SHAPE ANALYSIS BY SHADOW SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

3D reconstruction is a fundamental problem in computer vision, and the task is especially challenging when the object to reconstruct is partially or fully occluded. We introduce a method that uses the shadows cast by an unobserved object in order to infer the possible 3D volumes under occlusion. We create a differentiable image formation model that allows us to jointly infer the 3D shape of an object, its pose, and the position of a light source. Since the approach is end-to-end differentiable, we are able to integrate learned priors of object geometry in order to generate realistic 3D shapes of different object categories. Experiments and visualizations show that the method is able to generate multiple possible solutions that are consistent with the observation of the shadow. Our approach works even when the position of the light source and object pose are both unknown. Our approach is also robust to real-world images where ground-truth shadow mask is unknown.

1 INTRODUCTION

Reconstructing the 3D shape of objects is a fundamental challenge in computer vision, with a number of applications in robotics, graphics, and data science. The task aims to estimate a 3D model from one or more camera views, and researchers over the last twenty years have developed excellent methods to reconstruct visible objects (Horry et al., 1997; Hoiem et al., 2005; Ye et al., 2021; Mescheder et al., 2019; Mildenhall et al., 2020; Hartley & Zisserman, 2003; Agarwal et al., 2010). However, objects are often occluded, with the line of sight obstructed either by another object in the scene, or by themselves (self-occlusion). Reconstruction from a single image is an under-constrained problem, and occlusions further reduce the number of constraints. To reconstruct occluded objects, we need to rely on additional context.

One piece of evidence that people use to uncover occlusions is the shadow cast on the floor by the hidden object. For example, figure 1 shows a scene with an object that has become fully occluded. Even though no appearance features are visible, the shadow reveals that another object exists behind the chair, and the silhouette constrains the possible 3D shapes of the occluded object. What hidden object caused that shadow?

In this paper, we introduce a framework for reconstructing 3D objects from their shadows. We formulate a generative model of objects and their shadows cast by a light source, which we use

Input View
(Shadow shows occlusion)

Our Result
(New camera view)

Original Scene

Figure 1: We introduce a method to perform 3D reconstruction from the shadow cast on the floor by occluded objects. In the middle, we visualize our reconstruction of the occluded chair from a new camera view.

to jointly infer the 3D shape, its pose, and the location of the light source. Our model is fully differentiable, which allows us to use gradient descent to efficiently search for the best shapes that explain the observed shadow. Our approach integrates both learned empirical priors about the geometry of typical objects and the geometry of cameras in order to estimate realistic 3D volumes that are often encountered in the visual world.

Since we model the image formation process, we are able to jointly reason over the object geometry and the parameters of the light source. When the light source is unknown, we recover multiple different shapes and multiple different positions of the light source that are consistent with each other. When the light source location is known, our approach can make use of that information to further refine its outputs. We validate our approach for a number of different object categories on a new ground truth dataset that we will publicly release.

The primary contribution of this paper is a method to use the shadows in a scene to infer the 3D structure, and the rest of the paper will analyze this technique in detail. Section 2 provides a brief overview of related work for using shadows. Section 3 formulates a generative model for objects and how they cast shadows, which we are able to invert in order to infer shapes from shadows. Section 4 analyzes the capabilities of this approaches with a known and unknown light source. We believe the ability to use shadows to estimate the spatial structure of the scene will have a large impact on computer vision systems' ability to robustly handle occlusions.

2 Related Work

We briefly review related work in 3D reconstructions, shadows, and generative models. Our paper combines a model of image formation with generative models.

Single-view 3D Reconstruction and 3D Generative Models: The task of single-view 3D reconstruction – given a single image view of a scene or object, generate its underlying 3D model – has been approached by deep learning methods in recent years. This task is related to unconditional 3D model generation; while unconditional generation creates 3D models a priori, single-view reconstruction can be thought of as generation a posteriori where the condition is the input image view. Given the under-constrained nature of the problem, this is usually done with 3D supervision. Different lines of work address this by generating 3D models in different types of representations (Shin et al., 2018): specifically, whether they use voxels (Brock et al., 2016), point cloud representations (Fan et al., 2016), meshes (Groueix et al., 2018; Pontes et al., 2018), or the more recently introduced *occupancy networks* (Mescheder et al., 2019).

