3D Garments Crawler from large-scale fashion images

Authors
Daniel Gudmundsson (dgudmundsson)
Marion Barrau-Joyeaux (mbarrau)
Arthur Collette (acollette)
Amalie Kjaer (akjaer)
ETH Zürich

Supervisors
Yuliang Xiu
Songyou Peng

Abstract

The widespread adoption of deep learning techniques and implicit shape learning has led to remarkable progress in single-image human digitalization. In particular, it is now possible to recover subtle details such as clothing folds and wrinkles. Although the implicit-based methods show promising results, it remains challenging to get a separated and topology-consistent mesh for each garment piece. Yet, those digital human assets are crucial for the current 3D content creation pipelines. To address this issue we implemented an extension to the current ICON [1] model so that 3D clothes mesh can be extracted using single-view images and their corresponding clothes segmentation. With this extension we then create a large and diverse 3D fashion clothes dataset.

1. Introduction

There is an increasing demand for high-quality human-related 3D contents in various fields including, but not limited to, virtual live-streaming, gaming, filming or digital communication. A strong limitation of this is the tediousness of producing realistic 3D digital human assets like clothes, as it may take hours even for an expert modeller. By comparison, capturing in-the-wild images is now easily done and large databases are already available. This led researchers to develop new methods to generate visually plausible 3D virtual humans and its assets from single in-the-wild images.

Yet, although major breakthroughs have been achieved in the past years in single image body and clothed human reconstruction, research focused on garment reconstruction remains limited. The two key aspects when reconstructing garments with high-fidelity are the generation of the garment styles and the reproduction of the surface details. Im-

plicit methods show good reconstruction quality, but the diversity of the real-world clothes styles and geometry makes it difficult to delineate the garment on the implicit clothed human. Due to these limitations, there are therefore few large-scale and good quality 3D mesh clothes dataset that can be found without involving designers or extensive prior inputs from 3D scans.

In our approach, based on ICON [1] model, we first explore various methods to extract the garments meshes from the clothed human body. Our main contribution is then the design and implementation of a non-learning based pipeline to extract the garments meshes from the clothed human body of ICON model. We then tested our approach on the DeepFashion2 dataset [2] and create a large and diverse dataset of 3D garments.

2. Related Work

With the recent improvements in 3D deep learning, there have been new breakthroughs in creating 3D clothes and human bodies from 2D pictures. Having the body and clothes reconstructed on different layers, and not only as one 3D clothed body, provides easy and ready-to-use objects for diverse tasks in animation and content creation.

Closely related to our work, Zhu et. al. [3] developed a novel learning-based method that predicts the implicit garment boundary fields based on pixel-aligned features, curve-aligned features and manually created collar warehouse.

Cloth3D by Bertiche et. al. [4], which also creates a large garment dataset, uses a segmentation mask predicted by the network itself to extract the garment by removing body vertices. They use a Conditional Variational Auto-Encoder (CVAE) based on graph convolutions (GCVAE) to learn garment latent spaces, using template garments that have been manually created by designers from real patterns.

Finally DeepFashion3D by Zhu et. al. [5], that also provides a large 3D clothes dataset, based their work on 3D reconstruction of existing clothes by first building a dataset of 3D cloud points of clothes. They then use the latter as ground truths to train their model creating 3D clothes meshes. DeepFashion3D [5] consists of 2078 3D models reconstructed from real garments. It is built from 563 diverse garment instances, covering 10 different clothing categories. They also added detailed annotations tailored for 3D garment – 3D feature lines. The MGN model uses a similar initial approach, but in comparison has a smaller initial ground truth dataset composed of 712 digital garments.

Compared to these related works, our approach is non-learning based, and doesn't require any 3D cloth points clouds, which are usually complex to acquire. Instead, it combines ICON performance and its robustness to out-of-distribution poses, the well-defined SMPL [6] labels and the segmentation of a big and diverse dataset - DeepFashion2 [2]. Therefore, our approach adds the functionality of creating clothing meshes separated from the human body to the pre-trained ICON, either from an existing dataset including clothes segmentations, or by combining them with an already existing segmentation algorithm. Additionally, our work results in a large and diverse 3D cloth dataset.

