

# [Supplementary Materials] IMFine: 3D Inpainting via Geometry-guided Multi-view Refinement

Zhihao Shi<sup>\*1</sup>

Dong Huo<sup>\*o2</sup>

Yuhongze Zhou<sup>1</sup>

Yan Min<sup>o3</sup>

Juwei Lu<sup>1</sup>

Xinxin Zuo<sup>4</sup>

<sup>1</sup>Huawei Canada Research Institute

<sup>2</sup>University of Alberta

<sup>3</sup>McMaster University

<sup>4</sup>Concordia University

zhihaoshi2022@gmail.com, dhuo@ualberta.ca, yuhongze.zhou@mail.mcgill.com

miny13@mcmaster.ca, juwei.lu@huawei.com, xinxin.zuo@concordia.ca

## 1. Network Architecture

The network input follows the same structure as Stable Diffusion inpainting, where an encoded warped image, a down-sampled binary mask, and a random noise map are concatenated. Multiple views are processed in parallel by concatenating them along the batch dimension. Additionally, features from different views interact through our updated attention layers, referred to as space-time attention. These layers enable each view to attend to neighboring views and a specific reference view, ensuring view consistency.

## 2. Fine-tuning Data Synthesis

In Fig. 2(b) of the main paper, we illustrate that the data used to fine-tune the multi-view refinement model is synthesized from the randomly selected view. Specifically, as illustrated in Fig. 9, we begin with the extracted 3D mesh from the unedited GS scene. An arbitrary view is chosen, and a random mask is generated around the target object. This view and the corresponding mask are then warped to other views, guided by the extracted mesh.

To promote cross-view learning, we also generate an additional set of paired images by independently applying image augmentations to each view. As shown in Fig. 10, an irregularly shaped mask is randomly generated on a randomly selected view. Image-based augmentations—such as elastic transformations and color jittering, are then applied to the masked region to simulate warping artifacts.

## 3. Dataset Description

The proposed dataset comprises 20 scenes with diverse characteristics. Specifically, there are 2, 4, 4, and 10 scenes where the camera trajectory spans 90°, 120°, 180°, and

360°, respectively. Additionally, the dataset includes 4, 5, 5, and 6 scenes featuring single-plane, curved-plane, multi-plane, and irregular interfaces. Each scene consists of 175 images, along with corresponding camera poses and object masks. The first 50 images are used for evaluation purposes with the object removed, while the remaining 125 images containing the object are utilized to reconstruct the original Gaussian Splatting scene (GS). More data visualization is shown in Fig. 11 to 14.

Object masks for the training views are generated using SAM2, with human-provided masks for the first frame. For the testing views, object masks are derived through a three-step process: (1) reconstructing a GS scene using the training images, (2) segmenting object Gaussians based on the training object masks, and (3) rendering the object masks for the test views.

## 4. More Qualitative Evaluations

We provide additional qualitative evaluations and failure cases in the attached video. Our method has limitations when the original 3D scene is poorly reconstructed. Under such conditions, significantly anisotropic or oversized Gaussian primitives may become exposed in the pruned scene, negatively impacting subsequent processing.

\* denotes equal contribution.

o Work done during an internship at Huawei Canada Research Institute

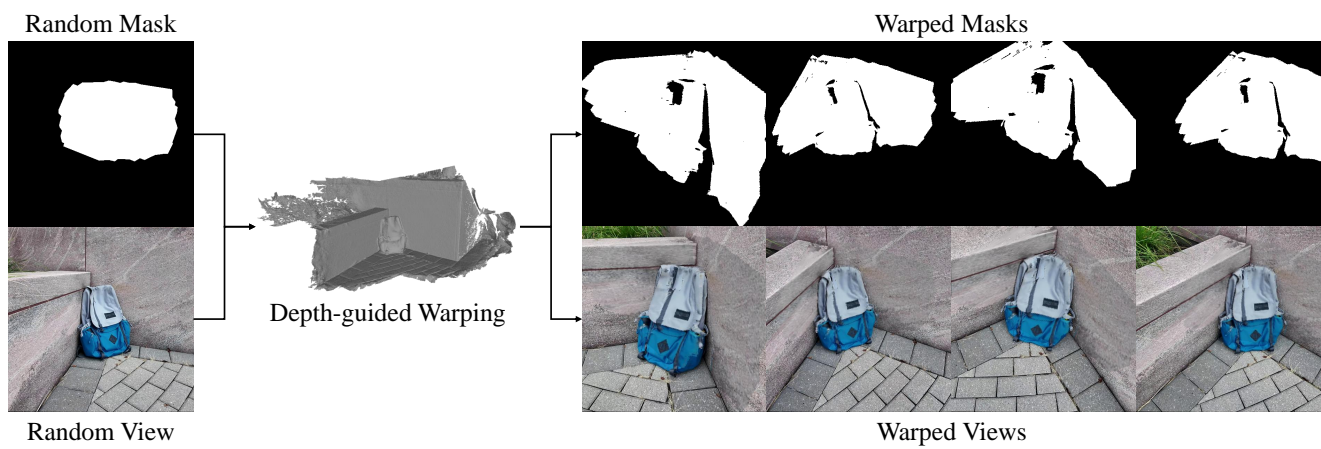


Figure 9. Data synthesis based on depth-guided warping.

