arXiv:2410.11520v1 [cs.CV] 15 Oct 2024

Look Ma, no markers: holistic performance capture without the hassle

CHARLIE HEWITT, FATEMEH SALEH, SADEGH ALIAKBARIAN, LOHIT PETIKAM, SHIDEH REZAEIFAR, LOUIS FLORENTIN, ZAFIIRAH HOSENIE, THOMAS J. CASHMAN, JULIEN VALENTIN, DARREN COSKER*, and TADAS BALTRUŠAITIS, Microsoft, United Kingdom, Switzerland and France



Fig. 1. Results of our holistic performance capture pipeline for the body, face, and hands. We fit our parametric model to image data using neural networks for dense landmark, pose and shape prediction followed by optimization, providing robust *and* accurate reconstructions. Our approach supports single- or multiple-camera setups without requiring calibration or specialist hardware and tracks the whole body including the eyes and tongue.

We tackle the problem of highly-accurate, holistic performance capture for the face, body and hands simultaneously. Motion-capture technologies used in film and game production typically focus only on face, body or hand capture independently, involve complex and expensive hardware and a high degree of manual intervention from skilled operators. While machine-learning-based approaches exist to overcome these problems, they usually only support a single camera, often operate on a single part of the body, do not produce precise world-space results, and rarely generalize outside specific contexts. In this work, we introduce the first technique for marker-free, high-quality reconstruction of the complete human body, including eyes and tongue, without requiring any calibration, manual intervention or custom hardware. Our approach produces stable world-space results from arbitrary camera rigs as well as supporting varied capture environments and clothing. We achieve this through a hybrid approach that leverages machine learning models trained exclusively on synthetic data and powerful parametric models of human shape and motion. We evaluate our method on a number of body, face and hand reconstruction benchmarks and demonstrate state-of-the-art results that generalize on diverse datasets.

CCS Concepts: • **Computing methodologies** → **Computer vision**; **Motion capture**; *Shape modeling*; Rendering.

Additional Key Words and Phrases: 3D reconstruction, body pose

*Also with University of Bath.

Authors' address: Charlie Hewitt, chewitt@microsoft.com; Fatemeh Saleh, fatemehsaleh@microsoft.com; Sadegh Aliakbarian, saliakbarian@microsoft.com; Lohit Petikam, lohitpetikam@microsoft.com; Shideh Rezaeifar, srezaeifar@microsoft.com; Louis Florentin, lflorentin@microsoft.com; Zafiirah Hosenie, zhosenie@microsoft.com; Thomas J. Cashman, tcashman@microsoft.com; Julien Valentin, juvalen@microsoft.com; Darren Cosker, coskerdarren@microsoft.com; Tadas Baltrušaitis, tabaltru@microsoft.com, Microsoft, United Kingdom, Switzerland and France.

1 INTRODUCTION

High-quality performance capture is a challenging problem, which traditional approaches simplify by introducing additional constraints. These take the form of expensive specialist hardware such as optical markers, precisely calibrated multi-camera rigs, or specific lighting setups. Additionally, the scope of capture is often limited; just the face or hands are targeted, further simplifying the problem. These methods achieve high-quality results, but only under certain circumstances and at significant cost. They usually require experts to configure and operate them, and the results need artist work to clean up or combine feeds from different capture technologies. In short, the traditional approach has exchanged rigidity for quality, and, in the process, has limited high-quality performance capture to the worlds of film, television and game production.

In this work, we tackle the problem of accurate, robust and adaptable performance capture of the whole human. Our approach eliminates the need for specialist hardware, supports any number of cameras without needing calibration, can achieve state-of-the-art results fully automatically, and significantly lowers the amount of expert manual labor required to reach production-grade performance capture results. We demonstrate holistic performance capture of the face, hands and body simultaneously in Figure 1.

Recent academic work relies heavily on ML-based approaches and Move AI [2024] introduced an ML capture solution for body pose aimed at production settings. These are less dependent on specialist hardware, and function from a single image taken with any camera [Goel et al. 2023; Wood et al. 2022; Zhang et al. 2023b]. Some also provide holistic capture of the body, hands and face [Pavlakos et al. 2019; Zhang et al. 2023b]. ML approaches which directly regress

body pose and shape can be highly robust, as a result of training on large and diverse datasets [Goel et al. 2023; Zhang et al. 2023b]. However, they are less precise than traditional approaches (e.g., markers or photogrammetry) and deep neural networks (DNNs) which predict image-space features such as 2D landmarks. They are also not easily adapted to arbitrary multi-camera rigs and are often unable to produce world-space results. Our method exploits the benefits of ML to provide robustness, but improves accuracy and adaptability by incorporating model fitting.

Many ML approaches share a challenge in ensuring the privacy and diversity of their training data. Human-annotated datasets are also limited in the fidelity of annotations that can be achieved and often include errors and noise. We choose instead to use *only* synthetic data to train the ML components of our method. This allows us to track very dense 2D landmarks, as well as predict precise model parameters without label noise. It also eliminates privacy concerns and gives us explicit control over the data in order to mitigate fairness disparities.

In summary, we combine the benefits of model fitting and ML-based methods to achieve: *(1) Robustness* of DNNs which directly regress body pose and shape. *(2) Accuracy* of DNNs predicting image-space features. *(3) Adaptability* of model fitting to support arbitrary cameras. A comparison of the capabilities of our method with other performance capture technologies is shown in Table 1. Specifically, in this work we present the following:

(1) A system providing state-of-the-art holistic 3D human reconstruction including body shape and pose, face shape and expression, hand and tongue articulation and eye gaze prediction from monocular images.
(2) Extension of this system to the multi-view scenario, providing improved accuracy and world-space 3D results. The system is adaptable to arbitrary camera rigs without explicit calibration and does not require any specialized hardware or manual intervention.
(3) A synthetic data pipeline for generating realistic images of humans with rich ground-truth annotations using a parametric model of the whole human with diverse face shape and tongue motion. We release the datasets of the body, face and hands which we use for training our method.

Our method shows that it is possible to achieve high-fidelity, holistic performance capture fully automatically from image inputs alone. This demonstrates that wider access to very high-quality performance capture is possible at lower cost. It can also accelerate the workflow for production performance capture significantly, as calibration is not required and manual work is reduced. More information and instructions for accessing the datasets can be found at https://aka.ms/SynthMoCap.

Table 1. Comparison of existing performance capture technologies with our proposed method.

System	Marker-free	Body Pose	Hand Pose	Facial Expression	Shape	Calibration Free	Commodity Hardware	Arbitrary Rig	World-space
Optical Markers (e.g., Vicon, Optitrack)	×	✓	✓*	✓†	×	×	×	✓	✓
Move AI [2024]	✓	✓	✓	×	×	×	✓	✓	✓
Faceware Technologies [2024]	✓	×	×	✓	×	×	×	×	×
Smart Suits (e.g., XSens, Manus)	✓	✓	✓§	×	×	×	×	-	×
SMPLify-X [Pavlakos et al. 2019]	✓	✓	✓	✓	✓	✓	✓	×	×
Wood et al. [2022]	✓	×	×	✓	✓	✓	✓	✓	✓
BEDLAM [Black et al. 2023]	✓	✓	×	×	✓	✓	✓	×	×
HMR2 [Goel et al. 2023]	✓	✓	×	×	×	✓	✓	×	×
PyMAF-X [Zhang et al. 2023b]	✓	✓	✓	✓	✓	✓	✓	×	×
TokenFace [Zhang et al. 2023a]	✓	×	×	✓	✓	✓	✓	×	×
TEMPO [Choudhury et al. 2023]	✓	✓	×	×	×	×	✓	✓	✓
WHAM [Shin et al. 2024]	✓	✓	×	×	×	✓	✓	×	✓
Ours	✓	✓	✓	✓	✓	✓	✓	✓	✓

* often requires extensive manual cleanup, † difficult to capture in the same shoot as body/hands, § typically requires manual alignment between body and hands.

2 RELATED WORK

2.1 Traditional Performance Capture

Most performance capture technologies used in production settings are able to produce high quality results due to additional constraints imposed on the problem; limiting to a single body part, using optical markers, requiring precise calibration or enrollment data. This usually increases cost and the amount of manual intervention required.

For example, Faceware Technologies [2024] consider only the face from a head-mounted camera and track a small number of landmarks, usually relying on manual annotation of a number of 'training frames' and actor-specific enrollment sequences. Optical marker systems (Vicon Motion Systems [2024], Optitrack [Natural Point 2024]) are popular for body-pose capture. These use specialist camera rigs that are expensive and require expert operators. The data often requires manual clean-up to resolve noisy, missing or swapped markers and to facilitate fitting a parametric model. XSens [Movella 2024], StretchSense [2024] and Manus-Meta [2024] use wearable 'smart suits' to determine motion. These rely heavily on enrollment sequences, can be expensive, and manual effort is often needed to combine the signals from different body parts.

Our system exploits advances in ML to move away from specialist hardware and allow anyone with a camera to achieve state-of-the-art performance capture. Moreover, our unified model allows for holistic capture of facial expression, body pose and hand pose in one, without manual work to combine separate streams.

2.2 Machine Learning Approaches

Recent academic work focuses on the application of ML techniques to solve the problem of performance capture. Pavlakos et al. [2019] use a DNN to regress skeletal 2D landmarks in an image and optimization to fit their parametric model to these observations. The sparse nature of the landmarks mean they are unable to recover accurate body shape and facial expression. Wood et al. [2022] employ dense 2D landmarks to achieve significantly higher-fidelity shape reconstruction and expression tracking, but only support the face and without tongue articulation. Zhang et al. [2023b] exploit image-space information to produce holistic reconstructions at high fidelity from an image, but do not support video or arbitrary camera rigs. Other ML approaches directly regress pose from monocular images, providing robust but less-precise results [Black et al. 2023; Goel

et al. 2023] or only targeting specific body parts [Zhang et al. 2023a]. Shin et al. [2024] also predict global trajectory and Choudhury et al. [2023] support arbitrary viewpoints, both providing world-space results but only reconstructing skeletal body pose. Move AI [2024] introduced an ML-based production performance capture system supporting arbitrary cameras. It uses skeletal landmark tracking, so is unable to recover facial expression or detailed body shape, and requires a calibration step. We find the combination of image-space features (dense landmarks) and direct pose regression lets us achieve accurate *and* robust results, while the combination of DNNs with conventional model-fitting makes our system more adaptable than pure ML approaches.

2.3 Models and Data

Parametric models. Many works consider different body parts in isolation, such as the face [Li et al. 2017; Wood et al. 2021], hand [Khamis et al. 2015; Romero et al. 2017] and body [Allen et al. 2003; Anguelov et al. 2005; Loper et al. 2015]. Some have combined hands and body into a single model, such as SMPL-H [Romero et al. 2017] and FrankMocap [Rong et al. 2021], with Adam [Joo et al. 2018] and SMPL-X [Pavlakos et al. 2019] also including the face. These are good statistical models of humans, but most do not lend themselves well to computer graphics applications for generating synthetic data. For example, FLAME [Li et al. 2017], used in SMPL-X [Pavlakos et al. 2019], has no teeth or tongue and the mesh does not subdivide well for close-up renders of the face. We build on the work of Romero et al. [2017] (SMPL-H) and Wood et al. [2021] to produce a parametric model of the full human including hand articulation and facial expression that is well suited to synthetic data generation. We also improve the shape model for the face and add tongue articulation.

Synthetic Data. Recent research has shown that synthetic data can be an effective means of achieving high-quality, robust reconstruction of humans. Wood et al. [2021] demonstrate the possibility of performing face-related computer vision in the wild using synthetic data alone. Similarly, Black et al. [2023] and Yang et al. [2023] introduce the BEDLAM and SynBody datasets respectively, both containing monocular RGB videos of synthetic humans using the SMPL-X model [Pavlakos et al. 2019]. They also demonstrate state-of-the-art accuracy on real-image benchmarks despite training with synthetic data alone, but their pipelines do not include high quality face shape and expression. We also choose to use synthetic data, given that it is now possible to achieve parity with real data in some tasks, it significantly reduces concerns around privacy, and can easily include complex ground-truth annotations without error. We extend the pipeline of Wood et al. [2021] using our parametric human model to be able to generate data including the full human, rather than just the face.

3 SYNTHETIC DATA

We exclusively use synthetic data to train the DNNs used in our performance capture system. The use of the same parametric model throughout the entire stack is critical to its success. As the 2D landmarks correspond directly to 3D points on the mesh generated by the model, we can fit to them directly. This eliminates the need for error-prone manual tuning of marker mappings [Loper et al. 2014] and prevents sliding of predicted landmarks relative to the mesh, as would be expected with DNNs trained on human-annotated landmark datasets or traditional contour tracking.

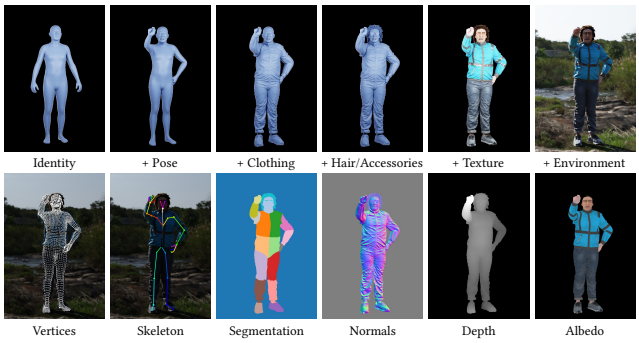


Fig. 2. Construction of synthetic image by sampling random identity and pose for the SOMA parametric model, clothing, hair and accessory assets, textures, shaders and HDRI environments from our asset library (top row). From this we can produce a highly realistic image and a number of corresponding ground-truth annotations (bottom row).

