, indoor settings, etc., do not work well on images of people. On the other hand methods trained on synthetic data like SPEC-camcalib [26] do not generalize well to in-the-wild data. This highlights the need for a robust camera calibration model for images containing people to achieve accurate 3D human pose and shape estimation.

To address this problem, we collect a dataset of about 500K images predominantly comprising people, to train a field of view (FoV) prediction model. The human body provides useful information for camera estimation. While it would be going too far to call the body a "calibration object," bodies have highly regular proportions and a limited range of heights. When projected into images, this regular structure systematically varies with focal length and perspective projection. To exploit this fact, we train a direct FoV regressor, **HumanFoV**, which generalizes well on benchmarks featuring humans compared to other state-of-the-art (SOTA) camera calibration methods. HumanFoV can be directly incorporated into HPS methods that use a full perspective camera model, enabling accurate 3D reconstruction. Using an accurate camera not only helps HPS regressors infer the 3D location of the people in camera space but it also improves alignment of the inferred body with image features, especially for wide angle images and extreme viewing angles.

While incorporating a more accurate camera model into HPS methods is important, we still need high-quality training data that is as diverse as possible. To that end, we use HumanFoV to improve the real-image pGT in the 4DHumans dataset that is used to train HMR2.0 [13]. The original dataset uses SMPLify [6] to fit SMPL to 2D keypoints under a weak-perspective assumption. Instead, we use a full perspective camera model in SMPLify and exploit HumanFoV to estimate the FoV of the training images.

Additionally, the original dataset is created by fitting SMPL to only 17 sparse 2D joints; these lack the detail necessary for accurate 3D shape reconstruction. To improve this, we train a keypoint detector (**DenseKP**) on the BEDLAM [5] dataset to estimate *138 dense surface keypoints*. We modify SMPLify to use these together with the original 17 2D joints. This results in significantly more realistic body shapes. Qualitatively, the improved camera model and dense keypoints lead to good 2D image alignment and more plausible 3D pGT compared to original dataset (Fig. 2).

With this, we generate high-quality 3D pGT for a large-scale real image dataset comprising approximately 3.2 million cropped images. Importantly, the dataset includes the camera intrinsics estimated by HumanFoV; these are crucial for HPS methods. We further modify the HMR2.0 architecture to incorporate camera parameters from HumanFoV in training. We iterate training this new **CameraHMR** model and refining the pGT with SMPLify initialized with the previously trained model. This significantly improves performance, with CameraHMR achieving state-of-the-art accuracy on multiple HPS benchmarks. See Fig. 1.

In summary, we (1) collect a dataset of varied images of humans with known FoV, (2) using this dataset, we train HumanFoV to regress FoV from images of people, (3) update SMPLify with a full perspective model that uses the HumanFoV output, (4) introduce a dense surface keypoint regressor and incorporate these keypoints into SMPLify, (5) improve the 4DHumans training set using the new version of SMPLify, (6) incorporate the FoV estimation in HMR2.0, (7) train a new model, CameraHMR, with SOTA accuracy. Code and pGT dataset is available for research purposes.

2. Related Work

3D Human Pose and Shape Regression. The field of mononcular 3D human pose and shape (HPS) estimation

has made rapid advances. The improvement of the backbone has played an important role, beginning with ResNet architectures pre-trained on the ImageNet dataset [20, 21, 28], then the HRNet architecture pre-trained on the COCO dataset [7, 11, 30, 47], and more recently Transformer-based models [10, 13, 42]. These changes have led to significant improvements in accuracy on standard benchmarks.

HMR [21] introduced a simplified weak perspective camera model to facilitate training with pseudo ground truth datasets. The availability of increasingly accurate 3D ground truth datasets, enables methods to be trained using the more complex full perspective camera model [5, 25, 31]. This evolution in camera modeling and training backbone has contributed to improvements in the accuracy of 3D pose and shape estimations. Despite advancements in achieving accurate 3D pose estimates, aligning these poses accurately with 2D image features remains challenging. This issue has been recently highlighted by TokenHMR [10], which attributes the misalignment to the use of incorrect camera models during prediction. Even methods that incorporate a full perspective camera model during training [25, 31, 42], encounter alignment issues due to the absence of camera intrinsics during inference, leading to inaccurate projection into 2D.

Regressing Camera Intrinsics. Several approaches have been developed to regress camera intrinsics from monocular images [18, 27, 29, 48]. However, these methods are often trained on datasets focused on indoor [39, 45], driving [4, 12], or object-centric [2] images, which typically feature consistent vanishing points and geometric cues. These images are quite different from images of people. Models trained on indoor and outdoor scenes struggle when presented with images of people, particularly portrait images where vanishing points are often unclear or absent. We argue that the human body itself offers essential cues for camera parameter estimation. To leverage this, we train our model on a dataset composed predominantly of images featuring people. Our experiments demonstrate that this approach significantly enhances the generalization of camera estimation across various human-centric benchmarks.

