# Cue3D: Quantifying the Role of Image Cues in Single-Image 3D Generation

**Xiang Li*** **Zirui Wang*** **Zixuan Huang** **James M. Rehg**

University of Illinois at Urbana-Champaign

https://ryanxli.github.io/cue3d

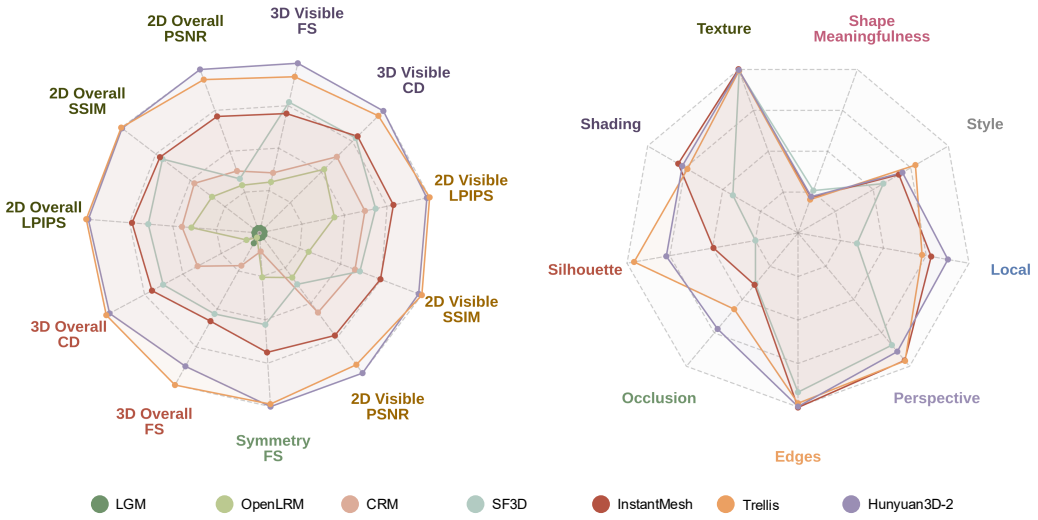


Figure 1: We present Cue3D, the first comprehensive, model-agnostic framework for quantifying the influence of individual image cues in single-image 3D generation. Left: Our unified evaluation of single-image 3D generation methods. Right: Performance robustness to the perturbation of each cue, lower values indicate higher importance. We show representative methods on Toys4K dataset for clarity; additional figures are available in the Appendix.

## Abstract

Humans and traditional computer vision methods rely on a diverse set of monocular cues to infer 3D structure from a single image, such as shading, texture, silhouette, etc. While recent deep generative models have dramatically advanced single-image 3D generation, it remains unclear which image cues these methods actually exploit. We introduce Cue3D, the first comprehensive, model-agnostic framework for quantifying the influence of individual image cues in single-image 3D generation. Our unified benchmark evaluates seven state-of-the-art methods, spanning regression-based, multi-view, and native 3D generative paradigms. By systematically perturbing cues such as shading, texture, silhouette, perspective, edges, and local continuity, we measure their impact on 3D output quality. Our analysis reveals that shape meaningfulness, not texture, dictates generalization. Geometric cues, particularly shading, are crucial for 3D generation. We further identify over-reliance on provided silhouettes and diverse sensitivities to cues such

*Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

as perspective and local continuity across model families. By dissecting these dependencies, Cue3D advances our understanding of how modern 3D networks leverage classical vision cues, and offers directions for developing more transparent, robust, and controllable single-image 3D generation models.

## 1 Introduction

Generating a 3D model from a single 2D image is a long-standing goal in computer vision, with broad applications in content creation, AR/VR, and graphics. Humans effortlessly recover 3D shape from a single view by exploiting a variety of monocular cues [3, 14, 31, 40]. Decades of research in classical computer vision studied these explicit monocular cues for shape inference, including shading patterns [19, 59], texture cues [38], silhouette contours [25], and many more. Recently, a new generation of single-image-to-3D methods has dramatically advanced the state of the art, fueled by large datasets [9] and advances in deep generative models [17]. These approaches can be grouped into three prominent categories: (i) Regression-based models that directly predict a 3D representation from the input image via a feed-forward network (*e.g.*, LRM [18], SF3D [6]), (ii) Multi-view methods that generates novel views consistent with the input image, then regress to a 3D model (*e.g.*, CRM [52], LGM [46], InstantMesh [56]), and (iii) Native 3D generative models that treat single-image-to-3D as a conditional generation problem in a learned 3D latent space (*e.g.*, Trellis [53] and Hunyuan3D-2 [60]). These approaches have enabled fast generation of textured 3D meshes from a single image, with impressive fidelity and generalization far beyond earlier methods.

Despite this rapid progress, the interpretability of single-image 3D networks remains largely underexplored. Current models are learned end-to-end on 3D supervision, and they operate as complex black boxes: we have little understanding of what information they rely on to infer 3D shape from a single image. Do these networks internally exploit the same set of visual cues as classical methods [19, 25, 38, 59], or do they rely on other information such as high-level semantics? Improving transparency in this process is important both scientifically, to connect with vision science and inform future model design, and practically, to diagnose failure modes and biases of these 3D generators.

To address this gap, we present Cue3D, the first comprehensive, model-agnostic framework for quantifying the influence of individual image cues in single-view 3D generation. We begin by establishing a unified benchmark covering seven state-of-the-art methods spanning regression-based, multi-view, and native 3D generative paradigms. We evaluate them on two standard datasets (GSO [10], Toys4K [45]). For each predicted mesh, we assess (1) both 2D appearance and 3D geometric quality for the entire shape, (2) 2D and 3D quality of the visible surface from the input viewpoint, and (3) the agreement between output and ground-truth symmetry. As summarized in Figure 1 left, native 3D generative models consistently outperform other approaches across all metrics.

