Monocular Expressive Body Regression through Body-Driven Attention
Supplemental Material

Vasileios Choutas[1,2], Georgios Pavlakos[3], Timo Bolkart[1], Dimitrios Tzionas[1], and Michael J. Black[1]

[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2] Max Planck ETH Center for Learning Systems
[3] University of Pennsylvania, Philadelphia, USA
{vchoutas, gpavlakos, tbolkart, dtzionas, black}@tuebingen.mpg.de

https://expose.is.tue.mpg.de

Here we provide addition details and visualizations of our results. Please also see the narrated **video** on our website for a summary of the method and results.

1 Training details

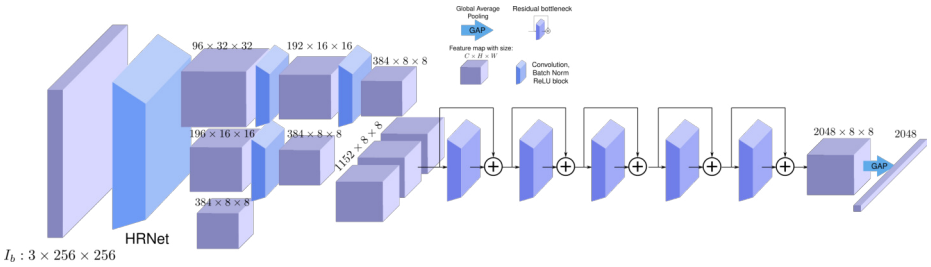


Fig. 1: Structure of the feature extractor used by the body prediction network. The image I_b is fed to HRNet [9] to extract multi-scale feature maps. These are then processed by extra convolutional blocks and downsampled to the same spatial resolution. All feature maps are subsequently concatenated and fed to 5 residual blocks [1], followed by a global average pooling operation that produces the final feature vector $\mathcal{F}_b$.

Architecture: The features $\mathcal{F}_b$ are extracted from the body image I_b using the architecture of Figure 1. The parameters $\Theta = \{\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, s, \boldsymbol{t}\}$ are predicted by feeding the features $\mathcal{F}_b$ and the mean parameters $\bar{\Theta}$ to an iterative regression network, whose structure follows [2]. The composition of the feature extraction network of Figure 1 and the iterative regressor forms the body network g.

Fig. 2: **Illustrative examples.** The default global rotation of the hand is replaced by a random rotation with angle $r_{\text{global}} \sim \mathcal{U}(r_{\min}, r_{\max})$ around the ground truth axis of rotation given by the training data. We selected a range $(r_{\min}, r_{\max})_{\text{hand}} = (-90, 90)$ degrees. Blue is the ground-truth mesh used as a target for training, while gray is the starting point of the iterative hand regressor with a perturbed global rotation.

Training: We pre-train the body network until validation performance on 3DPW [6] saturates, using Adam [4], with batch size 48. The hand and head sub-networks are pre-trained as well on the FreiHand [10] and FFHQ [3] data, again with Adam [4] and a batch size of 64. Once validation performance saturates, we freeze the body network and fine-tune the hand and head sub-networks with all available training data to produce ExPose. The exact hyper-parameters will be included in the released code. The entire pipeline is implemented in PyTorch.

2 Data augmentation

For hand and face-only data, shape and pose regression is done following the iterative scheme of [2], which computes offsets from a set of mean parameters. When we have access to full body information, we wish to condition the part specific sub-networks on the output of the body network. However, naively adding this conditioning is not enough, as this creates a domain gap between hands and face-only images and those coming from the body attention mechanism. To bridge this, we augment the training data by modifying the initial mean point to some random point. In this way, the part sub-network will be forced to learn to predict the correct offsets, no matter the initial point, that lead to the pose and shape that matches the image. As described in the main text, we randomly perturb the global rotation of the hand and face data around the ground-truth axis of rotation, as illustrated in Figures 2 and 3 respectively. We also modify the shape of the hand and the face by randomly sampling from normal distributions over the hand and face shape parameters, as illustrated in Figures 6 and 5 respectively. For the face-only data, we also augment the rotation of the jaw, by replacing the default value with a random rotation around the x-axis, seen in Figure 4. Finally, we replace the default mean expression with a sample drawn from a standard normal distribution, as seen in Figure 7.

