Supplementary Material for Robust 3D Garment Digitization from Monocular 2D Images for 3D Virtual Try-On Systems

Sahib Majithia
Myntra Designs Pvt. Ltd.
Bangalore, India
sahib.majithia@myntra.com

Sandeep N. Parameswaran
Myntra Designs Pvt. Ltd.
Bangalore, India
sandeep.narayan@myntra.com

Sadbhavana Babar
Myntra Designs Pvt. Ltd.
Bangalore, India
sadbhavana.babar@myntra.com

Vikram Garg
Myntra Designs Pvt. Ltd.
Bangalore, India
vikram.garg@myntra.com

Astitva Srivastava
IIIT-Hyderabad,
India
astitva.srivastava@research.iiit.ac.in

Avinash Sharma
IIIT-Hyderabad,
India
asharma@iiit.ac.in

1. Pose transfer for synthetic data generation

In order to replicate the pose variations present in real-world catalog images into our synthetic data, we first estimated the 3D human pose on the real catalog images using [1]. The skeleton information is retrieved from the estimated 3D pose and is converted to joint angles in MoCap format [2]. The human avatar is further re-target to the estimated 3D pose by generating the rotation and translation parameters for each bone present in the armature of the rigged skeleton with respect to the saved skeleton present in the form of motion data. The Figure 1 shows the results of pose transfer from real images to a 3D human avatar. Figure 2 shows the visual similarity in the pose variations of the real and synthetic datasets.

2. Shortcomings of DeepFashion2 Dataset

In this section, we explain the reason for curating our own real-world catalog dataset for finetuning the landmark detection module, instead of finetuning with the publicly available Deepfashion2 dataset [3] which has ground truth annotations for in-the-wild images. Our initial findings suggested that the networks trained on Deepfashion2 dataset don't perform well on standard catalog images. We tested the landmark estimation capability of HRNet [4] trained on Deepfashion2 for "DeepFashion2 Challenge 2020 : Track-1 (Clothes Landmark Estimation)." The landmarks predicted by HRNet are not reliable and sometimes missing as shown in Figure 3.

3. Additional Results on Inpainting

In this section, we present additional results for the texture inpainting module. Figures 4 and 6 show the inpainting results on our synthetic data for t-shirts and trousers respectively. In these figures, column (a) denotes the masked input image, column (b) denotes the occluded mask (area to be inpainted is in white), column (c) denotes the output of the texture inpainting network [5] and column (d) denotes the ground truth. The inpainting network which is trained only on synthetic images performs well on unseen real-world images as well. This generalizability is depicted by the qualitative results in Figures 5 and 7 for t-shirts and jeans respectively. In these figures, column (a) denotes the masked input image, column (b) denotes the occluded mask, column (c) denotes the output of the TPS [6] transform on clothing panels and column (d) denotes the output of texture inpainting network [5]. As shown in columns (d) of figures 5(i), 5(ii) and 7, the network is able to inpaint the information missing due to occlusion and around the borders of the panels.

If the backside image of the garment is available in the catalog, then our approach for generating texture for the front panel can be borrowed for the back panel of the template also, starting by detecting landmarks on the backside image and followed by texture transfer by TPS and inpainting. In the absence of the backside image, we identify a patch having a uniform gradient in the front panel and then extend this patch to the entire back panel. The patch-based approach is preferred for T-shirts only as for trousers the result doesn't look natural. For trousers, we use the original back view present in the catalog and process it similarly to the front view panels.

4. Qualitative Results of our Pipeline

Figure 8 shows the end-to-end results of our pipeline. Given the input image we detect the landmarks which are used to designate the panels on the input image. Then, TPS

