Collaborative Regression of Expressive Bodies using Moderation
Supplemental Material

Yao Feng* Vasileios Choutas* Timo Bolkart Dimitrios Tzionas Michael J. Black
Max Planck Institute for Intelligent Systems, Tübingen, Germany
{yfeng, vchoutas, tbolkart, dtzionas, black}@tuebingen.mpg.de
* Equal contribution

1. Implementation Details

Data augmentation: For training data, we use image crops around the body, face and hands. We augment our training image crops, following mainly [1], as described below. First, we use standard techniques, namely random horizontal flipping, random image rotations, color noise addition and random translation of the crop's center. However this is not enough, as there is a significant domain gap between face-only and hand-only datasets, and the respective image crops extracted from full-body images; the former have significantly higher resolution. To account for this, we also randomly down-sample and up-sample the head and hand image crops, to simulate various lower resolutions. Finally, inspired by [5], we add synthetic motion blur to face and hand crops, to simulate the motion blur that is common in full-body images. Exact augmentation parameters can be found in our code website.

Training details: We use PyTorch [3] to implement our pipeline. We follow a three-step training procedure: (1) We pre-train the model with body-only, face-only and hand-only datasets; for each dataset we train only the respective parameters. Since these datasets are captured independently, there is no body image that corresponds to a face-only or hand-only image. Consequently, for this step we cannot apply feature fusion, and body-part features go directly to the respective regressor(s) (bypassing the moderators), to estimate the respective body-part parameters. Similar to existing work, we train only a right hand regressor; for images of a left hand, we flip the image horizontally to use the right hand regressor, and mirror the predictions to get a left hand. (2) Then, using the same data, we freeze the feature encoders and proceed with training the regressors and extractors (see Fig. 3 of the paper the linear layers $\mathcal{L}^h$ and $\mathcal{L}^f$ between the body encoder E_b and moderators $\mathcal{M}_h$ and $\mathcal{M}_f$ respectively). This step encourages features F_b^h and F_b^f from body images to be in the same space as features F_h and F_f from part-only images, so that regressors $\mathcal{R}_f^{\text{fused}}$ and $\mathcal{R}_h^{\text{fused}}$ work for both feature types. (3) Finally, we train the full network, including the moderators $\mathcal{M}_h$ and $\mathcal{M}_f$, but this time using training images with full SMPL-X ground truth, to extract part crops from full-body images as well. However, there are two problems. First, for these images there is no skin mask available, consequently we remove the loss for body shape $\boldsymbol{\beta}$ and do not apply a photometric and identity loss on head crops. Second, localizing the hands with body-driven attention is much harder compared to the head, due to the longer kinematic chain, consequently we freeze the hand regressor $\mathcal{R}_h$ to avoid fine-tuning it with invalid inputs.

All parameters are optimized using Adam [2] with a learning rate of 0.0001. For training the body, hand and face sub-networks, we use a batch size of 16 , 16, and 8, respectively. The moderator is a fully connected network with the following structure: FC (2048, 1024), ReLU, FC (1024, 1). All input images are resized to 224×224 pixels before feeding them to our network. During inference, we extract the hand/face crops using the hand and face locations from R_b's output. Hand and face cameras are ignored when estimating full body pose.

Global to relative pose: The regressors $\mathcal{R}_f^{\text{fused}}$ and $\mathcal{R}_h^{\text{fused}}$ estimate the absolute head and wrist orientation $\boldsymbol{\theta}_{\text{g}}$, i.e. irrespective of the (parent) main body's pose. However, to "apply" these $\boldsymbol{\theta}_{\text{g}}$ estimates on a SMPL-X body that is already posed by $\mathcal{R}_b$ with $\boldsymbol{\theta}_b$ (up to the wrist and neck, excluding them), we need to express them relative to their parent in the kinematic skeleton:

$$\boldsymbol{\theta}_{\text{relative}} = \boldsymbol{\Gamma}(\boldsymbol{\theta}_{\text{g}}, \boldsymbol{\theta}_b), \tag{1}$$

where $\boldsymbol{\Gamma}$ is the chain transformation function according to SMPL-X's kinematic skeleton hierarchy.

