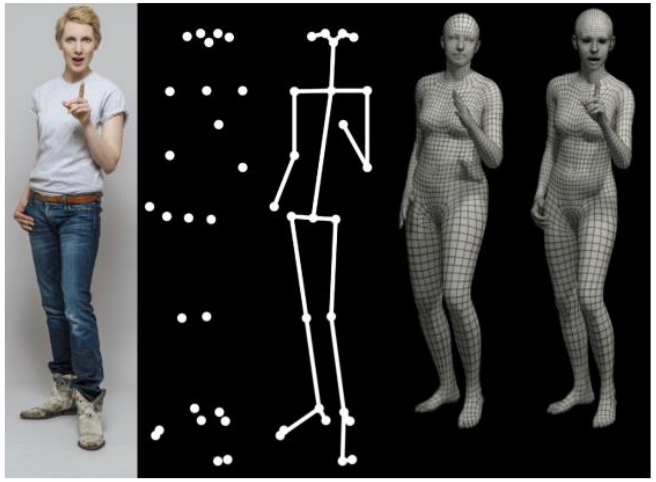
Expressive Body Capture: 3D Hands, Face, and Body from a Single Image

Thai Thanh Tuan, YoungSik Yun

corrected: 27th July 2021

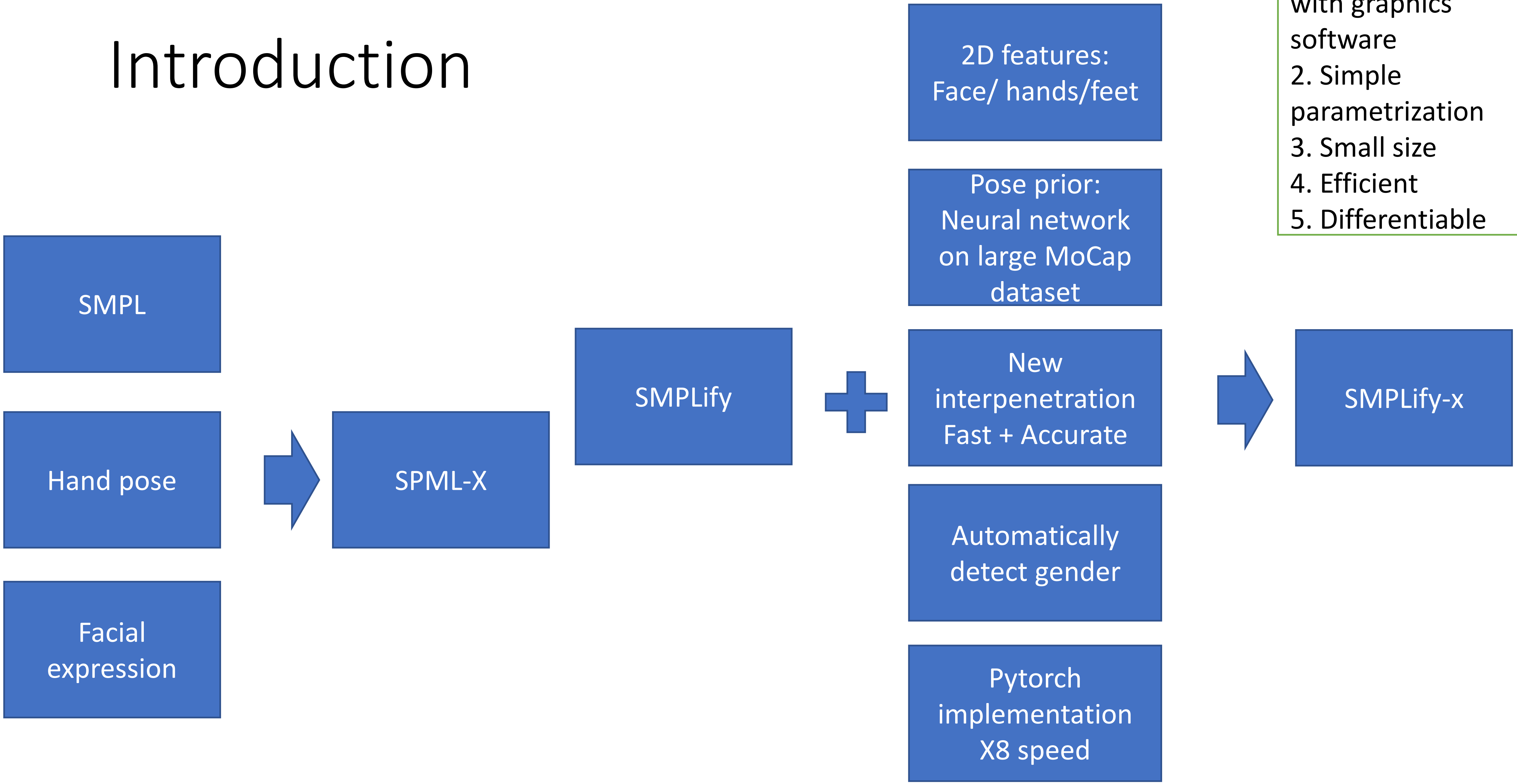


Introduction

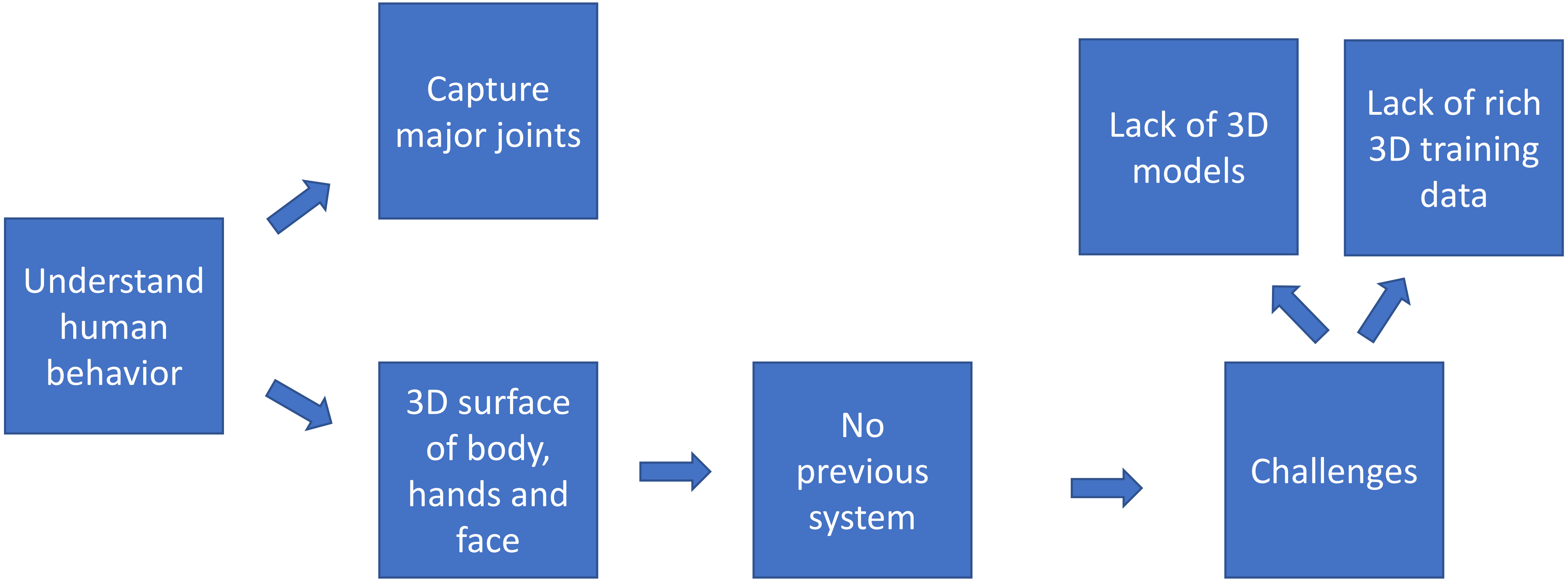
- Code: https://github.com/vchoutas/smplify-x
- Project page: https://smpl-x.is.tue.mpg.de/
- Paper: Expressive Body Capture: 3D Hands, Face, and Body from a Single Image
- Conference: CVPR 2019
- Paper: https://ps.is.tuebingen.mpg.de/uploads_file/attachment/attachment/497/SMPL-X.pdf
- Supplementary: https://ps.is.tuebingen.mpg.de/uploads_file/attachment/attachment/497/SMPL-X.pdf