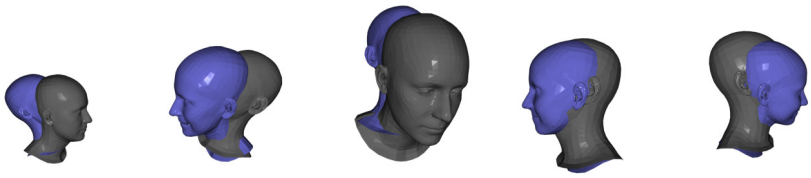


Fig. 3: The default global rotation of the head is replaced by a random rotation with angle $r_{\text{global}} \sim \mathcal{U}(r_{\min}, r_{\max})$ around the ground truth axis of rotation given by the training data. We selected a range $(r_{\min}, r_{\max})_{\text{head}} = (-45, 45)$ degrees. Blue is the ground-truth mesh used as a target for training, while gray is the starting point of the iterative face regressor with a perturbed global rotation.

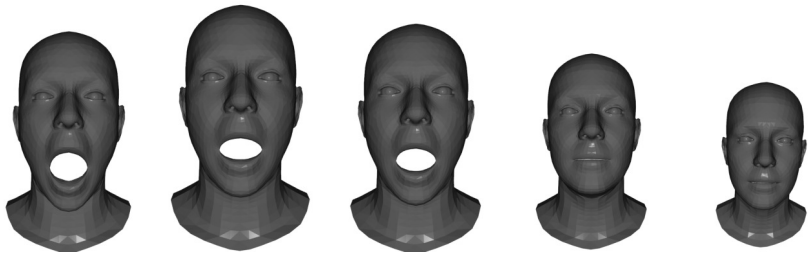


Fig. 4: The default rotation of the jaw, which corresponds to a closed mouth, is replaced by a random rotation around the x-axis. The angle of rotation is sampled randomly from the uniform distribution $r_{\text{jaw}} \sim \mathcal{U}(0, 45)$.

3 Converting SMPL to SMPL-X

There exist a wide variety of SMPL annotations for training 3D body pose and shape estimation methods. It is therefore important to create an automated method to convert them to the corresponding SMPL-X parameters, to use them as training data. To achieve this, we leverage the relation between SMPL and SMPL-X to build a correspondence map between the two models. SMPL and SMPL-X are articulated models of the human body that produce 3D triangle meshes:

$$M_{\text{SMPL}} = (V_{\text{SMPL}}, F_{\text{SMPL}}), \; V_{\text{SMPL}} \in \mathbb{R}^{6890 \times 3}, F_{\text{SMPL}} \in \mathbb{N}^{13776 \times 3} \tag{1}$$

$$M_{\text{SMPL-X}} = (V_{\text{SMPL-X}}, F_{\text{SMPL-X}}), \; V_{\text{SMPL-X}} \in \mathbb{R}^{10475 \times 3}, F_{\text{SMPL-X}} \in \mathbb{N}^{20908 \times 3} \tag{2}$$

We start by registering the SMPL template mesh to the SMPL-X template. Given the registered meshes, we compute for each SMPL-X vertex $\boldsymbol{v}_i$ its nearest point $\boldsymbol{p}_i$ on the SMPL mesh and store the index of the nearest SMPL triangle t_i, its vertex indices $\boldsymbol{f_i} = [f_0^i, f_1^i, f_2^i]$ and the barycentric coordinates $[\alpha_i, \beta_i, \gamma_i]$ of $\boldsymbol{p}_i$ with respect to triangle t_i. We also store a binary mask m_i for each vertex

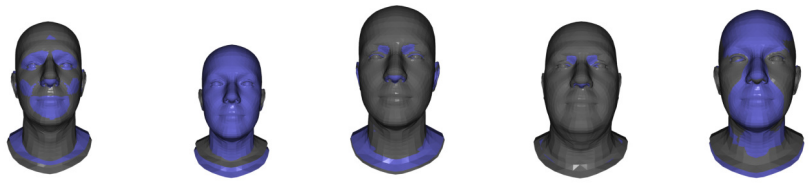


Fig. 5: The default mean shape of the head is replaced with a random vector $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, I), I \in \mathbb{R}^{10\times10}$ The blue mesh represents the mean shape, while the gray mesh has a random shape drawn from the above distribution.