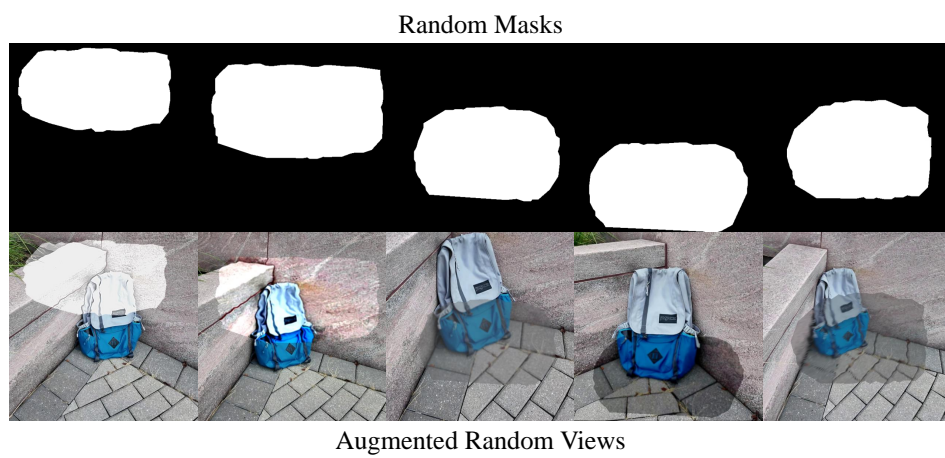


Figure 10. Data synthesis based on image augmentation.





Figure 11. Dataset visualization - 1.