We then systematically quantify the significance of each image cue. Building on meaningful perturbations [12], we disable or modify specific cues, such as silhouette shape, shading, texture semantics, perspective, and local continuity, and measure the resulting degradation in 3D output quality. Our perturbation analysis uncovers how modern single-image 3D models leverage image cues, revealing the following key insights. (1) **Meaningfulness of shape, not texture, dictates generalization.** For models to generalize, the input image must indicate a meaningful shape that does not significantly deviate from the training distribution. When we disrupt this cue by providing the models with a stochastic combination of textured primitive shapes [54], every method collapses with distinct failure modes. In contrast, the models perform surprisingly well on meaningless or random textures, with the best-performing models showing near perfect generalization. (2) **Semantics alone are not enough; Geometric cues are crucial.** Using a state-of-the-art style-transfer method [55], we convert images into artistic styles that preserve high-level semantics but often disrupts geometric cues like realistic shading and texture, as shown in Figure 2. We observe a significant drop on the performance compared to the original images, underscoring the continued importance of geometric cues. (3) **Shading is more important than texture.** To dissect the contribution of different geometric cues, we dive deeper into the image formation process. Surprisingly, even when all recognizable textures are replaced by procedural noise, natural patterns, or flat gray, for several methods, the quality of the 3D outputs remain almost unchanged, as long as the shading is kept intact. However, removing shading causes a noticeable performance decline. We further discover an interplay between shading and texture cues: intact shading alone suffices to uphold performance regardless of texture content, but when shading is removed, preserving the original texture yields better results than substituting with

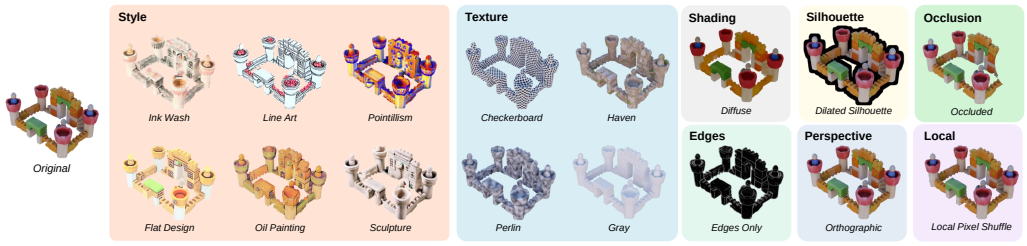


Figure 2: Overview of perturbations for analyzing individual image cues in single-image 3D generation. Starting from the original image, we systematically perturb specific visual cues. These targeted perturbations reveal the extent to which each cue influences model performance.

procedural textures or uniform color. (4) **Models are overly sensitive to silhouette and occlusion.** Dilating the object mask (without altering interior pixels) inflicts severe errors on regression-based and multi-view methods, whereas one native 3D generator remains relatively robust. In contrast, occlusion of both silhouette and image content dramatically degrades all approaches. (5) **Other cues have diverse impact.** Perturbing perspective, edge, and local-continuity signals produce measurable performance drops that vary across model categories, which we provide thorough analysis in the experiments section.

Cue3D establishes the first unified, model-agnostic framework for dissecting how modern single-image 3D generators exploit individual visual cues. Our perturbation study reveals that shape, rather than texture, meaningfulness dictates generalization. Geometric cues, especially shading, contribute significantly the 3D generation process. These models may overly rely on the provided silhouette. Meanwhile, edges, perspective, and local continuity each have distinct effects on different model families. By quantifying these dependencies across state-of-the-art approaches, Cue3D deepens our scientific understanding of image-based 3D generation, and provides potential guidance for designing more transparent, robust, and controllable single-image 3D generation methods.

## 2 Related Work

**Single-Image to 3D.** Recent advances in single-image-to-3D generation have converged on three principal paradigms. (1) Regression-based methods [6, 18, 20, 49, 51] employ neural networks to directly predict a 3D representation, such as voxels, deformed meshes, or implicit fields, from encoded image features in a single forward pass. For example, LRM [18] and its successors [6, 49] utilize transformer backbones to learn triplane representations, which are then rendered volumetrically, achieving both high fidelity and efficient inference. (2) Multi-view approaches [2, 36, 43, 46, 52, 56] follow a two-stage pipeline: first synthesizing multi-view RGB images [36], depth or coordinate maps [33, 52], normal maps [35, 37], or Gaussian splats [46], and then reconstructing 3D structure from these intermediate multi-view representations. Decoupling view synthesis from geometry enables the reuse of powerful 2D generative priors trained on billions of images [41], providing an especially strong texture prior. (3) Native 3D generative models [23, 29, 30, 50, 53, 58, 60] combine a VAE-based latent encoding of 3D data [24, 26] with a diffusion or flow model to generate high-quality and diverse 3D samples. Methods differ in their latent structures, input formats, and output representations: for instance, Hunyuan3D-2 [60] encodes point clouds to produce texture-free signed distance fields, while Trellis [53] proposes a sparse structured latent combining geometric and visual features, allowing flexible decoding into radiance fields, Gaussian splats, or meshes. Despite the iterations of approaches, it remains unclear what image cues these models rely on when producing the 3D output. In this paper, we systematically investigate how different single-image to 3D frameworks extract and transform visual signals from images cues into 3D representations.

**Image Cues.** Humans infer 3D structure from single images by integrating multiple monocular cues. Studies in developmental psychology and psychophysics show that the human visual system encodes properties like surface depth and orientation [8, 25, 44], and that internal object representations adhere to 3D constraints [42]. In contrast to humans' seamless cue integration, classical computer vision approaches explicitly leverage specific visual cues for shape inference—such as shape-from-shading [19, 21], texture gradients [22, 39], silhouettes [27, 32], contours and junctions [7], perspective effects

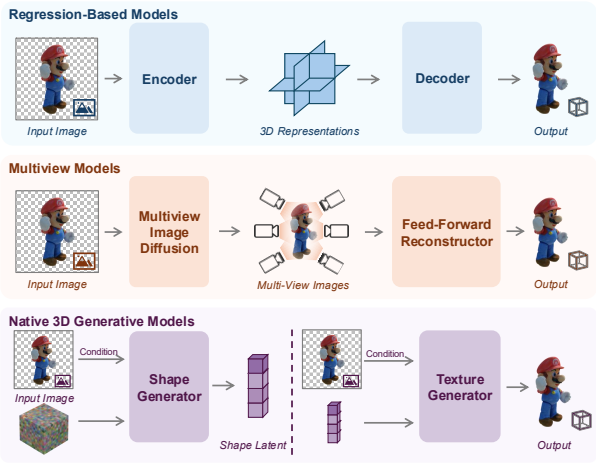


Figure 3: Illustration of the three single-image 3D generation paradigms evaluated in this paper: regression-based methods (OpenLRM [16], SF3D [6]), multi-view approaches (CRM [52], LGM [46], InstantMesh [56]), and native 3D generative models (Trellis [53], Hunyuan3D-2 [60]).

