Physio Motion Transfer: A retargeting approach to transfer human motion and appearance in monocular videos for physiotherapy applications

1st Fábio Oliveira
2Ai - School of Technology, IPCA, Barcelos, Portugal
fabiodiogo29@gmail.com

2nd José Brito
2Ai - School of Technology, IPCA, Barcelos, Portugal
https://orcid.org/ 0000-0002-4544-4698

***Abstract*—Transferring human motion and appearance between human bodies remains one of key challenges in Computer vision due to the complexity of the human body. Despite the advances from recent research projects in this fields, past methods still perform poorly but the results are evolving fast. Using this technology for physiotherapy purposes has not been done yet so we thought that it might be the right application for this kind of technology. In this paper, we propose a retargeting approach to transfer human motion and appearance in monocular videos for physiotherapy purposes. Given one or more patient(source) images and one physiotherapy movement video, we firstly create a 3D digital avatar of the patient and then, we retarget it based on the pose extracted from the movement video frames. The code for the project is open-source and available at *github.com/fabioo29*.**

***Index Terms*—Computer Vision, Human Image Synthesis, Motion Retargeting**

I. Introduction

Creating realistically-looking and articulated human avatars is a challenging task with a vast potential applications in character animation, reenactment, virtual clothes try-on, movie or game making, etc. In the past years we've seen this kind of technology being used in a large amount of these fields [15], [22], [42], [56] but never on rehab/physiotherapy applications. The methods used for those applications are mainly generative adversarial network(GAN) based [24], and because of that the output can be quite random without a good loss control approach which can be difficult to get, and the runtime can be really slow due to the complexity of neural network used in *GANs* but it's getting faster as computers power increases along the years. In this kind of approaches(appearance and motion transfer) without using *GANs*, there's a need to create a digital avatar to retarget it later. In 2015, *Matthew Loper et al.* changed the digital human animation field introducing the *SMPL* model, a 3D digital model controlled by 82 parameters [45] and from that moment on, researchers are using it in any application where there's a need of animating a human body.

Nowadays researchers upgraded the *SMPL* model into others *SMPL variations* like *SMPL-X* [53] which has new face and hands

parameters added to the original *SMPL* model for a better info quality in those spots. We've seen these human body models being used in motion imitation [8], [48], [67], [73], appearance transfer [59], [82] and novel view synthesis [74], [85] applications. In this paper we propose a shape-aware appearance and body motion transfer with a interface to help on rehab/physiotherapy purposes. To reach this outcome we firstly create a 3D digital avatar from monocular shots of a person, after that we extract the colored texture from the body, then we deform the model to match the body clothes and hair, and finally we retarget it to match other person movement. For the frontend we present a user friendly interface for the user to carefully follow up his avatar movement.

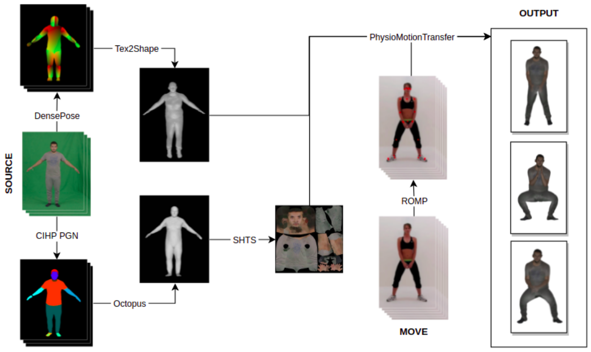

Fig. 1. PhysioMotionTransfer state-of-art method for human motion and appearance transfer for human bodies in monocular videos.

