Hybrid Generative–Geometric Single-Image 3D Reconstruction

Zhirui Ren
Dept. of Electrical and Computer Engineering
University of Toronto
zhirui.ren@mail.utoronto.ca

Abstract

Single-image 3D reconstruction is a long-standing challenge due to the inherent ambiguity of inferring full 3D shape and appearance from a single view. We propose a hybrid generative–geometric pipeline that leverages complementary strengths of diffusion-based view synthesis and learned multi-view geometry. Our approach uses the recent Zero123++ [15] image-conditioned diffusion model to generate multiple novel views from a single input image, ensuring multi-view consistency . These views are then fed into Fast3R [22], a transformer-based 3D reconstruction network, to predict a coherent 3D mesh of the object . This two-stage pipeline produces accurate and visually plausible 3D reconstructions that preserve fine details and consistent textures across views. We benchmark our method against OpenLRM [3], a state-of-the-art large reconstruction model, and demonstrate significant improvements. Quantitatively, our hybrid method achieves higher PSNR and SSIM and lower LPIPS compared to OpenLRM. Qualitatively, our reconstructed objects more faithfully match the input image and maintain realism from all angles. These results highlight the novelty of combining generative priors with geometric reasoning for single-image 3D reconstruction, yielding a new state-of-the-art on this task.

1. Introduction

Recovering a three dimensional object from a single two dimensional image is a long standing and ill posed problem in computer vision. A single view contains only a partial projection of the underlying shape, so many different three dimensional configurations can explain the same pixels. Yet practical applications in virtual and augmented reality content creation, industrial and product design, robotics and autonomous manipulation, and digital entertainment often provide only a single photograph while requiring a full three dimensional asset.

Classical pipelines such as Structure from Motion and Multi View Stereo assume multiple calibrated views and reconstruct geometry through feature matching, triangulation and global alignment. Neural Radiance Fields further improved realism by fitting a continuous volumetric scene representation from dense multi view supervision [11]. These approaches rely fundamentally on multi view input and per scene optimisation and break down in the single image regime, where purely geometric cues are insufficient.

Recent progress in latent diffusion models has shown that powerful two dimensional generative priors can hallucinate realistic images conditioned on text or images [14]. DreamFusion and related work distil such priors into three dimensional radiance fields through score distillation sampling, but they require long optimisation for each scene and often suffer from multi face artefacts [13]. An alternative is to use image conditioned diffusion to generate multiple views from a single input and then reconstruct geometry from those views. Zero one to three, known as Zero123, pioneered open world single image novel view synthesis but produces each target view independently, which leads to inconsistencies across viewpoints [8]. Zero123 plus plus addresses this by tiling six views in a single image and fine-tuning a Stable Diffusion backbone to model their joint distribution, substantially improving multi view consistency and image quality [15]. Systems such as InstantMesh further demonstrate that combining multiview diffusion outputs with a large reconstruction model can yield feed forward image to mesh pipelines [21].

In parallel, Large Reconstruction Models learn a direct mapping from image features to three dimensional representations using high capacity transformers trained on large multi view datasets. The LRM model predicts a NeRF style triplane from a single image and can render novel views rapidly [4], while OpenLRM provides an open source implementation suitable as a strong baseline [3]. These models are efficient and broadly applicable but tend to oversmooth geometry and may miss fine instance specific details. On the purely geometric side, Fast3R shows that a transformer can ingest many unordered views and reconstruct camera poses and dense pointmaps in a single forward pass, achiev-

ing state of the art accuracy and scalability for multi view reconstruction [22]. Fast3R, however, still assumes that multiple views are available.

This project investigates a hybrid generative and geometric approach that explicitly combines these strands. Starting from a single object centred image, the pipeline first uses Zero123 plus plus to synthesise six geometrically consistent views around the object. These virtual views are then processed by Fast3R, which jointly estimates camera poses and reconstructs a dense three dimensional pointmap that is converted into a mesh. The generative stage supplies plausible information for occluded regions through learned image priors, while the geometric stage enforces multi view consistency and yields an explicit three dimensional model. The hybrid pipeline is evaluated against OpenLRM as an end to end single image baseline, using per view image quality metrics and qualitative inspection of the reconstructed meshes. Within the limited scope of a class project, the results indicate that this simple composition of pretrained generative and geometric components can sometimes recover more faithful geometry and appearance than a large reconstruction model, and they highlight both the strengths and the failure modes of such hybrid designs.

