HumanOrbit: 3D Human Reconstruction as 360° Orbit Generation

Supplementary Material

Figure 11. Example of the initial frames in the videos used for training. For each 3D scan, we generate 6 different orbit videos.

7. Omni-View Image Generation

7.1. Training Data

To train HumanOrbit, we render orbit videos using 500 3D human scans from the PosedPro dataset. For each scan, we render both a full body and head orbit video. In each case, we render from the initial front view, but also render with slight rotations such that the subject is facing the left and right. An example of the different initial frames is shown in Figure 11. This results in a total of $500 \times 2 \times 3 = 3000$ videos using during training.

7.2. Pre-processing

Our HumanOrbit model not only outputs view consistent frames of the subject, but also tries to generate consistent background as well. This can be slightly problematic for some images taken in the wild where there are objects near the person. As our model tries to do a full circle around the person, the object can potentially occlude parts of the body from certain views. To mitigate this, we have an optional preprocessing where we apply rembg library's 'birefnet-general' model to segment the person and replace the background with a solid gray background. We used this preprocessing when generating results for the CCP dataset where it contains images of people taken in cities with potential occlusions. We did not apply it for the CelebA dataset as the photos are taken close-up without any nearby objects.

7.3. Post-processing

As mentioned before, HumanOrbit can synthesize a background for each generated view. However, this is unnecessary when reconstructing the target human. Additionally, the optimization for the mesh carving requires a ground truth silhouette mask for each view. Therefore, we conduct segmentation of the person in the generated video and create a video with a solid white background. For the segmentation, we segment the initial frame, similar to the pre-processing strategy, and then apply SAM 2.1's 'hiera_tiny' model to propagate the masks to the rest of the video frames.

8. 3D Reconstruction

8.1. Camera Pose Estimation with VGGT

After running VGGT for the initial 3D scene estimation, we conduct bundle adjustment to further refine the camera pose and point cloud. This is done in a similar manner that the original authors provide in their code, but we make two simple adjustments. First, to determine the query frames for computing the 2D query points, the original method used DINOv2 features to rank the representativeness of each frame to cover diverse appearances. However, this does not always get a full coverage of the different views. Therefore, we simplify this process by selecting evenly spaced out frames in the orbit video. Second, to compute the query points to track across frames, the original method applies a keypoint estimator. This method puts a focus on edges and areas with texture, resulting in a 3D point cloud with holes in low texture areas. To obtain a more complete coverage, we remove the keypoint estimator and uniformly sample pixels within the segmentation mask which are then tracked across frames.

9. Results

9.1. Multi-view Image Generation

9.1.1. Our Results as Videos

Compared to previous multi-view generation methods, our method creates much denser views. However, this is hard to visualize as 2D images in a limited amount of space. We provide video files to show example orbit videos generated by our method from a single image for both the CCP ('our_orbit_video_fullbody.mp4') and CelebA ('our_orbit_videos_head.mp4') datasets.

9.1.2. VGGT Results on Different Methods

As mentioned in the main paper, VGGT does not create a complete 3D reconstruction for multi-view images generated by competing methods. This is visualized in Figure 12. When VGGT is applied to the results from other methods, it can be seen that the predicted point cloud is incomplete, showing only the front half of body. VGGT filters out