Introduction

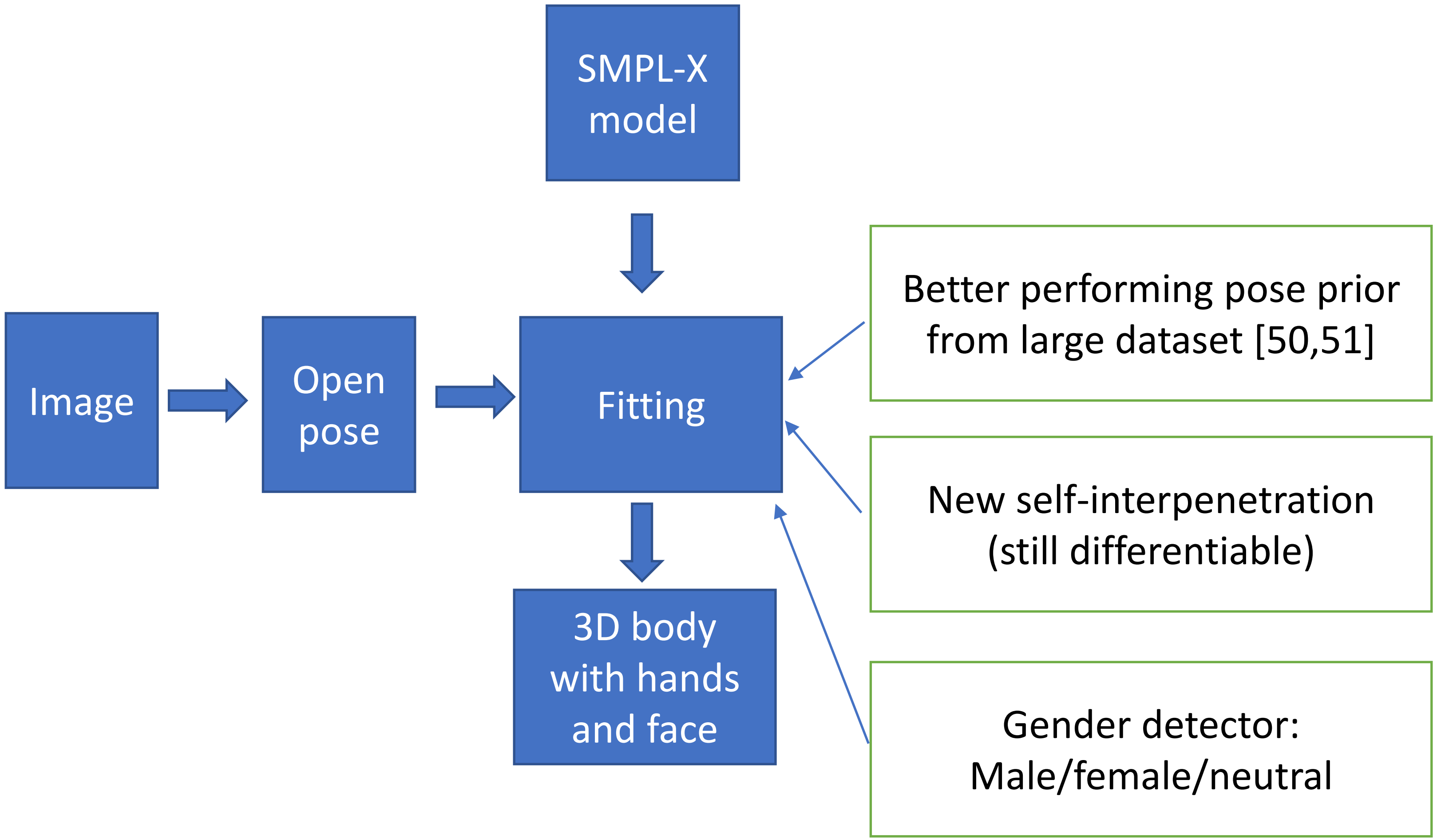
1. Compatibility with graphics software
2. Simple parametrization
3. Small size
4. Efficient
5. Differentiable



Introduction

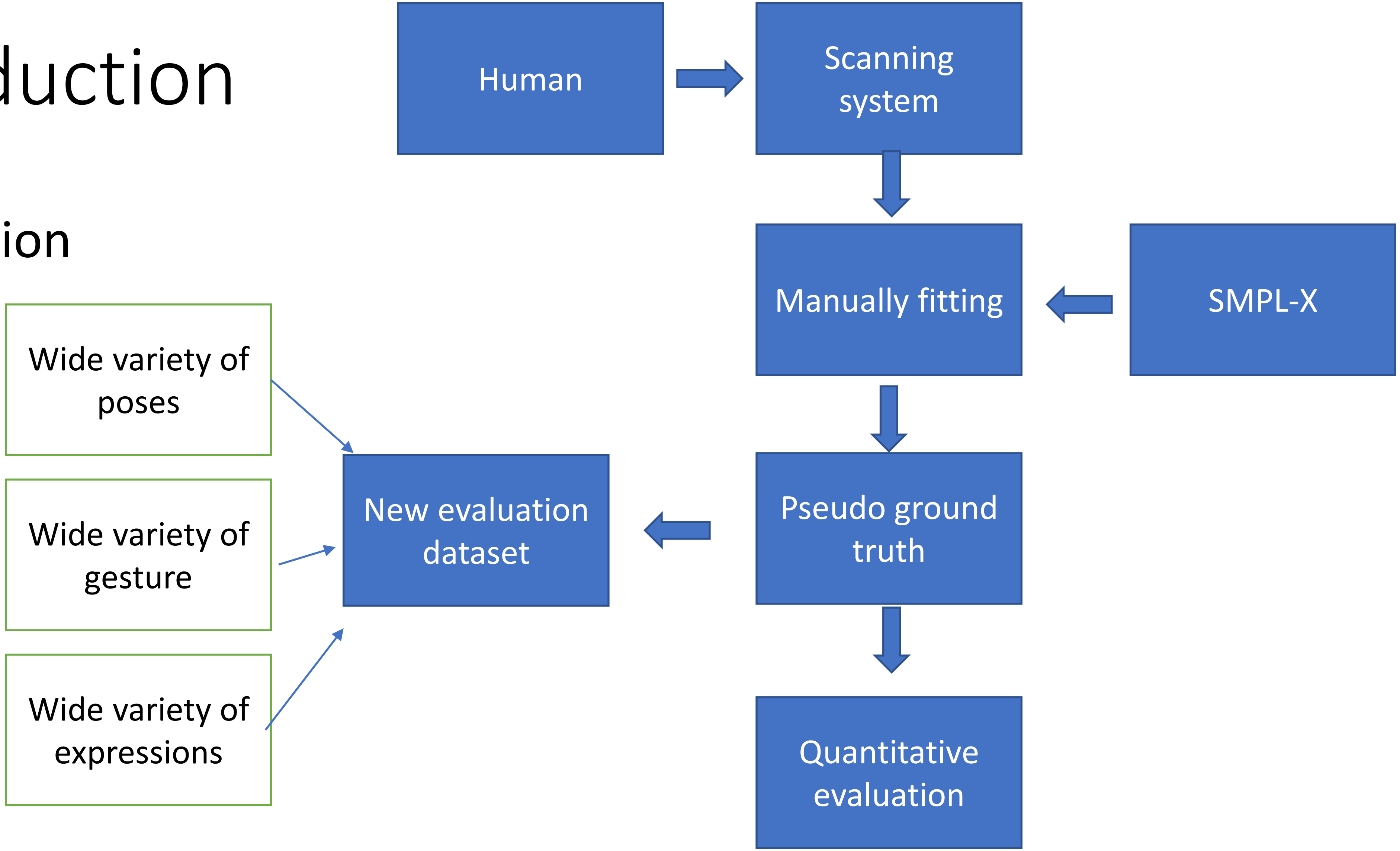


SPMLify-X


SMPL-X model
Image
Open pose
Fitting
3D body with hands and face
Better performing pose prior from large dataset [50,51]
New self-interpenetration (still differentiable)
Gender detector: Male/female/neutral