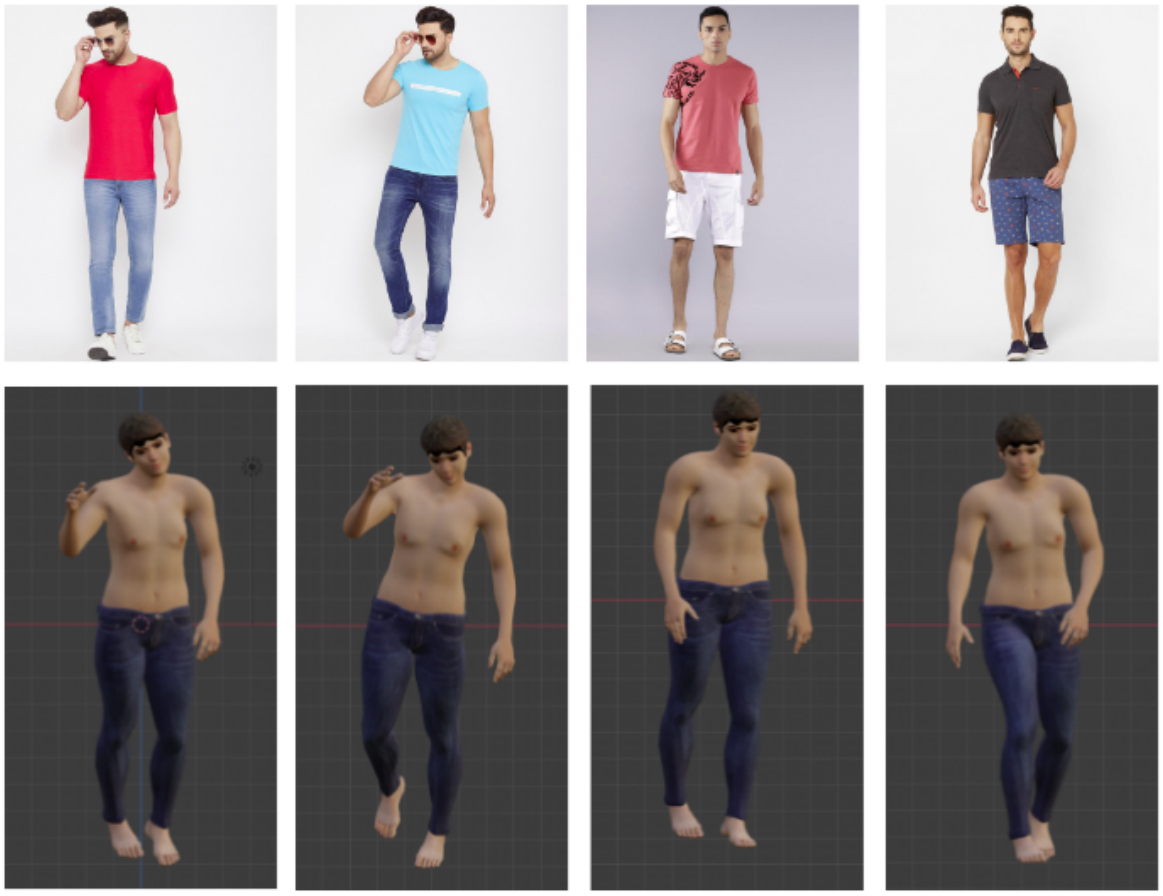

Figure 1: Results of the pose transfer from real catalog images to 3D human avatars.

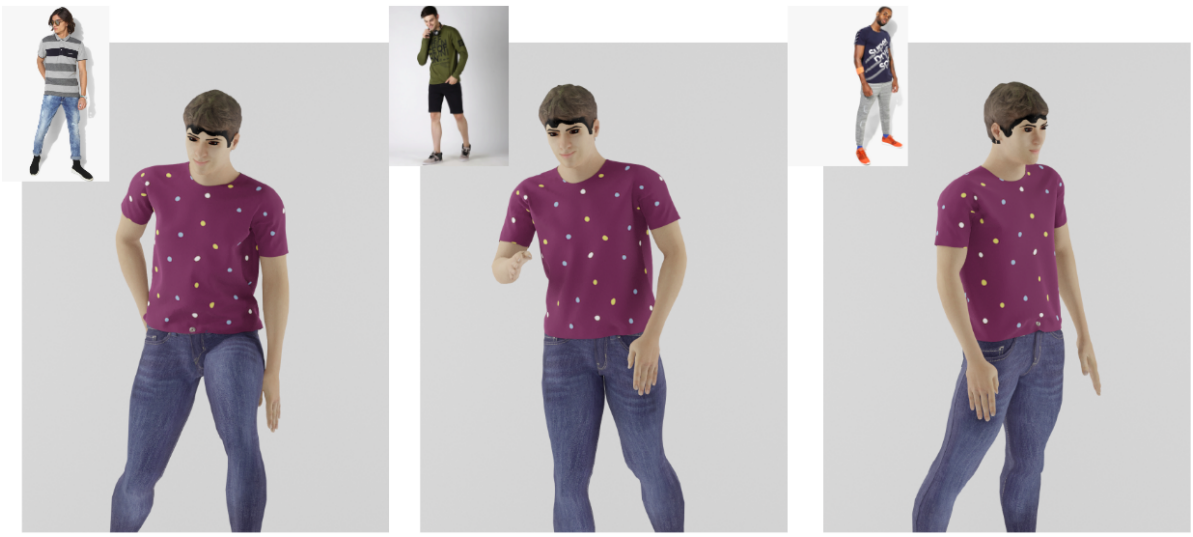

Figure 2: The 3D human avatar is draped with synthetic garment templates. The distribution of the poses is similar to that of an e-commerce catalog.