The cost to obtain 3D ground truth for supervision (Ionescu et al., 2014) poses a great limitation to the single-view 3D reconstruction. To scale up the applications, another line of work uses multi-view 2D images as supervision (Niemeyer et al., 2020; Kanazawa et al., 2019; Yu et al., 2021), or even only single image as supervision (Kanazawa et al., 2018; Liu et al., 2019; Goel et al., 2020; Wu et al., 2020; Li et al., 2020; Ye et al., 2021). More classically, approaches using Multi-View Stereo (MVS) reconstruct 3D object by combining multiple views (Bleyer et al., 2011; De Bonet & Viola, 1999; Broadhurst et al., 2001; Galliani et al., 2016; Schönberger et al., 2016; Seitz et al., 2006; Seitz & Dyer, 1999).

Occlusions and Shadows: Shadows present a naturally-occurring visual signal that can help to clarify uncertainty caused by occlusion. By observing the shadows cast by what we cannot see, we gain insight into the 3D structure of the unseen portion. Previous work has considered the use of shadows towards elucidating structure in a classic vision context. Waltz (1975) first applied shadows to determine shapes in 3D line drawings. This was extended by Shafer & Kanade (1983), who determined surface orientations for polyhedra and curved surfaces with shadow geometry. Shadows can also be used more actively to recover 3D shape. Bouguet & Perona (1999) shows how shadows can help infer the geometry of a shaded region. Savarese et al. (2007) also propose *shadow carving*, a way of using multiple images from the same viewpoint but with different lighting conditions to discover object concavities. Meanwhile, Troccoli & Allen (2004) use shadows as cues to determine parameters for refining 3D textures. Recent work has leveraged deep learning tools to enable detection of shadows from realistic images Wang et al. (2020), making it possible to extend the use of shadows to realistic settings. Thus far, shadows have not seen much application in determining structure using the tools afforded to vision by the latest deep learning techniques. Tiwary et al. (2022) is a concurrent work marrying implicit neural fields with shadows. However, the paper requires shadow masks from many light

sources with known locations, while our method only requires a single light sources with unknown location. This is made possible by leveraging the priors from a generative model of 3D objects.

Generation Under Constraints: Generation under constraints appears throughout the literature in many forms. It falls under the general framework of analysis by synthesis (Krull et al., 2015; Yuille & Kersten, 2006). Tasks such as super-resolution, image denoising, and image inpainting, begin with an incomplete image and ask for possible reconstructions of the complete image (Ongie et al., 2020). In other words, the goal is to generate realistic images that satisfy the constraint imposed by the given information. Typical approaches consider this as conditional generation, where a function (usually, a neural network) is learned to map from corrupted inputs to the desired outputs (Dong et al., 2015; Kim et al., 2016; Ongie et al., 2020). More recently, Menon et al. (2020) propose using search rather than regression to address these types of tasks in the context of the super-resolution problem. Recent work by Sadekar et al. (2022) uses differentiable rendering to deform an icosphere to generate targeted shadow art sculptures, with interesting results. Unlike their work, ours focuses on generating a set of *plausible* objects which could explain a given shadow in a real scene.

3 METHOD

We represent the observation of the shadow as a binary image $\mathbf{s} \in \mathbb{R}^{W \times H}$. Our goal is to estimate a set of possible 3D shapes, their poses, and corresponding light sources that are consistent with the shadow $\mathbf{s}$. We approach this problem by defining a generative model for objects and their shadows. We will use this forward model to find the best 3D shape that could have produced the shadow.

3.1 EXPLAINING SHADOWS WITH GENERATIVE MODELS

Let $\Omega = G(\mathbf{z})$ be a generative model for 3D objects, where Ω parameterizes a 3D volume and $\mathbf{z} \sim \mathcal{N}(0, I)$ is a latent vector with an isotropic prior. When the volume blocks light, it will create a shadow. We write the location of the illumination source as $\mathbf{c} \in \mathbb{R}^3$ in world coordinates, which radiates light outwards in all directions. The camera will observe the shadow $\hat{s} = \pi(c, \Omega)$, where π is a rendering of the shadow cast by the volume Ω onto the ground plane.

To reconstruct the 3D objects from their shadow, we formulate the problem as finding a latent vector $\mathbf{z}$, object pose ϕ, and light source location $\mathbf{c}$ such that the predicted shadow $\hat{\mathbf{s}}$ is consistent with the observed shadow $\mathbf{s}$. We perform inference by solving the optimization problem:

$$\min_{\mathbf{z},\mathbf{c},\phi} \mathcal{L}\left(s, \pi(\mathbf{c}, \Omega)\right) \quad \text{where} \quad \Omega = \mathcal{T}_\phi\left(G(\mathbf{z})\right) \tag{1}$$

The loss function $\mathcal{L}$ compares the candidate shadow $\hat{s} = \pi(\mathbf{c}, \Omega)$ and the observed shadow $\mathbf{s}$, and since silhouettes are binary images, we use a binary cross-entropy loss. We model the object pose with an SE(3) transformation $\mathcal{T}$ parameterized by quaternions ϕ. In other words, we want to find a latent vector that corresponds to an appropriate 3D model of the object that, in the appropriate pose, casts a shadow matching the observed shadow. Consequently, we can freely choose the location of the camera; we do not need to model the camera extrinsic parameters. Since $\mathbf{z}$ is normally distributed, we constrain the norm of $\mathbf{z}$ to be within $\delta \in \mathbb{R}$ distance to the surface of a unit hyper-sphere.

