Towards Robust 3D Body Mesh Inference of Partially-observed Humans

Xiyi Chen
ETH Zürich
xiychen@ethz.ch

Abstract

With the development of unified body models for expressive representations of pose, shape, and facial expressions of human, previous works have achieved robust and occlusion-aware 3D body mesh inference on human images. However, existing methods only focus on fully-observed human captures. It is unclear if their robustness could remain under partially-observed settings where parts of the body are out of the frame, which could be common for video captures and images from the social media. In this work, we follow the optimization-regression hybrid manner and extensively investigate the effectiveness of various modifications to the optimization pipeline using predictions from regression-based methods. Starting from the SMPLify-X pipeline, we make several important modifications to improve its robustness under partially-observed settings: keypoint blending with confidence calibration, thresholding on confidence values, and adopting a stronger body prior pose from the predictions of regression methods. We test our approach as well as the state-of-the-art methods on the upper-body crops of a benchmark dataset with pseudo ground-truth 3D mesh. Qualitative and quantitative results show that our method outperforms all existing methods with respect to the main body and entire observed mesh, although it comes with a slight compromise on face inference accuracy. Our implementation is modified based on SMPLify-X and is publicly available at `https://github.com/xiyichen/smplify-x-partial`.

1. Introduction

Human pose estimation is a key topic in computer vision that estimates the joint representation of human body pose from an input human image. Previous works have achieved decent accuracy on real-time pose estimation for both 2D [3, 4, 8, 14] and 3D [6, 15, 29] representations. However, only modeling the body skeletons could be insufficient to understand human behaviors. With the help of unified, parametric body models [17, 27, 28, 36], recent works start to tackle the more challenging 3D body mesh inference problem which reconstructs the body structure in a more holistic and expressive way. These methods could be classified into 3 categories: regression-based methods [5, 9, 19, 22, 32] and optimization-based methods [1, 28, 35], and optimization-regression hybrids [18, 23, 24]. Given a human capture, these methods could optimize or infer body model parameters for an accurate body mesh that aligns well with the human's pose, shape, and facial expressions.

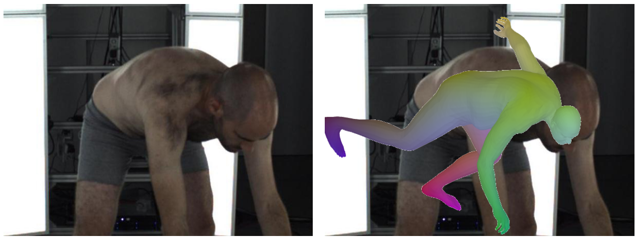

Figure 1. An example of optimization-based methods' inability to fit accurate 3D mesh on partially-observed human images. Left: a partially-observed input from upper-body crops on the benchmark EHF dataset [28]. Right: SMPLify-X fitting results.

Although existing mesh inference methods can perform robustly on fully-observed human images, some are even occlusion-aware [22], their performances have not been evaluated under partially-observed settings where certain parts of the body are out of the frame, which could be common for user inputs. To evaluate the performance under such settings, we test the existing methods on upper-body crops of the benchmark EHF dataset [28] using strict vertex-to-vertex error with Procrustes Alignment. We observe both quantitatively and qualitatively that optimization-based methods completely fail to fit accurate meshes due to their reliance on accurate keypoint detection. With important keypoints missing, corresponding body parts could be fitted to random positions, as shown in Figure 1. Some regression-based methods are able to maintain the robustness on the main body pose, but have problems fitting partially-observed hands and limbs accurately.

In this work, we aim to extensively investigate the effectiveness of various modifications to the state-of-the-art optimization pipeline [28] to improve its robustness under partially-observed settings. As suggested in [5, 21], optimization and regression can form a strong collaboration by initializing the optimization pipeline with predictions from regression networks. Starting with the SMPLify-X [28] pipeline, we make several new contributions: 1) perform keypoints blending with confidence calibration using heuristic statistics from a fully-observed person dataset to improve the accuracy of 2D body joints, and set a threshold empirically to ignore the potentially incorrect keypoints 2) adopt combined body prior poses from the combination of predictions of 2 regression-based methods 3) modify the optimization objective for the body pose prior and penalize the whole body pose towards the combined prior, and fine-tune the optimization weights accordingly. Quantitative and qualitative results show that our method outperforms the state-of-the-art methods in terms of the main body reconstruction of the observed areas, although it is slightly compromised on reconstruction at the face and hands regions. We believe our work is one of the first steps towards more robust 3D body mesh inference of partially-observed humans.