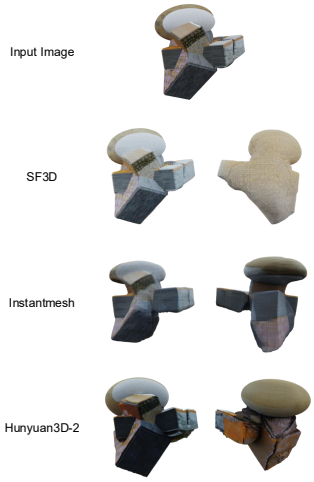


Figure 4: Qualitative comparison on the Zeroverse dataset of shapes without semantic meaning. We show one methods representative of each paradigms.

[15], and symmetry priors [4, 48]. Modern deep models instead learn these visual priors implicitly in an end-to-end manner, inspring us to explore the role of these cues in state-of-the-art models.

**Visual Cue Interpretability.** Interpreting the decision-making process of black-box models, especially with respect to the visual cues they exploit, remains an open challenge. A common strategy is input perturbation, where carefully crafted modifications are applied to input data to observe the resulting changes in model output [12, 13, 47]. For example, Geirhos et al. [13] generate images in which object shape and surface texture are semantically misaligned, revealing the relative importance of each cue in image classification models. Other approaches include latent-space probing, which trains auxiliary networks to investigate whether the internal representations of a pre-trained model fit certain downstream tasks [5, 11], and metric-based benchmarking, where models are compared across artificially curated datasets designed to emphasize specific visual attributes or cues [61]. Our work is inspired by these cue-based analysis methods but differs in key ways. Rather than solely focusing on classification or probing general features for diverse downstream tasks, we focus on presenting an in-depth analysis within the scope of single-image 3D generation. We introduce a model-agnostic framework that systematically applies controlled perturbations to distinct image cues and quantifies their effect. By evaluating a range of recent 3D architectures and employing both 2D and 3D performance metrics, our approach provides a faithful and comprehensive view of how state-of-the-art models leverage visual cues during singele-image 3D generation.

# 3 Cue3D

## 3.1 Evaluation Settings

**Methods.** We compare seven recent single-image-to-3D methods that collectively cover all three prevailing paradigms. In particular, we select OpenLRM [16] and SF3D [6] from regression-based networks, CRM [52], LGM [46] and InstantMesh [56] from multi-view reconstruction approaches, and Trellis [53] and Hunyuan3D-2 [60] from native 3D generative methods. We use the official implementation for all methods and evaluate mesh outputs in a unified way. We use 8 NVIDIA L40S GPU for all our experiments.

**Datasets.** We select two standard evaluation datasets for all methods: GSO [10], a dataset of high-quality scanned household items, and Toys4K[45], a collection of user-created 3D toy objects. We manually remove geometrically trivial objects (e.g., boxes) and balancing over-represented categories from these datasets. Our final evaluation sets contains 412 objects from the cleaned GSO dataset and 500 randomly sampled objects from the cleaned Toys4K dataset. Each object is rendered in Blender from a random camera pose (azimuth/elevation sampled within fixed limits) under a random Poly

Haven [1] HDRI lighting. More implementation details are in the appendix. To probe performance on shapes without semantic meaning, we additionally use Zeroverse [54], a procedurally generated dataset built from random assemblies of textured primitive. Zeroverse exhibits rich local geometric detail, but the shapes themselves are not meaningful, as it significantly deviate from the training distribution of single-image-to-3D methods.

**Overall Quality.** We evaluate both 2D appearance fidelity and 3D geometric quality of the 3D mesh results. We align the output mesh to the groundtruth following [6]. For appearance fidelity, we report PSNR, SSIM and LPIPS between rendered output meshes and groundtruth meshes. We render 16 views for each object with 8 uniform azimuth and 2 elevations. For geometry quality, we report the Chamfer Distance (CD) and F-scores at different thresholds to quantify the overall shape quality. The Chamfer distance between two point clouds $P_1 = \{x_i \in \mathbb{R}^3\}_{i=1}^n$ and $P_2 = \{x_j \in \mathbb{R}^3\}_{j=1}^m$. is defined as:

$$\text{chamfer}(P_1, P_2) = \frac{1}{2n}\sum_{i=1}^{n} |x_i - \text{NN}(x_i, P_2)| + \frac{1}{2m}\sum_{j=1}^{m} |x_j - \text{NN}(x_j, P_1)| \tag{1}$$

where $\text{NN}(x, P) = \arg\min_{x' \in P} \|x - x'\|$ denotes the nearest neighbor of source point $x$ in point cloud $P$.

**Visible Surface Quality.** Beyond assessing overall mesh quality, we specifically evaluate how accurately the predicted mesh aligns with the ground truth at the input image's viewpoint. We render RGB images of the output meshes from this viewpoint, obtain the corresponding depth map, and back-project the depth map into point clouds using the ground truth camera parameters. Subsequently, we employ the previously described 2D and 3D metrics on these rendered images and point clouds to quantitatively measure the quality of visible surfaces.

**Symmetry.** We further analyze the predicted object's symmetry agreement with the ground truth. Adopting the symmetry groundtruth generation procedure from [34], we apply a fixed threshold to identify planes of reflection symmetry in both predicted and ground truth meshes. For each method, we compute a binary symmetric-or-not F1 score across all predicted objects relative to their groundtruth counterparts.

### 3.2 Perturbations

We assess the importance of individual image cues through targeted perturbations. By selectively removing one cue while preserving others, significant performance degradation indicates the model's reliance on that cue. Conversely, minimal performance changes suggest the model's invariance to that cue. Additionally, preserving a single cue while removing most others can highlight its information contribution in the model's inference process. Below, we introduce the cues and their corresponding perturbations examined in this paper, illustrated in Figure 2. Additional perturbation analyses are detailed in the appendix.

