LYCB: Leave Your Clothes Behind

Keifer Lee,[1] **Shubham Gupta,**[1] **Karan Sharma**[1]

[1] New York University
kl3866@nyu.edu, sg7761@nyu.edu, ks6421@nyu.edu

Abstract

The demand for assets in the virtual world as recently gained a lot of attention. We present a novel framework dubbed LYCB: Leave Your Clothes Behind that allows users to directly generate a 3D mesh object of garments from a monocular video using Neural Radiance fields. The proposed method fills the gap in literature by addressing issues such as accurately fitting on to foreign bodies for virtual try-ons and model complex cloth properties using Neural Radiance fields and physics-based simulations. This lets us transfer clothing type and bodies easily. Our code is made available at `https://github.com/IamShubhamGupto/LYCB`

Introduction

The virtual world like Metaverse is a rapidly growing space, with more and more people spending time in it for work, play, and education. As the virtual world becomes more immersive, there is a growing demand for real world objects to be brought into it. This is especially true for clothes, as people want to be able to express themselves through their fashion choices, even in the virtual world.

Virtual try-on is a growing market in the fashion industry and users being able to satisfactorily wearing their garment virtually is directly responsible with the number of returns made to the vendors. We aim to provide a framework that accurately fits captured garments on to foreign bodies using physics based cloth simulations. This could successfully reduce losses made by vendors due to returns.

Users of the virtual world could generate virtual assets from scratch however, it may not be accurate and is very time consuming. This is necessarily true clothes due to their intrinsic properties which makes generating and simulating them virtually very difficult. In this work, we attempt to address this issue by providing a framework to directly generate cloth assets from a monocular RGB video.

Neural radiance fields have recently been adopted to generate 3D representations by combining multiple views. It is done by simulating light interacting with the object and hence creating a 3D depth map of the object from each view. In this work, we will be utilizing NeRF2Mesh (Tang et al. 2022) to create the meshes with texture. Using Blender, we apply physics based cloth simulations to fit on mannequins.

Related Work

Neural Radiance Fields (NeRF)

Neural Radiance Fields (Mildenhall et al. 2020) introduced the method of generating novel 3D views from sparse inputs. The input for NeRF can be defined as a single continuous 5D coordinate (location and viewing angle) and its output is the volume density and view dependent emitted radiance at that spatial location. However, while the original technique can be used to export to mesh, it fails to capture the texture, an important property of clothes.

NeRF2Mesh (Tang et al. 2022) address' the above issue by first initializing the geometry and appearance of the mesh using a Neural Radiance Field (NeRF). Next, it performs an iterative surface refinement process that adaptively adjusts the vertex positions and face density based on re-projected rendering errors. Finally, the appearance is jointly refined with the geometry and baked into texture images.

While NeRF produces great results, its performance is directly related to the amount of the training data and compute resources allocated. The authors of Instant NGP (Müller et al. 2022) propose a new method to train Neural Graphics Primitives using an encoding technique that applies a multiresolution hash table to store the input data, which allows the network to disambiguate hash collisions, making for a simple and efficient training algorithm. It provides a new way to train NGPs that is significantly faster and more efficient than existing methods (Mildenhall et al. 2020).

Radiance Fields Based Reconstruction

The authors of SCARF (Feng et al. 2022) address the problem of extracting a 3D clothed avatar using a neural implicit reconstruction process from monocular videos. The demonstrated method allows users to transfer clothing between avatars and also animate their movements. This makes it ideal for virtual applications such as try-ons.

Another method, PERGAMO (Casado-Elvira, Comino Trinidad, and Casas 2022) can also extract and fit 3D garments from monocular videos. The method demonstrated generating 3D representations from a single view. However, it suffers from drawbacks like un-modelled self-collisions and being only able to reconstruct close to the underlying deformable mesh. This creates a dependency and works only when we have a matching deformable mesh.

