Supplementary Material for 'MVDiffusion++: A Dense High-resolution Multi-view Diffusion Model for Single or Sparse-view 3D Object Reconstruction'

Shitao Tang[1*], Jiacheng Chen[1*], Dilin Wang[2*], Chengzhou Tang[2], Fuyang Zhang[1], Yuchen Fan[2], Vikas Chandra[2], Yasutaka Furukawa[1†], and Rakesh Ranjan[2†]

[1] Simon Fraser University
{shitaot, jca348, fuyangz, furukawa}@sfu.ca
[2] Meta Reality Labs
{wdilin, chengzhout, ycfan, vchandra, rakeshr}@meta.com

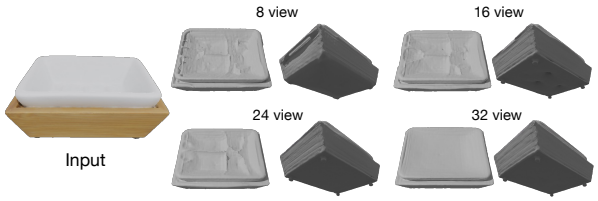
This supplementary document provides 1) Training details and 2) an Ablation study on the number of generated views. We also provide additional resources in the same folder as this document: 1) a short video (./demo.mp4) for quickly demonstrating the performance of MVDiffHD and 2) a local web page (./page) that contains many qualitative examples for single or sparse-view 3D object reconstruction. Please open the file "./page/index.html" with a browser to view our web page.

1 Training details

After initializing the UNet model weights by a pre-trained latent diffusion inpainting model, we train the proposed system in three stages. First, we train as an ϵ-prediction model only with single-view conditioning cases, because our pre-trained model was trained as ϵ-prediction. Second, we fine-tune as a v-prediction model [3] still with single-view conditioning cases. Third, we fine-tune as a v-prediction model with both single and sparse-view conditioning cases. Half the samples are single-view conditioning, and the other half are sparse-view conditioning, where the number of condition images is uniformly sampled between 2 and 10. We employ an AdamW optimizer with a learning rate of 7e-5 and use a cosine learning rate scheduler similar to Zero123++. We also incorporate the Zero-SNR fix [2] commonly adopted in video generation models.

2 Ablation study on the number of generated views

Figure 1 varies the number of generated images. The 32 views are divided into four elevation-based groups with the order 30°, 0°, -30°, 60°. The 8-view setting only generates the first group, and the 16-view setting generates the first two groups, and so on. The results suggest that fewer views cannot cover the entire object, leading to worse reconstruction quality. While the metric scores are almost



Views	CD ↓	IoU↑
8	0.0262	0.6020
16	0.0216	0.6532
24	0.0208	0.6657
32	0.0208	0.6689

Fig. 1: Ablation study on different numbers of generated views. **Left**: a qualitative comparison shows that denser views produce mesh with better quality. **Right**: quantitative results of 3D reconstruction using different numbers of generated views, evaluated on 30 GSO [1] objects.

the same between 24 views and 32 views, the mesh reconstructions from 32 views look smoother without artifacts.

Bibliography

[1] Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: ICRA (2022) 2

[2] Lin, S., Liu, B., Li, J., Yang, X.: Common diffusion noise schedules and sample steps are flawed. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5404–5411 (2024) 1

[3] Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022) 1