Multistable Shape from Shading Emerges from Patch Diffusion

Xinran Nicole Han
Harvard University
xinranhan@g.harvard.edu

Todd Zickler
Harvard University
zickler@seas.harvard.edu

Ko Nishino
Kyoto University
kon@i.kyoto-u.ac.jp

Abstract

Models for inferring monocular shape of surfaces with diffuse reflection—shape from shading—ought to produce distributions of outputs, because there are fundamental mathematical ambiguities of both continuous (e.g., bas-relief) and discrete (e.g., convex/concave) types that are also experienced by humans. Yet, the outputs of current models are limited to point estimates or tight distributions around single modes, which prevent them from capturing these effects. We introduce a model that reconstructs a multimodal distribution of shapes from a single shading image, which aligns with the human experience of multistable perception. We train a small denoising diffusion process to generate surface normal fields from 16×16 patches of synthetic images of everyday 3D objects. We deploy this model patch-wise at multiple scales, with guidance from inter-patch shape consistency constraints. Despite its relatively small parameter count and predominantly bottom-up structure, we show that multistable shape explanations emerge from this model for ambiguous test images that humans experience as being multistable. At the same time, the model produces veridical shape estimates for object-like images that include distinctive occluding contours and appear less ambiguous. This may inspire new architectures for stochastic 3D shape perception that are more efficient and better aligned with human experience.

1 Introduction

From chiaroscuro in Renaissance paintings to the interplay of light and dark in Ansel Adams' photographs, humans are masters at perceiving three-dimensional shape from variations of image intensity—shading—from a single image alone. Even though our visual experience is dominated by everyday objects, our perception of shape from shading generalizes to many synthetic "non-ecological" images invented by vision scientists. Some of these images have ambiguous (e.g., convex/concave) interpretations and lead to multistable perceptions, where one's impression of 3D shape alternates between two or more competing explanations. Figure 1 shows an example adapted from [31], which is alternately interpreted as an indentation or a protrusion. Both explanations are physically correct because the same image can be generated from either shape under different lighting conditions.

How can a computational model acquire this human ability to capture multiple underlying shape explanations? An algorithmic suggestion comes from Marr's principle of least commitment [34, 35], which requires not doing anything that may later have to be undone. But this is in contrast to many computer vision methods for shape from shading, including SIRFS [4] and recent neural feed-forward models [53, 56], which are deterministic and produce a single, best estimate of shape. These types of models commit to one explanation based on priors that are either designed or learned from a dataset, and they cannot express multiple interpretations of an ambiguous image. They are unlikely to be good models for the mechanisms that underlie multistable perception.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

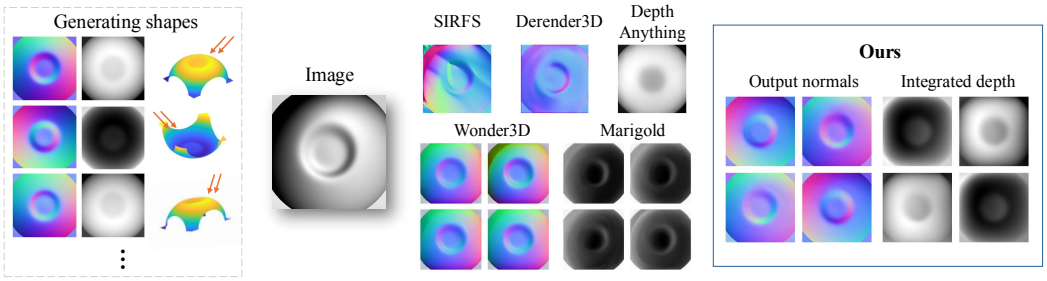


Figure 1: Many shapes (left) can explain the same image (middle) under different lighting, including flattened and tilted versions and convex/concave flips. The concave/convex flip in this example is also perceived by humans, often aided by rotating the image clockwise by 90 degrees. Previous methods for inferring either surface normals (SIRFS [4], Derender3D [53], Wonder3D [33]) or depth (Marigold [27], Depth Anything [56]) produce a single shape estimate or a unimodal distribution. Ours produces a multimodal distribution that matches the perceived flip. (Image adapted from [31].)

Instead, we approach monocular shape inference as a conditional generative process, and inspired by a long history of shape from shading with local patches [18, 25, 9, 29, 54, 20, 17], we present a bottom-up, patch-based diffusion model that can mimic multistable perception for ambiguous images of diffuse shading. Notably, our model is trained using images of familiar object-like shapes and has no prior experience with the ambiguous images that we use for testing. It is built on a small conditional diffusion process that is pre-trained to predict surface normals from 16×16 image patches. When we apply this patch process at multiple scales with inter-patch shape consistency constraints, and when we coordinate the sampling across patches, the model ends up capturing global ambiguities that are very similar to those experienced by humans.

An important attribute of our model is the way it handles lighting. It builds on the mathematical observation that shape perception can precede lighting inference [30]. It also adheres to the philosophy that inferred lighting cannot, and should not, be precise because of spatially-varying effects like global illumination [50]. Our model achieves these aims by guiding its diffusion sampling process with a very weak constraint on lighting consistency, where each patch nominates a dominant light direction and then all patches enact their own concave/convex flips in response to those nominations.

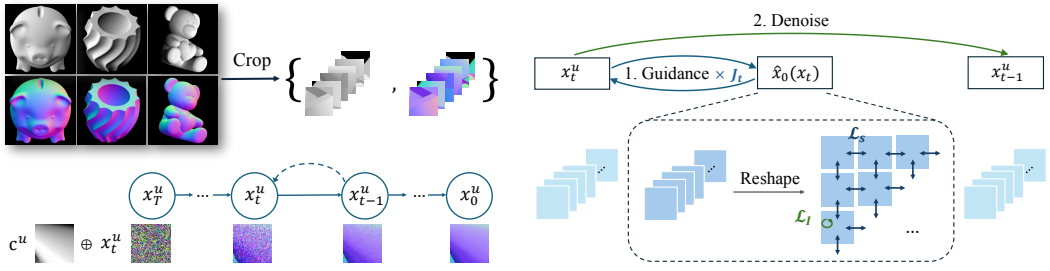
Another critical aspect of our model is a diffusion sampling process that is coordinated across multiple scales. It involves spatially resampling the normal predictions at intermediate diffusion time steps and then adding noise before resuming the diffusion at the resampled resolutions. Our approach is inspired by previous work that uses a "V-cycle" (fine-coarse-fine) to avoid undesirable local extrema during MRF optimization [36]. Our ablation experiments show that multi-scale sampling is crucial for finding good shape explanations that are globally consistent.

We train our patch diffusion model on images of objects like those in Fig. 2a, and we find experimentally that it can generalize to new objects as well as to images like Fig. 1 which are quite different from the training set and appear multistable to humans. This is in contrast to previous diffusion-based monocular shape models [27, 33] which cannot capture multistability and produce output that is much less diverse. Our model is also extremely efficient, only requiring a small pixel-based diffusion UNet that operates on 16×16 patches. Our total model weights require only 10MB of storage, much less than the 2-3GB required by some of the previous (and more general) models we compare to.

2 Background

2.1 Ambiguities in Shape from Shading

Shape from shading is a classic reconstruction problem in computer vision. Since being formulated by Horn in the 1970s [23] there have been many approaches to tackle it, often by assuming diffuse Lambertian shading and uniform lighting from either a single direction (e.g.,[43, 16]) or as low-order spherical harmonics (e.g., [4, 59]). Almost all methods either require the lighting to be known (e.g., for natural illumination [41]) or try to estimate it precisely via inverse rendering during the optimization process [54, 59, 53], and many approaches rely on a set of priors to constrain the possible



(a) Training (b) Inference (single scale)

Figure 2: Training patches are cropped from synthetic images of ordinary diffuse objects, and during training, a small diffusion model learns to denoise the normal field x_0^u for patch u from a random sample x_T^u conditioned on the patch intensities c^u. During inference, the model is applied in parallel to non-overlapping patches, with guidance from inter-patch shape-consistency constraints to minimize the curvature smoothness loss $\mathcal{L}_S$ and integrability loss $\mathcal{L}_I$.

search space. For instance, SIRFS [4] uses different lighting priors for natural versus laboratory conditions and a surface normal prior along occluding contours. Recent deep learning approaches like Derender3D [53] have demonstrated impressive results without being limited to Lambertian reflectance, but they similarly rely on priors internalized from their training sets and have trouble generalizing to new conditions.