Fig. 6: The default mean shape of the hand is replaced with a random vector $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, I), I \in \mathbb{R}^{10\times10}$ The blue mesh represents the mean shape, while the gray mesh has a random shape drawn from the above distribution.

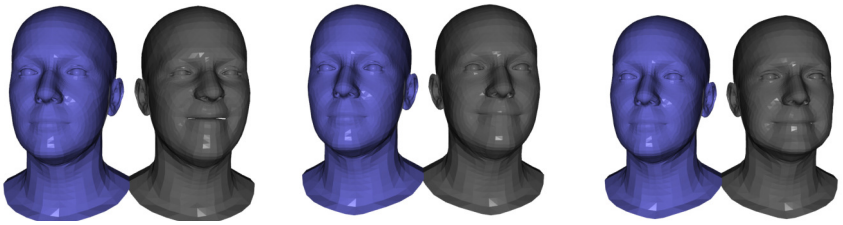


Fig. 7: The default neutral expression of the head is replaced with a random vector $\boldsymbol{\psi} \sim \mathcal{N}(\mathbf{0}, I), I \in \mathbb{R}^{10\times10}$. The blue mesh represents the neutral expression, while the gray mesh has a random expression drawn from the above distribution.

that is used to mask invalid correspondences between the two models, such as the eyes, inner lip region, etc.

Given a posed SMPL mesh $M' = (V', F')$, e.g. one sample from the fit data of SPIN [5], we build a mesh $\hat{M}$ in SMPL-X topology. Vertex $\hat{v}_i$ of the mesh $\hat{M}$ is computed as:

$$\hat{v}_i = \alpha_i v'_{f_0^i} + \beta_i v'_{f_1^i} + \gamma_i v'_{f_2^i} \tag{3}$$

where $v_{f_0^i}\boldsymbol{'}$ is the SMPL vertex with index f_0^i. We now have a mesh in SMPL-X topology, which we will use to find the corresponding pose $\boldsymbol{\theta}$, shape $\boldsymbol{\beta}$, expression ψ and translation $\boldsymbol{t}$ parameters. Let v_i be the i-th vertex returned by posing SMPL-X using the current values of the parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{t})$. We start by

optimizing only over the pose $\boldsymbol{\theta}$ using the following loss:

$$L_1(\boldsymbol{\theta}) = \sum_{(i,j)\in\mathcal{E}} m_i m_j \left\| (v_i - v_j) - (\hat{v}_i - \hat{v}_j) \right\|_2^2 \tag{4}$$

where $\mathcal{E}$ is the set of 3D edges of the SMPL-X mesh. We use the binary masks m_i, m_j to compute the loss only on valid vertices. For the second stage, we optimize the translation vector $\boldsymbol{t}$ using a vertex-to-vertex loss:

$$L_2(\boldsymbol{t}) = \sum_i m_i \left\| v_i - \hat{v}_i \right\|_2^2 \tag{5}$$

By this point, we have rigidly aligned the two meshes and matched the articulation of the original SMPL mesh. All that remains is to also match the shape, to get the best possible fit. The final step is to optimize over all parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{t})$ using again a vertex-to-vertex loss:

$$L_3((\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{t})) = \sum_i m_i \left\| v_i - \hat{v}_i \right\|_2^2 \tag{6}$$

We use a Trust Region Newton Conjugate Gradient optimizer [7] to search for minimize the objectives. The implementation for the transfer process can be in found on our website: https://expose.is.tue.mpg.de.