2. Related Works

2.1. From Pose Estimation to 3D Mesh Inference

Human pose estimation is a key topic in computer vision with wide applications in human-computer interaction, autonomous driving, video surveillance, bio-mechanics and medication, etc. [34]. It involves estimating the joint representation of human body pose in 2D or 3D space from an input 2D RGB human image. With the help of deep learning, earlier works have achieved decent accuracy and real-time inference on holistic keypoint detections for body, hands and face in both 2D [3, 4, 8, 14] and 3D [6, 15, 29] space. However, joint representation only gives a rough skeleton of the body and ignores the body shape or facial expressions of the person. While studying human behaviors, people are not only interested in the body pose but also their feelings, appearances, and interactions with the environments. Therefore, the skeleton representations are insufficient and the community moves forward for more expressive representations by reconstructing the 3D mesh that contains pose, shape, and facial expressions from the input human image.

2.2. Unified Body Modeling

To allow more expressive representations of human bodies, unified models that combine accurate modeling on body, hands, and face are first developed. These models either model the entire body parametrically or combine multiple models for different parts of the body. SMPL [27] holistically models shape and pose parameters that could accurately represents a wide variety of body shapes in natural human poses, using a rest pose template, blend weights, pose and identity dependent blend shapes, and a regressor from vertices to joint locations. Despite producing accurate body pose, SMPL has limited expressiveness for hands and face. In an attempt to track body, face, and hands simultaneously, later works builds on or extend SMPL. The Adam model [17] combines SMPL for body, an artist-created rig for hands, and the FaceWarehouse [2] model for face. SMPL+H [31] combines SMPL for body and a 3D hand model. SMPL-X [28] extends on SMPL with fully articulated hands from MANO [31] and expressive face from FLAME [25]. GHUM(L) [36], on the other hand, achieves a small but expressive parameterization of human with non-linear shape and facial expression spaces based on variational autoencoders (VAE) [20], pose-space deformation correctives, skeleton joint center predictors, and blend skinning functions.

2.3. Body Mesh Inference

With accurate pre-trained unified body models, methods are developed to holistically infer parameters that define a 3D body model from an input RGB image. These methods are categoried as regression-based, optimization-based, and optimization-regression hybrids.

2.3.1 Regression-based Methods

Regression-based methods approaches the mesh inference task by predicting body model parameters and camera parameters (usually weak-perspective) through deep learning. Once trained on large-scale datasets, these methods generalize well and are very efficient for inference. HMR [19] gives a canonical example on human mesh recovery. It develops an iterative regression module that infers mesh parameters directly from image features with feedbacks from a discriminator that acts as weak supervision. The SMPL based, occlusion-aware PARE [22] uses a part-guided attention mechanism to exploit information about the visibility of individual body parts and leverage the information from neighboring body parts to predict parts with occlusions. The SMPL-X based ExPose [5] and FrankMocap [32] develop sub-networks for hands, face, and body and merge them to form the entire mesh. PIXIE [9] builds on ExPose's body-driven attention and sub-networks approach, but it evaluates the confidences of each sub-networks and fuses body/face and body/hand features by weighting the confidences.

2.3.2 Optimization-based Methods

Optimization-based methods minimize an objective that contains a 2D keypoint constraint term to fit the body model towards the keypoint detections, prior terms on component parameters to prevent pose, shape, and facial expressions from deviating too much from the templates, and other relevant loss terms, balanced with tunable optimization weights. SMPLify [1] is the first method to estimate both pose and shape from a single human image. Based on SMPL, it fits SMPL pose and shape parameters to 2D body keypoints detected by DeepCut [14]. SMPLify-X [28] fits the more expressive SMPL-X body model to the image and make several significant improvements over SMPLify, including a VAE based pose prior, a more accurate interpenetration penalty, and a gender detector. It also replaces DeepCut with OpenPose [3] for more accurate keypoint detections. MTC [35] proposes an optimization pipeline to fit a deformable human model on 3D Part Orientation Fields and 2D keypoint measurements for total body pose estimation based on Adam model [17] and OpenPose. Although optimization-based methods are robust and accurate when keypoint detections are reliable, the optimization process is generally slow, which makes it difficult to apply them to large-scale datasets.

2.3.3 Optimization-Regression Hybrids

Other works combine regression and optimization collaboratively. ProHMR [24] builds on HMR and regresses a distribution of poses which can then be used as a prior term for a fitting process. SPIN [23] predicts SMPL parameters from a CNN-based deep network and iteratively fit the regressed mesh to OpenPose-detected 2D keypoints for more accurate meshes. The optimized fittings can meanwhile act as a strong supervision for the network. EFT [18] builds on SPIN and HMR and updates the networks weights during fitting. It is also suggested in [5] and [21] that initializing SMPLify(-X) with predicted body pose and camera parameters from predictions of regression methods could produce more accurate fittings with potentially faster convergence. Our approach follows the hybrid manner to initialize the SMPLify-X pipeline with prediction results from ExPose and PIXIE, and achieve more accurate fittings with faster convergence.

