ETH zürich

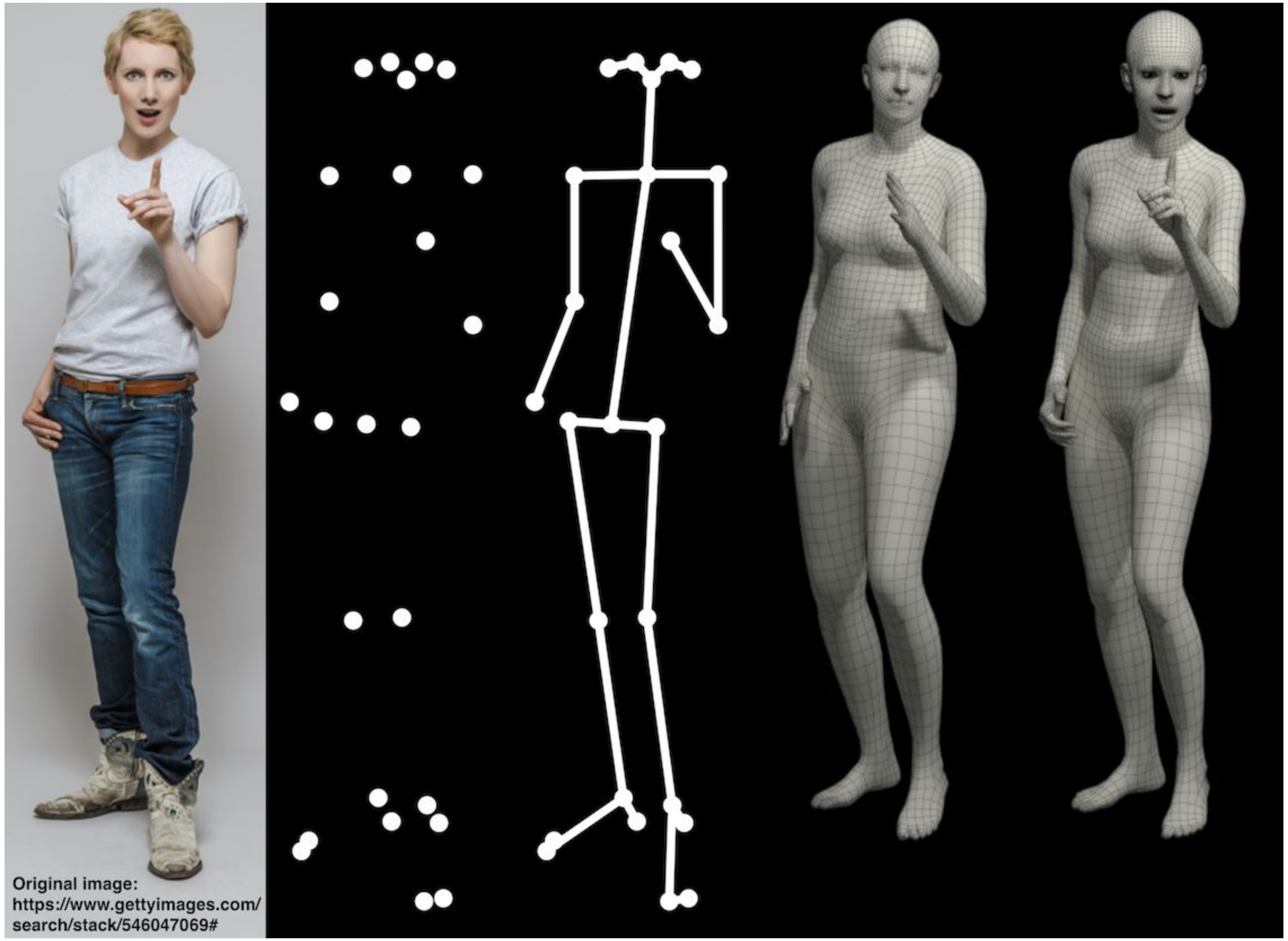
Towards Robust 3D Body Mesh Inference of Partially-observed Humans

