AUTOMATIC 3D CHARACTER RECONSTRUCTION FROM FRONTAL AND LATERAL MONOCULAR 2D RGB VIEWS

Alejandro Beacco, Jaime Gallego and Mel Slater

EventLab, Department of Clinical Psychology and Psychobiology,
University of Barcelona, Spain

ABSTRACT

We present an automatic animatable 3D character reconstruction method from frontal and lateral RGB pictures of the person to reconstruct. We propose a method where, after obtaining a 3D reconstruction from a frontal view, we adapt the template of the Skinned Multi-Person Linear Model (SMPL) to it. This frontal modified SMPL is then used in a second step as an input to reconstruct the character from the lateral view, thus obtaining a correct person reconstruction in all dimensions. Our method enables the creation of 3D animatable characters focusing on two main aspects: shape and texture. The obtained results show that the proposed method achieves an accurate automatic 3D character reconstruction from only two views that enhances current methods while offering a reproducible work-flow.

***Index Terms*—** 3D character reconstruction, 3D reconstruction from in the wild pictures, Animatable 3D reconstruction

1. INTRODUCTION

Animatable 3D character reconstruction from unconstrained 2D views is a central process for high level applications that need to recreate 3D models of people in pictures or video sequences. In this paper we address the challenge of reconstructing a realistic 3D virtual character from only two RGB pictures, frontal and lateral, of the target person. The goal is to generate the 3D model from unconstrained pictures or video frames, therefore no specific pose or scenario is required. The generated 3D model of the character is completely animatable to produce new content, for example, for virtual reality.

This challenge has been addressed in the literature with various techniques that are more complex than using 2D monocular RGB images. 3D scanners [1] and multi-view set-ups correctly calibrated and synchronized [2, 3] working in controlled scenarios are examples of the expensive available methods to obtain a high quality 3D model of the characters. Stereo and, more recently, depth cameras have allowed the capture of 3D models with high degree of accuracy [4, 5]. Although these set-ups are becoming more accessible to the general public, they are not as universal and affordable as RGB cameras. The specificity of these techniques and the high degree of preparation required to use them make them only suitable for planned and designed sequences beforehand. Therefore, they cannot be applied to RGB pictures recorded on the spur of the moment, or pictures from the past.

In contrast to the previous methods, using 2D monocular RGB images to obtain a 3D scene reconstruction is the most practical, although challenging, solution. Despite the difficulty of extracting information from RGB images, several proposals in the literature have dealt with 2D RGB image analysis to detect various characteristics and parameters related to the surrounding background and the human body.

Many researchers have developed methods related to 2D human pose detection by obtaining the joint positions from any human detected in a picture [6, 7], or to establish correspondences between 2D human pictures and the corresponding 3D texture model [8]. Some authors propose inferring the 3D pose and/or shape from a single 2D RGB human picture [9, 10, 11]. All these methods are constrained to the possible deformations allowed by the generic skinned animatable models used in the pose estimation. More specifically, these methods use the Skinned Multi-Person Linear Model (SMPL) [12], which is the main reference animatable model used in human pose and shape detection proposals.

In the recent years, some proposals obtained a detailed shape of the characters that appear in the scene based on the deformation of the Skinned Multi-Person Linear Model (SMPL) in the camera view plane [13, 14, 15, 16]. These proposals allow the inclusion of the hair and clothes into the 3D model. The methods presented in [14, 15], need several images from different points of view with the same pose to create the 3D model of the character. Other approaches like [13, 16], have focused on obtaining a 3D model approximation from a frontal image of the person under reconstruction.

