Hi3D: Pursuing High-Resolution Image-to-3D Generation with Video Diffusion Models —ACMMM 2024 Supplementary Material

Anonymous Authors

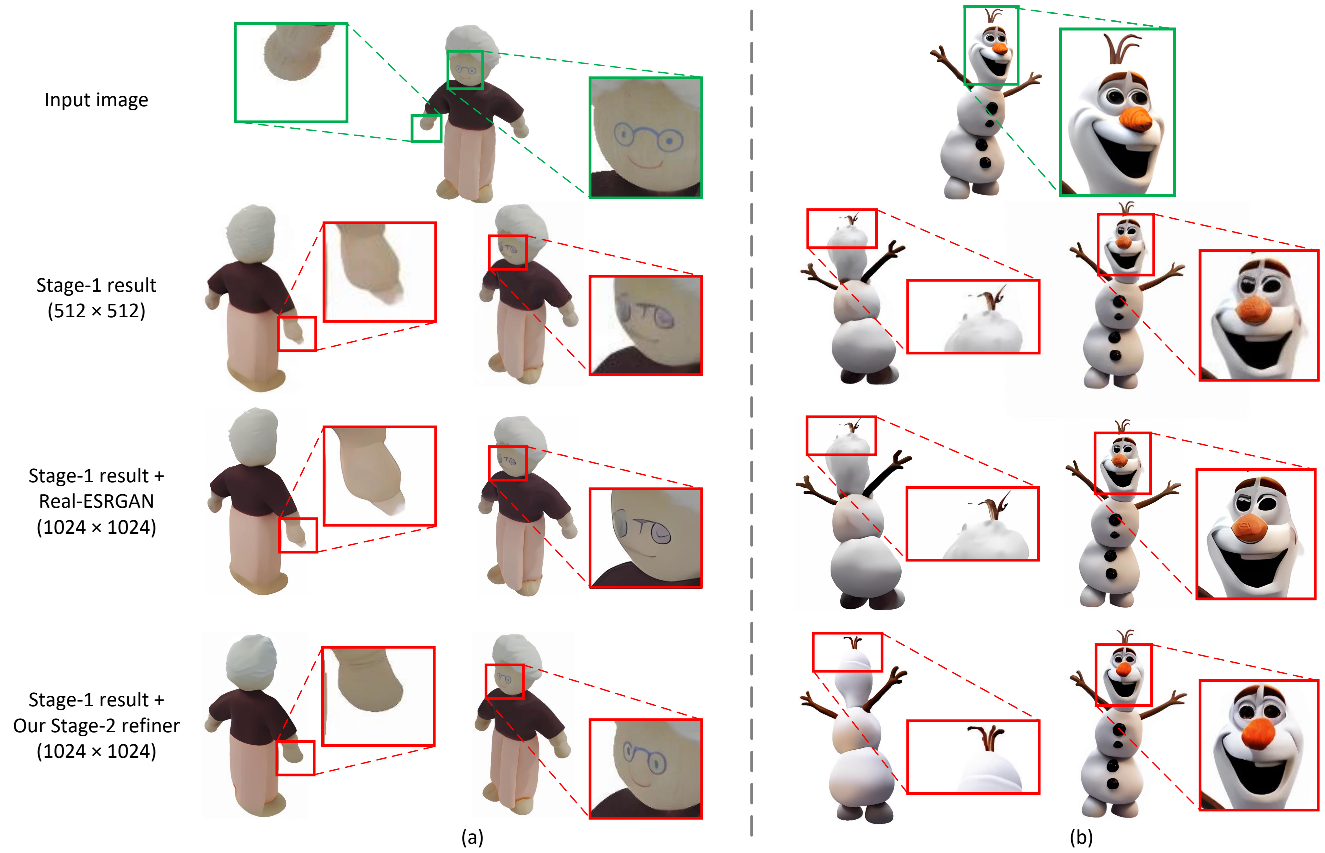

Figure 1: Comparing our 3D-aware video-to-video refiner with typical super-resolution method (Real-ESRGAN [1]) in Stage-2.

This supplementary material contains: 1) comparison of the proposed 3D-aware video-to-video refiner with conventional super-resolution, and 2) more image-to-3D generation results.

1 COMPARISONS WITH SUPER-RESOLUTION

Recall that in Stage-2 (Sec. 4.3 in the main paper), we devise a new 3D-aware video-to-video refiner to further scale up the low-resolution (512×512) outputs of Stage-1 to higher resolution ($1,024\times 1,024$). An alternative solution is to use a super-resolution (SR) model to directly upscale the generated multi-view images in Stage-1 into $1,024\times 1,024$ resolution. Here we adopt a typical SR method (Real-ESRGAN [1]) for comparison.

Figure 1 showcases the comparison results. The SR method can only eliminate the blurriness and produce sharp outputs, but fails to alleviate the geometry and appearance distortions in the input multi-view images. Taking Figure 1 (a) as an example, compared with the input image, the “hands” and “face” in the generated images of Stage-1 are distorted. The SR method Real-ESRGAN cannot correct these distortions as it was primarily trained to produce high-resolution images strictly consistent with the input. Thus these distortions inevitably remained in the SR outputs. In contrast, our devised 3D-aware video-to-video refiner not only produces clear results with no blur, but also generates correct “hand” and “face” that are consistent with the input image. This comparison clearly demonstrates the effectiveness of our proposed 3D-aware video-to-video refiner for generating high-resolution ($1,024\times 1,024$) multi-view images with finer 3D details and consistency.

2 MORE IMAGE-TO-3D RESULTS

Please refer to the “Hi3D.mp4” file for additional visual results of our Hi3D model and baseline methods.

REFERENCES

[1] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCVW*.