Hybrid Generative-Geometric Single-Image 3D Reconstruction

Zhirui Ren

University of Toronto

Motivation

- 3D reconstruction aims to recover the shape and appearance of real-world objects or scenes from images. It underpins many applications: robotic navigation and manipulation, AR/VR content creation, cultural heritage digitization, and product visualization, where a faithful 3D model enables new viewpoints, measurements, and realistic interaction.
- Classical methods assume multiple overlapping views of the same object. Structure-from-motion, multi-view stereo, and modern neural methods like NeRF or 3D Gaussian Splatting all rely on a set of calibrated images with diverse viewpoints. From these, they jointly estimate camera poses and a 3D representation that reprojects consistently into all views.
- In many practical cases, however, we only have a single image—such as a product shot or a social media photo—so dense multi-view capture is not available. This motivates single-image 3D reconstruction: given one image and its camera, infer a plausible 3D shape that both explains the input and produces realistic novel views. Because infinitely many shapes can match the same picture, this inverse problem is severely ill-posed and must rely on strong learned priors over object geometry and appearance. Recent advances in generative AI, especially large diffusion and flow-based models trained on massive image–text and 3D datasets, provide exactly such priors: they can hallucinate consistent novel views or directly predict 3D structure from a single photograph. This project explores how these generative priors, combined with geometric reconstruction tools, can turn one image into a usable 3D asset.

Related Work

Classical 3D Reconstruction:

- Traditional pipelines recover geometry from multiple overlapping views using structure-from-motion and multi-view stereo, followed by meshing and refinement. Neural methods such as NeRF [1] and 3D Gaussian Splatting [2] extend this idea by fitting a continuous radiance or Gaussian field that reprojects consistently into many calibrated images, but they still require dense multi-view supervision.

Single-Image 3D Prediction:

- Earlier deep-learning approaches map one image to voxel grids, point clouds, or meshes using encoder–decoder architectures. These models learn category-level priors but often struggle with fine details and out-of-distribution shapes.

Generative Priors for 3D:

- Recent diffusion- and flow-based models use massive 2D/3D datasets to hallucinate geometry or views. Text-to-3D and image-to-3D methods (e.g., LRM [3]) directly optimize or predict a 3D representation from a single image or prompt. View-conditioned diffusion models like Zero123 / Zero123++ [4] generate novel viewpoints from a single image, providing synthetic multi-view data for downstream reconstruction.

References

[1] Mildenhall, Srinivasan, Tancik, Barron, Ramamoorthi, Ng, *NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis*, ECCV, 2020.
[2] Kerbl, Kopanas, Leimkühler, Drettakis, *3D Gaussian Splatting for Real-Time Radiance Field Rendering*, ACM Transactions on Graphics (SIGGRAPH), 2023.
[3] Xie et al., *LRM-Zero: Training Large Reconstruction Models with Synthesized Data*, NeurIPS, 2024.
[4] Shi et al., *Zero123++: a Single Image to Consistent Multi-view Diffusion Base Model*, arXiv preprint, 2023.
[5] Yang, Sax, Liang, Henaff, Tang, Cao, Chai, Meier, Feiszli, *Fast3R: Towards 3D Reconstruction of 1000+ Images in One Forward Pass*, CVPR, 2025.

Method

- Finetunes Stable Diffusion v-prediction for tiled six-view generation with a linear noise schedule and global CLIP-image conditioning. Reference Attention runs a parallel denoising branch on the input view, then injects its self-attention keys/values—scaled latents—into target views, enforcing local texture and geometry consistency.