3. Method

In this section, we detail our approach for estimating camera intrinsics from images captured in uncontrolled settings. We further present enhancements to the HMR2.0 architecture by incorporating bounding box and camera parameter tokens into the vision transformer and adopting a perspective camera model for projection, in contrast to the previously used weak perspective model. We designate the two models as HumanFoV and CameraHMR. Furthermore, we describe the generation of improved pseudo-ground truth (pGT) by using a modified SMPLify fitting process, **CamSMPLify**.

3.1. Preliminaries

We use a simple perspective camera model for our experiments. In this model, a 3D point (X, Y, Z) in camera coordinate space is projected onto the image plane, at a point (x, y), using the camera intrinsic matrix parameterized by focal length f (assuming $f_x = f_y$) and principal point, (c_x, c_y) in pixels coordinates. We simplify the model by assuming no radial distortion and that the principal point is the center of the image. For CameraHMR, we assume the body is always predicted in camera space, i.e. the rotation of the camera is $R = I$.

As is common practice [27, 29, 48], we estimate the field of view from images instead of the focal length. The reason for this is that different focal lengths can produce images with same field of view if the camera sensor size is different. For example, a 50mm lens on a full-frame camera (35mm sensor) has a wider field of view than a 50mm lens with a crop sensor. While the focal length is an important characteristic of a lens, the field of view is a more direct measure of what will be captured in an image. Hence, we use vertical field of view υ as the primary output of HumanFoV. The focal length f_y used in the camera intrinsics can be derived from υ using the image height H.

$$f_y = \frac{H}{2 \cdot \tan\left(\frac{\upsilon}{2}\right)}. \tag{1}$$

To represent the 3D human, we use the SMPL parametric human body model [33] controlled by parameters (θ, β), where $\theta \in \mathbb{R}^{72}$ represents pose and $\beta \in \mathbb{R}^{10}$ represents identity shape. SMPL is a function that outputs a body mesh, $\mathcal{M}$ with vertices $V \in \mathbb{R}^{6890 \times 3}$. The 3D joints $J_{3d} \in \mathbb{R}^{K \times 3}$ with K joints are obtained using a pre-trained joint regressor. We use a gender-neutral SMPL model with 10 shape components. We use a cropped bounding box of size 256×256 as input to our models.

3.2. HumanFoV

Given an image $I \in \mathbb{R}^{W \times H \times 3}$ where W and H are the width and height of the original image respectively, we preprocess it by resizing it to a square resolution of 256×256 pixels. To achieve this, we resize the longer side to 256 and zero-pad the smaller side to maintain the aspect ratio. This preprocessing ensures uniform input dimensions for the network while preserving the aspect ratio integrity of the original image, which is important for field of view estimation. We train a deep neural network architecture with HRNet [40] as the backbone and an MLP head for direct estimation of the vertical field of view υ_{pred}. The HRNet backbone is pretrained on ImageNet [8]. Based on insights from previous work [25, 27], underestimating the FoV has less negative impact on reconstructed 3D poses compared to overestimating the FoV. Therefore, an asymmetric loss