II. PREVIOUS WORK

A. Human Motion Imitation

Significant advances have been recently developed to estimate the human skeleton and consequently the body pose. Often pose-estimation is posed as the estimation of 2D or 3D keypoints, corresponding to anatomical joints or landmarks [13], [14], [69]. In contrast, recent advances use richer representations of the 3D body surface in the form of parametric [12], [33], [52], [55] or non-parametric [37], [65], [77] models. To estimate bodies from images, many methods break the problem down into stages. First, they estimate some intermediate representation such as 2D joints [12], [25], [26], [28], [33], [49], [55], [76], [86], silhouettes [1], [28], [55], part labels [52], [64] or dense correspondences [27], [63]. Then, they reconstruct the body pose out of this proxy information,

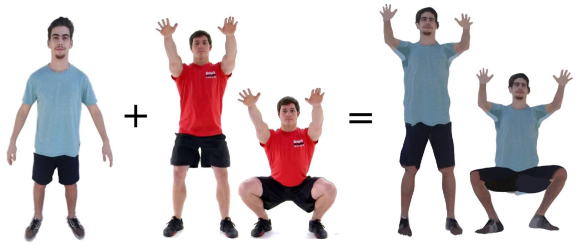

Fig. 2. Animated avatar(right) given source image(left) and movement video(middle).

by either using it in the data term of an optimized energy function [12], [28], [81] or "lifting" it using a trained regressor [33], [49], [52], [76]. Due to ambiguities in lifting 2D to 3D, such methods use various priors for regularization, such as known limb lengths [38], a pose prior for joint angle limits [2], or a statistical body model [12], [28], [52], [55] like SMPL [45]. The above 2D proxy representations have the advantage that annotation for them is readily available. Their disadvantage is that the eventual regressor does not get to exploit the original image pixels and errors made by the proxy task cannot be overcome. Other methods predict 3D pose directly from RGB pixels. Intuitively, they have to learn a harder mapping, but they avoid information bottlenecks and additional sources of error. Most methods infer 3D body joints [43], [54], [71], [72], [75], parametric methods estimate model parameters [33], [34], [36], while non-parametric methods estimate 3D meshes [37], depth maps [20], [70] voxels [77], [87] or distance fields [65], [66]. Datasets of paired indoor images and MoCap data [29], [68] allow supervised training, but may not generalize to in-the-wild data. To account for this, *Rogez and Schmid* [62] augment these datasets by overlaying synthetic 3D humans, while *Kanazawa et al.*, [33] include in-the-wild datasets [6], [30], [31], [44] and employ a re-projection loss on their 2D joint annotations for weak supervision.

B. Human Appearance Transfer

Human appearance modeling or transfer is a vast topic, especially in the field of virtual try-on applications, from computer graphics pipelines [58] to learning based pipelines [60], [83]. Graphics based methods first estimate the detailed 3D human mesh with clothes via garments and 3D scanners [84] or multiple camera arrays [41], and then human appearance with clothes is capable of being conducted from one person to another based on the detailed 3D mesh. Modeling the human body and generating the novel view texture are the two main components for this process. Modeling the human body is a long-standing problem in computer vision. Given a densely distributed multicamera system, one can make use of multi-view stereo methods [35] for reconstructing the human body [19], [21], [35]. More advanced systems allow reconstruction of body shape under clothing [79], [80], [84], joint shape, pose and clothing reconstruction [58], or capture body pose and facial expressions [32]. However, such setups are expensive and require complicated calibration. Texture generation is an essential task for modeling a realistic virtual character, since a texture image can describe the material properties that cannot be modeled by the surface geometry. The key of a texture generation method is how to combine texture fragments created from different views. Many early works blend the texture fragments using weighted averaging across the entire surface [10], [16], [51], [57]. Others make use of mosaicing strategies, which yields sharper results [9], [40], [50], [61]. [39] is the first to formulate texture stitching as a graph cut problem. Such formulation has been commonly used in texture generation for multi-view 3D reconstruction. However, without accurately reconstructed 3D geometry and registered images, these methods usually suffer from blurring or ghosting artifacts. To this end, many methods focus on compensating registration errors such as [11], [17], [18], [78].

