All-Age Human Mesh Recovery

Supplementary Material

7. Implementation details

7.1. Anny model mapping

We utilize the semantic shape space of the Anny model [7] to propose a direct mapping from shape attribute descriptors to normalized shape values. This mapping inherently accounts for the body model interpolation described in the same work. Figure 9 illustrates our complete mapping scheme for all experiments. To increase discriminative power, we supplement the age categories proposed in [52] by adding more resolution in the earlier ages where body shape changes rapidly. For gender, we set three equidistant anchors to better capture the continuous spectrum of visual genders.

7.2. Pre-processing feedforward initializations

Multi-HMR. As a bottom-up feed-forward model, Multi-HMR often struggles detecting certain people in the image. We particularly note how the non-maximum suppression process can remove highly occluded predictions. To recover these lost instances, we use the predicted nose keypoints from ViTPose to generate a mesh estimate for them. We perform a mutual best assignment matching between nose keypoints and head detections. If a nose keypoint lacks a corresponding head prediction, we select the closest patch available. Overall, this strategy improves the recall on the test set from 74.0% to 83.0%. We note that when the initial head prediction is missing, the mesh estimate can be highly inaccurate, forcing the final optimization to rely heavily on other cues.

From SMPL to Anny fits. Given a SMPL(X) mesh, we first convert it to the Anny topology using a vertex regressor [7]. We then estimate the corresponding Anny pose and shape parameters using a fast fitting algorithm similar to NLF [45]. The method alternates for 5 iterations between two steps: we independently fit the global orientation of each body part via a weighted Kabsch algorithm using Anny's blend skinning weights, and then solve for the shape parameters with linear least squares. Finally, we optionally refine the part rotations in a single forward pass along the kinematic tree. The procedure is highly efficient, supports batch GPU execution, and processes 128 meshes in about 100 ms.

7.3. Vision-Language Model prompting

Prompts. For estimating the shape attributes with a VLM we use the names of the anchors described in 7.1, below we show the prompts:

Gender estimation prompt

```
"Look at the cropped image of
a person and estimate their
gender.  Choose exactly one of
the following categories:  male,
female, unknown.  Respond with
only the category name.  For
example, if the person is female,
respond with 'female'."
```

Age estimation prompt

```
"Look at the cropped image of
a person and estimate their
age group.  Choose one of the
following categories according
to the age in years:  baby (0
to 1), toddler (2 to 3), kid
(4 to 8), teen (8 to 16), adult
(16+).  Respond with only the
category name.  For example, if
the person is a toddler, respond
with 'toddler'."
```

Table 5. **Prompting strategies.** Results of varying prompting strategies for the VLM model (Qwen 2.5 VL) on the Relative Human validation 'has child' subset. indicates the default setting.

Method	Age F1					Gender
	overall	adult	teen	kid	baby	F1
grounded	40.42	21.96	45.83	52.53	41.35	70.38
overlay	57.41	82.56	38.46	64.47	44.16	89.85
person crop	66.69	85.77	50.63	70.25	60.12	91.22
head crop	67.15	85.98	51.57	70.43	60.61	91.22

Ablation on prompting strategies. Given our multi-person image setting, we studied the effect of varying the prompting strategy to the VLM when specifying individual subjects. Table 5 shows the performance of different methods for utilizing the person's bounding box. The least effective strategy is the grounded approach, which provides the bounding box coordinates directly in the text prompt. Performance improves when we visually overlay the box in red directly onto the image. The most effective approaches involve spatially focusing the VLM, specifically by cropping the person's full body or cropping only their head. The head crop proved particularly useful for images featuring crowds

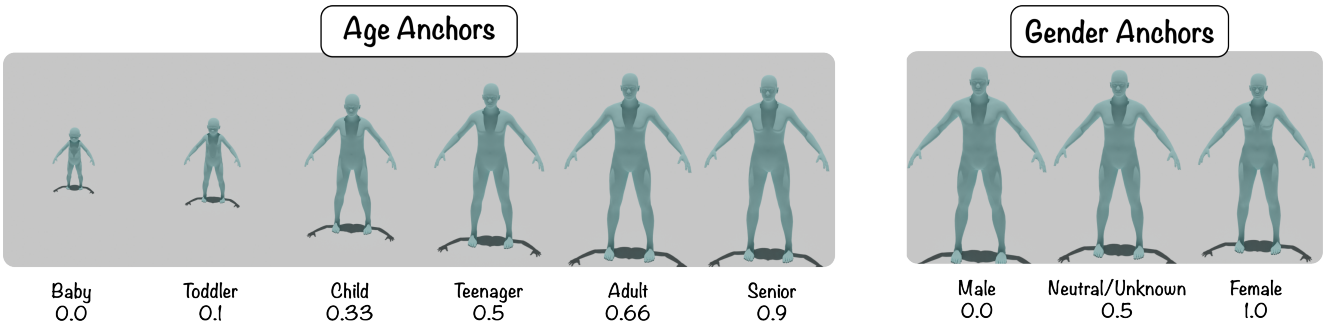

Figure 9. **Shape parameter anchors.** Examples of the text descriptors and Anny shape parameter values for each anchor. Left: Age mapping. Right: Gender mapping. Other shape parameters are set to 0.5 here for visualization. (Age set to 0.66 for gender).