3. Methods

3.1. Initial Research

This section outlines various methods explored to produce 3D garment models. This research phase was a big part of our project and led to the final pipeline presented in the next section.

Boolean Difference. To extract the clothes, we first used the boolean difference from Trimesh library. The idea was to remove the SMPL skinned body mesh from its ICON clothed body mesh output to keep only the 3D meshes of the clothes. However, although boolean operations are very fast, it is a known issue that they are often not robust and not deterministic on the type of 3D meshes we have. We therefore had very disparate results ranging from almost empty mesh to identical to initial ICON's output, with no strong difference of treatment between the clothes and the body.

Erosion. In order to improve the boolean difference performance, we tried to use the mesh erosion in order to minimise the impact of the different topology between the SMPL body mesh and ICON mesh. The idea was to erode the SMPL mesh by a small parameter and then apply the difference with the clothed mesh. The main issues with this method were that it was still not deterministic and the parameter had to be adjusted for every image. Therefore, it was in most cases still not outputting the right mesh.

Mesh registration and SMPL labels. We then want to have the possibility to use the SMPL labels in our pipeline, as some body parts aren't relevant for a certain type of clothes. So we looked into doing mesh registration of the ICON's [1] clothed mesh to its related SMPL body mesh. In order to do this we implemented a 1-NN algorithm. We use using sklearn library as it selects the best algorithm fit between BallTree, KDTree and brute-force search algorithm. It was easy to add to our pipeline and met our needs. We also explored the registration function from Trimesh library which is using the iterative closest point technique, which is another nearest-neighbour-search strategy.

Colour matching segmentation. As the SMPL labels were too coarse, we tried to use colour matching segmentation to separate the body from the clothes. We projected the