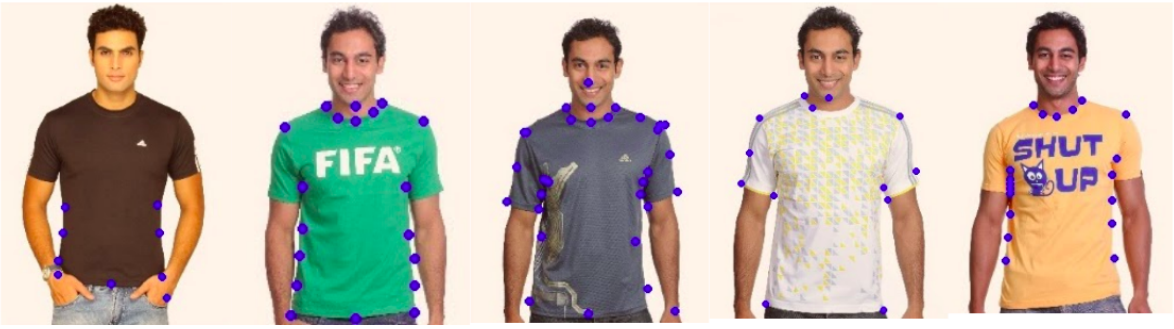

Figure 3: Unreliable landmarks predicted on real catalog images by HRNet (trained on Deepfashion2 dataset).