or significant occlusions. We hypothesize that combining both the full body crop and the head crop would likely overcome the limitations of each approach, but at an increase in computational cost.

7.4. Optimization parameters

For our experiments we set all parameter learning rates to 0.01 except for the full body rotation θ and shape (β) which are set to 0.001. Across all stages $\lambda_{shape} = 10.0, \lambda_{2D} = 0.01$ and $\lambda_{dense} = 0.001$ with $\sigma = 100$. The first stage where we update only τ runs for 50 iterations for Multi-HMR and 200 iterations for CameraHMR with $\lambda_{init} = 10.0, \lambda_{depth} = 10.0$. The second stage, that updates $\{\tau, \phi, \beta\}$, runs for 100 iterations for both initializations, with $\lambda_{depth} = 50.0$, and we split λ_{init} into the more granular $\lambda_{init_\beta} = \lambda_{init_{verts}} = 0.01, \lambda_{init_\phi} = 10.0$. The final stage (where all parameters are optimized) runs for 200 iterations with $\lambda_{depth} = 0.0$, $\lambda_{init_\phi} = \lambda_{init_{verts}} = 0.01, \lambda_{init_\beta} = 5.0$.

7.5.

8. Experiment details

8.1. Metrics

Percentage of Correct Keypoints (PCK). To robustly account for missing keypoint detections, prior work often assigns a fixed "punishment value" to unmatched predictions when computing the MPJPE. However, such heuristics distort the numerical scale of the evaluation and can introduce undesirable incentives—e.g., a missed detection may be penalized less than an inaccurate prediction. For this reason, we complement MPJPE with the PCK metric, which measures the proportion of predicted keypoints falling within a predefined distance threshold from the ground truth. This formulation naturally handles both poor localization and missing detections without relying on arbitrary penalty terms, yielding a more interpretable and principled performance measure. Note that we follow the same evaluation procedure on Relative Human, but extend prior implementations to account for non-detected individuals.

8.2. Relative Human

Ablation 'has child' subset. For the ablation experiments we report results on the validation 'has child' subset, composed by all the images that have at least one non-adult example. Figure 10 show the resulting class distribution, where the 'has child' validation subset balances corrects the over-representation of adults.

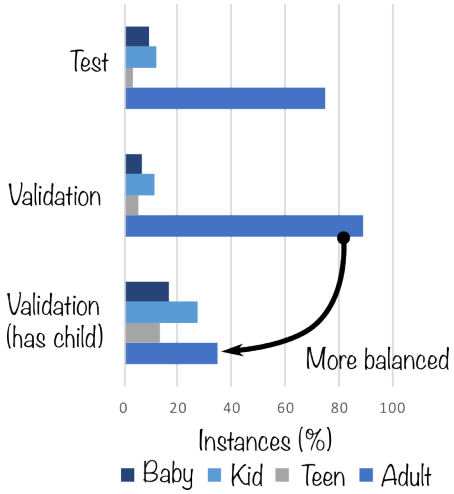

Figure 10. Age distribution across different subsets of the Relative Human dataset. We use the 'has child' subset for our ablation experiments.

Impact of detection. As an additional experiment we run Anny-Fit on the test set using ground-truth bounding boxes. Table 6 shows how the improvements from Anny-Fit are consistent when having ground-truth bounding boxes. We find that the ground-truth bounding boxes are helpful to recover occluded, back-facing or far away people in the image

Additional results. Anny-Fit also improves on the scene-aware method PromptHMR [56]. We use as prompts bound-

Table 6. **Reconstruction with ground-truth bounding boxes.** Δ: improvement over initialization. †: ground-truth boxes used for inference.

Method	2D (↑)	PCRD$^{0.2}$ (↑)					Age (↑)	Gender (↑)
	$mPCKh^{0.6}$	overall	adult	teen	kid	baby	F1	F1
Multi-HMR [7]†	65.43	60.74	61.10	68.36	57.67	44.99	24.01	35.40
+ **Ours**	79.37	67.78	68.20	71.80	67.04	57.55	49.00	84.45
Δ	+13.94	+7.04	+7.1	+3.44	+9.37	+12.56	+24.99	+45.05
CameraHMR [38]†	64.47	59.7	59.86	67.49	47.21	31.16	0.00	0.00
+ **Ours**	82.02	66.32	66.45	72.96	61.94	53.20	44.08	83.34
Δ	+17.55	+6.62	+6.59	+5.47	+14.73	+22.04	+44.08	+83.34

Table 7. **Additional Anny-Fit reconstruction results on Relative Human test**.