2.3.4 Partially-observed Settings

All of the aforementioned methods only target on the robustness when the human captures are fully-observed. None of them have been extended to settings where part of the bodies are invisible. Such partial observations could be common for video captures and images from the social media. It is yet unclear how well the existing methods could perform under such settings quantitatively and qualitatively. [30] proposes a self-training framework that adapts 3D mesh inference systems to consumer videos with unusual camera viewpoints and aggressive truncations. However, their main goal is to get the full-body awareness of the truncations rather than reconstructing the visible parts as accurately as possible, and the quantitative evaluations are only done by human judgements. Therefore, it is necessary to develop new approaches or modify the existing ones for partially-observed settings, and to evaluate the robustness of existing methods under such settings. Due to the lack of suitable training dataset for partially-observed 3D human mesh inference, it could be challenging to adapt regression methods to such settings. Optimization-based methods, however, have the potential to be adapted by modifying the pipeline to resolve the limitations on partial observations. We modify the SMPLify-X optimization scheme and integrate information from the predictions of regression-based methods. In addition, we perform upper-body cropping on the benchmark EHF dataset and demonstrate through quantitative and qualitative evaluations that our method outperform the existing methods on the main body and the whole mesh.

3. Methods

3.1. Keypoints Blending and Thresholding

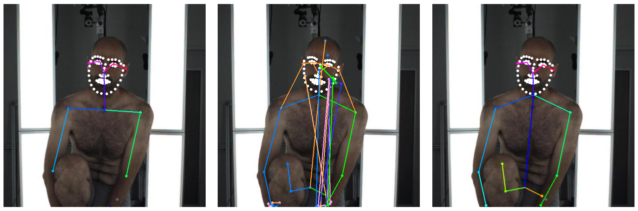

Figure 2. An example of our keypoints blending. Left: OpenPose BODY_25 format model [3] detections. Middle: MMPose Halpe format model [7, 8, 26] detections. Right: blending results.

The performance of optimization based methods are highly sensitive to the accuracy of keypoints detection. Any missing keypoints, especially the keypoints to define the body skeleton, could substantially harm the accuracy of mesh inference. Therefore, we attempt to blend results of multiple keypoint detection methods, considering different models could be better at detecting different parts of the body. However, the major challenge of this blending procedure is that confidence scores are defined differently for different detection methods. To tackle this problem, we need to know how confident the detection methods usually are within a large-scale human dataset of various poses and facial expressions. Therefore, we first run these methods on the SHHQ dataset [10] with 40,000 images, a fully-observed human dataset for fashion poses, to get heuristics on the statistical features of the confidence score for each keypoint. We then perform standardization on the confidence scores using the learned statistical features to calibrate the confidence scores of all methods towards one. For each keypoint, we select the detection with the highest calibrated score. We find it the most beneficial to blend results from two detection models: 1) BODY_25 format model from OpenPose [3] and 2) Halpe full-body format [8, 26] model from MMPose [7] trained on HRNet [33] with DarkPose [37]. Blending is performed only on hand and body keypoints, as we observe that blending face keypoints could sometimes break the face contours and lead to a performance loss (more about this is explained in Section 4.4).

In addition, we observe that some detection methods tend to give incorrect keypoint detections with relatively lower confidence scores. To prevent from optimizing the mesh towards these keypoints, we apply thresholds on confidence scores to ignore them. The thresholds are only applied to the body keypoints since we observe that keypoints with low confidences for face and hands could still be beneficial. We empirically select the threshold as 0.2 for OpenPose and 0.5 for the MMPose Halpe format model.

While the keypoints detection accuracy is worse with OpenPose, we find that its confidence scores are more discrimitive and could be better differentiated with a threshold. Therefore, we perform standardization on the results of the the MMPose Halpe format model and de-standardize them to align them with OpenPose's confidences. For each confidence score c_i of a keypoint detected by the MMPose Halpe format model, its calibrated confidence c_i' is:

$$c_i' = clip(\frac{c_i - \mu_{i,MMPose}}{\sigma_{i,MMPose}} * \sigma_{i,OpenPose} + \mu_{i,OpenPose}), \quad (1)$$

where μ_i and σ_i are heuristic per-keypoint statistics calculated on the SHHQ dataset, and *clip*() clips the calibrated value into a probability range [0, 1]. The blended keypoints follows the BODY_25 format, containing 135 keypoints (25 body keypoints shared by both methods, 21 keypoints for each hand, and 68 keypoints for face) for each human image. Figure 2 shows an example of our blending procedure. The blended results improves upon the results of both of the detectors: adding missing keypoints to OpenPose detections and thresholding incorrect ones from MMPose successfully.

3.2. Combined Body Pose Prior

The missing keypoints of the observed region usually cause SMPLify-X to fit the corresponding body parts to random poses and positions if the neutral pose is used as the pose prior. One potential solution is to adopt a stronger body pose prior to replace the neutral pose. We observe that regression methods could give more accurate fittings on partially-observed human images even for the unobserved parts. In total, 7 methods that regress body model parameters are tested: PARE [22], SPIN [23], ProHMR [24], EFT [18], FrankMocap [32], ExPose [5] and PIXIE [9]. We find that EFT, PARE, PIXIE, and ExPose provide comparably robust inferences on the body pose under partially-observed settings, while the results for SPIN, FrankMocap, and ProHMR are relatively worse. Quantitatively, ExPose produces the most accurate inference on main body and face. PIXIE, on the other hand, produces one of the most accurate hand and wrist poses, while being robust for all other body parts (see Table 1). Therefore, we build a combined body pose prior θ_{b_R} with the first 19 joints from ExPose and last 2 wrist joints from PIXIE.