Figure 2 illustrates an overview of this setup. The solution $\mathbf{z}^*$ of the optimization problem will correspond to a volume that is consistent with the observed shadow. We can obtain the resulting shape through $\Omega^* = \mathcal{T}_{\phi^*}\left(G(\mathbf{z}^*)\right)$. By solving Equation 1 multiple times with different initializations, we obtain a set of solutions $\{\mathbf{z}^*\}$ yielding multiple possible 3D reconstructions.

3.2 $G(\mathbf{z})$: GENERATIVE MODELS OF OBJECTS

To make the reconstructions realistic, we need to incorporate priors about the geometry of objects typically observed in the visual world. Rather than searching over the full space of volumes Ω, our approach searches over the latent space $\mathbf{z}$ of a pretrained deep generative model $G(\mathbf{z})$. Generative models that are trained on large-scale 3D data are able to learn empirical priors about the structure of objects; for example, this can include priors about shape (e.g., automobiles usually have four wheels) and physical stability (e.g., object parts must be supported). By operating over the latent space $\mathbf{z}$, we can use our knowledge of the generative model's prior to constrain our solutions to 3D objects that match the generative model's output distribution.

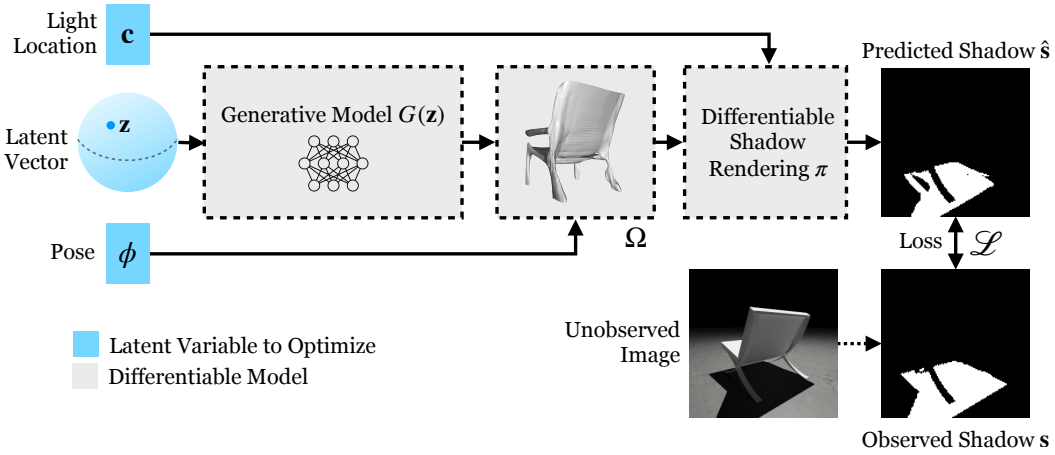

Figure 2: **Overview of our method.** Given an observation of a shadow $\mathbf{s}$, we optimize for an explanation jointly over the location of the light $\mathbf{c}$, the pose of the object $\mathcal{T}_\phi$, and the latent vector of the object 3D shape $\mathbf{z}$. Since every step is differentiable, we are able to solve this optimization problem with gradient descent. By starting the optimization algorithm with different initializations, we are able to recover multiple possible explanations Ω for the shadow.

Our approach is compatible with many choices of 3D representation. In this implementation, we choose to model our 3D volumes with an occupancy field (Mescheder et al., 2019). An occupancy network $\mathbf{y} = f_\Omega(\mathbf{x})$ is defined as a neural network that estimates the probability $\mathbf{y} \in \mathbb{R}$ that the world coordinates $\mathbf{x} \in \mathbb{R}^3$ contains mass. The generative model $G(\mathbf{z})$ is trained to produce the parameters Ω of the occupancy network.

3.3 π: Differentiable Rendering of Shadows

To optimize Equation 1 with gradient descent, we need to calculate gradients of the shadow rendering π and its projection to the camera. This operation can be made differentiable by max-pooling the value of the occupancy network along a light ray originating at the light source. Although integrating occupancy along the ray may be more physically correct to deal with partially transmissive media as in NeRF (Mildenhall et al., 2020), since we are primarily concerned with solid, opaque objects and binary shadow masks, we find max-pooling to be a useful simplifying approximation.