The main challenge of shape from diffuse shading comes from the many levels of inherent ambiguity. At a single Lambertian point, when lighting is unknown, there is a multi-dimensional manifold of surface orientations and curvatures that are consistent with the spatial derivatives of intensity at the point [29, 20]. Even when lighting and surface albedo are known at a point, there is a cone of possible normal directions. At the level of a quadratic surface patch, when lighting is unknown, there is a discrete four-way ambiguity corresponding to convex, concave, and saddle shapes [54, 29]. At a global level, when lighting and surface albedo are known, ambiguities arise from interpreting the Lambertian shading equation as a PDE (e.g., [6]) or as a system of polynomial equations [13]. And when lighting and albedo are unknown, there is an additional three-parameter global ambiguity that corresponds to flattenings and tiltings of the global shape [5]. Finally, when lighting is unknown, a global shape has a discrete counterpart that corresponds to a global convex/concave flip.

It is important to note that all of these mathematical ambiguities are based on certain idealized models for the image formation process, such as exact Lambertian shading, perfectly uniform albedo, and most commonly, perfectly uniform lighting that ignores global illumination effects such as interreflections and ambient occlusion, which in reality have substantial effects [50]. An advantage of a stochastic, learning-based approach, like the one presented here, is the potential to capture all of these ambiguities as well as others that have not yet been discovered or characterized.

2.2 Denoising Diffusion with Guidance

Diffusion probabilistic models [22](DDPM) generate data by iteratively denoising samples from a Gaussian (or other) pre-determined distribution. We build on a conditional denoising diffusion model, where the condition c is a grayscale image patch, and the model is designed to approximate the distribution $q(x_0|c)$ on 3-channel normal maps x_0 with a tractable model distribution $p_\theta(x_0|c)$. A 'forward process' adds Gaussian noise to a clean input x_0 and is modeled as a Markov chain with Gaussian transitions for timesteps $t = 0, 1, \cdots, T$. Each step in the forward process adds noise according to $q(x_t|x_{t-1}, c) := \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$, where $\{\beta_t\}$ is a predetermined noise variance schedule. The intermediate noisy input x_t can be written as

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\omega, \quad \omega \sim \mathcal{N}(0, \mathbf{I}), \quad \text{where } \alpha_t := \prod_{s=1}^{t}(1-\beta_s)\,. \tag{1}$$

The 'reverse process' $q(x_{t-1}|x_t, c)$ is modeled by a learned Gaussian transition $p_\theta(x_{t-1}|x_t, c) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t; c), \sigma_t^2\mathbf{I})$. The mean value $\mu_\theta(x_t, t; c)$ can be expressed as a combination of the noisy image x_t and a noise prediction $\epsilon_\theta(x_t, t; c)$ from a learned model. The noise prediction model

θ can be trained by minimizing the prediction error

$$L(\theta) := \mathbb{E}_{x_0,\omega\sim\mathcal{N}(0,\mathbf{I})}\left[||\omega - \epsilon_\theta(x_t, t; c)||_2^2\right], \tag{2}$$

as shown in [22]. To sample noiseless data using the learned model, we start from an initial random Gaussian noise seed and use the learned denoiser ϵ_θ to compute $x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t; c)) + \sigma_t z$ iteratively, where $z \sim \mathcal{N}(0, \mathbf{I})$, which is a stochastic process.

The denoising diffusion implicit model (DDIM) shows that the reverse procedure can be made *deterministic* by modeling it as a non-Markovian process with the same forward marginals [48]. This approach helps to accelerate the sampling process by using fewer steps and also provides an estimate of the predicted $\hat{x}_0$ at each timestep with x_t. Each denoising step combines noise and a *foreseen* denoised version

$$x_{t-1} = \sqrt{\alpha_{t-1}} f_\theta(x_t, t; c) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(x_t, t; c)\,, \tag{3}$$

where

$$f_\theta(x_t, t; c) = \hat{x}_0(x_t) = \frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t; c)}{\sqrt{\alpha_t}} \tag{4}$$

is the predicted $\hat{x}_0$ at reverse sampling step t.

The DDIM formulation provides a way to 'guide' the process of sampling x_{t-1} from x_t by applying additional constraints or losses to the predicted $\hat{x}_0(x_t)$ at intermediate sampling steps. Previous work has used similar approaches to solve inverse problems [8, 47] or to combine outputs from multiple diffusion models for improved perceptual similarity [32]. Guided denoising is achieved with

$$x_t' = x_t - \eta_t \nabla_{x_t}\mathcal{L}(\hat{x}_0(x_t))\,, \tag{5}$$

$$x_{t-1} = \sqrt{\alpha_{t-1}}\hat{x}_0(x_t') + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(x_t, t; c)\,, \tag{6}$$

in each sampling subroutine from time t to $t-1$, where η_t is the (possibly time-dependent) step size of the guided gradient update. The first step (5) can be repeated multiple times before applying the denoising step (6).

3 Multiscale Patch Diffusion with Guidance

Consider a surface represented by a differentiable height function $h(x, y)$ over image domain (x, y) viewed by parallel projection from above. The image plane is sampled on a square grid (pixels). Image patches have size $d \times d$ and are indexed by u, and we denote them as $c^u \in [0, 1]^{d\times d}$.

Training occurs on patches extracted from images of everyday objects, as depicted in Fig. 2a. For each image patch c^u there is a corresponding patch normal field $x_0^u \in [-1, 1]^{3\times d\times d}$, whose (i, j)th spatial element represents a surface normal vector via

$$\frac{x_0^u(i,j)}{\|x_0^u(i,j)\|} = \frac{(-p_{i,j}, -q_{i,j}, 1)}{||(-p_{i,j}, -q_{i,j}, 1)||} = n_{i,j} \in \mathbb{S}^2, \tag{7}$$

where $(p, q) = (\partial h/\partial x, \partial h/\partial y)$ are the surface derivatives. Training proceeds as described in Sec. 2.2, with a dataset of patch tuples (c^u, x_0^u) and a UNet $\epsilon_\theta(x_t^u, t; c^u)$ similar to that in [22] whose four-channel input is the concatenation $c^u \oplus x_t^u$. (Model and training details are in Appendix A.2.)

As depicted in Fig. 2b, inference occurs over images c of size $H \times W$ which are divided into their collections of non-overlapping patches c^u. We assume that the underlying global normal field $x_0 \in [-1, 1]^{3\times H\times W}$ is continuous including at most locations on the seams between the patch normal fields x_0^u. We formulate the prediction of global field x_0 as reverse conditional diffusion on an undirected, four-connected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$. Each patch $x_0^u, u \in \mathcal{V}$ is a node, and there are edges $\{u, v\} \in \mathcal{E}$ between pairs of patches that are horizontally or vertically adjacent.

To encourage the patch fields to form a globally coherent prediction, we use guidance as described by Eqs. 5 and 6. Our guidance includes two terms:

$$\mathcal{L}(\hat{x}_0) = \frac{1}{|\mathcal{E}|}\sum_{\{u,v\}\in\mathcal{E}} \mathcal{L}_S(\hat{x}_0^u, \hat{x}_0^v) + \lambda\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathcal{L}_I(\hat{x}_0^v)\,, \tag{8}$$

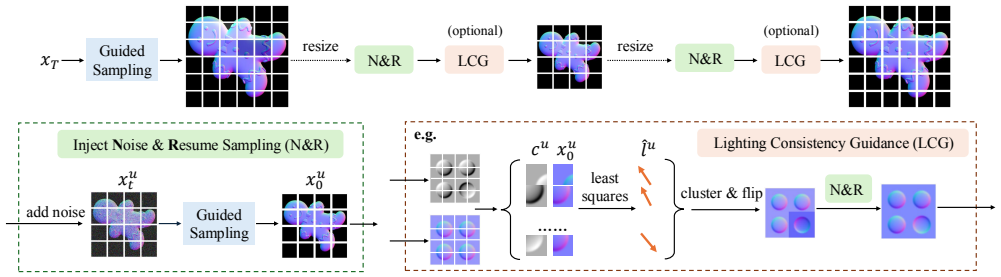


Figure 3: *Top*: Illustration of multiscale sampling across two scales in a fine-coarse-fine "V-cycle", with conditional images omitted for simplicity. In practice, our V-cycle covers more than two scales. *Left*: The $N\&R$ subroutine injects noise to an earlier timestep $0 < t < T$ and then resumes guided sampling (Fig. 2b) at that scale. *Right*: Optional intermediate guidance comes from lighting consistency (LCG), where each patch nominates a dominant light direction and then some patches flip in response to those nominations. Pseudocode is in the appendix.

where $\mathcal{L}_I$ is a within-patch continuity term that encourages integrability of the normal fields over small pixel loops, and $\mathcal{L}_S$ is an inter-patch spatial consistency term that encourages constant curvature across the seams between patches. Hyperparameter $\lambda \in (0, 1]$ controls the relative weighting of the two terms, and $\eta_t \geq 0$ in Eq. 5 determines the overall guidance strength.