Introduction

- Evaluation



Related work

- Modeling the body
  - Faces model
    - FLAME [43], models the whole head
    - Captures 3D head rotations
    - Models the neck region
    - No correlations in face shape and body shape
  - Hands model
    - MANO [68]
    - Rich shape and pose space using 3D scans of 31 subjects
    - 51 poses
    - Following SMPL formulation
  - Unified model
    - SMPL+H[68]
  - We start from
    - SMPL+H
    - FLAME
- Inferring the body:
  - Estimate SMPL model from single image: [37,41,59,62]
  - In [36]:
    - Capture environment is complex:
      - 140 VGA cameras for the body
      - 480 VGA cameras for the feet
      - 31 HD cameras for the face and hand keypoints

  ➔ use a single RGB image

Technical approach:

- Unified model: SMPL-X
- SMPLify-X: SMPL-X from a single image
- Variational Human Body Pose Prior
- Collision penalizer
- Deep Gender Classifier
- Optimizatiton

Neck
Jaw
Eyeballs
Fingers

Unified model: SMPL-X

N = 10475 vertices

K = 54 joints

$$M\left(\beta,\theta,\psi\right)=W\left(T_p\left(\beta,\theta,\psi\right),J\left(\beta\right),\theta,\mathcal{W}\right) \tag{1}$$

$$T_P\left(\beta,\theta,\psi\right)=\bar{T}+B_S\left(\beta;\mathcal{S}\right)+B_E\left(\psi;\mathcal{E}\right)+B_P\left(\theta;\mathcal{P}\right) \tag{2}$$

SMPL-X

$M\left(\theta,\beta,\psi\right)$ $\mathbb{R}^{|\theta|\times|\beta|\times|\psi|} \rightarrow \mathbb{R}^{3N}$

$\psi\in\mathbb{R}^{|\bar{\psi}|}$ The facial expression parameters

$\beta\in\mathbb{R}^{|\beta|}$ body, face and hands shape parameters

$\theta \in\mathbb{R}^{3(K+1)}$ θ_f for the jaw joint

θ_h for the finger joints

θ_b for the remaining body joints

K joints + global rotation

Unified model: SMPL-X

the expression blend shape function

$$B_E(\psi;\mathcal{E}) = \sum_{n=1}^{|\psi|} \psi_n \mathcal{E}$$

PCA coefficients

the expression blend shape function

$$M(\beta,\theta,\psi) = W(T_p(\beta,\theta,\psi), J(\beta), \theta, \mathcal{W}) \quad (1)$$

$$T_P(\beta,\theta,\psi) = \bar{T} + B_S(\beta;\mathcal{S}) + B_E(\psi;\mathcal{E}) + B_P(\theta;\mathcal{P}) \quad (2)$$

the shape blend shape function

the pose blend shape function

$$B_S(\beta;\mathcal{S}) = \sum_{n=1}^{|\beta|} \beta_n \mathcal{S}_n$$

β are linear shape coefficients.

$$\mathcal{S} = [S_1, \ldots, S_{|\beta|}] \in \mathbb{R}^{3N\times|\beta|}$$

$\mathcal{S}_n \in \mathbb{R}^{3N}$ Orthonormal principle components of vertex displacements capturing shape variations due to different person identity

θ^* the pose vector of the rest pose

$$B_P(\theta;\mathcal{P}) : \mathbb{R}^{|\theta|} \to \mathbb{R}^{3N}$$

$$B_P(\theta;\mathcal{P}) = \sum_{n=1}^{9K} (R_n(\theta) - R_n(\theta^*))\mathcal{P}_n$$

the n^{th} element of $R(\theta)$

$$R : \mathbb{R}^{|\theta|} \to \mathbb{R}^{9K}$$

$\mathcal{P}_n \in \mathbb{R}^{3N}$ Orthonormal principle components of vertex displacements

$$[P_1, \ldots, P_{9K}] \in \mathbb{R}^{3N\times 9K}$$

blend weights $\mathcal{W} \in \mathbb{R}^{N \times K}$

$$M(\beta, \theta, \psi) = W(T_p(\beta, \theta, \psi), J(\beta), \theta, \mathcal{W}) \quad (1)$$

$$T_P(\beta, \theta, \psi) = \bar{T} + B_S(\beta; \mathcal{S}) + B_E(\psi; \mathcal{E}) + B_P(\theta; \mathcal{P}) \quad (2)$$