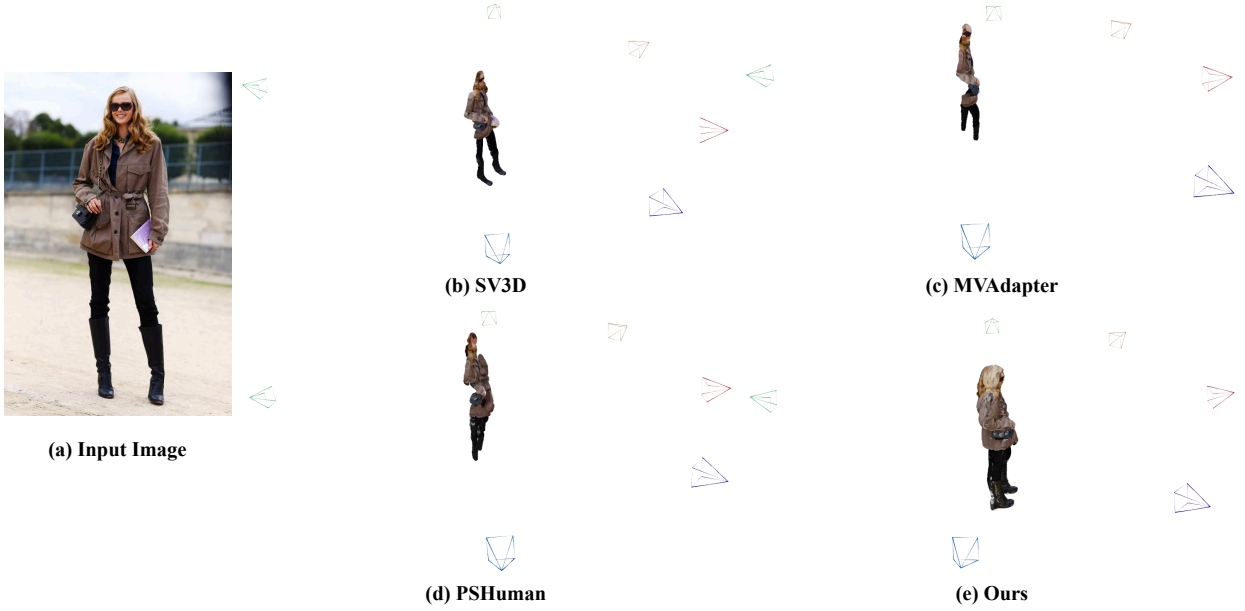

Figure 12. Comparison of the point cloud predicted by VGGT on multi-view images generated by various methods.

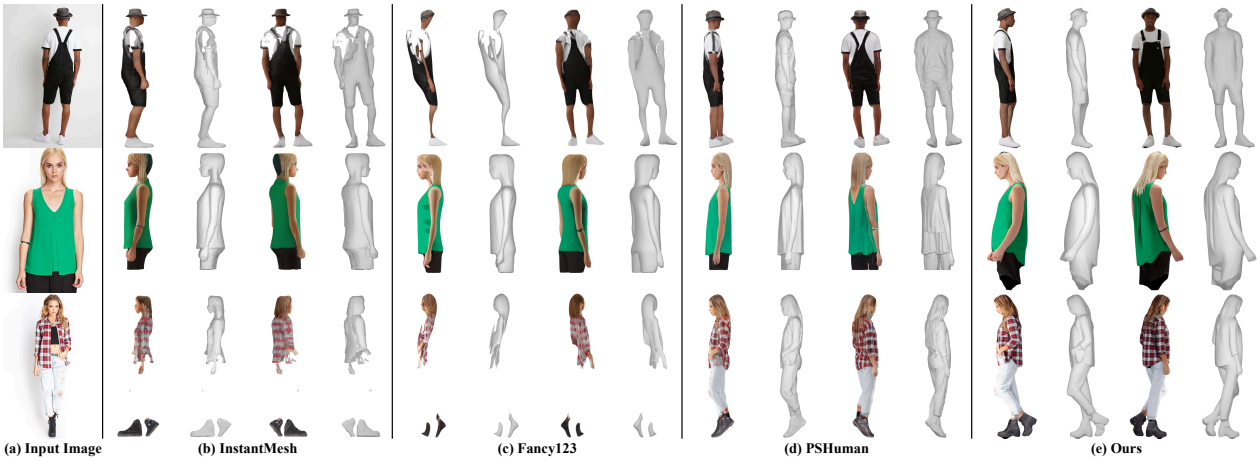

Figure 13. Additional visual comparison of the appearance and geometry of 3D mesh reconstruction methods on a body images.

points with low confidence suggesting that the multi-view images generated by the competing methods contains more inconsistencies than those created by our method.

9.2. Mesh Generation

Additional visual comparisons of the mesh reconstructed on full body and partial body images are shown in Figure 13. In the first example of a photo of a person from behind, none of the methods are able to reconstruct the facial features resulting in a person with no face. In contrast, our method is able to generate a plausible face with consistent clothing and pose even for images taken from behind. In the second example of the upper body photo, Fancy123 and PSHuman reconstruct a flat body with thin arms. Furthermore, the mesh generated by PSHuman has a large part of the lower back missing. Compared to these methods, our method reconstructs a more complete mesh with realistic appearance and shape. In the third example, it can be seen that InstantMesh and Fancy123 creates a mesh where most of the lower half of the body is missing. In addition, PSHuman merges the two shoes together, while our method is able to separate the shoes. We also provide a RGB image and normal rendering comparison with PSHuman as a video (‘fullbody_rendering_comparison_with_pshuman.mp4’).