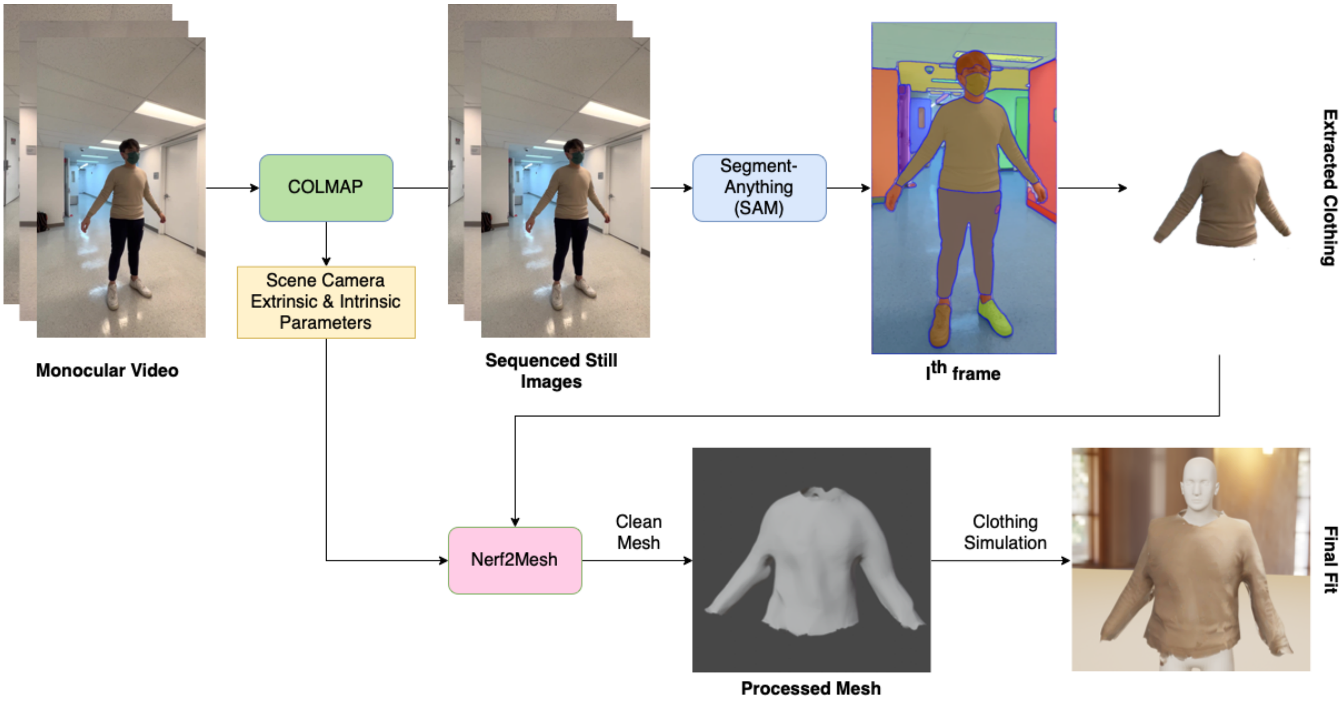

Figure 1: Overview of proposed framework, LYCB.

Non-Radiance Field Based Reconstruction

Besides Radiance Fields based methods, researchers have proposed alternatives such as JFNet (Xu et al. 2019) that utilizes multi task learning to produce 3D garments using just the front and back view. However, it suffers from the same issue as PERGAMO (Casado-Elvira, Comino Trinidad, and Casas 2022) where the template limits the range of garments that can be modelled.

Another approach is to use multi-layer perceptrons which can model pose, shape and style of the clothes avatar (Patel, Liao, and Pons-Moll 2020). The proposed method is computationally efficient but fails to generalize over different body types.

Methodology

The proposed framework of Leave of Cloths Behind (LYCB) is a multi-stage process that leverages on techniques of various domains, from Structure-from-Motion (SfM) to clothing simulation; the following section details the individual underlying components sequentially as outlined in Figure 1.

On a high-level, the main distinguishing factor of LYCB when compared to other existing methods (e.g. SCARF (Feng et al. 2022) and PERGAMO (Casado-Elvira, Comino Trinidad, and Casas 2022)) is its flexibility and the degree of control offered to the end-user. The aforementioned alternative have varying trade-offs such as, neural-implicit radiance field methods that has great flexibility in capturing complex garments (e.g. dress with frills) at the cost of sub-optimal physical response when fitting said garment on another body (Feng et al. 2022), or template-displacement with vertex regression methods that are able to accurately model said response (e.g. fine wrinkle from soft-body deformation) but is incapable of capturing and reconstructing a complex garment mesh (Casado-Elvira, Comino Trinidad, and Casas 2022). LYCB aims to fit the gap in between by leveraging the flexibility of geometry-agnostic neural implicit radiance fields for modelling and well-conditioned physics-based clothing simulation algorithms for accurate any-body garment fitting.