2. Evaluation

2.1. Body-face correlations discussion

PIXIE gives more realistic body shapes, not only due to its gendered shape loss, but also thanks to the shared body, hand and face shape space of SMPL-X. This allows PIXIE's

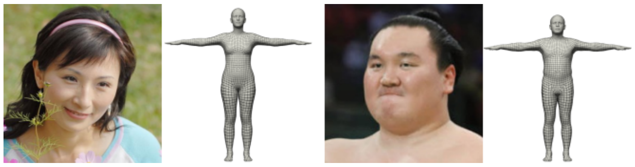

Figure 1: Whole-body shape estimation from *only* our face expert, using SMPL-X's joint shape space for all body parts.

face expert to – uniquely – contribute to whole-body shape. To verify this, we apply our face expert on face-only images and get the whole-body shapes of Fig. 1. These are not only correctly "gendered", but also have a plausible BMI. For the sumo wrestler in Fig. 1, PIXIE predicts a body with higher BMI (26.9) than the mean shape (26.1). PIXIE is the only 3D whole-body estimation method that explores such face-body shape correlations explicitly. We believe that this is a useful insight and points the community towards a new direction.

2.2. Qualitative Evaluation

Comparison with MTC: In Fig. 2 we compare PIXIE with MTC [6]. PIXIE is two orders of magnitude faster and predicts more accurate 3D body shapes. However, when 2D joint estimations are accurate, optimization-based methods, such as MTC [6] and SMPLify-X [4], tend to estimate bodies that are better aligned with the image.

Expressive body reconstruction: We compare our method, PIXIE, with other state-of-the-art expressive body reconstruction methods in Fig. 3. PIXIE is more robust to challenging ambiguities (blur, occlusion) than existing whole-body regressors [1, 5], since its moderators fuses "global" body and "local" part.

Qualitative results: Finally, in Fig. 4, 5 and 6 we provide more standalone PIXIE results. Overall, PIXIE produces visually plausible body shapes with detailed facial expressions.

Failure cases: Although the gender prior loss and the shared whole-body shape space result in better 3D shape predictions, they are not sufficient for perfectly estimating full-body 3D shape. Furthermore, the employed photometric term often causes the model to prefer to explain image evidence using lighting, rather than albedo, which leads to incorrect skin tone predictions. These points highlight important directions for improving PIXIE. Representative failure cases can be seen in Fig. 7.