The pipeline starts with a parametric model which captures face and body shape, body and hand pose and facial expression, including articulation of the tongue. To construct this, we build on the popular SMPL-H body model [Romero et al. 2017] and the face model of Wood et al. [2021]. We then add texture, hair and clothing from a large and diverse asset library. Finally, we situate the body in an indoor 3D scene or an outdoor scene with environmental lighting from an HDR image [Debevec 2006]. We render the scene using the Cycles rendering engine [Blender Foundation 2021]. Figure 2 shows how we construct a render, as well as some example ground-truth annotations output by our pipeline. Examples of images rendered using our pipeline can be seen in Figure 3. All data originating from captures of real people used in the pipeline is obtained with explicit consent. Below we give a brief summary of the parametric model and data generation pipeline; for greater detail, please consult the supplementary material.

3.1 Parametric Human Model

We improve the face model of Wood et al. [2021] by refining and expanding the training data (as described below) and adding tongue articulation. We combine this updated face model with the SMPL-H body model [Romero et al. 2017] to produce a unified parametric model of the full human with a high degree of control and expressivity, we call this SOMA (from the Greek for 'body'). The SOMA mesh vertex positions are defined by mesh generating function $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \boldsymbol{\psi}, \boldsymbol{\theta})$ which takes parameters $\boldsymbol{\beta} \in \mathbb{R}^{300}$ for body shape, $\boldsymbol{\gamma} \in \mathbb{R}^{256}$ for face shape, $\boldsymbol{\zeta} \in \mathbb{R}^{9}$ for hand shape, $\boldsymbol{\psi} \in \mathbb{R}^{224}$ for expression, and $\boldsymbol{\theta} \in \mathbb{R}^{54\times3}$ for pose. A complete mathematical definition of the model can be found in the supplementary material.

Tongue articulation is achieved through addition of 12 blendshapes solely to control the movement of the tongue, covering a large proportion of possible articulations of the tongue. Given that the tongue is such a complex muscle, it would be almost impossible



Fig. 3. Example images from the *SynthBody*, *SynthFace* and *SynthHand* datasets, all of our neural networks are trained exclusively on these synthetic datasets.

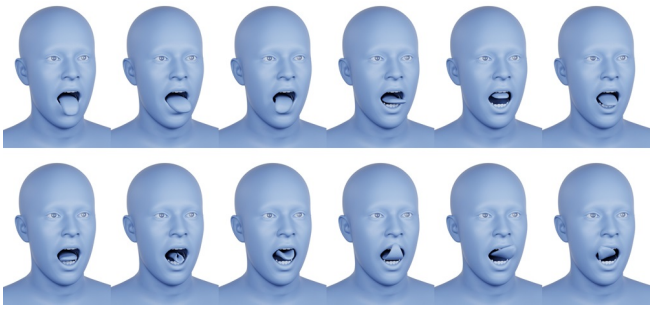


Fig. 4. Twelve tongue blendshapes included in the SOMA model, visualized with mouth opening blendshapes also activated. These blendshapes allow for significant coverage of the range-of-motion of the tongue.

to capture its *full* range of motion using a practical number of blendshapes. Rather than introducing a complex joint based rig which would involve a more challenging fitting procedure, we instead compromise on the expressivity of our model and select a minimal set of blendshapes that provide reasonable coverage, these are visualized in Figure 4.

Face Shape Basis. We retrain the model of Wood et al. [2021] with a training set of $N = 1040$ scans, selected to enhance the diversity of the data across multiple demographic axes. We find the model of Wood et al. [2021] does not generalize well to populations with thinner or thicker lips and analysis shows inconsistencies in the semantic correspondence of the topology around the lips in the original training data. As such, we reprocess the training data to enforce consistency in the mouth region using our method and commercial retopologization software [Russian3DScanner 2021] in a semi-automated approach. We then carry out photometric refinement [Nicolet et al. 2021] to better capture high-frequency features such as wrinkles. For further details of the improvements made to the training data, see the supplementary material. To obtain the shape basis we jointly fit the basis and parameters, $[\boldsymbol{\gamma}_0, ..., \boldsymbol{\gamma}_N]$, to

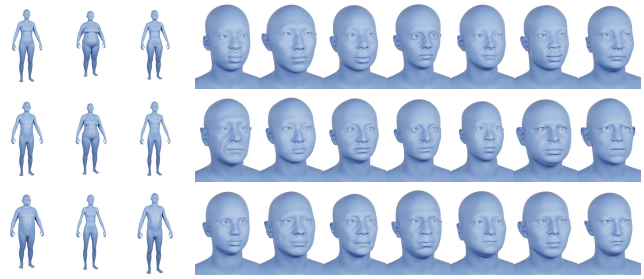


Fig. 5. Body and face shape samples. We sample at random from male, female and ungendered GMMs. The SOMA model itself has no explicit concept of gender.



Fig. 6. Pose and expression library samples. If not all data is present in the original capture, we splice together body, hand and face pose and expression from separate sequences.

the vertices of the training scans, while other components of the model are kept constant, using vector Adam [Ling et al. 2022].

3.2 Asset Library

As outlined in Figure 2, we build up a unique render by sampling shape, pose, hair and clothing assets, texture and environment. Face and body shape are sampled from GMMs fit to, or directly from, the face training data and a library of 3572 body scans from various sources [Mahmood et al. 2019; Renderpeople 2024; Ten24 Media 2024; Yang et al. 2014]. Example identities are shown in Figure 5.

For facial expression we re-register the library of Wood et al. [2021] to improve the quality and add tongue articulation, as well as adding additional artist generated sequences including the tongue. For body and hand pose we sample from a large library including AMASS [Mahmood et al. 2019], MANO [Romero et al. 2017] and other motion capture data processed using MoSh [Loper et al. 2014]. Example poses and expressions are shown in Figure 6.

Once we have a posed body mesh, we add hair, clothing and accessories and render the result in a realistic environment. For this, we construct a large library of artist-created assets, from which we can sample to produce a diverse range of realistic scenes. We use displacement maps to model basic clothing, and mesh assets for glasses and headwear. Hair is represented using strands, with separate assets for the scalp hair, facial hair, eyebrows and eyelashes. We use a mixture of HDRIs and 3D room environments to provide a wide variety of environmental lighting and background content.



Fig. 7. Selection of assets from our library of accessories, clothing and hair. Displacement-based assets automatically adapt to variation in body shape, mesh-based assets and hair are adapted using lattice warping.

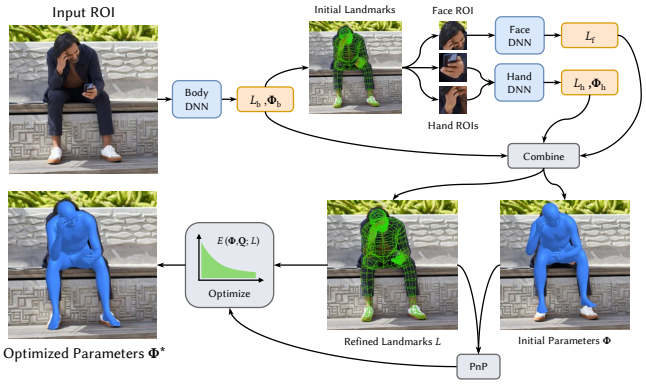


Fig. 8. Overview of our method architecture. From an input image we regress landmarks and initial model parameters, we refine these for the face and hands and minimize reprojection error to obtain the output parameters.

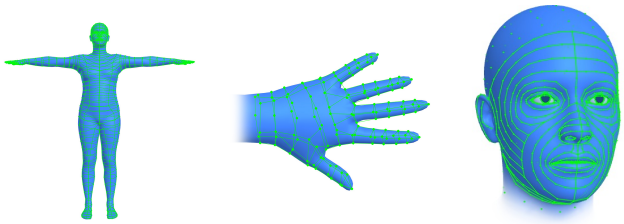


Fig. 9. Dense landmark definitions for the body, hands and face. We use 1428 landmarks for the full body, 141 for each hand and 744 for the face. Face landmarks include 16 points on the limbus boundary in each eye, 12 points on the teeth and 9 points on the tongue.

Example assets can be seen in Figure 7 and further details of the asset library are given in the supplementary material.

3.3 Training Datasets

To train the DNNs used in our method, we generate three datasets, each containing approximately 100,000 images, one featuring the entire body, one the face, and one the left hand. We call these datasets *SynthBody*, *SynthFace* and *SynthHand*. These datasets will be released publicly with sparse annotations upon publication. Example images from each dataset can be seen in Figure 3. *SynthBody* and *SynthFace* are rendered at 512×512 pixel resolution, both include 2D landmark and semantic segmentation annotations. *SynthBody* also includes 3D joint locations and SMPL-H pose and shape parameters while *SynthFace* also includes head pose annotations. *SynthHand* is rendered at 256 × 256 pixel resolution and includes 2D and 3D joint locations and MANO parameter annotations. All datasets also include camera parameters, download instructions can be found at https://aka.ms/SynthMoCap. Note that all results presented below are from DNNs trained *exclusively* on these synthetic datasets.

4 PERFORMANCE CAPTURE

We employ a multi-stage technique to fit the SOMA model to image or video data. The first stage regresses information including 2D landmarks from the input image(s) using DNNs trained on synthetic data generated using the pipeline described in Section 3. The second stage uses conventional optimization to fit our parametric model to the observed 2D landmarks by minimizing the reprojection error of the 3D mesh vertices into the image(s). The optimization process involves a series of energy terms to encode prior knowledge of the scenario, as detailed below. We find that using 2D landmarks alone does not lead to robust results, and that it is critical to initialize the optimization with a good approximation of the body shape and pose. As such, we also train the DNNs to predict body pose and shape from the input image(s). The method can therefore also be run in real-time without using optimization to provide a solution for previsualization of results. Figure 8 provides a graphical representation of our pipeline.

4.1 Neural Networks

Our model-fitting pipeline relies on full-body dense landmarks as well as an initialization for pose and shape. High density of the landmarks is critical to ensure accurate shape reconstruction of the surface of the human [Wood et al. 2022]; we show our landmark set on the template meshes in Figure 9. We observe that DNNs trained to directly regress parametric model parameters from images are very robust and learn a strong implicit prior over the body pose, though struggle to achieve very precise results on their own. Starting from a reasonable estimate of body pose, however, helps the optimizer to find a better solution, particularly in the under-constrained single-view case.

We find a single DNN is unable to capture adequately high-fidelity results for the face and hands, perhaps due to the limited resolution in a region of interest (ROI) including the whole body. We therefore train three separate DNNs for the body, hands, and face; each model takes ROIs cropped closely to the relevant body part as input. During training, the ROI is estimated based on the ground-truth landmarks, at inference time, a detector is used to estimate the full-body ROI and sub-ROIs are extracted based on the initial landmark prediction. For further details on detection and tracking, details of the training procedure and experiments supporting the above design choices, please see the supplementary material.

Architecture and Losses. Following Zhang et al. [2023b], we use a HRNet backbone [Sun et al. 2019], pretrained on ImageNet [Deng et al. 2009], to extract visual features from the input image. As shown in Figure 10, the compressed representation computed with HRNet then serves as input to three MLP-based heads to predict probabilistic landmarks, pose, and shape.

For predicted landmarks, $\mathbf{L}$, we model each as a circular Gaussian, $L_i = \{\boldsymbol{\mu}_i, \sigma_i\}$, where $\boldsymbol{\mu}_i = [x_i, y_i]$ is the expected 2D position of that landmark, and σ_i represents the uncertainty of the prediction [Wood et al. 2022]. We denote ground-truth landmark positions $\boldsymbol{\mu}'_i = [x'_i, y'_i]$. For training we use Gaussian negative log likelihood loss with per-landmark weights, λ_i,

$$\mathcal{L}_{\text{ldmk}} = \sum_{i=1}^{|\mathbf{L}|} \lambda_i \left(\log\left(\sigma_i^2\right) + \frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}'_i\|^2}{2\sigma_i^2} \right). \tag{1}$$

We represent pose parameters output by the pose head as 6D rotations following Zhou et al. [2019], and compute the L_1 loss $\mathcal{L}_{\text{pose}} = \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_1$ between the ground-truth rotations $\boldsymbol{\xi}'$ and the predicted rotations $\boldsymbol{\xi}$. We write the forward kinematics of the SOMA model as $\mathbf{J}_T, \mathbf{J}_R = H(\boldsymbol{\theta}; \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta})$, where H computes world-space joint translations $\mathbf{J}_T$ and rotations $\mathbf{J}_R$ for the given pose, $\boldsymbol{\theta}$, and shape parameters, $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}$. Then we have the ground-truth joint transforms $\mathbf{P}'_T, \mathbf{P}'_R = H(\boldsymbol{\theta}'; \boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\zeta}')$ and the predicted joint transforms $\mathbf{P}_T, \mathbf{P}_R = H(n(\boldsymbol{\xi}); \boldsymbol{\beta}, \boldsymbol{\gamma}', \boldsymbol{\zeta}')$ where n maps 6D rotations to axis-angles. We use these to compute a pose reconstruction loss decomposed into two parts:

$$\mathcal{L}_{\text{joint}_T} = \frac{1}{|\mathbf{J}|} \sum_{j=1}^{|\mathbf{J}|} \|\mathbf{P}_T^j - \mathbf{P}_T'^j\|_1 \tag{2}$$

$$\mathcal{L}_{\text{joint}_R} = \frac{1}{|\mathbf{J}|} \sum_{j=1}^{|\mathbf{J}|} \arccos\left(\frac{tr(\mathbf{P}_R^j (\mathbf{P}_R'^j)^T) - 1}{2} \right) \tag{3}$$

where $|\mathbf{J}|$ is the number of joints and $tr(.)$ is the trace operation.