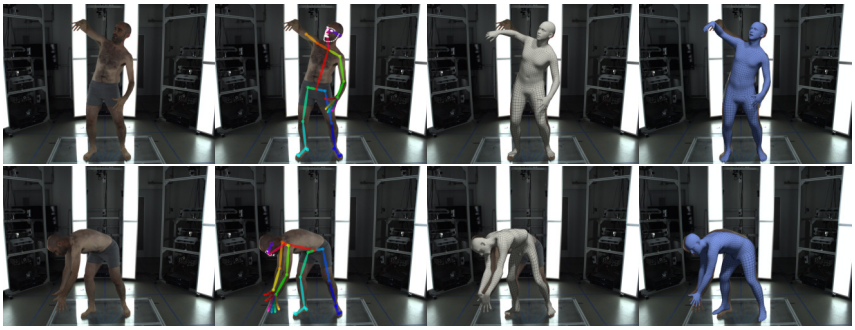
4 SMPLify-X qualitative comparison

As shown in Table 3 of the main manuscript, ExPose is almost 200× times faster compared to SMPLify-X [8], and provides qualitatively similar results to the latter, as seen in Figures 8a and 8b. Although the accuracy of ExPose is slightly lower than SMPLify-X, it can provide a better initialization to the latter, helping it overcome failures of its initialization heuristic and of the keypoint detector. Potentially, this could be done in a loop, similar to [5] to continuously improve the performance of ExPose using mode in-the-wild data.

5 In-the-wild qualitative results

A qualitative comparison of our method with the state-of-the-art SMPL regression methods shows the increase in expressivity offered by ExPose; see Figures 10 to 20. Figure 9 compares the output of the naive regression approach with the body-driven attention mechanism of ExPose. Finally, Figures 21 to 25 contain visualizations of ExPose predictions from multiple views.

(a) 1. The input image, 2. SMPLify-X (known gender), 3. naive regression from a single body image fails to capture detailed finger articulation and facial expressions, 4. ExPose is able to recover these details, thanks to its attention mechanism, and produces results of similar quality as SMPLify-X, while being 200 times faster.



(b) 1. The input image, 2. OpenPose detections, 3. SMPLify-X fitting, with the neutral model and default focal length, 4. ExPose. When 2D keypoint detections are missing or wrong, optimization based methods, such as SMPLify-X are unable to avoid implausible poses. Furthermore, they heavily depend on their initialization and can produce unnatural poses and shapes, when their initialization heuristic fails. Regression methods, such as ExPose, avoid these problems and can provide better initialization points, closer to the actual solution, and accelerate convergence.

Fig. 9: *Left:* The input image. *Middle:* Naive regression from a body crop. *Right:* ExPose. The attention mechanism helps capture detailed hand articulation and facial expression.

Fig. 10: Comparison of ExPose with the state-of-the-art body regression method: 1. RGB image, 2. SPIN [5], 3. ExPose. Our proposed method produces 3D body pose and shape results on par with SPIN [5] and captures more details for the hands and face.

Fig. 11: Comparison of ExPose with the state-of-the-art body regression method: 1. RGB image, 2. SPIN [5], 3. ExPose. Our proposed method produces 3D body pose and shape results on par with SPIN [5] and captures more details for the hands and face.

Fig. 12: Comparison of ExPose with the state-of-the-art body regression method: 1. RGB image, 2. SPIN [5], 3. ExPose. Our proposed method produces 3D body pose and shape results on par with SPIN [5] and captures more details for the hands and face.

Fig. 13: Comparison of ExPose with the state-of-the-art body regression method: 1. RGB image, 2. SPIN [5], 3. ExPose. Our proposed method produces 3D body pose and shape results on par with SPIN [5] and captures more details for the hands and face.

Fig. 14: Comparison of ExPose with the state-of-the-art body regression method: 1. RGB image, 2. SPIN [5], 3. ExPose. Our proposed method produces 3D body pose and shape results on par with SPIN [5] and captures more details for the hands and face.