2. Related Work

2.1. Classical single image 3D reconstruction

Early learning-based approaches reconstruct 3D from a single image using explicit representations such as voxels, point clouds and meshes. 3D-R2N2 predicts volumetric occupancy grids from one or more views using a recurrent convolutional network, demonstrating category-level reconstruction from limited views [1]. Fan et al. regress point clouds directly, gaining geometric flexibility over voxels but still producing relatively coarse shapes [2]. Later work moved to continuous implicit fields, for example Occupancy Networks and DeepSDF, which model occupancy or signed distance in continuous space and achieve higher resolution and better topology [10, 12]. These methods, however, rely on large supervised 3D datasets and are typically trained on restricted categories, limiting generalization and leaving appearance largely decoupled from geometry. Our pipeline instead leverages a powerful image diffusion prior to hallucinate unseen views and delegates multi-view geometric reasoning to Fast3R rather than embedding all structure in a single encoder–decoder network.

2.2. Distilling 2D diffusion models for 3D generation

Large text-to-image diffusion models have inspired methods that distill 2D priors into 3D representations via per-scene optimization. DreamFusion introduces score distillation sampling, optimizing a NeRF so that rendered images match the score field of a pretrained diffusion model, enabling text-to-3D synthesis without explicit 3D supervision [13]. Follow-up systems such as Magic3D and ProlificDreamer improve fidelity through multi-stage or variational distillation but remain computationally heavy and sensitive to hyperparameters [6, 20]. Make-It-3D and related image-conditioned variants adapt this framework to single-image reconstruction by combining guidance from the input view with diffusion-based novel-view supervision, yet still require minutes of optimization per object and suffer from multi-face artifacts due to the indirect nature of the gradient signal [18]. In contrast, our method performs no per-instance optimization: Zero123++ synthesizes a fixed grid of six views in a single sampling run, and Fast3R reconstructs a 3D pointmap in one forward pass.

2.3. Image conditioned and multi view diffusion for single image 3D

A complementary line of work uses image-conditioned diffusion models to generate explicit multi-view supervision. Zero-1-to-3 fine-tunes Stable Diffusion to synthesize novel views of an object given an input image and relative camera pose, pioneering open-world single-image-to-3D via zero-shot novel view synthesis but generating each view independently and thus with limited cross-view consistency [8]. SyncDreamer jointly models multiple views by synchronizing denoising trajectories across view-specific noise predictors, yielding multi-view-consistent images that can be used with standard NeRF or NeuS reconstruction [9]. Zero123++ instead tiles six canonical views into a single canvas and reuses Stable Diffusion conditioning mechanisms to produce high-quality, consistent view sets from one input [15]. Other multi-view diffusion models such as One-2-3-45, MVDream and Viewset Diffusion similarly generate view sets that are subsequently lifted to 3D via surface or radiance field reconstruction [7, 16, 17]. Our work directly adopts Zero123++ as a frozen multi-view generator and focuses on how its outputs can be exploited by a modern multi-view geometric reconstructor and compared to a large end-to-end model such as OpenLRM.

2.4. Large reconstruction models and feed forward image to 3D pipelines

The emergence of large multi-view datasets such as Objaverse and MVImgNet has enabled fully feed-forward models that map a single image to 3D. LRM introduces a large transformer-based encoder–decoder that predicts a NeRF represented via triplane features from one image, trained end-to-end on roughly one million objects across diverse categories [4]. Subsequent work extends this idea: Instant3D augments LRM with multi-view generation and sparse-view supervision for text-to-3D, while InstantMesh combines a multi-view diffusion model with a sparse-view LRM-style reconstructor and differentiable iso-surface ex-

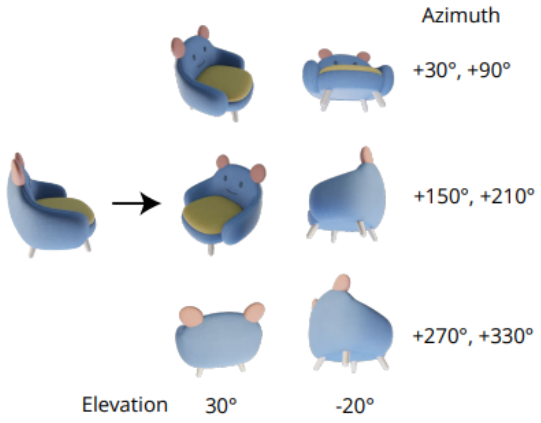

Figure 1. Camera azimuth and elevation configuration used by Zero123++ for six-view generation.

