Infer3D: Inverse Generative Modeling Enables Robust 3D Reconstruction

Amani Kiruga[1] **Mark Hamilton**[2] **Joshua B. Tenenbaum**[3] **William T. Freeman**[3]
Vincent Sitzmann[3] **Yilun Du**[1,†] **Ayush Tewari**[4,†]

[1]*Harvard University* [2]*Google DeepMind*
[3]*Massachusetts Institute of Technology* [4]*University of Cambridge*

Abstract

Humans can understand the 3D structure of scenes even from highly unfamiliar or extreme viewpoints. Cognitive science suggests this generalization stems from a top-down generative inference strategy. In contrast, state-of-the-art 3D reconstruction neural networks typically rely on amortized, feedforward inference. While fast, this pattern-matching approach degrades when faced with out-of-distribution (OOD) scene conditions unseen during training. Inspired by the generative nature of human vision, we propose Infer3D, an inverse generative modeling framework that recovers the latent variables of a generative model to best match the monocular input. To prevent degenerate flat outputs, we couple the generative image prior with a frozen 3D lifting network that constrains the optimization space to a valid, multi-view-consistent 3D manifold. Across a variety of controlled synthetic and real-world tasks, Infer3D demonstrates superior OOD reconstruction accuracy compared to feedforward networks trained on equal or larger datasets. Furthermore, mirroring how human perception requires more time to process confusing stimuli, we introduce an adaptive inference strategy that expends more test-time compute on ambiguous OOD examples while efficiently routing in-distribution inputs.

1 Introduction

The human visual system exhibits a remarkable ability to parse the 3D world from 2D projections. Crucially, this capability is not brittle: humans can readily understand the 3D structure of objects and scenes even when viewed from highly unconventional angles (Yildirim et al., 2015), under strange lighting (Adelson, 2000), or under different compositional arrangements (Lake et al., 2017). A central challenge in perception theory is explaining how human vision can be so robust. A rich body of literature in cognitive science and developmental psychology suggests that this robustness arises because human vision is not merely a passive, bottom-up filter but also includes an active, top-down generative process. By maintaining an internal mental model of the physical world, the mind can continuously predict and match sensory inputs, allowing it to reliably resolve ambiguities in unseen environments (Spagna et al., 2024; Clark, 2013; Friston, 2010; Tenenbaum et al., 2011).

This cognitive perspective heavily influenced classical computational theories, which formalized perception as "vision as inverse graphics" (Yuille & Kersten, 2006). In this paradigm, visual understanding is cast as Bayesian inference over a generative model of image formation and scene structure (Knill & Richards, 1996; Kersten et al., 2004). Rather than simply recognizing 2D patterns, the visual system performs analysis-by-synthesis, actively searching for the underlying 3D scene representation that best explains the observed 2D image.

In contrast, modern monocular 3D reconstruction systems have largely abandoned this generative top-down inference in favor of amortized, feedforward architectures (Szymanowicz et al., 2024; Yu et al., 2021; Charatan et al., 2024; Wewer et al., 2024; Hong et al., 2023) that directly map a 2D input image to a 3D output.

[†]Equal advising.

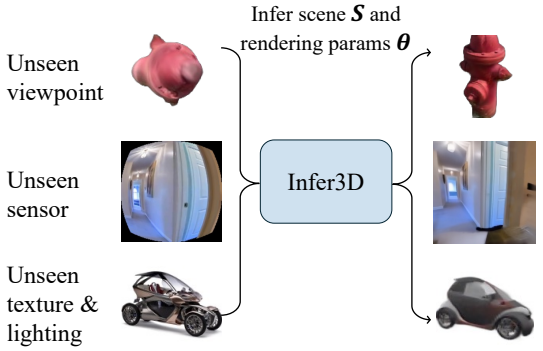


Figure 1: **Robust 3D reconstruction under unseen conditions.** Out-of-distribution (OOD) inputs categorized by domain shifts: unseen camera viewpoints (top), unseen sensors (middle), and real-world texture and lighting (bottom). Infer3D infers the underlying scene ($\boldsymbol{S}$) and rendering parameters ($\boldsymbol{\theta}$) despite these challenging OOD conditions.

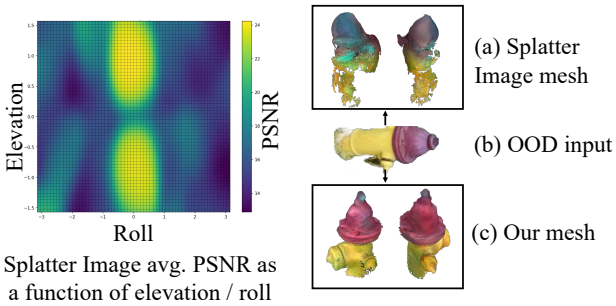


Figure 2: **Failure under Viewpoint Shift.** *Left:* Heatmap of Splatter Image (Szymanowicz et al., 2024) PSNR evaluated on input at different elevations (y-axis) and rolls (x-axis). Performance collapses rapidly in regions of OOD elevation or roll. *Right:* Comparison of extracted meshes (Marching Cubes) for an OOD view; the baseline exhibits strong artifacts, whereas our method recovers coherent 3D structure.

While this pattern-matching approach enables very fast inference, it fundamentally lacks the robustness of generative inference. As we demonstrate in this paper, feedforward networks degrade in unpredictable ways when faced with out-of-distribution (OOD) scene conditions unseen during training, such as unconventional camera viewpoints (see Fig. 2).

Inspired by the generative nature of human vision and classic analysis-by-synthesis methods (Blanz & Vetter, 2023a; Yuille & Kersten, 2006), we propose Infer3D. Rather than trusting a feedforward encoder on OOD data, Infer3D formulates perception as an inverse generative problem. We search for the scene geometry $\mathcal{S}$ and rendering parameters $\boldsymbol{\theta}$ that maximize the explanatory likelihood of the input observation. To ensure our optimization remains constrained to plausible, multi-view-consistent 3D structures, we couple a 2D generative image prior with a frozen 3D lifting network. This encourages the optimized latent variables to always decode to a valid representation on the training manifold, effectively synthesizing an internal hypothesis that best matches the confusing sensory input. Furthermore, mirroring how human perception requires more time to process ambiguous stimuli (Kar et al., 2019; Tang et al., 2018; Tschantz et al., 2023), we introduce an adaptive inference strategy. Because our generative framework leverages an encoder for initialization, we show that the initial reconstruction error provides a very reliable and efficient signal for OOD detection. This allows Infer3D to dynamically route standard inputs through fast feedforward passes while reserving expensive test-time optimization exclusively for OOD images.

Overall, our work makes the following contributions: (1) we introduce Infer3D, a modular analysis-by-synthesis framework that combines a learned 2D generative model with a feedforward 3D lifter to enable multi-view-consistent, OOD-robust inference from single images; (2) we propose an adaptive inference and OOD detection strategy that leverages the initial reconstruction error of the generative prior, requiring no external classifiers; and (3) through a controlled experimental setup, we demonstrate that Infer3D successfully generalizes to unseen viewpoints, non-linear lens distortions, and unobserved textures where state-of-the-art feedforward methods collapse (see Fig. 1).

2 Prior Work

Feedforward Inference. There has been a lot of success in models that can reconstruct 3D scenes from monocular images. Popular choices of 3D representation are meshes (Liu et al., 2019; Tewari et al., 2017; Chen et al., 2025), or neural radiance fields (Yu et al., 2021; Tewari et al., 2023; Szymanowicz et al., 2023; Hong et al., 2023; Xu et al., 2023b). Modern methods rely more on 3D Gaussian representations that also use volume rendering, but are significantly more efficient than neural radiance fields (Charatan et al., 2024; Szymanowicz et al., 2024; Xu et al., 2024; Tang et al., 2024). In addition, point-based representations

are also emerging as a popular choice (Wang et al., 2024; 2025a). These approaches use neural networks to directly transform the input image to the output 3D reconstruction and thus struggle to reason about out-of-distribution settings.

Several papers demonstrate lack of generalization with feedforward inference in out-of-distribution settings (Zhao et al., 2024; Madan et al., 2020; Alcorn et al., 2019; Abbas & Deny, 2023). Mitigating such failures is also a topic of active research. Several works have introduced equivariant neural network models that aim to generalize in OOD conditions. However, none of these approaches is as general as our method, e.g., we show that our method can be used for images captured under OOD camera poses as well as under OOD textures. In particular, equivariant approaches either assume 3D or multiview inputs (Xu et al., 2022; Deng et al., 2021; Esteves et al., 2018), or only include equivariance in parts of their network, e.g., equivariant intermediate representations without equivariant transformations between images and the corresponding representations (Dupont et al., 2020; Mitchel et al., 2024). Equivariant models that take 2D image inputs currently only guarantee equivariance under 2D or color transformations (Cohen & Welling, 2016; Yang et al., 2024). There is also a line of work in using 3D models for data augmentation (Zhou et al., 2023; Yang et al., 2025; Zhu et al., 2024). However, no such method has been proposed for monocular 3D reconstruction.