**Style.** We perturb geometric cues while preserving semantic content through reference-based style transfer [55]. We select six distinct styles: ink wash, line art, pointillism, flat design, oil painting, and sculpture. We manually curate five exemplar images per style. Each object image undergoes style transfer using a randomly selected style exemplar for each of the six styles. This approach preserves core 3D structure perceptually while altering geometric cues like shading and texture, as shown in Figure 2.

**Shading & Texture.** Given their prominence as geometric cues, we jointly analyze shading and texture perturbations within the rendering pipeline. We perturb shading by rendering diffuse maps in Blender. Specifically, since the groundtruth texture in the GSO dataset has baked-in lighting, we employ an image delighting method [60] to remove baked-in lighting for the GSO dataset. Texture perturbations involve swapping original textures with alternatives such as uniform checkerboards, Perlin noise, random textures from Poly Haven [1], and uniform gray. Each texture variant is rendered both with and without lighting (diffuse).

**Silhouette and Occlusion.** Silhouette captures global shape information, and many models explicitly takes object mask as input. We investigate its influence through mask dilation and occlusion. We first dilate the silhouette (alpha mask) of each object by a fixed pixel width, leaving other cues intact. Subsequently, we simulate occlusion by placing scaled masks of randomly selected objects from the dataset onto the original mask boundaries, creating weak, medium, and strong occlusion conditions. Though occlusion partially hides image content, humans typically can still mentally reconstruct the

(a) GSO

| Method | Overall 2D | | | Overall 3D | | Symmetry | Visible Surface 2D | | | Visible Surface 3D | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | $CD_{\times 1000}$ ↓ | FS ↑ | FS ↑ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | $CD_{\times 1000}$ ↓ | FS ↑ |
| LGM | 16.20 | 0.807 | 0.291 | 83.01 | 0.034 | 0.188 | 16.83 | 0.819 | 0.256 | 46.00 | 0.215 |
| OpenLRM | 17.09 | 0.820 | 0.245 | 80.89 | 0.033 | 0.391 | 17.48 | 0.824 | 0.218 | 33.00 | 0.297 |
| CRM | 17.68 | 0.833 | 0.232 | 68.07 | 0.043 | 0.285 | 18.49 | 0.845 | 0.193 | 31.10 | 0.298 |
| SF3D | 16.71 | 0.838 | 0.219 | 61.58 | 0.059 | 0.488 | 17.27 | 0.839 | 0.187 | 25.70 | 0.411 |
| InstantMesh | 19.01 | 0.849 | 0.192 | 54.54 | 0.072 | 0.715 | 19.21 | 0.853 | 0.168 | 24.00 | 0.424 |
| Hunyuan3D-2 | **19.98** | 0.862 | 0.159 | 41.82 | 0.087 | **0.894** | **20.08** | 0.863 | 0.143 | **19.10** | **0.497** |
| Trellis | 19.85 | **0.864** | **0.157** | **39.64** | **0.092** | 0.867 | 19.95 | **0.867** | **0.141** | 19.80 | 0.472 |

(b) Toys4K

| Method | Overall 2D | | | Overall 3D | | Symmetry | Visible Surface 2D | | | Visible Surface 3D | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | $CD_{\times 1000}$ ↓ | FS ↑ | FS ↑ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | $CD_{\times 1000}$ ↓ | FS ↑ |
| LGM | 16.76 | 0.833 | 0.272 | 77.01 | 0.051 | 0.270 | 17.42 | 0.843 | 0.243 | 41.50 | 0.259 |
| OpenLRM | 17.85 | 0.853 | 0.221 | 74.79 | 0.047 | 0.419 | 18.51 | 0.859 | 0.192 | 28.00 | 0.351 |
| CRM | 18.21 | 0.860 | 0.214 | 61.88 | 0.064 | 0.321 | 19.45 | 0.875 | 0.170 | 25.20 | 0.370 |
| SF3D | 18.01 | 0.875 | 0.186 | 52.78 | 0.094 | 0.600 | 18.69 | 0.876 | 0.162 | 21.00 | 0.512 |
| InstantMesh | 19.59 | 0.876 | 0.173 | 49.84 | 0.098 | 0.706 | 20.06 | 0.883 | 0.149 | 20.60 | 0.489 |
| Hunyuan3D-2 | **20.79** | 0.893 | 0.138 | 38.65 | 0.126 | **0.913** | **21.08** | 0.897 | 0.124 | **14.90** | **0.590** |
| Trellis | 20.53 | **0.893** | **0.136** | **37.78** | **0.137** | 0.904 | 20.85 | **0.898** | **0.122** | 16.00 | 0.563 |

Table 1: Unified evaluation results on the GSO and Toys4K datasets. Native 3D generative models achieve the highest overall performance across metrics.

| Method | Overall 2D | | | Overall 3D | | Symmetry | Visible Surface 2D | | | Visible Surface 3D | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | $CD_{\times 1000}$ ↓ | FS ↑ | FS ↑ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | $CD_{\times 1000}$ ↓ | FS ↑ |
| LGM | 16.15 | 0.754 | 0.321 | 86.19 | 0.019 | 0.000 | 16.98 | 0.767 | 0.289 | 54.50 | 0.134 |
| OpenLRM | 16.86 | 0.761 | 0.283 | 96.59 | 0.019 | 0.200 | 17.34 | 0.761 | 0.252 | 45.20 | 0.193 |
| CRM | 17.17 | 0.771 | 0.280 | 81.45 | 0.021 | 0.000 | **18.65** | **0.791** | **0.223** | 40.40 | 0.200 |
| SF3D | 15.11 | 0.767 | 0.276 | 90.34 | 0.021 | 0.267 | 15.83 | 0.768 | 0.231 | 38.60 | **0.249** |
| InstantMesh | 16.89 | 0.752 | 0.304 | 89.47 | 0.021 | 0.467 | 17.68 | 0.769 | 0.263 | 46.80 | 0.185 |
| Hunyuan3D-2 | **17.63** | 0.770 | **0.258** | **78.09** | 0.024 | **0.933** | 18.18 | 0.777 | 0.233 | **35.50** | 0.239 |
| Trellis | 17.29 | **0.773** | 0.264 | 78.14 | **0.024** | 0.467 | 17.75 | 0.781 | 0.248 | 43.20 | 0.181 |

Table 2: Evaluation results on the Zeroverse dataset of shapes without semantic meaning. Performance significantly drops compared to GSO and Toys4K, underscoring the significance of shape meaningfulness.

complete 3D shape by leveraging shape priors. These variants test the model's capability to infer complete 3D structures despite partial visibility. Additional perturbation scenarios to the silhouette are presented in the appendix.