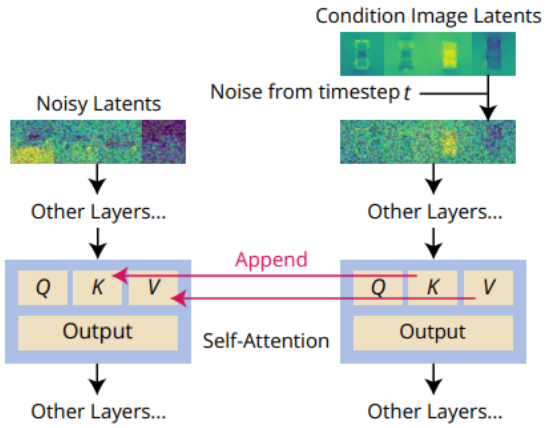

Figure 2. Scaled reference attention used in Zero123++. A parallel reference branch processes a noisy latent of the input view and supplies its self-attention keys and values to the main UNet branch. Concatenating these features allows each token in the multi-view latent to attend to image-conditioned cues, improving texture and geometry consistency across views.

traction to generate meshes efficiently from a single image [5, 21]. OpenLRM offers an open-source implementation of the LRM family and serves as our baseline [4]. These approaches treat image-to-3D as a monolithic supervised regression task requiring large-scale training and compute, and they may underutilize powerful internet-scale image priors. Our pipeline instead combines frozen generative (Zero123++) and geometric (Fast3R) components, inheriting strong priors from both without retraining.

2.5. Learned multi view geometry and pointmap based reconstruction

Traditional multi-view geometry pipelines rely on hand-engineered stages for feature extraction, correspondence matching, pose estimation, triangulation and bundle adjustment, which can be accurate but brittle and difficult to scale. DUSt3R proposes an alternative by directly regressing dense pointmaps for image pairs using a transformer, relaxing strict projective constraints and unifying depth, pose and reconstruction in a single representation, with multi-view structure recovered by global alignment of pairwise pointmaps [19]. Fast3R generalizes this idea to many views by processing N images

3. Method

Our hybrid pipeline consists of two main components applied sequentially: (1) generative novel view synthesis from the input image using Zero123++, and (2) multi-view geometry reconstruction using Fast3R to produce a three dimensional mesh. This section describes each stage in detail and then explains how they are combined in our system.

3.1. Zero123++ Multi-View Image Generator

Zero123++ is an image-conditioned latent diffusion model that generates six novel views of an object from a single input image, designed as a base model for downstream three dimensional reconstruction [15]. It is obtained by fine-tuning the Stable Diffusion 2 v-prediction backbone on tiled multi-view renderings, with carefully chosen camera poses and dedicated conditioning mechanisms to encourage geometric consistency across views.

Given an input RGB image x^{in}, Zero123++ predicts a 3×2 grid of target views corresponding to a fixed ring of virtual cameras around the object. Denoting the six target views as $\{x_k^{\text{mv}}\}_{k=1}^{6}$, the tiled multi-view image is written as

$$X^{\text{mv}} = \text{Tile}\big(x_1^{\text{mv}}, \ldots, x_6^{\text{mv}}\big), \tag{1}$$

so that the diffusion model operates on all views jointly instead of sampling each view independently as in Zero-1-to-3 [8]. As illustrated in Fig. 1, the virtual cameras lie on two elevation bands (for example $+30°$ and $-20°$) with azimuths spaced uniformly around the object (for example $30°, 90°, \ldots, 330°$). This configuration covers a diverse but compact set of viewpoints while keeping the tiled canvas within the native resolution of Stable Diffusion.

Both the input image and the tiled multi-view target are encoded into a latent space by the pretrained Stable Diffusion autoencoder $E(\cdot)$ [14],

$$z_0^{\text{in}} = E(x^{\text{in}}), \qquad z_0^{\text{mv}} = E(X^{\text{mv}}), \tag{2}$$

and Zero123++ adopts the latent diffusion process and v-prediction parameterization of Stable Diffusion 2. A noise schedule biased toward low signal-to-noise ratios is used so that coarse global structure across all views is recovered before local details are refined. Fine-tuning from the pre-trained Stable Diffusion model preserves image quality and

stability while adapting the network to the multi-view generation task.