Analysis-by-Synthesis. Rather than directly transforming an image into the corresponding 3D output with feedforward inference, analysis-by-synthesis solves an optimization problem for inference that searches for the optimal 3D solution that can explain the input observation. Because of the ambiguities in the problem, a good prior over 3D models is necessary. Many approaches use 3D training data, or pre-captured object geometries, to learn these priors (Blanz & Vetter, 2023b; Yuille & Kersten, 2006; Müller et al., 2023; Gothoskar et al., 2021). Recent papers instead rely on learning priors from 2D data, e.g., using a monocular image generative model as a prior over renderings of the 3D reconstruction (Poole et al., 2023; Xu et al., 2023a; Melas-Kyriazi et al., 2023; Wang et al., 2023). However, monocular priors are not sufficient to enforce multi-view consistency in output. Other approaches use multi-view priors (Lin et al., 2023; Xu et al., 2024; Long et al., 2024; Szymanowicz et al., 2025; Gao et al., 2024); however, these methods do not generalize to OOD images as they process input images with neural networks. The Bayesian approach of Anciukevicius et al. (2022) is somewhat related to us, but it makes strong assumptions by assuming known scene parameters for the OOD inputs. Recent work on inverse generative modeling for scene understanding (Wang et al., 2025b) is also closely related in spirit: it formulates perception as inverting a learned generative model and introduces compositional structure to handle complex 2D scenes. We are inspired by this view of inference as solving an inverse generative problem, but focus on multi-view-consistent 3D reconstruction and explicit control over camera parameters, rather than 2D compositional layouts.

Hybrid Inference. Hybrid inference methods have been explored in the literature (Tewari et al., 2018; Alldieck et al., 2019; Leung et al., 2022). However, they assume that the initial prediction is a reliable initialization—an assumption that breaks down under OOD conditions. In contrast, we propose an adaptive inference strategy that explicitly detects OOD inputs and selectively invokes robust optimization-based inference only when needed.

Optimization-Based Generative Inversion. pi-GAN (Chan et al., 2021) fits the latent conditioning of a 3D-aware GAN to an image but does not infer an unknown camera. BRFI (Pavllo et al., 2023) and 3D GAN Inversion (Ko et al., 2023) reconstruct 3D shape by jointly optimizing a generative latent code and camera pose. Both refine a single initialization obtained from learned pose or latent predictors, making them sensitive to initialization errors under distribution shift. FINV (Sun et al., 2024) filters multiple hypotheses using provided camera poses and additional observations, whereas Infer3D searches multiple joint latent and pose hypotheses from a single unposed image. These methods couple appearance and geometry within a monolithic 3D GAN. Infer3D instead searches a powerful 2D generative prior for appearance while a separately trained lifter supplies the 3D constraint, allowing the prior and lifter to be selected independently. We directly compare against BRFI and 3D GAN Inversion in Tables 5 and 10.

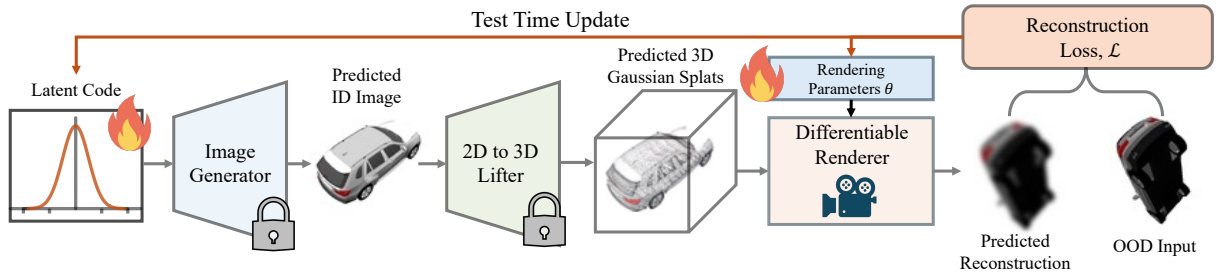


Figure 3: **Analysis-by-synthesis inference paradigm.** To handle OOD inputs, we employ a test-time optimization loop (red arrow). We freeze the **Image Generator** and **2D to 3D Lifter** (indicated by lock icons) to preserve the learned prior. Instead, we optimize the **Latent Code** and **Rendering Parameters** (indicated by flame icons) to minimize the reconstruction loss $\mathcal{L}$ between the rendered reconstruction and the OOD input. This ensures the input to the lifter remains a valid predicted ID image while matching the OOD observation.

3 Method

We provide an overview of our method in Fig. 3. We first describe the 3D scene representation and our general analysis-by-synthesis framework in Section 3.1. We then instantiate this framework across three distinct out-of-distribution (OOD) applications in Section 3.3: unseen camera poses, unseen sensor distortions, and unseen appearance. Finally, we discuss efficient OOD detection in Section 3.5.

3.1 Scene Representation and Differentiable Renderer

The 3D reconstructions are represented using a scene representation $\mathcal{S}$ which describes the underlying geometry and appearance of objects in the scene. This representation can be rendered into an image using a differentiable forward model $\mathcal{R}(\mathcal{S}, \boldsymbol{\theta})$, with rendering parameters $\boldsymbol{\theta}$. Depending on the scenario, $\boldsymbol{\theta}$ is parameterized to represent unknown extrinsic camera poses (viewpoint shifts) or expanded to include uncalibrated intrinsic sensor characteristics (e.g., non-linear fisheye lens distortions).

3D Gaussians. In practice, we use 3D Gaussians (Kerbl et al., 2023) as our 3D representation, where the scene is described using a collection of 3D Gaussians that represent its geometry and appearance. Each Gaussian is described with parameters that determine its shape, opacity, and color. We render the 3D Gaussians using the differentiable splatting-based renderer from 3DGS (Kerbl et al., 2023), which takes camera parameters as additional input.

3.2 Inference via Generative Decoder Inversion

Standard feedforward reconstruction networks function as amortized encoders, mapping observations directly to 3D representations. While efficient, these encoder-based methods lack the inductive biases required for generalization to out-of-distribution (OOD) scene conditions (Brady et al., 2025). To achieve robustness, we formulate perception as an inverse generative problem via maximum a posteriori (MAP) estimation. Given an OOD observation $\mathbf{I}_{\text{OOD}}$, we seek the scene parameters $\mathcal{S}$ and rendering parameters $\boldsymbol{\theta}$ that maximize the posterior:

$$p(\mathcal{S}, \boldsymbol{\theta} \mid \mathbf{I}_{\text{OOD}}) \propto p(\mathbf{I}_{\text{OOD}} \mid \mathcal{S}, \boldsymbol{\theta}) p(\mathcal{S}) p(\boldsymbol{\theta}). \tag{1}$$

Directly modeling a 3D prior $p(\mathcal{S})$ is intractable. Instead, we leverage the manifold of a 2D generative model $G(\mathbf{z})$ trained on in-distribution images, composed with a deterministic 2D-to-3D lifter $\Phi : \mathcal{I} \to \mathcal{S}$. This implicitly defines a structural prior where valid scenes are those generated by lifting valid in-distribution images: $\mathcal{S} = \Phi(G(\mathbf{z}))$, where $\mathbf{z}$ are samples from the latent space of the generative model. By substituting this parameterization into Eq. 1, the inference reduces to a search over the latent code $\mathbf{z}$ and parameters $\boldsymbol{\theta}$:

$$\hat{\mathbf{z}}, \hat{\boldsymbol{\theta}} = \arg\max_{\mathbf{z}, \boldsymbol{\theta}} \Big[\underbrace{\log p(\mathbf{I}_{\text{OOD}} \mid \mathcal{S}, \boldsymbol{\theta})}_{\text{Likelihood}} + \underbrace{\log p(\mathbf{z})}_{\text{latent prior}} + \underbrace{\log p(\boldsymbol{\theta})}_{\text{rendering prior}} \Big] \tag{2}$$

Algorithm 1 Analysis-by-Synthesis via Generative Inversion

Input: OOD Image $\mathbf{I}$, Generator G, Lifter Φ, Renderer $\mathcal{R}$
Input: Schedule $\sigma(t) \to (k_t, \lambda_t)$, Learning rate η
▷ *Initialize latent and camera populations*
$\mathcal{Z} \leftarrow \{\mathbf{z}_r \sim \mathcal{N}(\mathbf{0}, \mathbf{I})\}_{r=1}^{R}$
$\Theta \leftarrow \{\theta_r \sim \mathcal{U}(\mathrm{SO}(3))\}_{r=1}^{R}$
if Encoder E is available **then**
  ▷ *Warm-start the first particle*
  $\mathbf{z}_1 \leftarrow E(\mathbf{I}), \quad \theta_1 \leftarrow \mathbf{I}_{3\times 3}$
end if
$\bar{\mathbf{z}}_{\text{init}} \leftarrow \frac{1}{|\mathcal{Z}|} \sum \mathbf{z}_r$ ▷ *Compute regularization anchor*
for $t = 1, \ldots, N$ **do**
  $k_t, \lambda_t \leftarrow \sigma(t)$
  ▷ *Render hypotheses and compute loss for all particles*
  $\hat{\mathbf{I}}_r \leftarrow \mathcal{R}\big(\Phi(G(\mathbf{z}_r)), \theta_r\big)$
  $\mathcal{L}_r \leftarrow \mathcal{L}_{\text{recon}}(\hat{\mathbf{I}}_r, \mathbf{I}) + \lambda_t \|\mathbf{z}_r - \bar{\mathbf{z}}_{\text{init}}\|^2$
  ▷ *Gradient update*
  $(\mathbf{z}_r, \theta_r) \leftarrow (\mathbf{z}_r, \theta_r) - \eta \nabla \mathcal{L}_r$
  ▷ *Prune worst-performing particles*
  Keep the top k_t particles in $(\mathcal{Z}, \Theta)$ with the lowest $\mathcal{L}_r$
end for
return Best scene $\mathcal{S}^\star = \Phi(G(\mathbf{z}^\star))$ and pose $\theta^\star$

In practice, the (unscaled) negative log likelihood is implemented as $\mathcal{L}\Big(\mathcal{R}\big(\Phi(G(\mathbf{z})), \boldsymbol{\theta}\big),\ \mathbf{I}_{\text{OOD}}\Big)$, where $\mathcal{R}$ is the differentiable renderer from 3DGS and $\mathcal{L}$ is a reconstruction loss. We approximate the prior term $-\log p(\mathbf{z})$ as a quadratic penalty $\|\mathbf{z} - \bar{\mathbf{z}}_{\text{init}}\|_2^2$ centered on the mean of the initial latent population, serving to discourage the optimization from drifting too far from the valid manifold initialization. The final term defines the prior over rendering parameters, described further in Sec. 3.3.

