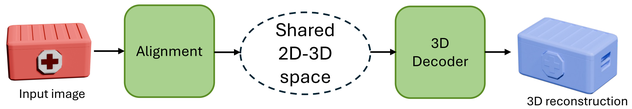
LAM3D: Large Image-Point-Cloud Alignment Model for 3D Reconstruction from Single Image

Ruikai Cui [1] Xibin Song [2,✉] Weixuan Sun [2] Senbo Wang [2] Weizhe Liu [2] Shenzhou Chen [2] Taizhang Shang [2] Yang Li [2] Nick Barnes[1] Hongdong Li [1] Pan Ji [2]

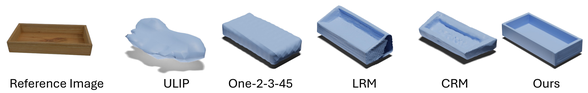
[1]Australian National University [2]Tencent XR Vision Labs

Introduction

In this worl, we achieve high-fidelity 3D reconstruction by aligning single-view images to a shared 2D-3D latent space, then decoding them into 3D shapes, addresses camera distortions caused by lack of explicit 3D prior.



(a) Illustration of the inference pipeline



(b) Result of different training scheme

Motivation

- **Image-Point-Cloud Alignment** Recent large-scale reconstruction models rely on volumetric rendering for supervision, which lacks explicit 3D information and leads to geometric distortions. We propose an image-point-cloud alignment model to address these distortions and improve reconstruction quality.
- **Independent Diffusion Processes** Treating the tri-plane as three independent feature maps, we introduce three separate diffusion networks to align them with 3D point cloud features, enhancing feature alignment accuracy and effectiveness.
- **A Novel Reconstruction Pipeline** Our diffusion-based module aligns 3D point cloud features and 2D image features into a unified latent space, enabling high-fidelity 3D reconstruction.

Contribution

Our main contributions are summarized below:

- An effective model for 3D mesh reconstruction from a single image using point cloud priors. LAM3D enables precise tri-plane feature transformation, significantly enhancing 3D mesh quality and accuracy.
- An independent diffusion processes that transmit image features to each respective tri-plane for accurate 3D mesh reconstruction.
- Extensive experiments demonstrate that our method achieves state-of-the-art results across multiple 3D shape reconstruction benchmarks.

Image Point Cloud Alignment

Our method contains two training stage. **Stage 1**: we train an encoder-decoder structure to compress point clouds to a latent tri-plane representation; **Stage 2**: we employ diffusion to align image modality to latent tri-planes obtained in stage 1. The diffusion step takes an initial noise and an image feature from a freezed DINO feature encoder and progressively align the image feature to the latent tri-plane. **Inference**: To reconstruct a 3D mesh from a single-view image, we use the alignment step, following the decoder (Plane Decoder, Plane Refiner) from the compression step, to predict a tri-plane. Then, we can use marching cubes to extract 3D meshes from the reconstructed tri-plane.