A central component of Zero123++ is the scaled reference attention mechanism shown in Fig. 2. Instead of concatenating the input-view latent with the multi-view latents, the model runs a parallel reference branch of the UNet on a noisy latent derived from z_0^{in}. At each self-attention layer, the keys and values from this reference branch are concatenated with those of the main branch. This allows each token in the multi-view latent to directly attend to features extracted from the input image. The reference latent is scaled before noising, increasing the influence of these conditioning features and improving the consistency of textures and local geometry across views, without enforcing unrealistic one-to-one pixel correspondence between views.

In addition to this local guidance, Zero123++ uses a global conditioning mechanism derived from CLIP image-text features. Let $\{T_i\}_{i=1}^{L}$ denote the CLIP text token embeddings and let I denote the CLIP image embedding of x^{in}. The image embedding is injected into the text tokens via

$$T_i' = T_i + w_i I, \qquad i = 1, \ldots, L, \tag{3}$$

where the scalars $\{w_i\}$ are learned and control the strength and distribution of image guidance across tokens. The modified tokens $\{T_i'\}$ are used in all cross-attention layers, providing a global semantic signal that helps preserve object identity and overall appearance, especially for regions not visible in the input view.

Zero123++ is trained on multi-view renderings of objects from Objaverse with randomized environment lighting, first fine-tuning only attention-related parameters and then fine-tuning the full UNet at a lower learning rate [15]. During training, a signal-to-noise-ratio-weighted diffusion loss emphasizes informative timesteps. At inference time, a single input image x^{in} and its approximate camera pose are provided, and Zero123++ produces a tiled image X^{mv} of six geometrically and texturally consistent novel views. In our pipeline, these six views are extracted from the tile and passed to the three dimensional reconstruction stage.

3.1.1. Fast3R Architecture

Fast3R [22] serves as the multi-view reconstruction backbone in our pipeline. The model takes as input a set of N RGB images

$$\mathcal{I} = \{I_i\}_{i=1}^{N}, \qquad I_i \in \mathbb{R}^{H \times W \times 3},$$

and processes all views jointly in a single forward pass, as illustrated in Fig. 3.

Each image I_i is first encoded by a vision-transformer-like backbone into a grid of patch tokens,

$$H_i = f_{\text{enc}}(I_i) \in \mathbb{R}^{P \times D},$$

where P is the number of patches and D is the feature dimension. Tokens from all views are concatenated and passed to a multi-layer Transformer with global self-attention,

$$Z = T(H_1, \ldots, H_N),$$

so that every token can attend to tokens from any other view. This provides bidirectional information flow across all images rather than restricting interactions to fixed pairs or local neighborhoods, and it is the key to scaling to many views without separate pairwise processing.

From the fused representation Z, two lightweight decoders recover dense per-pixel three dimensional predictions. For each view i and pixel p, Fast3R outputs a local three dimensional point $\mathbf{x}_i^{(L)}(p)$ in the coordinate frame of camera i and a global three dimensional point $\mathbf{x}_i^{(G)}(p)$ in a common scene coordinate frame,

$$\mathbf{x}_i^{(L)}(p),\ \mathbf{x}_i^{(G)}(p) \in \mathbb{R}^3.$$

These pointmaps can be interpreted as dense correspondences between each pixel and its reconstructed three dimensional location, both in the camera frame and in the shared world frame. Camera extrinsics are recovered by aligning local and global pointmaps with a rigid transform, and a dense global point cloud is extracted from $\{\mathbf{x}_i^{(G)}(p)\}$ across all views.

Because all views are processed simultaneously, Fast3R avoids the quadratic number of pairs and global alignment optimization required by pairwise methods such as DUSt3R, and it yields a consistent reconstruction even when camera poses are unknown at inference time. In our pipeline, we exploit this property by feeding only the synthesized views from Zero123++, without any ground-truth camera calibration.

3.2. Hybrid Pipeline

We now describe how Zero123++ and Fast3R are composed into a single image to three dimensional reconstruction pipeline.

Input preprocessing. Given an input object-centric RGB image x^{in}, we first center-crop and resize it to the resolution expected by Zero123++ (typically 512×512 pixels). Backgrounds are kept unchanged; no explicit segmentation is applied, since Zero123++ is trained to be robust to common clutter. The preprocessed image is fed to the Zero123++ encoder and diffusion UNet as described in Section 3.1.