Let $\mathbf{r}_\theta \in \mathbb{R}^3$ be a unit vector at an angle θ, and let $\mathbf{n} \in \mathbb{R}^3$ be a vector normal to the ground plane. We need to calculate whether the ray from the light source $\mathbf{c}$ along the direction of $\mathbf{r}_\theta$ will intersect with the ground plane $\mathbf{n}$, or whether it will be blocked by the object Ω. The shadow will be an image $\pi(\mathbf{c}, \Omega)$ formed on the ground plane, and the intensity on the plane at position $\mathbf{p}$ is given by:

$$\pi(\mathbf{c}, \Omega)[\mathbf{p}] = \max_{d \in \mathbb{R}} f_\Omega(\mathbf{c} + d\mathbf{r}_\theta) \quad \text{s.t.} \quad \mathbf{p} = \mathbf{c} - \frac{\mathbf{c}^T\mathbf{n}}{\mathbf{r}_\theta^T\mathbf{n}}\mathbf{r}_\theta \tag{2}$$

where we use the notation $\pi(\mathbf{c}, \Omega)[\mathbf{p}]$ to index into $\pi(\mathbf{c}, \Omega)$ at coordinate $\mathbf{p}$. The right-hand constraint between $\mathbf{p}$ and $\mathbf{r}_\theta$ is obtained by calculating the intersection of the light ray with the ground plane.

For the light ray $\mathbf{r}_\theta$ landing at $\mathbf{p}$, the result of π is the maximum occupancy value f_Ω along that ray. Since $\pi(\mathbf{c}, \Omega)$ is an image of the shadow on a plane, we use homography to transform $\pi(\mathbf{c}, \Omega)$ into the perspective image $\hat{\mathbf{s}}$ captured by the camera view. A figure 10 illustrating the image formation model can be found in appendix.

3.4 Optimization

Given a shadow $\mathbf{s}$, we optimize $\mathbf{z}$, $\mathbf{c}$, and ϕ in Equation 1 with gradient descent while holding the generative model $G(\mathbf{z})$ fixed. We randomly initialize $\mathbf{z}$ by sampling from a multivariate normal distribution, and we randomly sample both a light source location $\mathbf{c}$ and an initial pose ϕ. We then calculate gradients using back-propagation to minimize the loss between the predicted shadow $\hat{\mathbf{s}}$ and the observed shadow $\mathbf{s}$.

During optimization, we need to enforce that $\mathbf{z}$ resembles a sample from a Gaussian distribution. If this is not satisfied, the inputs to the generative model will no longer match the inputs it has

Algorithm 1 Inference by Inverting the Generative Model

Input: Shadow image $\mathbf{s}$, step size η, number of iterations K, and generator G.
Output: Parameters of a 3D volume Ω
Inference:
Randomly initialize $\mathbf{z} \sim \mathcal{N}(0, I)$
for $k = 1, ..., K$ **do**
  $J(\mathbf{z}, \mathbf{c}, \phi) = \mathcal{L}\left(\mathbf{s}, \pi(\mathbf{c}, \mathcal{T}_\phi(G(\mathbf{z})))\right)$
  $\mathbf{z} \leftarrow \mathbf{z} - \eta \cdot (\nabla_{\mathbf{z}} J(\mathbf{z}, \mathbf{c}, \phi) + \mathcal{N}(0, \sigma I))$ where $\sigma = \frac{K-1-k}{K}$
  $\mathbf{z} \leftarrow \mathbf{z}/||\mathbf{z}||_2$
  $\mathbf{c} \leftarrow \mathbf{c} - \eta \nabla_{\mathbf{c}} J(\mathbf{z}, \mathbf{c}, \phi)$
  $\phi \leftarrow \phi - \eta \nabla_{\phi} J(\mathbf{z}, \mathbf{c}, \phi)$
end for
Return parameters of 3D volume $\Omega = \mathcal{T}_\phi(G(\mathbf{z}))$

seen during training. This could result in undefined behavior and will not make use of what the generator has learned. We follow the technique from Menon et al. (2020), which made the observation that the density of a high-dimensional Gaussian distribution will condense around the surface of a hyper-sphere (the 'Gaussian bubble' effect). By enforcing a hard constraint that $\mathbf{z}$ should be near the hyper-sphere, we can guarantee the optimization will find a solution that is consistent with the generative model prior.