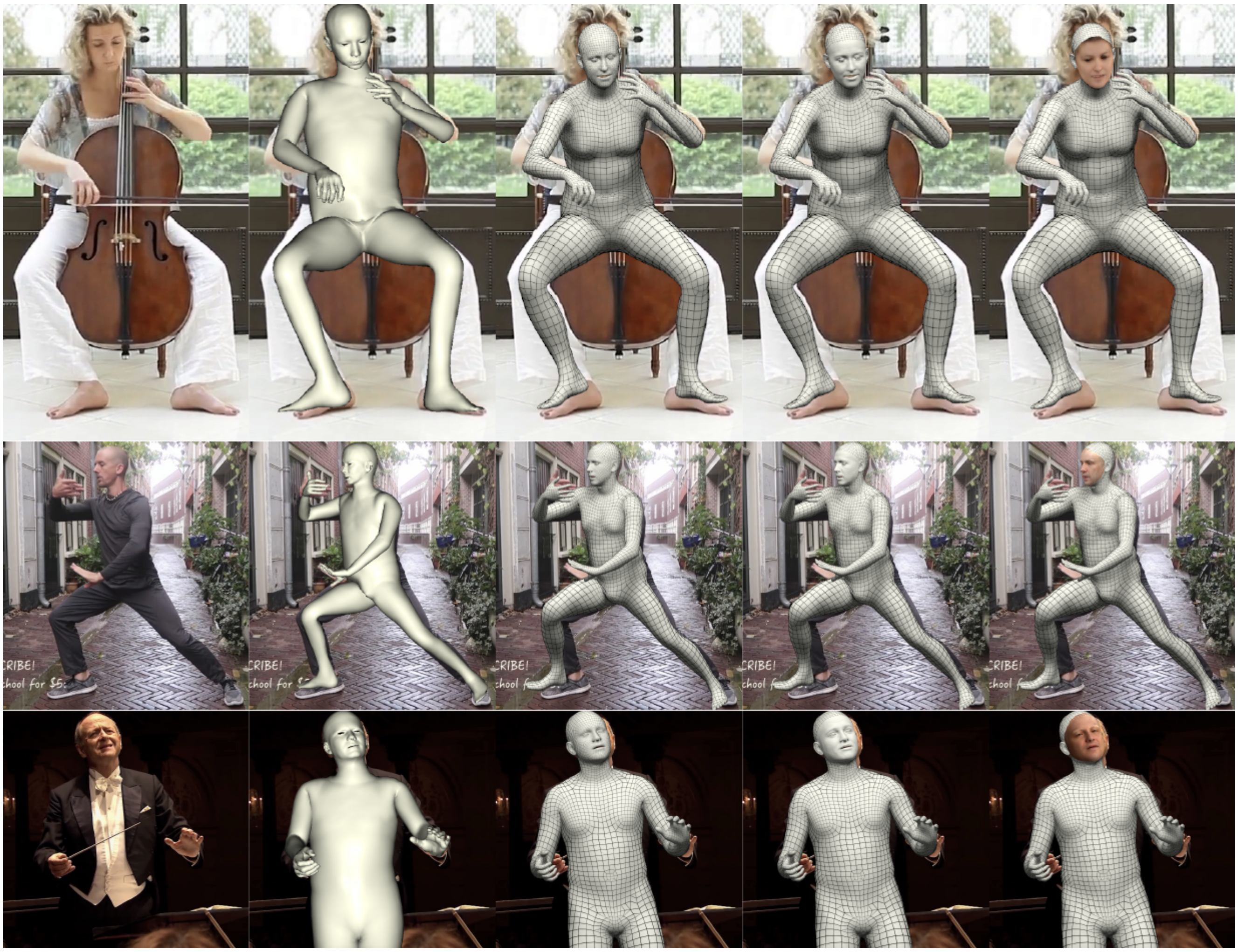

Figure 2: Qualitative PIXIE results and comparison to MTC [6]. From left to right: (1) RGB image, (2) MTC [6], (3) PIXIE, (4) PIXIE with facial geometric details, (5) PIXIE with estimated face albedo and lighting. Overall, PIXIE produces more visually plausible body shapes and more detailed facial expressions.