Camera Parameter Estimation with COLMAP

To begin, LYCB takes as input a 360° monocular sequence of the target garment for reconstruction via NeRF2Mesh (Tang et al. 2022), a neural implicit radiance field method with simultaneous texture recovery. However, in order to perform said reconstruction, the camera parameters (e.g. intrinsic and extrinsic) needs to be first determined since it will need to know how each individual pixel across frames corresponds spatially to the actual 3D coordinates of the target.

One method of such camera parameter estimation is COLMAP (Schönberger et al. 2016; Schonberger and Frahm 2016), a SfM algorithm designed to recover 3D structure from a collection of 2D views. On a high-level, COLMAP achieves this by first extracting a set of SIFT features (Lowe 1999) for each individual frame, and then trying to triangulate the global and relative position of each view relative to a common coordinate system via iterative feature matching and bundle adjustments. Therefore, for the input sequence, LYCB expects and inherits the assumptions from both COLMAP and its underlying SIFT feature extractor, such as scene intensity consistency, having high degree of iter-frame overlaps, and each frame having determinable ge-

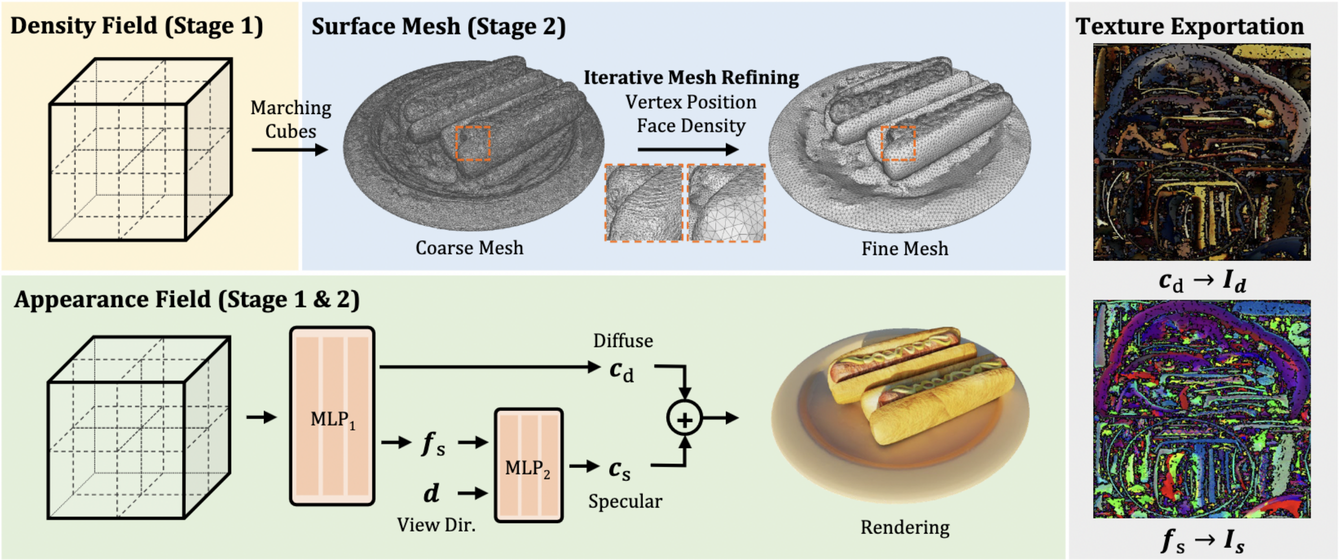

Figure 2: Overview of NeRF2Mesh's pipeline; adapted from (Tang et al. 2022).

ometric description for meaningful SIFT feature extraction (e.g. corners not edges).

Specifically, LYCB uses COLMAP to determine camera intrinsic parameters (e.g. K matrix) such as the horizontal f_x and vertical f_y focal lengths, and camera extrinsic parameters via the 4x4 transformation matrix which characterizes the rotation, translation and scaling of the camera relative to global reference coordinate; we shall denote said combined parameters for the i-th frame as P_i.