Fig. 15: Comparison of ExPose with the state-of-the-art body regression method: 1. RGB image, 2. SPIN [5], 3. ExPose. Our proposed method produces 3D body pose and shape results on par with SPIN [5] and captures more details for the hands and face.

Fig. 16: Comparison of ExPose with the state-of-the-art body regression method: 1. RGB image, 2. SPIN [5], 3. ExPose. Our proposed method produces 3D body pose and shape results on par with SPIN [5] and captures more details for the hands and face.

Fig. 17: Comparison of ExPose with the state-of-the-art body regression method: 1. RGB image, 2. SPIN [5], 3. ExPose. Our proposed method produces 3D body pose and shape results on par with SPIN [5] and captures more details for the hands and face.

Fig. 18: Comparison of ExPose with the state-of-the-art body regression method: 1. RGB image, 2. SPIN [5], 3. ExPose. Our proposed method produces 3D body pose and shape results on par with SPIN [5] and captures more details for the hands and face.

Fig. 19: Comparison of ExPose with the state-of-the-art body regression method: 1. RGB image, 2. SPIN [5], 3. ExPose. Our proposed method produces 3D body pose and shape results on par with SPIN [5] and captures more details for the hands and face.

Fig. 20: Comparison of ExPose with the state-of-the-art body regression method: 1. RGB image, 2. SPIN [5], 3. ExPose. Our proposed method produces 3D body pose and shape results on par with SPIN [5] and captures more details for the hands and face.