**Edges.** Edges are analyzed due to their role in separating surfaces and indicating curvature. We first extract edge maps using the Canny algorithm from input images. Two perturbation strategies are used: one replaces all internal object cues (except silhouette) with edge maps alone, evaluating if edges can sufficiently provide information for shape inference. The other softens edges by applying Gaussian blurring only in the local neighborhood of detected edges, merging adjacent surface regions visually. Significant performance drops under these perturbations would highlight the model's reliance on precise edge information, while negligible drop would indicate invariance. Additional edge extraction methods and results are included in the appendix.

**Perspective.** Perspective cue could indicate vanishing points and spatial relationships. This cue is perturbed by switching the rendering camera to an orthographic projection. Eliminating perspective effects enables evaluation of the model's dependence on perspective cues.

**Local Continuity.** To assess sensitivity to local structural details, we perturb local continuity by splitting image foreground into grids of $n \times n$ pixels and shuffling pixels within each grid cell independently. This maintains global structure while disrupting local detail continuity. Greater performance degradation under this perturbation reflects higher sensitivity to local information.

# 4 Results

## 4.1 Unified Evaluation

We begin by conducting a unified evaluation of all seven methods on the GSO and Toys4K datasets. We present the summary of the results in Figure 1 (left), and the full evaluation details in Table 1.

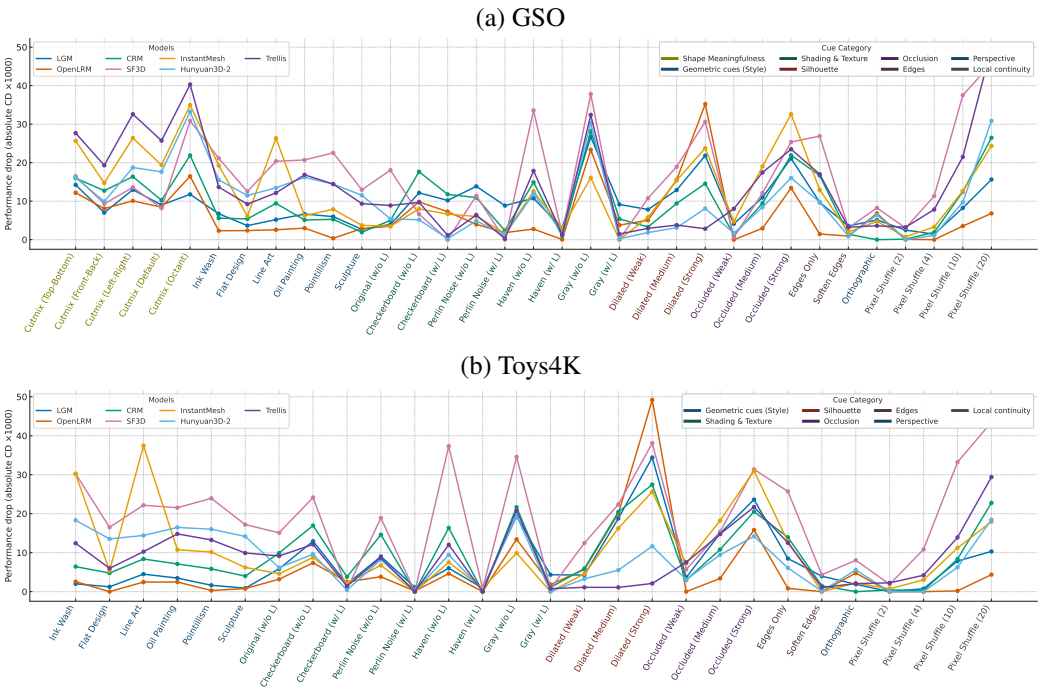


Figure 5: Quantitative analysis of image cue perturbations on single-image 3D generation. We report Chamfer Distance (CD $\times 1000$ for clarity; lower is better) for each model under different perturbations. A larger increase in CD indicates greater performance degradation, revealing the model's reliance on the perturbed cue. A tabulated version of these results is available in Appendix.

Our results show that the native 3D generative methods, see Hunyuan3D-2 and Trellis in Table 1, clearly outperform other methods across both datasets. Generally, these two methods are closely followed by InstantMesh, the best-performing multi-view method, and SF3D, the leading regression-based method, followed by the remaining methods.

Regarding 2D appearance quality, as shown in Table 1 Overall 2D and Visible Surface 2D columns, native 3D generative methods have only marginal improvements in terms of PSNR and SSIM compared to other methods. However, they substantially outperform the alternatives in terms of LPIPS scores. This suggests that, although pixel-level statistics appear similar across methods, the native 3D generative methods more accurately capture higher-level visual information encoded by deep features.

The most substantial advantage of native 3D generative methods emerges in their 3D geometry quality, see Table 1 Overall 3D and Visible Surface 3D columns. These methods exceed the next-best method by over 10 points in overall geometry evaluation and by more than 4 points on visible surfaces under our CD$\times$1000 metric. The visible surface quality assessments closely align with the overall geometry evaluations. Additionally, native 3D generative methods excel significantly in modeling symmetry, see Table 1 Symmetry column. They closely match the ground-truth symmetry across both datasets.

Comparing the two top methods, Hunyuan3D-2 and Trellis achieve similar 2D quality despite differences in their texture modeling approaches. Trellis demonstrates slightly better overall 3D quality, whereas Hunyuan3D-2 slightly excels in symmetry and visible surface quality. These insights provide valuable guidance for selecting the most appropriate method for practical, real-world applications.