3D joint locations

$W(.)$

$$J(\beta) = \mathcal{J}(\bar{T} + B_S(\beta; \mathcal{S}))$$

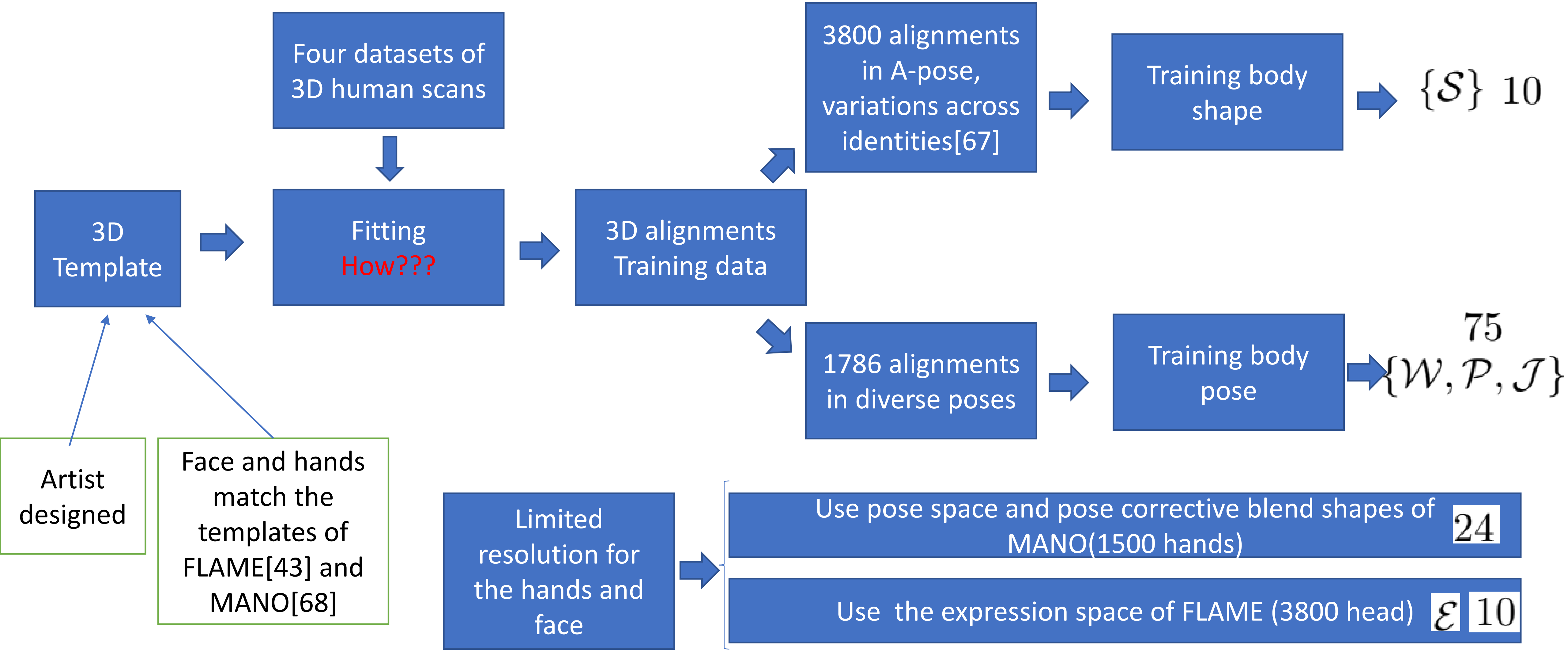
linear blend skinning function

sparse linear regressor

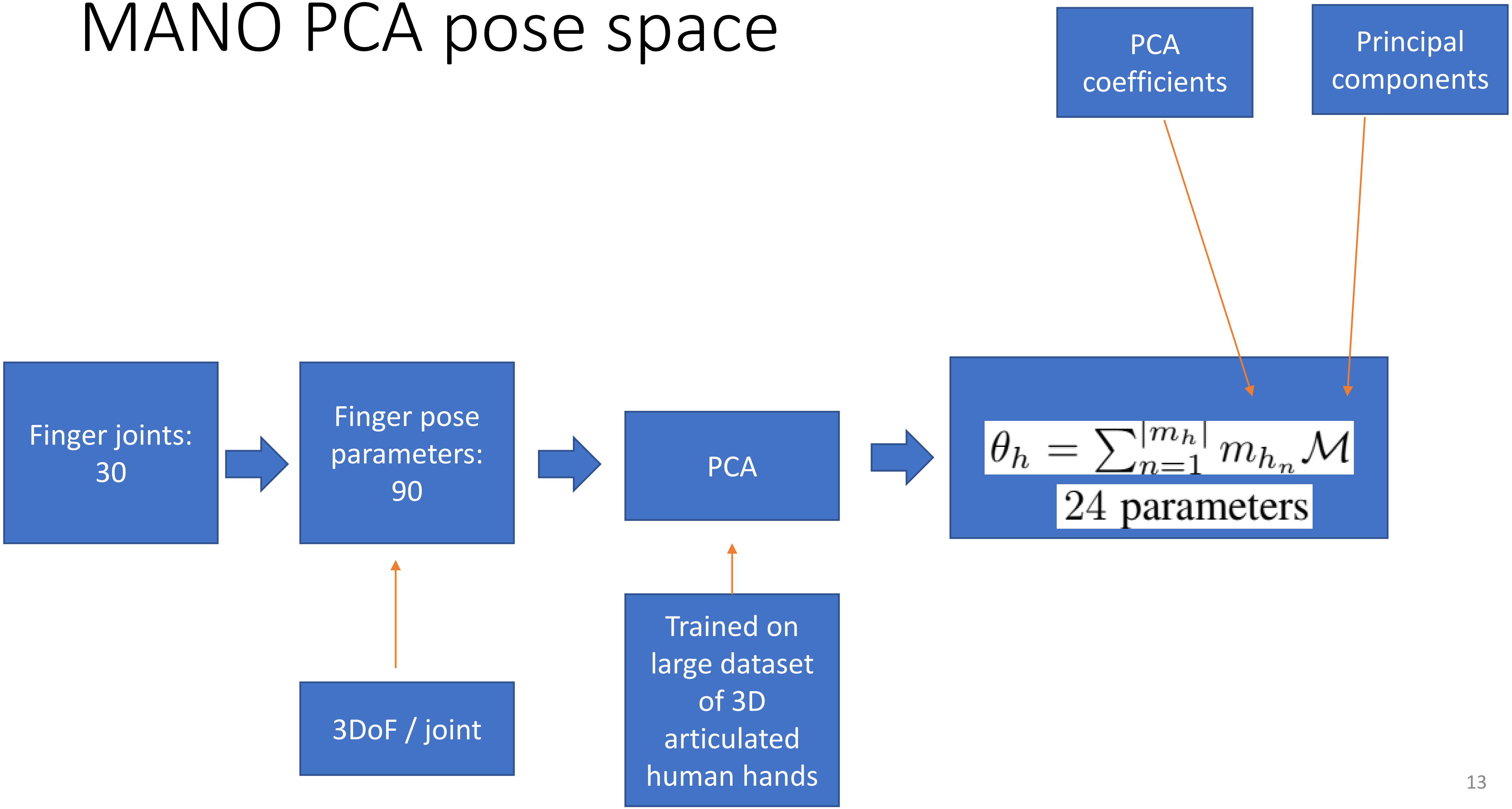
A standard linear blend skinning function $W(.)$ [42] rotates the vertices in $T_p(.)$ around the estimated joints $J(\beta)$ smoothed by blend weights $\mathcal{W} \in \mathbb{R}^{N \times K}$.

SMPL-X

parameters in SMPL-X is 119:



MANO PCA pose space



SMPLify-X: SMPL-X from a single image

the full set of optimizable pose parameters

$E_J(\beta, \theta, K, J_{est})$ is the data term

$E_{\theta_b}(\theta_b)$ VAE-based body pose prior

priors for the facial pose

priors for the hand pose

$$E(\beta, \theta, \psi) = E_J + \lambda_{\theta_b} E_{\theta_b} + \lambda_{\theta_f} E_{\theta_f} + \lambda_{m_h} E_{m_h} + \lambda_\alpha E_\alpha + \lambda_\beta E_\beta + \lambda_\mathcal{E} E_\mathcal{E} + \lambda_\mathcal{C} E_\mathcal{C} \quad (4)$$

priors for for unnatural bending only for elbows and knees.

$E_\alpha(\theta_b) = \sum_{i \in (elbows, knees)} \exp(\theta_i)$

priors for the body shape

$E_\beta(\beta) = \|\beta\|^2$

Facial expressions

$E_\mathcal{C}(\theta_{b,h,f}, \beta)$

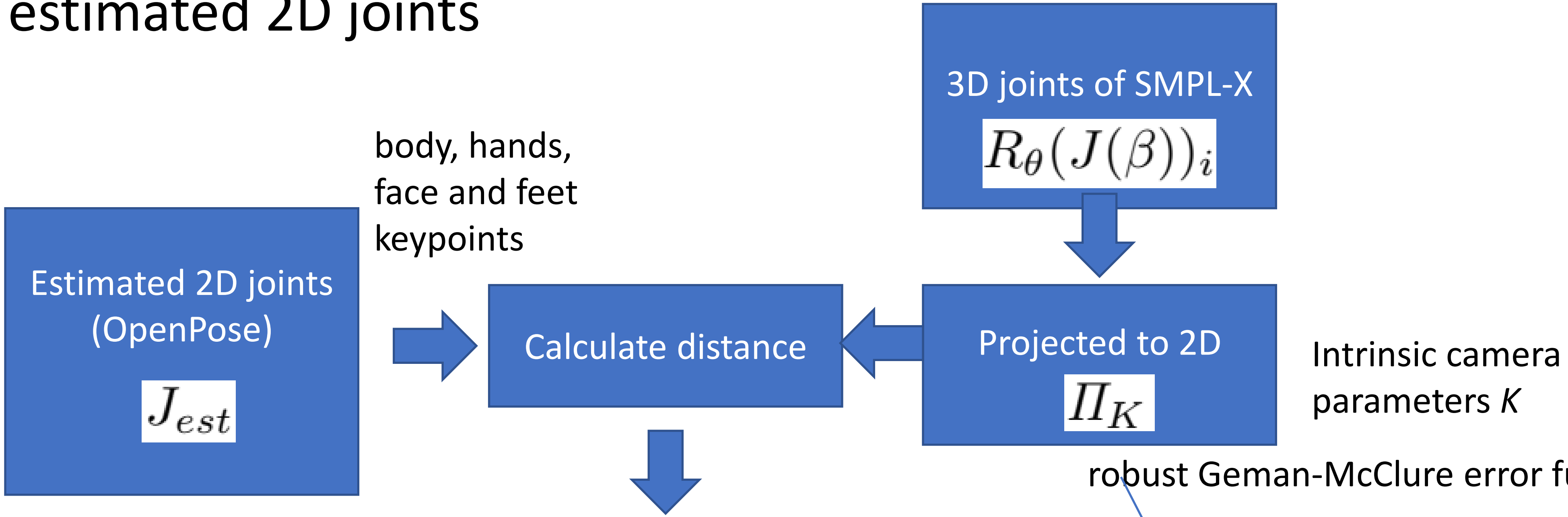
an interpenetration penalty

$\theta_b(Z)$ The body pose parameters are a function $Z \in \mathbb{R}^{32}$

θ_b, θ_f and m_h the pose vectors for the body, face and the two hands

Data term

- Re-projection loss to minimize the weighted robust distance between estimated 2D joints