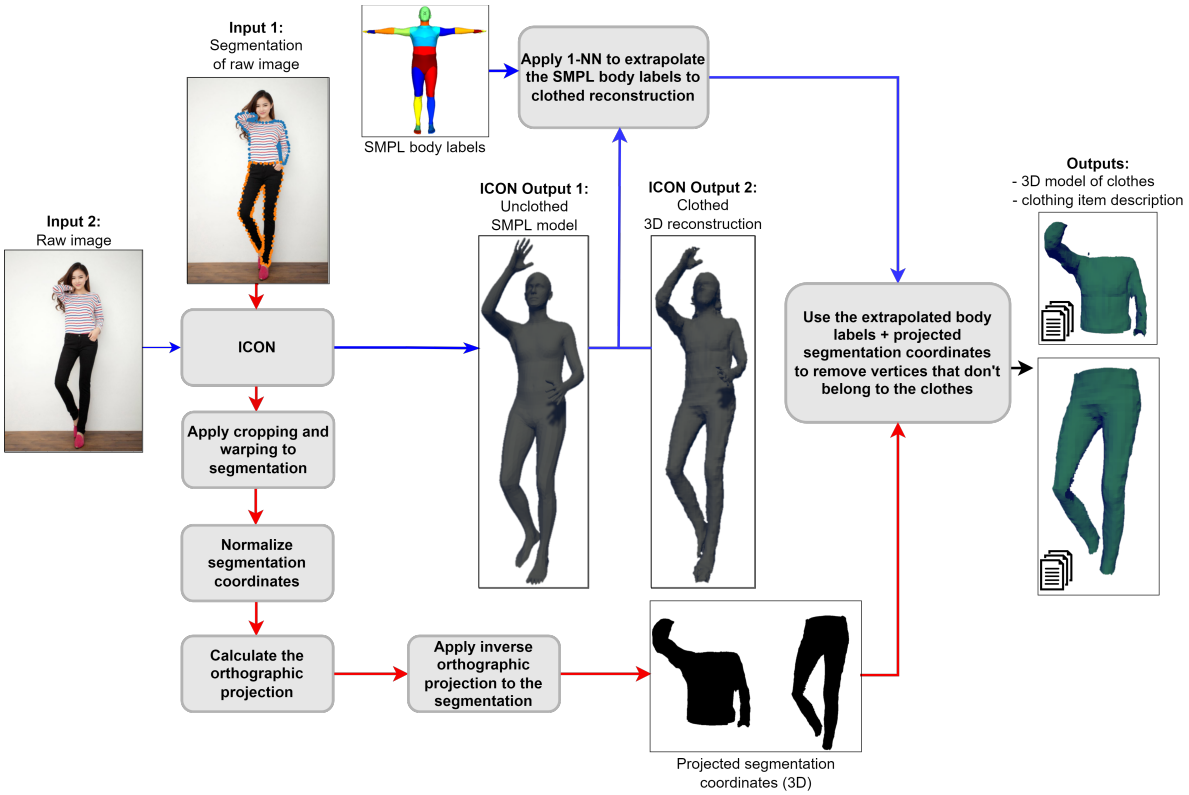

Figure 1: **Pipeline for generation of 3D cloth models.** We combine the projected garment segmentations (obtained via red arrows) with the body-labelled 3D reconstruction (obtained via blue arrows) to generate 3D garment models with garment description.

3D clothed mesh with the colours onto a 2D plane. Then, we used the pre-trained model from Levin Dabhi using the U2-Net [4] to apply colour matching segmentation. Then we used the result as a stencil to remove the body parts from the clothed mesh. A limitation of this technique is that it was very sensitive to any kind of occlusion and body pose, whether it is from the initial image or from ICON body pose prediction. Moreover, this technique relies entirely on the quality of the colours reconstruction.

3.2. Proposed Pipeline

This section presents the final pipeline implemented to produce labelled 3D garment models. The implemented pipeline is illustrated in Figure (1).

In essence, we utilise the orthographic projection generated in ICON (relating an input image to its 3D reconstruction) to project garment segmentation labels (given in the image dataset) onto the 3D reconstruction. This enables us to identify garment delimitations on the 3D reconstruction, and hence to create 3D garment models. We combine this method with the SMPL body labels to remove any unwanted vertices and improve the quality of 3D garment models. Finally, 3D garment models are labelled using the clothing item description provided in the image dataset.

To obtain the clothed 3D reconstruction with body labels, we first input a raw image to ICON and obtain an unclothed SMPL model along with a clothed 3D reconstruction. During this process, we establish a 3D to 2D orthographic projection, relating the input image to the 3D reconstruction. Then, SMPL body labels (eg. left hand, head, right leg) are extrapolated to the 3D reconstruction using 1-Nearest-Neighbour (1NN). This results in a clothed 3D reconstruction with body labels.

To obtain the projected garment segmentations, we first crop, warp and normalise the segmented image to match the image processing performed in ICON. Next we transform the coordinates of the segmentation to Normalized device coordinates (NDC) to match the coordinate system used by ICON. The previously obtained orthographic projection is applied to the segmentation to obtain 3D projected segmentation coordinates. We treat these coordinates like vertices in a polygon and use it to keep only vertices in the reconstruction mesh from ICON that lie inside this polygon by ignoring the z-axis. The resulting mesh is only the vertices

and faces within the segmentation.

We combine these results to improve the quality of 3D garment models. The quality of the 3D garment models obtained using the projected garment segmentation is improved by using the extrapolated body labels. For example, if a hand occludes a leg in an image, it is likely to be included in the trousers segmentation given in the image dataset. As a result, the hand will appear in the 3D cloth model, but is removed using the extrapolated body labels. For each type of clothing, we specify which body parts should systematically be removed. For example, no feet, legs or heads should appear in the 3D model of a t-shirt. However, torso and upper arms are maintained.

Note that we only use the SMPL labels to remove vertices that we can almost guarantee are not supposed to be there (e.g. hand is definitely not part of trousers) but we don't use the labels to add vertices to the mesh. This is because clothes differ so much in general and the labels only give us a rough segmentation of the body parts. For example, we might want to try to improve the quality of a sleeve of a T-shirt by using the SMPL labels, but we then immediately face the problem of how long should the sleeve actually be? It might be very short or rather long, but the SMPL labels only give us a single cutoff for the sleeve which is the same for all SMPL body meshes. Therefore, using the SMPL labels in this scenario is not useful.