The **integrability** term follows Horn and Brooks [24] by penalizing deviation from a discrete approximation to the integrability of surface normals, i.e., $\partial p/\partial y = \partial q/\partial x$, over 2×2 loops of pixels. We write this as

$$\mathcal{L}_I(\hat{x}_0^u) = \sum_{i,j} (p_{i,j+1} - p_{i+1,j+1} + p_{i,j} - p_{i+1,j} + q_{i,j+1} + q_{i+1,j+1} - q_{i,j} - q_{i+1,j})^2, \quad (9)$$

where the summation is over the i, j grid-indexed pixels in patch u, and p, q are the components of $\hat{x}_0^u$ implied by Eq. 7.

The **spatial consistency** term penalizes deviation from constant surface curvature across each seam $\{u, v\} \in \mathcal{E}$ in the direction perpendicular to the seam. Consider four consecutive normals in the perpendicular direction n_1, n_2, m_1, m_2 where n_i belong to patch u and m_i belong to patch v. We penalize the absolute angular difference between m_1 in v and its extrapolated estimate using normals in u, i.e., $n_2 + (n_2 - n_1)$. Making this symmetric gives

$$\left\| \cos^{-1}\left(m_1 \cdot \big(n_2 + (n_2 - n_1)\big)\right) \right\| + \left\| \cos^{-1}\left(n_2 \cdot \big(m_1 - (m_2 - m_1)\big)\right) \right\|, \quad (10)$$

which we sum over the length of the seam to obtain $\mathcal{L}_S(\hat{x}_0^u, \hat{x}_0^v)$.

3.1 Dominant Global Lighting Constraint

Our guidance so far enforces global coherence, but even globally coherent surfaces can contain regions that independently undergo convex/concave flips without affecting their surround [31]. The top row of Fig. 5 provides a familiar example. One can hide any three of the bumps/dents and then perceive the fourth as being either concave or convex. Yet, when one is allowed to examine the image as a whole and rotate it upside down, instead of perceiving $2^4 = 16$ interpretations, one sees only two: the bottom-right element is a bump (or dent) and the other three are its opposite. This behavior is explained by lighting. If one expects the dominant light direction to be similar everywhere on the surface, the four flips become tied together.

We can incorporate this notion of dominant light consistency into our model using an additional discrete guidance step, as depicted in the bottom right of Fig. 3. This step can be applied to any global sample $\hat{x}_0$ and has three parts: (*i*) patches $\hat{x}_0^u$ independently nominate dominant light directions $\hat{l}^u$; (*ii*) we identify a single direction $\hat{l}$ that is most common among these nominations; and (*iii*) some patches perform a concave/convex flip to become more consistent with $\hat{l}$.

Specifically, each patch $\hat{x}_0^u$ that is not too close to being planar (i.e., that has non-constant $\hat{x}_0^u(i)$) nominates its dominant light direction by computing the least-squares estimate according to shadowless

Lambertian shading:

$$\hat{l}^u = \arg\min_{l \in \mathbb{R}^3} \sum_i \left(c^u(i) - \frac{\langle \hat{x}_0^u(i), l \rangle}{\|\hat{x}_0^u(i)\|} \right)^2 . \tag{11}$$

We create two clusters in the set $\{\hat{l}^u\}$ using k-means and choose the center of the majority cluster as the dominant global direction $\hat{l}$. Each patch u in the minority cluster undergoes a concave/convex flip $(p, q) \to (-p, -q)$ by multiplying (-1) with the first two channels of $\hat{x}_0^u$. Since the independent flips can cause discontinuities at patch seams, we always follow this discrete lighting guidance step by an *inject Noise & Resume sampling* (N&R) subroutine, where we add noise to an intermediate timestep via Eq. 1 and resume spatially-guided denoising from that timestep. Pseudocode is provided in Appendix A.1.

This approach to lighting consistency has several advantages. Unlike many previous computational approaches to shape from shading, it does not assume the lighting to be known beforehand. Nor does it require the lighting to be exactly spatially uniform across the surface, which provides some resilience to global illumination effects. It imposes no prior on the dominant light direction (e.g., 'lighting from above'), but one can imagine extending it to do so. And because our patch-based framework can be applied with or without lighting consistency guidance (see Appendix A.3), it may provide a mechanism in the future for modeling the way in which humans selectively enforce lighting consistency across an image [7, 40].

3.2 Multiscale Optimization

Since each local image patch can be explained by either concave or convex shapes, the terms in the spatial guidance energy (Eq. 8) are multimodal, and finding a global minimum is computationally difficult. Our experiments, like the one in Fig. 4, show that optimization at a single scale with random initial noise and gradient-descent guidance often gets trapped in poor local minima. To overcome this, and also to fully leverage shading information from various spatial frequencies and scales, we draw inspiration from work on Markov random fields [36] and introduce a multiscale optimization scheme. This is possible because our patch diffusion UNet and guidance can be applied to any resampled resolution $(sH) \times (sW)$ that is divisible by patch size d.

Our multiscale optimization occurs in a "V-cycle", a sequence of fine-coarse-fine resolutions. We begin by applying guided denoising at the highest image resolution. Then, we downsample the predicted global field to a lower resolution before injecting noise and resuming reverse sampling ($N\&R$) at that lower resolution. As depicted in the left of Fig. 3, this has the effect of generating a random sample at the lower resolution that is informed by a previous sample at the higher resolution. A similar process occurs when going from coarse to fine, but with the global field being upsampled before applying the $N\&R$ subroutine.

To further reduce discontinuities at seams, we find it helpful to store and fuse the global field estimates from the final few resolutions of the fine-coarse-fine cycle. We do this by computing their (p, q) fields via Eq.7, resampling them to the highest resolution and averaging them, and then converting the average back to a normal field.

4 Experimental Results

The input to our patch UNet is the concatenation of c^u and x_t^u. It has 4 channels and spatial dimension $d \times d$ with $d = 16$. We train it using patches of size $d \times d$ extracted from rendered images of the 3D objects in [26] curated from Adobe Stock. We use Lambertian shading from random light directions, with a random albedo in $[0.5, 1]$ and without cast shadows or global illumination effects. Our dataset contains around 8000 images (256×256) of 400 unique objects. Some examples are shown in the left of Fig. 2. The images contain occluding contours, and for empty background pixels i, j we set $x_0(i, j) = (-1, -1, -1)$. We augment the training data by creating two convex/concave copies of each patch field x_0^u that does not contain any background. At inference time, we use the DDIM sampler [48] with 50 sampling steps and with guidance. Additional details are in the appendix.

For comparison, we consider three deterministic approaches and two stochastic models. SIRFS [4] and Derender3D [53] are deterministic inverse-rendering models that estimate lighting and reflectance

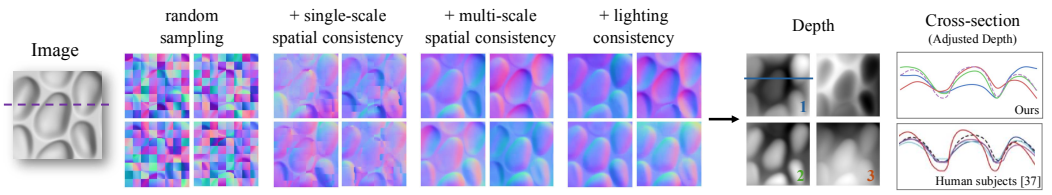


Figure 4: Ablations, and comparison to human subjects using image and psychophysics data from [37]. *Left:* Ablations demonstrate the importance of each component. *Right*: Depth cross-sections extracted from four (integrated) samples from the convex mode of our full model exhibit relief-like variations similar to those reported across human subjects. (The dashed line is the depth that was used to render the input image.)