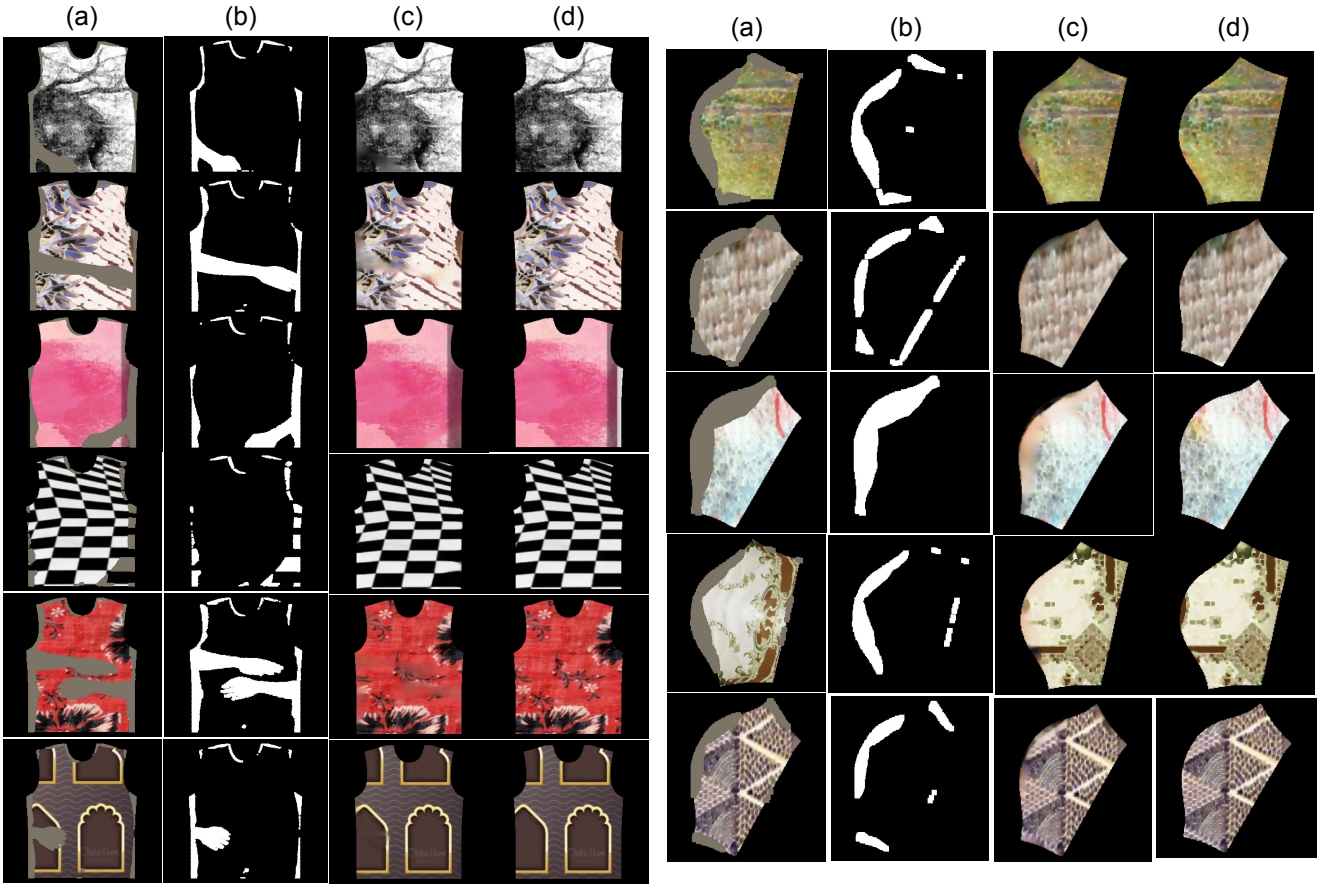

Figure 4: **Qualitative results of inpainting network on synthetic data for t-shirts.**

based texture transfer is done to transfer the texture information from the input image to the UV map of the template mesh. Missing texture information due to occlusions and on the border of the panels due to missing side view is filled using inpainting network. Since the back view of T-shirts was not available, we used uniform gradient patch based approach to transfer texture to the back panel in the UV map.

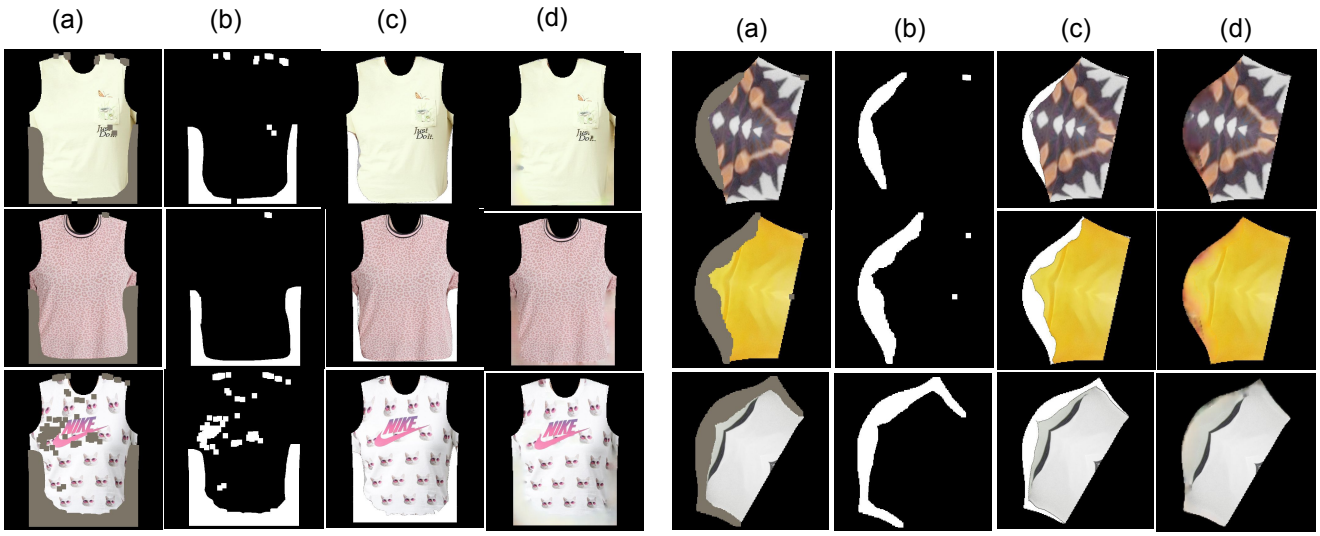

Figure 5: **Qualitative results of inpainting network on real-world data for t-shirts.**

Finally, the textured 3D template mesh is shown.