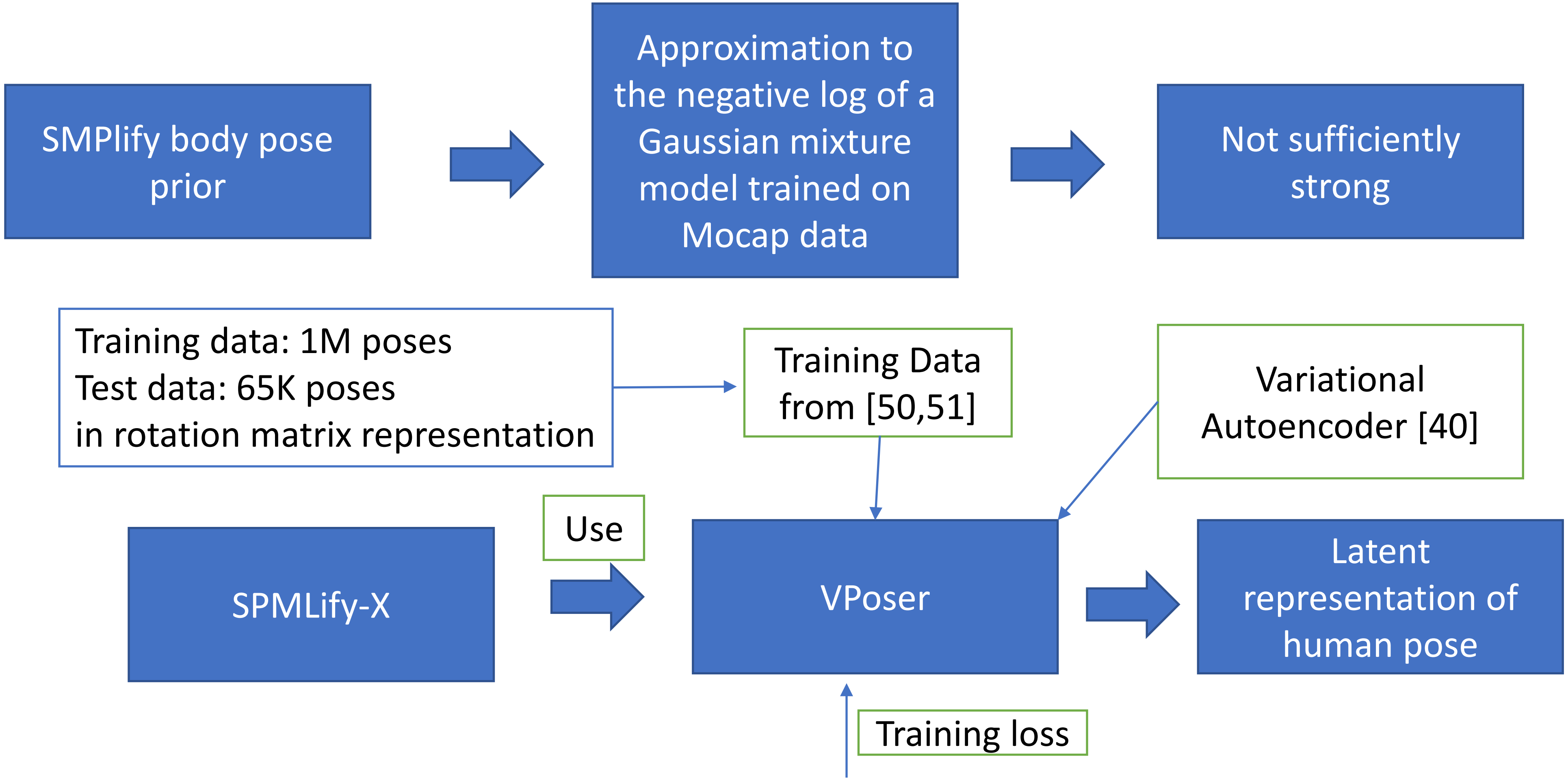
robust Geman-McClure error function

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

per-joint weights

confidence score
from OpenPose

Variational human body pose prior



$$\mathcal{L}_{total} = c_1\mathcal{L}_{KL} + c_2\mathcal{L}_{rec} + c_3\mathcal{L}_{orth} + c_4\mathcal{L}_{det1} + c_5\mathcal{L}_{reg}$$

Variational human body pose prior

Training data: 1M poses
Test data: 65K poses
in rotation matrix representation

$R \in SO(3)$ are 3×3 rotation matrices for each joint

Training Data from [50,51]

Variational Autoencoder [40]

VPoser

Latent representation of human pose $Z \in \mathbb{R}^{32}$

Training loss

$$\mathcal{L}_{total} = c_1\mathcal{L}_{KL} + c_2\mathcal{L}_{rec} + c_3\mathcal{L}_{orth} + c_4\mathcal{L}_{det1} + c_5\mathcal{L}_{reg}$$

$$\mathcal{L}_{KL} = KL(q(Z|R)||\mathcal{N}(0,I))$$

$$\mathcal{L}_{rec} = ||R - \hat{R}||_2^2$$

follow the VAE formulation in [40]

$\hat{R}$ is a similarly shaped matrix

encourage a normal distribution on the latent space, and to make an efficient code to reconstruct the input with high fidelity

$$\mathcal{L}_{orth} = ||\hat{R}\hat{R}' - I||_2^2$$

$$\mathcal{L}_{det1} = |det(\hat{R}) - 1|$$

encourage the latent space to encode valid rotation matrices

$\mathcal{L}_{reg} = ||\phi||_2^2,$ prevent over-fitting by encouraging smaller network weights ϕ

Variational human body pose prior

- Still not clear this.

To employ VPoser in the optimization, rather than to optimize over θ_b directly in Eq. 4, we optimize the parameters of a 32 dimensional latent space with a quadratic penalty on Z and transform this back into joint angles θ_b in axis-angle representation. This is analogous to how hands are treated except that the hand pose θ_h is projected into a linear PCA space and the penalty is on the linear coefficients.

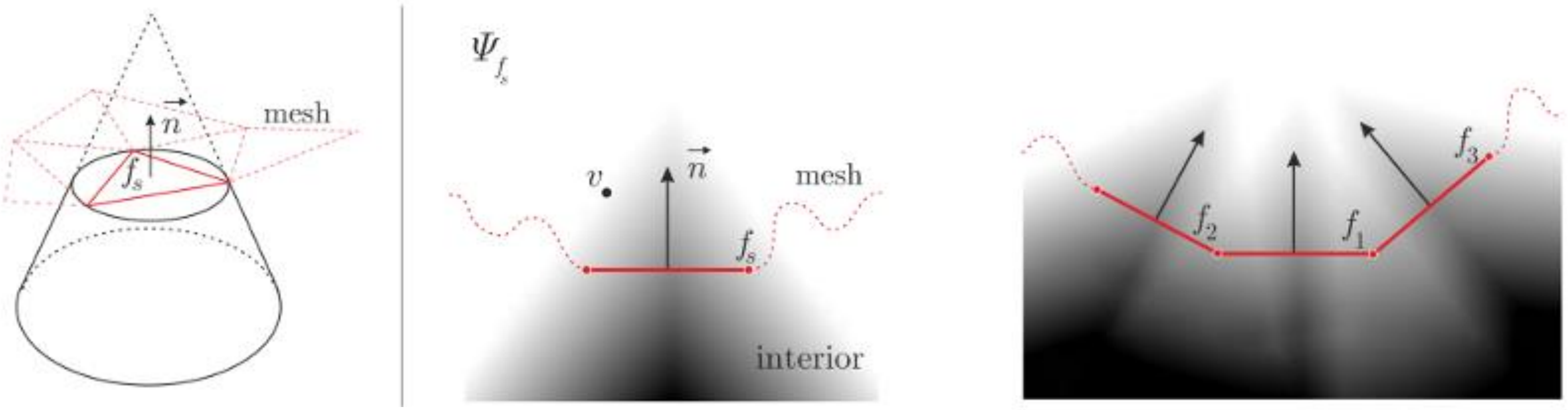
Collision penalizer

- Prevent the body penetration
- Developed from SMPLify and can apply in finger and facial expression