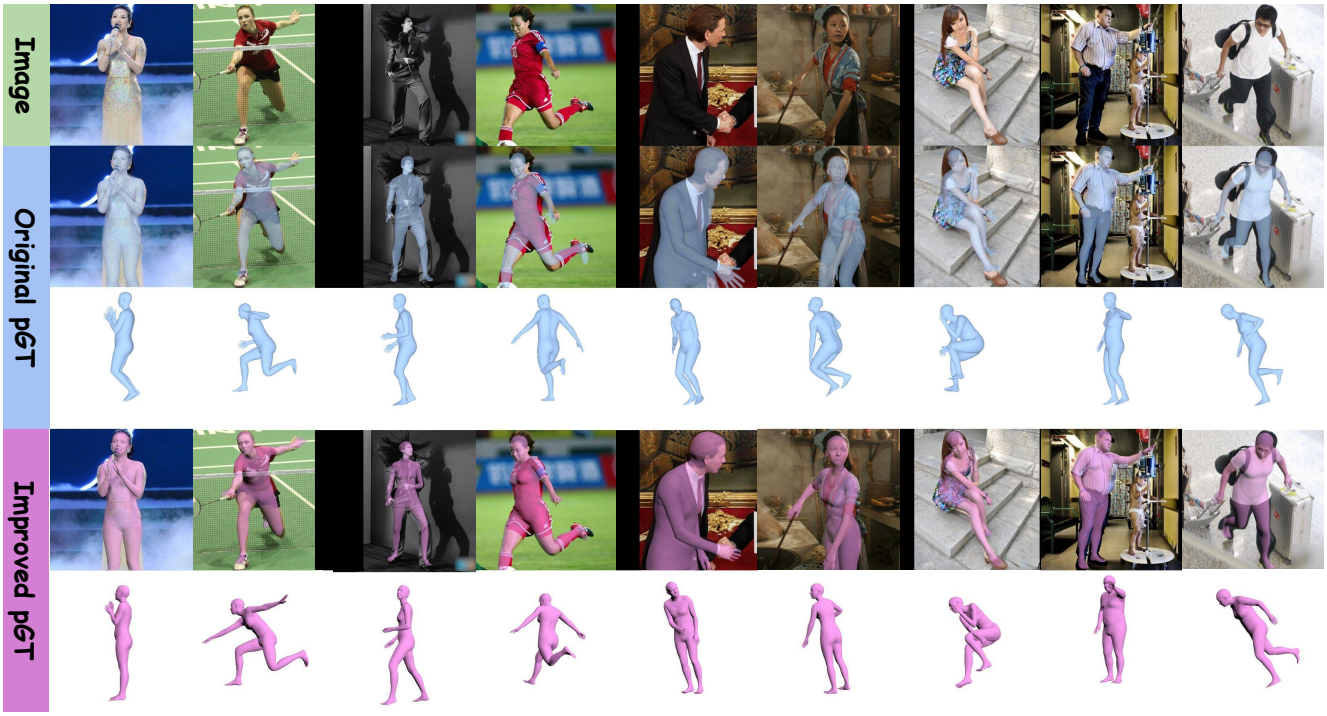

Figure 2. **Pseudo-Ground-Truth (pGT) training data.** Row 1: example images from the 4DHumans dataset. Rows 2 and 3: original pGT overlaid and viewed from a different perspective. Rows 4 and 5: our improved pGT using CamSMPLify. Note that our approach reduces the bias towards bent knees (columns 1, 5, 6), improves 3D pose and image alignment when there is foreshortening (Column 2, 4, 7, 9), and estimates more realistic body shape (columns 1, 3, 7, 8).

function is incorporated to penalize overestimation more heavily than underestimation. We define a loss, L_v, on the vertical field of view in radians as:

$$L_v = \begin{cases} 3\left\|v_{\text{gt}} - v_{\text{pred}}\right\|_2^2 & \text{if } v_{\text{pred}} > v_{\text{gt}} \\ \left\|v_{\text{gt}} - v_{\text{pred}}\right\|_2^2 & \text{if } v_{\text{pred}} \leq v_{\text{gt}}. \end{cases} \tag{2}$$

The training set for HumanFoV is described in Sec. 4.1. We train HumanFoV for around 16 epochs with a batch size of 64 and learning rate of 5×10^{-5}. We use an Adam optimizer [24] with no weight decay. We use different data augmentation techniques during training to ensure the model's robustness. This includes center-cropping of images to generate different aspect ratios. This augmentation helps the model become robust against variations in image cropping during inference. Images are also randomly flipped horizontally with a probability of 0.2, providing additional diversity to the training dataset.

3.3. CameraHMR

HPS methods have evolved to use progressively more powerful backbones from ResNet [15] to HRNet [40], and most recently ViT [9], resulting in improved performance. Here we use a ViTPose [46] backbone pretrained on COCO [32] to extract features from cropped images. Specifically, we adopt the HMR2.0 architecture, which employs a ViT backbone, and modify it to support training with a full perspective camera instead of a weak perspective camera.

The ViT backbone processes images by dividing them into patches, converting these patches into feature embeddings known as tokens, and utilizing self-attention mechanisms to capture the relationships among them. Along with the image tokens, we also provide bounding box information of the cropped region and the focal length as tokens. We follow CLIFF [31] and compute the bounding box token T_{bbox}, using the bounding box center c_x, c_y, scale s, and the focal length f of the full image.

$$\mathbf{T}_{\text{bbox}} = \left(\frac{c_x}{f}, \frac{c_y}{f}, \frac{s}{f}\right) \tag{3}$$

The ground truth focal length is known during training and predicted using our HumanFoV during inference.

The decoder in our modified architecture cross-attends to both image-derived tokens and the supplementary bounding box and focal length tokens. This approach enables the decoder to generate features essential for accurately regressing 3D rotations and human mesh parameters while accommodating camera perspective. We explain the losses used in training the model in Sup. Mat.