Optimization Dynamics. We solve Eq. 2 via gradient descent. To avoid local minima in the non-convex energy landscape, we employ a multi-start strategy with particle filtering. We initialize a population of latent codes $\{\mathbf{z}_i\}$ and parameters $\{\boldsymbol{\theta}_i\}$. The population is iteratively refined, with high-error particles culled periodically, yielding a set of plausible 3D interpretations for the ambiguous OOD input. This procedure is detailed in Algorithm 1.

The formulation in Eq. 2 ensures that the input to the lifting network Φ is always an in-distribution image $G(\mathbf{z})$, strictly adhering to the manifold where the feedforward model is valid. It is also highly flexible: we can adapt to distinct domain shifts merely by redefining the optimized rendering parameters $\boldsymbol{\theta}$, their associated prior $p(\boldsymbol{\theta})$, and the likelihood. We instantiate this approach for three specific OOD applications below.

3.3 Adapting to Out-of-Distribution Shifts

Out-of-Distribution Camera Poses. For unseen camera viewpoint shifts, we jointly optimize the scene representation alongside the extrinsic camera parameters. Specifically, we optimize the latent code $\mathbf{z}$ and the camera pose $\boldsymbol{\theta} \in \mathrm{SE}(3)$. We use a uniform prior over the camera pose parameters, continuously projecting rotation matrices onto the SO(3) manifold via SVD (Levinson et al., 2020) to ensure stable gradient descent. The likelihood is computed as a combination of MSE and LPIPS losses (Zhang et al., 2018) to align low-frequency geometry and high-frequency textures.

Out-of-Distribution Sensor Shifts. Standard feedforward models assume strict pinhole projection and inherently fail under severe intrinsic parameter shifts, misinterpreting non-linear fisheye lens warping as curved scene geometry. To address this, we optimize the latent code $\mathbf{z}$, camera intrinsics, i.e., focal lengths

and principal points (f_x, f_y, c_x, c_y), and complex non-linear distortion parameters $\boldsymbol{\phi}$. The distortion prior is parameterized by a Hybrid Neural Distortion Field (Deng et al., 2025), which is a learnable polynomial fisheye model coupled with an iResNet on a sparse 2D grid. To evaluate the likelihood term, we first warp the perspective-rendered hypothesis generated by our generative model into the fisheye space using the learned forward distortion field ($D_{\boldsymbol{\phi}}$). We then compute MSE and LPIPS losses between this distorted rendering and the input OOD fisheye observation, effectively combining 3D Gaussian Splatting with a hybrid fisheye forward model.

Out-of-Distribution Appearance. When transitioning from synthetic training data to real-world captures, standard pixel-level reconstruction collapses due to large domain gaps caused by complex reflections, unobserved lighting, and novel materials. We can bridge this synthetic-to-real domain gap by leveraging the inherent modularity of our inverse framework. We replace the pixel-level likelihood in the previous cases with a deep semantic feature loss, maximizing the cosine similarity of DINOv2 (Oquab et al., 2023) patch-level features extracted from both the OOD input and the rendered 3D hypothesis. As semantic features lack fine-grained spatial precision, we couple this with an explicit MSE depth consistency loss to constrain the geometric scale, all rendered via standard pinhole 3D Gaussian Splatting renderer. Similar to the OOD camera poses setting, we search for the latent code $\mathbf{z}$ and camera pose $\boldsymbol{\theta} \in \mathrm{SE}(3)$.

3.4 Lifting and Generative Networks

3D Lifting Network (Φ). Our framework is agnostic to the specific architecture of the network Φ. In our experiments, we instantiate Φ using either:

- Splatter Image (Szymanowicz et al., 2024), which predicts pixel-aligned 3D Gaussians using a U-Net architecture.
- CATSplat (Roh et al., 2024), which utilizes transformer-based context aggregation for more complex geometries.

Crucially, Φ is frozen during inference; we rely on the optimization of $\mathbf{z}$ to adapt the geometry.

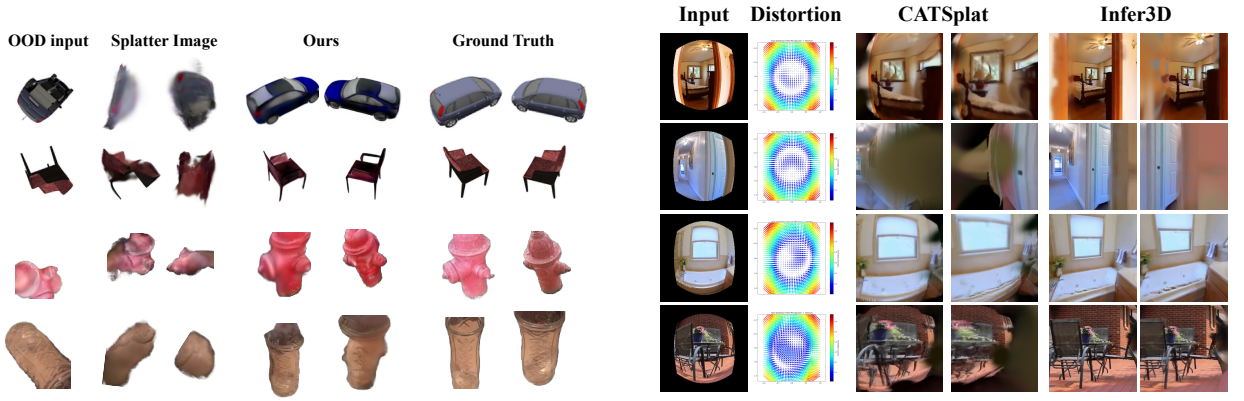
Generative Image Priors (G). We explore two classes of generative models to parameterize the image manifold, see appendix Fig. 8. The choice of G dictates the structure of the latent space $\mathcal{Z}$:

- GAN-based: We employ StyleGAN2 (Karras et al., 2020). The latent space is defined as $\mathbf{z} \in \mathcal{W}$, where $\mathcal{W}$ refers to the intermediate 1D representation in the StyleGAN architecture. This generator provides a compact, smooth manifold that is amenable to gradient-based optimization.
- Diffusion-based: We also utilize Diffusion Autoencoders (DiffAE) (Preechakul et al., 2022). DiffAE encodes an image into a semantic latent $\mathbf{z}_{sem}$ and a stochastic latent $\mathbf{x}_T$. This offers a distinct advantage: unlike GANs, DiffAE includes an explicit encoder E that can initialize $\mathbf{z} = E(\mathbf{I})$ for in-distribution data, significantly accelerating convergence (see Sec. 3.5).

3.5 OOD Detection via Generative Prior

A key capability of our method is that it naturally exposes signals for out-of-distribution (OOD) detection, without requiring the training of additional classifier models.

Inference dynamics as an OOD signal. Given an input image, analysis-by-synthesis optimizes latent variables and rendering parameters to minimize a reconstruction loss. For in-distribution inputs $x \in \mathcal{X}_{\text{ID}}$, the generative encoder provides an initialization that lies close to an optimal solution. Consequently, the initial reconstruction error of the generated hypothesis is inherently low. In contrast, OOD inputs $x \in \mathcal{X}_{\text{OOD}}$ lie completely outside the support of the learned image distribution. The encoder fails to provide a near-optimal initialization, resulting in a severe structural mismatch and a high initial reconstruction loss.



(a) (b)

Figure 4: **Robustness to geometric and sensor domain shifts. (a) Viewpoint shift.** Columns show the OOD input, Splatter Image baseline, Infer3D (Ours), and ground truth. Unlike the baseline, our method recovers view-consistent geometry without collapsing or hallucinating structure. **(b) Sensor shift (fisheye distortion).** Left to right: (1) OOD fisheye input, (2) our learned displacement field, (3) CATSplat baseline (incorrectly modeling distortion as curved geometry), and (4) our disentangled 3D geometry.

We leverage this asymmetry to build an efficient OOD detector. Specifically, we run one forward pass through our method and use the initial reconstruction error (MSE and LPIPS) against the input image as our classification score. If this initial loss exceeds a calibrated threshold, the sample is flagged as OOD and routed to the full optimization loop (Algorithm 1); otherwise, we bypass the optimization and trust the feedforward lifter.

Automatic Threshold Calibration. Importantly, this detection signal requires no additional trained classifiers. To set the decision threshold, we perform an automatic calibration step on a small held-out validation set (5 ID and 5 OOD objects), selecting the threshold that maximizes the F1 score. As we batch all objects and run a single encode-decode pass, this entire one-time calibration process is fast, taking a maximum of 500 milliseconds on CO3D. At test time, we can accurately detect OOD inputs in roughly 286 milliseconds without constructing expensive backpropagation graphs. We quantitatively evaluate our fast OOD detection and calibration on CO3D and RE10K (Zhou et al., 2018) in Section 4.3.

4 Experiments

We evaluate Infer3D's capabilities across three core dimensions: (1) generalization to geometric and camera-level domain shifts, (2) robustness to appearance and environmental domain gaps, and (3) efficacy of our adaptive inference strategy. We report Peak Signal-to-Noise Ratio (PSNR, ↑), Structural Similarity Index Measure (SSIM, ↑), and Learned Perceptual Image Patch Similarity (LPIPS, ↓) (Zhang et al., 2018) for novel-view fidelity, Chamfer Distance (CD, ↓) for 3D geometry error, and the Area Under the Receiver Operating Characteristic Curve (AUROC, ↑) and F1 score (↑) for OOD detection.