C. Human Pose Classifier

Quantifying the level of correctness in completing prescribed exercises is important for the development of tools and devices in support of home-based rehabilitation. The movement assessment in existing studies is typically accomplished by comparing a patient's performance of an exercise to the desired performance by healthy participants. Several studies in the literature on exercise evaluation employed machine learning methods to classify the individual repetitions into correct or incorrect classes of movements. Methods used for this purpose include Adaboost classifier, k-nearest neighbors, Bayesian classifier, and an ensemble of multi-layer perceptron NNs [7]. The outputs in these approaches are discrete class values of 0 or 1 (i.e., incorrect or correct repetition). However, these methods do not provide the capacity to detect varying levels of movement quality or identify incremental changes in patient performance over the duration of the rehabilitation program. Our approach relies on the ability to detect pose correctness calculating the formed degrees between some important human pose keypoints for our application.

III. IMPLEMENTATION

This work aims to create a digital avatar of a given human body(monocular shots) and add a movement to it. To create a digital avatar there's a few steps we need to follow. We will split our approach in three main steps which are **Avatar colored texture**, **Avatar textured mesh**, **Retargeting avatar** and **User interface**.

A. Avatar Colored Texture

As we mentioned before there are several methods to extract and save the color from a body in a monocular image to the avatar. The method used in this paper work as follows.

Person detector. In the first place there's a need to identify the person and the pose of the body in the picture. There are several methods to do this. In this paper we use *OpenPose* [14] to detect the human body on the picture and estimate

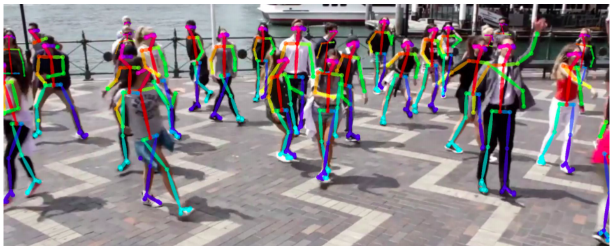

Fig. 3. OpenPose state-of-art method - 25 multi-body keypoints estimation

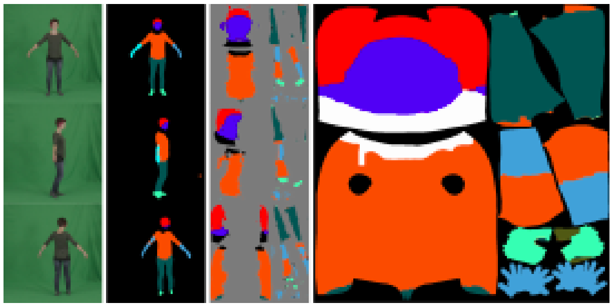

Fig. 4. Alldieck1 et al. in Pic2Tex - Fully textured instance segmentation UV map given monocular images(left), body orientation(Octopus) and instance segmentation correspondent images(2nd collumn)

its current pose. Each detected pose has 95 labeled keypoints including face keypoints.

SMPL body model. After extracting the pose keypoints from a given source picture we then feed the data into *Octopus* [3] where the *OpenPose* data its converted into a *SMPL* [46] body model. As we mentioned before the *SMPL* body model is really useful way to recreate a naked body digitization. There are several parameters to control the model itself but at this point we only have interest in saving the body orientation parameter to check to where the body is facing.

Body instance segmentation. This is the last data component needed for *Tex2Shape* to successfully extract the colored UV map for the body model.

Body instance segmentation is the process of labeling different parts of the body given a picture of a human body. There's a few methods to to this. In this paper we used *Instance-level Human Parsing via Part Grouping Network* [23] for this purposes that can label up to 20 categories.

Body color extraction. Of all the methods we've seen for body color extraction purposes, we've found out that the one from *Alldieck et al.* [4] its the way to go because of it's simplicity and efficiency. Because we feed the model with the body orientation, the body pose and all the body parts labeled as input data, it can take different partial views from the same person and merge all extracted portions for each view into one fully colored UV map.

B. Avatar Textured Mesh

The goal of this step is to create an animatable 3D model of a subject from a single photograph. The model should reflect the subject's body shape and contain details such as hair and clothing with garment wrinkles. Details should be present also on body parts that have not been visible in the input image, e.g. on the back of the person.