4. Experiments

4.1. Dataset

The DeepFashion2 [2] dataset is a rich fashion dataset created by Yuying Ge et al. from the Chinese University of Hong Kong. It contains 491K images of 13 popular clothing categories (short sleeve top, long sleeve top, short sleeve outwear, long sleeve outwear, vest, sling, shorts, trousers, skirt, short sleeve dress, long sleeve dress, vest dress and sling dress). Compared to previous fashion datasets, such as DeepFashion[7], it proposes a large-scale benchmark with comprehensive tasks and annotations of fashion image understanding. Each image is paired with a JSON file that contains information such as:

- The category of the item(s) presents in the image
- The bounding box of the clothes
- The landmarks, 2D cloth segmentation labels
- The scale level
- The occlusion level
- The zoom level

We use these features to preprocess the dataset and only keep useful images. Indeed, a lot of them are too occluded or zoomed-in to be utilised for our problem. In order to have a robust way of selecting the images, we use landmarks, a list of 2D coordinates along with a visibility parameter. Each point can be either visible, occluded, or not present in the image. We put a threshold of 0.6 on the ratio of visible points compare to the total number of points. As such, we reduced the dataset to around 100 000 images that we used to create our 3D clothes dataset. The distribution of images of each garment type after preprocessing is shown in Figure (2).

4.2. Results

The resulting dataset contains around 160K samples. Figure (3) shows examples of outputs for four distinct categories (trousers, shorts, short sleeve tops and dresses). In most cases, the outputs were of high quality (no holes, high resemblance to input image). More examples are available in the provided sample dataset.

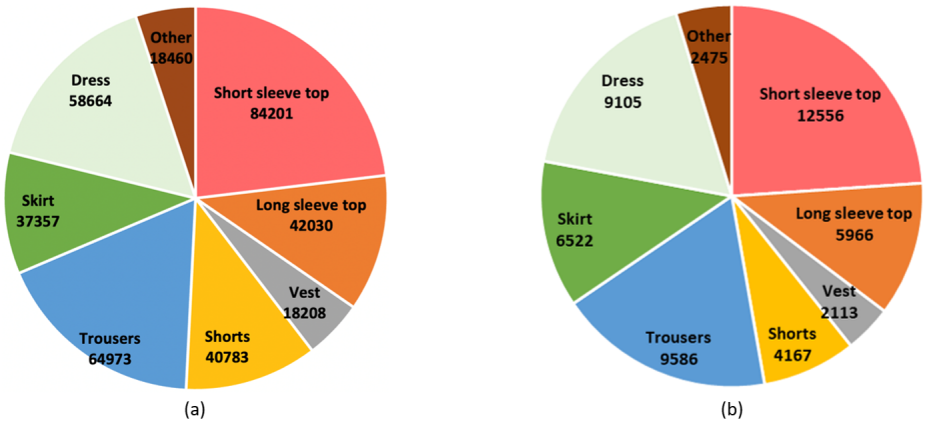

Figure 2: Distribution of images of each garment type before (a) and after (b) preprocessing.

Figure 3: Example outputs for different types of garments.

4.3. Discussion

The results shown in Figure (3) showcase the high quality of obtained outputs on trousers, shorts, short sleeve tops and dresses. However, quality output is highly dependent on the quality of ICON reconstructions. In particular, due to the strong body prior used by ICON model, loose clothing were difficult to reconstruct as we can already observe on some dresses. Other garment types, that were even looser fitting or with poor initial picture quality, lead to reconstruction failures. ICON also often fails to create high quality meshes for images with strong perspective effects (as shown in Figure 4). Better results could be achieved by a using a more strict selection of the input pictures.