To improve the speed of convergence and accuracy, we initialize the SMPL-X model with global orientation $\mathcal{G}$ from its prediction with ExPose and body pose θ_b with θ_{b_R}.

3.3. Camera Optimization

The first step of the SMPLify-X optimization pipeline is to estimate the camera parameters to project the SMPL-X 3D joints and vertices into 2D space. We follow [21] to initialize the focal length and the camera translation vector with more accurate approximations. For a point $[x, y, z]$, its 3D-2D projection Π_K is computed as

$$[\tilde{x}, \tilde{y}, \tilde{z}]^T = \Pi_K([x, y, z]) = K \cdot ([x, y, z] + T)^T, \quad (2)$$

where K =

$$\begin{bmatrix} f & 0 & C_x \\ 0 & f & C_y \\ 0 & 0 & 0 \end{bmatrix} \quad (3)$$

is the intrinsic matrix for an input image of size $W \times H$ with bounding box (predicted by RCNN-like networks [12] as a part of the regression methods) of the detected human centered at $[C_x, C_y]$, f is the focal length and is approximated as $f \approx \sqrt{W^2 + H^2}$, and T is the learnable translation parameters. The projected 2D point is then normalized as $[\tilde{x}/\tilde{z}, \tilde{y}/\tilde{z}]$.

We extract the weak-perspective camera parameters $[s, t_x, t_y]$ predicted by the regression method, and initialize T as $T_0 = [t_x, t_y, \frac{2 \cdot f}{s \cdot b}]$, where b is the size of the bounding box.

We also include all keypoints in the upper-body joints (hip joints and above, excluding wrist keypoints) into optimization and incorporate corresponding confidence values. The overall objective for camera initialization is:

$$E(T, \mathcal{G}) = ||(R_\theta(J_u) - J_{est,u}) \odot \omega_u^{\geq\tau}||_2^2 + \lambda_T ||T_z - T_{0_z}||_2^2, \quad (4)$$

where $R_\theta(\cdot)$ is a function that transforms the joints along the kinematic tree according to the pose θ, J_u denotes the upper-body joints of the current SMPL-X model before transformation, element-wisely multiplied with corresponding confidence scores $\omega_u^{\geq\tau}$ (with the ones below the threshold τ set to 0), and $J_{est,u}$ denotes their pseudo ground-truth keypoint detections. λ_T represents the regularization weight for the depth parameter in the translation vector.

3.4. Full Model Optimization

The second step of the pipeline is to optimize the full body model. The overall objective of the optimization pipeline

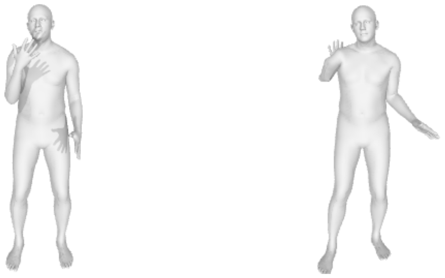

Figure 3. An example of erroneous arm pose after reconstruction with VPoser. Left: pose prediction from regression methods. Right: reconstruction with VPoser.

that optimizes SMPL-X shape parameters β, pose parameters θ, and facial expression parameters ψ follows that of SMPLify-X:

$$E(\beta,\theta,\psi) = E_J + \lambda_{\theta_b} E_{\theta_b} + \lambda_{\theta_f} E_{\theta_f} + \lambda_{m_h} E_{m_h} + \lambda_{\alpha} E_{\alpha} + \lambda_{\beta} E_{\beta} + \lambda_{\varepsilon} E_{\varepsilon} + \lambda_{\mathcal{C}} E_{\mathcal{C}}, \quad (5)$$

The data term E_J is defined as $E_J(\beta, \theta, K, J_{est}) =$

$$\sum_{joint\ i} \gamma_i \omega_i^{\geq \tau_i} \rho(\Pi_K(R_\theta(J(\beta))_i - J_{est,i}), \quad (6)$$

where γ_i denotes the per-joint weight, $\omega_i^{\geq \tau_i}$ denotes the per-joint confidence (set to 0 if below its corresponding threshold τ_i). Note that $\tau_i = 1$ for $\forall\ i \in \{J_{face}, J_{hands}\}$ since no thresholding is applied to face and hands. ρ denotes a robust Geman-McClure error function [11].