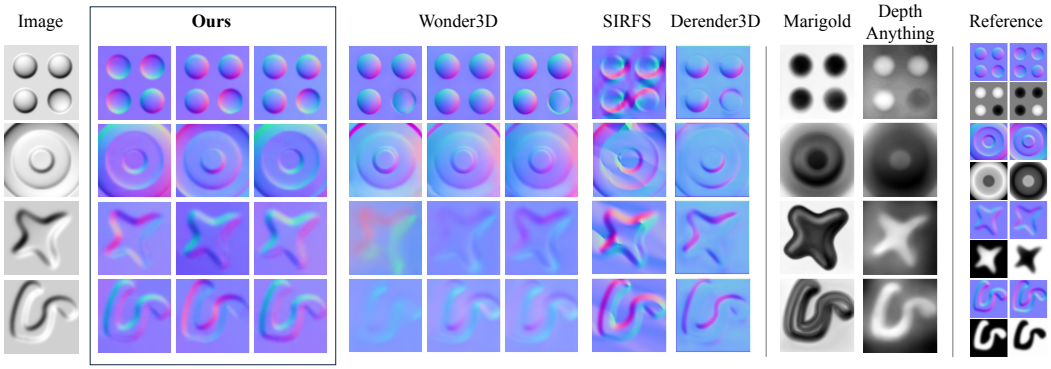


Figure 5: Normals produced by our model for various synthetic test surfaces rendered with directional light sources. For depth maps, brighter is closer. "Reference" depicts the shapes—each with a convex/concave counterpart—that were used to render the input images. We find that our reconstructions are more accurate and diverse than other methods.

together with shape. They are among the few recent models that do not rely critically on having input occluding contour masks. Depth Anything [56] is a recent learning-based deterministic model for monocular depth estimation. It leverages a DINOv2 encoder [39] and a DPT decoder [44] and is trained for depth regression using 62M images. For comparisons to stochastic models, we include Marigold [27] which is derived from Stable Diffusion [45] and is fine-tuned for depth estimation. We also include Wonder3D [33], which likewise leverages a prior based on Stable Diffusion. Wonder3D is trained to generate consistent multi-view normal estimates on more than 30k 3D objects, and it achieves state-of-the-art results on 3D reconstruction benchmarks [12].

4.1 Ablation Studies

Figure 4 analyzes the key components of our model using a crop of a shape and image from the lab of James Todd [37] (the complete image is in Appendix A.5). The left of the figure shows that when each patch is reconstructed independently, the resulting normals are inconsistent, because each patch may choose a different concave/convex mode as well as its various flattenings and tiltings. When spatial consistency guidance is applied at one scale, the global field is more consistent but suffers from discontinuous seams due to poor local minima. With multiscale sampling the seams improve, but separate bump/dent regions can still choose different modes without being consistent with any single dominant light direction. Finally, when lighting consistency is added, the output fields become more concentrated around two global modes—one that is globally convex (lit primarily from below) and another that is globally concave (lit primarily from above).

In the right of the figure, we compare samples from our model to depth profiles that were labeled by humans on the same image [37]. (These results have appeared previously in [31].) We generate four samples from our model's globally-convex mode and integrate them to depth maps using [15]. Their cross-sections exhibit variations with similar qualitative structure.

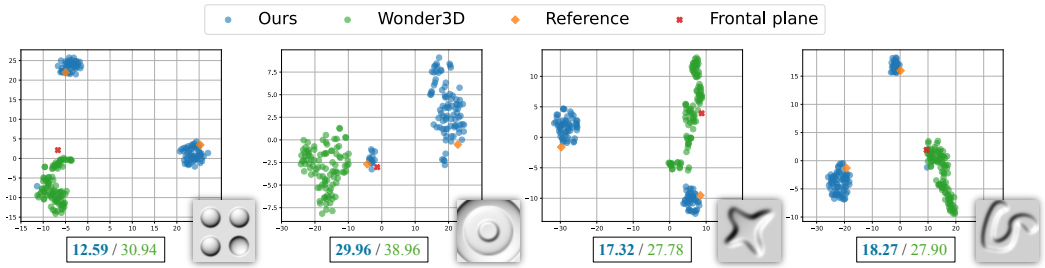


Figure 6: t-SNE visualizations of normal field samples produced by our model and by Wonder3D. Plots depict 100 samples from each model, along with the two mathematical possibilities (under directional light) and the normals of a trivial frontal plane. For each model we report the Wasserstein distance (smaller is better) between its samples and the reference distribution, which is uniform over two possibilities. Our model is more accurate and in all cases covers both possibilities.

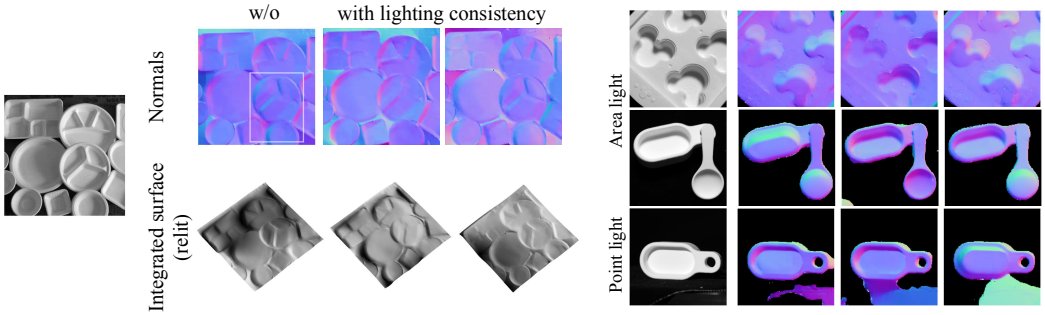


Figure 7: Sampled reconstructions for real images. (a) For the 'plates' image from [57], regions such as the indicated box can exhibit independent convex/concave flips when lighting consistency is not used; but when lighting consistency is enforced, only two global modes emerge. (b) Sampled reconstructions for some multistable images we captured with illumination from a point or area light. (Rotate them by 180° to enhance the alternative experience.) Note that half of the object in the first row was painted matte, and its other half was left glossy. Despite being trained entirely on synthetic data under idealized lighting, the model exhibits some generalization by producing plausible multistable outputs for these captured scenes.

4.2 Ambiguous Images

Figure 5 shows results from our full model for images and shapes that we intentionally design to be ambiguous, using insight from [31]. Each one can be perceived as either convex or concave, as shown in the right-most column (Reference). Samples from our model clearly demonstrate the effectiveness of our model in terms of both coverage and accuracy of the possible shapes. In contrast, we find that the two models derived from Stable Diffusion (Wonder3D and Marigold) provide less accuracy on these images, and that, on average, they tend to have a 'lighting from above' prior baked in. For instance, they tend to interpret the third and fourth row as concave, while Depth Anything [56] interprets them as convex. Additional results are included in the appendix.

Figure 6 visualizes distributions of shape reconstructions as 2D t-SNE plots (with perplexity equal to 30) by sampling 100 random seeds for our model and for Wonder3D. For reference, we also plot the t-SNE embeddings of a frontal plane, $\hat{x}_0(i, j) = (0, 0, 1)$ and of the two reference shapes. Our model covers both reference shapes whereas Wonder3D either covers only one or is close to a plane. These differences in coverage and accuracy are also apparent in terms of Wasserstein distance.

4.3 Real Images

We evaluate our model using a few different categories of captured images. In each case, we resize the image to 256×256 (to accommodate the restrictions of some of the previous models), and we use the multiscale schedule described in Appendix A.10.

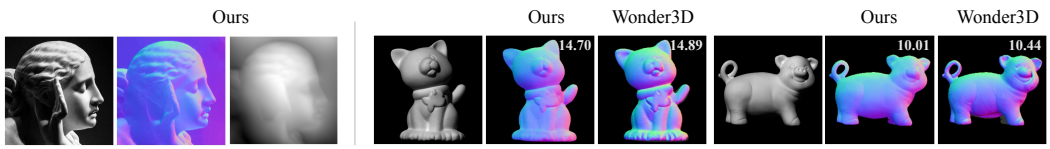


Figure 8: *Left*: Reconstructed normals and integrated depth for an image taken from the web. *Right*: Reconstructions for images in the SfS dataset of [54] with median angular error from the ground truth normals (lower is better). Our model's accuracy is on par with the best existing methods. Additional quantitative results are in the appendix.

Captured ambiguous images. Inspired by the 'plates' image in [58], we captured a handful of images that induce multistable perceptions for human observers. We captured these images with a Sony α7SIII camera under lighting from area or point light sources, and Fig. 7b shows some examples. We find that our model's multimodality is qualitatively well aligned with perceptual multistability.