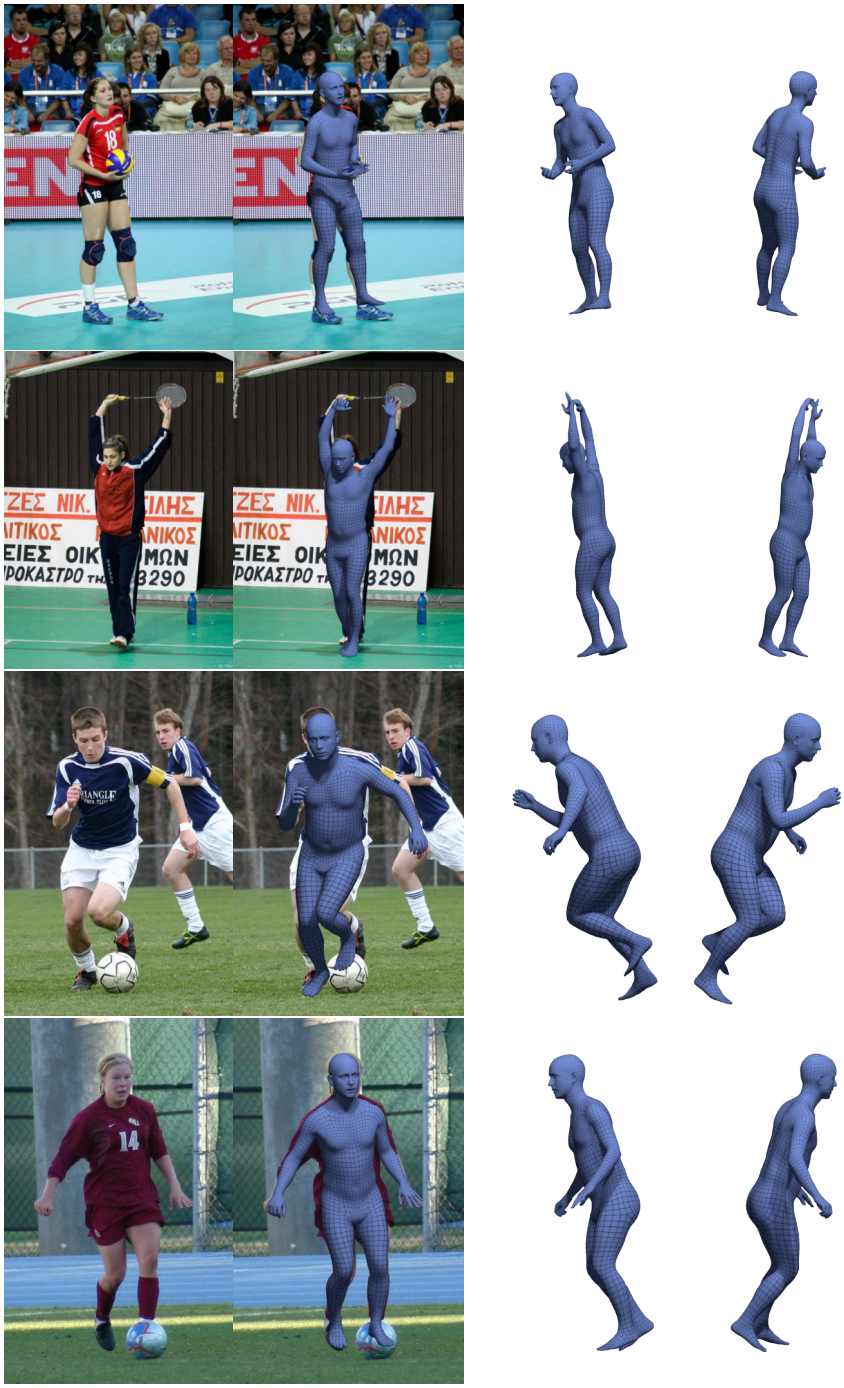


Fig. 21: ExPose results visualized from multiple views. 1. RGB image, 2. overlay, 3. , 4. rotations around the vertical axis

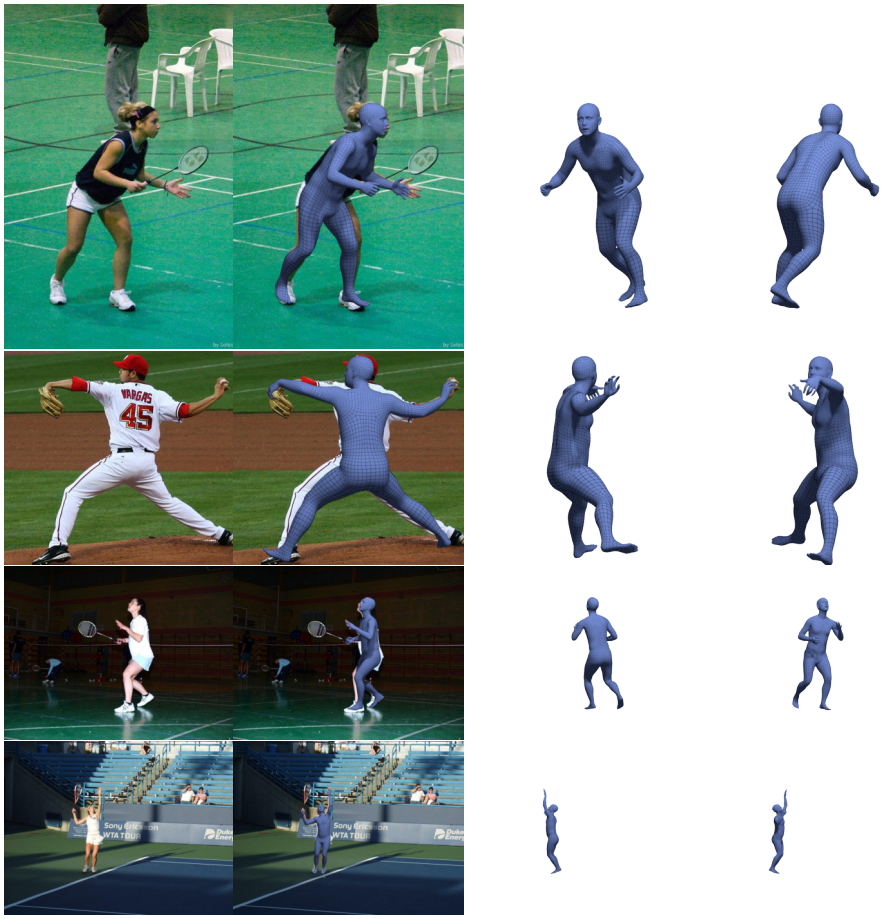


Fig. 22: ExPose results visualized from multiple views. 1. RGB image, 2. overlay, 3. , 4. rotations around the vertical axis