4.1 Geometric and Sensor Shift Generalization

We first evaluate how well our approach recovers consistent 3D geometry under extreme geometric perturbations, specifically out-of-distribution camera viewpoints and non-linear lens distortions.

Viewpoint Robustness. We evaluate on Multi-Category ShapeNet (Chang et al., 2015) (trained on a fixed $\pm 15^\circ$ elevation ring) and CO3D (Reizenstein et al., 2021) (Hydrants and Vases), see appendix Fig. 9. We test under random $SO(3)$ rotations and $SE(3)$ translations that severely deviate from the training manifold. We compare against the state-of-the-art feedforward models Splatter Image (Szymanowicz et al., 2024),

Table 1: **CO3D Evaluation (Hydrants and Vases).** Metrics: CD ↓, PSNR ↑, SSIM ↑, LPIPS ↓. **Bold** = best.

		In-distribution				OOD SE(3)			
Dataset	**Method**	**CD**	**PSNR**	**SSIM**	**LPIPS**	**CD**	**PSNR**	**SSIM**	**LPIPS**
Hydrants	Splatter Image	**0.052**	**21.48**	**0.785**	**0.168**	0.733	15.73	0.665	0.266
	Viewset Diffusion	0.145	20.95	0.724	0.216	0.782	14.88	0.631	0.321
	CATSplat	0.287	18.78	0.645	0.285	1.120	12.65	0.589	0.384
	Infer3D (DiffAE)	0.102	20.55	0.761	0.182	**0.461**	**18.46**	**0.714**	**0.220**
	Infer3D (StyleGAN)	0.220	17.73	0.722	0.224	0.637	17.08	0.694	0.251
Vases	Splatter Image	**0.061**	**20.82**	**0.751**	**0.179**	0.775	15.01	0.661	0.315
	Viewset Diffusion	0.138	19.98	0.684	0.245	0.802	14.12	0.628	0.368
	CATSplat	0.295	17.85	0.612	0.305	0.945	13.65	0.552	0.412
	Infer3D (DiffAE)	0.119	19.65	0.715	0.215	**0.552**	**17.65**	**0.678**	**0.261**
	Infer3D (StyleGAN)	0.245	16.92	0.675	0.258	0.725	16.27	0.658	0.292

Table 2: **ShapeNet (11 Categories).** Comparison of Splatter Image vs. Infer3D (Ours) (StyleGAN) across difficulty settings.

Setting	Method	CD	PSNR	SSIM	LPIPS
In-dist.	Splatter Img.	**0.122**	**24.27**	**0.911**	**0.107**
	Infer3D (SG)	0.147	21.61	0.878	0.133
OOD SO(3)	Splatter Img.	0.181	18.12	0.822	0.208
	Infer3D (SG)	**0.137**	**20.67**	**0.865**	**0.148**
OOD SE(3)	Splatter Img.	0.323	14.27	0.771	0.308
	Infer3D (SG)	**0.306**	**16.30**	**0.795**	**0.238**

Table 3: **RE10K Evaluation.** Comparison of CATSplat vs. Infer3D (DAE) across in-distribution and extreme focal distortion

Setting	Method	PSNR	SSIM	LPIPS
In-dist.	CATSplat	**25.41**	**0.840**	**0.149**
	Infer3D (DAE)	24.03	0.801	0.186
OOD Fisheye	CATSplat	16.25	0.610	0.327
	Equidistant	19.24	0.688	0.253
	Infer3D (DAE)	**23.33**	**0.775**	**0.198**

CATSplat (Roh et al., 2024), and Viewset Diffusion (Szymanowicz et al., 2023). As shown in Table 1 and Table 2, feedforward baselines degrade sharply under OOD shifts, often hallucinating or flattening geometry (Figure 4a). By explicitly inverting generative models, Infer3D successfully matches the observation while maintaining view-consistent structure, recovering performance levels comparable to ID inference.

We further evaluate viewpoint robustness against BRFI (Pavllo et al., 2023) and 3D GAN Inversion (Ko et al., 2023) on ShapeNet Cars, for which pretrained generators are available. We use BRFI's released model, which is trained on the full SRN pose distribution, whereas Infer3D's prior and lifter only see restricted elevation rings of ShapeNet Cars. We adapt 3D GAN Inversion using the official EG3D (Chan et al., 2022) Cars checkpoint, following its initialization protocol to train latent and pose encoders on SRN training renders with known poses. On cars with OOD SO(3) viewpoints, Infer3D achieves 20.29 dB, outperforming BRFI (14.30 dB) and 3D GAN Inversion (15.23 dB), even when the baselines receive ground-truth poses (17.70 and 17.37 dB, respectively). Infer3D also preserves more of its ID performance than BRFI, with a drop of 1.14 versus 5.55 dB (Table 10).

For CO3D where scale-depth ambiguity is significant, we also incorporate a geometric consistency loss $\mathcal{L}_{\text{depth}}$ to constrain the optimization to plausible absolute depth scales. We minimize the Mean Squared Error (MSE) between the rendered predicted depth map $\hat{\mathbf{D}}$ and the ground-truth depth $\mathbf{D}$ as $\frac{1}{N}\sum_{i=1}^{N}(\hat{\mathbf{D}}_i - \mathbf{D}_i)^2$. Without extra guidance, the optimization struggles to resolve absolute scale in CO3D (as also reported in related works (Szymanowicz et al., 2024; Tewari et al., 2023; Szymanowicz et al., 2023)).

Sensor and Field-of-View Shift. Standard feedforward models assume strict pinhole camera projection. We evaluate robustness to intrinsic parameter shifts by simulating realistic wide-angle and radial/tangential fisheye distortions (Lensfun Developers, 2024; Deng et al., 2025) on RealEstate10K (RE10K). To handle this, we augment our optimization loop with a Hybrid Neural Distortion Field. As shown in Table 3 and Figure 4b, the baseline (CATSplat) misinterprets lens warping as scene geometry (e.g., curved walls). As a stronger baseline, *Equidistant* is given privileged access to the ground-truth source and target focal lengths and undistorts the input under an equidistant fisheye model before applying CATSplat. Because this model



Figure 5: **Robustness to Unseen Texture and Lighting.** Despite training exclusively on SRN Cars (Sitzmann et al., 2019), our method (**Infer3D**) integrates semantic features (DINOv2) to infer plausible 3D structures unlike feedforward baseline Splatter Image (2nd column) or generative baseline LGM (3rd column) trained on equal or larger (Objaverse) synthetic datasets.

Table 4: **Adaptive Inference in Mixed ID/OOD Streams.** Performance (PSNR) and amortized runtime using our adaptive strategy across simulated real-world distributions on CO3D Hydrants.

ID/OOD Split	F1	Splatter Image	Opt. Only	Adaptive	Avg. Time
99% / 1%	0.943	21.42	20.53	**21.45**	2.1s
90% / 10%	0.955	20.91	20.34	**21.16**	19s
70% / 30%	0.971	19.76	19.92	**20.54**	55s
50% / 50%	0.986	18.61	19.50	**19.94**	1.4m

Table 5: **3D reconstruction accuracy on unseen texture and lighting.** Chamfer Distance (m^2, ↓) on ARKit car scenes from the RealCars dataset.

Method	Mean CD ↓	Median CD ↓
Splatter Image (Szymanowicz et al., 2024)	0.219	0.179
Splatter Image + DINOv2 + depth	0.152	0.134
LGM (Tang et al., 2024)	0.405	0.307
SF3D (Boss et al., 2024)	0.242	0.206
BRFI	0.491	0.502
3D GAN Inversion	0.387	0.373
Infer3D	**0.131**	**0.093**

cannot capture the non-linear distortion in our inputs, residual warping remains. Infer3D explicitly models the forward imaging process, successfully disentangling the non-linear lens characteristics from the 3D geometry to recover a clean structural output.

4.2 Appearance Shift and In-the-Wild Generalization

Unseen Texture and Lighting. We evaluate on the RealCars dataset (Du et al., 2025) (in-the-wild captures with complex reflections), restricting the training of our generative prior exclusively to synthetic SRN Cars (Sitzmann et al., 2019). For the 3D lifter, we use official Splatter Image weights pretrained on SRN Cars. Because standard pixel-level reconstruction fails under this synthetic-to-real gap, we integrate DINOv2 semantic feature optimization into Infer3D. While Splatter Image collapses (Chamfer Distance: 0.219) and generative baselines like Large Multi-View Gaussian Model (LGM) (Tang et al., 2024), trained on a much larger Objaverse (Deitke et al., 2023) dataset, hallucinate artifacts, our analysis-by-synthesis bypasses the texture nuisance and reconstructs the correct 3D pose and geometry, achieving a significantly lower mean Chamfer Distance of 0.131 (Table 5, Figure 5). For fairness, we also include a finetuned Splatter Image baseline where we provide it with the same DINO features we use as input (PCA-reduced features of channel-32) and pseudo-ground-truth depth from Depth Anything 3 (Lin et al., 2025). This improves the robustness of Splatter Image (Mean CD of 0.152 from 0.219) but still significantly underperforms Infer3D, demonstrating that inverse generative modeling remains essential for generalization.

The improvement also holds against optimization-based inversion methods trained on synthetic SRN Cars. Infer3D achieves a mean Chamfer Distance of 0.131, compared with 0.491 for BRFI (Pavllo et al., 2023) and 0.387 for 3D GAN Inversion (Ko et al., 2023), reducing error by 73% and 66%, respectively (Table 5).