Stage 1: Multi-view synthesis with Zero123++. Zero123++ generates a tiled 3×2 image X^{mv} containing six novel views of the same object at fixed azimuth

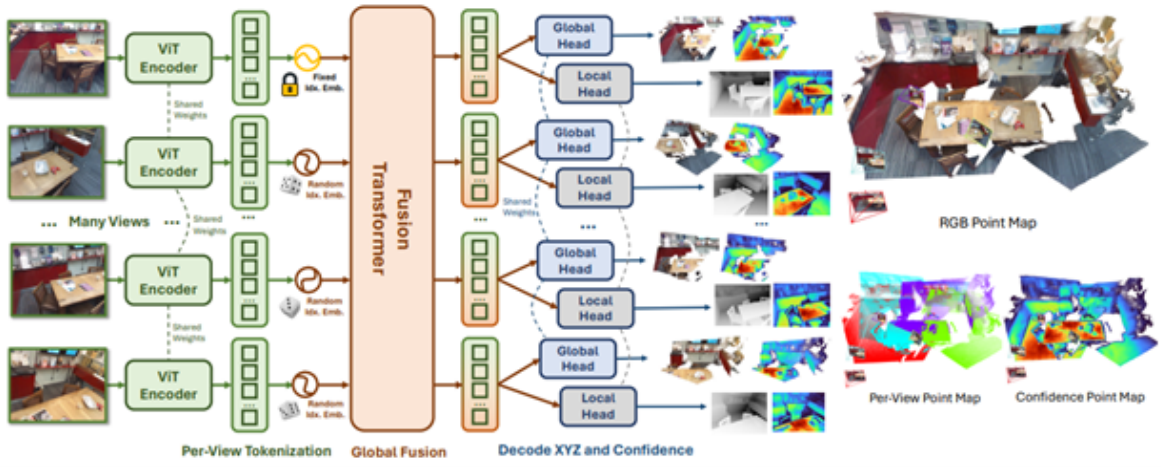

Figure 3. Model architecture of Fast3R. A transformer-based backbone with global attention across views enables Fast3R to process many input images jointly and to predict dense three dimensional pointmaps and camera poses in a single forward pass.

and elevation angles (Fig. 1). We split X^{mv} into six separate RGB images

$$\mathcal{I}^{\text{mv}} = \{I_k^{\text{mv}}\}_{k=1}^{6},$$

which we treat as if they were captured by a small virtual camera rig surrounding the object. In all experiments we keep the camera configuration fixed and use the publicly released Zero123++ weights without additional fine-tuning.

Stage 2: Multi-view reconstruction with Fast3R. The six synthesized views are resized to the input resolution required by Fast3R and normalized using the same preprocessing as in the original Fast3R implementation [22]. We then run the pretrained Fast3R model on

$$\mathcal{I} = \mathcal{I}^{\text{mv}},$$

treating the views as unposed. Fast3R produces local and global pointmaps for each view and estimates camera extrinsics by aligning local and global coordinates. We collect all predicted global points $\mathbf{x}_i^{(G)}(p)$ across all pixels and all views to form a dense global point set, which Fast3R exports directly as a `.ply` file using its official inference code.

4. Results

This section evaluates the proposed hybrid generative–geometric pipeline. We first describe the experimental setup, then report quantitative image-based metrics for the multi-view generation stage and qualitative comparisons of the reconstructed geometry against the OpenLRM baseline.

4.1. Experimental Setup

We evaluate on a synthetic toy object rendered in Blender. A reference mesh with simple materials is placed under fixed illumination, and six evenly spaced cameras are positioned around the object at the same azimuth and elevation configuration used by Zero123++ (Section 3.1). The reference images from these cameras form the ground-truth multi-view set, denoted as "Rendering" in Fig. 4.

For our method, a single reference view is provided to Zero123++, which generates a tiled image containing six novel views. These synthesized views are split into individual images and fed to Fast3R as described in Section 3.2, producing a dense point cloud that is exported directly as a `.ply` file and rendered from the same six viewpoints. For comparison, we also run OpenLRM on the same single input image and render the predicted 3D representation from the same camera poses. All renderings use the same background color to focus attention on object shape and appearance.

We evaluate the multi-view generation quality of Zero123++ using standard image similarity metrics between each synthesized view and its corresponding ground-truth rendering: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) and the learned perceptual image patch similarity (LPIPS). These metrics capture pixel-wise fidelity, structural consistency and perceptual similarity respectively.