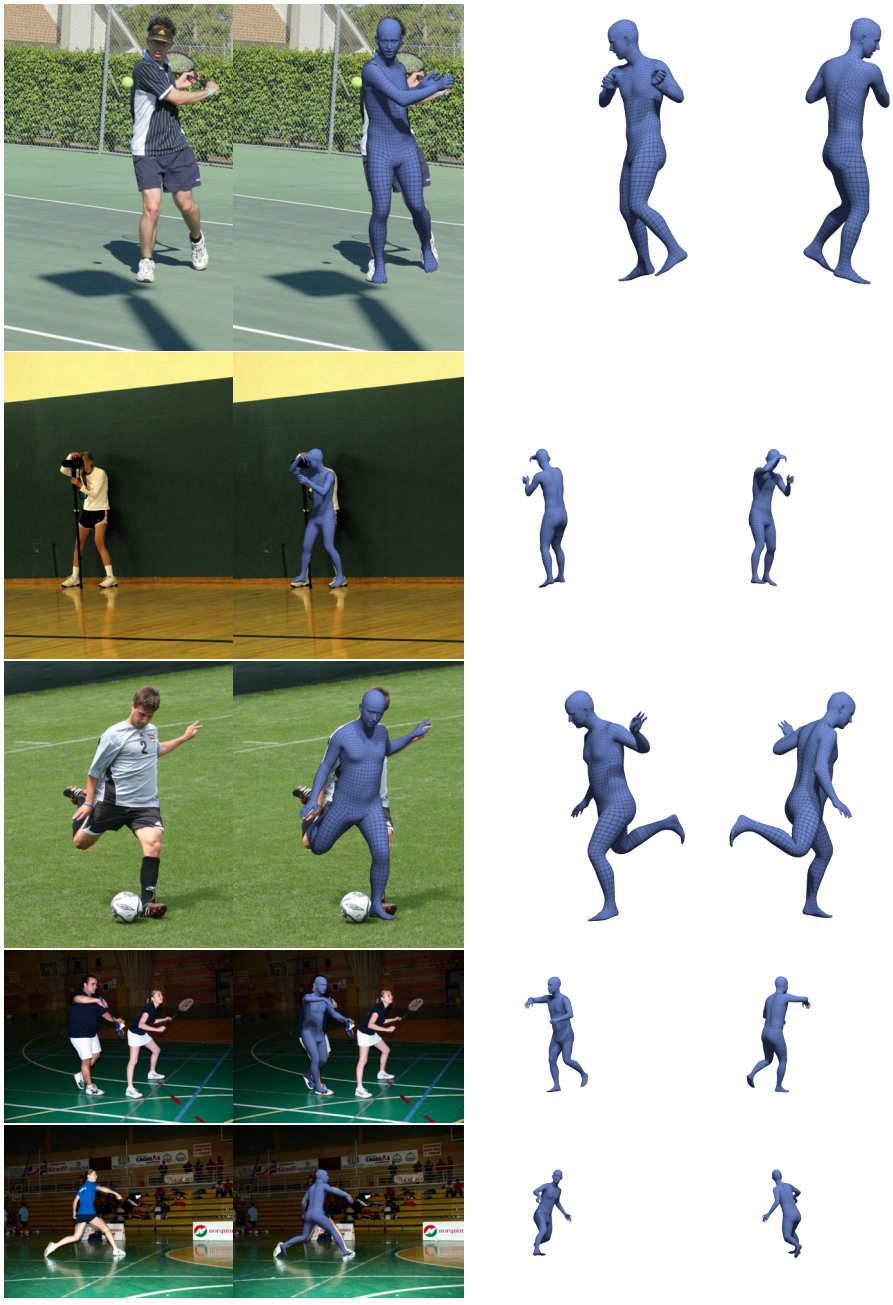


Fig. 23: ExPose results visualized from multiple views. 1. RGB image, 2. overlay, 3. , 4. rotations around the vertical axis