Zero-Shot In-the-Wild Captures. Figure 6 demonstrates Infer3D on casual smartphone photos with varying lighting, partial occlusions, and non-standard cropping. Feedforward baselines fail catastrophically as

Table 6: **Out-of-Distribution Detection AUROC.** We evaluate the ability to distinguish between ID and OOD camera shifts.

Method	Hydrants ↑	Vases ↑	RE10K ↑
Infer3D	**0.972**	**0.961**	**0.997**
Image alignment	0.554	0.621	0.748

Table 7: **Calibration Sensitivity.** The detection threshold is highly stable and requires very few validation objects to calibrate (CO3D Hydrants).

Num Val Objects	Threshold (ID %)	F1 Score
10	97.7	0.965
20	97.1	0.944
30	97.3	0.955

Table 8: **Ablation of Inference Losses.** Evaluation of our objective components on CO3D Hydrants. Combining perceptual losses with latent and depth regularization is critical for recovering accurate OOD geometry.

Loss Function Configuration	PSNR ↑	LPIPS ↓	CD ↓
$\mathcal{L}_{\text{MSE}}$	14.85	0.482	0.954
$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{LPIPS}}$	16.20	0.315	0.725
$+\mathcal{L}_{\text{prior}}$ (Mean Latent Dist. Reg.)	17.45	0.255	0.580
$+\mathcal{L}_{\text{depth}}$ (Full Objective)	**18.51**	**0.218**	**0.458**

Table 9: **Pose and latent optimization ablation.** Evaluation on OOD SO(3) ShapeNet Cars using the StyleGAN2 prior. Changes are relative to full Infer3D.

Configuration	PSNR ↑	Δ PSNR
Full Infer3D (StyleGAN 2)	**20.29**	–
Without pose optimization	17.19	−3.10
Without latent optimization	18.02	−2.27
Without both	16.21	−4.08

these images violate strict training preprocessing pipelines. By searching for the camera parameters that best align the generative prior to the noisy image, our compositional approach robustly recovers plausible shapes.

4.3 Adaptive Inference and System Analysis

While optimization-based inference is highly robust, it requires more compute than feedforward mapping. We analyze the initial reconstruction error from our generative prior to validate an adaptive system that selectively routes inputs to reduce the effective amortized time per sample.

OOD Detection and Adaptive Routing. We follow the method of Section 3.5 to perform OOD detection. We use the encoder of Diffusion Autoencoders (DiffAE) (Preechakul et al., 2022) to provide a near-optimal initialization for ID inputs, which exhibits a low initial reconstruction error. As shown in Table 6, our initial reconstruction error metric achieves strong detection performance with an AUROC of 0.972 on CO3D Hydrants, 0.961 on Vases, and 0.997 on RE10K. We compare this with an image alignment baseline: the OOD image is fed directly to Splatter Image and the output 3D Gaussians are rendered at identity pose $\mathbf{I}$. We then build a classification score using the reconstruction accuracy of this render and the OOD image. Because encoder-based 3D reconstruction has a bias of copying the image at identity pose, this performs poorly and our method significantly outperforms it while requiring just roughly 286ms of latency. We also apply this adaptive strategy to a mixed stream of ID and OOD inputs (Table 4). In a stream with 10% OOD data, our system routes standard inputs through the fast feedforward lifter and reserves optimization for outliers, achieving better performance than Splatter Image alone, or optimization alone. The amortized runtime is just 19 seconds per sample. Finally, calibration of our adaptive threshold is sample-efficient and stable across sample sizes. Fitting on merely 10 held-out objects takes ≈ 500ms and yields an F1 score of 0.965 (see Table 7).

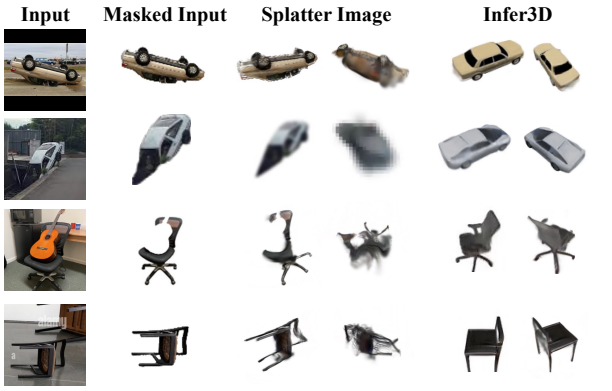


Figure 6: **Reconstruction from in-the-wild real input images.** Each row shows: the original photo; the binary mask fed to the model; the Splatter Image feed-forward prediction; and our analysis-by-synthesis reconstruction. Our method is robust to inputs of varying quality, including low-resolution photos, cluttered backgrounds, and masks that partially occlude the object.

Method	ID SRN Cars PSNR ↑	OOD SO(3) PSNR ↑
BRFI	19.85	14.30 (17.70)
3D GAN Inversion	16.22	15.23 (17.37)
Infer3D (StyleGAN2)	**21.43**	**20.29**

Table 10: **Viewpoint generalization on ShapeNet Cars.** We compare our method's performance against other 3D inversion methods: BRFI (Pavllo et al., 2023) and 3D GAN Inversion (Ko et al., 2023). All 3 methods are given unposed input. Performance with oracle pose is reported in parentheses. Oracle-pose variants receive the ground-truth test camera pose.

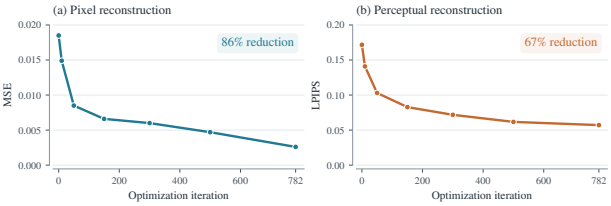


Figure 7: **Optimization convergence on OOD ShapeNet Cars.** Unweighted MSE and LPIPS reconstruction terms on OOD ShapeNet Cars objects rapidly decrease by 86% and 67%, respectively.

Ablations on Compute Tradeoffs. We run inference for 783 iterations, initializing 600 hypotheses (30 rotations × 20 latent codes) and pruning to 32, 10, and 5 at iterations 3, 122, and 302. Where available, an encoder provides one warm-start hypothesis. Full optimization settings are given in Section B.3.

We find that the computational cost of Infer3D can be explicitly scaled to prioritize speed without sacrificing robustness. By parallelizing particle optimization ($B = 10$), full convergence is achieved efficiently in roughly 2.8 minutes. For applications requiring even faster turnaround, early stopping the optimization loop after just 100 steps yields a rapid 1.2-minute runtime per OOD sample. Notably, this early-stopping configuration still achieves a PSNR of 16.98 which still outperforms the baseline feedforward model Splatter Image (15.73) on OOD CO3D Hydrants. In practice, data streams mix standard inputs with occasional anomalies; pairing rapid OOD detection with analysis-by-synthesis (without early stopping) allows Infer3D to process mixed streams efficiently without bottlenecking pipelines (Table 4).

Objective and Optimization Ablations. Finally, we validate our optimization objective and the necessity of our constrained approach. As detailed in Table 8, relying solely on pixel-wise MSE for CO3D yields poor geometry. Sequentially integrating perceptual penalties ($\mathcal{L}_{\text{LPIPS}}$), latent regularization ($\mathcal{L}_{\text{prior}}$), and depth consistency ($\mathcal{L}_{\text{depth}}$) is critical for recovering accurate out-of-distribution structures, improving Chamfer Distance from 0.954 to 0.458. Furthermore, we test whether simply updating the weights of the feedforward lifter on the OOD input (trivial test-time training) could achieve similar robustness. Table 13 demonstrates that this naive adaptation collapses the geometry and performs worse than the frozen baseline (12.41 vs. 15.73 PSNR), indicating that optimizing over an internal generative model is more effective than naive test-time adaptation of the lifter.

We further examine how pose and latent optimization contribute to our method. On OOD SO(3) ShapeNet Cars, removing pose or latent updates reduces PSNR by 3.10 and 2.27 dB, respectively, while removing both reduces it by 4.08 dB (Table 9). We do not change the initialization in these ablations, only whether the target variable is optimized further or selected among initialization candidates. These results support jointly refining shape and viewpoint during inference and show the gain is spread across the two optimization regimes.

The optimization also progressively improves the fit to the input image, reducing MSE by 86% and LPIPS by 67% over the reported checkpoints on OOD ShapeNet Cars (Fig. 7).

5 Limitations and Conclusion

Limitations. While Infer3D significantly improves robustness to OOD shifts, several limitations remain. While our method excels at OOD recovery, purely optimization-based inference can result in slightly lower quantitative metrics on in-distribution data compared to a specialized feedforward network. However, our proposed adaptive inference strategy effectively neutralizes this issue by classifying inputs based on convergence dynamics, ensuring that optimization is only triggered when necessary. We further discuss this drop in appendix Section C.6. Additionally, the visual sharpness of our OOD reconstructions is bound by the capabilities of the underlying 2D generative prior. In this work, we utilized lighter-weight generators (e.g., DiffAE, StyleGAN2) which can sometimes struggle with high-frequency details. We believe that

integrating state-of-the-art diffusion models as the generative backbone would resolve this, albeit at a higher computational cost.

Conclusion. We presented a framework for robust monocular 3D reconstruction that treats inference as an inverse generative problem. We demonstrated that standard feedforward networks overfit to the training distribution, failing catastrophically under viewpoint, sensor, or appearance shifts. By combining a learned 3D lifter with a test-time optimization loop that enforces a 2D generative prior, we recover accurate geometry in scenarios where baselines produce collapsed artifacts. Furthermore, we showed that the dynamics of this optimization process serve as a strong signal for OOD detection, enabling an adaptive system that is both efficient on familiar data and robust to the unexpected. We hope this work encourages a shift from purely amortized inference toward hybrid methods for reliable 3D perception.