Shape from shading dataset [54]. This dataset contains diffuse objects captured with directional light sources and a dark background, along with ground truth surface normals for measuring accuracy. The right of Fig. 8 shows that our model produces normal maps that are on par with previous methods, even without knowledge of light source directions. Additional quantitative results are in Appendix A.8

Internet and astronomical images. The left of Figure 8 shows that our model can produce a detailed and plausible shape estimate for a tone-mapped sRGB image taken from the web [2]. Appendix A.6 includes two satellite images of the surface of Mars and shows that our model reproduces the so-called crater illusion.

5 Related Work

Given the recent success of diffusion models in generating realistic images [22, 49, 28], many works have explored the power of patch-based diffusion, including for generating high-resolution panoramic images [3, 32]. Our method also leverages patch diffusion, but it departs from these work in two key ways. Unlike [3, 32], we do not generate patches auto-regressively or require an anchor patch. Instead, we simultaneously guide all patches (e.g., 100 patches for a 160×160 image) toward a coherent output using spatial consistency guidance. A second distinction is that we do not provide a global condition such as text to each individual patch. Instead, each patch is conditioned only on the corresponding crop of the input grayscale image, which is why we call it a bottom-up architecture. It shares the same spirit as previous work on inverse lighting [14], which also uses a bottom-up architecture to produce a variety of explanations that can then be integrated with top-down information.

Recently, Wang et al. [52] introduced a patch-based diffusion training framework that incorporates patch coordinates to reduce training time and storage cost. Patch-based diffusion has also been used for other tasks. Ozdenizci et al. [42] use overlapping patch diffusion to restore images in adverse weather, and Ding et al. [11] use it to synthesize images in higher resolution. All of these works use fairly large patches (e.g., 64×64) and some of their components, such as feature-averaging or noise-averaging, are not appropriate for our shape from shading problem because of its inherent multi-modality. A convex sample and a concave sample cannot simply be averaged to improve the output. These challenges motivate the novel features of our model, including global consistency guidance and multiscale sampling.

6 Conclusion

Inspired by the multistable perception of ambiguous images, and by mathematical ambiguities in shape from shading, we introduce a diffusion-based, bottom-up model for stochastic shape inference. It learns exclusively from observations of everyday objects, and then it produces perceptually-aligned multimodal shape distributions for images that are different from its training set and that appear ambiguous to human observers. A critical component of our model is a sampling scheme that

operates across multiple scales. Our model also provides compositional control: global lighting consistency can be turned on or off, thereby controlling whether regional bumps/dents can each undergo concave/convex flips independently. Our findings motivate the exploration of other multiscale stochastic architectures, for a variety of computer vision tasks. They may also help improve the understanding and modeling of human shape perception.

Limitations A key limitation is our restriction to textureless and shadowless Lambertian shading. While this restriction is common in theoretical work [29, 54, 20] and useful for creating ambiguous images, it is well-known that many ambiguities disappear in the presence of other cues such as glossy highlights, cast shadows, and repetitive texture. Also, since our model is predominantly bottom-up, it suffers when large regions of an image are covered by cast shadows (e.g., the shoulder region in Fig. 8). These types of regions often require non-local context like object recognition in order to be accurately completed. Incorporating more diverse materials (e.g., as in [14]) and top-down signals into our model are important directions for future research.

Another limitation of our model stems from its sequential V-cycle approach to multiscale sampling. It scales linearly with the number of resolutions, which is likely to be improved by optimization or parallelization that increases runtime efficiency. Also, since our multiscale approach is training free, it requires a manual search to identify a good schedule. Similar to previous work that restarts the sampling process from intermediate timesteps [55], ours also require choosing the timestep at which to resume sampling. Overall, further analysis is needed to better understand the structure of our model's latent space, and to discover more efficient and general approaches to multiscale generation.

Acknowledgments

We thank Jianbo Shi for helpful discussions and Steven Zucker for suggesting the crater illusion. We also thank James Todd, Benjamin Kunsberg, Steven Zucker and William Smith for kindly sharing their images and perceptual stimuli. This work was in part supported by JSPS 20H05951, 21H04893, and JST JPMJAP2305. It was also in part supported by the NSF cooperative agreement PHY-2019786 (an NSF AI Institute, http://iaifi.org).

References

[1] Adobe Stock. `https://stock.adobe.com/`.

[2] Nydia, the Blind Flower Girl of Pompeii. Randolph Rogers. Photo by Zack Jarosz. `https://www.pexels.com/photo/woman-statue-1727658/`.

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In *International Conference on Machine Learning*, pages 1737–1752. PMLR, 2023.

[4] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687, 2014.

[5] Peter N Belhumeur, David J Kriegman, and Alan L Yuille. The bas-relief ambiguity. *International Journal of Computer Vision*, 35(1):33–44, 1999.

[6] Michael Breuß, Emiliano Cristiani, Jean-Denis Durou, Maurizio Falcone, and Oliver Vogel. Perspective shape from shading: Ambiguity analysis and numerical approximations. *SIAM Journal on Imaging Sciences*, 5(1):311–342, 2012.

[7] Patrick Cavanagh. The artist as neuroscientist. *Nature*, 434(7031):301–307, 2005.

[8] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023.

[9] Forrester Cole, Phillip Isola, William T Freeman, Frédo Durand, and Edward H Adelson. Shapecollage: occlusion-aware, example-based shape interpretation. In *ECCV 2012: 12th European Conference on Computer Vision*, 2012.

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[11] Zheng Ding, Mengqi Zhang, Jiajun Wu, and Zhuowen Tu. Patched denoising diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2023.

[12] 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 *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022.

[13] Ady Ecker and Allan D Jepson. Polynomial shape from shading. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 145–152. IEEE, 2010.

[14] Yuto Enyo and Ko Nishino. Diffusion reflectance map: Single-image stochastic inverse rendering of illumination and reflectance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

[15] Robert T. Frankot and Rama Chellappa. A method for enforcing integrability in shape from shading algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 10(4):439–451, 1988.

[16] John Haddon and David Forsyth. Shape representations from shading primitives. In *Computer Vision—ECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June 2–6, 1998 Proceedings, Volume II 5*, pages 415–431. Springer, 1998.

[17] Xinran Han and Todd Zickler. Curvature fields from shading fields. In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*, 2023.

[18] Tal Hassner and Ronen Basri. Example based 3D reconstruction from single 2D images. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, 2006.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] Kathryn Heal, Jialiang Wang, Steven J Gortler, and Todd Zickler. A lighting-invariant point processor for shading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 94–102, 2020.

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[23] Berthold KP Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. 1970.

[24] Berthold KP Horn and Michael J Brooks. *Shape from shading*. MIT press, 1989.

[25] Xinyu Huang, Jizhou Gao, Liang Wang, and Ruigang Yang. Examplar-based shape from shading. In *Sixth International Conference on 3-D Digital Imaging and Modeling (3DIM 2007)*, 2007.

[26] Satoshi Ikehata. Universal photometric stereo network using global lighting contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12591–12600, 2022.

[27] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.

[28] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. DiffusionCLIP: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022.

[29] Benjamin Kunsberg and Steven W Zucker. Characterizing ambiguity in light source invariant shape from shading. *arXiv preprint arXiv:1306.5480*, 2013.

[30] Benjamin Kunsberg and Steven W Zucker. How shading constrains surface patches without knowledge of light sources. *SIAM Journal on Imaging Sciences*, 7(2):641–668, 2014.

[31] Benjamin Kunsberg and Steven W Zucker. From boundaries to bumps: when closed (extremal) contours are critical. *Journal of Vision*, 21(13):7–7, 2021.

[32] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. SyncDiffusion: Coherent montage via synchronized joint diffusions. *Advances in Neural Information Processing Systems*, 36, 2024.

[33] 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*, pages 9970–9980, 2024.

[34] David Marr. Early processing of visual information. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 275(942):483–519, 1976.

[35] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

[36] Omer Meir, Meirav Galun, Stav Yagev, Ronen Basri, and Irad Yavneh. A multiscale variable-grouping framework for MRF energy minimization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1805–1813, 2015.

[37] Makaela Nartker, James Todd, and Alexander Petrov. Distortions of apparent 3D shape from shading caused by changes in the direction of illumination. *Journal of Vision*, 17(10):324–324, 2017.

