

CameraHMR: Aligning People with Perspective

Supplementary Material

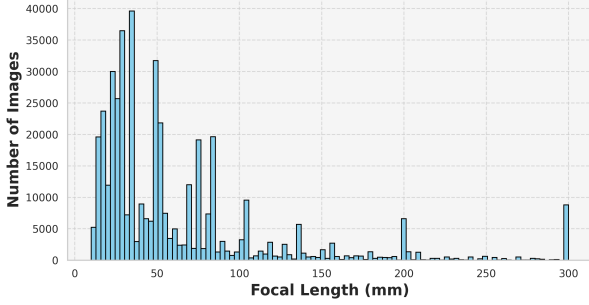


Figure 1. Focal Length distribution of images used in training HumanFoV.

1. Focal length distribution

We plot the distribution of focal lengths used in training HumanFoV model in Figure 1. The distribution shows notable peaks corresponding to the focal lengths of lenses most frequently used in photography, e.g. 24, 28, 35, 50, 85, 105, 135, 200, 300,

2. CameraHMR

2.1. Losses

We use several loss functions to ensure accurate 3D human pose and shape estimation. We minimize the L2 norm distance between the ground truth 3D joint locations $\hat{\mathbf{J}}_{3d} \in \mathbb{R}^{44 \times 3}$ and the predicted 3D joint locations $\mathbf{J}_{3d} \in \mathbb{R}^{44 \times 3}$ centered around pelvis joint using $\mathcal{L}_{J_{3d}}$.

$$\mathcal{L}_{J_{3d}} = \|\hat{\mathbf{J}}_{3d} - \mathbf{J}_{3d}\|_2^2 \quad (1)$$

We also penalize the deviation of the ground truth SMPL shape and pose parameters $\hat{\beta}$ and $\hat{\theta}$ from the predicted parameters β and θ , respectively using $\mathcal{L}_{\text{SMPL}}$. The network predict θ in a 6D rotation format [12], which is later converted to a rotation matrix representation to facilitate loss calculation. This avoids the discontinuities and ambiguities associated with angle-based representations, such as Euler angles or quaternions.

$$\mathcal{L}_{\text{smpl}} = \|\hat{\beta} - \beta\|_2^2 + \|\hat{\theta} - \theta\|_2^2 \quad (2)$$

Since we have accurate body shape labels we also minimize the difference between the ground truth 3D body mesh vertices $\hat{\mathbf{V}}_{3d} \in \mathbb{R}^{6890 \times 3}$ and the predicted vertices $\mathbf{V}_{3d} \in \mathbb{R}^{6890 \times 3}$, using $\mathcal{L}_{v_{3d}}$

$$\mathcal{L}_{v_{3d}} = \|\hat{\mathbf{V}}_{3d} - \mathbf{V}_{3d}\|_2^2 \quad (3)$$

For the 2D keypoint loss, we project \mathbf{J}_{3d} and $\hat{\mathbf{J}}_{3d}$ onto the 2D image plane using the camera intrinsic matrix \mathbf{K} . For

BEDLAM [2] and AGORA, the ground truth intrinsic matrix \mathbf{K} is provided. For 4DHumans dataset, where ground truth camera parameters are unavailable, we estimate \mathbf{K} using our HumanFoV model. The 2D projection \mathbf{J}_{2d} of the 3D joints is computed as $\Pi(\mathbf{J}_{3d}; \mathbf{K})$ where $\Pi(\cdot)$ denotes the projection using the intrinsic matrix \mathbf{K} .

Given that J_{2d} represents points in the original image coordinates where the horizontal and vertical coordinates x and y satisfy $x \in [0, W]$ and $y \in [0, H]$ respectively, we need to normalize these coordinates before calculating the loss. Therefore, we first transform the point from the full image coordinates to the cropped coordinates, based on the center and scale of the bounding box. The cropped image coordinates are then resized to a fixed resolution and normalize between -1 to 1. Finally, the 2D keypoint loss, $\mathcal{L}_{j_{2d}}$, is computed based on the normalized 2D keypoints, $\mathbf{J}_{2d}^{\text{norm}}$.

$$\mathcal{L}_{j_{2d}} = \|\hat{\mathbf{J}}_{2d}^{\text{norm}} - \mathbf{J}_{2d}^{\text{norm}}\|_2^2 \quad (4)$$

3. CamSMPLify

Here we provide more details about the optimization procedure used for generating our pseudo ground truth data for 4DHumans dataset. As describe in Eq. ?? from the main paper, we minimize the energy term $E(\beta, \theta, t^{\text{full}})$ by optimizing for SMPL shape β and pose θ as well as the camera translation t^{full} in 2 steps.