The terms E_{θ_f}, E_{m_h}, E_β, E_ε are simple L_2 priors for facial pose, hand pose, body shape, and facial expressions, penalized towards their neutral states. We leave them unmodified instead of penalizing them towards regression predictions because 1) these predictions are sometimes inaccurate and could thus mislead the optimization process; 2) these parts are easier to be optimized than the body pose since they are less susceptible to the missing keypoints of unobserved body parts. $E_\alpha(\theta_b)$ is the angle prior that penalizes extreme bendings for elbows and knees, and $E_{\mathcal{C}}$ is the interpenetration penalty as in SMPLify-X on a list of self-colliding triangles of the body parts that are physically impossible.

E_{θ_b} represents the body pose prior. In SMPLify-X, a Variational Human Body Pose Prior (VPoser) is proposed and E_{θ_b} is defined as a L_2 prior that penalizes its latent space $Z \in \mathbb{R}^{32}$ towards the mean. Although this term works well in the original pipeline when penalized towards the neutral pose, we find that the reconstruction error of VPoser is considerable when applied to pose priors from regression-based methods. Especially, the arm and hand poses could become erroneous when we encode a pose prior and decode it back, as shown in Figure 3. Although this could be corrected if the body pose is penalized towards the neutral pose, we need to penalize the latent vector towards the latent representation of the predicted body pose from the regression methods to guide the optimization process for unobserved parts. In such cases, the incorrectly reconstructed poses are sometimes unable to be corrected and thus could lead to undesirable fitting results. Therefore, we remove the usage of VPoser and directly regularize the 21-joint pose parameter θ_b as:

$$E_{\theta_b} = ||\theta_b - \theta_{b_R}||_2^2 \quad (7)$$

Since the body poses are initialized with and penalized towards more accurate priors, we find that it is now enough to only perform 3 stages of optimization instead of 5 as in the original pipeline, which reduces the run-time by about half. In the first stage, our pipeline aligns the initialized body model more closely with the 2D keypoints. In the next 2 stages, it gradually fits the face and hands. We start with the released optimization weights of the last 3 stages of SMPLify-X with an annealing scheme for λ and apply several changes to λ_{θ_b}, λ_{θ_f}, and λ_β on each stage. The entire set of modified weights is released with our implementation.

4. Experiments

4.1. Cropped EHF Dataset

To evaluate our modified modification pipeline quantitatively, we prepare a partially-observed human dataset with pseudo ground truth 3D mesh scans by cropping the 100 fully-observed human images from the benchmark EHF dataset experimented in [28]. For each image with detected keypoints in BODY_25 format, we center the bounding box at the neck keypoint for x-axis, and the mid-point between the highest keypoint (with smallest y axis) and the mid-hip keypoint for y-axis. We then crop a 800 × 600 area around the center, which roughly crops the upper body.

Since we are mostly interested in how accurate the mesh reconstruction is within the observed area, we use the ground-truth camera parameters of EHF to project all vertices of the pseudo ground truth 3D mesh into 2D space. We then record the indices of vertices within the boundary for each ground-truth mesh and subset vertices with the same indices from our fitting results.

4.2. Evaluation Metrics

4.2.1 Mesh Alignment

To align a fitted 3D mesh with the ground truth, Procrustes Alignment [13] is performed to solve for scale, rotation, and translation. When reporting the errors for different parts of the mesh (whole observed mesh, body, face, and left/right hands), we align them separately.

4.2.2 Mean Per-Joint Position Error (MPJPE)

We first report the mean Euclidean distance between the joints of the fitted mesh and the ground truth, and denote this as PA-MPJPE with Procrustes Alignment. Since the EHF dataset only provides vertices of the ground truth 3D mesh, we use a linear regressor to predict the 14 LSP-common [16] joints from the ground-truth vertices as in [5], [9], and [22], and do the same on the fitted vertices of our modified optimization pipeline and methods to compare. Although a common metric, MPJPE could not capture the error of the entire 3D mesh well due to the sparseness of body joints.

4.2.3 Vertex-to-vertex Error (V2V):

V2V error computes the mean Euclidean distance for all pairs of vertices. It is stricter than MPJPE since it also captures 3D shape errors and unnatural limb rotations. Note that aligning a SMPL based mesh with a SMPL-X based one is only supported for the main body part where SMPL and SMPL-X share the same topology, and it is done by matching a subset of 4410 vertices of the body part shared by both types of models.

4.3. Quantitative and Qualitative Evaluation

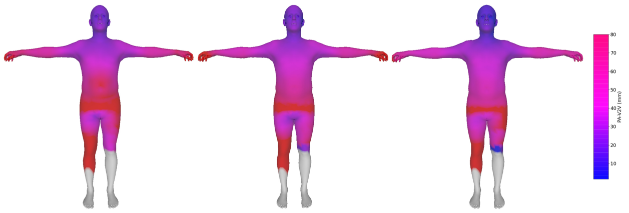

Figure 4. Heatmaps of per-vertex PA-V2V errors averaged on all 100 images in the cropped EHF dataset. Left: PIXIE, Middle: ExPose, Right: ours. The gray parts denote body parts that are never observed in any image. The errors reported are calculated by performing Procrustes Alignment using the entire observed meshes.