[38] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[40] Yuri Ostrovsky, Patrick Cavanagh, and Pawan Sinha. Perceiving illumination inconsistencies in scenes. *Perception*, 34(11):1301–1314, 2005.

[41] Geoffrey Oxholm and Ko Nishino. Shape and reflectance from natural illumination. In *European Conference on Computer Vision (ECCV 2012)*, 2012.

[42] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[43] Alex Pentland. Shape information from shading: a theory about human perception. In *[1988 Proceedings] Second International Conference on Computer Vision*, pages 404–413. IEEE, 1988.

[44] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021.

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[46] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient Attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3531–3539, 2021.

[47] Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. *arXiv preprint arXiv:2307.08123*, 2023.

[48] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *International Conference on Learning Representations*, 2021.

[49] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[50] James T Todd, Eric JL Egan, and Flip Phillips. Is the perception of 3D shape from shading based on assumed reflectance and illumination? *i-Perception*, 5(6):497–514, 2014.

[51] Johan Wagemans, Andrea J Van Doorn, and Jan J Koenderink. The shading cue in context. *i-Perception*, 1 (3):159–177, 2010.

[52] Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang Wang, Weizhu Chen, Mingyuan Zhou, et al. Patch Diffusion: Faster and more data-efficient training of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[53] Felix Wimbauer, Shangzhe Wu, and Christian Rupprecht. De-rendering 3D objects in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[54] Ying Xiong, Ayan Chakrabarti, Ronen Basri, Steven J Gortler, David W Jacobs, and Todd Zickler. From shading to local shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):67–79, 2014.

[55] Yilun Xu, Mingyang Deng, Xiang Cheng, Yonglong Tian, Ziming Liu, and Tommi Jaakkola. Restart sampling for improving generative processes. *Advances in Neural Information Processing Systems*, 36: 76806–76838, 2023.

[56] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024.

[57] Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3164, 2019.

[58] Ye Yu and William AP Smith. Outdoor inverse rendering from a single image using multiview self-supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3659–3675, 2021.

[59] Daniel Zoran, Dilip Krishnan, Jose Bento, and Bill Freeman. Shape and illumination from shading using the generic viewpoint assumption. *Advances in Neural Information Processing Systems*, 27, 2014.

A Appendix / supplemental material

A.1 Algorithms pseudocode

We provide the pseudocode for the single-scale spatial consistency guided sampling (Alg. 1) and the lighting consistency guidance (Alg. 2). Here, we have hyperparameters λ for weighting the smoothness and integrability loss, η_t as guidance update weight and J_t as the number of noise update steps. The results in our paper use $\lambda = 0.5$ and $J_t = 3$. The parameter η_t is resolution-dependent and is included with the schedule specification in Appendix A.10.

A.2 Experimental setup details

Model Architecture. We use a conditional UNet architecture similar to [22] with input spatial dimensions of 16×16 and 4 channels. The input to the UNet is a concatenation of the grayscale shading image and a 3-channel normal map. We use a linear attention module [46] for better time and memory efficiency. The UNet consists of 4 downsampling and upsampling stages composed of the commonly used ResNet [19] blocks, group normalization layers, attention layers, and residual connections.

Dataset and Training Details. We train the pixel-space conditional diffusion model on a dataset that we build from the UniPS dataset [26]. It contains about 8000 256×256 synthetic images of 400 unique objects from the Adobe3D Assets [1] rendered from different viewing directions. We render the objects with the shadow-less Lambertian model using the provided ground truth normal maps in [26], a randomly sampled directional light source within 60 degrees of the z-axis, and an albedo value in $[0.5, 1]$. The surface normal values outside of the objects are set to $(-1, -1, -1)$.

We subdivide the image into non-overlapping patches of size 16×16 and train our model to predict the noise at each sampled timestep using a smooth L1 loss. To train the diffusion UNet, we use the cosine variance schedule [38] with 300 timesteps. The model is trained using the AdamW optimizer for 500 epochs with learning rate 2e-4. It takes about 40 hours using one Nvidia A100 GPU.

Algorithm 1: Spatial Consistency Guided Sampling at a Single Scale

Data: $\{x_t^u, c^u\}_{u\in\mathcal{V}}; \epsilon_\theta, \{\eta_t, \lambda\}$
while $t > 0$; // Parallel guidance
do
  for $j = 1, 2, \cdots, J_t$; // Gradient descent for multiple steps
  do
    $\hat{x}_0^u = \text{Pred}(x_t^u; c^u, \epsilon_\theta)$; // Predict $\hat{x}_0^u$ using Eq.4
    $\hat{x}_0 = \text{Reshape}(\{\hat{x}_0^u\})$; // Patch to global layout
    $\mathcal{L}(\hat{x}_0) = \frac{1}{|\mathcal{E}|}\mathcal{L}_S(\hat{x}_0^u, \hat{x}_0^v) + \lambda\frac{1}{|\mathcal{V}|}\mathcal{L}_I(\hat{x}_0^v)$; // Eq. 8
    $x_t^u \leftarrow x_t^u - \eta_t \cdot \nabla_{x_t^u}\mathcal{L}(\hat{x}_0)$
  end
  $x_{t-1}^u \leftarrow \text{Denoise}(x_t^u; c^u, \epsilon_\theta)$; // Parallel denoising (Eq.3) for each patch
end
return $\{x_0^u\}_{u\in\mathcal{V}}$; // Denoised normal prediction

Algorithm 2: Lighting Consistency Guidance

Data: $\{x_0^u, c^u\}_{u\in\mathcal{V}}$
$\hat{l}^u = \text{Robust Infer}(x_0^u, c^u)$; // Infer light source direction as in Eq. 11
Cluster center $\{L_1, L_2\}$, assignment k^u = K-Means Clustering($\{\hat{l}^u\}$, # clusters = 2)
for $u \in \mathcal{V}, k^u = 2$; // Assume that $|k^u = 1| \geq |k^u = 2|$
do
  $(x_0^u)_\mathbf{x} \leftarrow -(x_0^u)_\mathbf{x}, (x_0^u)_\mathbf{y} \leftarrow -(x_0^u)_\mathbf{y}$; // Convex/concave flip of the normals
end
return $\{x_0^u\}_{u\in\mathcal{V}}$

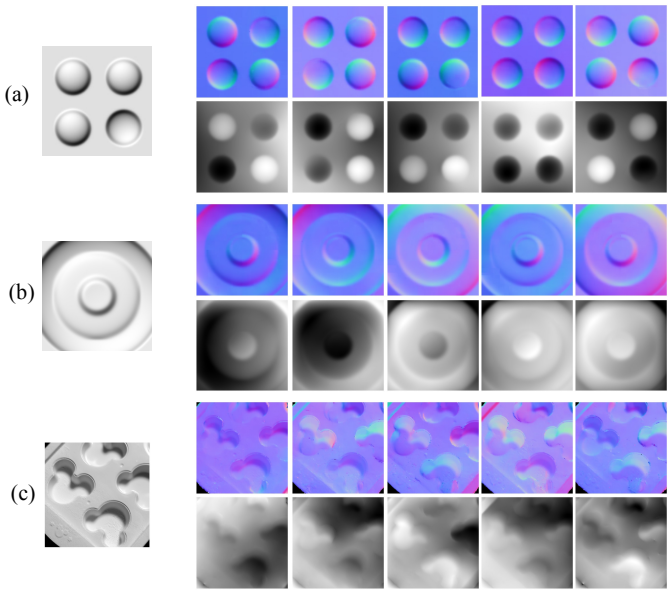


Figure 9: Sampled output normals and integrated depth maps when lighting consistency is not enforced. In contrast to the samples in Fig. 5, regions within the same image can undergo independent convex/concave flips.

Data Augmentation. During training, we augment the dataset with convex/concave flips of interior patches, meaning those that do not include any occluding contour and background. In Sec. A.7, we compare two models trained with and without such augmentation. While the augmentation leads to a more balanced multimodal output distribution and thus smaller Wasserstein distance, our model trained without augmentation is already capable of producing multistable outputs on test images.

Sampling. We use the DDIM sampler [48] with 50 sampling steps with guidance. Details of the multiscale sampling schedule and guidance learning rate η_t can be found in Appendix A.10.

Visualization and Evaluation Metric. We visualize the output normal map from multiple samples to show the multistable reconstructions. We also show depth maps by integrating the normals using the method by Frankot & Chellappa [15].