Tex2Shape. To reach this results *Alldieck et al.* [5] converts a monocular human body into a UV unwrap partial map which is then feed into a GAN that tries to guess the remaining texture parts that are not seen in the UV map. After that, the UV map is converted into two UV maps (normal map and vector displacement map) which is then equivalent to the 3D output mesh of the initial body mesh but with clothes and hair.

UV map unwrap. In 2018 *Guler et al.* launched *DensePose* [27] which is a R-CNN type network that can unwrap any human body into a plain texture map with coordinates (UV map). This is really useful to match colored textured maps into 3D meshes using UV map coordinates.

C. Retargeting Avatar

Retargeting an avatar its the name called to the process of giving movement to a static 3D avatar. We approach this process as follows.

ROMP. It adopts a simple multi-head design with a backbone and three head networks. Given a single RGB image as input, it outputs a Body Center heatmap, Camera map, and SMPL map, describing the detailed information of the estimated 3D human mesh. In the Body Center heatmap, *ROMP* predict the probability of each position being a human body center. At each position of the Camera/SMPL map, it predict the camera/SMPL parameters of the person that takes the position as the center. From this method *PhysioMotionTransfer* saves the pose estimation for each frame in the movement video and saves also the rendering process.

Posed deformed mesh. Because *ROMP* uses purple SMPL model, there's a lack of information about the clothes, hair features and also the texture color of each material in an pictured body. Because of that we take the data from early steps (*Body color extraction* and *Avatar Textured Mesh*) and the pose given in ROMP to merge it together to what we call a 3D realistic avatar using human motion and appearance transfer.

D. User Interface

The user interface has a huge impact on a application like this where the user needs to replicate the movement

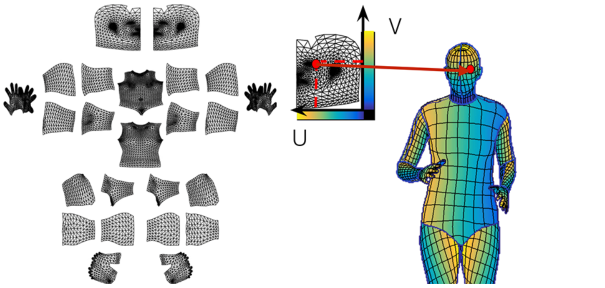

Fig. 5. Guller et al. in DensePose - UV texture mesh unwrap with coordinates given a human body monocular image

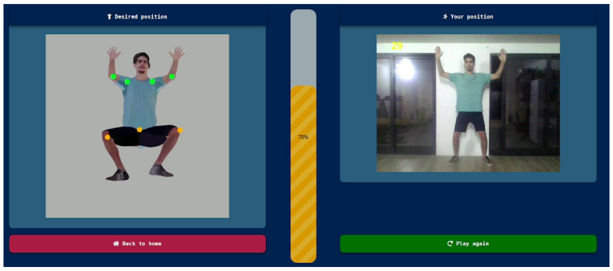

Fig. 6. User interface Web app for physiotherapy purposes. Rendered video from PhysioMotionTransfer method(left), User trying to replicated his movement on avatar(right)

on screen but also need to keep track of his pose while doing the movement.With that being said we decided to developed a web application using Django framework which is the Web App framework compatible for python. The advantage of it being a web app is that it can be hosted in a server and accessible from any device even a mobile device.

Starting the inference. In the web app home page the user is asked for the source images (input images for the avatar digitization) and one movement video (desired movement for the avatar retargeting). After that the *PhysioMotionTransfer* state-of-art method will run and the user will be redirected to the playground page.

Playground page. This page was designed to fit the Physiotherapy/Rehab needs. The user can see the output video (generated from PhysioMotionTransfer) and the real time camera output. The goal is for the body on the right to match the body on the left and whenever both bodies match within a 80 percentage rate the video on the left will play and stop at the next position (30 frames after) and both bodies have to match again.