The shape head predicts SOMA shape parameters for which we compute the L_1 error between the ground truth shape parameters and the prediction, $\mathcal{L}_{\text{shape}} = \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_1$. Note that shape parameters are also implicitly optimized in Equation 2. Our final loss function is therefore

$$\mathcal{L} = \alpha_l \mathcal{L}_{\text{ldmk}} + \alpha_p \mathcal{L}_{\text{pose}} + \alpha_t \mathcal{L}_{\text{joint}_T} + \alpha_r \mathcal{L}_{\text{joint}_R} + \alpha_s \mathcal{L}_{\text{shape}} \tag{4}$$

where $\alpha_l, \alpha_p, \alpha_t, \alpha_r, \alpha_s$ are the weights assigned to each loss term.

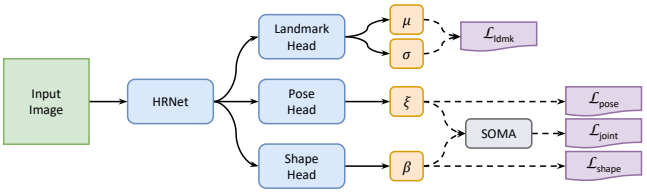


Fig. 10. Landmark, pose, and shape estimator DNN architecture. Used in full for the body DNN, without shape head for the hand DNN, and with only landmark head for the face DNN. $\boldsymbol{\theta} = n(\boldsymbol{\xi})$ is input to the SOMA model.

Models. We train the body DNN on the *SynthBody* dataset using the full architecture and losses described above. The model predicts 1428 landmarks on the body (see Figure 9), body pose, $\boldsymbol{\xi} \in \mathbb{R}^{23\times6}$, (excluding the pelvis and hands) and the first 10 components of the SOMA body shape parameters which explain the majority of shape variance. The hand DNN is trained on the *SynthHand* dataset with the shape head and corresponding loss omitted ($\alpha_s = 0$). We train only a single DNN for the left hand. During inference, we mirror the input image and predicted landmarks in the x-direction and mirror the predicted pose around the y-axis for the right hand. The hand DNN predicts 141 dense landmarks on the hand (see Figure 9) and 15 joints for the hand pose (excluding the wrist). The face DNN is trained on the *SynthFace* dataset and directly regresses 744 landmarks on the face (see Figure 9). We omit the pose and shape heads ($\alpha_s = \alpha_p = \alpha_t = \alpha_p = 0$) as variation in the facial appearance due to shape and expression is of far lower magnitude than for the body, and the optimizer is able to fully determine these in the model fitting stage.

4.2 Model Fitting

We use L-BFGS [Liu and Nocedal 1989] to find the optimal parameters for the SOMA model to explain the observed landmarks, starting from our pose and shape initialization. If multiple views are present then we regress pose and shape from the view with the lowest mean sigma of predicted probabilistic landmarks. We first initialize the global position and orientation using PnP [Fischler and Bolles 1981], then employ a three-stage fitting procedure: (1) finalizing the PnP initialization by optimizing only the translation and rotation (2) fitting the SOMA model by optimizing pose, expression and shape parameters, and optionally (3) refining the camera extrinsics (multi-view) or focal length (single-view), assuming camera calibration is not provided.

In multi-view scenarios where no camera calibration is provided, the camera positions and orientations can be estimated accurately using PnP with the template 3D head mesh and the 2D face landmarks for each view [Szymanowicz et al. 2022]. The head is used instead of the full body due to the minimal impact on overall shape caused by variation in pose and expression, leading to more reliable camera positioning. Unless provided, we assume no camera distortion is present and that the optical center is aligned to the image center.

The objective function for the optimization is:

$$E(\mathbf{\Phi}, \mathbf{Q}; \mathbf{L}) = \underbrace{E_{\text{ldmks}}}_{\text{Data term}} + \underbrace{E_{\text{shape}} + E_{\text{exp}} + E_{\text{pose}} + E_{\text{temp}} + E_{\text{intsct}} + E_{\text{cam}}}_{\text{Regularizers}} \quad (5)$$

where $\mathbf{\Phi} = \{\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \boldsymbol{\psi}, \boldsymbol{\theta}\}$ are the SOMA parameters, $\boldsymbol{Q}$ are the camera extrinsics in the multi-view case, or focal length in the single-view case, which are constant through the sequence, and $\mathbf{L}$ are the predicted probabilistic landmarks.

The only data term is the re-projection error for $\mathbf{L}$. Critically, we exploit the predicted uncertainty, σ, to weight the loss for each landmark, so for F frames over C cameras:

$$E_{\text{ldmks}} = \sum_{i,j,k}^{F,C,|\mathbf{L}|} \frac{\|\mathbf{x}_{ijk} - \boldsymbol{\mu}_{ijk}\|^2}{2\sigma_{ijk}^2} \quad (6)$$

where $[\boldsymbol{\mu}_{ijk}, \sigma_{ijk}]$ is the 2D location and uncertainty predicted for the k^{th} landmark, and $\boldsymbol{x}_{ijk}$ is the 2D projection of that landmark on the posed SOMA mesh, $\mathcal{M}(\mathbf{\Phi})$, as seen by the j^{th} camera in the i^{th} frame [Wood et al. 2022]. This weighting of the loss based on uncertainty of the predicted landmarks is particularly valuable in the full-body scenario where body parts can be easily self-occluded based on the pose or viewing angle. In the multi-view case, this lets the view with the best visibility of a certain body part take precedence over those with occlusions. In the single-view case it allows the priors to exert greater influence for occluded body parts, while the landmarks take precedence for unoccluded regions.

Regularizers. We use a number of priors to regularize the optimization [Pavlakos et al. 2019; Wood et al. 2022]. E_{shape} is a combination of GMM priors; minimizing the log probability of the given body and face shape, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. For faces we fit a GMM to the shape parameters of the training scans described in Section 3.1, for bodies we fit a GMM to a library of SMPL-H identities (Section 3.2). E_{exp} is an L_1 loss on expression parameters to promote sparsity, combined with quartic barrier loss to keep coefficients in the range $[0, 1]$. E_{temp} is an L_2 loss on frame-to-frame distance between 3D vertex locations, this helps to reduce jitter that might arise from noise in the predicted 2D landmarks. E_{intsct} is to prevent intersection of the eyeballs, teeth and tongue with the surface of the face [Wood et al. 2022]. E_{cam} is an L_2 loss on the position of the cameras compared to their initial location, to prevent them moving too far from the initialization. It is trivial to extend the system with additional priors in the case where we have further information on the scene. For example we can add an L_2 loss on the height of the subject to reduce ambiguity in scale, or restrict the shape prior based on the gender of the subject.

Neural Pose Prior. A key challenge is to ensure the plausibility of the estimated body and hand poses, particularly in single-view scenarios where the problem is ill-posed. The level of variation and self-occlusion due to body and hand pose is much higher than for facial expression, which we find does not require complex priors. To address this, we use a neural pose prior, specifically a Normalizing Flow [Rezende and Mohamed 2015].

Our prior consists of a known base distribution and a sequence of bijective transformations that enable the model to map the data to the latent space and vice versa. The body (or hand) poses are modeled as a result of applying a transformation T to a latent code z sampled from a base distribution $p_z(z)$, which is a multivariate normal in our case. The transformation T is invertible, which allows us to use the change of variable formula to compute the density of the poses as

$$p_q(q) = p_z\big(T^{-1}(q)\big)\left|\det \mathbb{J}_{T^{-1}}(q)\right| \quad (7)$$

where $\mathbb{J}_T$ is the Jacobian of the transformation and q is the pose. More specifically $q = [\boldsymbol{\xi}, \mathbf{P}_T, \mathbf{P}_R]$, the concatenation of local joint rotations and global joint transforms of a normalized pose (front facing root orientation positioned at $[0, 0, 0]$).

To capture the complexity of the body and hand pose distributions, we use a series of simple transformations $T = T_K \circ T_{K-1} \circ \ldots \circ T_1$ to form a complex one, where each T_i maps z_{i-1} to z_i, z_0 is the latent variable in the base distribution, and z_K is the data q. This composition is modeled by a neural network (with Real-NVP [Dinh et al. 2017] architecture) and is trained to maximize the data log-likelihood, as

$$\log p_q(q) = \log p_z(z_0) - \sum_{i=1}^{K} \log \det \left|\frac{\partial T_i}{\partial z_i}\right| \quad (8)$$

For model fitting, we define the pose prior energy term as $E_{\text{pose}} = -\alpha_b \log p_{q_b}(q_b) - \alpha_h \log p_{q_h}(q_h)$, wherein q_b and q_h are the body and hand pose representations, respectively, and α_b and α_h are the corresponding negative log-likelihood weights for body and hands.

5 EXPERIMENTS

Given that holistic performance capture is not a common technique, there are few datasets available to evaluate the full method. As such, we use a number of different benchmarks to evaluate subtasks in isolation. Taken together, and in combination with our qualitative results, these demonstrate the high quality achieved by our method, and its robustness across diverse input data and scenarios. Further results, details of the benchmark datasets and experimental procedure, and comprehensive ablation experiments are provided in the supplementary material.

It is important to note that while many benchmarks provide both training and test sets, we use the same DNNs (trained on the *SynthBody*, *SynthHand* and *SynthFace* datasets only) for all experiments; we do not use *any* of the benchmark training sets. It is very difficult to achieve comparable metrics when evaluating cross-dataset [Choudhury et al. 2023], though cross-dataset evaluation provides a much better estimate of generalization of the approach. For benchmarks with training sets, the best results using the training set are denoted in **bold** and the best results not using the training set are shown underlined. The ability of our method to generalize well across *all* of these benchmarks suggests high diversity and fairness in our synthetic data, and that we are not biased to any particular scenario.

Hands. To evaluate performance capture for the hands we use the FrieHAND dataset [Zimmermann et al. 2019]. We fit just the MANO hand model [Romero et al. 2017] using the prediction of the hand DNN and E_{pose} only. Quantitative results are shown in Table 2, where our method performs comparably to recent approaches. Note that ours is the only method that is *not* trained on the FreiHAND

Table 2. Single-view hand reconstruction errors on the FreiHAND dataset [Zimmermann et al. 2019], *ours is the only method not trained on FreiHAND*. Qualitative results are shown in Figure 11.

Method	PA-MPVPE ↓	PA-MPJPE ↓	F-Scores ↑ 5 mm / 10 mm
Non-parametric			
I2L-MeshNet [Moon and Lee 2020]	7.6	7.4	0.681 / 0.973
METRO [Lin et al. 2021a]	**6.7**	**6.8**	**0.717** / **0.981**
Parametric			
FrankMocap [Rong et al. 2021]	11.6	9.2	0.553 / 0.951
PIXIE [Feng et al. 2021]	12.1	12.0	0.468 / 0.919
Hand4Whole [Moon et al. 2022]	7.7	7.7	0.664 / 0.971
PyMAF-X [Zhang et al. 2023b]	8.1	8.4	0.638 / 0.969
PyMAF-X (extra data) [Zhang et al. 2023b]	7.5	7.7	0.671 / 0.974
Ours	8.1	8.6	0.653 / 0.967

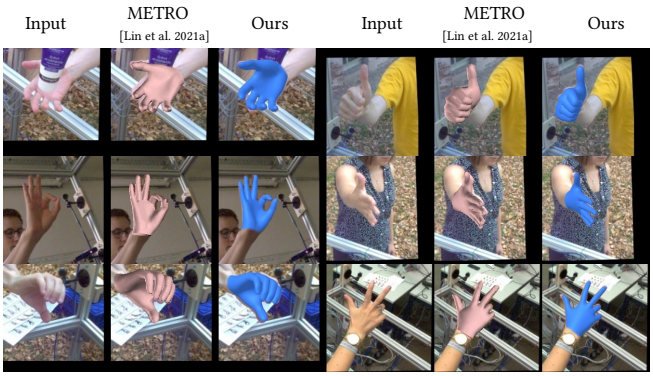


Fig. 11. Qualitative comparison to METRO [Lin et al. 2021a] for hand reconstruction, on images from the FreiHAND dataset [Zimmermann et al. 2019]. *FreiHAND © LMB, University of Freiburg.*

Table 3. Face shape reconstruction results on the NoW Challenge [Sanyal et al. 2019] validation and test sets. Qualitative results are shown in Figure 12.

Method	Val Error (mm)			Test Error (mm)		
Single-view	Median	Mean	Std	Median	Mean	Std
PyMAF-X [Zhang et al. 2023b]	-	-	-	1.13	1.42	1.20
Wood et al. [2022]	0.95	1.17	0.97	1.02	1.28	1.08
AlbedoGAN [Rai et al. 2024]	0.90	1.12	0.96	0.98	1.21	0.99
MICA [Zielonka et al. 2022]	0.91	1.13	0.95	0.90	1.11	0.92
TokenFace [Zhang et al. 2023a]	-	-	-	**0.76**	**0.95**	**0.82**
Ours	**0.87**	**1.08**	**0.89**	0.89	1.11	0.93
Multi-view	Median	Mean	Std	Median	Mean	Std
Bai et al. [2021]	-	-	-	1.08	1.35	1.15
Wood et al. [2022]	0.75	0.90	0.73	0.81	1.01	0.84
Ours	**0.68**	**0.83**	**0.68**	**0.70**	**0.88**	**0.74**

training set and that recent non-parametric methods [Lin et al. 2021b; Moon and Lee 2020] lead to superior metrics but are not useful for performance capture. Qualitative results and comparison to recent work are shown in Figure 11 and the supplementary material.

