VTON 360: High-Fidelity Virtual Try-On from Any Viewing Direction

Supplementary Material

Appendix A introduces the preliminaries of 3DGS. The detailed formulations of the two quantitative metrics are presented in Appendix B. Additionally, Appendix C outlines the post-processing techniques applied to ensure the preservation of human characteristics in image editing. Appendix D elaborates on the failure cases and proposes a mitigation strategy to address it. Finally, Appendix E showcases additional VTON results, including those from a real 3D scene used in GaussianVTON [6].

A. 3D Representation: Gaussian Splatting

3D Gaussian Splatting (3DGS) [25] has emerged as a prominent technique in 3D reconstruction due to its ability to render high-quality scenes in real-time. Unlike traditional point cloud based methods, which directly represent scenes as discrete points, 3DGS models each point as a continuous Gaussian function g_i:

$$g_i(x;\mu_i,\Sigma_i) = e^{-\frac{1}{2}(x-\mu_i)^\top \Sigma_i (x-\mu_i)}, \tag{7}$$

where x is the position vector of g_i, $\mu_i \in \mathbb{R}^3$ and $\Sigma_i \in \mathbb{R}^{3\times3}$ are g_i's mean and covariance matrix, respectively. Then, g_i is projected onto a 2D image plane to facilitate rendering. This projection yields a new mean vector ${\mu_i}' \in \mathbb{R}^2$ and an updated covariance matrix $\Sigma_i' \in \mathbb{R}^{2\times2}$ defined as:

$${\mu_i}' = KT[{\mu_i}^\top, 1]^\top, \Sigma_i' = JT\Sigma_i T^\top J^\top, \tag{8}$$

where J is the Jacobian matrix derived from the affine approximation of the perspective projection, T and K denote the extrinsic and intrinsic matrices, respectively. Given the color c_i and opacity α_i at the Gaussian center point, the rendered color at a 2D pixel p is calculated as follows:

$$\begin{aligned} C_p &= \sum_{i=1}^{N} \alpha_i c_i T_i g_i(p;\mu_i',\Sigma_i') \\ T_i &= \prod_{j=1}^{i-1} (1-\alpha_j g_j(p;\mu_j',\Sigma_j')), \end{aligned} \tag{9}$$

where T_i denotes the cumulative transmission along the ray.

B. Metrics

In the quantitative evaluation, we employ two metrics:

- Average DINO Similarity [63], which measures the alignment between the garment image and the edited 3D human.
- CLIP Directional Consistency Score [17], which evaluates multi-view consistency.

Specifically, given an edited 3D human (after VTON), 120 views are uniformly projected around its central axis. These views are divided into three categories based on orientation: S_f, S_b, and S_s, corresponding to 40 front views, 40 back views, and 40 side views, respectively. Let $D(\cdot)$ represent the normalized DINO embedding and $C(\cdot)$ denote the normalized CLIP embedding. Using these, we formally define the two metrics as follows:

$$\begin{aligned} \text{DINO}_{sim} &= \frac{1}{80}\Big(\sum_{i\in S_f} D(g_f)\cdot D(e_i) + \sum_{i\in S_b} D(g_b)\cdot D(e_i)\Big) \\ \text{CLIP}_{cons} &= \frac{1}{120}\sum_{i}(C(e_i)-C(o_i))\cdot(C(e_{i+1})-C(o_{i+1})) \end{aligned} \tag{10}$$

where e_i, e_{i+1} and o_i, o_{i+1} denotes the two consecutive novel views from the edited 3DGS and the original 3DGS, respectively.

C. Post-processing

The clothing-agnostic maps $\mathbf{A}$ often mask parts of the face and hair, particularly for females. Due to the inherent properties of the diffusion model, it is unable to fully restore the intricate details of these masked regions. To ensure high-fidelity preservation of human characteristics, we apply a post-processing step where, after editing the rendered views, we "copy" the face and hair from the original image o onto the edited image e. Specifically, let m represent the region corresponding to the face and hair, which can be extracted from the parsed map during pre-processing, we implement post-processing as:

$$e = (1-m)\cdot e + m\cdot o \tag{11}$$

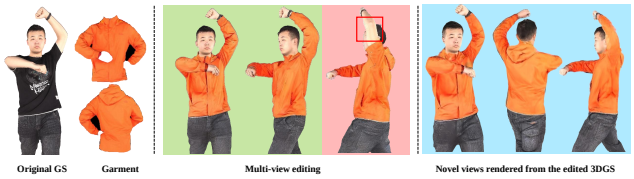

Figure 8. Our multi-view editing may fail in certain views with complex poses (red box in pink background) but these views can be automatically discarded to mitigate their impact on 3D VTON (blue background).