Garment Selection via Segmentation Masking

Next, in order to isolate the target garment from other surrounding scene objects and characters, LYCB utilizes a pre-trained Segment-Anything (SAM) segmenter (Kirillov et al. 2023) to perform zero-shot semantic segmentation.

SAM is state-of-the-art segmentation model that could be positively or negatively prompted with keypoints, bounding boxes or masks to generate a multi-level semantic mask of a specified object; as seen in Figure 1 SAM could segment individual apparel with great accuracy. Underlying SAM is an encoder-decoder architecture driven by Vision Transformers (ViT) (Dosovitskiy et al. 2020) with an additional CLIP prompt-encoder (Radford et al. 2021); the rest of the architectural details of SAM shall be omitted here for brevity.

Since SAM could generate multi-level output per instance, one would need to determine which output corresponds the the target garment for proper masking and extraction. By default, LYCB picks the final mask for the i-th frame M_i by selecting for maximum the objectness score, $S_{i,obj}$. It needs to be noted that this could vary for different input sequence and therefore may require fine-tuning to ensure a consistent and accurate extraction across all frames. The entire process can be described by the following equation where x_i is the i-th RGB frame, Z_i is the set of prompts for the i-th frame and N is the total number of frames in the sequence.

$$M_i = \underset{S_{i,obj}}{\operatorname{argmax}} \; \mathit{SAM}(x_i, Z_i) \quad , \mathit{for}\ i \in [1, N] \qquad (1)$$

To simplify the extraction process, it would be useful to ensure that the target stays centered in the frame across the entire sequence and then assignment the frame center with a positive keypoint and the edges with negative keypoints correspondingly in Z_i.

Garment Reconstruction with Neural Radiance Fields

For garment modelling and reconstruction, LYCB leverages NeRF2Mesh (Tang et al. 2022). Unlike NeRF (Mildenhall et al. 2020), NeRF2Mesh is different as it is able to recover the texture of the object, crucial for virtual try-on applications, and reconstruct detailed non-watertight and hollow surfaces - a key characteristic of garments. Figure 2 illustrates the key procedure of NeRF2Mesh end-to-end.

Succintly, neural radiance field methods attempt to learn how to represent a target object as a 3D volumetric density, given a collection of multi-view 2D images of said target with the corresponding camera parameters; e.g. given a ray projected from point Q of R^3 in a certain direction θ, return the corresponding density and chromatic values of a set of points sampled along said ray. In addition to the default density and chromatic values, NeRF2Mesh further decomposes the chromatic component into 2 sub-components - namely the diffuse and specular values with separate MLPs (dubbed the *Appearance Field*). By learning to represent the diffuse color and specular component, it enables NeRF2Mesh to generate the corresponding diffuse and specular texture map via exhaustive sampling and UV unwrapping after training.

Since the learned radiance fields represents only the volumetric density and appearance information, additional steps for surface reconstruction is required in order to transform said volumetric information into an actual 3D mesh. In the original NeRF (Mildenhall et al. 2020), Marching Cube is used for surface reconstruction, but the resulting mesh is inefficient and often contains undesirable artifacts for our application - blocky and admits non-hollow surfaces. NeRF2Mesh resolves these issues by adopting a course-to-fine mesh refinement schema with additional supervision to optimize for more accurate vertex displacements and reconstructed mesh faces on top of Marching Cube.

Overall, LYCB with NeRF2Mesh is able to generate high quality textured mesh model of the target garment in a neural implicit manner. The key loss functions of NeRF2Mesh are as follow, where r is the ray index, $C(.)$ the pixel color function, c_s the specular color value, q_k the k-th query point along ray r, w_i is the point-wise rendering weight of the k-th query point and $\mathbf{v}_i$ the vertex offsets array. All together, L_{total} denotes the total loss, L_{nerf} the conventional radiance field photometric loss, L_{etp} the entropy loss for better surface details, L_{offset} the loss term to regularize vertex displacements during mesh refinement, and L_ϵ the catch-all for optional loss terms not covered explicitly here for brevity. Finally, it's worth mentioning that for LYCB, the PyTorch NeRF2Mesh implementation from (Ashawkey 2023) is utilized.