SMPLify	To express the finger and facial expressions	SMPLify-X
Capsule & Sphere	→	Triangle (mesh-face)
Too approximate		More exact

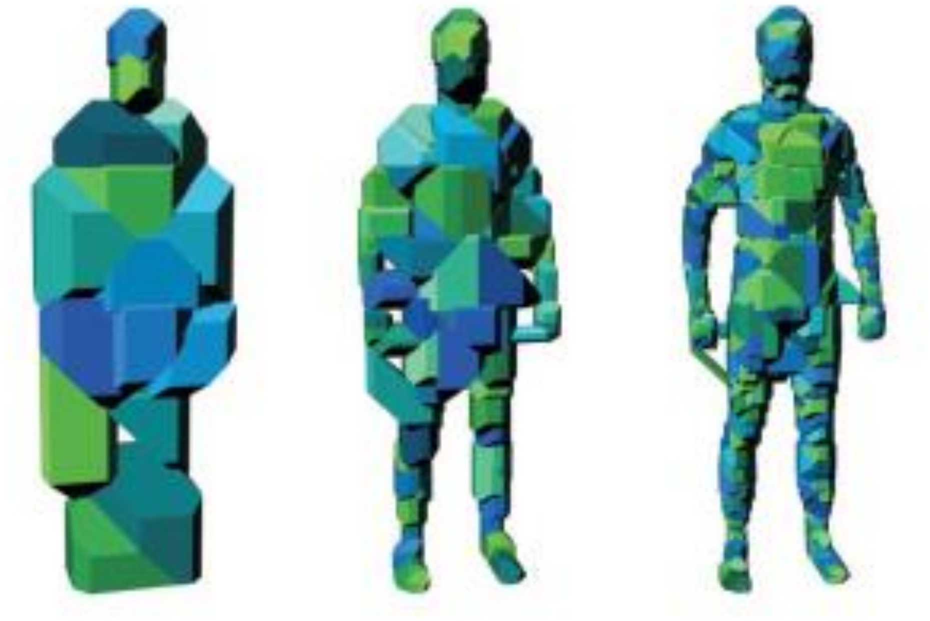
Collision penalizer

Search colliding triangle by BVH → List of Colliding triangle → Get distance field (draw conic) → Calculate collision term



BVH

- Place the bounding volume of the object in a tree called BVH.
- Time complexity is logarithmic.
- When the bounding box collide, the objects in the box collide.
- Check the collide while increasing the depth.
- Don’t check the hierarchy that collide in the previous depth.



Depth:1 depth:2 depth:3

Formular

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

Collision term
- Because the intrusion is bi-directional, they have two terms.
- Fs can be intruder and receiver

$$\Psi_{f_s}(\mathbf{v}_t) = \begin{cases} |(1-\Phi(\mathbf{v}_t))\Upsilon(\mathbf{n}_{f_s}\cdot(\mathbf{v}_t-\mathbf{o}_{f_s}))|^2 & \Phi(\mathbf{v}_t) < 1 \\ 0 & \Phi(\mathbf{v}_t) \geq 1 \end{cases}$$

Distance field

-if $\Phi(\mathbf{v}_t)$ is smaller then 1, the vertex is in the cone. Else it is out the cone, so we don't give a penalty

$$\Phi(\mathbf{v}_t) = \frac{\|(\mathbf{v}_t-\mathbf{o}_{f_s}) - (\mathbf{n}_{f_s}\cdot(\mathbf{v}_t-\mathbf{o}_{f_s}))\mathbf{n}_{f_s}\|}{-\frac{r_{f_s}}{\sigma}(\mathbf{n}_{f_s}\cdot(\mathbf{v}_t-\mathbf{o}_{f_s})) + r_{f_s}}$$

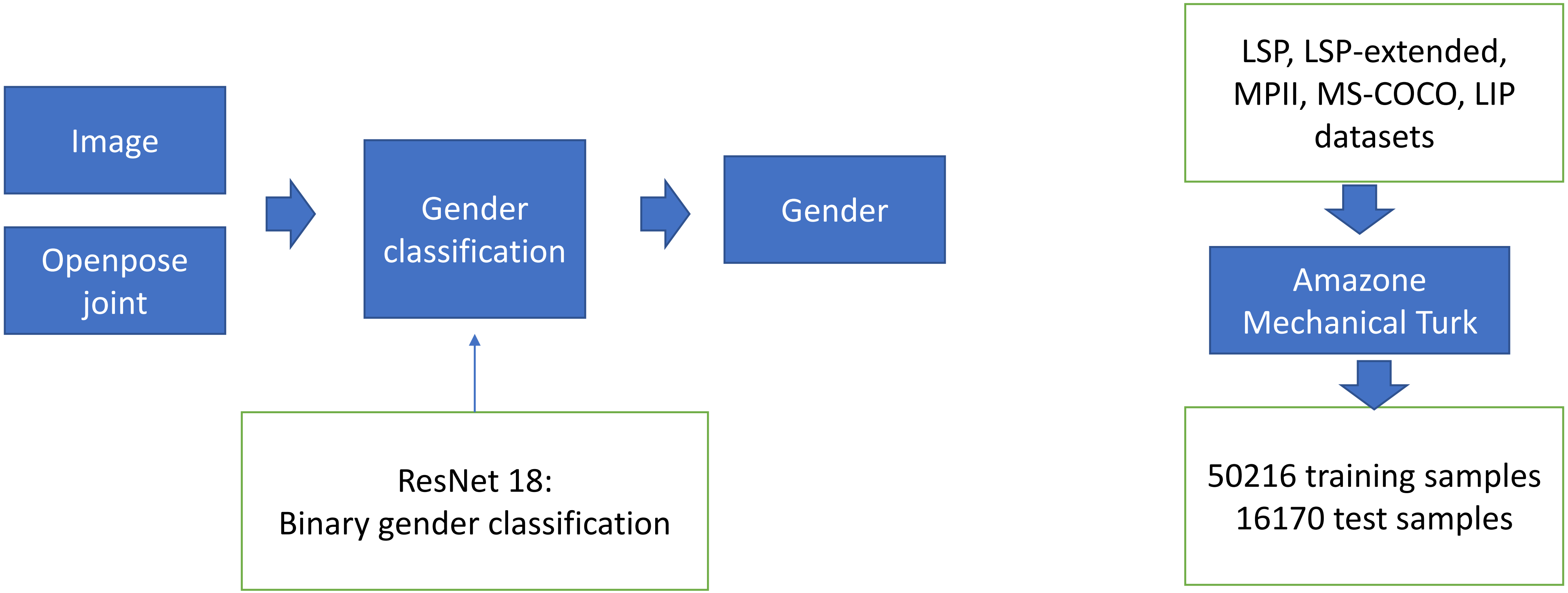
A measure of whether a vertex is in a cone.
- Molecule means the distance between vertex to axis.
- The denominator approximates the radius of the cone at the height of the vertex.