References

Amro Abbas and Stéphane Deny. Progress and limitations of deep networks to recognize objects in unusual poses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 160–168, 2023. 3

Edward H Adelson. Lightness perception and lightness illusions. In Michael S Gazzaniga (ed.), *The new cognitive neurosciences*, pp. 339–351. MIT press, 2000. 1

Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4845–4854, 2019. 3

Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1175–1186, 2019. 3

Titas Anciukevicius, Patrick Fox-Roberts, Edward Rosten, and Paul Henderson. Unsupervised causal generative understanding of images. *Advances in Neural Information Processing Systems*, 35:37037–37054, 2022. 3

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 157–164. 2023a. 2

Volker Blanz and Thomas Vetter. *A Morphable Model For The Synthesis Of 3D Faces*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023b. ISBN 9798400708978. URL https://doi.org/10.1145/3596711.3596730. 3

Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. *arXiv preprint arXiv:2408.00653*, 2024. 9

Jack Brady, Bernhard Schölkopf, Thomas Kipf, Simon Buchholz, and Wieland Brendel. Generation is required for data-efficient perception. *arXiv preprint arXiv:2512.08854*, 2025. 4

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. 22

Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proc. CVPR*, 2021. 3

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proc. CVPR*, pp. 16123–16133, 2022. 8

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 7

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19457–19467, 2024. 1, 2

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025. 2

Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013. 1

Taco S Cohen and Max Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016. 3

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023. 9, 20

Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for so (3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12200–12209, 2021. 3

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. 20

Youming Deng, Wenqi Xian, Guandao Yang, Leonidas Guibas, Gordon Wetzstein, Steve Marschner, and Paul Debevec. Self-calibrating gaussian splatting for large field of view reconstruction. *arXiv preprint arXiv:2502.09563*, 2025. 6, 8

Xiaobiao Du, Yida Wang, Haiyang Sun, Zhuojie Wu, Hongwei Sheng, Shuyun Wang, Jiaying Ying, Ming Lu, Tianqing Zhu, Kun Zhan, and Xin Yu. 3DRealCar: An in-the-wild RGB-D car dataset with 360-degree views. In *Proc. ICCV*, 2025. 9

Emilien Dupont, Miguel Bautista Martin, Alex Colburn, Aditya Sankar, Josh Susskind, and Qi Shan. Equivariant neural rendering. In *International Conference on Machine Learning*, pp. 2761–2770. PMLR, 2020. 3

Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so (3) equivariant representations with spherical cnns. In *Proceedings of the european conference on computer vision (ECCV)*, pp. 52–68, 2018. 3

Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010. 1

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *Advances in Neural Information Processing Systems*, 2024. 3

Nishad Gothoskar, Marco Cusumano-Towner, Ben Zinberg, Matin Ghavamizadeh, Falk Pollok, Austin Garrett, Josh Tenenbaum, Dan Gutfreund, and Vikash Mansinghka. 3dp3: 3d scene perception via probabilistic programming. *Advances in Neural Information Processing Systems*, 34:9600–9612, 2021. 3

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. *arXiv preprint arXiv:2311.04400*, 2023. 1, 2

Kohitij Kar, Jonas Kubilius, Kailyn Schmidt, Elias B Issa, and James J DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition. *Nature neuroscience*, 22(6): 976–983, 2019. 2

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020. 6

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/. 4

Daniel Kersten, Pascal Mamassian, and Alan Yuille. Object perception as bayesian inference. *Annual review of psychology*, 55:271–304, 2004. 1

David C Knill and Whitman Richards. *Perception as Bayesian inference*. Cambridge University Press, 1996. 1

Jaehoon Ko, Kyusun Cho, Daewon Choi, Kwangrok Ryoo, and Seungryong Kim. 3D GAN inversion with pose optimization. In *WACV*, 2023. 3, 8, 9, 11

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. 1

Lensfun Developers. Lensfun: A library for lens distortion correction. https://lensfun.github.io/, 2024. 8

Brandon Leung, Chih-Hui Ho, and Nuno Vasconcelos. Black-box test-time shape REFINEment for single view 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4080–4090, June 2022. 3

Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, Angjoo Kanazawa, Afshin Rostamizadeh, and Ameesh Makadia. An analysis of SVD for deep rotation estimation. *arXiv preprint arXiv:2006.14616*, 2020. 5

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 300–309, 2023. 3

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 9, 22

Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3d supervision. *Advances in Neural Information Processing Systems*, 32, 2019. 2

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9970–9980, 2024. 3

Spandan Madan, Timothy Henry, Jamell Arthur Dozier, Helen Ho, Nishchal Bhandari, Tomotake Sasaki, Fredo Durand, Hanspeter Pfister, and Xavier Boix. On the capability of cnns to generalize to unseen category-viewpoint combinations. *arXiv preprint arXiv:2007.08032*, 2020. 3

Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. RealFusion: 360° reconstruction of any object from a single image. *Proc. CVPR*, 2023. 3

Thomas Mitchel, Michael J Taylor, and Vincent Sitzmann. Neural isometries: Taming transformations for equivariant ml. *Advances in Neural Information Processing Systems*, 37:7311–7338, 2024. 3

Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4328–4338, 2023. 3

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, and Armand Joulin. Dinov2: Learning robust visual features without supervision. *arXiv*, 2023. doi: 10.48550/arxiv.2304.07193. 6

Dario Pavllo, David Joseph Tan, Marie-Julie Rakotosaona, and Federico Tombari. Shape, pose, and appearance from a single image via bootstrapped radiance field inversion. In *CVPR*, 2023. 3, 8, 9, 11

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *Proc. ICLR*, 2023. 3

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10619–10629, 2022. 6, 10

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 18

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 7

Wonseok Roh, Hwanhee Jung, Jong Wook Kim, Seunggwan Lee, Innfarn Yoo, Andreas Lugmayr, Seunggeun Chi, Karthik Ramani, and Sangpil Kim. Catsplat: Context-aware transformer with spatial guidance for generalizable 3d gaussian splatting from a single-view image. *arXiv preprint arXiv:2412.12906*, 2024. 6, 8

Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Proc. NeurIPS*, 2019. 9

Alfredo Spagna, Zoe Heidenry, Michelle Miselevich, Chloe Lambert, Benjamin E. Eisenstadt, Laura Tremblay, Zixin Liu, Jianghao Liu, and Paolo Bartolomeo. Visual mental imagery: Evidence for a heterarchical neural architecture. *Physics of Life Reviews*, 48:113–131, 2024. 1

Fan-Yun Sun, Jonathan Tremblay, Valts Blukis, Kevin Lin, Danfei Xu, Boris Ivanovic, Peter Karkus, Stan Birchfield, Dieter Fox, Ruohan Zhang, Yunzhu Li, Jiajun Wu, Marco Pavone, and Nick Haber. Partial-view object view synthesis via filtering inversion. In *International Conference on 3D Vision (3DV)*, 2024. 3

Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8863–8873, 2023. 2, 8

Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10208–10217, 2024. 1, 2, 6, 7, 8, 9, 20

Stanislaw Szymanowicz, Jason Y Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T Barron, and Philipp Henzler. Bolt3d: Generating 3d scenes in seconds. *arXiv preprint arXiv:2503.14445*, 2025. 3

Hanlin Tang, Martin Schrimpf, William Lotter, Charlotte Moerman, Ana Paredes, Josue Ortega Caro, Walter Hardesty, David Cox, and Gabriel Kreiman. Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, 115(35):8835–8840, 2018. 2

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024. 2, 9

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011. 1

Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 1274–1283, 2017. 2

Ayush Tewari, Michael Zollhoefer, Florian Bernard, Pablo Garrido, Hyeongwoo Kim, Patrick Perez, and Christian Theobalt. High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):357–370, 2018. 3

Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Josh Tenenbaum, Frédo Durand, Bill Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *Advances in Neural Information Processing Systems*, 36:12349–12362, 2023. 2, 8

Alexander Tschantz, Beren Millidge, Anil K Seth, and Christopher L Buckley. Hybrid predictive coding: Inferring, fast and slow. *PLOS Computational Biology*, 19(8):e1011280, 2023. 2

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a. 3

Runqian Wang and Yilun Du. Equilibrium matching: Generative modeling with implicit energy-based models. *arXiv preprint arXiv:2510.02300*, 2025. 20

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024. 3

Yanbo Wang, Justin Dauwels, and Yilun Du. Compositional scene understanding through inverse generative modeling. In *Proceedings of the 2025 International Conference on Machine Learning (ICML)*, 2025b. URL https://arxiv.org/abs/2505.21780. arXiv preprint arXiv:2505.21780. 3

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*, 36:8406–8441, 2023. 3

Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. *arXiv preprint arXiv:2403.16292*, 2024. 1

Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4479–4489, 2023a. 3

Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217*, 2023b. 2

Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In *European Conference on Computer Vision*, pp. 1–20. Springer, 2024. 2, 3

Yinshuang Xu, Jiahui Lei, and Kostas Daniilidis. Equivariant light field convolution and transformer. *arXiv preprint arXiv:2212.14871*, 2022. 3

Sizhe Yang, Wenye Yu, Jia Zeng, Jun Lv, Kerui Ren, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Novel demonstration generation with gaussian splatting enables robust one-shot manipulation. *arXiv preprint arXiv:2504.13175*, 2025. 3

Yulong Yang, Felix O'Mahony, and Christine Allen-Blanchette. Learning color equivariant representations. *arXiv preprint arXiv:2406.09588*, 2024. 3

Ilker Yildirim, Tejas D Kulkarni, Winrich A Freiwald, and Joshua B Tenenbaum. Efficient analysis-by-synthesis in vision: A computational framework, behavioral tests, and comparison with neural representations. In *Proceedings of the 37th annual conference of the cognitive science society*, pp. 2727–2732, 2015. 1

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proc. CVPR*, 2021. 1, 2

Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308, 2006. 1, 2, 3

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. CVPR*, 2018. 5, 7

Bingchen Zhao, Jiahao Wang, Wufei Ma, Artur Jesslen, Siwei Yang, Shaozuo Yu, Oliver Zendel, Christian Theobalt, Alan Yuille, and Adam Kortylewski. Ood-cv-v2: An extended benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

Allan Zhou, Moo Jin Kim, Lirui Wang, Pete Florence, and Chelsea Finn. Nerf in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17907–17917, 2023. 3

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 37, 2018. URL https://arxiv.org/abs/1805.09817. 7

Eric Zhu, Mara Levy, Matthew Gwilliam, and Abhinav Shrivastava. Nerf-aug: Data augmentation for robotics with neural radiance fields. *arXiv preprint arXiv:2411.02482*, 2024. 3

A Additional Figures

Here we provide some additional figures that we reference in the main body text.