Initially, we optimize for n iterations focusing solely on the parameters β (shape), θ_0 (global orientation), and t^{full} (translation), while excluding θ (pose parameters) from the optimization. During this stage, we set the pose prior weight λ_{int} to 0, as we are not yet optimizing the pose. This strategy prevents the model from distorting the body pose excessively to match the keypoints, ensuring that any discrepancies due to incorrect orientation, shape, or translation are addressed first. By refining these parameters initially, we avoid overcompensating for errors related to the pose. After this stage, we update V_{int} with the output vertices from this optimization stage. In the subsequent stage, we optimize all the parameters, including θ , and set λ_{int} to 1.0 to obtain our final pose, shape and camera translation. For more details on the hyperparameter settings, please refer to the code.

4. Shape Evaluation

Most HPS evaluation benchmarks primarily represent average body shapes and offer limited shape diversity, which restricts their effectiveness in assessing improvements in shape accuracy. To address this, we utilize the SSP-3D [8] dataset, which includes a broad spectrum of body shapes.

| Method | Model | PVE-T-SC ↓ |
|---------------------|--------|-------------|
| HMR [5] | SMPL | 22.9 |
| SPIN [6] | SMPL | 22.2 |
| SHAPY [3] | SMPL-X | 19.2 |
| STRAPS [8] | SMPL | 15.9 |
| Sengupta et. al [9] | SMPL | 13.6 |
| CLIFF [7] | SMPL | 18.4 |
| BEDLAM-CLIFF [2] | SMPL-X | 14.2 |
| CameraHMR (B) | SMPL | 13.3 |
| CameraHMR (B+4DH) | SMPL | 11.6 |

Table 1. Shape error evaluation on SSP-3D dataset. B means trained on BEDLAM and AGORA and 4DH means trained on 4DHumans.

SSP-3D contains 311 real-world images of 62 individuals in fitted clothing, along with estimated ground-truth body shape data.

We evaluate shape accuracy using the PVE-T-SC metric on the SSP-3D dataset. PVE-T-SC, or Per-Vertex Error in T-pose after Scale Correction, calculates the per vertex average error by comparing a reconstructed 3D body mesh in a standardized T-pose to the ground truth. Before computing this error, the scale of the predicted model is adjusted to match the ground truth, ensuring that the metric reflects inaccuracies in shape and pose, rather than differences in overall scale.

As demonstrated in Table 1, CameraHMR outperforms all other benchmarks in terms of shape accuracy, even surpassing methods specifically trained to enhance shape prediction. Additionally, incorporating our improved 4DHumans pGT into the training process, along with BEDLAM, further improves shape accuracy, highlighting the high quality of the shape information in our pseudo ground truth.

5. More Qualitative Results

In Fig. 2 and Fig. 3, we present qualitative results of CameraHMR applied to images downloaded from Pexels [1]. The results for multi-person images are obtained by first generating the bounding box for each person using Detectron2 [11] on the full image. These cropped bounding boxes are then fed into CameraHMR. The results demonstrate that CameraHMR effectively estimates both accurate body poses and detailed body shapes, even for complex body poses and camera angles.

In Figure 4, we compare the results of CameraHMR with HMR2.0 [4] and ReFit [10] on images from Pexels. HMR2.0 employs a weak perspective camera model, while ReFit, like CameraHMR, uses a full perspective camera model during training. Despite achieving good alignment on images with standard focal length, HMR2.0 often results in unrealistic body poses. 2D alignment with the image also gets worse as FoV of the image increases as shown in some of the images in Figure 4. ReFit [10], although producing a more accurate 3D pose, suffers from poor 2D alignment due

to reliance on default camera parameters during inference. In contrast, CameraHMR leverages robust camera intrinsics predicted by our HumanFoV model, resulting in accurate body poses and shapes as well as improved 2D alignment, even under extreme camera conditions.