$$\Upsilon(x) = \begin{cases} -x+1-\sigma & x \leq -\sigma \\ -\frac{1-2\sigma}{4\sigma^2}x^2 - \frac{1}{2\sigma}x + \frac{1}{4}(3-2\sigma) & x \in (-\sigma,+\sigma) \\ 0 & x \geq +\sigma. \end{cases}$$

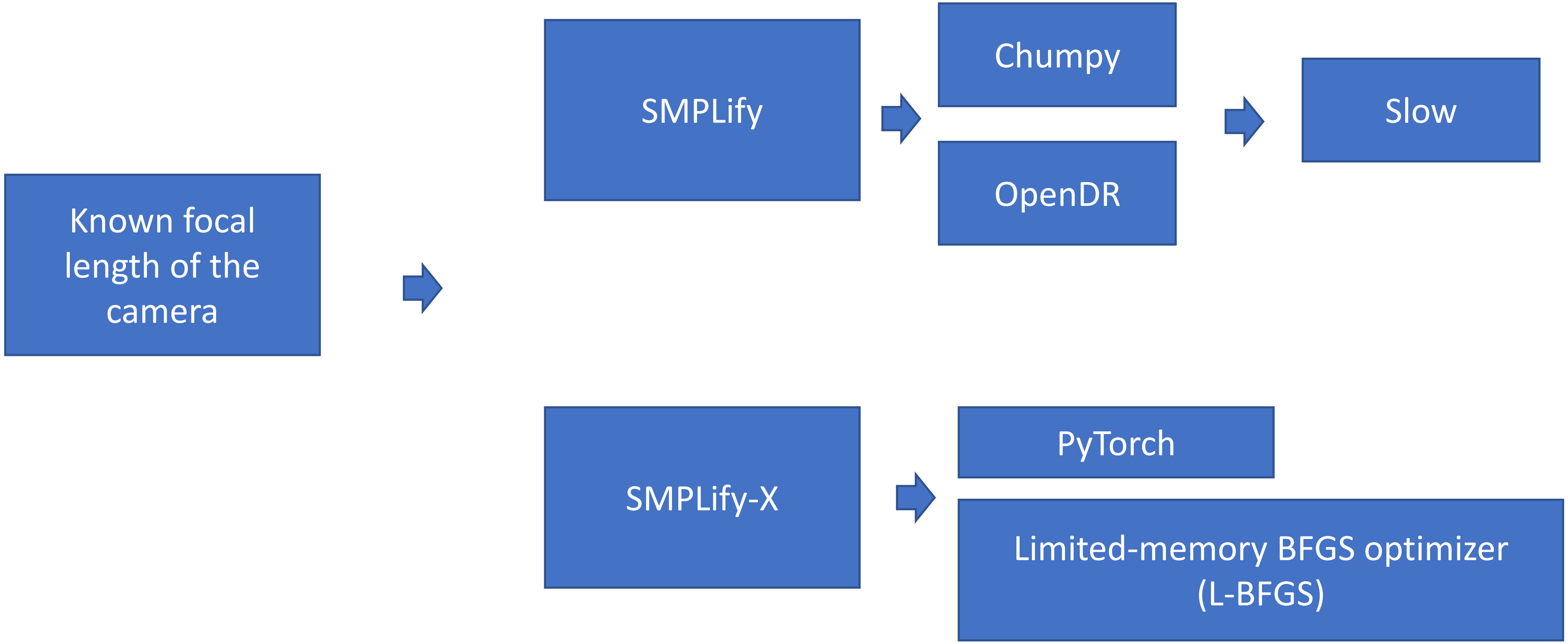
Intensity of repulsion
- x is the height of the vertex

Deep Gender Classifier

- No previous method that automatically takes gender into account.



Optimization



$$E_J(\boldsymbol{\beta}, \boldsymbol{\theta}; K, J_{\text{est}}) = \sum_{\text{joint } i} w_i \rho(\Pi_K(R_\theta(J(\boldsymbol{\beta})_i)) - J_{\text{est},i})$$

On torso joints

Optimization

Camera translation and body orientation are unknown

The camera focal length or its rough estimate is known.
Side view

Via the ratio of similar triangles

SMPL-X

Torso length of mean SMPL shape

Starting with high value then decreasing

λ_θ and λ_β

Assumption

Initialize the camera translation

Estimate depth

Estimating camera translation

Fitting by minimizing Eq. (1) Using Powell's dogleg method [31]

The person is standing parallel to the image plane

β fix to mean shape

Predicted 2D joints

$$E(\beta, \theta, \psi) = E_J + \lambda_{\theta_b} E_{\theta_b} + \lambda_{\theta_f} E_{\theta_f} + \lambda_{m_h} E_{m_h} + \lambda_\alpha E_\alpha + \lambda_\beta E_\beta + \lambda_\mathcal{E} E_\mathcal{E} + \lambda_\mathcal{C} E_\mathcal{C} \quad (4)$$

Optimization: hyper parameters

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

Small body parts like the hands and face → a lot of keypoints

Can dominated in Eq. 4

Local optimum (bad initializatiton)

Weights for joints

γ_b body keypoints,

γ_h hands

γ_f facial keypoints

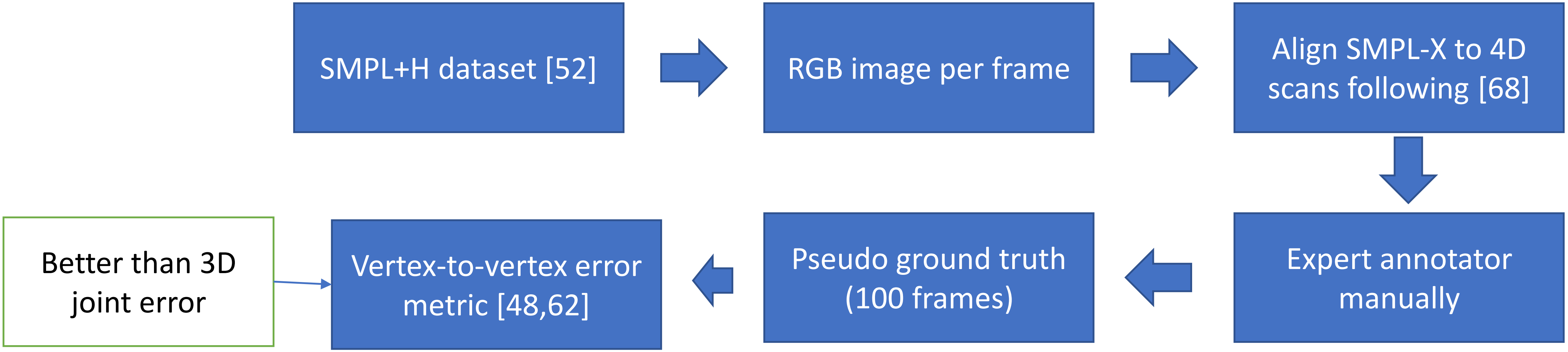
Focus on body pose

Increase influence of hands arm

Facial KJ.

Experiments

- Evaluation datasets
  - NO dataset with ground-truth shape for bodies, hands and face together.
  - ➔Create a dataset for evaluation.
  - Expressive hands and faces dataset (EHF)



Qualitative & Quantitative evaluations

Use less inputs than previous works

Model	Keypoints	v2v error	Joint error
“SMPL”	Body	57.6	63.5
“SMPL”	Body+Hands+Face	64.5	71.7
“SMPL+H”	Body+Hands	54.2	63.9
SMPL-X	Body+Hands+Face	52.9	62.6

Table 1: Quantitative comparison of “SMPL”, “SMPL+H” and SMPL-X, as described in Section 4.2, fitted with SMPLify-X on the EHF dataset. We report the mean vertex-to-vertex (v2v) and the standard mean 3D body (only) joint error in mm. The table shows that richer modeling power results in lower errors.