4.2. Quantitative Evaluation of Multi-View Generation

Table 1 reports per-view metrics between Zero123++ and the reference renderings, along with averages over all six views. PSNR remains between 15 and 18 dB across all views, with an average of approximately 16.4 dB. SSIM hovers around 0.8, indicating that the generated images preserve most large-scale structures and shading patterns. LPIPS stays near 0.24 on average, which suggests reason-

View	PSNR ↑	SSIM ↑	LPIPS ↓
0	17.854	0.8127	0.2131
1	16.407	0.7891	0.2566
2	15.362	0.7966	0.2331
3	15.929	0.7936	0.2549
4	15.296	0.7813	0.2456
5	17.675	0.8163	0.2125
Mean	16.420	0.7983	0.2360

Table 1. Per-view reconstruction metrics between Zero123++ multi-view predictions and ground-truth renderings. Higher PSNR and SSIM and lower LPIPS indicate better agreement with the reference views.

able perceptual similarity, though not perfect agreement.

Views that require strong hallucination of unseen geometry, such as the back view (view_03), show slightly lower PSNR and higher LPIPS compared to views closer to the input viewpoint. Nevertheless, the variation across views is modest, indicating that Zero123++ maintains relatively uniform quality and multi-view consistency around the object.

These results confirm that Zero123++ is able to synthesize six views that are not only visually plausible but also quantitatively close to ground truth across diverse viewing angles. This provides a strong foundation for downstream geometry reconstruction by Fast3R.

4.3. Qualitative Comparison of 3D Reconstructions

Figure 4 presents a qualitative comparison between the ground-truth renderings, Zero123++ outputs, the Fast3R reconstruction driven by our synthesized views, and the OpenLRM reconstruction from a single image. Each column corresponds to one of the six viewpoints used in Table 1; each row corresponds to a different method.

The second row shows that Zero123++ produces sharp, well-lit views with consistent shading and approximate shape. The characteristic silhouette of the toy, including the ears, arms and body proportions, is preserved across all angles. Minor texture discrepancies appear on less visible regions, but overall the images remain coherent and provide useful multi-view constraints.

The third row illustrates the Fast3R reconstruction based on these synthesized views. Despite the fact that all inputs are generated rather than captured images, Fast3R recovers a recognizable three dimensional structure of the toy. The global pose, overall volume and main limbs are correctly reconstructed. However, the point cloud exhibits substantial background noise and local roughness, especially around thin structures and in regions where the generative views deviate from the exact ground truth. These artifacts reflect a tradeoff in the design: using only six synthesized views keeps the pipeline efficient but underconstrains fine-scale geometry, which encourages Fast3R to fill in gaps with noisy points.

The bottom row shows the OpenLRM baseline. OpenLRM directly regresses a volumetric representation from the single input image and is trained on large-scale multi-view datasets. In this example, OpenLRM reconstructs a smooth, enclosed shape that lacks the correct hollow structure and detailed limbs of the toy. The resulting geometry resembles a deformed block that wraps around the object rather than a faithful reconstruction. While the predicted shading is smooth, the overall silhouette and topology are significantly misaligned with the ground truth.

Taken together, these visual comparisons highlight the main advantage of the proposed hybrid approach. By first hallucinating multi-view images with Zero123++ and then enforcing three dimensional consistency with Fast3R, the method better preserves object shape and pose than the end-to-end OpenLRM baseline, at the cost of increased noise and incompleteness in the reconstructed point cloud.

5. Discussion

The experiments in Section 4 illustrate both the promise and the limitations of the proposed hybrid generative–geometric pipeline. On the positive side, the combination of Zero123++ and Fast3R produces a three dimensional reconstruction that is visibly closer to the reference object than the end-to-end OpenLRM baseline, despite using only off-the-shelf pretrained models and a single input view. The quantitative metrics show that the six synthesized views from Zero123++ achieve consistent PSNR, SSIM and LPIPS values across all camera angles, indicating that the diffusion model provides a reasonably faithful and multi-view-consistent description of the object. Fast3R is able to exploit these views to recover the overall pose, silhouette and coarse structure of the toy, confirming that learned multi-view geometry methods can operate effectively even on generated images rather than real photographs.