Method	2D (↑)	PCRD$^{0.2}$ (↑)					Age (↑)	Gender (↑)
	$mPCKh^{0.6}$	overall	adult	teen	kid	baby	F1	F1
PromptHMR [56]	67.70	60.15	60.80	61.90	47.48	29.06	0	0
+ **Ours**	81.84	67.88	68.61	70.08	66.56	52.88	48.73	81.72
Δ	+14.14	+7.73	+7.81	+ 8.18	+19.08	+23.82	+48.73	+81.72

ing boxes, segmentation mask, and find no change when including text descriptions, likely because age modeling is not part of the text training corpus of PromptHMR.

Additional qualitative results. Figure 15 shows additional qualitative results on the Relative Human dataset.

8.3. CMU Toddler

To the best of our knowledge, this is the only dataset from real captures with 3D ground-truth annotations for non-adults. We select the 5 sequences from the CMU Panoptic dataset [21] that include a non-adult, specifically the sequences follow a toddler and 1 to 3 adults interacting, sometimes with objects. Following standard HMR evaluations on the 4 adult sequences [52, 62], we evaluate all frames at 1 FPS, on 2 selected cameras and on the valid 3D joints (those in frame). For most sequences, the toddler is out of frame for a large part of the videos because of their height. As such, we select the following cameras to ensure the subject is in frame: Ian 1: $\{16, 21\}$, Ian 2 and Ian 5: $\{11, 21\}$, Ian 3: $\{11, 23\}$, Toddler 5: $\{15, 23\}$.

Figure 11 shows examples of Anny-Fit applied to CMU Panoptic toddler sequences, using both initializations. Our observations are consistent with Table 1b: Multi-HMR shows strong baseline performance, where our method refines relative positioning and pose. Conversely, for CameraHMR, the primary gain lies in relative positioning and shape.

8.4. Training with pseudo-ground truth fits

Fitting on MS-COCO. We create pseudo- ground-truth fits for 30537 images in the MS-COCO train dataset using the ground-truth bounding boxes. Before training, we filter images with unsuccessful fits using the proxy of the 2D reprojection loss for all people. We select the images within the 3rd and 95th error percentiles, leaving 28039 after filtering. Figure 12 shows examples of the error percentiles. The lowest percentiles are likely due to not enough estimated keypoints, while the highest percentiles point to highly cropped people.

Gender prediction. Figure 13 shows the effect of retraining with our fits. Similarly to Figure 7, the improved performance exemplifies how Anny-Fit can effectively distill information onto a retrained model.

9. Beyond all-age shape estimation

While not our primary objective, our shape estimation formulation generalizes to account for attributes beyond age and gender. As a case study, we explore diverse body shapes. To manage a diverse range of shapes with a compact categorization for VLM querying, we define a discrete mapping to weight and muscle attributes using the following anchors: $slim = \{muscle : 0.3, weight : 0.3\}$, $average = \{muscle : 0.4, weight : 0.7\}$, $overweight = \{muscle : 0.1, weight : 0.8\}$, and $muscular = \{muscle : 0.9, weight : 0.7\}$. Figure 14 presents preliminary examples of how mapping the muscle and weight dimensions of the Anny body model improves shape estimation on in-the-wild images (sourced from Pexels [41]). These results suggest that Anny-Fit can be extended to align 3D reconstructions with other shape estimation settings.

10. Limitations

Our method integrates multiple expert predictions to guide the optimization for all-age human reconstruction, which enhances overall robustness. While this multi-expert strategy improves robustness, it also makes performance dependent on the accuracy of each expert. Errors in keypoints or depth under occlusion or extreme viewpoints can propagate through the optimization. Furthermore, age and gender estimates rely heavily on facial cues, making them unreliable for back-facing, occluded, or low-resolution subjects; inaccurate estimates can consequently bias the inferred shape. While the scarcity of child datasets (due to privacy and ethical constraints) limits comprehensive evaluation for younger age groups. Finally, while our method jointly optimizes all visible individuals in multi-person scenes, it does not explicitly handle interpenetration. Accounting for this could provide a valuable physical prior for the optimization and help prevent implausible overlaps between people.