The output quality also relies on the quality of the segmentation labels: poor cloth segmentation will likely lead to garment models with holes, or covering wrong regions of the body. High quality segmentation does not however guarantee high quality garment models, as strong perspective effects in the image most often leads to garment models with holes in its sides, as shown in Figure (5). This is due to the fact that the cloth segmentation on a rotated body will be smaller than the actual garment size, leading to truncated garments when projected onto the front of the ICON reconstruction.

Furthermore, this approach does not work well on e.g. outwear clothes like jackets. Since the segmentation of a jacket is usually split in (at least) two parts this will result in the back of the jacket to be missing since technically it's not part of the segmentation (6). A possible solution could be to segment jackets differently than other clothes to differ between sections of the mesh where the front should be removed while the back is kept. Another solution could be to use the current approach but modify in the following way: Create a large segmentation out of the two segmentation parts, apply the same procedure to extract the mesh and then use already implemented functions in ICON to get the visibility of each vertices in the additional segmentation created and remove vertices that are only in the front but keep the vertices in the back.

The currently proposed pipeline can be directly integrated with ICON and is scalable. It can also be combined with segmentation algorithms to create 3D garment meshes from any (non-labelled) input image.

Future work should include developing evaluation methods to assess the quality of 3D garment meshes, such as training a classification neural network, in order to identify flawed meshes and improve the overall quality of the dataset. An even better solution would be to figure out a way to refine the results e.g. by using the SMPL shape and pose parameters along with the fact that clothes are generally symmetrical to fill in gaps in the mesh. A next step would then be to use the cleaned dataset to train a Generative Adversarial Network.

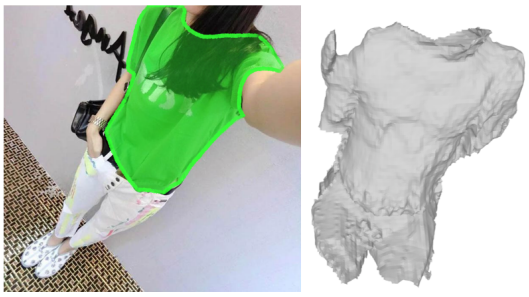

Figure 4: Example of poor reconstruction due to strong perspective effects (the trousers appear in the 3D cloth model of the top due to an overhead perspective).

Figure 5: Example of failed side reconstruction due to poor segmentation.

Figure 6: Example of poor reconstruction due to symmetry effects (back is assumed to be of same geometry as front).

5. Conclusion

We have presented a non-learning based pipeline to generate separable, high quality 3D garment meshes from ICON reconstructions. The main idea is to project garment segmentation labels onto the ICON reconstruction, combined with SMPL body labels. We have implemented this to obtain a large and diverse dataset of 3D garment meshes. Limitations are mainly tied to the segmentation quality, ICON reconstruction quality and input image quality (occlusions, cropping, strong perspective effects). Future work should explore evaluation methods to assess the quality of outputs and potentially methods to improve imperfect meshes. A key application of this dataset could be to train a Generative Adversarial Network to generate new garment meshes. This could advance research in garment mesh generation.

Code The code behind this project was integrated into the ICON Github repository which can be found at https://github.com/YuliangXiu/ICON.

Contribution All authors contributed equally to this project.

Acknowledgements. We would like to thank Yuliang Xiu and Songyou Peng for the research topic, and for their valuable inputs and discussions. We also thank Yao Feng for the presentation feedback and Denys Rozumnyi for technical help with Euler clusters.

References

[1] Y. Xiu, J. Yang, D. Tzionas, and M. J. Black, "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.

[2] Y. Ge, R. Zhang, L. Wu, X. Wang, X. Tang, and P. Luo, "A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," *CVPR*, 2019.

[3] H. Zhu, L. Qiu, Y. Qiu, and X. Han, "Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images," 2022.

[4] H. Bertiche, M. Madadi, and S. Escalera, "Cloth3d: Clothed 3d humans," 2019.

[5] H. Zhu, Y. Cao, H. Jin, W. Chen, D. Du, Z. Wang, S. Cui, and X. Han, "Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images," 2020.

[6] 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)*, vol. 34, pp. 248:1–248:16, Oct. 2015.

[7] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.