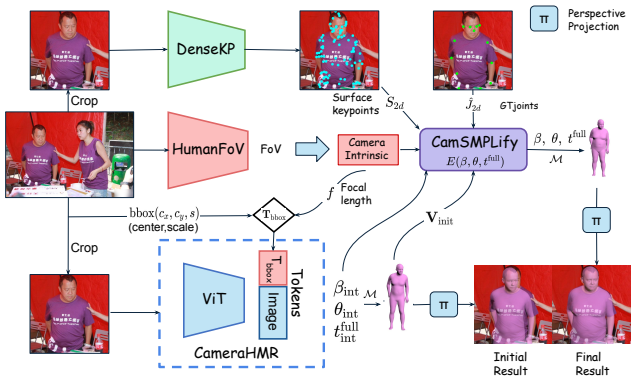

Figure 3. **Overview of CamSMPLify**: The DenseKP module processes cropped images to produce dense surface keypoints, while the HumanFoV module uses full images to estimate camera intrinsics. The output from these are used by CamSMPLify to optimize the SMPL model parameters, β, θ, and the global translation t^{full}. Our iterative training strategy starts with initial estimates, V_{init} from CameraHMR, which are used to regularize the CamSMPLify estimates. CameraHMR is then iteratively refined based on the improved pGT from CamSMPLify.

3.4. CamSMPLify

To train CameraHMR, we modify the original 4DHumans [13] dataset, upgrading it from a weak perspective to a full perspective camera format. This dataset includes images from the InstaVariety [22], COCO [32], MPII [3], AI Challenger [43] and AVA [14] datasets. Note that AVA [14] contains many movie clips where the aspect ratio is stretched horizontally (e.g. from 4:3 to 16:9), violating the assumption that $f_x = f_y$. These clips are hard to detect automatically so we exclude AVA from our improved pGT. We make several improvements to the fitting process to enhance the dataset, including better initialization and priors, more accurate camera intrinsics, dense surface 2D keypoints, and multiple fitting iterations. The overview is shown in Fig. 3. This comprehensive approach results in better pseudo ground truth, significantly improving both 3D pose and shape accuracy, as well as 2D alignment as shown in Fig. 2. In the following section, we explain each step in greater detail.

Camera Intrinsics. One major challenge with in-the-wild image datasets is the absence of ground truth intrinsic camera parameters, which makes it difficult to accurately fit a 3D body to the 2D keypoints while ensuring plausible 3D poses and precise 2D alignment. To address this issue, we employ our HumanFoV model to estimate the vertical field of view v for all images in the 4DHumans dataset. The corresponding focal length f is then calculated from v and image height H using Eq. 1. This approach enables us to infer the necessary camera intrinsics for the dataset, enabling us to project the 3D joints using a full perspective camera in contrast to the weak perspective camera used in the original dataset. As a result, we achieve accurate 2D alignments without compromising 3D pose accuracy during the fitting.

Surface Keypoints. Another challenge that we face is that the original 4DHumans dataset is annotated with 17 sparse 2D body joints. This sparse annotation lacks sufficient detail to accurately reconstruct the 3D body shape. To address this, we train a dense surface keypoint detector, DenseKP, using the synthetic datasets BEDLAM and AGORA, which offer diverse body shapes and precise ground truth annotations. We use ViTPose [46] pretrained on COCO to extract features from the cropped images. The model takes a centered crop of the person, resized to 256×256 pixels, as input and outputs 2D dense surface keypoints $S_{2d} \in \mathbb{R}^{138\times2}$. To generate the 138 surface keypoint ground truth labels, we down-sample the SMPL ground truth vertices and project them onto the image. The down-sampling approach, adapted from COMA [36], focuses on sampling vertices in high-curvature regions, which effectively preserves the body shape and key structural details. To train the model we use an L_2 loss on ground truth $\hat{S}_{2d}$ and predicted keypoints S_{2d}. We use this model to generate 138 dense keypoints for all images in the 4DHumans dataset. We modify the fitting by combining the 17 provided joints with the estimated 138 dense surface keypoints. This, along with improved camera intrinsics, helps achieve better shape alignment in 2D during projection.

Initialization. Like with any optimization-based method, the choice of initialization can significantly affect the performance of the fitting process, as poor initialization can lead to suboptimal convergence. To ensure a robust 3D initialization, we train our CameraHMR on the BEDLAM and AGORA dataset, which provides accurate ground truth annotations, including camera intrinsics. Using this pretrained model, we generate initial 3D pose θ_{int} and shape β_{int} predictions for the 4DHumans dataset, resulting in bodies with relatively accurate pose and shape. Additional refinement through surface keypoint and joint fitting improve both the body shape and pose, leading to even more precise results. We also predict the initial global translation of the mesh $t^{\text{full}}_{\text{int}}$ relative to the optical center of the full image. Note that instead of the standard pose prior used in SMPLify [6], we regularize the solution to this initial prediction.