Pose classifier. To compare both bodies there was a need of adding a pose classifier to the app. The way we approach this need is by extracting the most important pose keypoints for this situation and compare the formed degrees between some of them and use it to compare both poses. To extract the pose keypoints we've now used *MediaPipe* [47] by Google because it's one of the fastest production ready ML pipeline for human pose estimation and according to our tests it was running 30*fps* in a *GTX 1050 ti* which was ideal for our application.

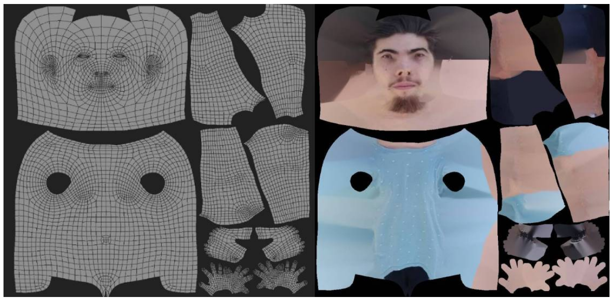

Fig. 7. PhysioMotionTransfer body colored texture extraction to UV map using Pic2Tex

IV. RESULTS

A. Discussion

Because *Tex2Shape* and *Pic2Tex* used on *Body color extraction* were trained using *People Snapshot Public*, we can ensure the quality of the outputs/predictions could be better. The dataset used to train these models is called *People Snapshot Public* and it contains 24 samples of people rotating in A-Pose. Because of the lack of samples in the dataset the results are good but not too realistic yet. In a future work we would add more samples to this dataset to improve the quality of the model deformation to match clothes and hair with a higher quality, and also to improve the colored texture extraction in the UV map.

B. Limitations

Due to our limitations in computational power sometimes we can't run some of the models present in *PhysioMotionTransfer* so whenever it happens we can't really get an output. With that being said we could fix this issue by replacing current models with lighter version of them by training each model just for the purpose of this application and nothing more. In the case of *OpenPose* as an example, we could build and train a lighter neural network just for A-Poses as we are expecting only this kind of poses as input and don't really need the model to be trained on any other poses.

As we said before that's a lot that can be done to improve the models quality like training these networks using a bigger dataset but because of our computational power limitations we couldn't really do much about that since we can't train them.

In the beginning we were also limited to *Python2.7* but after a few upgrades we could successfully upgrade all the code to *3.7* version and it's now running fine on a *Docker container* using *Ubuntu 16.04* and *Python 3.7*.

C. Future Work

As we tackled before, there's a huge need to create a bigger and better dataset and we could work on that in a future work. For the dataset we would need more people rotating in A-Pose, every person mesh and also the UV textured colored map of each person. This kind of work could consume a lot of time but it's completely doable.

There's also a lack of facial expression and hands information in the current *SMPL* model so we encourage future researchers to upgrade from this mesh model to the *SMPL-X* one which already has hands and facial expression in it.

In this work we used *Tex2Shape* state-of-art model to generate a deformed model given a input picture but the problem with this method is that we can't merged all deformed models into one, given multiple pictures so in our work we choose the front view model all the times, excluding any other also good deformations.

V. CONCLUSIONS

In this work we've taken past works/methods on human body reconstruction to successfully create a 3D digital representation of a human body and then animate it based on a given movement. We used this approach for physiotherapy/rehab purposes to help any person with difficulties through his rehabilitation process by copying his avatar movements. All the code for this method is open source in ***Github.com*** and the code is adapted so the current models can be changed over the time so if there's a new pose estimation in the future it could be easily be adapted into our code.

References

[1] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):44–58, 2006.

[2] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1446–1455, 2015.

[3] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera, 2019.

[4] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Detailed human avatars from monocular video, 2018.

[5] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2shape: Detailed full human body geometry from a single image, 2019.

[6] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[7] I. ar and Y. Akgul. A computerized recognition system for the home-based physiotherapy exercises using an rgbd camera. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22, 05 2014.

[8] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag. Synthesizing images of humans in unseen poses, 2018.