The PhotoWakeUp method [13], achieves a realistic and accurate reconstruction of a character from a frontal view picture, obtaining a complete deformation of the SMPL [12] shape. Depths, normals and skinning maps of the SMPL are warped in 2D to match the real silhouette of the person. The problem with this method, requiring a frontal view, is that it is only accurate in two dimensions (the warping ones), with a Z or depth dimension of the generated character completely

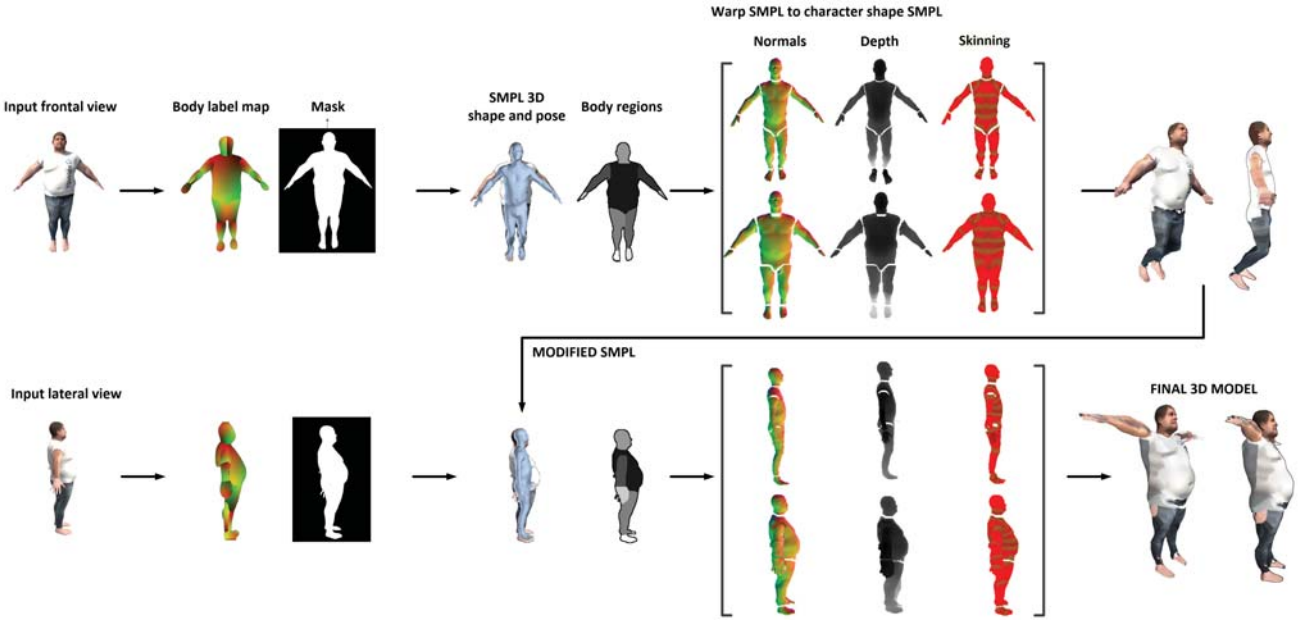

Fig. 1. Work-flow of the proposed method. First, the initial 3D reconstruction from an unconstrained frontal view is processed. Next, the person 3D model is complete with the lateral image of the character. Note that normal maps and depth maps from SMPL are warped to the real shape of the person.

invented based on the SMPL model.

In this work, we address automatic 3D animatable character reconstruction from 2D RGB frontal and lateral pictures with no specific pose requirements. We define a method based on [13] to extract all the necessary information from the RGB pictures to reconstruct a realistic model of the person in the three spatial dimensions.

The results shown in the paper demonstrate that our proposal achieves a 3D character reconstruction, correctly combining information from the frontal and lateral views with no pose constraints, thus creating 3D character models true to the real ones that can be fully animated.

The remainder of the paper is organized as follows: The description of the reconstruction method is explained in Section 2. Section 3 shows the results obtained. Finally, conclusions are presented in Section 4.

