Avatarify Yourself: 3D Animatable Mesh From a Single RGB Image

Vera Milovanović, Agniv Sharma, Anna Manasayan

1 Abstract

Abstract. In this project, we aim to create an animatable avatar of a person using a single RGB image as input. Additionally, we make a comparison between some of the existing 3D pose and texture reconstruction methods, as well as animation methods to find failure cases and define future research problems. The avatar is animated using motion data from the AMASS dataset [4]. The resulting avatar is textured and capable of realistic movement.

2 Introduction

Human avatars have many applications such as games,movies AR/VR, virtual try-ons, Metaverse. Inspired by this, in the scope of this project we create an animatable avatar of a person using a single RGB image, taken by a mobile phone camera. This comes with many challenges: (1) *3D mesh reconstruction from a single image* due to the lack of depth data, self-occlusion, and out-of-distribution poses, (2) *animation* as the obtained mesh is not animatable, so fitting e.g. SMPL[3] is needed, (3) *texture reconstrucion* as the backside of a person is not available and it should be seamless. In our project we tackle all of these challenges and discuss limitations encountered during the experimentation phase, along with potential avenues for future research and improvement. We also showcase the capabilities of our avatarification pipeline and provide insights into the current state-of-the-art methods. The code and animation videos are available here.

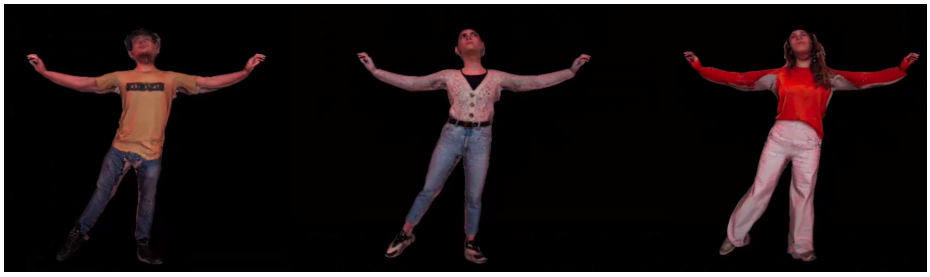

Fig. 1: Textured animated avatar from single RGB image

3 Related work

Regarding geometric representation, we categorize the mainstream clothed human reconstruction approaches into (1) *implicit* and (2) *explicit*. One of the most prominent models form the fist category is ICON [8] that reconstructs clothed human surfaces in different postures by utilizing initial normal maps generated from the predicted SMPL/SMPL-X model [3] [5]. It achieves robustness to unfamiliar poses, but sacrifices its capacity for generalization across different clothing topologies, particularly loose ones. A number of recent approaches reconstructs high-quality 3D texture or geometry from a single RGB image by making use of a texture map representation, which is used to estimate geometric or color details. One such work is Tex2Shape [1], that reconstructs high quality 3D geometry by regressing displacements in an unwrapped UV space. However, this type of approach is constrained by the topology of the template mesh and the topology chosen for the UV parameterization. A constraint on topology of the template mesh leads to failure in generalization to different types of hair or challenging clothing such as skirts and using UV parametrization can cause for example visible texture seams artifacts. The topology constraint problem can be avoided by using Implicit Function-based methods that are topology agnostic, so they recover free-form geometry. One of the most distinguished examples is PIFuHD [7], that serve us as a baseline for 3D reconstruction. We discuss it in the Method section.

4 Method

4.1 ECON: Explicit Clothed humans Optimized via Normal integration

Given single RGB image ECON performs human reconstructs. The model is structured in three sequential steps: (1) *front and back normal reconstruction*, (2) *front and back surface reconstruction*, (3) *full 3D shape completion*. Methods prior to ECON such as ICON or PIFuHD accurately estimate front normals, however because of a lack of cues back normal estimation is a challenge for those methods leading to artifacts such as over-smoothed surfaces. To tackle the issue ECON using ICONs back normal predictor adds an extra loss term called MRF loss which enhances the local details by minimizing the difference between the predicted normals and ground truth in feature space. To make the normal maps more robust to various poses the method also minimizes 2D body landmarks and SMPL-X body. After obtaining the normals in the next step ECON reconstructs front and back surfaces. It is expected for the reconstructed surface to have three properties such as high-frequency surface details agree with predicted clothed normal maps, low-frequency surface variations, including discontinuities, agree with SMPL-X's ones, and the depth of the front and back silhouettes are close to each other. Front and back reconstruction is done combined by a new method called d-BiNI which solves the optimisation problem by taking normals and front and back coarse body depth image rendered from the SMPL-X mesh. Finally,

the last step is shape completion. Combining the reconstructed surfaces from the front and back is straightforward when there are no occlusions. However, when occlusions are present, additional techniques are required to fill in the missing parts. The proposed method called $ECON_{IF}$ combines PSR completion with inpainting using IF-Nets+.