Face. We evaluate facial performance capture on the NoW benchmark [Sanyal et al. 2019] by fitting just the face of the SOMA model, omitting any unnecessary energy terms; results are shown in Table 3. Most full-body benchmarks have low-quality face ground truth, so this lets us compare our face model with other recent methods, including the model of Wood et al. [2021], more directly. We also evaluate a variant of the NoW benchmark where we reconstruct the face shape of each subject using *all* images of that subject together as input to our system, i.e., the multi-view scenario. While our method does not outperform Zhang et al. [2023a] in the single-view case, our approach extends to multi-view input which provides more accurate reconstructions. Qualitative results and comparison to recent work are shown in Figure 12 and the supplementary material. The NoW challenge only considers the face, not the full head, so improved quality of head/ear shape is not captured in the quantitative results.

Body. For the full body we evaluate on three datasets: EHF [Pavlakos et al. 2019] which evaluates both shape and pose reconstruction, SSP-3D [Sengupta et al. 2020] which evaluates shape reconstruction, and Human3.6M [Ionescu et al. 2014] which evaluates skeletal pose estimation. EHF is most directly applicable to the performance capture use case, though only covers the single-view scenario. Human3.6M does not allow for assessment of shape reconstruction, but enables us to test our multi-view effectiveness.

Results for the EHF dataset [Pavlakos et al. 2019] are shown in Table 4. We outperform recent methods in most metrics or are on par with concurrent works, and qualitatively achieve significantly more accurate results as shown in Figure 13. We can easily introduce additional supervisory signals to our method; for example if we introduce a single measurement of the actor's height we achieve a full-body MPVPE (*without* pelvis alignment) of 43.1mm, as the scale ambiguity inherent to the single-view scenario is removed.

Results for shape reconstruction on the SSP-3D dataset are shown in Table 5. Our method outperforms the closest method also trained exclusively on synthetic data [Black et al. 2023], as well as beating the overall state-of-the-art approach [Sengupta et al. 2021a].

Results for multi-view reconstruction on the Human3.6M dataset [Ionescu et al. 2014] are shown in Table 6. Our method significantly outperforms other multi-view methods which are also not trained on Human3.6M. Given the limited variety of the data there is a high chance that other methods over-fit; this is supported by the results of Choudhury et al. [2023] reproduced in Table 6. The Human3.6M benchmark evaluates only joint positions which do not fully specify human pose, leaving some degrees of freedom (e.g., rotation about a limb) undefined [Zhang et al. 2021]. Our method, which recovers the full body mesh, often leads to more compelling results, supported by the qualitative results shown in Figure 13.

6 DISCUSSION

6.1 Limitations

Our method has a number of limitations, both in the synthetic training data and model fitting approach. Due to the probabilistic nature of the landmarks we can ignore any frames where the average confidence of the landmarks is too low, resulting in smoothly interpolated results from the adjacent frames. This produces plausible results, but at significant cost to accuracy in those frames. The system as a whole is reliant on the quality of the predicted 2D landmarks which can lead to failures for interlocking hands or if poor sub-ROIs are selected. Tuning of priors to regularize the optimization can also be cumbersome, though in most cases these are transferable.

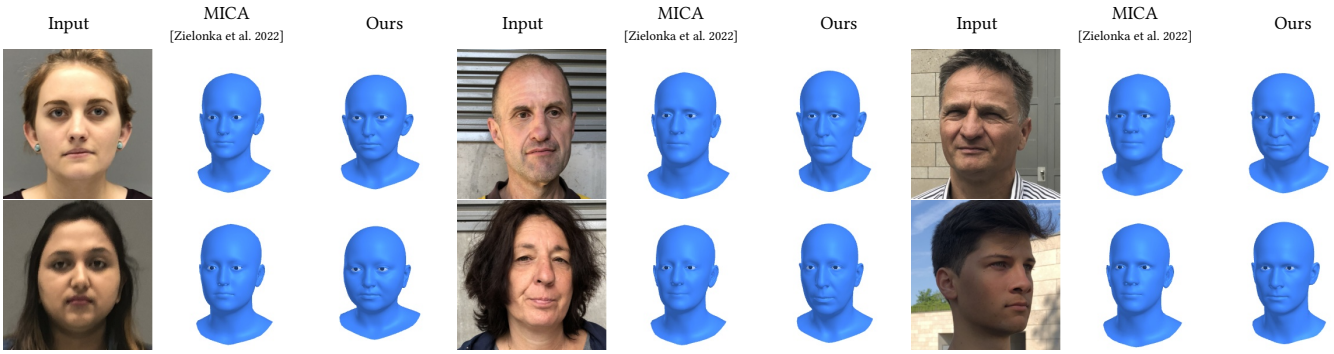


Fig. 12. Qualitative comparison to MICA [Zielonka et al. 2022] for face reconstruction on a number of images from the NoW challenge [Sanyal et al. 2019]. We observe greater variation in nose, lip, ear and head shape, along with more accurate representation of wrinkles, in our method. *NoW © MPI.*

Table 4. Single-view holistic reconstruction errors on the EHF dataset [Pavlakos et al. 2019]. MPVPE is pelvis aligned, PA-MPVPE is aligned with Procrustes analysis. Qualitative results are shown in Figure 13.

Method	MPVPE			PA-MPVPE				PA-MPJPE	
	Full-body	Hands	Face	Full-body	Body	Hands	Face	Body	Hands
PIXIE [Feng et al. 2021]*	89.1	42.8	32.6	55.0	53.8	11.2	5.5	65.5	11.1
HybrIK-X [Li et al. 2023]*	121.4	52.1	41.9	59.8	50.0	17.6	8.1	60.8	17.9
OSX [Lin et al. 2023]	70.8	53.7	26.4	48.7	-	15.9	6.0	-	-
PyMAF-X (Res50) [Zhang et al. 2023b]	68.0	29.8	20.5	47.3	45.9	10.1	5.6	52.9	10.3
PyMAF-X (HR48) [Zhang et al. 2023b]	64.9	29.7	19.7	50.2	44.8	10.2	5.5	52.8	10.3
SMPLer-X-L20 [Cai et al. 2024]†	65.4	49.4	17.4	37.8	-	15.0	5.1	-	-
SMPLer-X-L32 [Cai et al. 2024]†	62.4	47.1	17.0	37.1	-	14.1	5.0		
Multi-HMR-ViT-L/14 [Baradel* et al. 2024]†	**42.0**	28.9	**18.0**	**28.2**	-	10.8	5.3	-	-
Ours	51.0	**21.6**	18.4	34.2	**33.0**	**9.8**	**4.9**	**38.0**	**9.7**

* Results evaluated for this paper using official implementation - Feng et al. [2021] use a different evaluation protocol for the face, Li et al. [2023] do not report on EHF. † Results from concurrent works.

Table 5. Single-view body shape reconstruction errors on the SSP-3D dataset [Sengupta et al. 2020].

Method	PVE-T-SC
HMR [Kanazawa et al. 2018]	22.9
SPIN [Kolotouros et al. 2019]	22.2
SHAPY [Choutas et al. 2022b]	19.2
STRAPS [Sengupta et al. 2020]	15.9
Sengupta et al. [2021b]	15.2
Sengupta et al. [2021a]	13.6
BEDLAM-CLIFF [Black et al. 2023]	14.2
Ours	**11.8**

Table 6. Multi-view body pose errors on the Human3.6M validation set [Ionescu et al. 2014] excluding sequences with incorrect ground-truth annotations following Iskakov et al. [2019]. MPJPE is *not* pelvis aligned.

Multi-view Method	Trained on H36M	MPJPE
Pavlakos et al. [2017]	×	56.9
Iskakov et al. [2019]	✓	**17.7**
TEMPO [Choudhury et al. 2023]	✓	18.5
TEMPO [Choudhury et al. 2023]	×	63.0
Ours	×	27.9

In the future we can extend to an ML-based iterative optimization approach to help reduce the reliance on bespoke priors and increase robustness to errors [Choutas et al. 2022a]. We often find that failures can be traced back to the synthetic training data. Particular shortcomings of our synthetic data include absence of particularly unusual poses, loose-fitting clothing and hair simulation. Using our method, gaps in pose data can be quite easily addressed through capture and manual cleanup. Gaps in the asset library can be addressed through artist effort. While we can support moving cameras when provided with calibration, the current implementation does not support jointly optimizing for subject *and* camera positions that vary through time. This would be a powerful extension of the system to allow truly arbitrary camera input and has been shown to be effective in recent work [Shin et al. 2024].

6.2 Conclusion

We have demonstrated a system which enables holistic, high-quality performance capture of humans, including the body, face and hands. Our parametric model improves on models of the body and hands [Romero et al. 2017] and face [Wood et al. 2021] to represent the complete human at high fidelity. We use this model to generate synthetic data to train DNNs for 2D landmark regression and direct pose and shape prediction from images. We then use optimization to fit our model to images or videos of people, providing high-quality holistic performance capture. Our approach focuses on *(1) Robustness* enabled by DNNs directly regressing pose and shape; *(2) Accuracy*

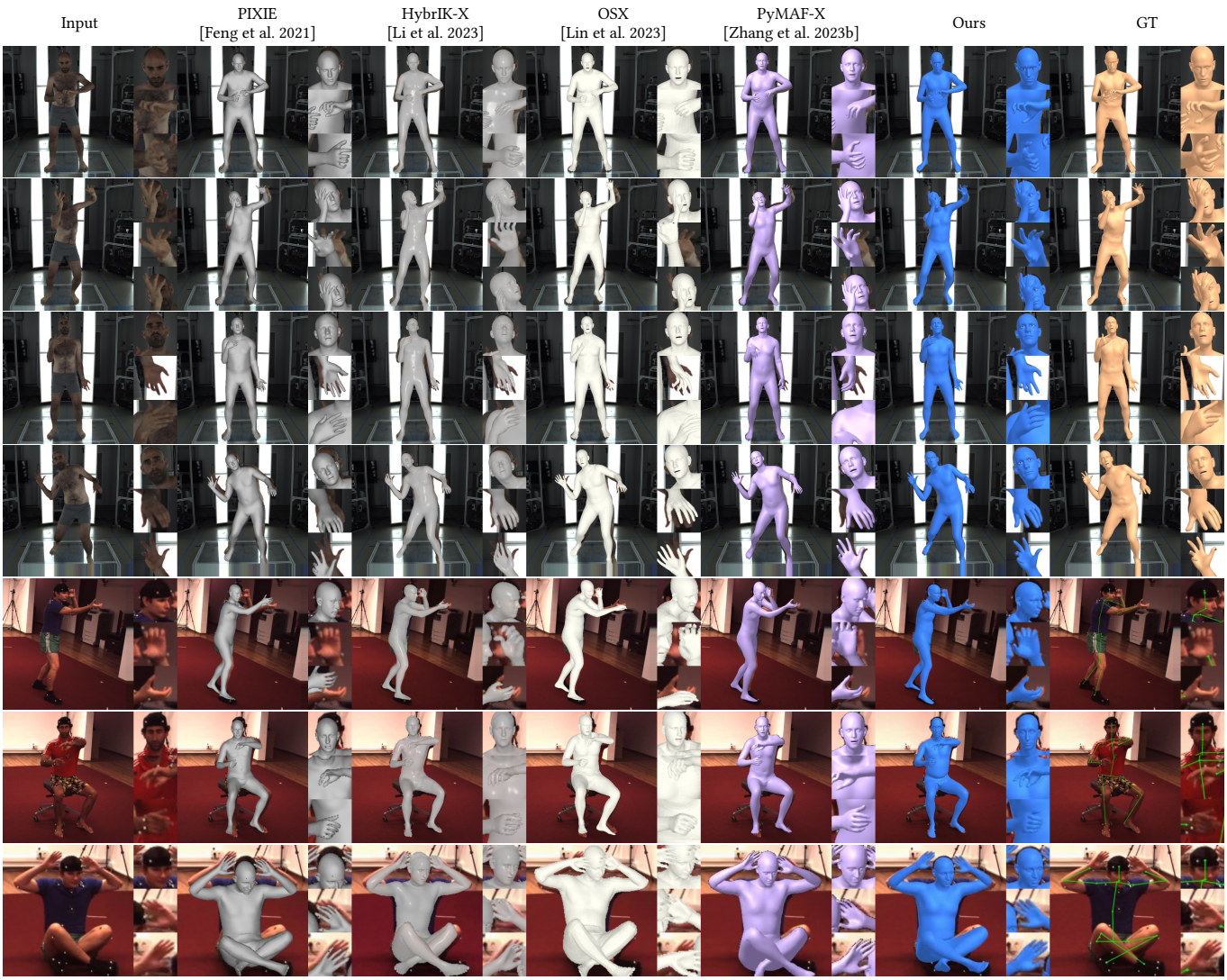


Fig. 13. Qualitative examples of single-view full body reconstruction for images from the EHF [Pavlakos et al. 2019] (top 4 rows) and Human3.6M [Ionescu et al. 2014] (bottom 3 rows) datasets compared with recent methods. Additional qualitative results including failure cases can be found in the supplementary material. Our method achieves more precise results and significantly higher fidelity reconstruction of the face. *EHF © MPI, Human3.6M © IMAR.*

enabled by DNNs predicting image-space features; *(3) Adaptability* of model fitting to support arbitrary numbers and positions of cameras. Because of this, our system can operate with no markers or proprietary hardware, arbitrary cameras, without calibration, and produces combined results for the body, hands and face from a single shot. We show state-of-the-art results on benchmarks for human reconstruction, demonstrating that it is possible to achieve high-quality, holistic performance capture without the hassle.