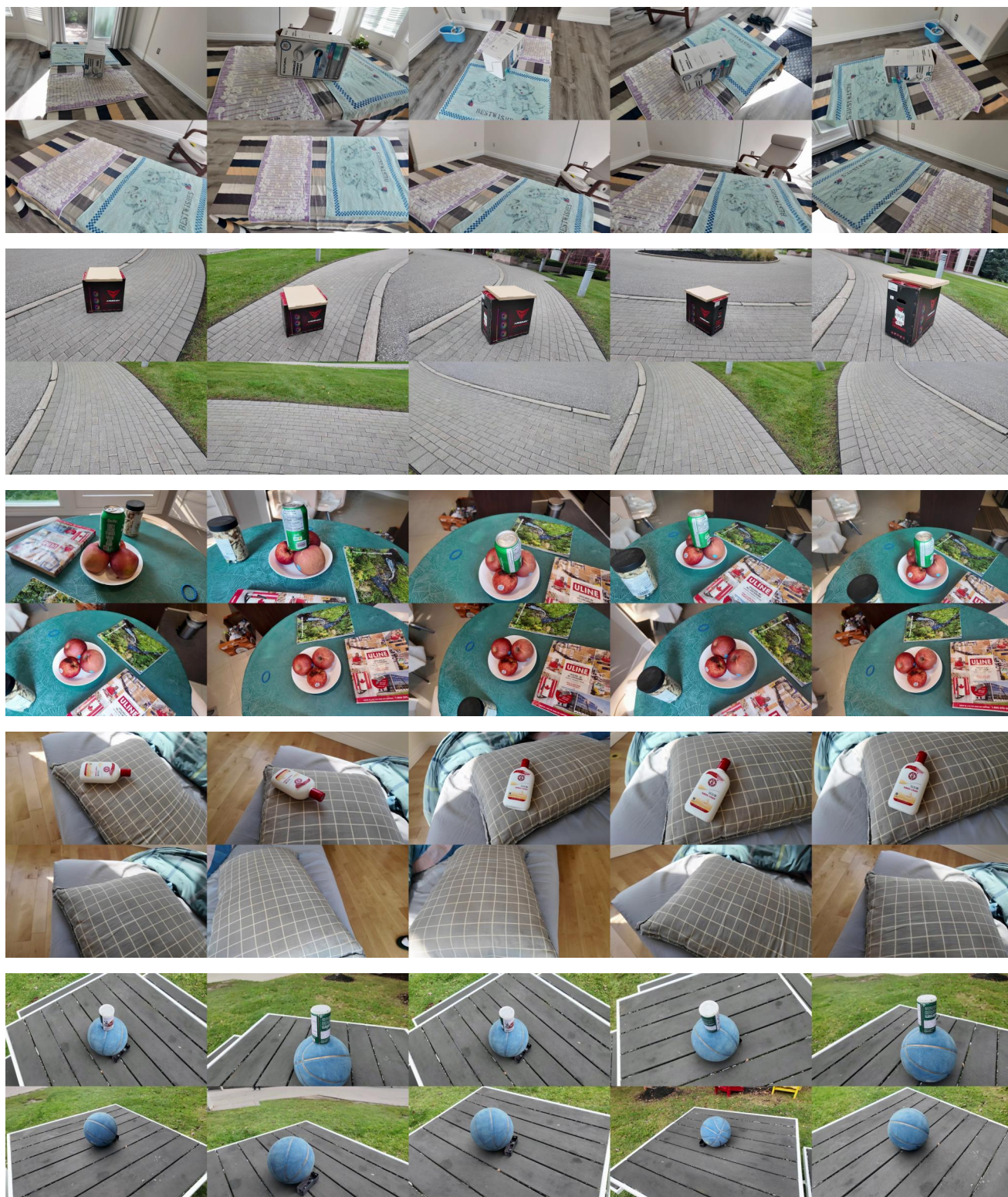


Figure 12. Dataset visualization - 2.



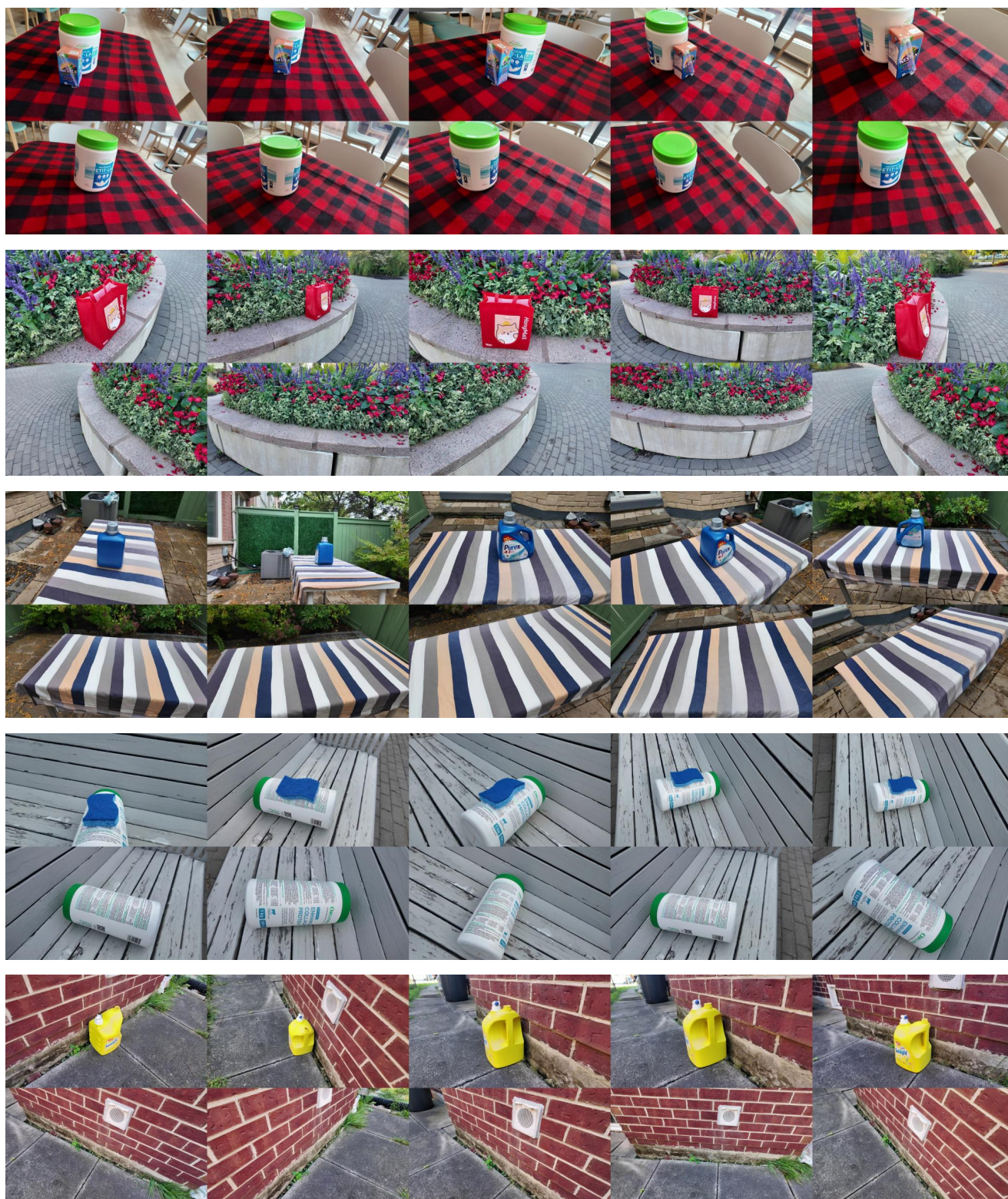


Figure 13. Dataset visualization - 3.





Figure 14. Dataset visualization - 4.