Quantitative evaluation is performed on our upper-body crops of the EHF dataset to compare our method with state-of-the-art methods, and the numerical scores can be seen in Table 1. FrankMocap produces one of the most accurate hand models when aligned separately, but its body pose predictions are not very accurate. PIXIE also produces accurate hand meshes while being overall robust for all other body parts, producing the smallest error on the entire observed mesh among all existing methods. Expose produces the smallest errors on body vertices and joints, verifying that it is the most robust method in terms of body pose prediction. Meanwhile, its face inference is also the most accurate among all methods. However, it produces an overall error 1.76 mm higher than PIXIE, mainly caused by the inaccurate wrist poses.

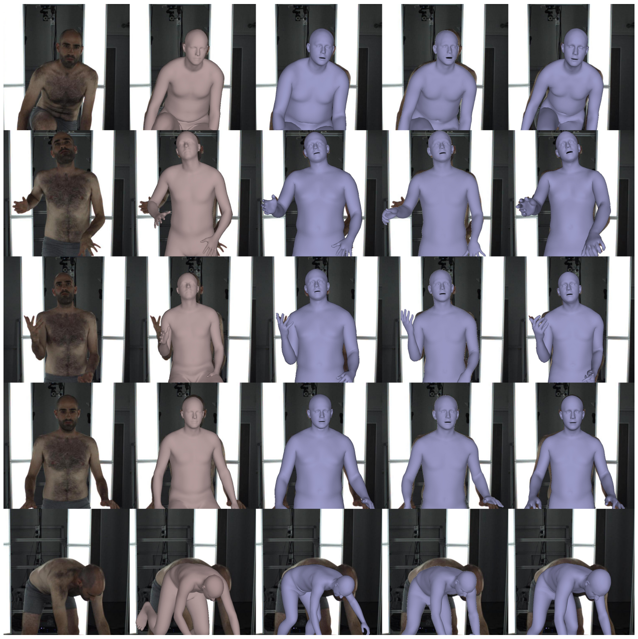

Figure 5. Qualitative evaluation of the 3 best-performing regression methods and our method on the cropped EHF dataset. From left to right: (1) input image, (2) PARE, (3) PIXIE, (4) ExPose, (5) ours. Note that camera projections for the regression methods are only a rough approximation based on the predicted weak-perspective camera parameters.

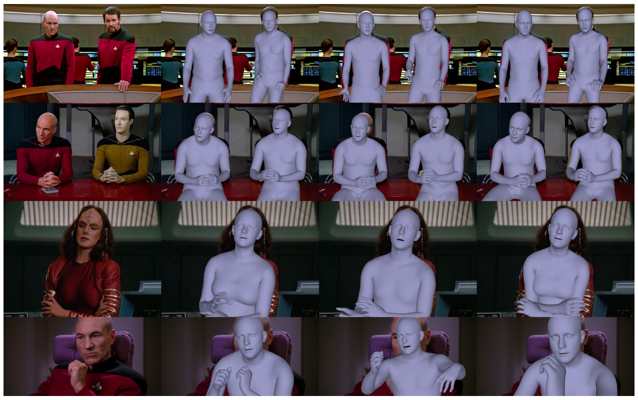

Figure 6. Qualitative evaluation on some screenshots of partially-observed person in *Star Trek: The Next Generation*. From left to right: (1) input image, (2) PIXIE, (3) ExPose, (4) ours. ExPose model is trained with the neutral shape spaces and could only produce gender neutral body models.

Taking advantages of both ExPose and PIXIE, our method produces the most accurate fitting with respect to the body pose and the entire observed mesh. Although we integrate hand poses from PIXIE, our hands fittings still

Method	Type	Body Model	Time (s)	PA-V2V (mm) ↓				PA-MPJPE (mm) ↓
				All	Body	Face	L/R Hand	14 Body Joints
SMPLify-X' [28]	O	SMPL-X	40-60	56.39	68.77	6.25	12.67/13.17	78.56
SMPLify-X [28]	O	SMPL-X	40-60	68.71	82.17	8.76	12.54/13.73	98.71
SPIN [23]	H	SMPL	<1	N/A	60.46	N/A	N/A	74.34
ProHMR [24]	H	SMPL	<1	N/A	52.10	N/A	N/A	60.69
EFT [18]	H	SMPL	<1	N/A	43.41	N/A	N/A	55.16
PARE [22]	R	SMPL	<1	N/A	40.33	N/A	N/A	49.15
FrankMocap [32]	R	SMPL-X	<1	54.59	53.02	5.50	**10.69**/11.79	66.61
PIXIE [9]	R	SMPL-X	<1	37.56	43.16	5.29	11.25/**10.58**	49.58
ExPose [5]	R	SMPL-X	<1	39.08	39.76	**5.13**	12.85/12.71	45.78
Ours	H	SMPL-X	10-30	**32.78**	**38.03**	7.03	12.23/12.76	**42.26**