ACKNOWLEDGMENTS

The authors would like to thank Rodney Brunet, Kendall Robertson and Jon Hanzelka for their work on the clothing asset library; Steve Hoogendyk for his work on the tongue blendshapes; and Ben Lundell and Erroll Wood for their comments and suggestions.

A MODEL DEFINITION

The SOMA mesh is made up of $N = 12943$ vertices and 12726 polygons with a skeleton of $K = 54$ joints: 22 for the body (the SMPL skeleton), 15 per hand (as in SMPL-H) and 2 for the eyes, from the face model of Wood et al. [2021], represented as axis-angles.

The mesh vertex positions are defined by mesh generating function $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \psi, \boldsymbol{\theta}) : \mathbb{R}^{|\boldsymbol{\beta}|+|\boldsymbol{\gamma}|+|\boldsymbol{\zeta}|+|\boldsymbol{\psi}|+|\boldsymbol{\theta}|} \rightarrow \mathbb{R}^{N\times 3}$ which takes parameters $\boldsymbol{\beta} \in \mathbb{R}^{300}$ for body shape, $\boldsymbol{\gamma} \in \mathbb{R}^{256}$ for face shape, $\boldsymbol{\zeta} \in \mathbb{R}^{9}$ for hand shape, $\boldsymbol{\psi} \in \mathbb{R}^{224}$ for expression, and $\boldsymbol{\theta} \in \mathbb{R}^{54\times 3}$ for skeletal pose.

$$\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \boldsymbol{\psi}, \boldsymbol{\theta}) = \mathcal{L}(\mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \boldsymbol{\psi}, \boldsymbol{\theta}), \boldsymbol{\theta}, \mathcal{J}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}); \mathbf{W}) \quad (9)$$

where $\mathcal{L}(\mathbf{X}, \boldsymbol{\theta}, \mathbf{J}; \mathbf{W})$ is a standard linear blend skinning (LBS) function that rotates vertex positions $\mathbf{X} \in \mathbb{R}^{N\times 3}$ about joint locations $\mathbf{J} \in \mathbb{R}^{K\times 3}$ by local joint rotations $\boldsymbol{\theta}$, with per-vertex hand-authored skinning weights $\mathbf{W} \in \mathbb{R}^{K\times N}$ determining how rotations are interpolated across the mesh.

$\mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \boldsymbol{\psi}, \boldsymbol{\theta}) : \mathbb{R}^{|\boldsymbol{\beta}|+|\boldsymbol{\gamma}|+|\boldsymbol{\zeta}|+|\boldsymbol{\psi}|+|\boldsymbol{\theta}|} \rightarrow \mathbb{R}^{N\times 3}$ constructs an unposed body mesh by adding displacements to the template mesh $\overline{\mathbf{T}} \in \mathbb{R}^{N\times 3}$, which represents the average body in T-pose with neutral expression:

$$\mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \boldsymbol{\psi}, \boldsymbol{\theta})^j_k = \overline{T}^j_k + \beta_i S^{ij}_k + \gamma_i U^{ij}_k + \zeta_i V^{ij}_k + \psi_i E^{ij}_k + P(\theta)^j_k \quad (10)$$

With linear body shape basis, $\mathbf{S} \in \mathbb{R}^{|\boldsymbol{\beta}|\times N\times 3}$, face shape basis, $\mathbf{U} \in \mathbb{R}^{|\boldsymbol{\gamma}|\times N\times 3}$, hand shape basis, $\mathbf{V} \in \mathbb{R}^{|\boldsymbol{\zeta}|\times N\times 3}$, expression basis, $\mathbf{E} \in \mathbb{R}^{|\boldsymbol{\psi}|\times N\times 3}$, and $P(\boldsymbol{\theta})$ which represents pose-dependent blendshape offsets for pose parameters $\boldsymbol{\theta}$. Finally, $\mathcal{J}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}) : \mathbb{R}^{|\boldsymbol{\beta}|+|\boldsymbol{\gamma}|+|\boldsymbol{\zeta}|} \rightarrow \mathbb{R}^{K\times 3}$ moves the joint locations to account for changes in shape:

$$\mathcal{J}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta})^j_k = J(\overline{T}^j_k + \beta_i S^{ij}_k + \gamma_i U^{ij}_k + \zeta_i V^{ij}_k) \quad (11)$$

Where J is a modified version of the SMPL-H joint regressor. The body shape basis, $\mathbf{S}$, is that from neutral SMPL-H [Romero et al. 2017] masked to only the body. The hand shape basis, $\mathbf{V}$, is that from MANO [Romero et al. 2017] excluding the first component which broadly represents scale, as this is already captured in $\mathbf{S}$. Both are PCA bases learned from scans of humans. The pose-dependent blendshapes, $P(\boldsymbol{\theta})$, are also taken from neutral SMPL-H. The face shape basis, $\mathbf{U}$, is learned from a library of face scans as described in the main paper. The face expression basis, $\mathbf{E}$, is taken from Wood et al. [2021], but with an additional 12 blendshapes added to provide articulation of the tongue.

To construct the template mesh, $\overline{\mathbf{T}}$, we manually align the template of Wood et al. [2021] to the head of the SMPL-H template. Once aligned, the head of SMPL-H and lower neck of the new head are removed and the two partial meshes merged. The topology around the join is hand-crafted to create a smooth transition given the different density of the two meshes, see Figure 14. A new UV-map is also hand-authored based on the SMPL-H UV space.

Given that the topology of the template mesh differs significantly from both Wood et al. [2021] and SMPL-H, the bases associated with these models must be adapted to work with the SOMA model. To do this, we calculate a mapping function from the respective source topologies to the SOMA topology. Specifically, we map data from the source model to the SOMA by iterating through each SOMA vertex and taking the sum of the values for each vertex in the closest triangle of the source mesh, weighted according to the barycentric coordinate of the closest surface point on the source mesh to the target SOMA vertex. We restrict the bases of Wood et al. [2021] to affect on the head and neck, and those of SMPL-H to affect only the body, using a hand-crafted mask.

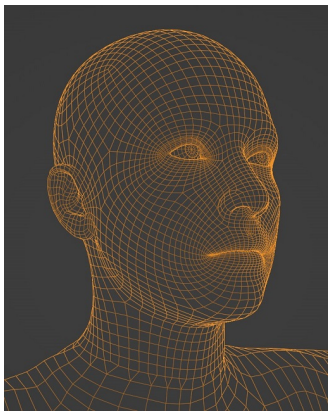


Fig. 14. Head and neck topology of the SOMA model.

B AUTOMATIC FACE RETOPOLOGIZATION

To add training data to the model of Wood et al. [2022], one would typically retopologize new head scans into the target topology using manual correspondence point selection in a tool such as Wrap [Russian3DScanner 2021]. This is slow and prone to human labeling inconsistencies. To reach our total of 1040 samples we automatically retopologized new face scans using the following method.

Firstly, due to poor consistency in the mouth region of the training data we annotate points on the lip margin of each scan. We then render each scan with uniform shading from many viewpoints. We use our multi-view fitting method to fit a 3D mesh that best reconstructs the scan in the face model's topology. We use this as an initialization to Wrap which can correct inaccuracies and improve quality of surface details compared to the scan. Supervising the Wrap retopologization process requires sparse correspondence points between the raw scan and the target face model topology for accurate surface alignment. We automatically select a sparse set on the fitted mesh's ears and eyes and add these to our manually annotated lip correspondences. Since the fitted mesh is close to the scan surface, we select corresponding points on the scan by finding the nearest scan vertices that have similar normals to those on the fitted mesh. To improve accuracy of back-of-ear correspondence points, we displace these points on the fitted mesh by 2cm behind the ear before finding the nearest point with similar normals on the scan mesh. Once the fitted mesh is retopologized onto the scan mesh using these correspondences, we use photometric refinement [Nicolet et al. 2021] to better capture high-frequency features such as wrinkles in the same topology.

Examples showing the poor semantic consistency in the mouth region described in the main paper, and improvements resulting from the above process, are shown in Figure 15.



Fig. 15. Improvements to semantic consistency in the mouth region as part of our updated face identity basis training. The same edge loop is shown in bold for the original retopologized scans and our updated versions. The new training data has topology that is significantly more consistent with the semantic region of the lips shown by the texture.

C SYNTHETIC DATA DETAILS

C.1 Pose Library

In order to pose the body realistically we collect a large library of poses and expressions from a variety of sources. For the facial expression and eye motion, we start with the library of Wood et al. [2021], which is captured using the technique of Wood et al. [2022]. Using the updated model and techniques outlined here we reprocess the raw multi-view video data to extract higher quality results including tongue motion. We also add a number of artist-animated expression sequences making use of the tongue articulation in order to boost representation in the generated data. In total this gives us almost 165,000 frames of expression and eye gaze data to sample from. For body pose data we combine the AMASS library [Mahmood et al. 2019] with a diverse library of additional pose sequences captured using conventional optical motion capture technologies and processed using MoSh [Loper et al. 2014]. Most of these libraries include only the body pose, but some also include articulated hands. In total this comprises over 2 million frames. We also use the MANO dataset [Romero et al. 2017] to provide additional hand poses in isolation.

We sample randomly from these libraries, combining face expression and body pose from independent sequences. If the body pose sequence lacks articulated hands then we splice these in from the hand pose library. For body pose, we weight the sampling by the log probability of an individual pose under a GMM fit to the entire library. For facial expression, we weight the sampling by the sum of the absolute value of the expression coefficients, and additionally randomize gaze direction and eye closing. This ensures that we pick difficult and diverse poses and expressions from the library.

C.2 Clothing

To model clothing in a way that adapts to a wide variety of body shapes and poses we use displacement maps [Ma et al. 2020]. The clothing assets are artist authored as textured meshes using Marvelous Designer [CLO Virtual Fashion Inc. 2024] on our template body mesh and then baked to displacement maps. We also author albedo, normal, metallic and roughness for each asset, providing a high level of realism. At render time we use adaptive subdivision to retain high frequency details of the clothing. Clothing assets are split into tops and bottoms, as well as a number of accessories such as bracelets, rings, gloves and watches. We also model prosthetic limbs using displacement. In total we have 384 displacement clothing assets and 148 displacement accessories.

We also adapt and extend some of the mesh clothing assets of Wood et al. [2021] to fit our new template mesh and use the same lattice-based deformation to fit these to different identities at render time. Our mesh clothing library includes 36 glasses assets, 56 headwear items and 7 masks and eyepatches.

While displacement clothing adapts very well to diverse body shapes and poses, it has significant limitations in terms of representation. It is not possible to represent loose-fitting clothing such as dresses and skirts using this technique, leading to a significant gap in our synthetic data. We hope to address this in the future using cloth simulation similar to that used by Black et al. [2023].

C.3 Hair

We adapt the hair library of Wood et al. [2021] and extend it with a particular aim towards diversity of hair types and styles. The hair grooms are artist-authored strand-based assets which we adapt to fit our new template mesh. At generation time the base assets are morphed to fit the head shape of the specific identity. In total our hair library includes 479 head hair styles, 81 eyebrows, 71 eyelashes and 91 facial hair styles.

Given the complexity of these assets, we currently do not perform any simulation to produce more realistic interaction of the hair with gravity or the body. We mostly observe that this isn't problematic in terms of our results as the DNNs are able to learn to generalize sufficiently, however it is undeniably a component of the domain gap between our synthetic and real images.

C.4 Textures

We extend the face texture library of Wood et al. [2021] to include 207 textures, enhancing the diversity of the selection in terms of ethnicity and age. To obtain pore-level detail bumpmaps we also high-pass filter the albedo texture, as in Wood et al. [2021]. However, this creates inconsistent appearance of skin specular roughness for different skin tones due to albedo textures differing in contrast. We weight the bumpmap strength by the skin tone brightness resulting in more natural and consistent appearance of skin specular highlights across skin tones. For the body we use a library of 25 high quality textures balanced by gender. When sampling textures we first select the face texture and then select a body texture with in some bound of perceptual similarity to the face texture. Following this we recolor the body texture to precisely match the face texture.

C.5 Environment

We use two techniques to provide lighting and background for our renders: 1) HDR-image-based lighting for indoor and outdoor environments [Debevec 2006; Wood et al. 2021], 2) and mesh-based scenes for indoor environments. For HDR environments we randomly sample from 914 images [Poly Haven 2024] and randomly rotate them and increase/decrease brightness to increase environment diversity. For mesh-based scenes we use 47 indoor environments [Chocofur 2024; Evermotion 2024]. The indoor environments are illuminated by area and point lights that simulate indoor lighting and an additional outdoor HDRI environment chosen from 477 outdoor HDRIs. To find a valid placement of a subject in the scene we randomly sample both the location (on predefined ground planes) and the pose of a clothed mesh until we find a combination without any intersections between the subject and the scene. As the mesh scene can cause occlusions, we randomly sample the camera until we find a position where at least 75% of joints are visible. We include secondary people to serve as occluders/distractors in a proportion of scenes.

D DETECTION AND TRACKING

For detection and tracking of the body in images we use a sliding-window approach similar to that described by Wood et al. [2022]. We train a light-weight DNN to detect 36 landmarks across the full body, with a ResNet18 [He et al. 2016] backbone.

For the first frame in a video, or for a single image, we use the sliding-window approach, executing this DNN on each window and taking the window with the highest confidence over a threshold. We then refine this candidate ROI by executing the DNN additional times to get a more precise bounding box around the predicted landmarks. This refined ROI then serves as input to the dense landmark DNN.