Semester project at VLG
Xiyi Chen

Supervisor: Dr. Sergey Prokudin

Introduction

Body modeling



RGB Image | Major joints | Skeleton | SMPL | SMPL-X

Pavlakos et al.,
SMPLify-X, CVPR 2019

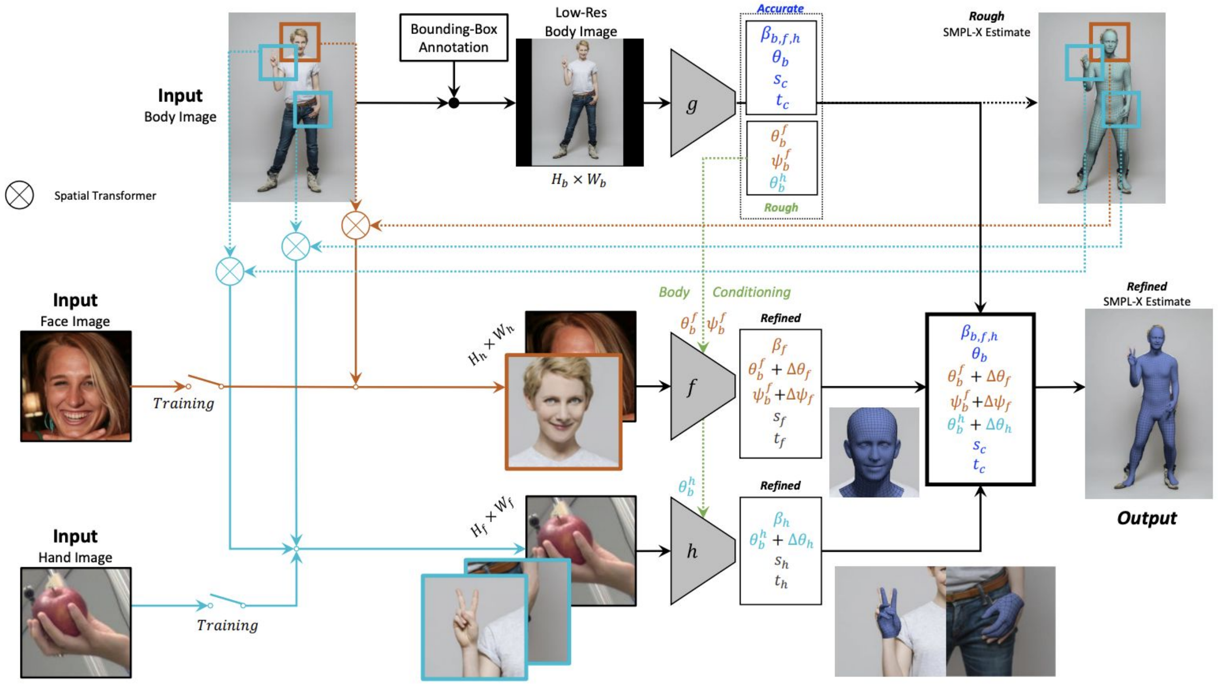
Body mesh inference

Optimization-based methods

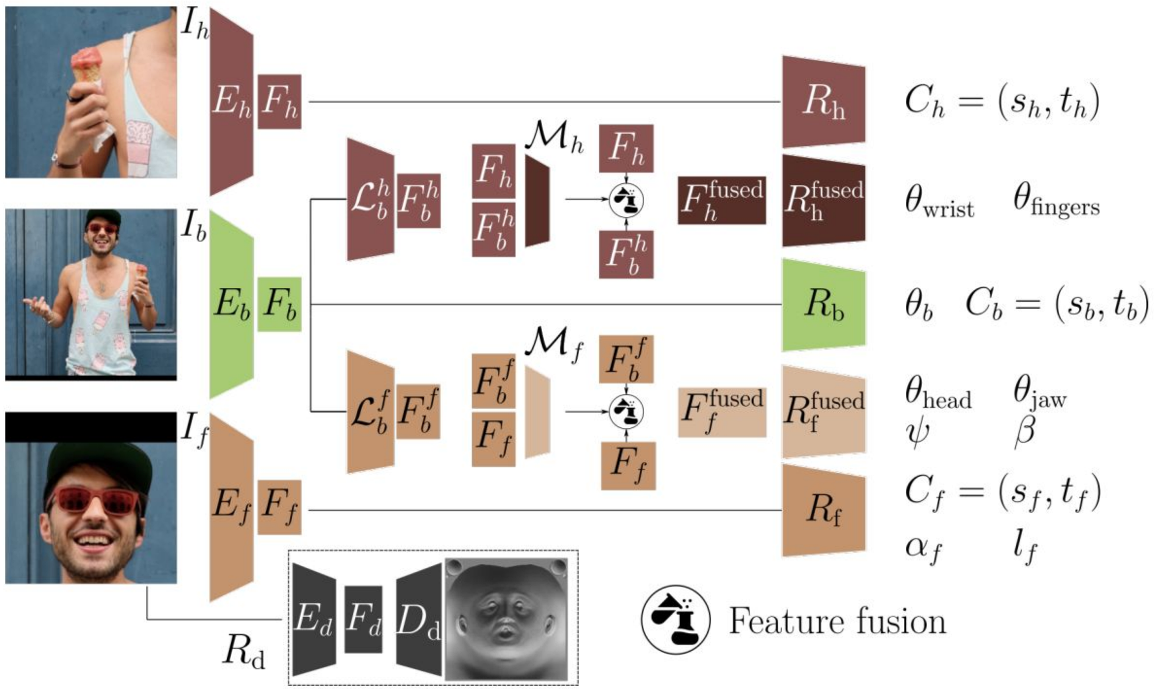
$$E(\beta,\theta,\psi) = E_J + \lambda_{\theta_b} E_{\theta_b} + \lambda_{\theta_f} E_{\theta_f} + \lambda_{m_h} E_{m_h} + \lambda_{\alpha} E_{\alpha} + \lambda_{\beta} E_{\beta} + \lambda_{\mathcal{E}} E_{\mathcal{E}} + \lambda_{\mathcal{C}} E_{\mathcal{C}}$$