The objective in Equation 1 is non-convex, and there are many local solutions for which gradient descent can become stuck. Motivated by Welling & Teh (2011), we found that adding linearly decaying Gaussian noise helped the optimization find better solutions. Algorithm 1 summarizes the full procedure.

4 Experimental Results

The goal of our experiments is to analyze how well our method can estimate 3D shapes that are consistent with an observed shadow. We first introduce a new 3D shadow dataset, then we perform two different quantitative experiments to evaluate the 3D reconstruction performance of our model. We further provide several visualizations and qualitative analysis of our method.

4.1 Common Experimental Setup

Evaluation Metric: We use volumetric IoU to evaluate the accuracy of 3D reconstruction. Volumetric IoU is calculated by dividing the intersection of the two volumes by their union. We uniformly sample 100k points in the bounding volume. We then calculate the occupancy agreement of the points between the candidate 3D volume and the original 3D volume.

Baselines: To validate our method quantitatively, we selected several baselines for comparison. Since we are analyzing how well generative models can explain shadows in images, we compare against the 3 approaches: *Regression*, *Nearest Neighbor*, *Random*. Detailes of these baseline methods can be found in section A.3 of appendix.

4.2 Reconstruction with Known Light and Object Pose

We first evaluate our method on the task of 3D reconstruction when the light position and pose are known. For each scene, we randomize the location of the light source, and put the objects in their canonical pose. Since the problem is under-constrained, there is not a single unique answer. We consequently run each method eight times to generate diverse predictions, and calculate the average volumetric IoU using the best reconstruction for each example.

Table 1 compares the performance of our approach versus baselines on this task. Our approach is able to significantly outperform the baselines on this task (by nearly 9 points), showing that it can effectively find 3D object shapes that are consistent with the shadows. Since our approach integrates empirical priors from generative models with the geometry of camera and shadows, it is able to

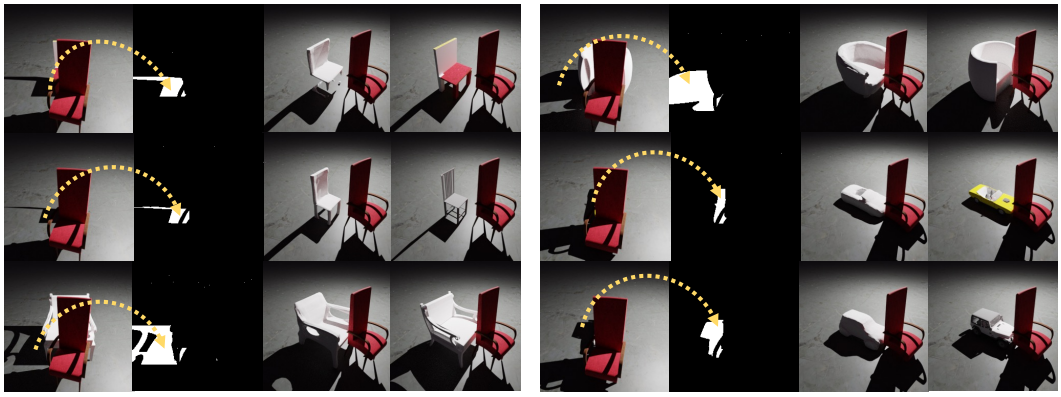

Figure 3: **3D Reasoning Under Occlusion.** We show several examples of 3D object reconstruction under occlusion. The **1st** column shows the original scenes including both objects. Shadow masks shown in the **2nd** column. The **3rd** and **4th** column are our reconstruction as seen from another camera view. Note that the red chair in the front is not being reconstructed by our model.

better generalize to the testing set. The regression baseline, for example, does not benefit from these inductive biases, and instead must learn them from data, which our results show is difficult.

When the full image is available, Table 1 shows that established 3D reconstruction methods are able to perform better, which is expected because more information is available. However, when there is an occlusion, the full image will not be available, and we instead must rely on shadows to reconstruct objects. Figure 3 shows qualitative examples where we were able to reconstruct objects that are occluded other objects. Although there is no appearance information, these results show that shadows allow our model to "see through" occlusions in many cases. The examples show that the method is able to reconstruct objects faithfully with diverse shapes and across different categories. We include more examples from all categories in the supplementary materials.

4.3 Reconstruction with Unknown Light and Object Pose

Since our approach is generative and not discriminative, a key advantage is the flexibility to adapt to different constraints and assumptions. In this experiment, we relax our previous assumption that the light source location and the object pose are both known. We evaluate our approach at reconstruction where all three variables (latent vector $\hat{\mathbf{z}}$, light source location $\hat{\mathbf{c}}$, and object pose parameters ϕ) must be jointly optimized by gradient descent to minimize the shadow reconstruction loss.