Losses. We optimize the SMPL model parameters β and θ to match the ground truth 2D body joints $\hat{J}_{2d}$ and 2D surface keypoints $\hat{S}_{2d}$ while also optimizing the global translation of the mesh t^{full}. The model parameters are initialized with β_{int} and θ_{int}, as determined during the initialization stage using CameraHMR. The output vertices V_{int} from the model $M(\beta_{\text{int}}, \theta_{\text{int}})$ serve as the prior in the regularization process. J_{2d} and S_{2d} are obtained from the 3D joints J_{3d} and surface keypoints S_{3d} using $\Pi(J_{3d} + t^{\text{full}})$, where Π represents the perspective projection with camera intrinsics obtained from

HumanFoV. The S_{3d} are derived from the SMPL vertices V using a downsampling matrix similar to BEDLAM [5]. Specifically the optimization minimizes the following objective function:

$$E(\beta, \theta, t^{\text{full}}) = \lambda_{S_{2d}} E_{S_{2d}} + \lambda_{J_{2d}} E_{J_{2d}} + E_{\text{reg}} \quad (4)$$

$$E_{\text{reg}} = \lambda_{\beta} \|\beta\|_2^2 + \lambda_{\text{int}} \|V - V_{\text{int}}\|_2^2, \quad (5)$$

where $E_{J_{2d}}$ and $E_{S_{2d}}$ are L_2 losses on 2D joints and surface keypoints, respectively. The λ terms denote the weights for each component of the objective function. We apply a threshold value τ to filter out results with $E > \tau$, thereby excluding pseudo ground truth samples with high convergence errors. For further details, refer to the Sup. Mat.

Iteration. We run the whole process for multiple iterations of refinement to ensure the pseudo ground truth is of high quality. This is similar in spirit to SPIN [28]. The θ and β parameters for CamSMPLify fitting are initialized using version **v1** of the CameraHMR trained on the BEDLAM dataset. After applying the filtering criteria based on the threshold τ in CamSMPLify fitting we obtain approximately 2.8 million crops out of 4 million crops from 4DHumans dataset. These crops, together with the BEDLAM dataset, are then utilized to train an enhanced version, **v2**, of CameraHMR. We further iterate and employ **v2** to generate improved initializations for the 4D Humans dataset, followed by another round of CamSMPLify fitting. This improved initialization substantially improves convergence, leading to a more accurate fitting with lower error. Applying the same filtering criteria, we are able to further expand our dataset to around 3.2 million high-quality annotations.

4. Datasets

4.1. HumanFoV

To train the HumanFoV model we use around 500K images collected from Flickr [1]. To get human-centeric data, we filter Flickr with keywords such as people, man, woman, kid, human, crowd etc. We use the Flickr API to download only the images that have associated EXIF information, which usually contains the focal length in mm. To calculate the vertical field of view υ from the focal length f in mm, we need to know the sensor height sh. Vertical FoV υ can be calculated as

$$\upsilon = 2 \cdot \arctan\left(\frac{sh}{2 \cdot \mathsf{f}}\right). \quad (6)$$

We use the field *FocalLengthIn35mmFormat* from the EXIF, which contains the focal length corresponding to a 36mm wide and 24mm high sensor. This allows us to use a sensor height of 24mm directly in our calculations. Note that if the aspect ratio of the image is less than 1 (i.e. portrait mode), we calculate υ using the sensor width instead of sensor height. Please refer to the Sup. Mat. to see the distribution of focal lengths in the dataset.

Note that the standard aspect ratios for images captured from a phone or camera are in the range of 16:9, 4:3, 3:2, 1:1, 5:4. If the aspect ratio of the image collected from Flickr is outside this range, we assume that the image is cropped. Cropped images, especially those not centered, could introduce inconsistencies with the camera model used and might confuse the neural network. We filter such images from the training data. Although we do not use cropped images directly in our training data, we extensively apply crop augmentation during training to ensure model's robustness.

To evaluate our HumanFoV model, unlike previous methods, we focus on benchmarks containing images of people. Consequently we use test-set images from SPEC [27], 3DPW [41] and EMDB [23], which all provide camera intrinsics. We also create a test set of around 10K images from Flickr that are similar to training set. We also include a scene from the BEDLAM [5] dataset, BEDLAM-Z, which contains a wide range of focal lengths because the camera is zooming from 28 to 80mm. This provides a more varied distribution over the intrinsics in the test set.

4.2. CameraHMR

For training CameraHMR along with the enhanced 4DHumans dataset (minus AVA) we also use the synthetic datasets AGORA [35] and BEDLAM [5], which contain accurate ground truth camera information. The combination of all these datasets is called "All" in Table 4. For evaluation we use the 3DPW [41] EMDB [23] and RICH [16] datasets. Additionally, we evaluate accuracy on the SPEC test set [27], which features multiple off-center individuals and more varied camera perspectives. To evaluate 3D shape accuracy, we utilize the SSP-3D [37] dataset. We also evaluate 2D alignment accuracy on the COCO validation set and perform qualitative evaluation using images from the LSP [19] and MPII [3] test sets in Fig. 4.