Pavlakos et al., SMPLify-X, CVPR 2019

Regression-based methods



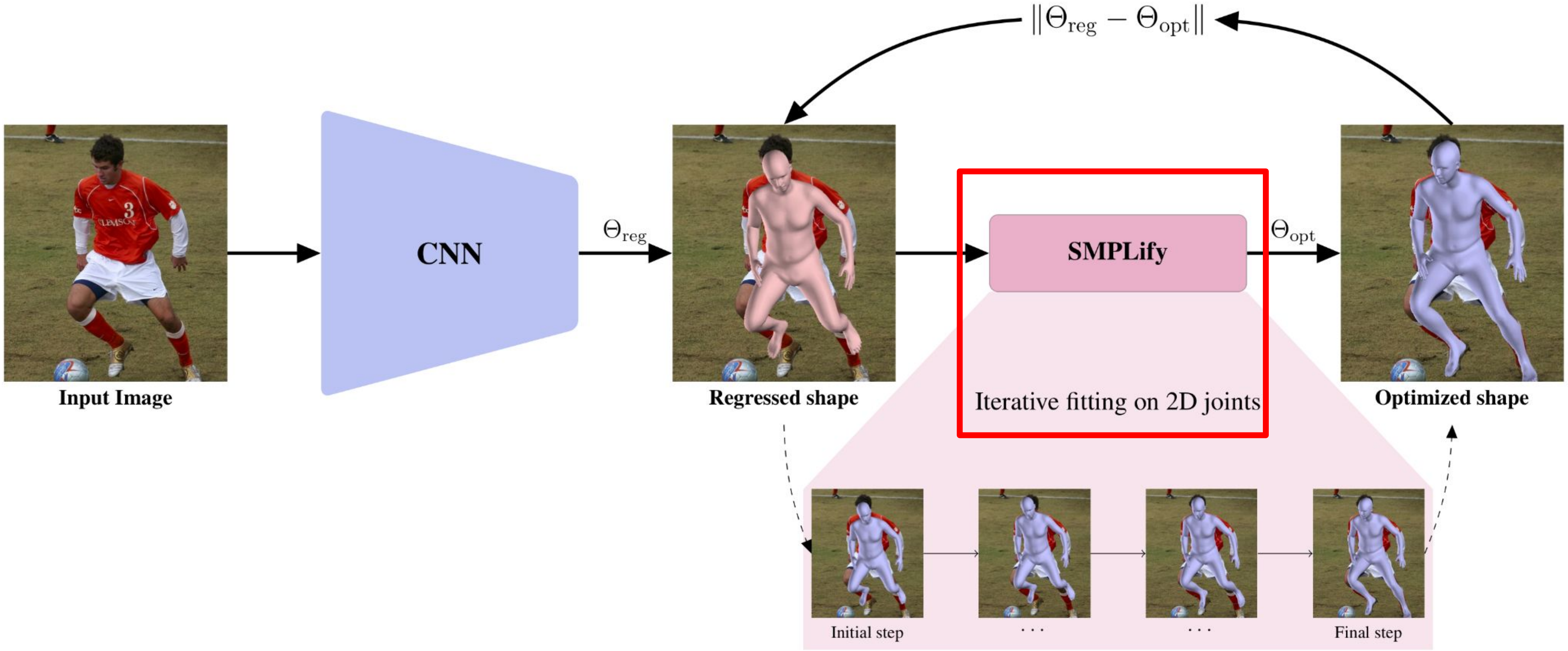
Choutas et al., ExPose, ECCV 2020



Feng et al., PIXIE, 3DV 2021

Body mesh inference

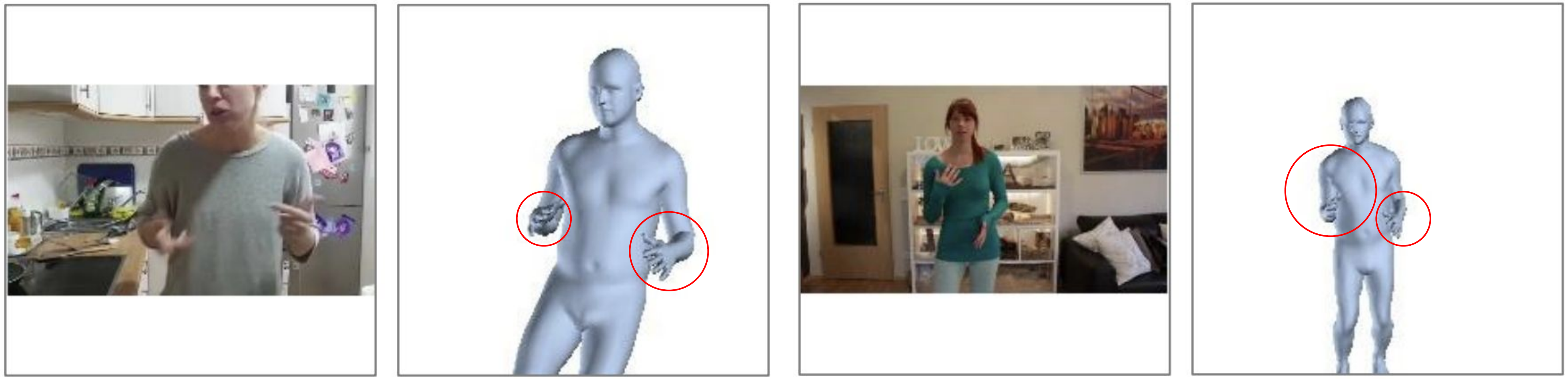
Optimization-regression hybrids



Kolotouros et al., SPIN, ICCV 2019

Partially-observed settings

?



Predictions classified as “Good”, identified by human

Rockwell et al., ECCV 2020

Partially-observed settings

We could start with optimization-based methods like SMPLify-X, but …

Main contributions

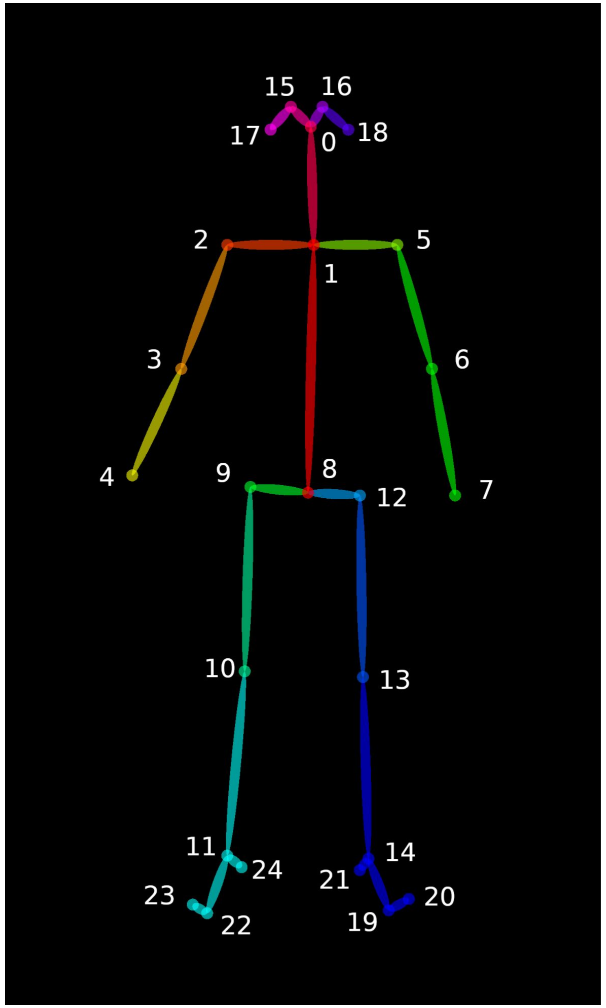