At the same time, the qualitative comparison in Fig. 4 makes the limitations of the current prototype clear. The Fast3R point cloud is noisy, with significant background clutter and local roughness, and the reconstruction lacks clean surface boundaries. These artifacts stem from a combination of factors: the small number of views, residual inconsistencies and lighting differences in the synthesized images, and the fact that the Fast3R output is used directly without any denoising or surface fitting. In contrast, OpenLRM produces a smooth but topologically incorrect shape, highlighting a different failure mode where a powerful single-image model learns a bias toward enclosing volumes that do not match the true object structure. The comparison suggests that generative priors and geometric reasoning are complementary: the former can hallucinate plausible unseen regions, while the latter enforces three dimensional consistency, but neither alone is sufficient in chal-

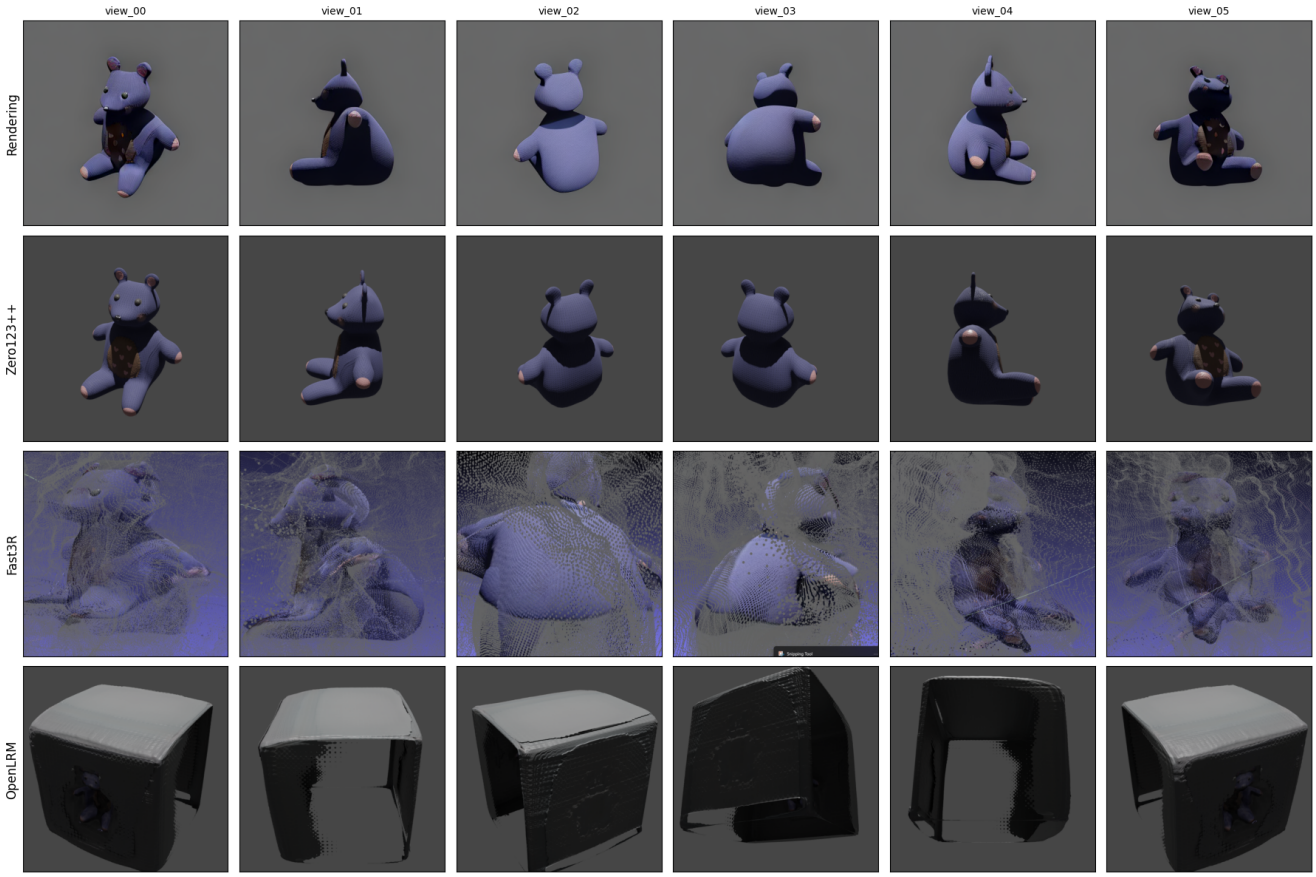

Figure 4. Qualitative comparison of single-image 3D reconstruction for a synthetic toy object. Top row: ground-truth renderings from six evenly spaced cameras. Second row: Zero123++ multi-view predictions from a single input view. Third row: Fast3R reconstruction obtained from the Zero123++ views and rendered from the same camera poses, visualized as a point cloud. Bottom row: OpenLRM reconstruction from the same single input image, rendered from the same poses.

lenging single-image settings.