5. Discussion regarding the video

In the video, first we have shown the comparison of our method with PIFu [7] and Pix2Surf [8] in 3D. Our method handles occlusions well and preserves fine texture details, but both Pifu and Pix2Surf can not deal with occlusions. Pix2Surf retains a fixed geometry but it does not retain texture associated with the occluded region. PIFu is noisy and does not retain even the geometry associated with the occluded region (refer to the video for a better understanding). In regard to preserving fine texture details, the sharpness of the text "WRANGLER" and "VACCINATED" in the input images is clearly preserved in the textured meshes obtained by our method. On the other hand, blurriness can be observed in the output of PIFu and Pix2Surf. Finally, we have shown that the textured template meshes can draped on 3D human avatars with varying body shapes without any unnatural deformation in the texture.

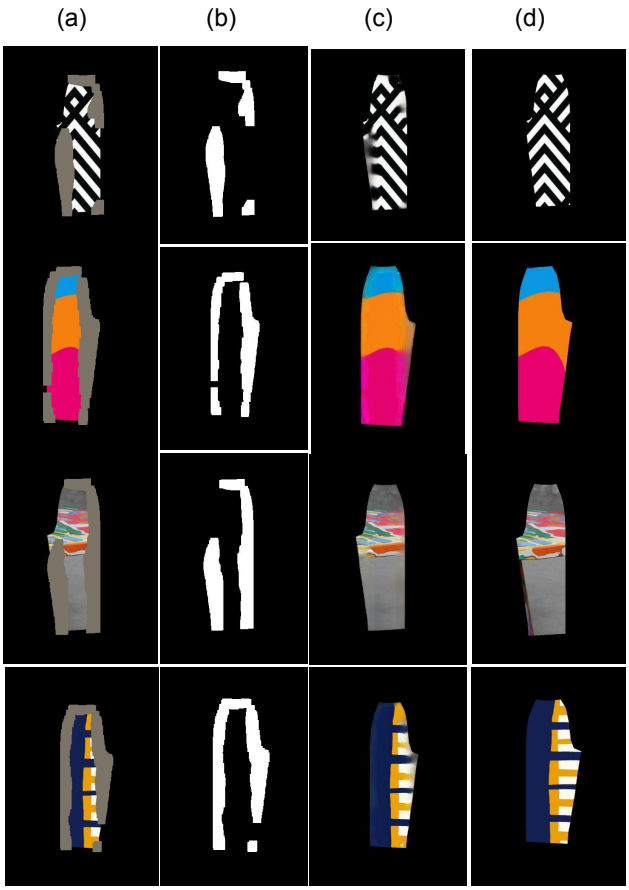

Figure 6: **Qualitative results of inpainting network on synthetic data for trousers.**

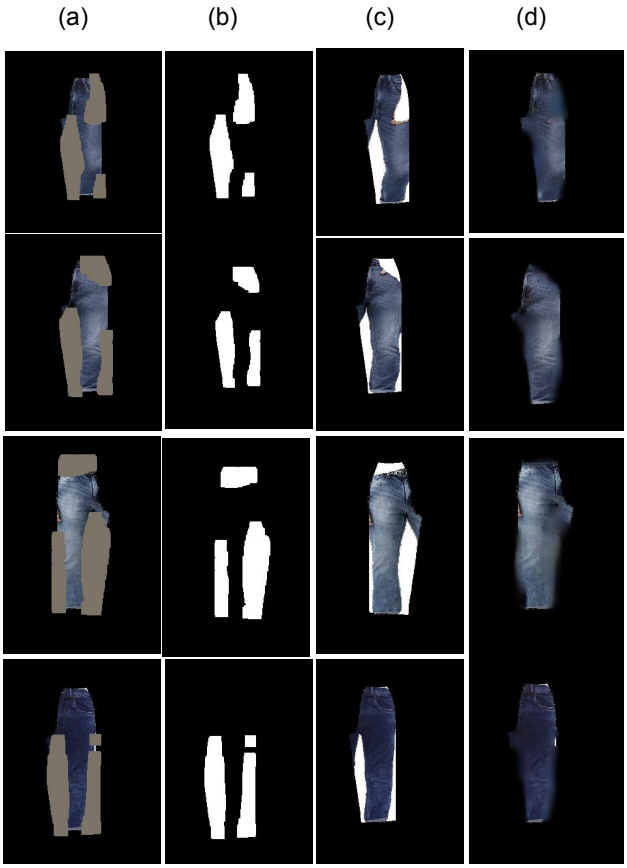

Figure 7: **Qualitative results of inpainting network on real-world data for trousers.**