In a video we use the bounding box of the dense landmarks from the previous frame as the candidate ROI instead of running the sliding-window method every frame. If the median confidence for the predicted landmarks in the new frame is too low then we re-run the detection process in full for that frame. We can optionally carry out the iterative refinement of the candidate ROI for each frame in the video. To extract the face and hand ROIs in a given frame we take bounding boxes around subsets of the predicted dense full-body landmarks for the frame.

E DNN TRAINING DETAILS

For training the DNNs we use single frames of our synthetic datasets, which are cropped closely to the relevant body part, i.e., body, hand, and face. To this end, we extract a squared region of interest (ROI) around the specific set of landmarks. The crop is aligned with the image axes and tightly contains the 2D landmarks. We extract the squared ROI of size 128×128 for the hand model and of size 256×256 for body and face models. For evaluating the SOMA model in $\mathcal{L}_{\text{joint}_T}$ and $\mathcal{L}_{\text{joint}_R}$ for partial predictions we use ground-truth values for shape, translation and any missing pose parameters.

Augmentation. During training, we enhance the ROI extraction process by applying rotation, scaling, and translation. This augmentation simulates scenarios where the ROI is not perfectly defined during inference. In addition to ROI augmentations, we incorporate appearance augmentations into the training pipeline. In particular, we randomly blur the image to simulate motion blur effects caused by camera motion. This is achieved using directional blur kernels. It is important to note that this simulated motion blur is not due to actual scene movement; rather, it is a post-processing effect. This operation is proportional to image size and the direction of blur is randomly chosen. We also adjust brightness by adding a constant offset to all channels within a specified range. Similarly, we randomly increase or decrease contrast via $\text{img} = (\text{img} - 0.5) \times (1 + \text{contrast}) + 0.5$. Additionally, we alter the hue and saturation of the image, apply JPEG compression, and convert the image from BGR to grayscale. The application of each of these augmentations depends on a given probability. Furthermore, for training our hand model, we introduce random occlusions by overlaying occluders from a library of images onto the original image. These occluders are not physically present in the scene; they are simply layered over the image. The occluders undergo random shifts, scaling, and rotations. Lastly, we observe that incorporating random ISO noise [Alvar and Bajić 2022] (inspired by real camera noise) into the input image benefits our body model training. This noise is approximated as a combination of image intensity-dependent Poissonian noise and image intensity-independent Gaussian noise.

Network architecture. Our backbones are constructed from models in the PyTorch Image Models library [Wightman 2019]. In particular we use *hrnet-w48* backbone pre-trained on ImageNet1k for our face and body DNNs and *hrnet-w18* backbone pre-trained on ImageNet1k for our hand DNN. We define the hidden dimension of size 512 to get the visual features from these backbones. For the landmark, pose, and shape heads, we use a simple feed-forward neural network. The network is composed of two fully-connected layers separated by a Leaky ReLU activation function. We use a hidden dimension of size 512 and the output dimension is defined separately based on the number of outputs for each task.

Training hyper-parameters. We train the body and hand DNNs for 600 epochs and the face DNN for 300 epochs with a batch size of (64×4), distributed over four A100 GPUs. All networks are trained using the AdamW [Loshchilov and Hutter 2017] optimizer with a learning rate of 10^{-4}. The learning rate is adjusted using a Cosine Annealing scheduler for improved convergence. For training body and hand models, we set the landmark loss weight (α_l) to 10.0 and the pose and shape weights to 1.0 where applicable. We set α_t to 2.0 and α_r to 1.0 for both body and hand DNNs. For the neural pose prior, we train the body and hand priors for 600 epochs with a batch size of 1024. For both models, we use the AdamW optimizer, and we use a learning rate of 10^{-4} and 10^{-5} for the body and hand priors, respectively. Similar to body and hand DNNs, the learning rate is adjusted using a Cosine Annealing scheduler for improved convergence.

F REGISTRATION DETAILS

F.1 Energy Terms

A number of energy terms not fully defined in the main paper are provided here. The shape prior is made of two parts, one for the

face and one for the body:

$$E_{\text{shape}} = E_{\text{face\_shape}} + E_{\text{body\_shape}}$$

For each of face and body shape we have a multivariate Gaussian Mixture Model (GMM) of G components:

$$E_{\text{part\_shape}} = -\log\left(p(\boldsymbol{\beta})\right)$$

where $p(\boldsymbol{\beta}) = \sum_{i=1}^{G} \alpha_i \ \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\nu}_i, \Sigma_i)$. $\boldsymbol{\nu}_i$ and Σ_i are the mean and covariance matrix of the i^{th} component, and α_i is the weight of that component. E_{exp} prevents excessive use of expression blendshapes

$$E_{\text{exp}} = \|\boldsymbol{\psi}\| + E_{\text{quartic}}$$

where

$$E_{\text{quartic}} = \begin{cases} (t_0 - t)^4 & t < 0 \\ 0 & 0 \leq t \leq 1 \\ (t - t_1)^4 & t > 1 \end{cases}$$

E_{temp} reduces jitter by encouraging face mesh vertices to remain still between neighboring frames $i-1$ and i in a sequence of length F with C cameras.

$$E_{\text{temp}} = \sum_{i=2,j,k}^{F,C,|\mathbf{L}|} \|x_{i,j,k} - x_{i-1,j,k}\|^2$$

E_{cam} penalizes movement of the cameras from their initial position:

$$E_{\text{cam}} = \|\mathbf{Q} - \mathbf{Q}_{\text{init}}\|^2$$

where $\mathbf{Q}$ is the current position of all cameras and $\mathbf{Q}_{\text{init}}$ is the initial position of all cameras. E_{intsct} prevents intersection of the eyes and teeth with the face mesh, following Wood et al. [2022].

$$E_{\text{intsct}} = E_{\text{eyes}} + E_{\text{teeth}}$$

E_{eyes} penalizes the squared distance between the vertices in the eyelid and the surface of the sphere, E_{teeth} instead uses a convex hull of the teeth geometry and penalizes the distance inside this.

$$E_{\text{teeth}} = \sum_{i \in I} D_i^2$$

where D_i measures the distance the i^{th} skin vertex is inside the convex hull,

$$D_i = \min_{h \in H} \left\{ d_{i,1}, \ldots, d_{i,H} \right\}$$

and $d_{i,h}$ measures the internal distance between the i^{th} skin vertex and the h^{th} plane of the convex hull, defined by position p_h and normal $\hat{n}_h$.

$$d_{i,j} = -\min\left(\hat{n}_h \cdot x_i + p_h, 0\right)$$

F.2 Pose Prior

Our method utilizes a flow-based model, with Real-NVP [Dinh et al. 2017] architecture, as the pose prior. We deliberately choose Normalizing Flows [Aliakbarian et al. 2022; Kolotouros et al. 2021] over other commonly used priors based on Variational Autoencoders (VAEs) [Kingma and Welling 2014] for two reasons: first, VAEs tend to require careful tuning of the trade-off between the KL divergence loss and the reconstruction loss, which may vary depending on the scenario [Pavlakos et al. 2019; Rempe et al. 2021]. Second, unlike VAEs, which rely on approximate inference of the posterior distribution, Normalizing Flows can perform exact likelihood evaluations, enabling more robust optimization when used as a prior [Aliakbarian et al. 2022]. This capability of Normalizing Flows (compared to VAEs) has been extensively evaluated in the literature [Aliakbarian et al. 2022] wherein Normalizing Flows not only provided more accurate likelihood estimates, but also more robust estimates when evaluated on in- and out-of-distribution pose samples.

G EXPERIMENTAL DETAILS

Face. The NoW benchmark contains 2054 2D images of 100 subjects, with ground-truth 3D head scan for each subject. The validation set consists of 20 subjects and the test set consists of 80 subjects. The objective is to reconstruct the 3D shape of the subjects' faces as accurately as possible from each single image. We use the official evaluation process to evaluate our output meshes.

Hands. The FreiHAND [Zimmermann et al. 2019] test set contains 3960 224×224 pixel RGB images of the right hands of a number of subjects in various poses and environmental conditions, with ground-truth MANO parameters. We use the official implementation to calculate the standard metrics for the 3D reconstruction benchmark on this dataset; PA-MPVPE, PA-MPJPE and F-scores at 5 and 10mm error.

Body. The EHF dataset [Pavlakos et al. 2019] contains 100 monocular images of a single subject in a studio setting, with ground truth SMPL-X [Pavlakos et al. 2019] meshes. We register our SOMA model to the data and then retopologize to enable us to evaluate against the SMPL-X ground-truth. This retopologization will itself introduce error as some geometry is not common between the two meshes, so the results we report here are a conservative approximation of the true quality. We report MPVPE and PA-MPVPE for the entire body and various body parts in isolation, as well as PA-MPJPE for body and hands independently.

The SSP-3D dataset [Sengupta et al. 2020] contains 311 single-view images of sportspeople in various poses with SMPL [Loper et al. 2015] ground-truth annotations. Again we register the SOMA model and retopologize to match the ground-truth, in this case there is negligible error in this process. We report the PVE-T-SC metric, vertex error for the scale-corrected T-pose mesh, which is standard for this dataset.

The Human3.6M dataset [Ionescu et al. 2014] contains captures of a number of subjects from four cameras with ground-truth skeletal 3D landmarks. We use every fifth frame for subjects 9 and 11 for our evaluation, both in the single-view and multi-view case. For multi-view evaluation, we omit some sequences from subject 9 which have invalid ground truth [Iskakov et al. 2019]. For the single-view case, MPJPE with pelvis alignment and PA-MPJPE are reported, while for multi-view we report MPJPE *without* pelvis alignment. We evaluate the standard 14 joint configuration for the dataset, but a mapping is required from the SOMA mesh to these 14 joints as the joints of the SOMA models are not consistent with the Human3.6M annotation scheme. We follow the approach of Tan et al. [2016] and solve for the affine combination of just four vertices of the mesh that best predict each joint on an equally-spaced 5% subset of frames. Note that the flexibility of this mapping is extremely limited and this procedure simply automates an otherwise manual and biased process.



Fig. 16. Results including visible tongue. In many cases we achieve accurate results (left), although we do see a number of failure cases (right). For example, where the landmark DNN misses the tongue or confuses it for thicker lips, or the tongue blendshapes can't capture extreme tongue shapes.

Performance. We run our experiments either locally on a machine with an NVIDIA RTX 2080 GPU or on a virtual machine with an NVIDIA A100 GPU. On the RTX 2080 machine execution of the hand, face and body DNNs for a single image takes approximately 100ms The model fitting process for the EHF dataset (100 single-view frames, 200 iterations) takes 10 minutes on an A100 VM. Model fitting is performed jointly on the whole sequence and a large amount of time is spent copying data to and from the GPU, so execution time does not scale linearly with the number of frames. For example, fitting a 3000 frame sequence from Human3.6M (200 iterations) takes 30 minutes on an A100 VM.

H ADDITIONAL RESULTS

Figure 16 shows some results for faces including visible tongues. In many cases the reconstruction is of high quality, though we note failures due to lack of expressivity of the blendshapes and failures of the landmark DNN to accurately regress landmarks on the tongue. Common failure cases include confusion of the tongue for thicker lips, and failure to identify the visibility of the tongue at all. Addition of the tongue blendshapes also demonstrates how we could quite easily add further blendshapes based on quality gaps noticed in reconstruction quality for a given capture, or based on an enrolment sequence for a specific actor with idiosyncratic facial motion.

Results for the single-view case on the Human3.6M dataset [Ionescu et al. 2014] are shown in Table 7. Our method performs comparatively to other methods despite ours being the only method *not* trained on the Human3.6M training set. As highlighted in the main paper, the qualitative results shown in the main paper and in Figure 17 demonstrate that our method recovers significantly higher-quality body meshes. This however, is not reflected in the MPJPE metric calculated on sparse ground-truth annotations.

Further qualitative comparisons for full-body reconstruction on the EHF [Pavlakos et al. 2019] and Human3.6M [Ionescu et al. 2014] datasets are shown in Figure 17. One failure case of our method is for complex interlocking hand poses. We suspect that with higher-quality pose data (e.g., Moon et al. [2020]) and consequently better synthetic data (e.g., Moon et al. [2023]) we might improve robustness in these cases. It is possible though that the model fitting method itself may need to be augmented to adequately deal with these scenarios, e.g., to account for intersection or contact points [Taheri et al. 2022, 2020].

Table 7. Single-view body pose errors on the Human3.6M validation set [Ionescu et al. 2014]. MPJPE is pelvis aligned, PA-MPJPE is aligned with Procrustes-analysis. *Note that ours is the only method not trained on the Human3.6M training set.* Qualitative results are shown in the main paper and Figure 17.

Single-view Method	MPJPE	PA-MPJPE
SPIN [Kolotouros et al. 2019]	62.5	41.1
I2L-MeshNet [Moon and Lee 2020]	55.7	41.1
HybrIK [Li et al. 2021]	54.4	34.5
Graphormer [Lin et al. 2021b]	51.2	34.5
PyMAF-X (Res50) [Zhang et al. 2023b]	58.8	40.2
PyMAF-X (HR48) [Zhang et al. 2023b]	54.2	37.2
HMR2 [Goel et al. 2023]	**45.3**	33.4
HybrIK-X [Li et al. 2023]	47.0	**29.8**
Ours	56.2	40.1

Qualitative results of our method for face reconstruction on the NoW benchmark [Sanyal et al. 2019] are shown in Figure 18. Qualitative results of our method for hand reconstruction on the FreiHAND dataset [Zimmermann et al. 2019], including some failure cases, are shown in Figure 19.

I ABLATIONS

In this section we present and discuss a number of ablation experiments to verify the different components of our approach.