4.3. Evaluation Metrics

We follow previous work [13, 27] and evaluate 3D reconstruction accuracy using MPJPE (Mean Per Joint Position Error), PA-MPJPE (Procrustes Analysis Mean Per Joint Position Error), and PVE (Per Vertex Error), which measures the Euclidean distance (in mm) between predicted and actual 3D vertices and joints after aligning the pelvis. PVE is useful for evaluating body shape accuracy. We also evaluate the 2D alignment using PCK (Percentage of Correct Keypoints) on COCO-val [32]. PCK measures the accuracy of 2D keypoint detection by calculating the percentage of predicted keypoints within a specified distance threshold from the ground truth. We use thresholds of 0.05 and 0.1, which correspond to 5% and 10% of the crop size, respectively.

	Flickr-test	SPEC	3DPW	EMDB	BEDLAM-Z
Perspective Fields [18]	15.3	*8.0*	14.0	12.8	18.0
Ctrl-C [29]	26.5	10.4	*5.6*	*5.4*	31.3
WildCamera [48]	*10.9*	17.8	8.2	**2.3**	**4.8**
SPEC-camcalib [27]	14.0	14.3	8.8	5.9	14.2
HumanFoV (Ours)	**7.3**	**7.9**	**5.0**	5.7	*5.4*

Table 1. Vertical field of view error in degrees. BEDLAM-Z stands for the "zoom" sequence from BEDLAM used for testing (see text). Bold and italics correspond to best and second best respectively.

Method	PCK 0.05 ↑	PCK 0.1 ↑
CLIFF [31]	0.66	0.84
BEDLAM-CLIFF [5]	0.62	0.80
HMR2.0a [13]	0.79	0.94
HMR2.0b [13]	**0.86**	**0.96**
TokenHMR [10]	0.80	*0.95*
ReFit [42]	0.74	0.84
CameraHMR (Ours)	*0.84*	0.94

Table 2. PCK on COCO-val dataset measures 2D alignment accuracy. Bold: most accurate. Italics: second most.

5. Experiments

5.1. Comparision to SOTA

HumanFoV. As shown in Table 1, HumanFoV generalizes well across all benchmarks with diverse fields of view. While WildCamera [48] excels on benchmarks with narrow to average field of view, it underperforms on those with wide field of view, such as SPEC. In contrast, our HumanFoV maintains consistent accuracy across all benchmarks.

CameraHMR. As shown in Table 4, CameraHMR outperforms the baselines on all three benchmarks by a large margin. For a fair comparison, we categorize the methods based on the major datasets that were used in training. STD refers to standard datasets comprising Human3.6M [17], COCO [32], MPII [3] and MPI-INF-3DHP [34] while 4DHumans comprises InstaVariety [22], COCO [32], MPII [3], AI Challenger [43] and AVA [14]. Note that CameraHMR training never uses AVA. Even when trained on similar datasets, CameraHMR consistently outperforms all other baselines, demonstrating notable improvements particularly on the EMDB and SPEC-SYN benchmarks, which feature a wide variety of cameras. Additionally, our estimated dense keypoints improve the pGT body shape accuracy and this translates into improved accuracy of CameraHMR on the SSP-3D [37] dataset; see Sup. Mat.

Table 2 also shows that CameraHMR is either comparable to, or better than, other baselines in terms of 2D alignment on the COCO-val dataset. CameraHMR is nearly identical to HMR2.0b in terms of 2D keypoint alignment, while having significantly better 3D accuracy. As shown qualitatively in Fig. 4, CameraHMR achieves not only better 3D reconstruction, but significantly better 2D alignment, especially in cases of foreshortening or for people with non-average body shapes.

Camera	EMDB [23]↓	RICH [16]↓	SPEC-SYN [27]↓
fixed	89.3	64.8	138.7
default	82.7	64.9	115.2
predicted	**81.7**	**64.4**	**72.9**

Table 3. Per-vertex error (PVE) in mm for different focal length used during inference of CameraHMR.

5.2. Ablation

To understand the effect of using the predicted camera intrinsics from our HumanFoV, we perform an ablation study in which we vary the focal length used during inference. We use a fixed focal length of 5000 pixels, the predicted focal length from HumanFoV, and the default focal length that is use by other HPS methods [25, 31], defined as $\sqrt{w^2 + h^2}$ where w and h are the width and height of the full image respectively. We evaluate per-vertex error on three different benchmarks: EMDB, RICH, and SPEC-SYN. As shown in Table 3, the impact of using the predicted focal length is modest on EMDB and RICH. However, there is a significant improvement on the SPEC-SYN benchmark. EMDB and RICH largely contain centered individuals with fixed camera intrinsics, resulting in similar performance whether using predicted or default focal lengths. In contrast, SPEC-SYN includes varied camera intrinsics and off-center subjects, resulting in foreshortening and perspective distortion. In such cases, the benefits of using predicted focal length over the default focal length is significant.