$$L_{nerf} = \sum_r ||C(r) - \hat{C}(r)||^2 \tag{2}$$

$$L_{spec} = \sum_k |c_s(q_k)| \tag{3}$$

$$L_{etp} = -\sum_k (w_k log(w_k) + (1 - w_k) log(1 - w_k)) \tag{4}$$

$$L_{offset} = \sum_k (\Delta \mathbf{v}_i)^2 \tag{5}$$

$$L_{total} = L_{nerf} + L_{spec} + L_{etp} + L_{offset} + L_\epsilon \tag{6}$$

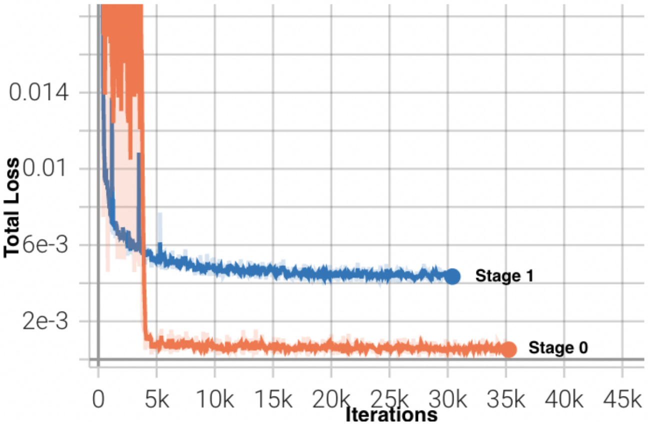

Figure 3: Training loss for Stage 0 and 1 respectively. *Lower* is better.

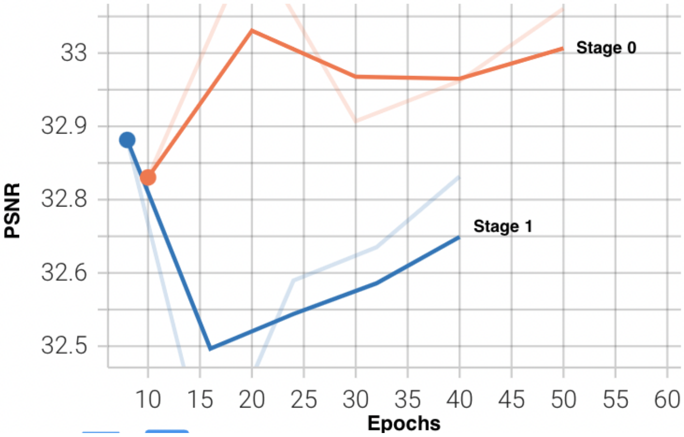

Figure 4: PSNR of synthesized views for Stage 0 and 1 respectively. *Higher* is better.

Results

Garment Fitting with Physics-Based Clothing Simulation

For virtual try-on, we have utilized Blender and its cloth physics simulation engine to fit our test garment on a non-descriptive mannequin as demonstration. Note that mesh cleaning is performed prior to the fitting step to ensure a clean and proper mesh.

We chose a physics based cloth simulator approach instead of an implicit alternative such as (Casado-Elvira, Comino Trinidad, and Casas 2022) since the former provides a much higher-degree of control over the garment characteristics and behaviour (e.g. stretchability, compression-resistance) and that it requires no additional training to do so; additionally the result of physics based simulator is well-behaved and better bounded as well. It is worth nothing that the target body for fitting can be freely swapped out since the reconstructed garment is a generic mesh object, and is amenable to all treatments and manipulation like any mesh.

For validation, sample data was collected as seen in Figure 5, depicting a 360° monocular sequence of a subject with the target garment(the beige top). As with NeRF, LYCB is not trained to generalize (e.g. learning to fit) but instead to represent the target garment as accurately as possible (e.g. overfit on the target garment). Therefore, a large scale dataset is not required since each training instance per new sample will be trained to overfit as much as possible.