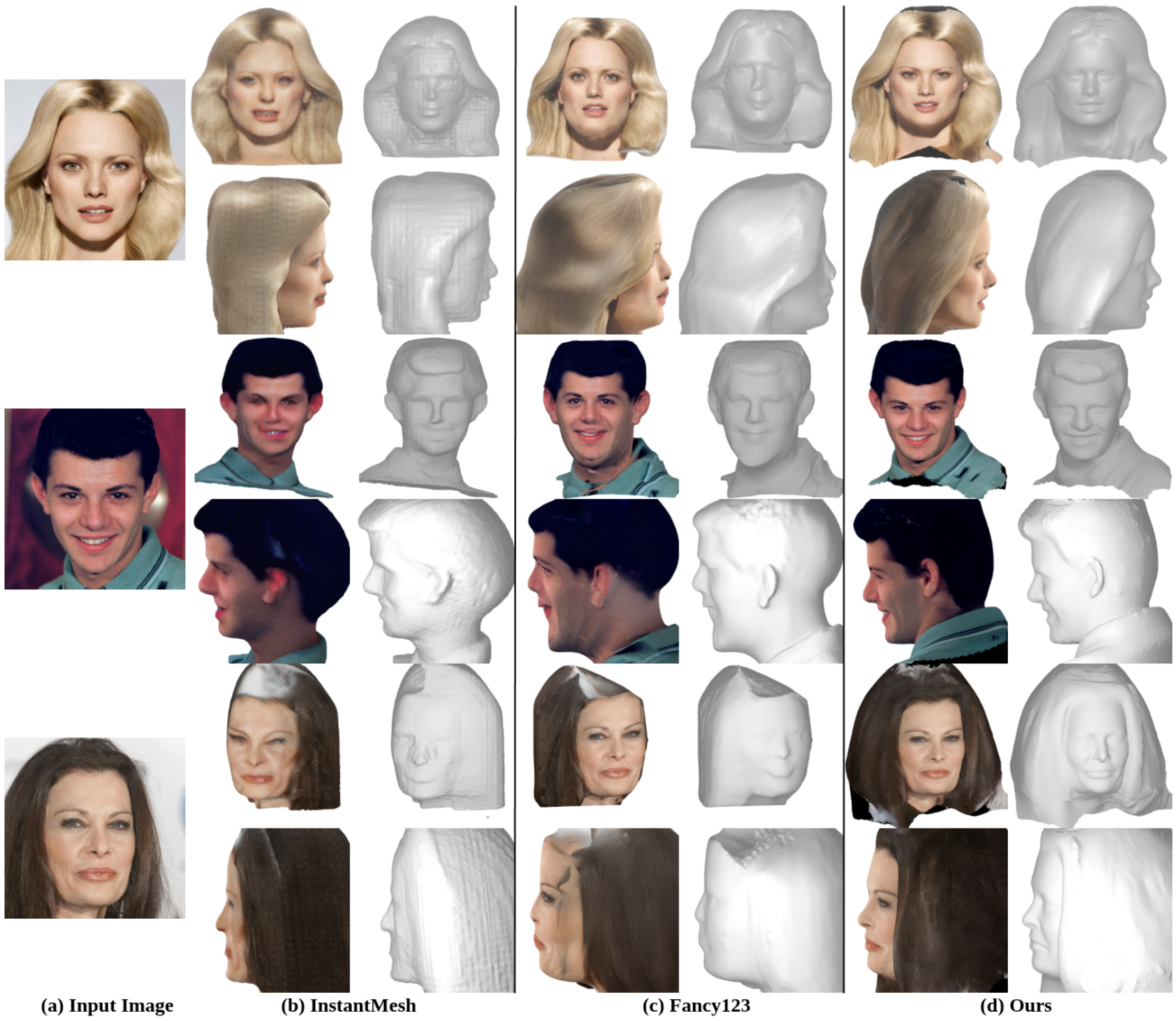

Figure 14. Additional visual comparison of the appearance and geometry of 3D mesh reconstruction methods on a head portrait.

Additional visual comparisons of the head mesh reconstructed by various methods are shown in Figure 14. For Fancy123, the appearance of the frontal view looks accurate as they can directly project the colors from the input image, but when looking at the novel views and the geometry it does not look as detailed. In contrast, our method creates a more detailed geometry of the nose, mouth, and ears with a realistic appearance.