Fig. 24: ExPose results visualized from multiple views. 1. RGB image, 2. overlay, 3. , 4. rotations around the vertical axis

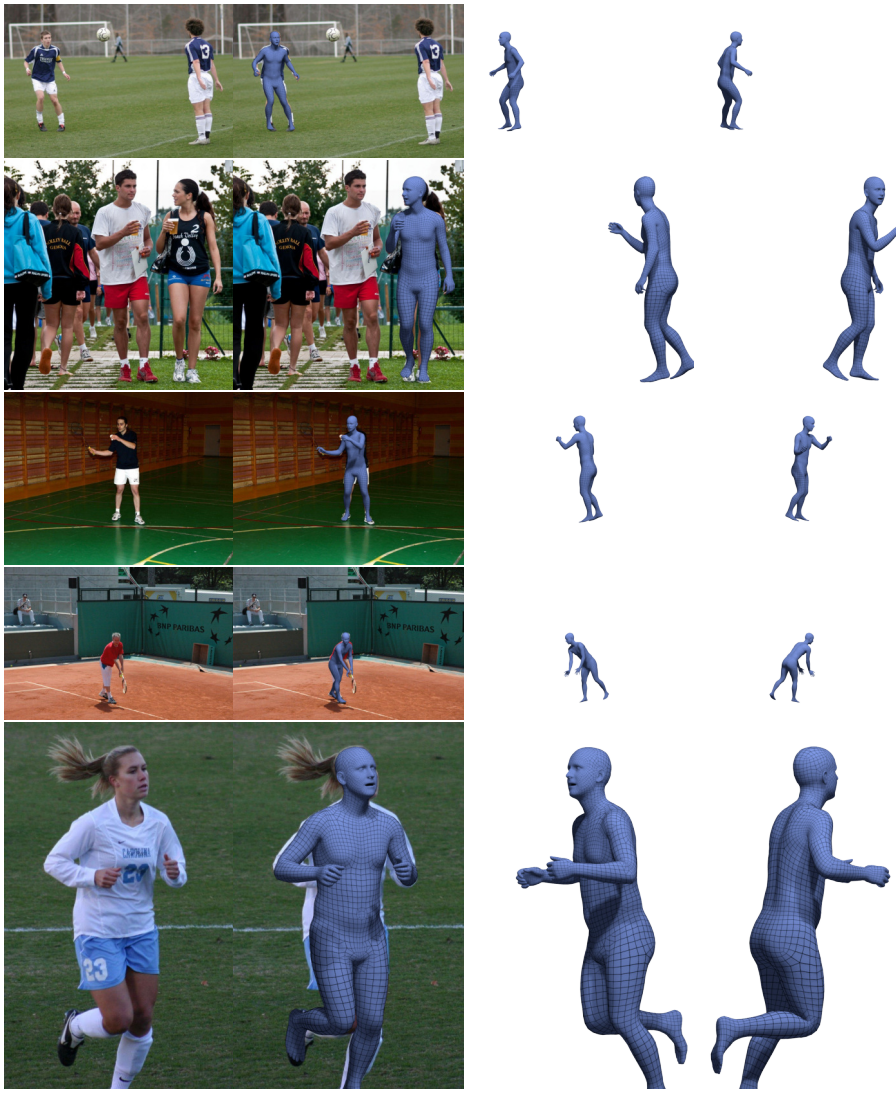


Fig. 25: ExPose results visualized from multiple views. 1. RGB image, 2. overlay, 3. , 4. rotations around the vertical axis

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
2. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7122–7131 (2018)
3. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4396–4405 (2019)
4. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
5. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2252–2261 (2019)
6. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In: European Conference on Computer Vision (ECCV). pp. 614–631 (2018)
7. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, New York, NY, USA, second edn. (2006)
8. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10967–10977 (2019)
9. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5686–5696 (2019)
10. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 813–822 (2019)