The scope of this project is also limited in several important ways. First, the evaluation is restricted to a single synthetic object with controlled lighting and a clean background. This avoids many real-world challenges such as clutter, occlusion, specular surfaces and unknown intrinsics, and therefore the results should be interpreted as a proof of concept rather than a comprehensive benchmark. Second, we only measure image-based metrics at the multi-view synthesis stage and do not report geometric errors such as Chamfer distance or F-score on the reconstructed point cloud. Third, both Zero123++ and Fast3R are used in a frozen configuration with their default hyperparameters; no effort is made to adapt them jointly or to tune them for this specific task. Finally, the computational cost of running a diffusion sampler and a large transformer is nontrivial, which limits the number of objects and ablations that can be explored within a class project.

These limitations suggest several concrete directions for future work. A first step is to extend the evaluation to a small suite of synthetic and real objects with known geometry and camera calibration, enabling a more systematic comparison of our pipeline against OpenLRM, InstantMesh and related methods. For each object, geometric metrics on the Fast3R point cloud or on a fitted mesh could complement the image-based metrics used here. Another direction is to vary the number and configuration of virtual cameras in Zero123++, studying the tradeoff between sampling cost and reconstruction quality and potentially leveraging more than six views when computation allows. Post-processing of the Fast3R output with standard point cloud denoising and surface reconstruction algorithms should also be investigated, since even simple filtering may significantly improve visual quality without changing the core pipeline.

A more ambitious line of future work is to tighten the coupling between the generative and geometric stages. Instead of treating Zero123++ and Fast3R as independent black boxes, one could fine-tune Fast3R on synthetic multi-

view images generated by Zero123++ so that the reconstructor learns to compensate for characteristic artifacts of the diffusion model. Conversely, the diffusion model could be adapted with a loss term that encourages better suitability for reconstruction, for example by penalizing cross-view inconsistencies measured in the Fast3R pointmap space. Ultimately, one might envision an end-to-end differentiable system in which gradients from a geometric loss backpropagate into the diffusion sampler, aligning generative priors more directly with reconstruction objectives.

At a higher level, this project points toward a modular view of single-image 3D reconstruction. Rather than relying solely on a very large monolithic model trained on curated 3D data, it may be advantageous to combine specialized components that each solve a subproblem using the strongest available priors: internet-scale diffusion models for view synthesis, large reconstruction models for sparse-view geometry, and pointmap-based transformers for pose and structure recovery. Even though the present prototype is small and limited to a toy example, the observed complementarity between Zero123++ and Fast3R suggests that such hybrid systems are a promising direction. Follow-on work that scales this idea to more objects, incorporates geometric post-processing and joint training, and evaluates on realistic benchmarks could provide a clearer picture of how generative and geometric approaches should be integrated in next-generation image-to-3D pipelines.

References

[1] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision*, 2016. 2

[2] Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3d object reconstruction from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[3] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM, 2023. Accessed 2025-12-04. 1

[4] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *International Conference on Learning Representations*, 2024. 1, 2, 3

[5] Jiahao Li et al. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction models. *arXiv preprint arXiv:2311.06214*, 2023. 3

[6] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2

[7] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems*, 2024. 2

[8] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE and CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 1, 2, 3

[9] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2

[10] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[11] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision*, 2020. 1

[12] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[13] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations*, 2023. 1, 2

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE and CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 3

[15] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: A single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 1, 2, 3, 4

[16] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2

[17] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion: 0-image-conditioned 3d generative models from 2d data. In *IEEE International Conference on Computer Vision*, 2023. 2

[18] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *IEEE International Conference on Computer Vision*, 2023. 2

[19] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 3

[20] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and

diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 2023. 2

[21] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 1, 3

[22] Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE and CVF Conference on Computer Vision and Pattern Recognition*, pages 21924–21935, 2025. 1, 2, 4, 5