Table 2 shows the performance of our model at reconstructing objects with an unknown illumination position and pose. In this under-constrained setting, our approach is able to significantly outperform the baseline methods as much as 29%. In this setting, the most difficult object to reconstruct is a chair, which often has thin structures in the shadow.

Discriminative regression models are limited to produce reconstructions that are consistent with their training conditions, which is a principle restriction of prior methods. As we relax the number of

Method	Car	Chair	Plane	Sofa	All
Random	.329	.203	.211	.209	.238
Nearest Neighbor (Chang et al., 2015)	.414	.299	.349	.352	.322
Regression (Mescheder et al., 2019)	.611	.274	.410	.524	.467
Latent Search (Ours)	**.706**	**.371**	**.537**	**.598**	**.553**
Im2Mesh (full image) Mescheder et al. (2019)	.737	.501	.571	.680	.622

Table 1: Results for 3D reconstruction from the shadows assuming the object pose and the light source position are both known. We report volumetric IoU, and higher is better. The Im2Mesh result shows the performance at 3D reconstruction when the entire image is observable, not just the shadows.

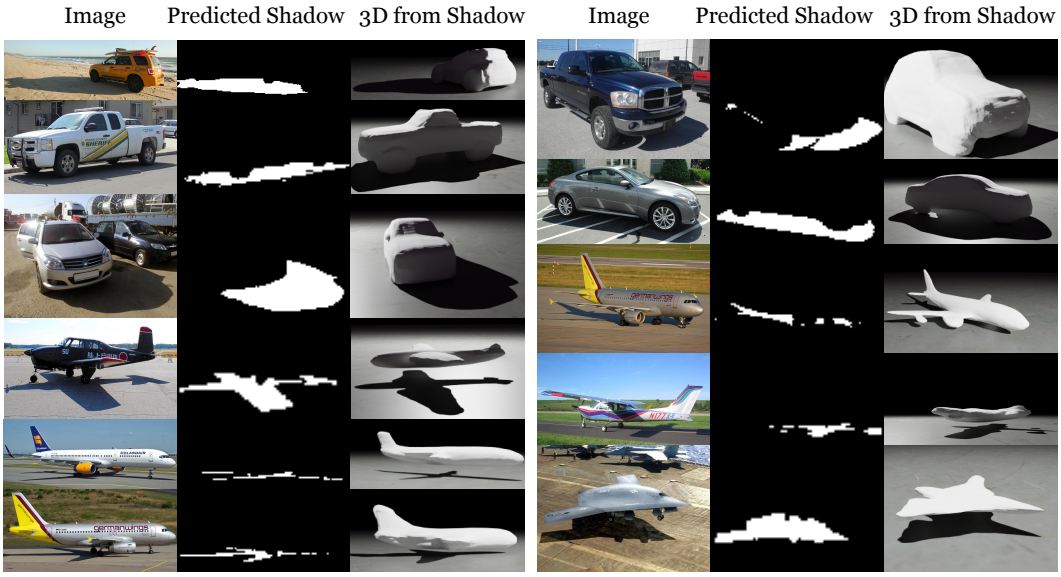

Figure 4: 3D reconstructions in natural images. We first automatically segment shadow mask with Wang et al. (2020). We then run our algorithm.

known variables, the size of the input space significantly increases, which requires the regression baseline to increasingly generalize outside of the training set. Table 2 shows that regression is only marginally better than a nearest neighbor search on average. However, since our approach is generative, and integrates inductive biases about scene illumination, it is able to better generalize to more unconstrained settings.

4.3.1 Natural Image

We applied our method to the real-world dataset in Wang et al. (2020), and automatically obtain shadow segmentations with the detector proposed by the same work. Fig. 4 shows our 3D reconstructions from just the estimated shadows. Our method remains robust both for real-world images and slightly inaccurate shadow masks. These results also show our method estimates reasonable reconstructions when the ground-truth camera pose or light source location are unknown. Our method also returns reasonable-looking models even if the floor is not flat (e.g. car on sand).

4.4 Diversity of Reconstructions

By modeling the generative process of shadows, our approach is able to find multiple possible 3D shapes to explain the observed shadow. When we sample different latent vectors as initialization, coupled with stochasticity from Gaussian noise in gradient descent, our method can generate a diverse set of solutions to minimize the shadow reconstruction loss. Estimation of multiple possible scenes is an important advantage of our approach when compared with a regression model. There are many correct solutions to the 3D reconstruction task. When a regression model is trained to make a prediction for these tasks, the optimal solution is to predict the average of all the possible shapes in order to minimize the loss. In comparison, our approach does not regress to the mean under uncertainty.