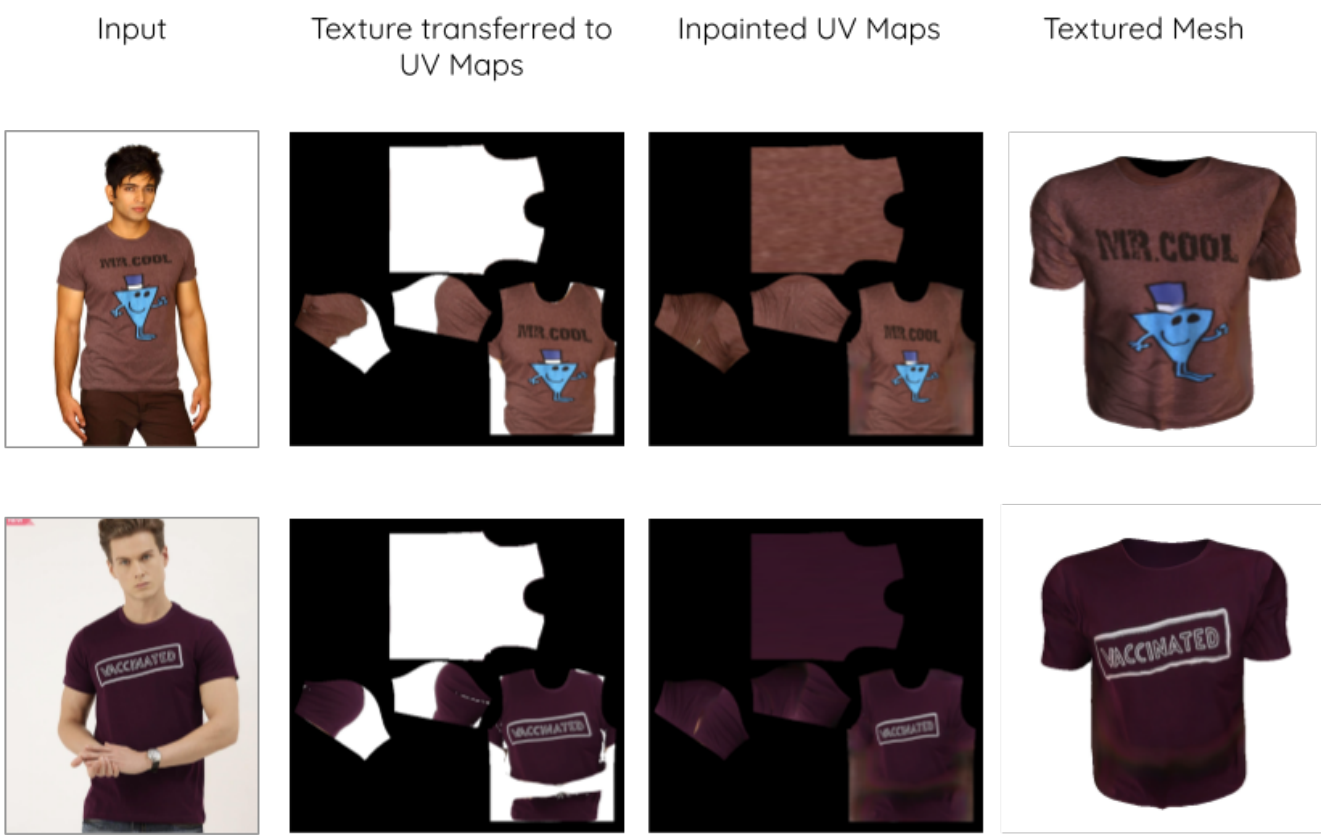

Figure 8: **Results of our pipeline.**

References

[1] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6173–6183, 2020.

[2] João Cartucho, Samyakh Tukra, Yunpeng Li, Daniel S. Elson, and Stamatia Giannarou. Visionblender: a tool to efficiently generate computer vision datasets for robotic surgery. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–8, 2020.

[3] Yuying Ge, Ruimao Zhang, Lingyun Wu, Xiaogang Wang, Xiaoou Tang, and Ping Luo. A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. *CVPR*, 2019.

[4] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition, 2020.

[5] Manyu Zhu, Dongliang He, Xin Li, Chao Li, Fu Li, Xiao Liu, Errui Ding, and Zhaoxiang Zhang. Image inpainting by end-to-end cascaded refinement with mask awareness. *IEEE Transactions on Image Processing*, 30:4855–4866, 2021.

[6] Alexander C. Berg and Jitendra Malik. *Shape Matching and Object Recognition*, pages 483–507. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[7] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019.

[8] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. Learning to transfer texture from clothing images to 3d humans. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 7021–7032. IEEE, 2020.