I.1 Initialization and Optimization

Our method uses direct regression by DNNs to initialize an optimization process, while many prior works use *either* DNNs to regress pose and/or shape, or optimization without dynamic initialization to reconstruct the human.

In Table 8, Table 9 and Table 10 we compare the performance without DNN initialization to the full method. When not using DNN initialization we initialize with T-pose and template shape. In the single-view case (Table 8, Table 9), for the same number of iterations, we see significantly worse results, demonstrating that the initialization is critical to enable to optimizer to find a good minimum. Results are far less stable in the absence of initialization, drastically affecting the robustness of the method. In the multi-view case (Table 10), the impact of initialization seems to be minor, likely because there is a great deal more data informing the optimization and therefore much less ambiguity for the optimizer to deal with.

In Table 8 and Table 9 we compare the performance of the DNN only (i.e., the method without optimization) to the full method. The impact of optimization is generally small as the DNNs are able to directly produce quite accurate results. The models are robust to diverse poses and image conditions, and appear to have learned a strong prior over both body pose and shape. The optimization serves to fine-tune the precision of the results in terms of image-space alignment, fine-grained pose and shape detail and temporal stability. This makes a large difference to perceived quality of the results, though this is not clearly expressed in the metrics used, see Figure 20.

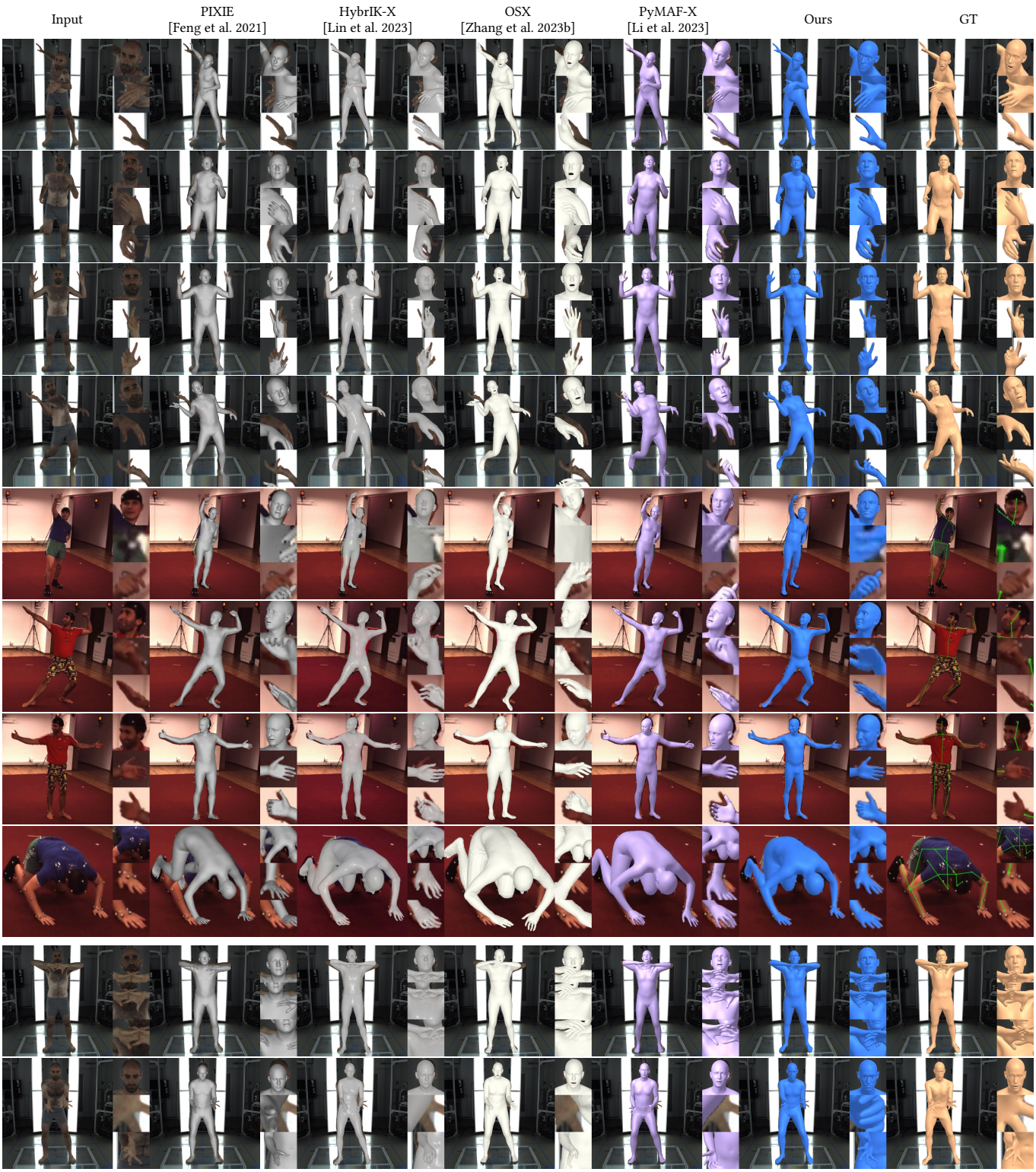


Fig. 17. Further qualitative comparisons for full-body reconstruction on the EHF [Pavlakos et al. 2019] and Human3.6M [Ionescu et al. 2014] datasets. Failure cases are shown in the bottom two rows; our method can perform poorly for interlocking hand poses, and in cases where the hand network fails, here due to a bad ROI. *EHF © MPI, Human3.6M © IMAR*

Table 8. Ablations on the EHF dataset [Pavlakos et al. 2019]. MPVPE is pelvis aligned.

Method	MPVPE			PA-MPVPE				PA-MPJPE	
	Full-body	Hands	Face	Full-body	Body	Hands	Face	Body	Hands
No DNN init.	105.5	60.5	20.1	100.1	63.4	16.6	5.2	95.0	16.5
No optimization	61.7	56.0	**17.9**	49.3	42.6	10.4	5.9	52.5	10.2
No prob. landmarks	67.4	25.6	18.2	37.0	34.9	11.8	5.2	40.7	11.7
Full	**51.0**	**21.6**	18.4	**34.2**	**33.0**	**9.8**	**4.9**	**38.0**	**9.7**



Fig. 18. Qualitative results for single-view face reconstruction on images from the NoW benchmark [Sanyal et al. 2019]. *NoW © MPI.*

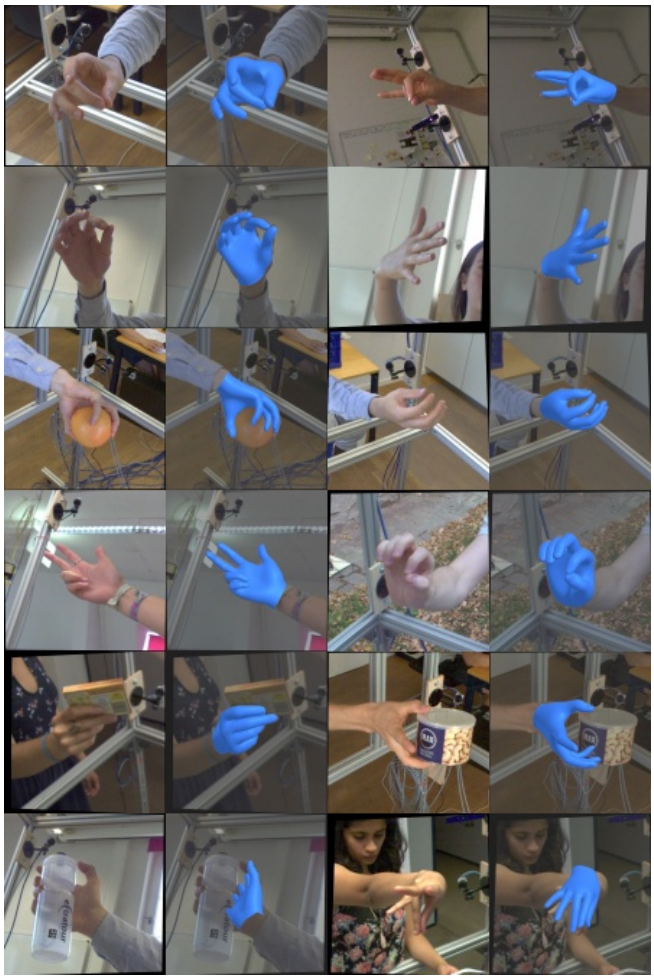


Fig. 19. Qualitative results for hand reconstruction on images from the FreiHAND dataset [Zimmermann et al. 2019]. Our method fails for some cases of significant occlusion and challenging poses (bottom row). *FreiHAND © LMB, University of Freiburg*

I.2 Hand and Face DNNs

Our method uses dedicated DNNs to refine the face and hand predictions. Figure 20 shows results when not performing this refinement and just using the body DNN in isolation. This significantly degrades performance for the face and hands.

I.3 Probabilistic Landmarks

We make use of dense probabilistic landmarks in our method to provide additional freedom to the optimizer based on the confidence of the DNN prediction. In Table 8, Table 9 and Table 10 we show the effect of ignoring these confidence predictions and treating all landmarks with uniform certainty during optimization. In all cases this has a negative impact on results, as the optimizer is 'distracted'

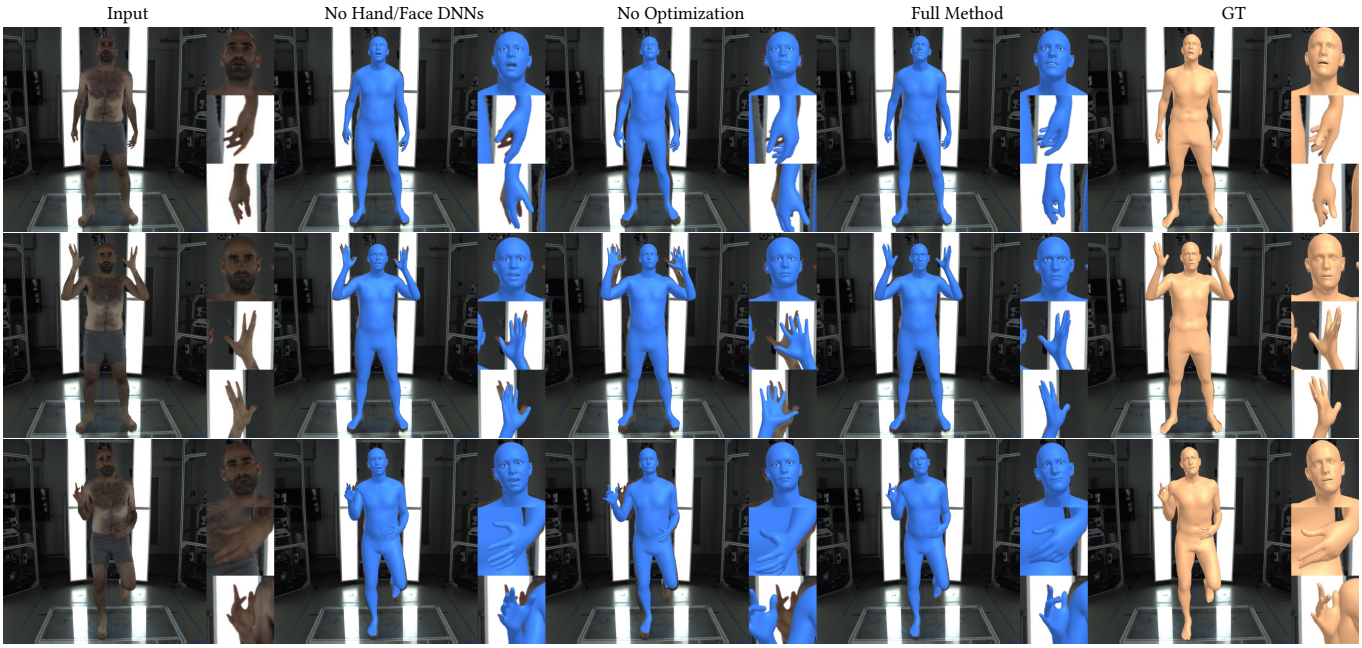


Fig. 20. Comparison of our method without hand or face DNNs, without optimization and with the full method on the EHF dataset [Pavlakos et al. 2019]. Without hand and face DNNs the quality of these areas is drastically reduced, without optimization image alignment is significantly poorer, and the face shape and expression is not captured at all as we do not initialize this using a DNN. *EHF © MPI.*

Table 9. Single-view ablations on the Human3.6M validation set [Ionescu et al. 2014]. MPJPE is pelvis aligned.

Single-view Method	MPJPE	PA-MPJPE
No DNN init.	92.8	78.6
No optimization	66.6	47.1
No prob. landmarks	59.6	41.4
No pose prior	56.5	40.6
No identity prior	56.8	41.0
No temporal prior	57.2	41.0
Full	56.2	40.1

Table 10. Multi-view ablations on the Human3.6M validation set [Ionescu et al. 2014] excluding sequences with incorrect ground-truth annotations following Iskakov et al. [2019]. MPJPE is *not* pelvis aligned.