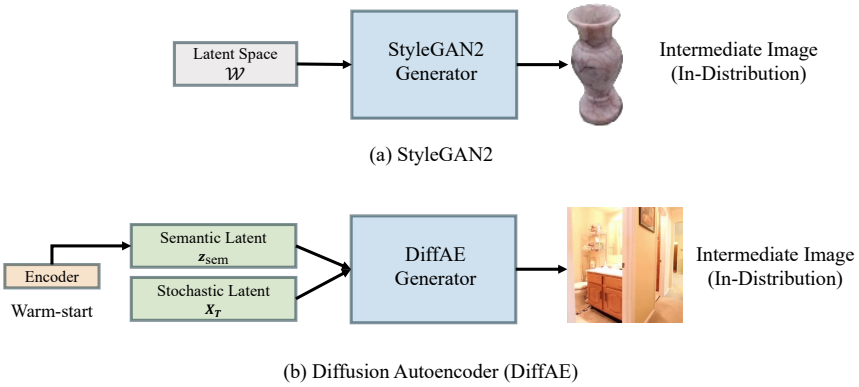


Figure 8: **Generative Backbone Architectures.** (a) **StyleGAN2:** Optimization occurs in the $\mathcal{W}$ latent space for semantic consistency. (b) **Diffusion Autoencoder (DiffAE):** Optimization leverages both a semantic latent $\mathbf{z}_{sem}$ and a stochastic latent $\mathbf{x}_T$. Its architecture allows for an encoder-based warm-start.

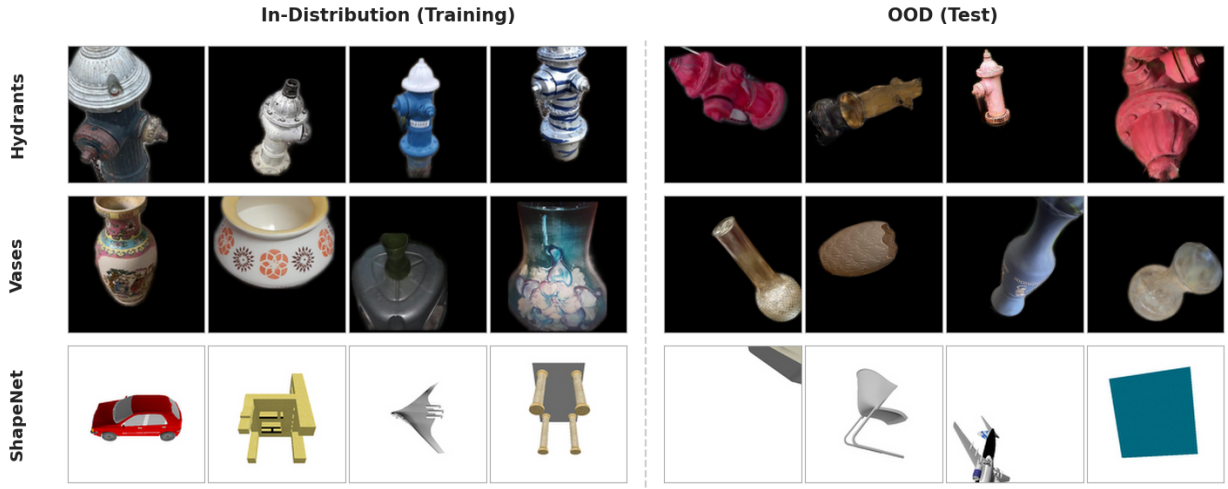


Figure 9: **Dataset splits for viewpoint robustness evaluation.** *Left:* In-distribution training images from CO3D (Hydrants, Vases) and ShapeNet. These observations are captured from standard, constrained viewpoints, such as fixed elevation rings. *Right:* Out-of-distribution (OOD) test images featuring severe camera pose shifts. These include random SE(3) translations and SO(3) rotations over the full viewing sphere, resulting in unconventional angles, extreme elevations, and partial occlusions that challenge standard feedforward priors.

B Additional Method Details

B.1 In-the-wild generalization

In order to support running our 3D reconstruction inference pipeline on in-the-wild real input images, we first acquire pseudo-GT segmentation masks using Segment-Anything 2 (Ravi et al., 2024) by initializing multiple point prompts, and selecting the largest mask that is close to the center. We then follow the method of Out-of-Distribution appearance to recover the geometry $\mathcal{S}$ and camera extrinsics $\boldsymbol{\theta}$ (See Section 3.3).

Table 11: **Optimization schedule.** Population sizes, reconstruction-loss weights, and learning rates used in the reported inference procedure.

Parameter	Value or schedule	Meaning
R	$30 \times 20 = 600$	Initial rotation–latent hypotheses
N	783	Total iterations
B	10	Maximum simultaneous hypotheses
k_t	$600 \to 32 \to 10 \to 5$	Pruning at $t = 3, 122, 302$
λ_{MSE}	$1 \to 10 \to 2$	Changes at $t = 302, 732$
λ_{LPIPS}	$0.5 \to 2$	Change at $t = 302$
λ_{prior} (DiffAE)	0.001	Constant latent-prior weight
η_z	0.01 with ramp-up and ramp-down	First 5% and final 25% of iterations
η_θ	0.01	Constant rendering-parameter learning rate

B.2 Compute Resources

To run all our experiments we use 2 80GB A100 NVIDIA GPUs. At proposal batch size = 10, our full inference occupies a Peak VRAM of 48GB GPU memory.

B.3 Optimization Settings

Table 11 summarizes the optimization schedule. Iterations are indexed from 0 to 782. Each iteration updates all active hypotheses, processing at most $B = 10$ simultaneously. Thus, B controls memory usage rather than the total population size.

Initialization. When an encoder E is available, we replace one random joint hypothesis with $\mathbf{z}_1 = E(\mathbf{I})$ and identity camera pose. This hypothesis undergoes the same optimization and pruning as the remaining hypotheses. Without an encoder, all hypotheses are randomly initialized.

Prior-specific regularization. For DiffAE, we apply a quadratic latent penalty centered on the mean of the encoder-embedded initial latent population, with weight 0.001. For StyleGAN2, we instead use noise-map regularization to penalize spatial correlations in the optimized noise maps.

Learning rates and batching. The latent learning rate uses a base value of 0.01 with a cosine ramp-up over the first 5% of iterations and a ramp-down over the final 25%. The rendering-parameter learning rate remains 0.01. When the active population exceeds $B = 10$, we process it in successive batches, completing an iteration after all active hypotheses have been updated.

C Additional Experiments

C.1 Component Ablations

We evaluate optimization variables, loss terms, and initialization counts on 25 OOD ShapeNet Cars using the StyleGAN2 prior. Component-removal experiments use identical examples and initializations. Unless specified otherwise, all settings follow the full configuration.

Removing MSE or LPIPS reduces PSNR by 1.20 and 0.53 dB, respectively, supporting their complementary contributions. Reducing the number of pose initializations from 30 to 10 costs 0.32 dB, while reducing latent initializations from 20 to 2 costs 0.80 dB. Noise-map regularization has a smaller effect of 0.12 dB.

Table 12: **Component ablations on OOD ShapeNet Cars.** Changes are relative to full Infer3D. The full configuration uses 30 pose and 20 latent initializations.

Configuration	PSNR ↑	Δ PSNR
Full Infer3D	**20.29**	–
Without pose optimization	17.19	−3.10
Without latent optimization	18.02	−2.27
Without both	16.21	−4.08
Without MSE	19.09	−1.20
Without LPIPS	19.76	−0.53
Without noise-map regularization	20.17	−0.12
Pose initializations: 30 → 10	19.97	−0.32
Latent initializations: 20 → 2	19.49	−0.80

C.2 Comparison with test-time training

We evaluate direct test-time training (TTT) on OOD CO3D Hydrants by updating the feedforward lifter's weights to match the input image. As shown in Table 13, this update scheme degrades novel-view reconstruction quality relative to the frozen lifter, whereas Infer3D improves all three metrics. For this setting, optimizing the generative latent while keeping the networks frozen is more effective than directly adapting the lifter's weights.