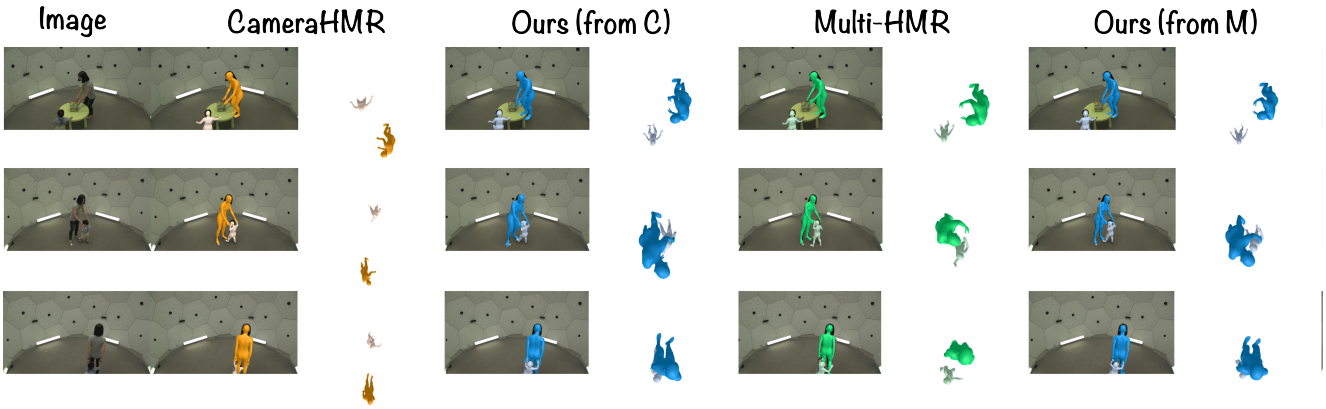

Figure 11. **Results on CMU toddler.** For each method we show the camera and top views.

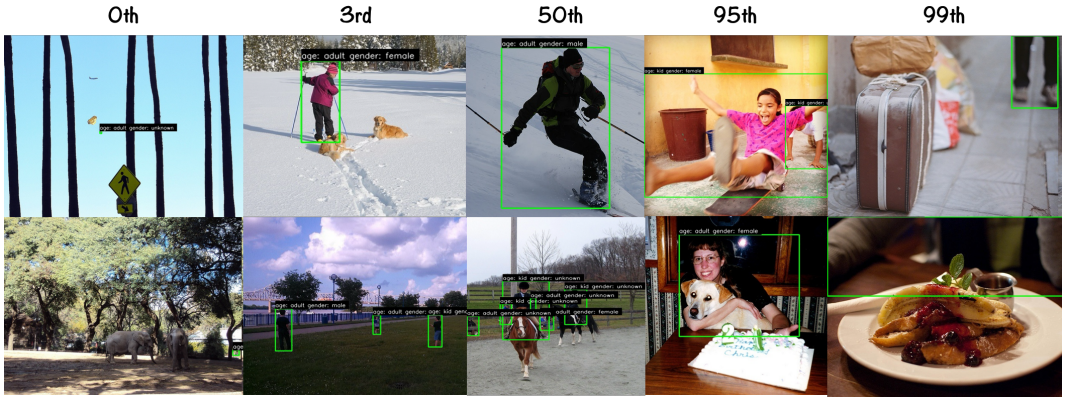

Figure 12. **Visualization of 2D Keypoint reprojection error percentiles.** We filter out fits that are not within the 3rd and 95th percentiles. We overlay the ground-truth bounding boxes and estimated age and gender.

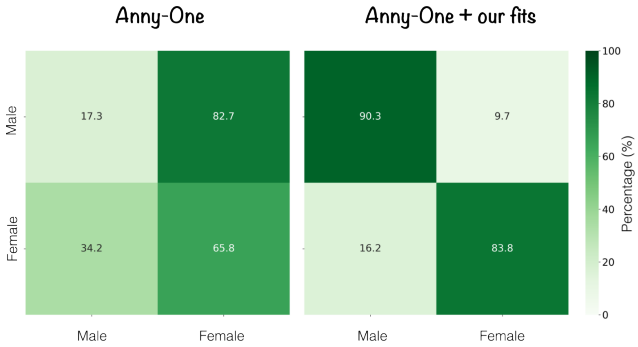

Figure 13. **Gender confusion matrix on Relative Human test.** Note how after retraining with our Anny-Fit fits, the model can accurately predict gender in an unseen dataset.

Figure 14. Optimization results from Anny-Fit demonstrating estimation of diverse weight and muscle shape attributes. Top: 'muscular' Mid: 'overweight'; and Bottom: 'overweight' and 'slim'.

Figure 15. Additional examples of reconstructions with Anny-Fit (Ours) compared to SOTA on the Relative Human dataset.