References

- [1] Pexels. <https://www.pexels.com/>, 2024.
- [2] 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.
- [3] Vasileios Choutas, Lea Müller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. Accurate 3D body shape regression using metric and semantic attributes. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2718–2728, 2022.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3D human shape and pose estimation from images in the wild. In *International Conference on Computer Vision (ICCV)*, pages 11219–11229, 2021.
- [10] Yufu Wang and Kostas Daniilidis. Refit: Recurrent fitting network for 3d human recovery. In *International Conference on Computer Vision (ICCV)*, 2023.
- [11] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019.
- [12] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019.



Figure 2. CameraHMR results on landscape images downloaded from Pexels [1].



Figure 3. CameraHMR results on portrait images downloaded from Pexels [1].

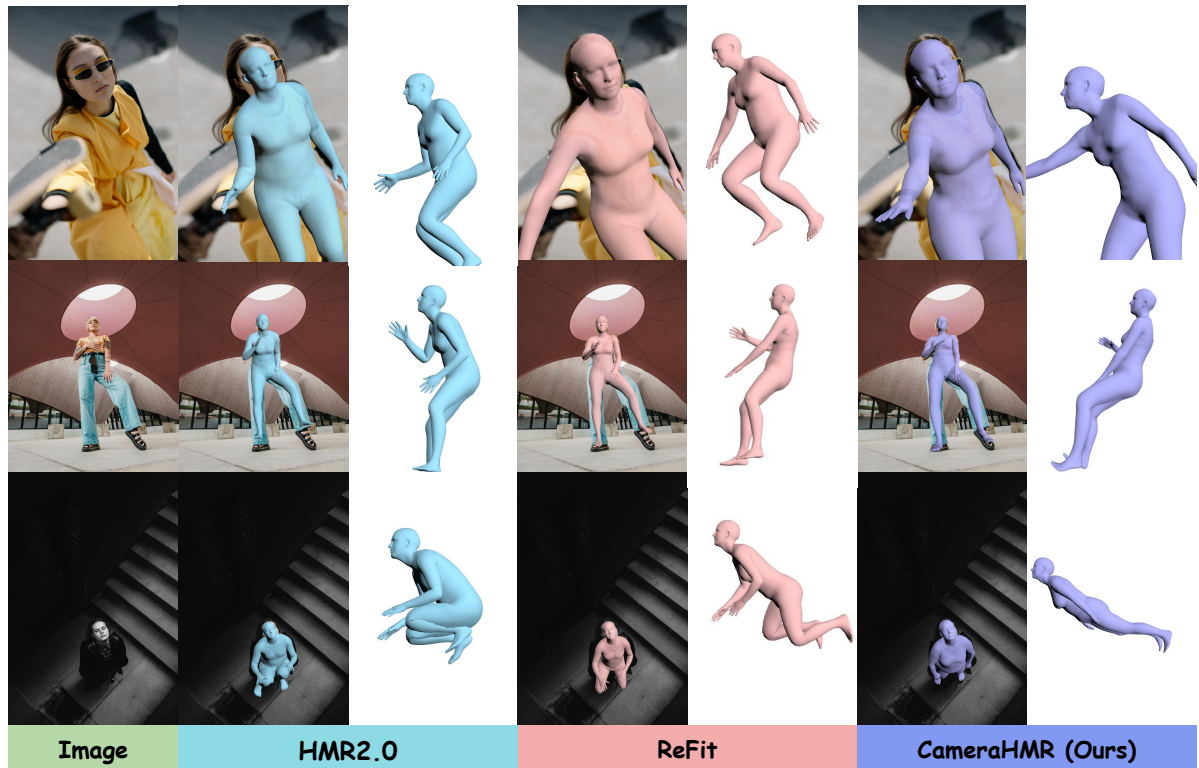


Figure 4. CameraHMR achieves more accurate 3D pose estimation, shape reconstruction, and 2D alignment with the image even for extreme camera angles, outperforming other methods in these challenging scenarios.