To produce the t-SNE visualizations in Figs 6 and 13, we draw 100 samples from each model, and then downsize the sampled normal maps to 64×64 resolution for computational efficiency. Then we project the samples to a 2-dimensional space using t-SNE with perplexity value of 30. To compute the 1-Wasserstein distance, we set the ground truth distribution to be a uniform distribution over the two reference normal maps: the one used to render the input image, and its convex/concave flip. When computing the 1-Wasserstein distance of model outputs to the ground truth distribution, all normal maps are first downsampled to 64×64 resolution.

Baseline models. For testing with Wonder3D [33], we extract the frontal view prediction and use the default setting in their online demo and code base with a crop size of 256 by 256, classifier-free guidance scale of 3, and 50 sampling steps.

A.3 Ablation study on lighting consistency guidance

In Fig. 9, we show additional samples from our model where lighting consistency guidance is turned off. Results show that in this case the output distribution of our model allows different regions within a surface (e.g., each dimple, ring or mouse-shape) to undergo an independent convex/concave flip. This is in contrast to Fig. 5 of the main paper, where we see only two global modes.

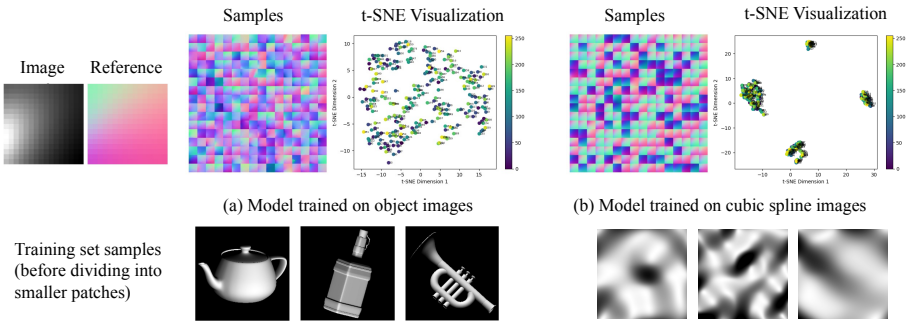


Figure 10: Output normal maps and their t-SNE visualizations when our model is applied to a 16×16 image of an exactly quadratic surface under directional lighting. When our model is trained using images of spline surfaces (b), the outputs cluster around the four mathematical interpretations from [54] (convex, concave, and two saddles). When it is trained using images of everyday objects (a), the outputs exhibit more diversity. In both scenarios, samples are drawn independently without guidance.

A.4 Relation to the four-way convex/concave/saddle ambiguity

Figure 10 examines the relationship between our model and the mathematical results from Xiong et al. [54], which show that, in general, an image of an exactly quadratic surface under unknown directional lighting can be explained by four quadratic shapes (convex, concave, and two saddles). If our model is consistent with the theory, we would expect its output shape distributions for such images to be concentrated around four distinct modes that match the four distinct possibilities. When we apply our model to images of exactly-quadratic surfaces like the one in the left of Fig. 10, we find that its output distribution is *not* concentrated near four modes (middle panel in the figure). However, we find that the four-mode behavior emerges when the model is retrained on a different dataset comprising random cubic splines (right panel), which by construction contain a much higher proportion of exactly-quadratic surface patches.

One potential explanation for this behavior is that exactly-quadratic surface patches are too rare in everyday scenes for a vision system to usefully exploit. This may relate to the perceptual experiments in [51] that suggest humans also struggle to perceive the four distinct shapes for such images.

A.5 Additional results on perceptual stimuli

Figure 11 shows our model's output for images that were used to study human perception in [31] and [37]. Our model produces plausible shapes and multimodal output distributions, while other models sometimes fail to recover a plausible shape or produce only one of the global concave/convex possibilities. The bottom row is an image of several bumps ('cobbles') lit from below. We observe that Wonder3D [33], Marigold [27] and SIRFS [4] interpret the bumps as concave, while Depth Anything [56] interprets them as convex.

A.6 Additional results on astronomical images; relation to the crater illusion

Figure 12 shows results for two satellite images of the surface of Mars. In images like these, humans often misperceive craters as mountains and vice versa, perhaps due to their bias toward lighting from above. (This is sometimes called the crater illusion.) We tested our model, the diffusion-based models [33, 27], and Depth Anything [56]. Our model sees both the crater and mountain possibilities, but the other models only see one of the two.

Data acquisition is expensive in these situations, so there could be benefit to having a monocular vision model that can automatically produce the multitude of explanations, thereby allowing all possibilities to be examined by gathering additional context. Similar to human perception, in astronomical imaging it is beneficial to have access to all of the possible "bottom-up" explanations, so that one can use context or "top-down" information as effectively as possible.

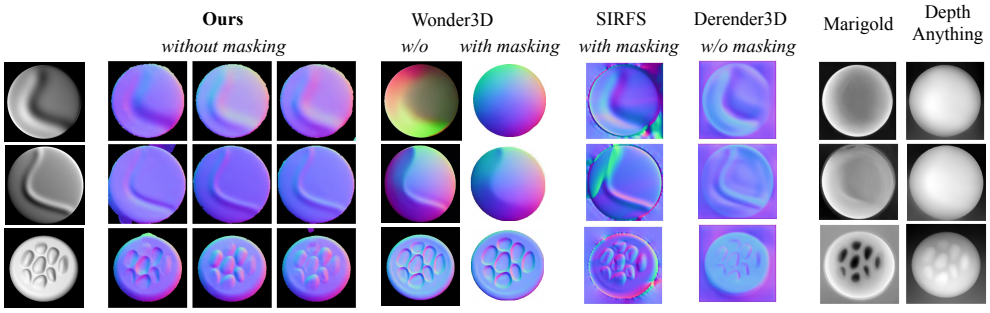


Figure 11: Inferred normal map samples from our model on ridge images taken from [31] and 'cobble' test image from [37].

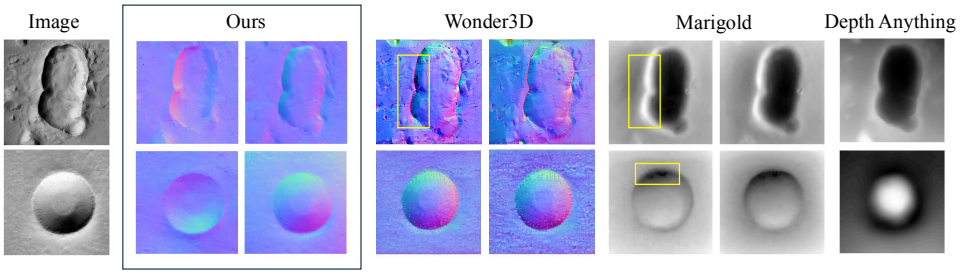


Figure 12: Shape from Martian crater images. For diffusion-based models we show two samples that qualitatively represent all of the 25 samples that we generated for that model. For depth models (Marigold and Depth Anything) brighter is closer. Our model outputs both possibilities, crater and mountain, while other models only see one of the two. Outputs from Wonder3D and Marigold also include artefacts (yellow boxes), with sharp spurious variations in normals or depth.

The two images are taken from:

(1) A triple crater in Elysium Planitia on Mars. Credit: NASA/JPL/University of Arizona. `https://www.universetoday.com/118581/amazing-impact-crater-where-a-triple-asteroid-smashed-into-mars/`

(2) A fresh impact crater, about 3 kilometers wide, gouged from a lava-covered plain in the Lunae Planum region of Mars. `https://skyandtelescope.org/astronomy-resources/astronomy-questions-answers/is-it-possible-that-photos-of-lunar-or-martian-landscapes-show-craters-as-blisters/`

A.7 Ablation of dataset augmentation

To explore the effect of convex/concave data augmentation during training (see the top of Section 4), we perform an ablation in which we train a model from the same set of patches but without the augmentation. Since the original patches from our training set tend to be dominated by convexity, we expect this to have an affect on the model's response to ambiguous images. Figure 13 and Table 1 bear this out. The figure and table show the same t-SNE visualizations and Wasserstein distances as in the main paper, but this time with without augmentation during training. The model without augmentation exhibits some multistability, especially for the last two images, but it tends to provide less diversity, especially for circular shapes. We hypothesize that this is caused by the existence of predominantly convex spherical shapes in the training set.