1) Perform keypoints blending with confidence calibration using heuristic statistics from a fully-observed person dataset to improve the accuracy of 2D body joints, and set a threshold empirically to ignore the potentially incorrect keypoints

2) Adopt combined body prior poses from predictions of 2 regression-based methods; initialize global orientation and camera parameters with ExPose predictions

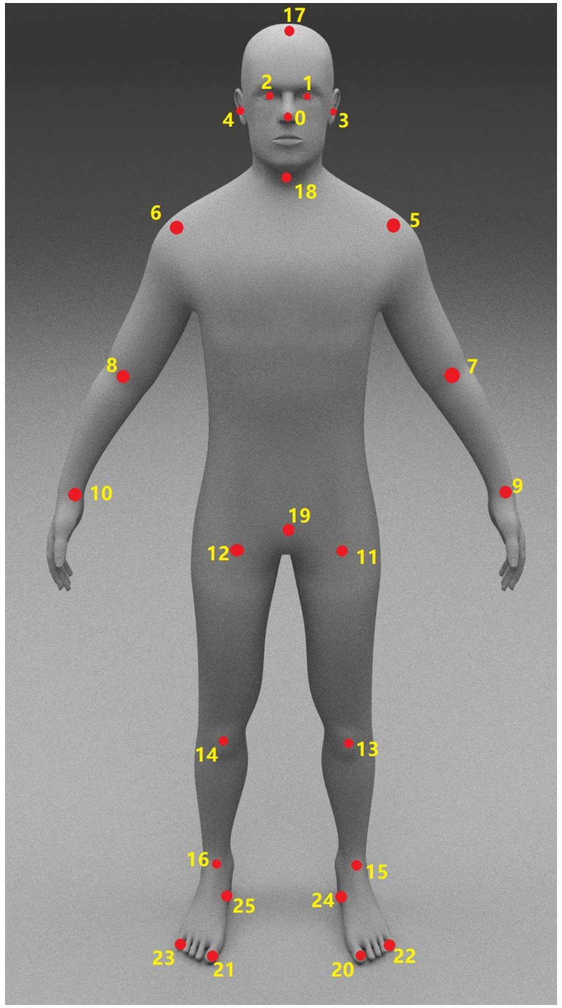
3) Modify the optimization objective for the body pose prior and penalize the whole body pose towards the combined prior, and fine-tune the optimization weights accordingly