D. Limitations

As shown in Fig. 8, our method may fail in certain views with complex postures. To address this, we use Z-Score

Normalization to automatically identify and discard problematic views based on the view reconstruction loss during the process of lifting multiple views to 3D space, mitigating their adverse impact.

E. Additional Visualization Results

Fig. 9 illustrates additional VTON results. The first two rows showcase results from the THuman2.0 dataset; the middle two rows showcase results from the MVHumanNet dataset. To further demonstrate the effectiveness of our method, we apply it on a real 3D scene used in GaussianVTON [6]. The last two rows in Fig. 9 illustrate these VTON results with the model trained on Thuman2.0 dataset. Despite the data gap, including w/wo background and unseen camera poses, our method exhibits robust performance and preserves the details of the clothing well.

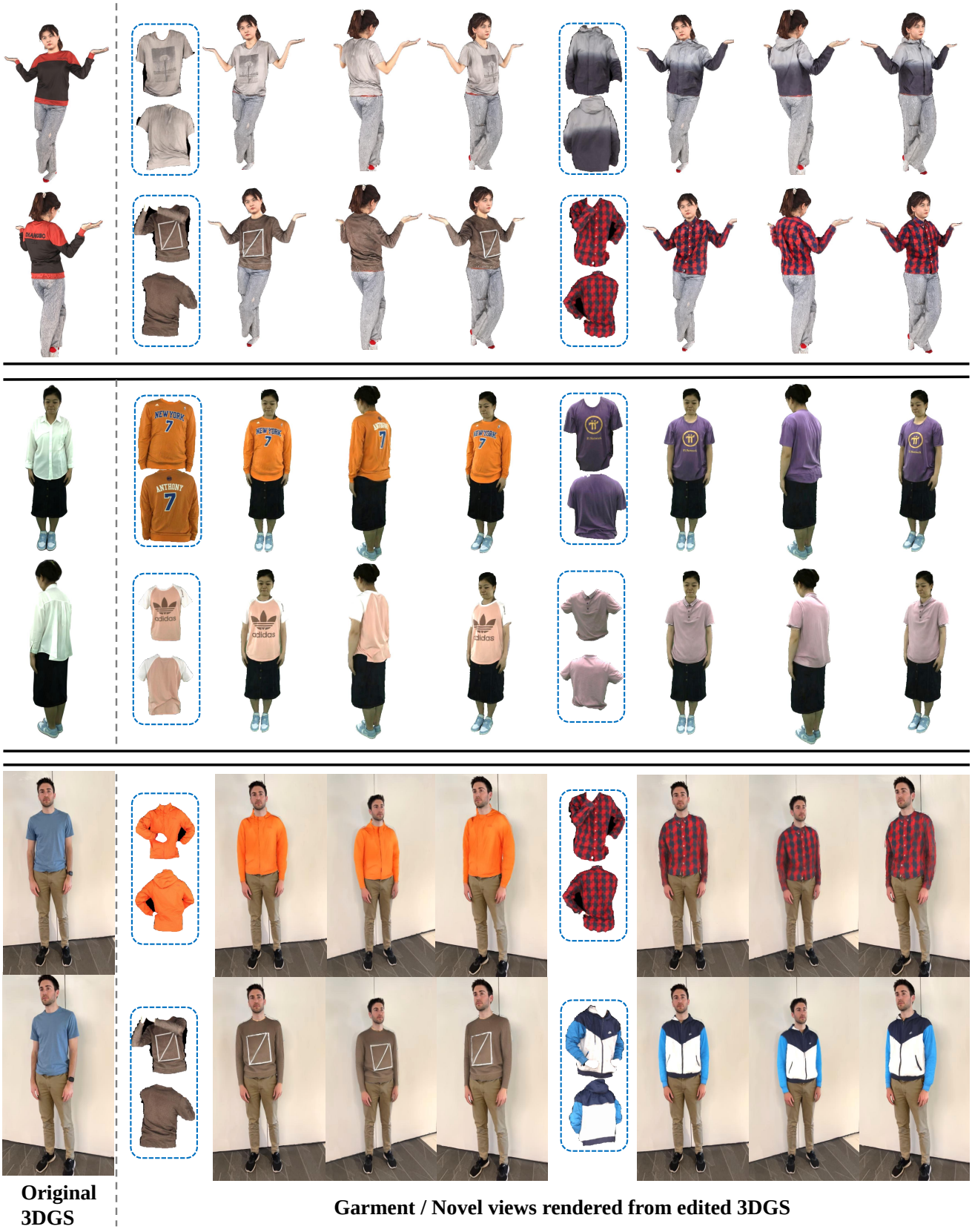

Figure 9. **Additional visualization results.** The first, middle, and last two rows show results on Thuman2.0, MVHumanNet, and a real 3D scene used in GaussianVTON, respectively.