In terms of metric, the total loss L_{total} as outlined in Equation 6, and the Peak-Signal-to-Noise-Ratio (PSNR) of the synthesized novel views are used to to evaluate the reconstruction performance. For the fitting of the reconstructed garment mesh on a foreign body (e.g. mannequin), there are unfortunately no direct way to evaluate the performance since there are no ground truths 3D mesh of how such garment would fit said foreign body without explicit modelling. Conceivably, in later work a more extensive dedicated dataset with known foreign bodies can be collected and constructed (e.g. let subjects swap clothing and take some as

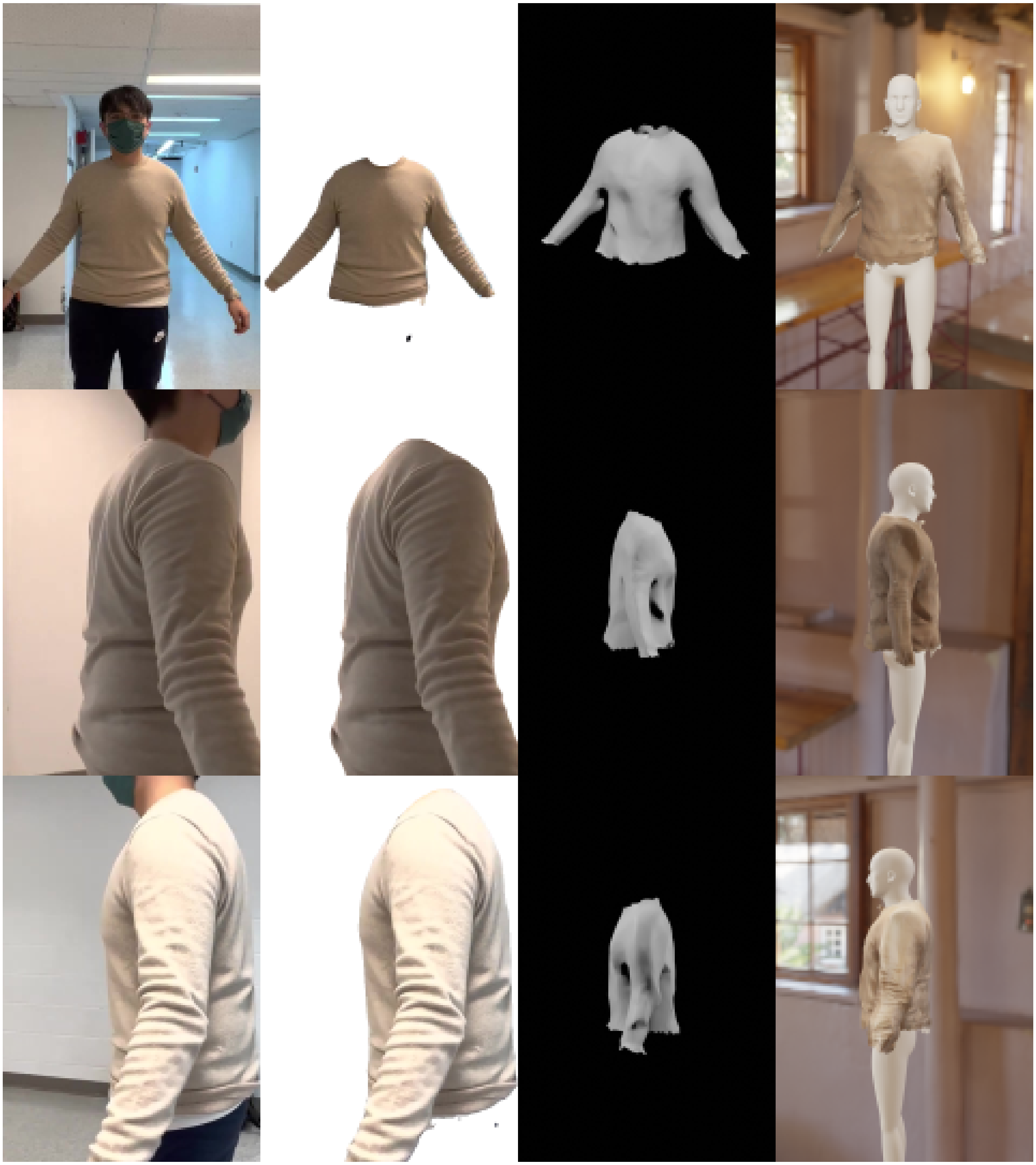

Figure 5: Sample output of each stage across 3 frames. From **left** to **right** - input sequence, extracted garment, reconstructed mesh and fitted garment on mannequin in Blender.

ground truths via explicit 3D scan or equivalent photogrammetry methods). However for now, the former evaluation metrics will be used for evaluation.