Methods

Keypoints blending and thresholding



OpenPose BODY_25 format



MMPose Halpe format

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

Keypoints blending and thresholding

use per-keypoint statistics of 40,000 fully-observed human images with various fashion poses in the SHHQ dataset (Fu et al., ECCV 2022)

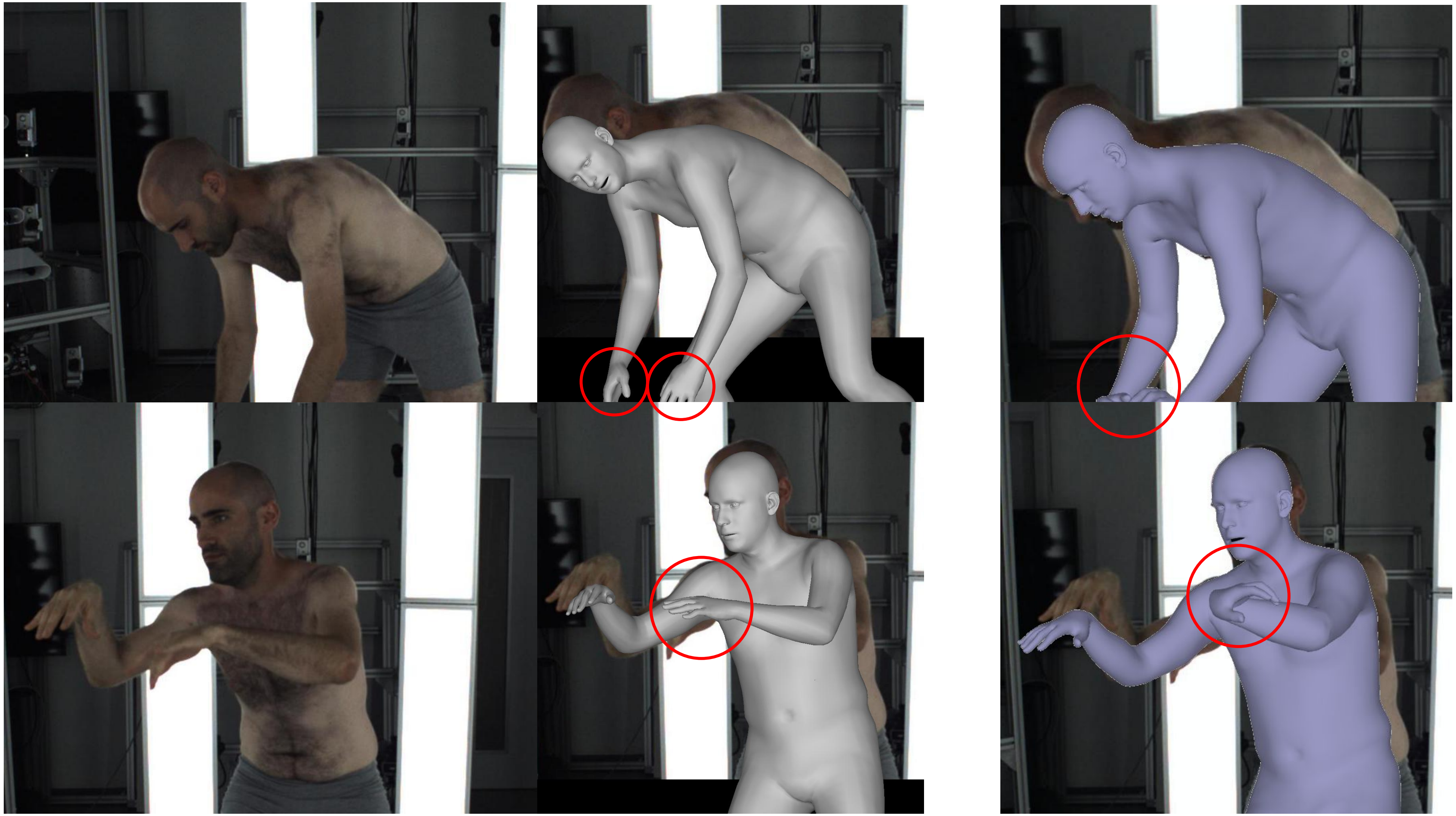
Keypoints blending and thresholding, visualization

OpenPose
(thr=0.2)

MMPose
(thr=0.5)