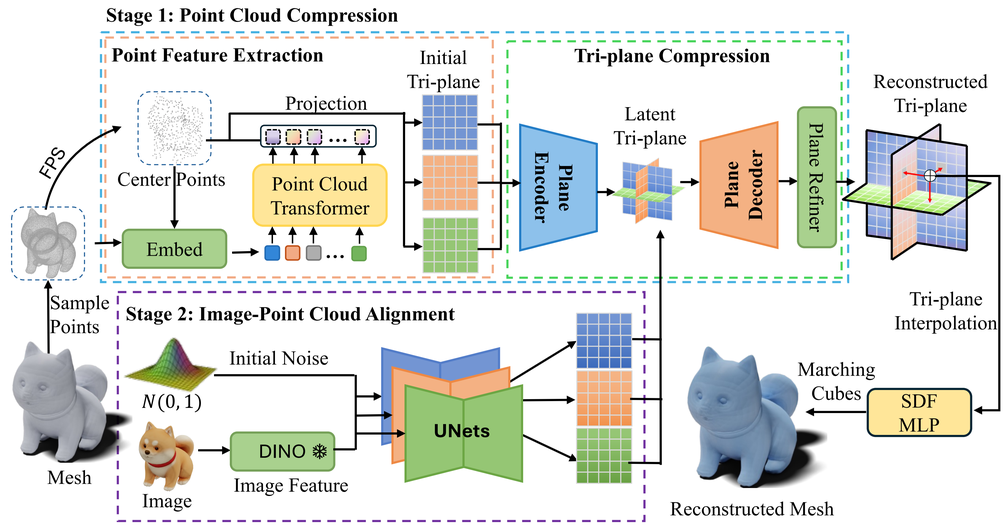


Figure 1. The Pipeline of *LAM3D*.

Loss Functions

Stage 1: Compression

$$\mathcal{L}_{comp} = \mathcal{L}_{sdf} + \mathcal{L}_{normal} + \mathcal{L}_{lsdf} + \mathcal{L}_{KL}$$

where

$$\mathcal{L}_{sdf} = \lambda_1 \sum_{p\in\Omega_0} ||\Phi_P(p)|| + \lambda_2 \sum_{p\in\Omega} ||\Phi_P(p) - d_p||$$

$$\mathcal{L}_{normal} = \lambda_3 \sum_{p\in\Omega_0} ||\nabla_p \Phi_P(p) - n_p||$$

$$\mathcal{L}_{lsdf} = \lambda_4 \sum_{p\in\Omega_0} ||\Phi_t(p)|| + \lambda_5 \sum_{p\in\Omega} ||\Phi_t(p) - d_p||.$$

Stage 2: Alignment

$$\mathcal{L}_{align} = ||\Psi(z_t, z_{img}, \gamma(t)) - z_0||^2$$

Experiments

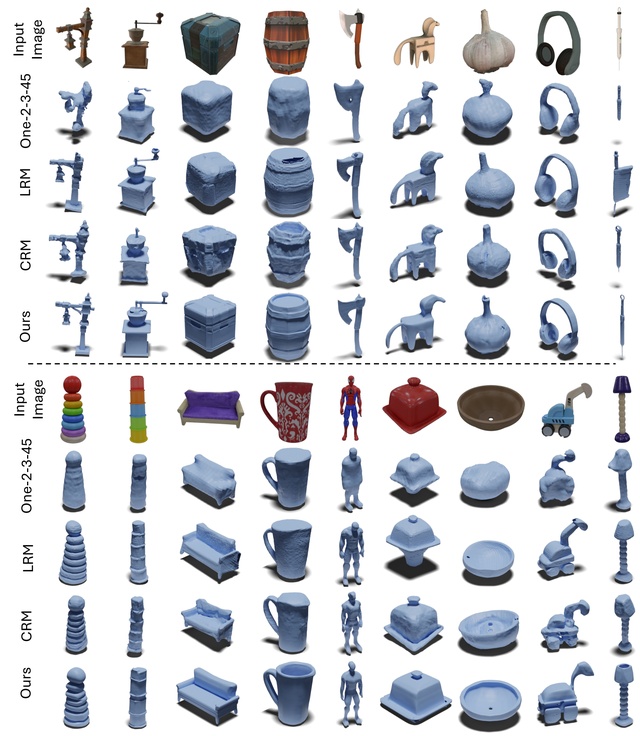


Figure 2. Rendered images of shapes reconstructed by various methods from single images. The upper samples are from Objaverse and the lowers are from Google Scanned Objects.

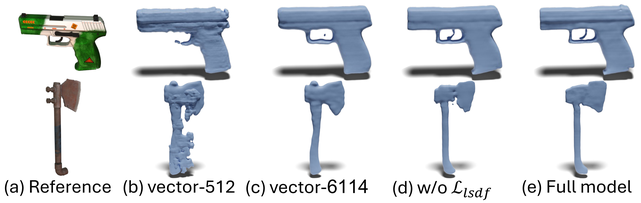


Figure 3. Qualitative comparisons of different latent representations.