Version	v2v error
SMPLify-X	52.9
gender neutral model	58.0
replace Vposer with GMM	56.4
no collision term	53.5

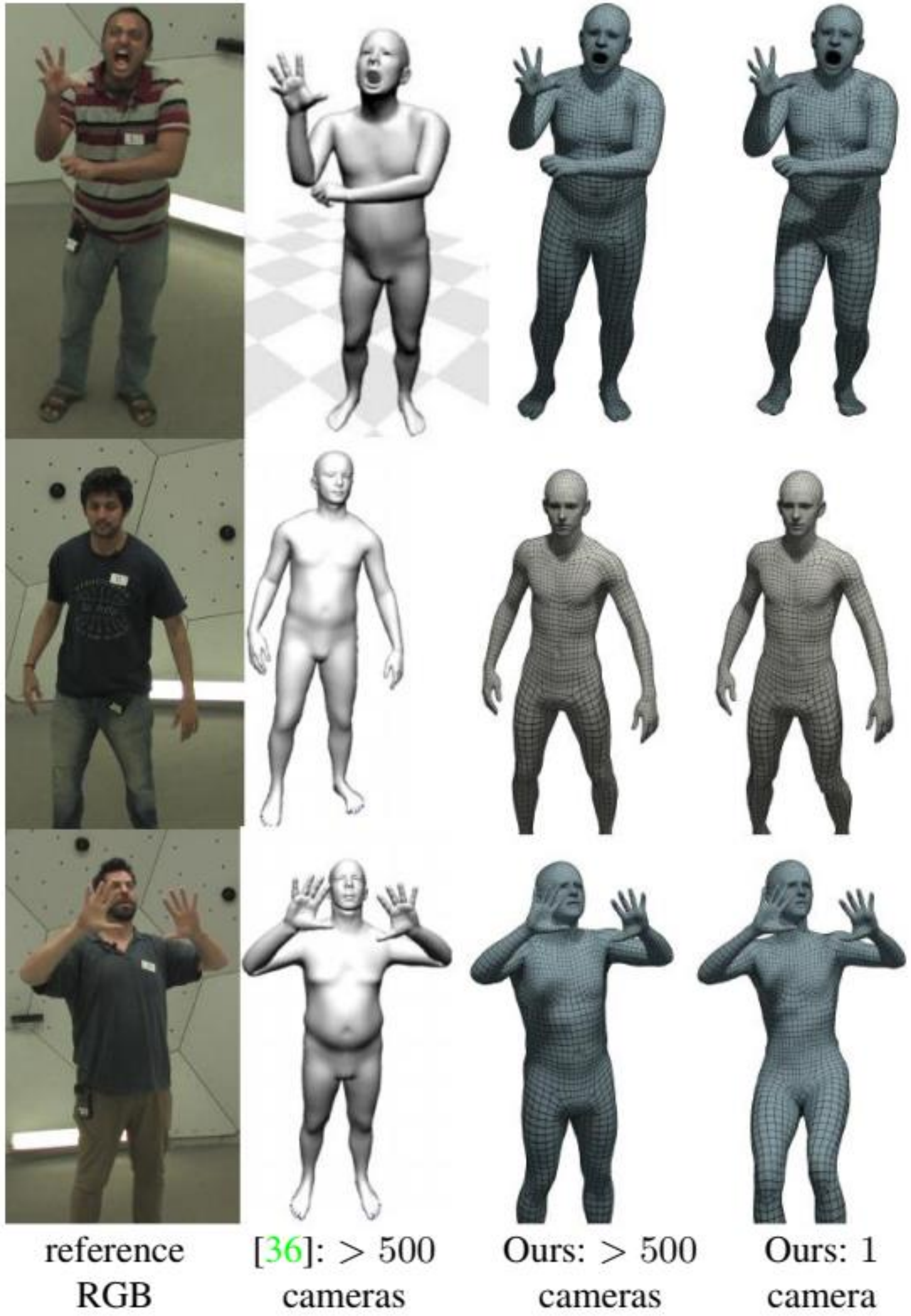
Table 2: Ablative study for SMPLify-X on the EHF dataset. The numbers reflect the contribution of each component in overall accuracy.

SMPL-X:
+No artifacts around the joints: elbows…
+Less inputs

Neutral model →

specific model →

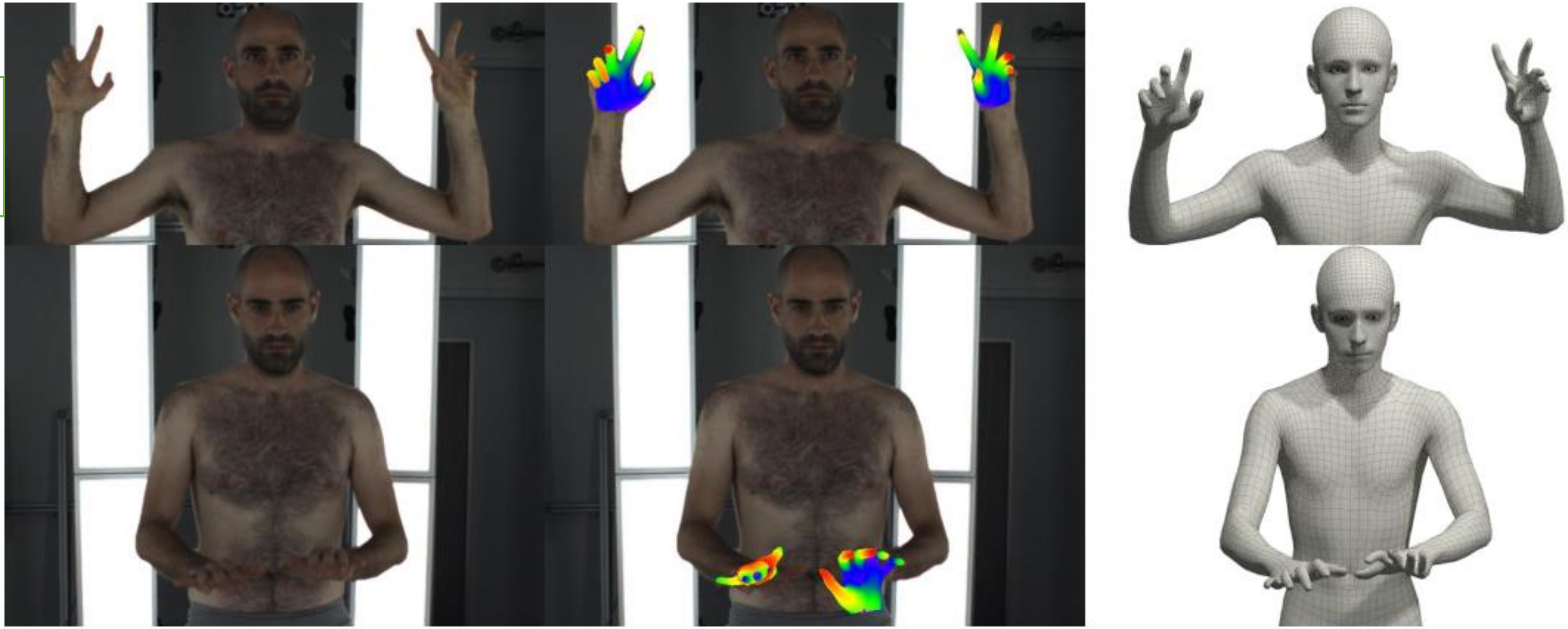
Neutral model →



Hand only

Good 2D Joints Detector→

Bad 2D Joints Detector→



Qualitative results

Figure 4: Qualitative results of SMPL-X for the in-the-wild images of the LSP dataset [33]. A strong holistic model like SMPL-X results in *natural* and *expressive* reconstruction of bodies, hands and faces. Gray color depicts the gender-specific model for confident gender detections. Blue is the gender-neutral model that is used when the gender classifier is uncertain.

Conclusion

- SMPL-X: model with body, face and hands.
- SMPLify-X: fit SMPL-X to single RGB image and 2D Open Pose joints.
- New body pose prior: fast and accurate
- Introduce a curated dataset with pseudo ground-truth.