Table 1. Quantitative evaluation results on the 100 images in the cropped EHF dataset. Our method outperforms the state of the art regression methods w.r.t main body and the whole 3D mesh, and reduces the runtime by about half compared to the original SMPLify-X pipeline, but is slightly worse w.r.t face and hand performance. Note that SMPLify-X' uses the ground-truth focal length and is not directly comparable with other methods. All error scores are recorded only on the observed parts of the images. "O/R/H" denotes Optimization/Regression/Hybrid.

come out less accurate when aligned separately with the ground truths, caused by errors in keypoint detections. We observe that keypoint detectors are the least robust in case of low degree of observation, extreme or unnatural hand poses, and when hands overlap with other body parts. In such scenarios, all keypoint detectors we have experimented with tend to make similar mistakes by ignoring important keypoints, hence this problem could still not be solved with keypoint blending. On the other hand, our method also produces underfitted face meshes, producing an even higher error in terms of face PA-V2V compared to the original SMPLify-X method with ground-truth focal length. We believe this is a compromise when we replace the latent representation of body pose using VPoser with the actual body pose vector whose size is about twice as big (more of this is discussed in Section 4.4.3).

Figure 4 shows, when aligning the whole meshes with the ground truths, the error distributions of all vertices fitted by the 3 best-performing methods that are based on SMPL-X body models, averaged on all 100 images in the cropped EHF dataset. It can be seen that PIXIE has the highest errors in the main body part, especially at the stomach and hips areas. ExPose performs better on the main body, but produces larger errors on hands. Our method is on par with ExPose in terms of the body, while slightly improved on hips, necks, head poses and hand/wrist poses.

In addition, we show some qualitative results on the cropped EHF dataset as well as some partially-observed human captures on the internet in Figure 5 and 6. Regarding body poses, the predictions with PARE and PIXIE tend to fit the partially-observed limbs inaccurately, bended instead of neutralized when some important keypoints are missing. ExPose gives the most robust body pose predictions, but less accurate wrist poses and hand inference compared to PIXIE. In some cases when hands are only partially visible, ExPose fits them as unnatural poses or even twisted/flipped from front to back (see image 1 and 3 in Figure 6). Our method, optimized on the combined prior, slightly improves the accuracy of body pose prediction from ExPose, and fuses the more accurate wrist and hand poses from PIXIE.

4.4. Ablation Studies

Version	PA-V2V (mm) ↓			
	All	Body	Face	L/R Hand
Ours	**32.78**	**38.03**	7.03	12.23/12.76
Keypoints Blending:				
Blend face keypoints	+0.72	+0.48	+0.06	+0.02/+0.04
Blend body only	+1.12	+0.89	-0.03	+0.20/-0.24
OpenPose only	+1.53	+0.68	-0.22	-0.03/-0.23
MMPose only	+2.79	+3.13	+0.90	+1.95/+2.07
Aligned to MMPose	+1.17	+2.64	+0.03	+1.39/+1.59
Thresholding:				
Without threshold	+0.85	+0.89	±0.00	-0.04/-0.10
Threshold on hands	+0.09	+0.22	-0.01	+0.08/+0.06
Threshold on face	+0.59	+0.90	±0.00	+0.36/-0.03
Body Pose Prior:				
PARE body pose	+3.33	+2.17	+0.29	+1.85/+1.60
PIXIE body pose	+0.85	+0.93	-0.22	**-0.43**/**-0.30**
ExPose body pose	+2.76	+1.18	+0.05	-0.12/+1.10
Use VPoser	+11.07	+11.12	**-0.56**	+0.76/+1.13

Table 2. Ablation studies on cropped EHF dataset using the stricter PA-V2V metric.

Next, we perform ablation studies by modifying one of the components of our proposed method. Table 2 compares each of the modifications with our proposed version using the stricter PA-V2V metric.

4.4.1 Keypoints Blending

We change the blending procedure in two ways: components to blend and calibration target of confidence scores. Not blending at all, i.e., only using OpenPose or MMPose results, increases the error with respect to main body and overall mesh, although using only OpenPose keypoints produces slightly better face and hands inference. Blending face keypoints could mix up the contours of different detectors since the keypoints are too close to each other and result in inaccurate face shapes. Excluding hand keypoints from blending would also lead to slightly higher errors since OpenPose cannot always detect hands accurately. When we calibrate the confidence scores towards MMPose instead of OpenPose, the errors also increase, but not as much as using MMPose keypoints only. It verifies our observation that OpenPose confidence scores are more discriminative. Overall, these results prove that keypoints blending is the most beneficial when blending the body and hand keypoints and calibrate confidence scores towards OpenPose.