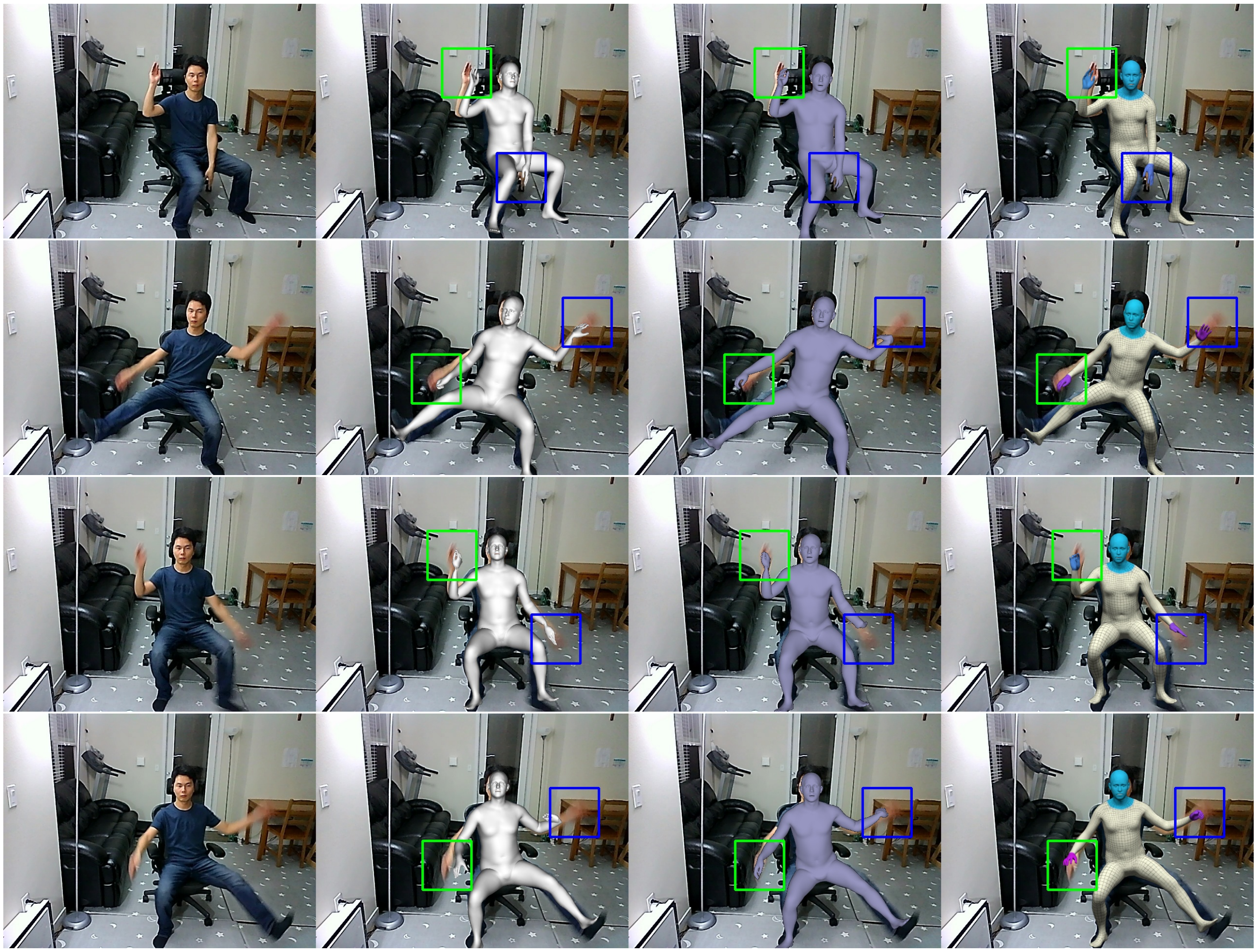

Figure 3: Qualitative PIXIE results and comparison to ExPose [1] and FrankMocap [5]. From left to right: (1) RGB images from video, (2) FrankMocap [5], (3) ExPose [1], (4) PIXIE 3D body predictions with color-coded part-expert confidence. Moderator predicts the confidence of body/face/hand experts, reder means higher confidence in the body expert rather than the results from face/hand experts. Thanks to the moderators, PIXIE is more robust to low quality part images. For example, when the hand is blurry, PIXIE still predicts a plausible wrist pose, instead of an unnatural twist.