2. METHOD

We propose a 3D reconstruction process where first, we adapt the SMPL [12] model to the real shape, pose and texture of the character in the frontal view. In a second step, we adapt the 3D model obtained from the frontal view, to the lateral view in order to use this information to complete the 3D model of the character concerning the shape and texture. To allow this processing, some further methods are used to obtain the complementary pre-processed data: Body label map is obtained by using densepose [8], and the mask of the person to reconstruct is computed processing human semantic segmentation method proposed in [17]. Figure 1 shows the work-flow of the system.

We define a character $C = \langle M, Sk, T \rangle$ by a set of mesh M, skeleton Sk and texture image T. A mesh M is defined by $M = \langle V, F, N, UV, W \rangle$, where V are the vertices, F the faces, N the normals, UV the texture coordinates and W the skinning weights and joint indexes of V. A skeleton Sk is a set $Sk = \langle J, A \rangle$, where J are the 3D joints positions at the current pose P and A the joint transformation matrices going from the binding pose of M to P.

Our method requires an initial frontal image of the character without any pose constraint, f_{init} , to obtain an initial animatable character $C_{\text{init}} = \langle M_{\text{init}}, Sk_{\text{init}}, T_{\text{init}} \rangle$. With a new lateral view $f_{lateral}$ we adapt M_{init} to also fit this view and improve T_{init} with the new color information. We process the two images sequentially, instead of using a global optimization, to allow for a progressive 3D character reconstruction that could be eventually enhanced with several images.

2.1. Initial frontal frame reconstruction

In order to obtain the initial frontal reconstruction we adapt the PhotoWakeUp method [13]. We first use [11] to obtain camera parameters, and shape and pose parameters of the SMPL model [12] that fit it to f_{init}. Note here that, since it only works by matching 2D joint positions, the shape parameters only adjust the body proportions of the SMPL, and not the silhouette. After doing this fitting, we render the depth, normal and skin maps of the obtained SMPL model with the

same camera parameters obtained from the frontal image, and with a back camera generated according to these estimations. According to [13], once the maps are computed, we warp them in the 2D image space to match the silhouette of the character in the picture. Hence, a grid mesh M_{init}^{grid} can be reconstructed by back-projecting the depth pixels with the camera parameters, and the original image projected as texture image T_{init}^{grid}.

Actually, PhotoWakeUp [13] divides the character into 6 body parts (torso, arms, legs and head) to avoid self-occlusion problems, and thus renders and warps each body part separately to finally assemble a final 3D mesh. We noticed that artifacts appeared caused by warping arms and hands together, as well as legs with feet. We therefore propose to use ten body parts instead, by detaching also arms and feet.

The body label map construction presented in [13] relies too much on a correct pose fitting of the SMPL model, but fails too much with poses that are not properly recognized. So we also propose to use [8] as an initial body label map, that we later refine in the same way, instead of the reference SMPL labels.

At this point, we have all the body part meshes assembled in M_{init}^{grid}, and from [11] we can also obtain the skeleton Sk_{init}^{P} in pose P. Therefore we can use linear blend skinning [18] with the inverse of A_{init}^{P} to put both J_{init}^{P} and M_{init}^{grid} into the bind pose of the SMPL model, which is in fact also a T-Pose. This results in M_{init}^{TP} and Sk_{init}^{TP} . Having our mesh and joints in T-Pose, we can align the original template of the SMPL model using Iterative Closest Points (ICP) [19, 20]. The next step consists on wrapping the SMPL template M_{SMPL} into our initial reconstruction in T-Pose M_{init}^{TP} : we find the optimal deformations to apply to M_{SMPL} so that it keeps its topology, but has its shape fitting M_{init}^{TP}. The wrapped template results in our final M_{init}, where we can also project and transfer the texture T_{init}^{grid}, keeping the texture coordinates of the SMPL template UV_{SMPL} . This gives us T_{init} .

To sum up, $C_{\text{init}} = \langle M_{\text{init}}, Sk_{\text{init}}, T_{\text{init}} \rangle$ where $M_{init} = \langle V_{init}, F_{SMPL}, N_{init}, UV_{SMPL}, W_{SMPL} \rangle$, $SK_{init} = \langle J_{init}^{TP}, A^I \rangle$ and T_{init} is obtained by transfering T_{init}^{grid} to M_{init}, where A^I are identity matrices.