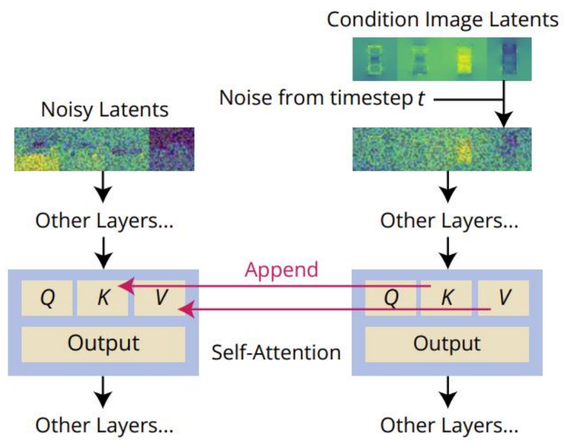


Figure 1. Zero123++ reference attention [4]

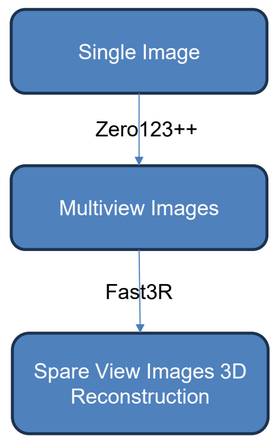


Figure 2. Flow chart

Fast3R extends DUSt3R to many views using a transformer that encodes each image into tokens, then applies global reference attention across all views, letting every pixel attend to others to infer consistent 3D pointmaps. It predicts local and global pointmaps in one pass, avoiding pairwise reconstruction and expensive alignment overhead.

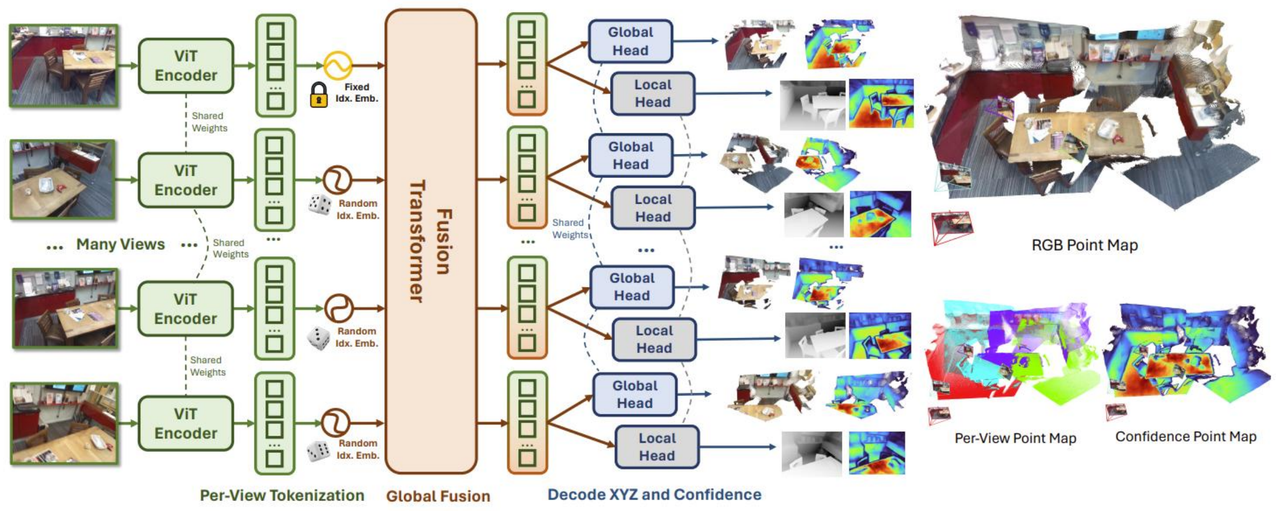
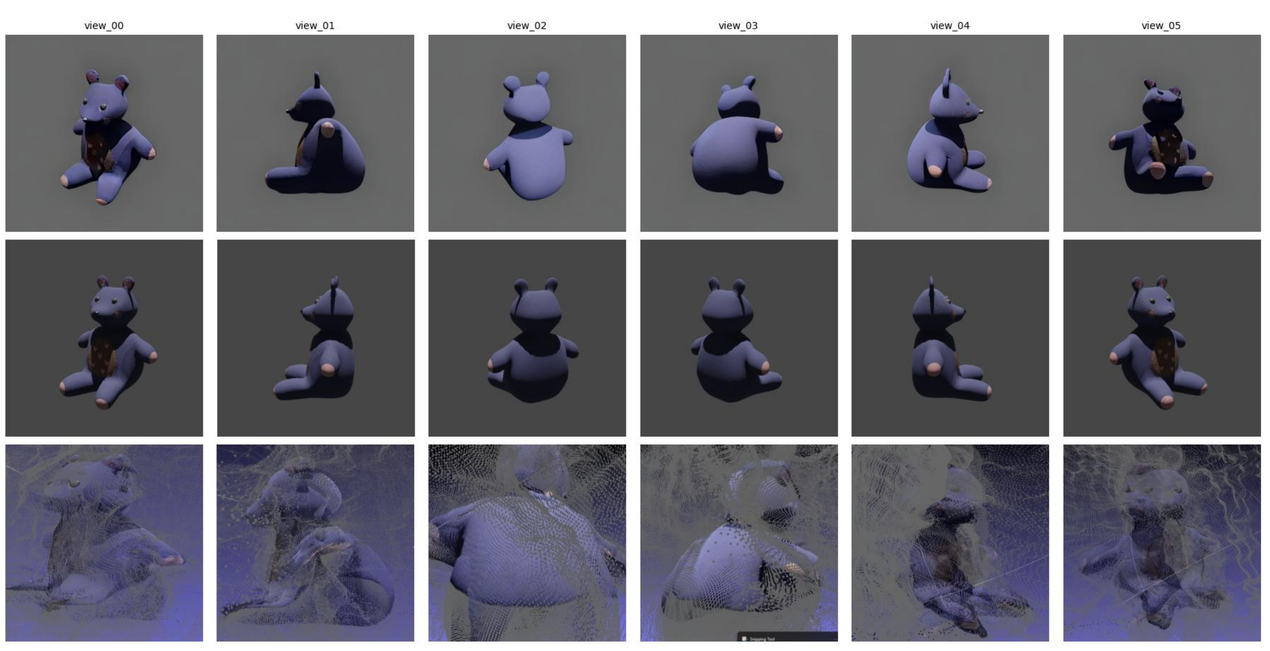