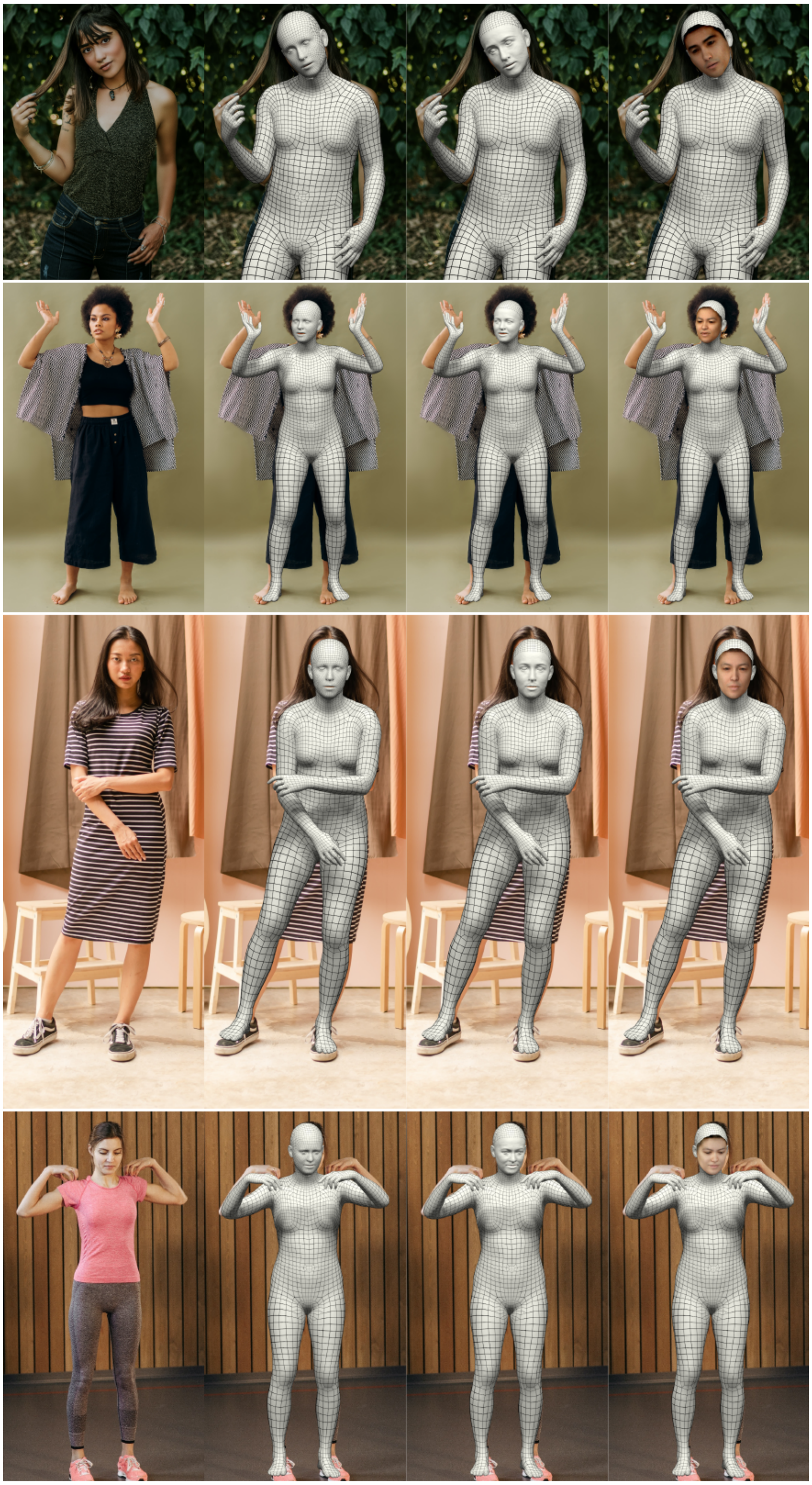

Figure 4: Qualitative PIXIE results. From left to right: (1) RGB image, (2) PIXIE, (3) PIXIE with facial geometric details, (4) PIXIE with estimated face albedo and lighting.