Table 13: **Comparison with direct test-time training on OOD CO3D Hydrants.** Best results are shown in bold.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Splatter Image	15.73	0.665	0.266
Test-Time Training	12.41	0.569	0.382
Infer3D (DAE)	**18.46**	**0.714**	**0.220**

C.3 Generative priors trained on diverse categories

To test compatibility with components trained on diverse categories, we combine a frozen class-conditional Equilibrium Matching (EqM) model (Wang & Du, 2025) trained on ImageNet (Deng et al., 2009) with a frozen Objaverse-trained Splatter Image lifter (Szymanowicz et al., 2024; Deitke et al., 2023). We evaluate on 25 partially occluded objects from a held-out set spanning six categories, using the same prior–lifter pair throughout. Infer3D-EqM reduces Chamfer Distance in every category and by approximately 30% overall (Table 14). Figure 10 illustrates improvements in object completeness, including the recovery of missing handles and body regions.

Table 14: **Reconstruction with an Equilibrium Matching prior.** Chamfer Distance (CD) and its relative reduction compared with applying the frozen lifter directly.

Category	Direct lifter CD ↓	Infer3D-EqM CD ↓	Reduction
Teapot	0.214	**0.157**	27%
Mug	0.225	**0.171**	24%
Vase	0.278	**0.180**	35%
Pineapple	0.147	**0.096**	35%
Wine bottle	0.145	**0.109**	25%
Helmet	0.214	**0.199**	7%
All	0.206	**0.145**	**30%**

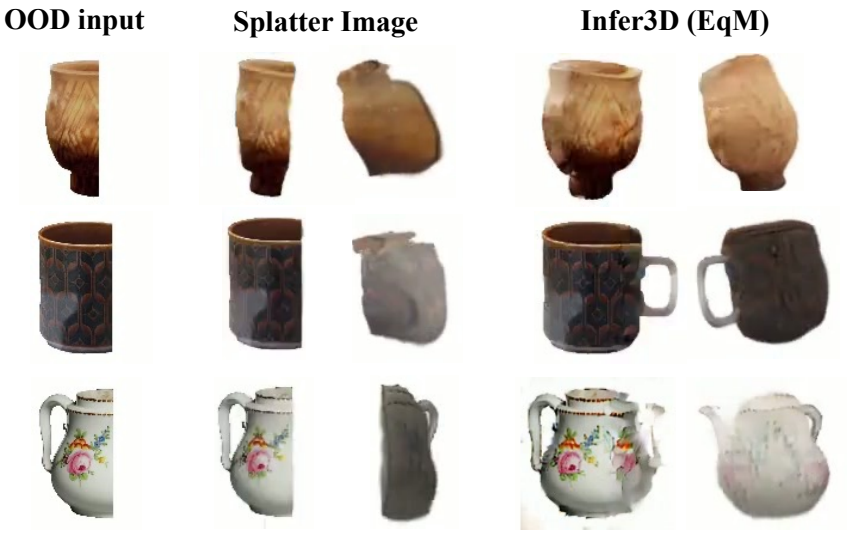


Figure 10: **Qualitative reconstruction with an EqM prior.** Left: partially occluded inputs. Middle: direct reconstructions from the frozen Splatter Image lifter. Right: Infer3D-EqM reconstructions. Two views are shown per reconstruction.

C.4 Comparison with SAM 3D

Table 15 compares novel-view synthesis quality on ShapeNet across 11 categories. Infer3D achieves the best results on all three metrics under OOD SO(3) input poses, while Splatter Image performs best on in-distribution inputs. Figure 11 presents selected SAM 3D failure cases, illustrating flattened or distorted geometry and missing structural components under challenging input orientations. We find that even in in-distribution settings, SAM3D does not accurately reconstruct the colors well and so the in-distribution NVS performance is also low. But the shape is not distorted in-distribution as it is OOD.

Table 15: **Comparison with SAM 3D on ShapeNet (11 categories).** Best results within each setting are shown in bold.

Setting	Method	PSNR ↑	SSIM ↑	LPIPS ↓
In-distribution	SAM 3D	16.00	0.79	0.181
	Splatter Image	**24.27**	**0.911**	**0.107**
	Infer3D (SG)	21.61	0.878	0.133
OOD SO(3)	SAM 3D	16.18	0.79	0.194
	Splatter Image	18.12	0.822	0.208
	Infer3D (SG)	**20.67**	**0.865**	**0.148**

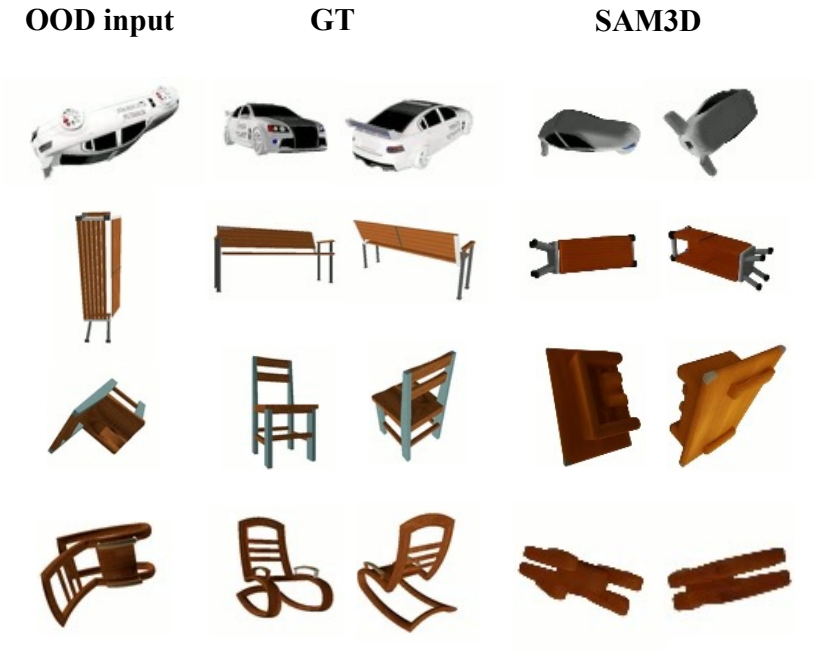


Figure 11: **SAM 3D failure cases under OOD input poses.** Each row shows an input image (left), two views of the ground-truth shape (middle), and two views of the SAM 3D reconstruction (right).

C.5 Potential application to robotic manipulation

Figure 12 illustrates a potential perception pipeline for robotic manipulation, where wrist-camera observations can involve unusual viewpoints and gripper occlusions. We estimate depth using Depth Anything 3 (Lin et al., 2025) and segment the target object using SAM 3 (Carion et al., 2025). Given the masked RGB image and estimated depth, Infer3D jointly optimizes the generative latent and camera parameters to recover the bowl geometry. This example demonstrates reconstruction from a partially occluded wrist-camera observation; its utility for downstream manipulation remains to be evaluated.

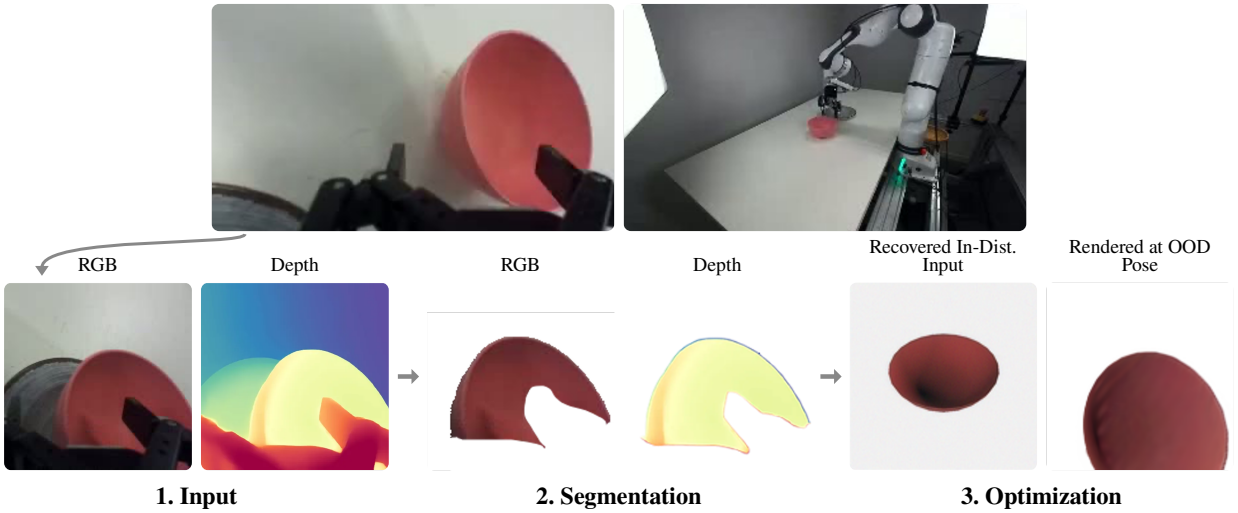


Figure 12: **Reconstruction pipeline for a robotic manipulation scene.** A wrist-camera observation undergoes depth estimation and object segmentation, followed by Infer3D optimization.

C.6 In-Distribution Trade-off

Run on every input, the optimizer alone is less accurate than the feedforward lifter on ID data: by 0.93 dB on CO3D Hydrants, 2.66 dB on ShapeNet, and 1.38 dB on RE10K (Tables 1 to 3). We perform analysis about this drop and find that the performance gap is concentrated near the input view. On ID CO3D Hydrants, the feedforward advantage is 5.7 dB for views close to the input but only 0.9 dB for the most occluded views. Unlike the feedforward lifter, the lifter inside Infer3D never receives the input image directly, so it cannot copy visible details into nearby views; for views that require inferring occluded 3D structure, Infer3D is nearly as accurate. In practice, an image-conditioned feedforward model is preferable for narrow-baseline rendering close to the input view, whereas Infer3D is more robust for wide-baseline views and 3D geometry. The detector-routed system (Table 4) avoids this cost on inputs classified as ID.