Blended
(thr=0.2)

Combined pose prior



Input image | PIXIE | ExPose

Combined pose prior

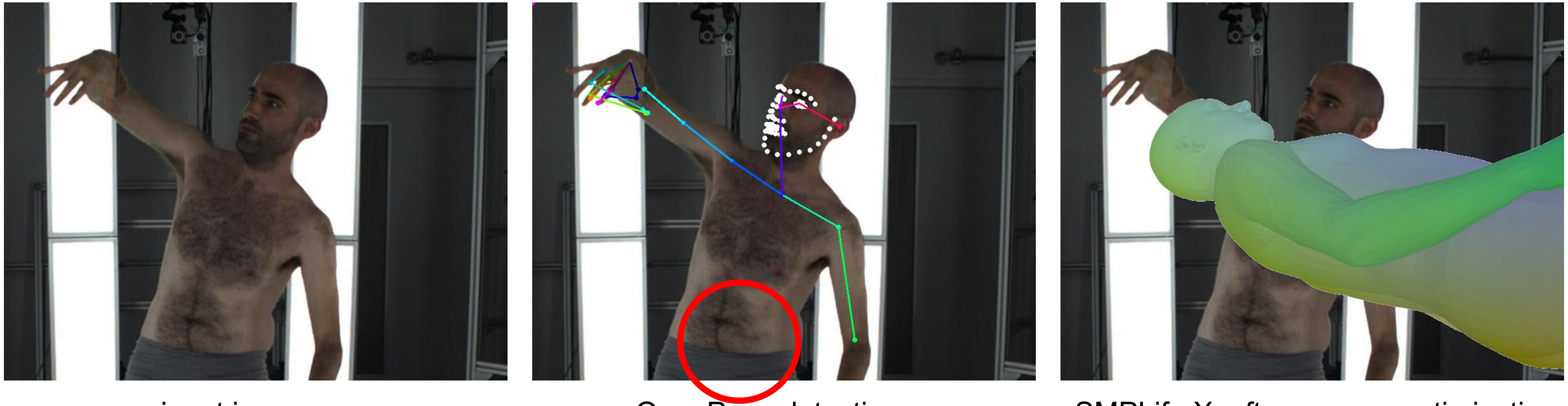
ExPose: the most accurate body pose predictions, but wrist poses are sometimes wrong.
PIXIE: more accurate wrist and hand poses
Therefore, we build a combined pose prior:

$$\theta_{b_R} = \theta_{b_{ExPose}}[:19] + \theta_{b_{PIXIE}}[19:]$$

Camera optimization

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

SMPLify-X starts with the default global orientation and initializes the depth parameter in the translation vector according to shoulder and hip keypoints. However, this is inaccurate and could be especially problematic when these keypoints are missing.



input image | OpenPose detections | SMPLify-X, after camera optimization

Camera optimization

We follow Kissos et al. [1], and initialize the camera matrix and parameters as:

bounding box center [Cx, Cy]

weak-perspective camera parameters[s, tx, ty], predicted by ExPose

$$\mathbf{K} = \begin{bmatrix} f & 0 & \boxed{\begin{matrix} C_x \\ C_y \end{matrix}} \\ 0 & f & \\ 0 & 0 & 0 \end{bmatrix} \qquad f \approx \sqrt{W^2 + H^2} \qquad T_0 = [\boxed{t_x, t_y, \frac{2 \cdot f}{s \cdot \boxed{b}}}]$$

area of bounding box b

We also include all upper-body keypoints and incorporate their corresponding confidence values. The overall camera optimization objective is:

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

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

Camera optimization, visualization

input image

Blended Keypoints

Ours, after camera optimization

Full model optimization

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

$E_J(\beta, \theta, K, J_{est}) = \sum_{joint\ i} \gamma_i \omega_i^{\geq \tau_i} \rho(\Pi_K(R_\theta(J(\beta))_i - J_{est,i})$: data term

E_{θ_f}, E_{m_h}, E_β, E_ε: simple L2 priors for facial pose, hand pose, body shape, and facial expressions

$E_\alpha(\theta_b) = \sum_{i \in (elbows, knees)} e^{\theta_i}$: angle prior that penalizes extreme bendings for elbows and knees

$$E_\mathcal{C}(\theta, \beta) = \sum_{(f_s(\theta), f_t(\theta)) \in \mathcal{C}} \left\{ \sum_{v_s \in f_s} || - \Psi_{f_t}(v_s) n_s ||^2 + \sum_{v_t \in f_t} || - \Psi_{f_s}(v_t) n_t ||^2 \right\}$$, interpenetration penalty