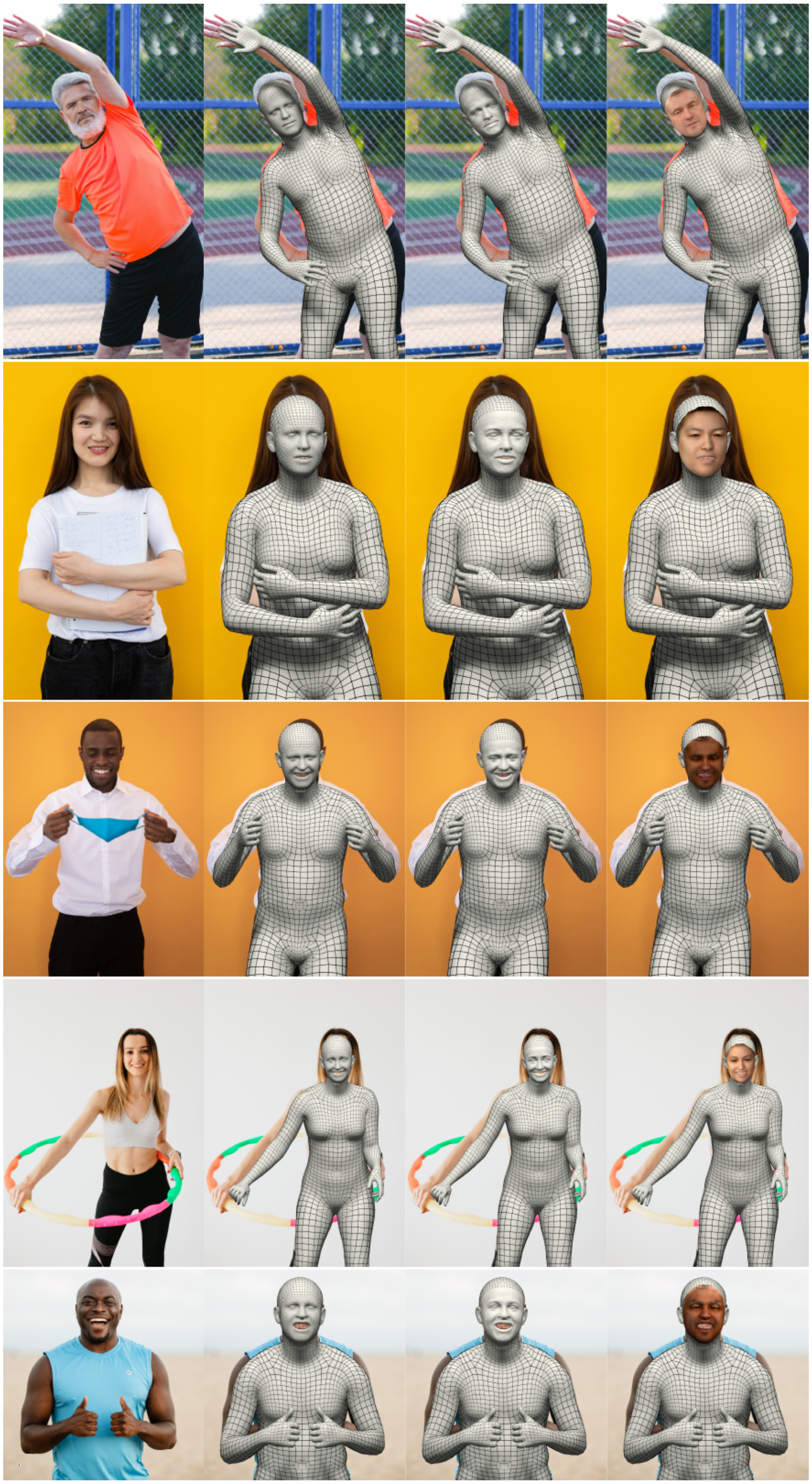

Figure 5: Qualitative PIXIE results. From left to right: (1) RGB image, (2) PIXIE, (3) PIXIE with facial geometric details, (4) PIXIE with estimated face albedo and lighting.