### 4.2 Image Cues Analysis

In this section, we quantitatively analyze the role of various image cues in single-image 3D generation. An overview of our key findings is illustrated in Figure 1 (right). Detailed results are presented in Figure 5. Performances in this table are measured by Chamfer Distance (CD, scaled $\times 1000$ for clarity), where lower values indicate better performance, thus a larger increase represents significantly

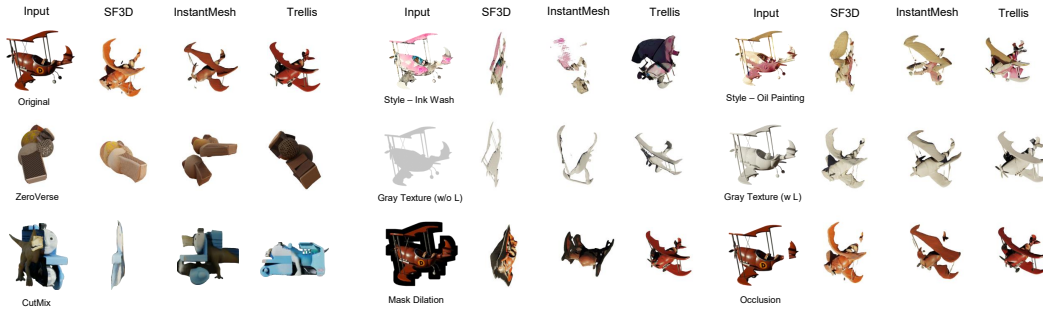


Figure 6: Qualitative example of our image cue perturbation analysis. More qualitative results are available on the project webpage and in the Appendix.

degraded performance. We show qualitative examples in Figure 6. More qualitative results are available on the project webpage and in the Appendix.

**Shape Meaningfulness.** We probe the role of shape meaningfulness in two complementary ways: (i) Zeroverse and (ii) CutMix on GSO. First, we use Zeroverse [54], a dataset comprising procedurally generated combinations of textured primitive shapes. We show qualitative results of representative methods in Figure 4, and the quantitative evaluations in Table 2. Figure 4 shows that the input image does not correspond to a meaningful shape, and has a significant gap to the training distribution. Performance notably declines across both 2D and 3D metrics compared to the more meaningful GSO and Toys4K datasets, confirming that meaningful shape in the input images are critical for the generalization of single-image 3D generation. Native 3D generative methods generally maintain the highest overall quality, while CRM best recovers visible surfaces in 2D, and SF3D and Hunyuan3D-2 perform best in visible surface 3D quality.

We further examine how the absence of shape meaningfulness influences different failure modes qualitatively in Figure 4 and quantitatively in the Appendix. Regression-based methods produce smooth, averaged back surfaces. We quantify this phenomenon by measuring the difference between each normal and the average normal in its local neighborhood, normalized against the ground truth, and we see a substantial drop of this metric on Zeroverse. Multi-view methods fail due to inconsistencies in synthesized views, as evidenced by decreased pairwise DINOv2 similarity scores, contributing to degraded 3D performance. Native 3D generative methods, lacking meaningful shape information, tend to hallucinate symmetrical completions, resulting in higher false-positive symmetry detections. See appendix for the detailed results of these experiments. These diverse failure modes underline the crucial role of meaningful shape cues, particularly for reconstructing occluded surfaces.

Meanwhile, to test shape meaningfulness with minimal domain shift, we introduce shape CutMix variants on GSO. We combine parts of different GSO meshes to perturb shape meaningfulness, while keeping appearance statistics similar. We conducted several shape CutMix experiments of varying difficulty, Given two meshes $M$ and $N$:

1) *Half-and-half*: We mix half of mesh $M$ with the other half of mesh $N$ to construct a new mesh from GSO meshes. This limits the distribution shift and preserves many local and global shape cues (e.g., surface smoothness, local symmetry), and also preserves a significant amount of shape meaningfulness to human perception. We show three variants: front-back, left-right, and top-bottom.

2) *Default CutMix*. We follow the CutMix [57] paper and randomly sample an axis-aligned 3D cube within the bounding cube of the object. We replace the part of mesh $M$ that falls into the cube with the part from mesh $N$ that falls into the same cube. When sampling the cube, we pin one of its corners at the corner of the object bounding cube to avoid significant discontinuity in the output shape. The length ratio ($\text{length}_{\text{sampled cube}}/\text{length}_{\text{bounding cube}}$) is uniformly sampled from $[0.4, 0.6]$. Most parts of the object $M$ are outside the chosen cube and remain intact. Meanwhile, the local shape cues are mostly preserved.

3) *CutMix by Octant*. We center each mesh and split it into 8 octants by the coordinate planes (XY, YZ, and XZ planes). Then we replace the part in each octant by the corresponding part from other random meshes from the same dataset. This variant still preserves the local shape cues, and it has a significantly smaller distribution gap than Zeroverse compared to our original evaluation data (GSO).

Results in Figure 5(a) show that all variants substantially hurt performance. Notably, CutMix by Octant causes a performance drop similar to Zeroverse despite a smaller domain gap. Standard CutMix, which modifies only about 1/8 of the mesh volume, still results in large drops (*e.g.*, 20 points for Hunyuan3D-2). Even minimal half-and-half perturbations, where shape meaningfulness is mostly preserved, typically cause performance drops of over 10 points, which is greater than for most other cues. This confirms that shape meaningfulness is a dominant cue for generalization.

**Geometric Cues (Style).** To explore geometric cues broadly, we apply style transfer to preserve semantics while altering geometric cues (Figure 2). Figure 5 summarizes the results on GSO and Toys4K. Performance significantly deteriorates under style perturbations. Sculpture-style images retain the most geometric information, thus yielding the smallest performance drops in general. Note that lower-performing methods show less degradation not because of their robustness, but due to metric saturation. Overall, semantics alone are insufficient; geometric cues are essential for reliable 3D inference.