[9] A. Baumberg. Blending images for texturing 3d models. *Proceedings of the British Machine Vision Conference*, 01 2003.

[10] F. Bernardini, I. Martin, and H. Rushmeier. High-quality texture reconstruction from multiple scans. *IEEE Transactions on Visualization and Computer Graphics*, 7(4):318–332, 2001.

[11] S. Bi, N. K. Kalantari, and R. Ramamoorthi. Patch-based optimization for image-based texture mapping. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2017)*, 36(4), 2017.

[12] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image, 2016.

[13] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d and 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[14] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019.

[15] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody dance now, 2019.

[16] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, page 11–20, New York, NY, USA, 1996. Association for Computing Machinery.

[17] M. Eisemann, B. De Decker, M. Magnor, P. Bekaert, E. de Aguiar, N. Ahmed, C. Theobalt, and A. Sellent. Floating textures. *Computer Graphics Forum (Proc. of Eurographics EG)*, 27(2):409–418, Apr 2008. Received the Best Student Paper Award at Eurographics 2008.

[18] Y. Fu, Q. Yan, L. Yang, J. Liao, and C. Xiao. Texture mapping for 3d reconstruction with rgb-d sensor. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4645–4653, 2018.

[19] S. Fuhrmann, F. Langguth, and M. Goesele. MVE - A Multi-View Reconstruction Environment. In R. Klein and P. Santos, editors, *Eurographics Workshop on Graphics and Cultural Heritage*. The Eurographics Association, 2014.

[20] V. Gabeur, J.-S. Franco, X. Martin, C. Schmid, and G. Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images, 2019.

[21] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 873–881, 2015.

[22] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, and P. Luo. Parser-free virtual try-on via distilling appearance flows, 2021.

[23] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin. Instance-level human parsing via part grouping network, 2018.

[24] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014.

[25] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3d structure with a statistical image-based shape model. volume 1, pages 641–647 vol.1, 11 2003.

[26] P. Guan, A. Weiss, A. Balan, and M. Black. Estimating human shape and pose from a single image. pages 1381–1388, 09 2009.

[27] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild, 2018.

[28] Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, I. Akhter, and M. J. Black. Towards accurate markerless human shape and pose estimation over time, 2018.

[29] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

[30] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, pages 12.1–12.11. BMVA Press, 2010. doi:10.5244/C.24.12.

[31] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR 2011*, pages 1465–1472, 2011.

[32] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies, 2018.

[33] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose, 2018.

[34] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video, 2019.

[35] R. Koch, M. Pollefeys, and L. Van Gool. Multi viewpoint stereo from uncalibrated video sequences. In H. Burkhardt and B. Neumann, editors, *Computer Vision — ECCV'98*, pages 55–71, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

[36] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop, 2019.

[37] N. Kolotouros, G. Pavlakos, and K. Daniilidis. Convolutional mesh regression for single-image human shape reconstruction, 2019.

[38] H.-J. Lee and Z. Chen. Determination of 3d human body postures from a single view. *Computer Vision, Graphics, and Image Processing*, 30(2):148–168, 1985.

[39] V. Lempitsky and D. Ivanov. Seamless mosaicing of image-based texture maps. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2007.

[40] H. Lensch, W. Heidrich, and H.-P. Seidel. A silhouette-based algorithm for texture registration and stitching. *Graphical Models*, 63:245–262, 07 2001.

[41] V. Leroy, J.-S. Franco, and E. Boyer. Multi-view dynamic shape refinement using local temporal integration. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3113–3122, 2017.

[42] K. Li, M. jin Chong, J. Zhang, and J. Liu. Toward accurate and realistic outfits visualization with attention to details, 2021.

[43] S. Li, W. Zhang, and A. B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation, 2015.

[44] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context, 2015.

[45] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. 34(6), 2015.

[46] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015.

[47] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann. Mediapipe: A framework for building perception pipelines, 2019.

[48] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. V. Gool. Pose guided person image generation, 2018.

[49] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation, 2017.