4.2 PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization

PIFuHD is a end-to-end trainable method for a high-fidelity 3D reconstruction of clothed humans from a single RGB image. It leverages full 1k-resolution input image, preserving details in the original input without any post-processing. The method consists of two modules: (1) a *coarse level* that captures global information by taking x0.5 downsampled input image and predicted frontside and backside normal maps, and producing a lower resolution 3D embedding as a intermediate step towards low-resolution occupancy and (2) a *fine level* that introduces details to the coarse representation by taking frontside and backside normal maps, 3D embedding from the coarse level module, and producing a high-resolution occupancy. Since the original input is a single RGB image, the method has to infer the backside bacause it is not observed in the input. It is done by utilizing the predicted normal maps in the image space at the start of the method's pipeline, which guides 3D geometry reconstruction described in (2). This method significantly improves the geometric details, such that it is able to recover fingers, facial features and clothing folds. However, this method doesn't recover texture. Another drawback of this method is that it overfits to the body poses in the training data, e.g. fashion poses. They fail to generalize to challenging out-of-distribution poses, producing disembodied limbs. In contrast to PIFuHD, our method recovers texture and produces animatable avatar. Additionally, by using ECON as a method for 3D reconstruction, we are able to get good 3D reconstruction for out-of-distribution poses as well.

4.3 Animation

To animate the mesh generated by the image-to-mesh pipeline, our first step is to convert it into an animatable avatar. This involves registering an SMPL model to the mesh. The registration process finds the nearest SMPL vertex to the given mesh output and uses these correspondences to create a linear blend skinning model. This allows for the transfer of SMPL deformation to the output mesh. Since ECON natively utilizes SMPL mesh as a prior in its pipeline, it is better suited for our avatarification process. Therefore, the animation experiments are limited to ECON.

4.4 Texture

Most image-to-mesh methods, other than recent ones like TeCH [2], only output a base mesh without any texture. While ECON does produce the frontal texture

as output, it is also unoptimized. Therefore, we decide to use TEXTure [6] to produce full-body texture. TEXTure paints the mesh from different viewpoints, and at each step, it uses stable diffusion conditioned on the user's text prompts to improve the texture quality for different views. In our experiments, we use our own text prompts, we also use ECON's frontal texture to condition the final generated texture.

4.5 Pose Interpolation

While datasets like AMASS provide SMPL poses for performing animation, a more general pipeline must be capable of animating by interpolating between various poses captured from different images. Towards this end, several pose interpolation methods are explored.

Linear Interpolation The simplest pose interpolation method is linear interpolation given as:

$$\theta_t = t \cdot \theta_{end} + (1 - t) \cdot \theta_{start} \qquad t \in [0, 1] \tag{1}$$

For SMPL models, the linear interpolation is done separately for each joint in the quaternion space.

Slerp Interpolation Linear interpolation can fail more sophisticated pose changes, and thus SLERP (Spherical Linear Interpolation) is also experimented with. The formula for SLERP is given as:

$$\text{slerp}(q_0, q_1, t) = (q_1 \cdot q_0^{-1})^t \cdot q_0 \tag{2}$$

PoseNDF based Interpolation In cases where poses are out of distribution, their interpolation can lead to unnatural poses, thus, we also use the PoseNDF model, to project the unnatural poses into a manifold of natural poses, to give smooth interpolation. For PoseNDF, we first obtain the interpolated poses by using Linear interpolation or SLERP, then based on its distance from the pose manifold, we can calculate the gradient to project it onto the pose manifold.

5 Experiments

Different parts of the pipeline are being investigated with various experiments. While our central model, ECON, works very well for many of the cases, we also observe defects in more involved settings.

5.1 ECON vs PIFuHD

ECON and PIFuHD are being compared for in-distribution and out-of-distribution poses, different clothes, and distinct image backgrounds. For in-distribution poses, both ECON and PIFuHD produce pretty good reconstructions (Fig2), while for out-of-distribution poses, ECON still performs quite well but PIFuHD reconstructions lead to artifacts like disembodied limbs and failure of back face reconstruction (Fig3). Both methods tend to work well with different cloth settings, but as noticed in the previous case, while ECON is robust to different image backgrounds, PIFuHD can exhibit artifacts if the background is not plain or if the clothing texture matches the background, causing ambiguity (Fig4).