**Shading and Texture.** We dissect geometric cues further by separately manipulating shading and texture, which are historically established cues for shape inference. Figure 5 presents evaluations across five texture conditions: original, checkerboard, Perlin noise, random Poly Haven texture [1], and solid gray, each tested with lighting (w/ L) and without lighting (w/o L). Surprisingly, altering texture while preserving shading minimally impacts performance for leading methods (SF3D, Hunyuan3D-2, Trellis). Multi-view approaches show slightly more sensitivity to texture changes but remain relatively robust overall. However, removing shading consistently decreases performance across methods, underscoring shading's significant role. Interestingly, there is an interplay between shading and texture cue: meaningful texture mitigates this drop due to removing lighting to some degree, especially on Toys4K. These results highlight that texture meaningfulness is not a necessary cue for generalizion. Meanwhile, shading is generally more influential than texture, with several top-performing methods exhibiting near texture invariance provided shading cues remain accurate.

**Silhouette and Occlusion.** Dilating object masks severely reduces performance despite unchanged interior pixels, indicating silhouette cues' importance. Trellis remains comparatively stable, suggesting a level of learned silhouette invariance. Occluding both silhouette and content dramatically reduces performance universally. This shows the combined importance of silhouette and interior visual cues.

**Edges.** We probe the role of edges cue in two ways, leaving only edges on silhouette and softening edges with localized gaussian filter. Edge-only input significantly degrade performance for most models except OpenLRM, suggesting edges alone may not provide sufficient shape information. Softening edges yields minor performance reductions, confirming edges are supportive but not primary cues.

**Perspective.** Switching from perspective to orthographic projection notably reduces performance, particularly for regression-based methods (OpenLRM, SF3D), indicating their reliance on perspective cues. CRM remains unaffected, since it uses orthographic images in training. Hunyuan3D-2 is more sensitive than Trellis, potentially due to its latent representation capturing perspective.

**Local Continuity.** Local cue scrambling significantly impacts regression-based SF3D, while other methods show varied but less severe sensitivity. Hunyuan3D-2 demonstrates the greatest robustness. However, all methods degrade substantially under extensive local scrambling, emphasizing the general importance of local continuity.

## 5 Discussions

**Correlation of Different Cues.** We choose our cues primarily based on their perceptual importance and interpretability to humans, rather than strict orthogonality. As noted in Section 2, our selected cues originate from psychological studies of human visual perception and have strong foundations in prior vision research, as they represent factors that humans typically find meaningful. While some cues naturally remain disentangled (*e.g.*, shading versus texture), others inherently overlap to some extent (*e.g.*, style with texture).

We assess whether cues impact objects similarly by calculating per-object performance drops in CD for each cue and then computing the Spearman rank correlation between pairs of cues. This produces a correlation matrix showing how similarly each pair of cues affects the same set of objects. We show

| | Texture | Shading | Silhouette | Occlusion | Edges | Local continuity | Style |
|---|---|---|---|---|---|---|---|
| Texture | 1.00 | 0.66 | 0.31 | 0.29 | 0.36 | 0.39 | 0.50 |
| Shading | 0.66 | 1.00 | 0.34 | 0.35 | 0.35 | 0.51 | 0.60 |
| Silhouette | 0.31 | 0.34 | 1.00 | 0.27 | 0.24 | 0.28 | 0.34 |
| Occlusion | 0.29 | 0.35 | 0.27 | 1.00 | 0.19 | 0.30 | 0.31 |
| Edges | 0.36 | 0.35 | 0.24 | 0.19 | 1.00 | 0.27 | 0.31 |
| Local cont. | 0.39 | 0.51 | 0.28 | 0.30 | 0.27 | 1.00 | 0.47 |
| Style | 0.50 | 0.60 | 0.34 | 0.31 | 0.31 | 0.47 | 1.00 |

Table 3: Analysis on the correlation of different cues. We present the Spearman rank correlations ($\rho$) between per-object performance drops in CD for each cue pair. Lower off-diagonal values indicate weaker similarity in object-wise effects; the diagonal is 1 by definition.

| Input Image | InstantMesh | SF3D | Trellis |
|---|---|---|---|
| Original | 54.5 | 61.6 | 39.6 |
| Line-art | 66.1 | 79.5 | 51.3 |
| Line-art + FLUX | 59.4 | 67.4 | 50.0 |

Table 4: Line-art-to-3D case study (lower is better). Adding explicit geometric cues to line-art narrows the gap to original images across three off-the-shelf image-to-3D models.

this result in Table 3. Importantly, this is a similarity-of-effect analysis. It does not test statistical independence nor guarantee disentanglement.

The result suggests that overall the correlation is low. Interestingly, texture and shading cues seem to affect a set of objects in similar ways, though they are inherently disentangled. These results also indicate that, while some appearance-related cues partially overlap, the cue effects are largely isolated at the level of object-wise impact.

**Practical Implications of Our Analysis.** To illustrate how our insights could inspire new research directions, we explore a line-art-to-3D problem: given a line-art image (in our case, extracted from GSO images), we aim to recover the underlying 3D shape. Pure line-art lacks surface appearance, and indeed leads to markedly worse 3D generation than original images. Inspired by our analysis, we enrich line-art with geometric cues by prompting an image diffusion model (Flux ControlNet [28]) conditioned on line-art to synthesize 3D rendering–style shading and texture. We then feed these cue-augmented images into off-the-shelf image-to-3D models (InstantMesh, SF3D, Trellis). As shown in Table 4, injecting geometric cues substantially improves performance, validating that our proposed insights could meaningfully contribute to future research in image-to-3D.

**Limitations.** While Cue3D provides a systematic and comprehensive analysis of cue importance across seven state-of-the-art methods and two widely used datasets, there remain several limitations. First, our study, though broad, is not exhaustive; evaluating a wider range of models and datasets would further strengthen our conclusions. Nevertheless, because our framework is both method and dataset-agnostic, it can be readily extended to additional settings. Second, our experiments focus on clean, object-centric datasets to minimize confounding factors, but extending the analysis to more diverse and nuanced real-world data could reveal additional insights. Third, although we primarily probe individual cues in isolation, understanding the interplay and correlation between multiple cues, beyond the initial shading-texture analysis presented here, remains an important direction for future work.