6. Conclusion

In this work, we address the limitations of using an incorrect camera model in 3D human pose and shape estimation. By developing HumanFoV, a robust FoV predictor trained on a diverse human-centric dataset, we significantly enhance the accuracy of 3D human pose and shape estimation. Our integration of a full perspective model and dense surface keypoints into the SMPLify process improves the quality of pseudo ground truth data for in-the-wild images. Incorporating these advancements into the training of CameraHMR results in state-of-the-art performance on various benchmarks, demonstrating the effectiveness of our approach in improving both 2D alignment and 3D reconstruction.

Disclosure. MJB has received research gift funds from Adobe, Intel, Nvidia, Meta/Facebook, and Amazon. MJB has financial interests in Amazon and Meshcapade GmbH. While MJB is a co-founder and Chief Scientist at Meshcapade, his research in this project was performed solely at, and funded solely by, the Max Planck Society.

	Method	3DPW [41]			EMDB [23]			SPEC-SYN [27]		
		PA-MPJPE ↓	MPJPE ↓	PVE↓	PA-MPJPE↓	MPJPE↓	PVE↓	PA-MPJPE ↓	MPJPE↓	PVE↓
STD	SPEC [27]	53.2	96.5	118.5	87.7	138.9	161.3	56.9	83.5	98.9
	CLIFF* [31]	43.0	69.0	81.2	68.3	103.5	123.7	55.8	128.5	139.0
	HMR2.0a* [13]	44.4	69.8	82.2	61.5	97.8	120.0	55.8	133.3	153.0
BEDLAM	TokenHMR [10]	43.8	70.5	86.0	49.8	88.1	104.2	51.8	110.5	127.6
	WHAM†* [38]	35.9	57.8	68.7	50.4	79.7	94.4	-	-	-
	ReFit* [42]	38.2	57.6	67.6	55.5	91.7	106.2	51.3	103.6	116.3
	BEDLAM-CLIFF [5]	46.6	72.0	85.0	61.3	97.1	113.2	55.6	109.9	124.6
	CameraHMR (Ours)	40.0	62.3	74.8	45.4	82.7	97.0	**31.8**	**58.9**	**70.0**
4DH	HMR2.0b [13]	54.3	81.3	93.1	79.2	118.5	140.6	67.6	150.7	172.9
	CameraHMR (Ours)	38.7	62.7	73.4	43.9	73.2	85.6	37.0	66.0	79.1
All	CameraHMR (Ours)	38.5	62.1	72.9	43.7	73.0	85.4	33.0	61.8	73.1
	CameraHMR *(Ours)	**35.1**	**56.0**	**65.9**	**43.3**	**70.3**	**81.7**	32.9	61.8	72.9

Table 4. **Reconstruction error comparison for HPS.** *denotes method is finetuned on 3DPW training data. †denotes video based method. Datasets used in training; STD: standard datasets, 4DH: 4DHumans dataset, All: BEDLAM + 4DHumans dataset.

