Easy3E: Feed-Forward 3D Asset Editing via Rectified Voxel Flow

Supplementary Material

A. Editing Efficiency

Table 3. Runtime comparison with different methods.

Method	Runtime
Vox-E [40]	37 min
MVEdit [6]	212 s
Instant3dit [2]	25 s
Ours	**75 s**

We compare the efficiency of our approach with three baselines: Instant3dit, MVEdit, and Vox-E, as shown in Tab. 3. Instant3dit is the fastest (25 seconds) due to its highly streamlined pipeline, but this compactness often limits its ability to handle complex disentanglement, resulting in lower fidelity. In contrast, MVEdit incurs a significantly higher computational cost of 212 seconds, as it relies on heavy iterative multi-view diffusion refinement. Vox-E is even more time-consuming, requiring full 3D optimization of the voxel grid, which takes approximately 37 minutes per edit. Our method operates in a sweet spot with a total runtime of 75 seconds. Adopting an efficient feed-forward design similar to Instant3dit, our approach eliminates the need for time-consuming per-scene optimization. However, unlike the unified pipeline of Instant3dit, we utilize a structured workflow decomposed into geometry editing (30 seconds), texture refinement (30 seconds), and back-projection (15 seconds). This 75 seconds duration allows us to achieve substantially better consistency and detail than Instant3dit, while remaining orders of magnitude faster than the optimization-heavy baselines.

B. Effectiveness of Texture Refinement

Fig. 7 presents the qualitative results obtained from our normal-guided Multi-view Diffusion Module. Conditioned on multi-view normal maps rendered from the input mesh and a single reference image, the module synthesizes a coherent sequence of multi-view images. As shown in the visualization, the generated images exhibit high-fidelity textures with intricate details. More importantly, benefiting from the structural guidance of surface normals, the results demonstrate rigorous cross-view consistency, where the object identity and geometric features remain stable across varying camera poses. This ensures that the subsequent texture back-projection step produces a seamless 3D model without alignment artifacts.

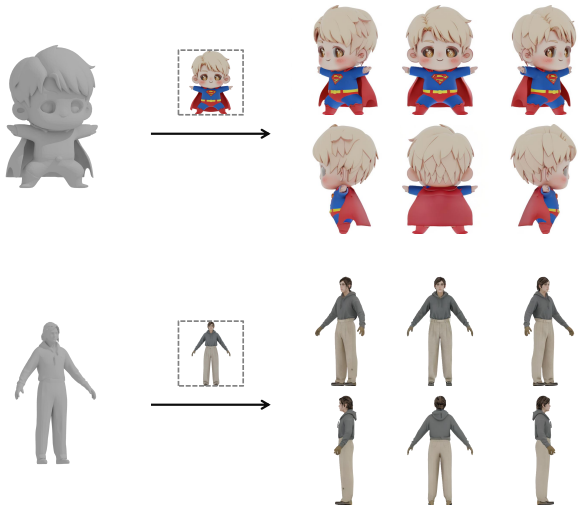

Figure 7. Visualizations of the normal-guided multi-view diffusion module. Taking rendered normal maps and a reference image as input, the module generates multi-view images that are both texture-rich and geometrically consistent, serving as robust priors for 3D texture recovery.

C. More Results

We present six supplementary examples in Fig. 8 to further validate the robustness of our approach. (a)-(c) demonstrate our capabilities on non-realistic objects; observe that the edited regions undergo significant geometric deformation, while the geometry of the unedited regions is strictly preserved. (d) illustrates the effectiveness of our method in scene-level editing. (e)-(f) showcase results on human subjects, where the clothing is successfully modified with high visual quality while maintaining the texture fidelity of the unedited body parts.

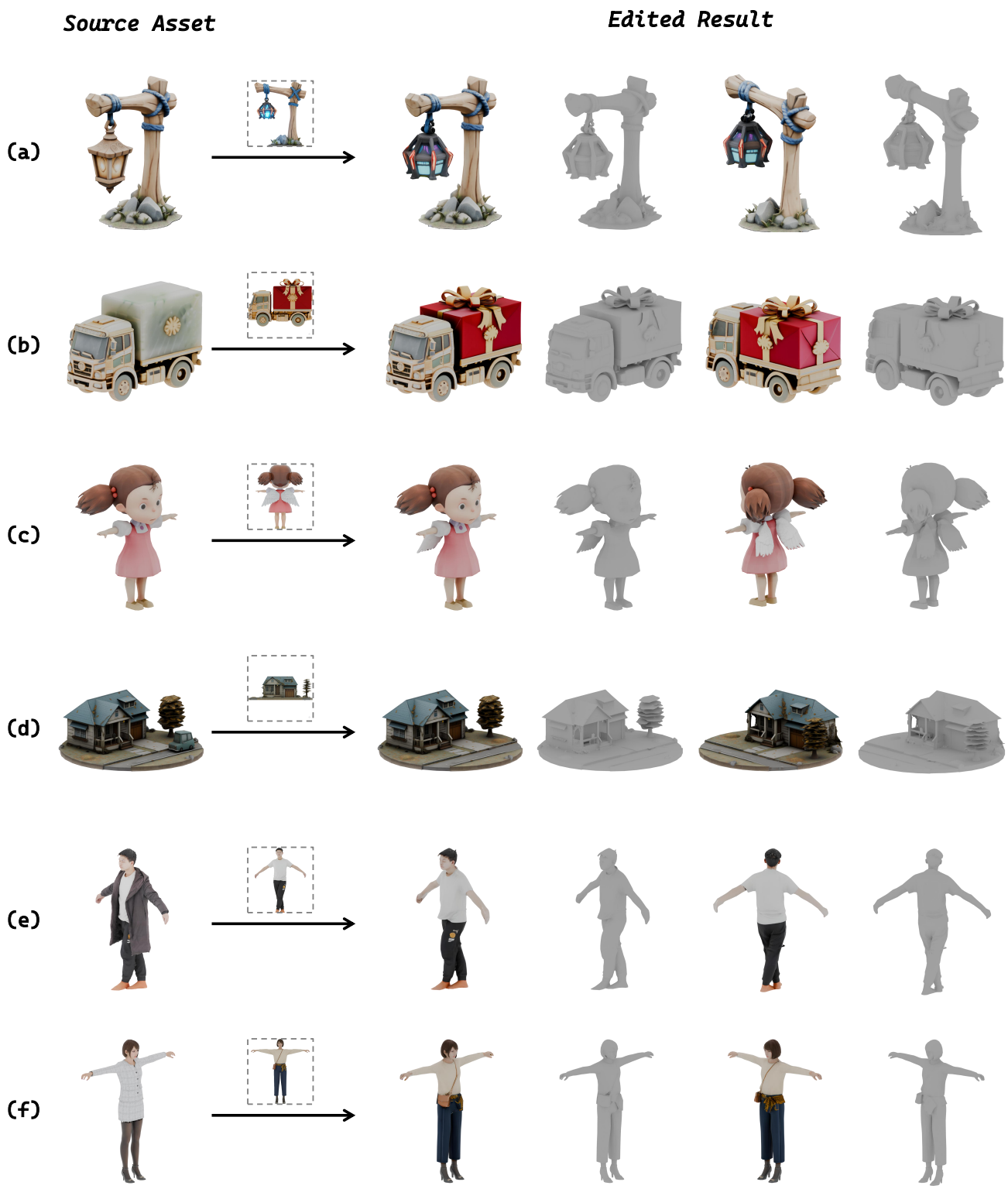

Figure 8. More visualization results.