Figure 3 and Figure 4 shows the L_{total} and PSNR value for stage 0 (course mesh) and stage 1 (iterative mesh refinement) respectively. As observed, the loss converges quite rapidly and starts to saturate around $5k$ iterations for stage 0 and $20k$ iterations for stage 1 respectively. Notably for both the loss and PSNR values, the result of stage 1 is higher than that of stage 0. We hypothesized that this is likely because in stage 1, the mesh refinement procedure attempts to generate a fine mesh with smooth, well defined surface (enforced via a Laplacian smoothness term) and removes appendages that the algorithm deems to be noisy; potentially at the cost of reduced PSNR and increased loss when compared to the projected 2D input references, which itself may contain noise from imperfect environment and noisy segmentation mask.

On the other hand, we can observe in Figure 5 that qualitatively speaking, the reconstructed garment has good fidelity, thus validating the approach; the overall geometric shape, structure and texture indeed closely resembles the garment as seen in the input sequence. Although there extracted masked garment in the second column is not perfect, it did not impact the reconstruction significantly and further indicates that the refinement process is capable to handling such noise. Additionally, it is also noteworthy to point out that the fitted garment as seen in the fourth column does seem to conform to the mannequin's body in a reasonable manner as expected (e.g. note the body-to-waist ratio of the fit between the first and fourth column).

Conclusion

In this work, we have put forth a framework for garment extraction, explicit mesh reconstruction and virtual fitting that requires only a monocular video as input - dubbed Leave Your Clothes Behind (LYCB).

Unlike existing alternatives that either suffers from lackluster physical compliance (e.g. inaccurate virtual fitting on foreign body) or is incapable of reconstructing complex garments, LYCB is able to fit the gap by providing both accurate and flexible virtual fitting, and capability to model complex garments via physics-based clothing simulation and neural radiance fields that are both model (e.g. any clothing type) and subject (e.g. could transfer between any foreign bodies for fitting) agnostic.

However, LYCB has its disadvantages as well, such as the potential need to fine-tune the garment extraction process with SAM depending on the input, and the long processing time of the entire pipeline; e.g. from end-to-end, the entire process could span hours depending on various factors such as the number of input frames, the input resolution, the complexity of COLMAP parameter estimation for the given sequence and the number of NeRF2Mesh fitting iterations. We believe that with the reconstruction quality and flexibility demonstrated by LYCB, if the aforementioned issues are alleviated, hybrid approaches such as LYCB that blends cutting-edge Deep Learning methods and well-established physics-based simulation could be a viable path forward in addressing useful tasks such as personalized virtual try-on and easy virtual asset generation in the future.

References

Ashawkey. 2023. nerf2mesh. https://github.com/ashawkey/nerf2mesh.

Casado-Elvira, A.; Comino Trinidad, M.; and Casas, D. 2022. PERGAMO: Personalized 3D Garments from Monocular video. *Computer Graphics Forum (Proc. of SCA), 2022*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Feng, Y.; Yang, J.; Pollefeys, M.; Black, M. J.; and Bolkart, T. 2022. Capturing and Animation of Body and Clothing from Monocular Video. In *SIGGRAPH Asia 2022 Conference Papers*, SA '22.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. *arXiv:2304.02643*.

Lowe, D. G. 1999. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, 1150–1157. Ieee.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.*, 41(4): 102:1–102:15.

Patel, C.; Liao, Z.; and Pons-Moll, G. 2020. TailorNet: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. *CoRR*, abs/2103.00020.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Schönberger, J. L.; Zheng, E.; Pollefeys, M.; and Frahm, J.-M. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*.

Tang, J.; Zhou, H.; Chen, X.; Hu, T.; Ding, E.; Wang, J.; and Zeng, G. 2022. Delicate Textured Mesh Recovery from NeRF via Adaptive Surface Refinement. *arXiv preprint arXiv:2303.02091*.

Xu, Y.; Yang, S.; Sun, W.; Tan, L.; Li, K.; and Zhou, H. 2019. 3d virtual garment modeling from rgb images. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 37–45. IEEE.