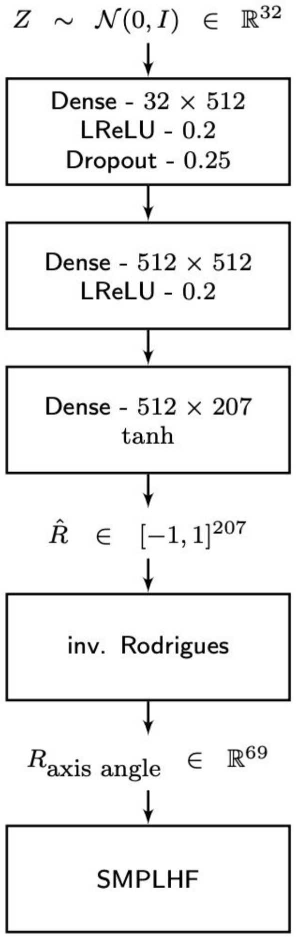
Body pose prior E_β

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

SMPLify-X:

L2 prior on the latent vector of VPoser:

$$||Z||_2^2$$



Input image | body pose prediction by regression methods | reconstruction with VPoser

Problem: erroneous arm pose due to reconstruction loss of VPoser

Ours:

$$E_{\theta_b} = ||\theta_b - \theta_{b_R}||_2^2$$

More accurate pose prior improves the performance and reduces the runtime by half!

Evaluation

Dataset and metrics

Dataset:

- Take upper-body crops of the benchmark EHF dataset
- Use the ground-truth camera parameters to project all vertices of the pseudo ground-truth 3D mesh into 2D space
- Record the indices of vertices within the boundary for each ground-truth mesh and subset vertices with the same indices from our fitting results

Metrics:

Procrustes Alignment on vertices (PA-V2V) and 14 LSP-common joints (PA-MPJPE)

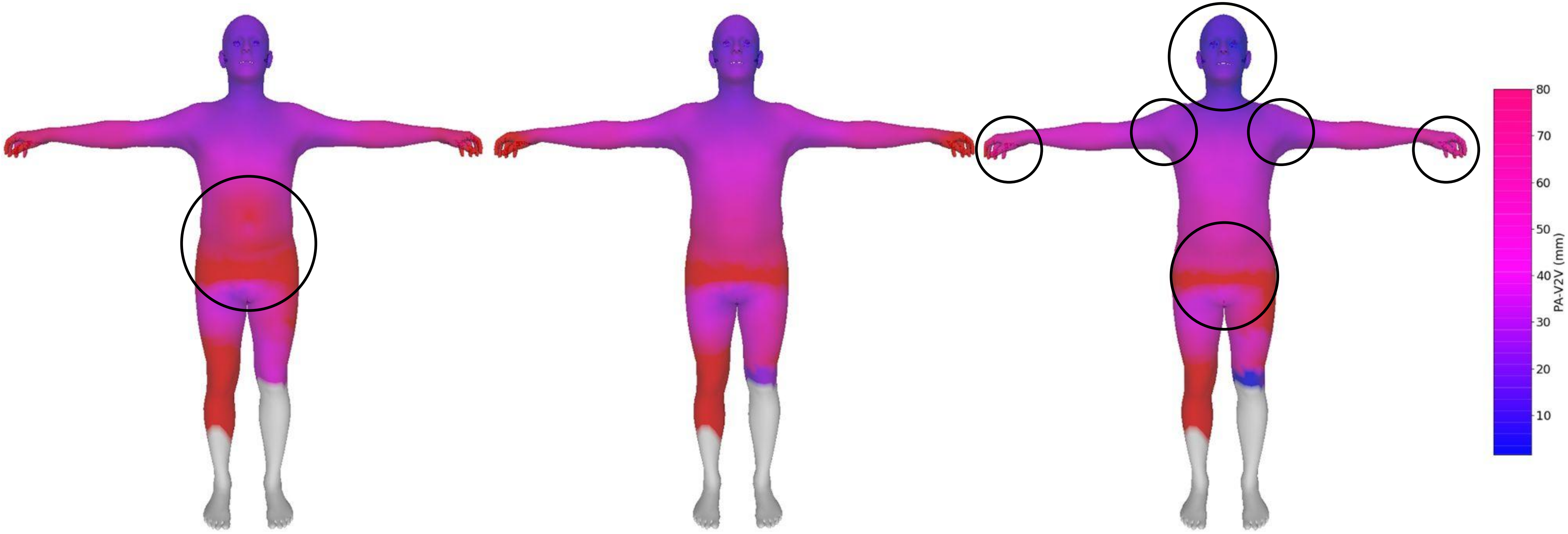
Align the whole observed mesh, body, face, and left/right hands separately, and report each loss

Quantitative evaluation

Method	Type	Body Model	Time (s)	PA-V2V (mm) ↓ All	Body	Face	L/R Hand	PA-MPJPE (mm) ↓ 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 [24]	H	SMPL	<1	N/A	60.46	N/A	N/A	74.34
ProHMR [25]	H	SMPL	<1	N/A	52.10	N/A	N/A	60.69
EFT [17]	H	SMPL	<1	N/A	43.41	N/A	N/A	55.16
PARE [23]	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.