2.2. Lateral view refinement

Given M_{init}, and assuming that it is accurate in the frontal view plane, we use the character image from the lateral plane f_L to improve the shape and texture in the invented third dimension. For the new frame f_L we apply almost the same process as for f_{init}, but with some important differences. We still use [11] to obtain camera and pose parameters that fit the SMPL model to f_L. However, we ignore shape parameters and instead, we replace the vertices of the SMPL template by V_{init} before applying linear blend skinning, to render the depth, normal and skin maps with the new pose. This way, the renders that we will warp in 2D as in [13], will not be of the SMPL template, but of our initial reconstruction, which has already the same topology than the SMPL model. Therefore the reconstructed mesh M_L^{grid} will not have an invented third dimension, but instead will have its depth obtained from our initial reconstruction M_{init}. As with the frontal view, we can obtain Sk_L and M_L^{TP} , but now, instead of wrapping M_{SMPL} to it, we wrap M_{init} to M_L^{TP} . This results into a character $C_L = \langle M_L, Sk_L, T_L \rangle$ that has its shape adapted to the two orthogonal views f_{init} and f_L , but not the texture, which is only from the second view f_L.

Although a global optimization of the SMPL could be applied, as in [15] but with less frames, we must keep in mind that the two frames can have the character appearing in different poses. The optimization would then need to take this into account, exponentially increasing the complexity of the search space.

2.3. Texture processing and combination

We still have to combine the two textures T_{init} and T_L into one. Following [13], when creating the image texture T^{grid} of a grid mesh M^{grid} , we either project the color of the frame picture, mirroring it for the back view, or inpaint it because it was occluded. We can therefore save a visibility texture Vis^{grid} storing how every texel was retrieved: 1 if the color was directly projected from the picture, or 0.5 otherwise. We can also generate a view angle texture α^{grid} storing for every texel the angle between the camera view vector and the normal of the surface. As these textures can be also transfered when wrapping the mesh, we end up with all of them in the same UV space of the SMPL model, giving us the oportunity to mix them (see Figure 2). We define a function that combines the two textures T_{init} and T_L using Vis_{init}, Vis_L, α_{init} and α_L in algorithm 1. The function applies symmetry to the head and the body in T_L , and then replaces the pixels that were not visible in T_{init} by the corresponding pixels in T_L, as well as the mirrored pixels from T_{init} that have a greater view angle than in T_L. Pixels are replaced using Poisson blending [21] to mitigate differences in lighting conditions between the two views.

Algorithm 1 $T_{final} = Combine(T_{init}, T_L, Vis_{init}, Vis_L, \alpha_{init}, \alpha_L)$

$T_L^S \leftarrow ApplySymmetryToBodyAndHead(T_L)$
$T_{final} \leftarrow T_{init}$
$indexs1 \leftarrow FindPixels((Vis_{init} \leq 0.5)$ **or** $(Vis_{init} = 0.0$ **and** $Vis_L = 1))$
$indexs2 \leftarrow FindPixels((Vis_{init} = 0.5)$ **and** $\alpha_{init} > \alpha_L)$
$T_{final}(indexs1) \leftarrow T_L^S(indexs1)$
$T_{final}(indexs2) \leftarrow T_L^S(indexs2)$

Finally, we obtain our final reconstructed character $C_{\text{final}} = \langle M_L, Sk_{\text{init}}, T_{\text{final}} \rangle$ where $M_L = \langle V_L, F_{SMPL}, N_L, UV_{SMPL}, W_{SMPL} \rangle$.