(a) ECON reconstruction in distribution pose

(b) PiFUHD reconstruction in distribution pose

Fig. 2: Comparison of reconstruction results ECON and PiFUHD

(a) ECON reconstruction out of distribution pose

(b) PIFuHD reconstruction out of distribution pose

Fig. 3: Comparison of reconstruction results ECON and PIFuHD

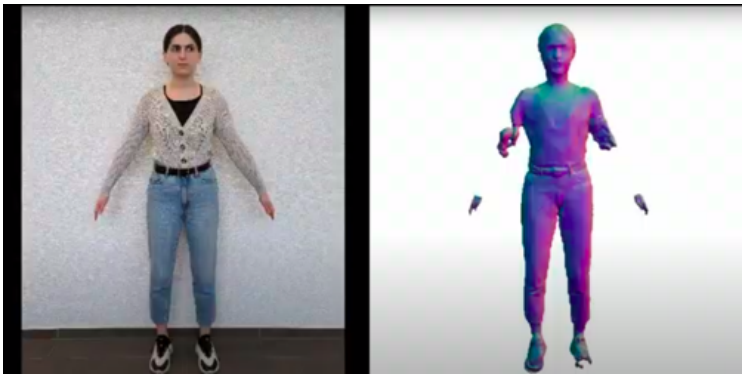

Fig. 4: PIFuHD reconstruction fails because of not enough background foreground(the cardigan) contrast

5.2 ECON for various viewpoints

ECON-based reconstruction is also being utilized for partially visible poses and non-frontal camera angles. In both scenarios, ECON yields high-quality reconstructions. However, one drawback observed with ECON is that it sometimes produces artifacts, such as feet at unequal heights or knee bends, even when the person is standing straight, due to its lack of consideration for environmental information.

5.3 TEXTure

For full-body texture reconstruction, we employ the TEXTure model. While this model provides full-body texture, its output is found to be highly unreliable. In Figure [Cite], the text prompt "High-resolution picture of a girl in a white cardigan and black t-shirt" results in a reconstruction of a white T-shirt with a black cardigan. Similarly, a "High-resolution picture of a girl in an orange shirt, white pants, and green sneakers" is misinterpreted as a girl wearing a "green sweater" (Fig5a).

The model also produces other artifacts, such as generating output with two faces on both sides of the body (Fig5b) and unnatural darkening of the skin. These experiments highlight the need for a better language model and dedicated fine-tuning with human body models. The base TEXTure model was trained on a diverse synthetic dataset featuring everyday objects like statues and artwork, underscoring the necessity for refinement when applied to human subjects.

(a) "White cardigan and black t-shirt" results in white T-shirt, black cardigan.

(b) Generated texture is good, however the virutal human has two faces.

Fig. 5: TEXTure results

5.4 Interpolation

The three methods tested for interpolation, namely Linear Interpolation, SLERP, and PoseNDF based interpolation, all give convincing interpolations. In most cases, linear interpolation and SLERP are enough to produce in-distribution poses which can't be further improved with PoseNDF. In more difficult cases, using PoseNDF leads to a more natural-looking interpolation. A common artifact seen with all three methods is the smoothing of the motion. For instance, while reconstructing jumping jacks, the jump part of the motion is smoothed out, which suggests as a future research direction conditioning the pose manifold on action can lead to interpolations that look like the actions.

6 Conclusion

In this project, we successfully reconstruct a textured, animatable mesh from a single RGB image. Throughout the process, we address the challenges associated with setting up the environment for various state-of-the-art libraries such as ECON, TEXTure, and PoseNDF. Additionally, we gain insights into 3D body model visualization pipelines and the transformation between SMPL and SMPL-X parameters. We also attain an understanding of the capabilities and limitations of different interpolation methods.

In conclusion, the ECON method provides stability in handling pose variations and diverse clothing styles, as well as accurate frontal texture reconstruction. However, it lacks the capability for full-body texture reconstruction and does not consider environmental factors. On the other hand, TEXTure generates full textures based on textual input but may exhibit poor generalization due to not being exclusively trained on human models, and its LLM network may lack robustness.

In terms of interpolation techniques, simple methods such as Linear Interpolation and Spherical Linear Interpolation (SLERP) are often sufficient. However, for interpolation that accurately represents the action, conditioning on the specific action is necessary.

References

1. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
2. Huang, Y., Yi, H., Xiu, Y., Liao, T., Tang, J., Cai, D., Thies, J.: TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In: International Conference on 3D Vision (3DV) (2024)
3. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
4. Mahmood, N., Ghorbani, N., F. Troje, N., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2019), `https://amass.is.tue.mpg.de`
5. 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 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019)
6. Richardson, E., Metzer, G., Alaluf, Y., Giryes, R., Cohen-Or, D.: Texture: Text-guided texturing of 3d shapes. arXiv preprint arXiv:2302.01721 (2023)
7. Saito, S., Simon, T., Saragih, J., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: CVPR (2020)
8. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: ICON: Implicit Clothed humans Obtained from Normals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13296–13306 (June 2022)