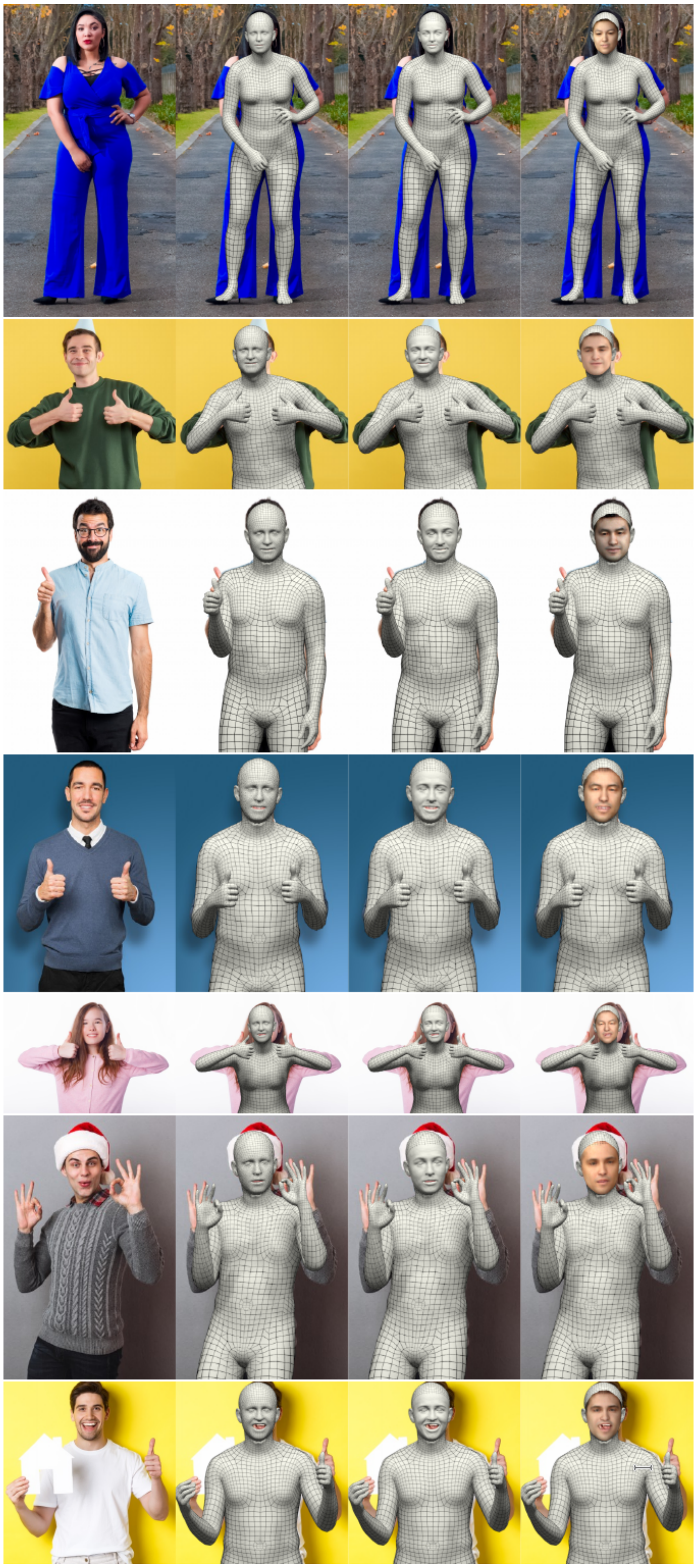

Figure 6: Qualitative PIXIE results. From left to right: (1) RGB image, (2) PIXIE, (3) PIXIE with facial geometric details, (4) PIXIE with estimated face albedo and lighting.