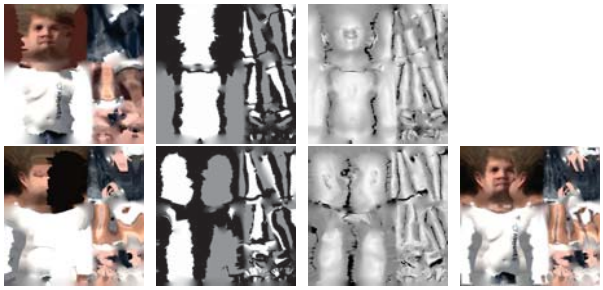

Fig. 2. Color, visibility and view angle textures obtained from a front view (top row) and a lateral view (bottom row). Color (1st column) has been directly projected in white zones of visibility textures (2nd column), mirrored or inpainted in grey zones, and not retrieved in black zones. On view angle textures (3rd column), the angle from the view camera and the normal is lower as the pixel is whiter. All data is used to combine color texture into one texture (4th column) using algorithm 2.3.

3. RESULTS

Figure 3 shows a comparison between a reconstruction with our method and previous methods [16, 15, 13]. Most of these have good results in the frontal view, but only our method and Octopus [15] are able to recover properly the third dimension, although Octopus uses different constraint views with contraint poses. Figure 4 shows four more qualitative results, where we can appreciate how we can achieve reconstructions that are accurate to the two input frames with arbitrary poses.

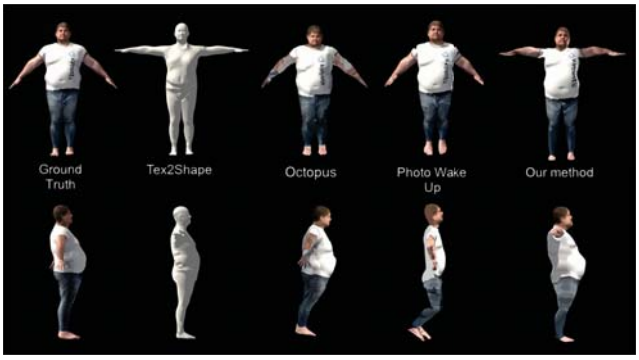

Fig. 3. Comparison results of our method against the ground truth model and three other methods.

Images were scaled to a maximum resolution of 350×450 for performance optimization, except for color texture projection, resulting in processing each character in less than 10 minutes. All experiments were carried out on a PC with an AMD Ryzen 7 3700X 8-Core @ 3.59 GHz CPU.

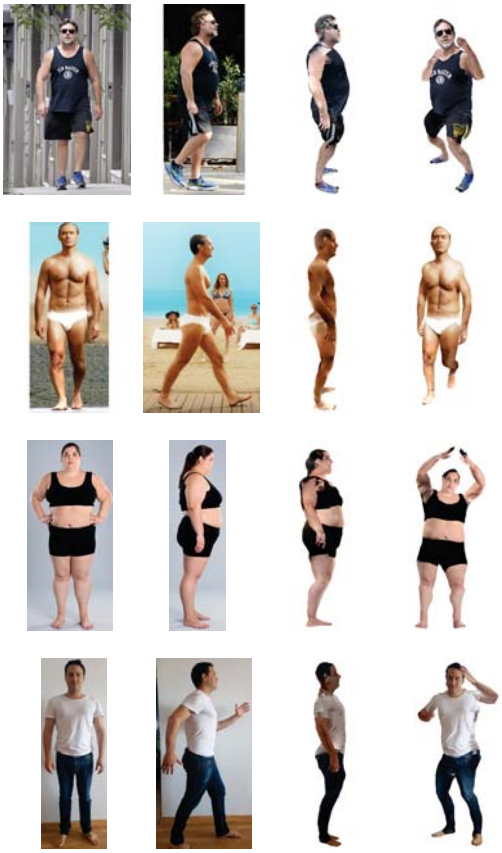

Fig. 4. Four example results. 1st and 2nd columns are input frontal and lateral images with unconstrained poses. 3rd and 4th columns show the resulting 3D character with a retarget animation in two different points of view.