Figure 3. Fast3R model architecture [5]

Experimental Results

- The six novel views generated by Zero123++ are visually very consistent with each other: the object silhouette and overall geometry remain stable across viewpoints, and the appearance is smooth and clean, although the exact camera pose is slightly misaligned compared to the ground-truth renders. Quantitatively, the per-view PSNR is around 16 dB and fairly uniform across all views, while the mean SSIM is ≈0.80 and the mean LPIPS is ≈0.236, indicating moderate pixel-level error but reasonably good structural and perceptual similarity.
- On the reconstruction side, the Fast3R mesh broadly matches the object's shape but appears rough and irregular, with noisy grey "blobs" around the object that likely come from background geometry or hallucinated structures in the generated views. In contrast, the LRM mesh is smooth but noticeably "sunken", with missing volume and distorted interior regions, suggesting that a pure single-image prior struggles to recover full, watertight geometry in this example.



Comparison result between rendered views, generated multi-views and generated dense model

```
Per-view metrics (Zero123 vs Rendered):
 view      PSNR      SSIM      LPIPS
    0    17.654    0.8137    0.2175
    1    16.771    0.8090    0.2432
    2    15.228    0.7928    0.2282
    3    16.099    0.8062    0.2472
    4    16.597    0.8108    0.2332
    5    16.896    0.7888    0.2491

Averages over all views:
  PSNR  mean: 16.540848814167585
  SSIM  mean: 0.80352646
  LPIPS mean: 0.23638228823741278
```

Generated view evaluation

LRM result