Figure 5 shows how the generative model is able to produce multiple, diverse samples that are all consistent with the shadow. For example, when given a shadow of a car, the method is able to produce

Method	Car	Chair	Plane	Sofa	All
Random	.283	.175	.177	.161	.199
Nearest Neighbor (Chang et al., 2015)	.346	**.233**	.241	.233	.264
Regression Mescheder et al. (2019)	.559	.116	.218	.317	.303
Latent Search (Ours)	**.618**	.187	**.343**	**.413**	**.390**

Table 2: Results for 3D reconstruction from the shadows assuming the object pose and the light source position are **both unknown**. We report volumetric IoU, and higher is better.

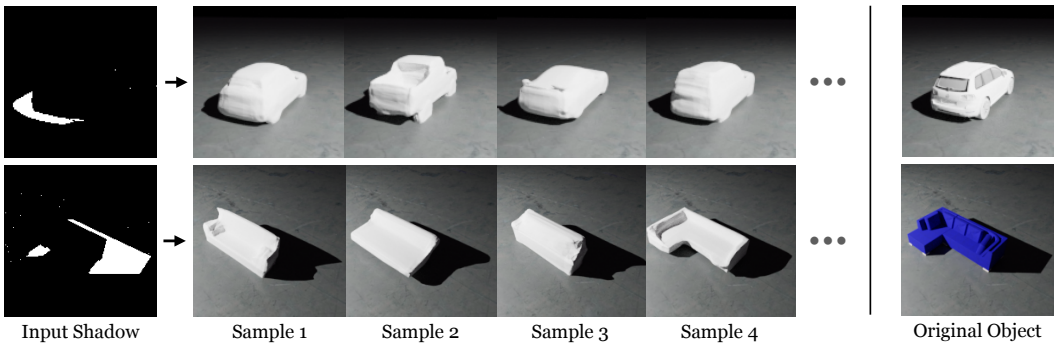

Figure 5: **Diversity of Reconstructions.** Given one shadow (**left**), our method is able to estimate multiple possible reconstructions (**middle**) that are consistent with the shadow. We show four samples from the model (columns). The **right** side shows the original object.

both trucks and sedans that might cast the same shadow. When given the shadow of a sofa, the latent search discovers both L-shaped sofas and rectangular sofas that are compatible with the shadow. Figure 6 quantitatively studies the diversity of samples from our method. As we perform latent search on the generative model with different random seeds, the likelihood of producing a correct prediction monotonically increases. This is valuable for our approach to be deployed in practice to resolve occlusions, such as robotics, where systems need to reason over all possible hypotheses for up-stream decision making.

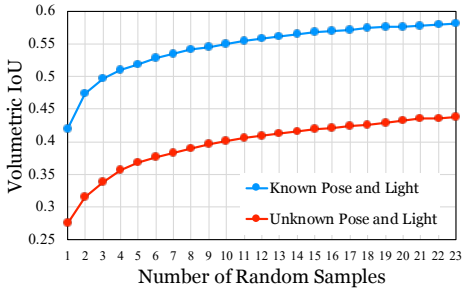

Figure 6: **Performance with Diverse Samples.** We show that our approach is able to make diverse 3D reconstructions from shadows. We plot best volumetric IoU versus the number of random samples from our method. The upward trends indicate the diversity of the prediction results from our method.

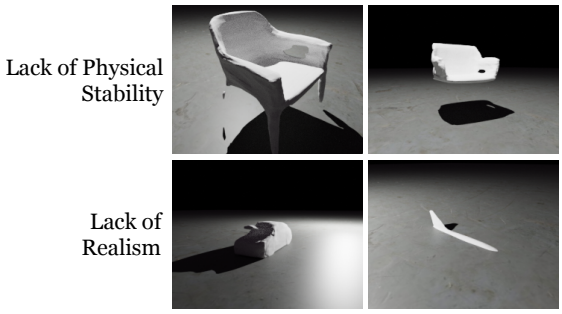

Figure 7: **Failures.** We visualize representative failures where the model produces incorrect shapes that still match the shadow. Our experiments suggest that results can be further improved with more priors, such as physical knowledge (**top**) and refined generative models (**bottom**).

4.5 Analysis

Optimization Process: To gain intuition into how our model progresses in the latent space to reach the final shadow-consistent reconstruction, we visualize in figure 8 the optimization process by extracting the meshes corresponding to several optimization iterations before converging at the end. Figure 8 shows a clear transition from the first mesh generated from a randomly sampled latent vector, to the last mesh that accurately cast shadows matching the input ones. The reconstructed meshes at the end also match the original objects.