4.4.2 Thresholding

Using threshold provides a 0.8+ mm performance gain with respect to main body and whole mesh. However, adding threshold to hands and faces leads to slightly higher errors. Therefore, thresholding is the most beneficial when applied only on the body keypoints.

4.4.3 Body Pose Prior

We replace our combined pose prior with solely using body poses from PARE, PIXIE, and ExPose, the 3 best-performing methods in terms of body pose. We observe that using any of them increases the error of the body mesh. Although only using PIXIE poses gives the best inference performance on hands, there's a greater performance loss on the main body and whole 3D mesh.

We also test the performance of using VPoser as our body pose prior as in the original SMPLify-X pipeline. Since the dimension of the pose vector is changed, we modify the corresponding optimization weights λ_{θ_b} accordingly. As described in Section 3.4, the reconstruction loss of VPoser is sometimes considerable, leading to incorrect decoded poses, especially with respect to arms and hands. Although we try to mitigate this issue by penalizing the latent body pose representation towards the mean in the first 2 stage and towards the encoded prior pose only in the last stage, the fitted arm poses could still sometimes be twisted undesirably. Quantitatively, using VPoser produces the largest errors with respect to main body and whole mesh. However, it is worth noticing that the face inference is the most accurate when using VPoser, since it's easier to fit the body pose as the latent space is about half the size of the full body pose space. Even if we try to modify optimization weights for the body pose prior λ_{θ_b} and facial pose prior λ_{θ_f}, as well as face joint weights γ_{face}, our proposed version still comes with a compromise with respect to face inference accuracy.

5. Conclusion

In this work, we propose several modifications to the SMPLify-X [28] optimization pipeline to make it more robust under partially-observed settings, including keypoints blending with confidence calibration, thresholding on confidence values, initializing the camera parameters and body pose with results of regression-based methods, and replacing the pose prior on latent representation with the actual pose space, penalized towards the combined body pose prior. Qualitative and quantitative evaluation results on the upper-body crops of the benchmark EHF dataset with ground-truth 3D mesh show that the fitting results of our proposed new pipeline are more accurate than state-of-the-art methods in terms of the main body and the entire mesh. Ablation studies show that our proposed version has the best overall performance, although it comes with a slight compromise with the performance of face mesh inference.

6. Future Works

To extensively evaluate the robustness of our method and existing methods under partially-observed settings, a larger, preferably in-the-wild dataset with pseudo ground truth 3D meshes will be needed. A potential one to use is the EgoBody dataset [38]. Although captured indoors, its first-person views in 125 sequences provide a wide variety of motions and facial expressions for research on partially-observed body mesh inference. Evaluation could be performed on different degrees of observation to test how the robustness is effected by the visibility of the body.

In addition, since SMPLify-X optimizes for the body shape only with respect to the keypoint locations, it could only give a rough approximation of the body shape. Future works could potentially incorporate silhouette into the objective and/or incorporate inverse rendering as part of the optimization scheme.

7. Acknowledgements

This work is a master's semester project at ETH Zürich. We thank our supervisor Dr. Sergey Prokudin for his helpful advise throughout the project and Muhammed Kocabas for the fruitful discussions.

References

[1] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*. Springer International Publishing, 2016. 1, 3

[2] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. 2

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1, 2, 3, 4

[4] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded Pyramid Network for Multi-Person Pose Estimation. 2018. 1, 2

[5] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3, 4, 6, 7

[6] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2262–2271, 2019. 1, 2

[7] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 3, 4

[8] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 1, 2, 3, 4

[9] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, 2021. 1, 2, 4, 6, 7

[10] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen-Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *European conference on computer vision (ECCV)*, 2022. 3

[11] Stuart Geman and Donald E. McClure. Statistical methods for tomographic image reconstruction. 1987. 5

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2013. 4

[13] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 5

[14] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schieke. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 2, 3

[15] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[16] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, pages 12.1–12.11. BMVA Press, 2010. doi:10.5244/C.24.12. 6

[17] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8320–8329, Los Alamitos, CA, USA, 2018. IEEE Computer Society. 1, 2, 3

[18] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In *3DV*, 2020. 1, 3, 4, 7

[19] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 1, 2

[20] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 2

[21] Imry Kissos, Lior Fritz, Matan Goldman, Omer Meir, Eduard Oks, and Mark Kliger. Beyond weak perspective for monocular 3d human pose estimation, 2020. 2, 3, 4

[22] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 1, 2, 4, 6, 7

[23] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 1, 3, 4, 7

[24] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021. 1, 3, 4, 7

[25] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6), 2017. 2

[26] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *CVPR*, 2020. 3, 4

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 1, 2

[28] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 5, 7, 8

[29] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[30] Chris Rockwell and David F. Fouhey. Full-body awareness from partial observations. In *ECCV*, 2020. 3

[31] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6), 2017. 2

[32] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021. 1, 2, 4, 7

[33] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[34] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3d human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225, 2021. 2

[35] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 3

[36] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 1, 2

[37] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[38] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision (ECCV)*, 2022. 8