A.8 Quantitative results on a shape from shading benchmark

Table 2 shows quantitative comparisons using photographs from the dataset in [54]. We follow prior practice and report the median angular error of the predicted normal field, where the angular error at each pixel i is $AE(\hat{n}_i, \hat{n}_i^{gt}) = \left|\cos^{-1}(\hat{n}_i \cdot \hat{n}_i^{gt})\right|$. Diffusion-based outputs are stochastic and

Table 1: Wasserstein distance on multistable perception stimuli

Model	(a) four circles	(b) nested rings	(c) star	(d) snake
Wonder3D[33]	<u>30.94</u>	<u>38.96</u>	27.78	27.90
Ours (w/o data augmentation)	33.79	39.56	<u>22.04</u>	<u>23.02</u>
Ours	**12.59**	**29.96**	**17.32**	**18.27**

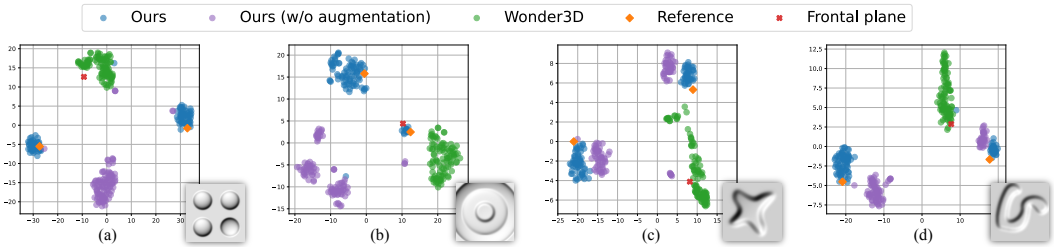


Figure 13: t-SNE plots showing that our model trained without flip augmentation already exhibits multistable outputs, especially on the last two images. Training with data augmentation improves the overall diversity on those test images.

potentially multi-modal (e.g., with global convex or concave possibilities), so for these we report the average error of the top-five predictions taken from 50 independent runs. Note that the method from [54] requires the true light direction to be provided as input, whereas ours and others do not.

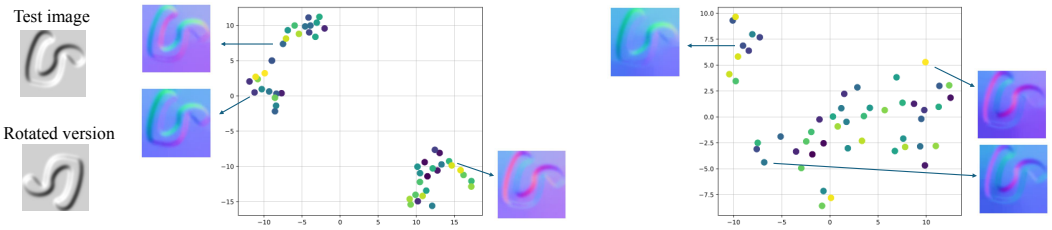
Overall, our model shows competitive performance on this benchmark compared with existing methods. We notice that our model trained without data augmentation via per-patch concave/convex flips (see the top of Section 4) performs better, while our model trained with data augmentation has slightly worse performance, likely due to its larger per-patch search space. This may be related to other quality-versus-diversity trade-offs that have been observed in previous conditional diffusion models [10, 21]. We leave for future research questions of how to achieve better combinations of quality and diversity, and how to incorporate other cues such as occluding contours and top-down recognition cues.

A.9 Ablation on lighting distribution in training set

Figure 14 shows an experiment where we change the distribution of light source directions in the training set. The two models A and B are trained on the same shapes without any data augmentation, but Model B has 80% of the images lit from above. Model A is trained with uniformly sampled lighting from above and from below. We test both models on an image that appears concave when 'lit from above'. From 50 random samples and their t-SNE projections, we see that Model B's distribution is biased towards the concave answer while Model A shows a more balanced distribution. This shows that lighting bias in the training set can have an effect on the output distribution.

Table 2: Shape from shading benchmark quantitative results. Errors are measured by median angular error of normals maps. Model performance for diffusion based models is averaged over the top 5 estimates over 50 random seeds.

Model	cat	frog	hippo	lizard	pig	turtle	scholar
Xiong et. al [54] (known lighting)	14.83	**11.80**	20.25	**12.70**	15.29	17.90	28.13
SIRFS Cross-Scale [4]	20.02	19.86	21.00	23.26	13.17	11.96	25.80
Wonder3D [33]	<u>14.29</u>	18.20	16.81	15.70	<u>10.10</u>	<u>9.59</u>	<u>25.32</u>
Ours (w/o data augmentation)	**11.49**	<u>15.56</u>	**14.17**	13.30	**9.27**	**8.72**	**22.01**
Ours	14.95	21.46	<u>15.94</u>	<u>12.82</u>	11.98	9.99	27.21



(a) Model A (b) Model B

Figure 14: Effect of lighting bias during training. Model A is trained using synthetic images with lighting directions that are uniformly sampled within a $60°$ cone around the view direction. Model B is trained using the same shapes but with lighting that is distributed non-uniformly within the cone, where 80% of images are lit from above. After training, we draw 50 samples from each model for a test image, using the same schedule and guidance hyperparameters. We project the results using t-SNE (dots are randomly colored for visual clarity) and show representative samples. Model A produces a balanced distribution across convex and concave explanations, whereas Model B produces concave predictions more often. (To humans, the test image usually appears concave, and it usually appears convex when rotated. These are both physically consistent with lighting from above.)

Table 3: Multiscale optimization schedule

	Perception Stimuli	Captured Photo
Resolution	[160, 128, 64, 80, 96, 112, 128, 144, 160]	[256, 160, 96, 128, 192, 224, 240, 256]
Guidance rate η	[20, 15, 10, 10, 10, 15, 15, 20, 20]	[30, 20, 12, 15, 20, 25, 28, 30]
Lighting guidance	[T] $\times$ 2 + [F] $\times$ 7	[F] $\times$ 3 + [T] $\times$ 2 + [F] $\times$ 3
$N\&R$ start t	[300] + [232] $\times$ 8	[300] + [238] $\times$ 7
Runtime (seconds)	105s (single Quadro RTX 8000)	125s (single Quadro RTX 8000)

A.10 Multiscale schedule specification

In Table 3 lists the scheduler hyperparameters for multiscale guided sampling. We use a notation for lists with repeating elements, where concatenation is represented as follows: for example, $[A] \times 2 + [B] \times 3$ denotes the list $[A, A, B, B, B]$.

When designing the multiscale schedule for inference, we find it helpful to have consecutive resolutions that are not integer multiples of the previous one, especially in the coarse to fine direction. This leads to improved quality because pixels that are adjacent to patch seams at one resolution become interior to a patch at another resolution.

For the initial resolution (our experiments use 160×160 or 256×256), guidance is applied only after the 8th DDIM denoising step since the predicted $\hat{x}_0$ at very early stages are often not informative enough for guidance. In terms of lighting consistency guidance, we find that it is often not necessary to apply at every resolution for a perceptually consistent normal estimation.

We apply normal field fusion using the last three resolutions, after the spatial predictions and choices of per-patch convex/concave modes have stabilized.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our experiments in Sec. 4 reflect the claims in abstract and introduction. We show that the model achieves multistable outputs on various perception stimuli and also produce veridical shape estimates for real objects.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include text in Sec. 6 dedicated for such discussions and possible future directions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not include new theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have included in the paper a detailed description of the model architecture, and pseudo-code for our algorithms. We also provide the hyperparameter and schedulers used in inference time in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We include in the supplemental material the implementation of the main algorithms in the paper. The data preparation instructions for our training dataset are also included in the appendix.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have included all necessary details about the experiment settings and training details in Appendix A.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We do not report statistical significance because the main observation in the paper is that there are many "ground truth" shapes for this monocular shape inference problem. They are related, for example, by convex/concave flips and bas-relief transformations. However, we use large numbers of samples to support the claims in our paper. Our quantitative results of the Wasserstein distance on perception stimuli is computed over 100 samples, and the angular errors for shape from shading benchmark is computed from 50 samples for all diffusion-based models.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute, memory and runtime requirement for reproducing the experiments are included in Appendix A.10.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We have discussed the potential positive impact of our model: (1) our model is compute-efficient and does not require many GPUs for training at large scale and (2) our work can help with understanding human vision. We do not perceive immediate negative societal impact from our work.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The data and model used in our work do not pose such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Our dataset is based on previous work *Universal Photometric Stereo Network using Global Contexts (CVPR 2022)* which uses GPL License and allows user to run, study, modify and share the project.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our work does not include any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.