## 6 Conclusion

We introduce Cue3D, a model-agnostic framework for quantifying the influence of individual image cues in single-image 3D generation. We benchmark seven state-of-the-art methods across three major paradigms and two datasets in a unified approach. Then we apply targeted perturbations to individual cues like shading, texture, silhouette, occlusion, perspective, edges, and local continuity. We reveal that shape meaningfulness is crucial to the generalization of single-image 3D generation, while texture meaningfulness is not a necessary condition. Geometric cues are crucial, especially shading. Our analysis further shows that the models might be overly relying on silhouette cues, while perspective, edge, and local continuity cues affect reconstruction to varying degrees. We hope Cue3D and the insights presented here will deepen our understanding of how deep 3D networks leverage classical vision cues, and inspire future work on cue-aware architectures, robust training, and diagnostic perturbation tests for more transparent and controllable single-image 3D generation.

## References


[1] Poly Haven. https://polyhaven.com/, 2025. 5, 9

[2] Stability AI. Stable Zero123, 2023. 3

[3] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 1987. 2

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 2023. 4

[5] Tyler Bonnen, Stephanie Fu, Yutong Bai, Thomas O'Connell, Yoni Friedman, Nancy Kanwisher, Josh Tenenbaum, and Alexei Efros. Evaluating multiview object consistency in humans and image models. *NeurIPS*, 2024. 4

[6] Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. *CVPR*, 2025. 2, 3, 4, 5

[7] Maxwell B Clowes. On seeing things. *Artificial intelligence*, 1971. 3

[8] James E Cutting and Peter M Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Perception of space and motion*. 1995. 3

[9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3D objects. In *CVPR*, 2023. 2

[10] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *ICRA*, 2022. 2, 4

[11] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *CVPR*, 2024. 4

[12] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, 2017. 2, 4

[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2018. 4

[14] James J Gibson. The perception of the visual world. 1950. 2

[15] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[16] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM, 2023. 4

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2

[18] 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 *ICLR*, 2024. 2, 3

[19] Berthold KP Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. *MIT Tech. Rep.*, 1970. 2, 3

[20] Zixuan Huang, Stefan Stojanov, Anh Thai, Varun Jampani, and James M Rehg. Zeroshape: Regression-based zero-shot shape reconstruction. In *CVPR*, 2024. 3

[21] Katsushi Ikeuchi and Berthold KP Horn. Numerical shape from shading and occluding boundaries. *Artificial intelligence*, 1981. 3

[22] Bela Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 1981. 3

[23] Heewoo Jun and Alex Nichol. Shap-E: Generating conditional 3d implicit functions. *ArXiv*, 2023. 3

[24] Diederik P Kingma, Max Welling, et al. Auto-encoding variational Bayes, 2013. 3

[25] Jan J Koenderink and Andrea J Van Doorn. Surface shape and curvature scales. *Image and vision computing*, 1992. 2, 3

[26] Adam R Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Sona Mokrá, and Danilo Jimenez Rezende. Nerf-vae: A geometry aware 3d scene generative model. In *ICML*, 2021. 3

[27] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *IJCV*, 2000. 3

[28] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024. 10

[29] Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3D generation. In *ECCV*, 2024. 3

[30] Yushi Lan, Shangchen Zhou, Zhaoyang Lyu, Fangzhou Hong, Shuai Yang, Bo Dai, Xingang Pan, and Chen Change Loy. Gaussiananything: Interactive point cloud latent diffusion for 3D generation. In *ICLR*, 2025. 3

[31] Michael S Landy, Laurence T Maloney, Elizabeth B Johnston, and Mark Young. Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision research*, 1995. 2

[32] Aldo Laurentini. The visual hull concept for silhouette-based image understanding. *TPAMI*, 1994. 3

[33] Weiyu Li, Rui Chen, Xuelin Chen, and Ping Tan. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d. *ArXiv*, 2023. 3

[34] Xiang Li, Zixuan Huang, Anh Thai, and James M Rehg. Symmetry strikes back: From single-image symmetry detection to 3D generation. In *CVPR*, 2025. 5

[35] Minghua Liu, Chong Zeng, Xinyue Wei, Ruoxi Shi, Linghao Chen, Chao Xu, Mengqi Zhang, Zhaoning Wang, Xiaoshuai Zhang, Isabella Liu, et al. Meshformer: High-quality mesh generation with 3d-guided reconstruction model. *ArXiv*, 2024. 3

[36] 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 *ICCV*, 2023. 3

[37] 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 *CVPR*, 2024. 3

[38] Jitendra Malik and Ruth Rosenholtz. Computing local surface orientation and shape from texture for curved surfaces. *IJCV*, 1997. 2

[39] Alex P Pentland. Shading into texture. *Artificial Intelligence*, 1986. 3

[40] Vilayanur S Ramachandran. Perception of shape from shading. *Nature*, 1988. 2

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[42] Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 1971. 3

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

[44] Elizabeth Spelke, Sang Ah Lee, and Véronique Izard. Beyond core knowledge: Natural geometry. *Cognitive science*, 2010. 3

[45] Stefan Stojanov, Anh Thai, and James M Rehg. Using shape to categorize: Low-shot learning with an explicit shape bias. In *CVPR*, 2021. 2, 4

[46] 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 *ECCV*, 2024. 2, 3, 4

[47] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, 2019. 4

[48] Sebastian Thrun and Ben Wegbreit. Shape from symmetry. In *ICCV*, 2005. 4

[49] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *ArXiv*, 2024. 3

[50] Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *NeurIPS*, 2022. 3

[51] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018. 3

[52] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3D textured mesh with convolutional reconstruction model. In *ECCV*, 2024. 2, 3, 4

[53] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *CVPR*, 2025. 2, 3, 4

[54] Desai Xie, Sai Bi, Zhixin Shu, Kai Zhang, Zexiang Xu, Yi Zhou, Sören Pirk, Arie Kaufman, Xin Sun, and Hao Tan. Lrm-zero: Training large reconstruction models with synthesized data. *ArXiv*, 2024. 2, 5, 8

[55] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *ArXiv*, 2024. 2, 5

[56] 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*, 2024. 2, 3, 4

[57] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 8

[58] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3D assets. *ACM TOG*, 2024. 3

[59] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *TPAMI*, 1999. 2

[60] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3D 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *ArXiv*, 2025. 2, 3, 4, 5

[61] Yefan Zhou, Yiru Shen, Yujun Yan, Chen Feng, and Yaoqing Yang. A dataset-dispersion perspective on reconstruction versus recognition in single-view 3D reconstruction networks. In *3DV*, 2021. 4