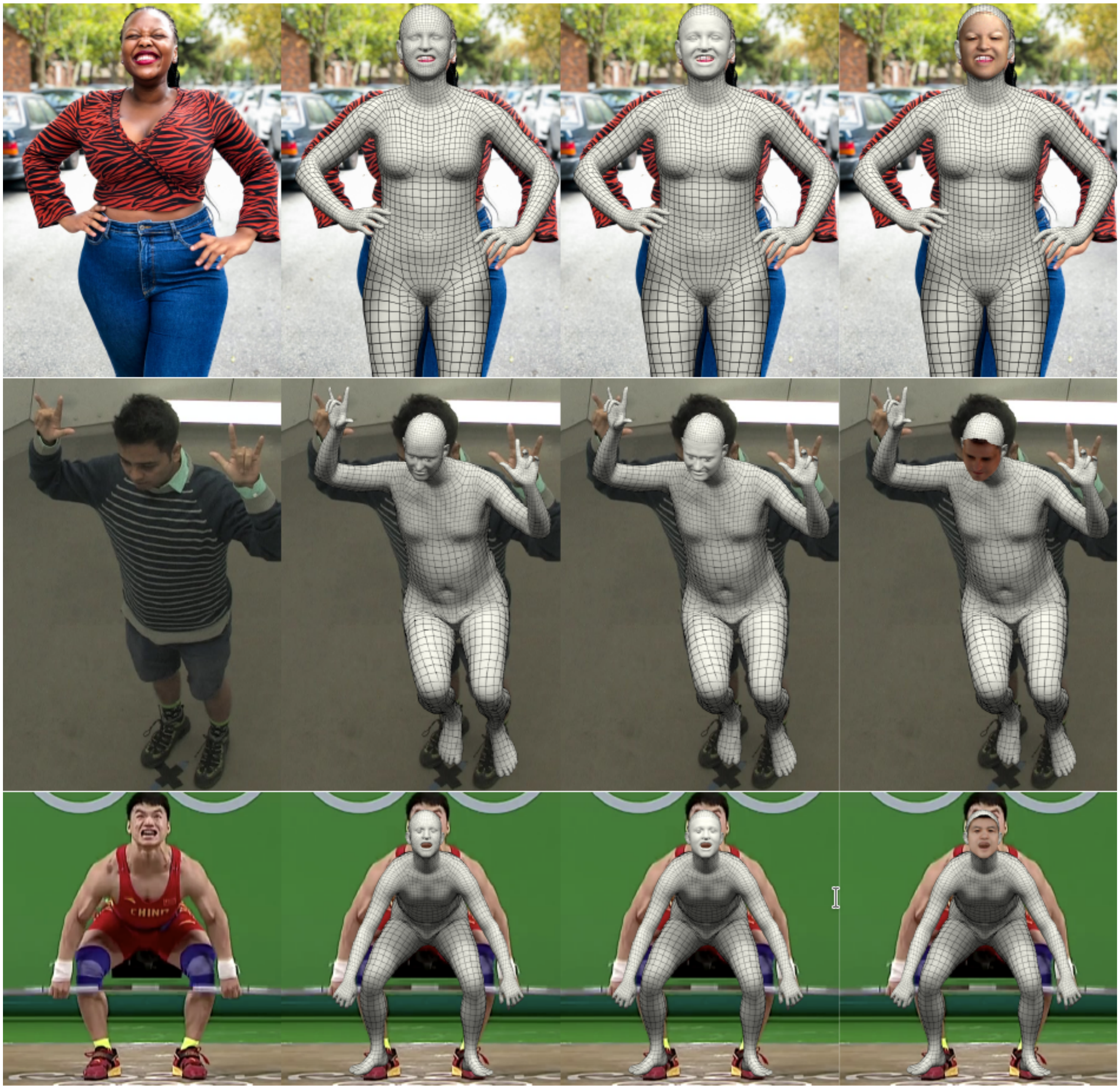

Figure 7: Failure cases for PIXIE. In these examples, the implicit reasoning about gender and the face shape information are not enough to correctly infer the body shape. Furthermore, due to the formulation of the photometric term the model prefers to explain image evidence using lighting, rather than albedo, which leads to wrong skin tone predictions. Finally, replacing the weak-perspective camera with a perspective model would make the model more robust to extreme viewing angles and perspective distortion effects. Future work should look into denser forms of supervision, formulating a better photometric term and integrating a perspective camera to resolve these issues.

References

[1] 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)*, pages 20–40, 2020. 1, 2, 4

[2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 1

[3] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, 2019. 1

[4] 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 *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2

[5] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration. In *International Conference on Computer Vision Workshops (ICCVw)*, 2021. 1, 2, 4

[6] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10974, 2019. 2, 3