[50] W. Niem and J. Wingbermuhle. Automatic reconstruction of 3d objects using a mobile monoscopic camera. In *Proceedings. International Conference on Recent Advances in 3-D Digital Imaging and Modeling (Cat. No.97TB100134)*, pages 173–180, 1997.

[51] E. Ofek, F. Shilat, A. Rappoport, and M. Werman. Multiresolution textures from image sequences. *IEEE Computer Graphics and Applications*, 17(2):18–29, 1997.

[52] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation, 2018.

[53] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[54] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose, 2017.

[55] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3d human pose and shape from a single color image, 2018.

[56] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang, S. Zhang, P. Wu, B. Zhou, and W. Zhang. Deepfacelab: Integrated, flexible and extensible face-swapping framework, 2021.

[57] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin. Synthesizing realistic facial expressions from photographs. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '98, page 75–84, New York, NY, USA, 1998. Association for Computing Machinery.

[58] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Trans. Graph.*, 36(4), July 2017.

[59] A. Popa, A. Zanfir, M. Zanfir, and C. Sminchisescu. Human appearance transfer. 2018.

[60] A. Raj, P. Sangkloy, H. Chang, J. Hays, D. Ceylan, and J. Lu. Swapnet: Image based garment transfer. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 679–695, Cham, 2018. Springer International Publishing.

[61] C. Rocchini, P. Cignoni, C. Montani, and R. Scopigno. Multiple textures stitching and blending on 3d objects. 07 1999.

[62] G. Rogez and C. Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[63] Y. Rong, Z. Liu, C. Li, K. Cao, and C. C. Loy. Delving deep into hybrid annotations for 3d human recovery in the wild, 2019.

[64] N. Rueegg, C. Lassner, M. J. Black, and K. Schindler. Chained representation cycling: Learning to estimate 3d human pose and shape by cycling between representations, 2020.

[65] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization, 2019.

[66] S. Saito, T. Simon, J. Saragih, and H. Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.

[67] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe. Deformable gans for pose-based human image generation, 2018.

[68] L. Sigal, A. Balan, and M. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87:4–27, 03 2010.

[69] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping, 2017.

[70] D. Smith, M. Loper, X. Hu, P. Mavroidis, and J. Romero. Facsimile: Fast and accurate scans from an image in less than a second, 2019.

[71] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression, 2017.

[72] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei. Integral human pose regression, 2018.

[73] Y. Sun, Q. Bao, W. Liu, Y. Fu, B. Michael J., and T. Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021.

[74] Y.-T. Sun, Q.-C. Fu, Y.-R. Jiang, Z. Liu, Y.-K. Lai, H. Fu, and L. Gao. Human motion transfer with 3d constraints and detail enhancement, 2021.

[75] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured prediction of 3d human pose with deep neural networks, 2016.

[76] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[77] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3d human body shapes, 2018.

[78] M. Waechter, N. Moehrle, and M. Goesele. Let there be color! large-scale texturing of 3d reconstructions. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 836–850, Cham, 2014. Springer International Publishing.

[79] S. Wuhrer, L. Pishchulin, A. Brunton, C. Shu, and J. Lang. Estimation of human body shape and posture under clothing. *Computer Vision and Image Understanding*, 127:31–42, Oct 2014.

[80] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Estimation of human body shape in motion with wide clothing. volume 9908, pages 439–454, 10 2016.

[81] A. Zanfir, E. Marinoiu, and C. Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes: The importance of multiple scene constraints. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018.

[82] M. Zanfir, A.-I. Popa, A. Zanfir, and C. Sminchisescu. Human appearance transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[83] M. Zanfir, A.-I. Popa, A. Zanfir, and C. Sminchisescu. Human appearance transfer. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5391–5399, 2018.

[84] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences, 2017.

[85] B. Zhao, X. Wu, Z.-Q. Cheng, H. Liu, Z. Jie, and J. Feng. Multi-view image generation from a single-view, 2018.

[86] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.

[87] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. Deephuman: 3d human reconstruction from a single image, 2019.