Multi-view Method	MPJPE
No DNN init.	28.7
No prob. landmarks	32.9
No pose prior	27.9
No identity prior	28.1
No temporal prior	27.8
Full	27.9

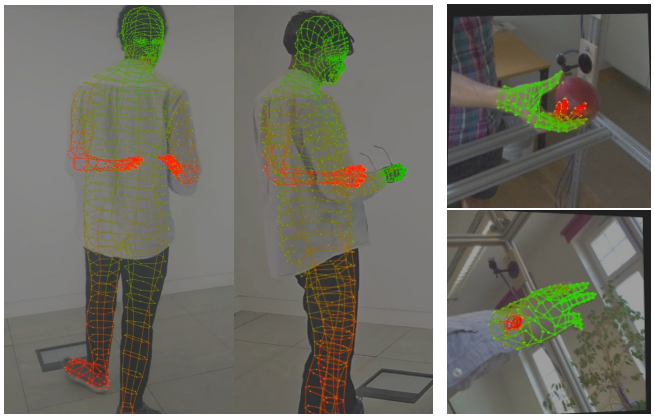


Fig. 21. Predicted dense landmarks for a multi-view, full-body scenario (left) and monocular hand scenario (right), landmarks are colored from least confident (red) to most confident (green). Occluded parts are less confident than visible parts meaning during model-fitting landmarks from views without occlusion (multi-view) or priors can be more heavily relied upon. *FreiHAND © LMB, University of Freiburg.*

by poor landmark predictions and unable to fine-tune for those landmarks that are highly accurate. In the multi-view case (Table 10) the benefit of probabilistic landmarks appears to be greater, allowing the optimizer to use the best available viewpoint, while in the single-view case (Table 9) the benefit is less significant.

Figure 21 shows the benefit of probabilistic landmarks in dealing with self-occlusions due to variation in body pose. Parts that are occluded can be informed more strongly by observations from other views, or priors, leading to more plausible results in these cases.

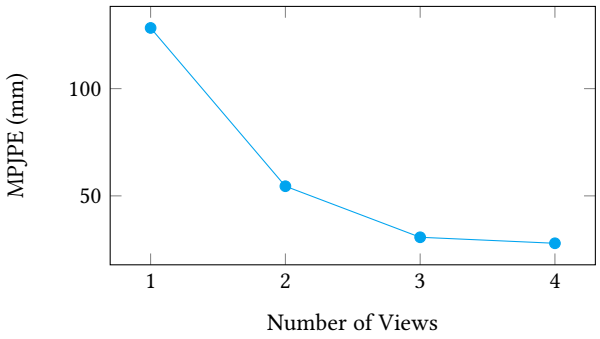


Fig. 22. Variation in error for the Human3.6M benchmark [Ionescu et al. 2014] with different numbers of cameras used as input. The more cameras the better the results, though with diminishing gains. MPJPE is *not* pelvis-aligned.

I.4 Priors

Table 9 and Table 10 show the impact of the pose, shape and temporal priors on registration quality in monocular and multi-view scenarios. In the monocular case all of these priors have a small beneficial effect on results. In the multi-view case, however, the impact is negligible and in the case of the temporal prior it actually increases the MPJPE very slightly. This is not unexpected; in the multi-view case there is sufficient data from the dense landmarks to almost completely specify the result, and we need to rely very little on priors. In the monocular case there is often self-occlusion or depth ambiguity where we rely more heavily on the pose prior, or scale ambiguity where we rely more on the shape prior. The temporal prior promotes smoothness of the motion, something that is perceptually important for performance capture, but not reflected at all in the MPJPE metric.

I.5 Number of Views

We use the Human3.6M benchmark [Ionescu et al. 2014] to evaluate the impact of number of views on our method. As our approach can take arbitrary input viewpoints, we run the reconstruction process with subsets of the four views available in the dataset. As shown in Figure 22, results improve with each camera added, although the benefit diminishes with greater numbers of cameras. In the monocular case there is a significant scale/translation ambiguity present, so it is no surprise that world-space (not pelvis-aligned) MPJPE is high in this case. Adding a second view gives a strong cue for scale so drastically improves results. After this each camera can help to reduce the impact of self-occlusions and further refine results, but with far more limited effect. Going beyond four cameras we would expect the impact to further diminish as it is largely redundant information that is being added to the objective function.

REFERENCES

Sadegh Aliakbarian, Pashmina Cameron, Federica Bogo, Andrew Fitzgibbon, and Thomas J Cashman. 2022. Flag: Flow-based 3d avatar generation from sparse observations. In *CVPR*. 13253–13262.

Brett Allen, Brian Curless, and Zoran Popović. 2003. The space of human body shapes: reconstruction and parameterization from range scans. *ACM TOG* 22, 3 (2003), 587–594.

Saeed Ranjbar Alvar and Ivan V Bajić. 2022. Practical noise simulation for rgb images. *arXiv preprint arXiv:2201.12773* (2022).

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: Shape Completion and Animation of People. *ACM TOG* 24 (2005), 408–416.

Ziqian Bai, Zhaopeng Cui, Xiaoming Liu, and Ping Tan. 2021. Riggable 3D Face Reconstruction via In-Network Optimization. In *CVPR*.

Fabien Baradel*, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas*. 2024. Multi-HMR: Multi-Person Whole-Body Human Mesh Recovery in a Single Shot. In *ECCV*.

Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. 2023. BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion. In *CVPR*. 8726–8737.

Blender Foundation. 2021. Cycles Renderer. https://www.cycles-renderer.org/.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. 2024. Smpler-x: Scaling up expressive human pose and shape estimation. *NeurIPS* 36 (2024).

Chocofur. 2024. Chocofur Interior Scene. https://store.chocofur.com/scenes-interior.

Rohan Choudhury, Kris M Kitani, and László A Jeni. 2023. TEMPO: Efficient multi-view pose estimation, tracking, and forecasting. In *ICCV*. 14750–14760.

Vasileios Choutas, Federica Bogo, Jingjing Shen, and Julien Valentin. 2022a. Learning to fit morphable models. In *ECCV*. 160–179.

Vasileios Choutas, Lea Müller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. 2022b. Accurate 3D Body Shape Regression using Metric and Semantic Attributes. In *CVPR*. 2718–2728.

CLO Virtual Fashion Inc. 2024. MarvelousDesigner. https://marvelousdesigner.com/.

Paul Debevec. 2006. Image-based lighting. In *ACM SIGGRAPH 2006 Courses*. 4–es.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*. 248–255.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2017. Density Estimation Using Real NVP. In *ICLR*.

Evermotion. 2024. Evermotion. https://evermotion.org/.

Faceware Technologies. 2024. Faceware. https://facewaretech.com/.

Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. 2021. Collaborative Regression of Expressive Bodies using Moderation. In *3DV*.

Martin A Fischler and Robert C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (1981), 381–395.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. 2023. Humans in 4D: Reconstructing and tracking humans with transformers. In *CVPR*. 14783–14794.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI* 36, 7 (2014), 1325–1339.

Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. 2019. Learnable Triangulation of Human Pose. In *ICCV*.

Hanbyul Joo, Tomas Simon, and Yaser Sheikh. 2018. Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies. In *CVPR*. 8320–8329.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2018. End-to-end Recovery of Human Shape and Pose. In *CVPR*. 7122–7131.

Sameh Khamis, Jonathan Taylor, Jamie Shotton, Cem Keskin, Shahram Izadi, and Andrew Fitzgibbon. 2015. Learning an Efficient Model of Hand Shape Variation from Depth Images. In *CVPR*. 2540–2548.

Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. 2019. Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop. In *ICCV*. 2252–2261.

Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. 2021. Probabilistic modeling for human mesh recovery. In *ICCV*. 11605–11614.

Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. 2023. HybrIK-X: Hybrid Analytical-Neural Inverse Kinematics for Whole-body Mesh Recovery. *arXiv preprint arXiv:2304.05690* (2023).

Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. 2021. HybrIK: A Hybrid Analytical-Neural Inverse Kinematics Solution for 3D Human Pose and Shape Estimation. In *CVPR*. 3383–3393.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM TOG* 36, 6 (2017), 194–1.

Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. 2023. One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer. In *CVPR*. 21159–21168.

Kevin Lin, Lijuan Wang, and Zicheng Liu. 2021a. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*. 1954–1963.

Kevin Lin, Lijuan Wang, and Zicheng Liu. 2021b. Mesh Graphormer. In *ICCV*.

Selena Zihan Ling, Nicholas Sharp, and Alec Jacobson. 2022. VectorAdam for Rotation Equivariant Geometry Optimization. *NeurIPS* 35 (2022), 4111–4122.

Dong C Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical programming* 45, 1-3 (1989), 503–528.

Matthew Loper, Naureen Mahmood, and Michael J Black. 2014. MoSh: Motion and Shape Capture from Sparse Markers. *ACM TOG* 33, 6 (2014), 1–13.
Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM TOG* 34, 6 (2015).
Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. 2020. Learning to dress 3d people in generative clothing. In *CVPR*. 6469–6478.
Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *ICCV*.
Manus-Meta. 2024. Manus. https://www.manus-meta.com/.
Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. 2022. Accurate 3D hand pose estimation for whole-body 3D human mesh estimation. In *CVPR*. 2308–2317.
Gyeongsik Moon and Kyoung Mu Lee. 2020. I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image. In *ECCV*. 752–768.
Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mize Mallorie, Philippe Bree, Tomas Simon, Bo Peng, Shubham Garg, Kevyn McPhail, and Takaaki Shiratori. 2023. A Dataset of Relighted 3D Interacting Hands. In *NeurIPS Track on Datasets and Benchmarks*.
Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. 2020. InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image. In *ECCV*.
Move AI. 2024. Move AI. https://www.move.ai/.
Movella. 2024. XSens. https://www.movella.com/products/motion-capture.
Natural Point. 2024. OptiTrack. https://www.optitrack.com/.
Baptiste Nicolet, Alec Jacobson, and Wenzel Jakob. 2021. Large Steps in Inverse Rendering of Geometry. *SIGGRAPH Asia* 40, 6 (Dec. 2021).
Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*. 10975–10985.
Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. 2017. Harvesting multiple views for marker-less 3d human pose annotations. In *CVPR*. 6988–6997.
Poly Haven. 2024. Poly Haven. https://polyhaven.com/hdris.
Aashish Rai, Hiresh Gupta, Ayush Pandey, Francisco Vicente Carrasco, Shingo Jason Takagi, Amaury Aubel, Daeil Kim, Aayush Prakash, and Fernando De la Torre. 2024. Towards realistic generative 3d face models. In *WACV*. 3738–3748.
Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. 2021. HuMoR: 3D Human Motion Model for Robust Pose Estimation. In *ICCV*.
Renderpeople. 2024. Renderpeople. https://renderpeople.com/.
Danilo Rezende and Shakir Mohamed. 2015. Variational Inference with Normalizing Flows. In *ICML*. 1530–1538.
Javier Romero, Dimitrios Tzionas, and Michael J Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM TOG* 36, 6 (2017), 1–17.
Yu Rong, Takaaki Shiratori, and Hanbyul Joo. 2021. FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration. In *ICCVW*.
Russian3DScanner. 2021. Wrap3. https://www.russian3dscanner.com/.
Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J. Black. 2019. Learning to Regress 3D Face Shape and Expression From an Image Without 3D Supervision. In *CVPR*. 7763–7772.
Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. 2020. Synthetic Training for Accurate 3D Human Pose and Shape Estimation in the Wild. In *BMVC*.
Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. 2021a. Hierarchical Kinematic Probability Distributions for 3D Human Shape and Pose Estimation from Images in the Wild. In *ICCV*. 11219–11229.
Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. 2021b. Probabilistic 3D Human Shape and Pose Estimation from Multiple Unconstrained Images in the Wild. In *CVPR*. 16094–16104.
Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. 2024. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *CVPR*. 2070–2080.
StretchSense. 2024. StretchSense. https://stretchsense.com/.
Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. *CVPR* (2019).
Stanislaw Szymanowicz, Virginia Estellers, Tadas Baltrušaitis, and Matthew Johnson. 2022. Photo-Realistic 360° Head Avatars in the Wild. In *ECCV*. 660–667.
Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. 2022. GOAL: Generating 4D Whole-Body Motion for Hand-Object Grasping. In *CVPR*. 13263–13273.
Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. 2020. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*. 581–600.
David Joseph Tan, Thomas Cashman, Jonathan Taylor, Andrew Fitzgibbon, Daniel Tarlow, Sameh Khamis, Shahram Izadi, and Jamie Shotton. 2016. Fits like a glove: Rapid and reliable hand shape personalization. In *CVPR*. 5610–5619.
Ten24 Media. 2024. 3D Scane Store. https://www.3dscanstore.com/.
Vicon Motion Systems. 2024. Vicon. https://www.vicon.com/.
Ross Wightman. 2019. PyTorch Image Models. https://github.com/rwightman/pytorch-image-models.
Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. 2021. Fake it till you make it: face analysis in the wild using synthetic data alone. In *ICCV*. 3681–3691.
Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 2022. 3d face reconstruction with dense landmarks. In *ECCV*. 160–177.
Yipin Yang, Yao Yu, Yu Zhou, Sidan Du, James Davis, and Ruigang Yang. 2014. Semantic parametric reshaping of human body models. In *3DV*, Vol. 2. 41–48.
Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, et al. 2023. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *ICCV*.
Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. 2023b. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *PAMI* (2023).
Tianke Zhang, Xuangeng Chu, Yunfei Liu, Lijian Lin, Zhendong Yang, Zhengzhuo Xu, Chengkun Cao, Fei Yu, Changyin Zhou, Chun Yuan, et al. 2023a. Accurate 3D Face Reconstruction with Facial Component Tokens. In *ICCV*. 9033–9042.
Yan Zhang, Michael J Black, and Siyu Tang. 2021. We are more than our joints: Predicting how 3d bodies move. In *CVPR*. 3372–3382.
Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the Continuity of Rotation Representations in Neural Networks. In *CVPR*. 5745–5753.
Wojciech Zielonka, Timo Bolkart, and Justus Thies. 2022. Towards metrical reconstruction of human faces. In *ECCV*. 250–269.
Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. 2019. FreiHAND: A Dataset for Markerless Capture of Hand Pose and Shape from Single RGB Images. In *ICCV*. 813–822.