4. CONCLUSIONS

In order to obtain a 3D character from frontal and lateral views, we have presented a workflow that builds upon existing techniques that work with only one view, and then combines these in a novel way. This results in a more accurate and enhanced reconstruction in all dimensions. Using only a frontal view has insufficient information, whereas adding information from an orthogonal lateral view is enough to complete the missing parts.

In future work, we would intend to cancel lighting from the input images to have a better mixing of textures and final lighting of the character. We would also like to replace [11] by SMPL-X [22], which is a newer method that also retrieves hands and face poses while being more accurate.

Acknowledgement

This work is funded by the European Research Council (ERC) Advanced Grant Moments in Time in Immersive Virtual Environments (MoTIVE) number 742989.

5. REFERENCES

[1] Brett Allen, Brian Curless, Brian Curless, and Zoran Popović, "The space of human body shapes: reconstruction and parameterization from range scans," in *Transactions on Graphics*. ACM, 2003, vol. 22, pp. 587–594.

[2] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Computer Vision and Pattern Recognition*. IEEE, 2006, vol. 1, pp. 519–528.

[3] Vladimir Kolmogorov and Ramin Zabih, "Multi-camera scene reconstruction via graph cuts," in *European conference on computer vision*. Springer, 2002, pp. 82–96.

[4] Alexander Weiss, David Hirshberg, and Michael J Black, "Home 3d body scans from noisy image and range data," in *International Conference on Computer Vision*. IEEE, 2011, pp. 1951–1958.

[5] Jing Tong, Jin Zhou, Ligang Liu, Zhigeng Pan, and Hao Yan, "Scanning 3d full human bodies using kinects," IEEE, 2012, vol. 18, pp. 643–650.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proc. Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 7291–7299.

[7] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu, "Rmpe: Regional multi-person pose estimation," in *Proc. Int. Conf. on Computer Vision*. IEEE, 2017, pp. 2334–2343.

[8] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proc. Conf. on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 7297–7306.

[9] 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 *European Conference on Computer Vision*. Springer, 2016, pp. 561–578.

[10] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler, "Unite the people: Closing the loop between 3d and 2d human representations," in *Proc. of Conf. on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 6050–6059.

[11] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7122–7131.

[12] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black, "SMPL: A skinned multi-person linear model," ACM, 2015, vol. 34, pp. 248:1–248:16.

[13] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman, "Photo wake-up: 3d character animation from a single photo," in *Proc. Conf. on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 5908–5917.

[14] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll, "Video based reconstruction of 3d people models," in *Proc. Conf. on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 8387–8397.

[15] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll, "Learning to reconstruct people in clothing from a single RGB camera," in *Proc. Conf. on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 1175–1186.

[16] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor, "Tex2shape: Detailed full human body geometry from a single image," in *Proc. Int. Conf. on Computer Vision*. IEEE, 2019, pp. 2293–2303.

[17] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin, "Instance-level human parsing via part grouping network," in *Proc. Eur. Conf. on Computer Vision (ECCV)*. Springer, 2018, pp. 770–785.

[18] J. P. Lewis, Matt Cordner, and Nickson Fong, "Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation," in *Proc. Conf. on Computer Graphics and Interactive Techniques*. ACM, 2000, pp. 165–172.

[19] Yang Chen and Gérard Medioni, "Object modelling by registration of multiple range images," Butterworth-Heinemann, 1992, vol. 10, pp. 145–155.

[20] Paul J. Besl and Neil D. McKay, "A method for registration of 3-d shapes," in *Trans. Pattern Analysis and Machine Intelligence*. IEEE, 1992, vol. 14, pp. 239–256.

[21] J. Matías Di Martino, Gabriele Facciolo, and Enric Meinhardt-Llopis, "Poisson Image Editing," *Image Processing On Line*, vol. 6, pp. 300–325, 2016.

[22] 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 *Proc. Conf. on Computer Vision and Pattern Recognition*. IEEE, 2019.