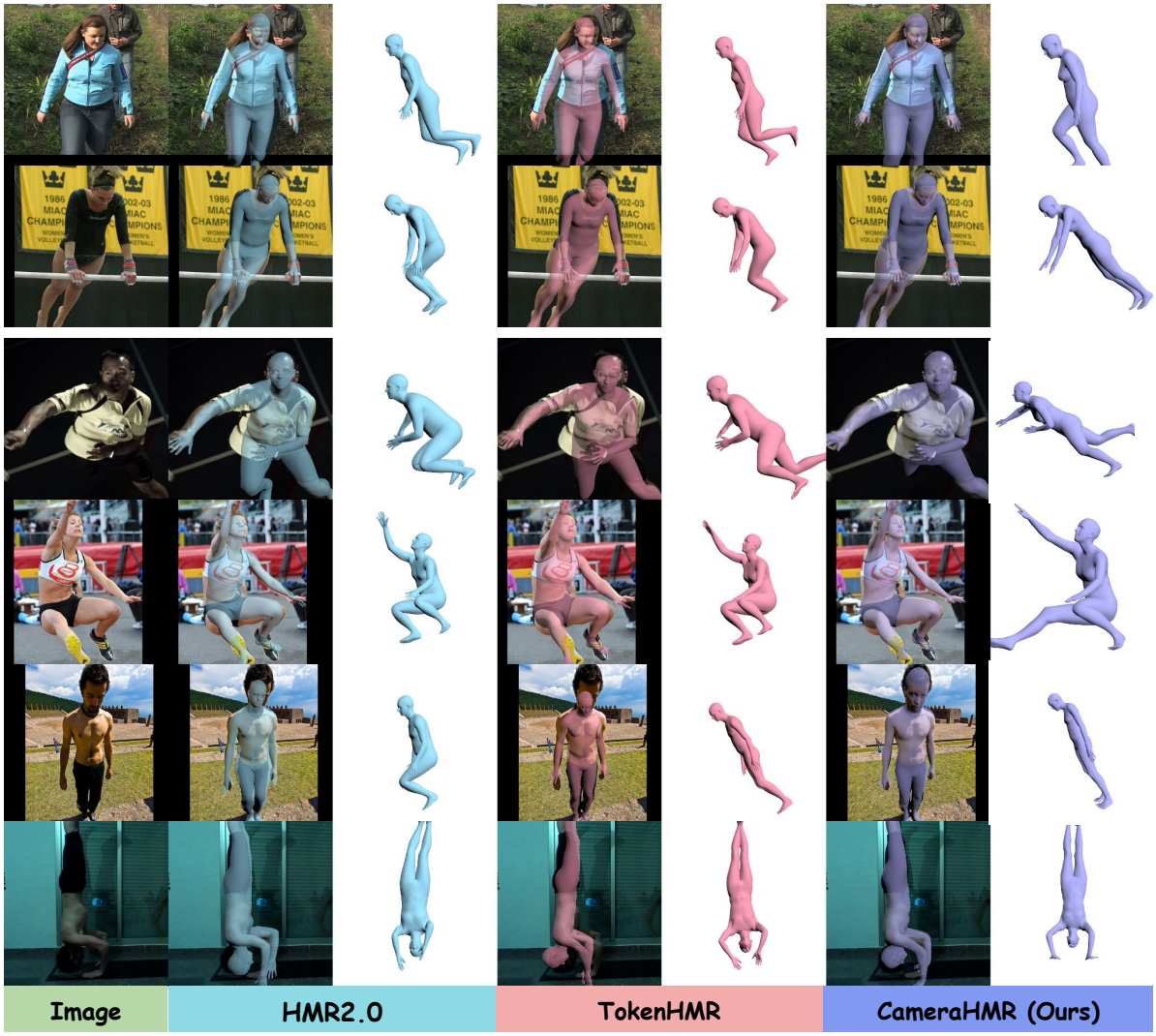

Figure 4. **Qualitative results of different baselines on LSP [19] and MPII [3] test images**. CameraHMR achieves better 3D pose and shape reconstruction while also achieving more accurate 2D alignment compared to other SOTA methods trained on comparable datasets.

References

[1] Flickr. https://www.flickr.com, 2024.

[2] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7822–7831, 2021.

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[4] Dragomir Anguelov, Carole Dulong, Daniel Filip, Christian Frueh, Stéphane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent, and Josh Weaver. Google street view: Capturing the world at street level. pages 32–38, 2010.

[5] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, 2023.

[6] 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 *European Conference on Computer Vision (ECCV)*, 2016.

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

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[10] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[11] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, 2021.

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, pages 1231 – 1237, 2013.

[13] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023.

[14] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6047–6056, 2018.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[16] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[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. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2014.

[18] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F. Fouhey. Perspective fields for single image camera calibration. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[19] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[20] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In *International Conference on 3D Vision (3DV)*, 2020.

[21] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018.

[22] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5614–5623, 2019.

[23] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *International Conference on Computer Vision (ICCV)*, 2023.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[25] Imry Kissos, Lior Fritz, Matan Goldman, Omer Meir, Eduard Oks, and Mark Kliger. Beyond weak perspective for monocular 3d human pose estimation. In *European Conference on Computer Vision Workshops (ECCV-W)*, 2020.

[26] 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 International Conference on Computer Vision (ICCV)*, 2021.

[27] 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 *International Conference on Computer Vision (ICCV)*. IEEE, 2021.

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

[29] Jinwoo Lee, Hyunsung Go, Hyunjoon Lee, Sunghyun Cho, Min-Hyuk Sung, and Junho Kim. CTRL-C: camera calibration transformer with line-classification. In *International Conference on Computer Vision (ICCV)*, 2021.

[30] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3383–3393, 2021.

[31] 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*, 2022.

[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 *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *Transactions on Graphics (TOG)*, 34 (6):248:1–248:16, 2015.

[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. 2017.

[35] Priyanka Patel, Chun-Hao Paul Huang, Joachim Tesch, David Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, 2021.

[36] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European conference on computer vision (ECCV)*, pages 704–720, 2018.

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

[38] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[39] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, pages 746–760, 2012.

[40] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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

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

[43] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017.

[44] Jianxiong Xiao, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[45] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. SUN3D: A database of big spaces reconstructed using sfm and object labels. In *International Conference on Computer Vision (ICCV)*, pages 1625–1632, 2013.

[46] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022.

[47] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[48] Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera: In-the-wild monocular camera calibration. *Advances in Neural Information Processing Systems*, 36, 2024.