Quantitative evaluation, heatmaps on PA-V2V



PIXIE ExPose Ours

Qualitative evaluation

Input image PARE PIXIE ExPose Ours

ETH zürich

Qualitative evaluation

Input image PIXIE ExPose Ours

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.

Using VPoser

Ours (using the full pose space)

Using VPoser could produces inaccurate arm poses, but it fits the face slightly more accurately

Conclusion

We modify the SMPLify-X pipeline to improve its robustness under partially-observed settings.

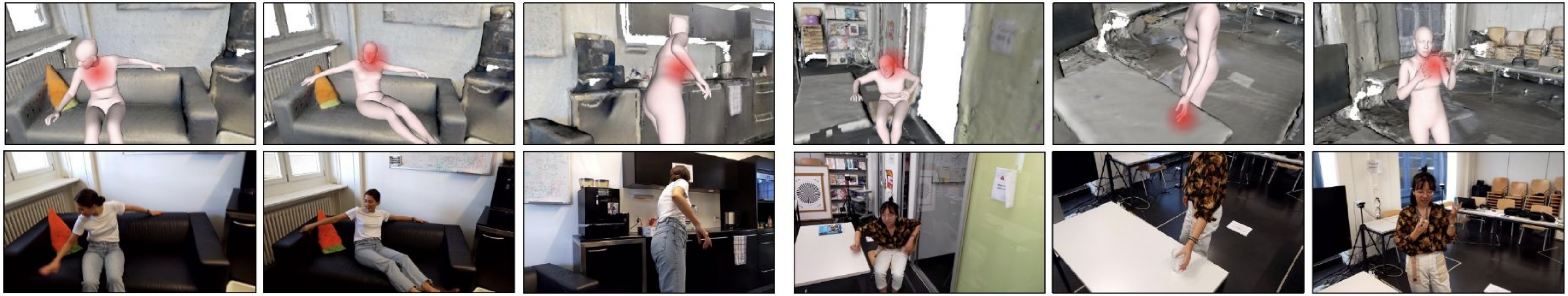
Following the optimization-regression hybrid manner, we make several contributions:

- keypoints blending with confidence calibration, and thresholding on confidence values
- initializing the camera parameters and body pose with results of regression-based methods
- replacing the pose prior on latent representation with the actual pose space and modify the optimization weights accordingly

However,

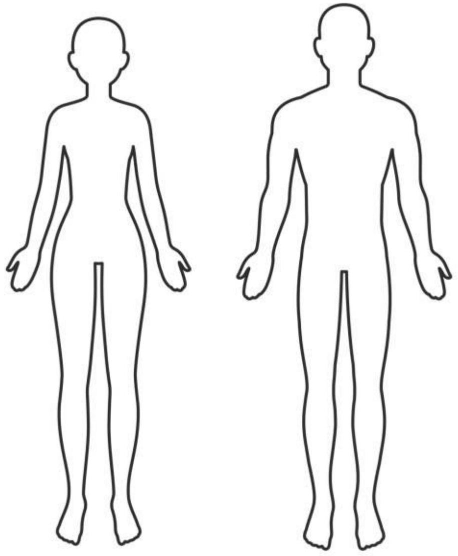
- quantitative evaluation is only performed on 100 images, not extensive enough
- face mesh reconstruction accuracy is compromised
- body shape is only a rough approximation based on 2D keypoints

Future works



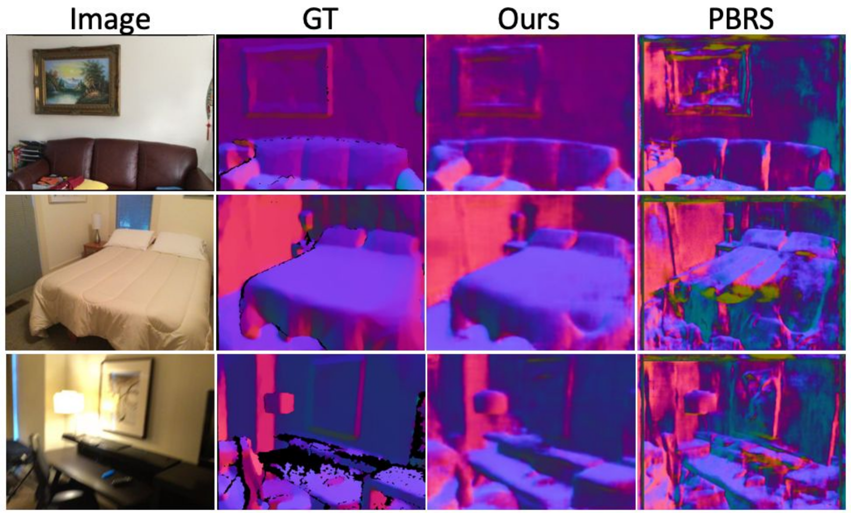
larger partially-observed human dataset

Zhang et al., EgoBody Dataset, ECCV 2022



incorporate silhouette into objective

incorporate inverse rendering into the optimization scheme



Thank you!