Reconstructions of Edited Shadows: We found that our approach is able to exploit subtle details in shadows in order to produce accurate reconstructions. To study this behavior, we manually made small modifications to some of the shadow images, and analyzed how the resulting reconstructions changed. Figure 9 shows two examples. In the example on the left, we modified the shadow of a chair to add an arm rest in the shadow image. In the comparison between the original reconstruction and modified reconstruction, we can see an arm rest being added to the reconstructed chair. In the example on the right, we take a shadow image of a sedan and make the shadow corresponding to the

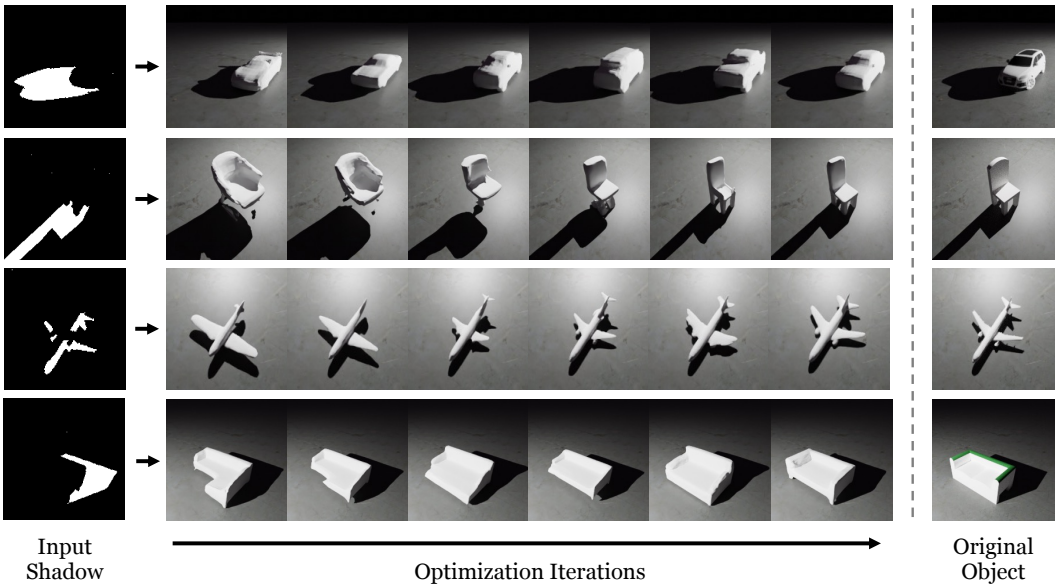

Figure 8: **Visualizing Optimization Iterations.** Visualizing the process of our model searching for 3D shapes that cast a shadow consistent with the input. The **1st column** shows the shadow used as a constraint for searching. **The middle** sequence of figure shows the process of searching in the latent space. The **last column** shows the original object as a reference, which is unseen by our model.

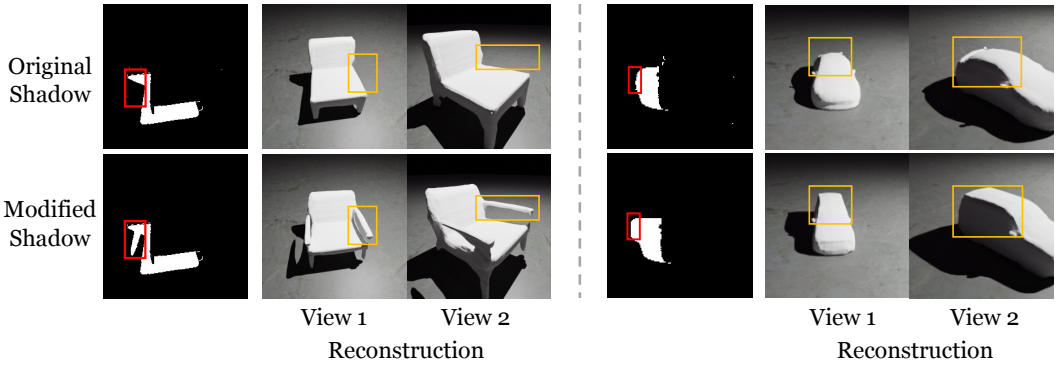

Figure 9: **Reconstructing Edited Shadows.** We manually modify a shadow mask and comparing the reconstructed 3D object between the original and modified shadows. View 1 is the same as the original shadow image. View 2 is a second view for visualizing more details.

rear trunk part higher. The reconstructed car from the modified image becomes an SUV to adapt to the modified shadow.

Analysis of Failures: We show a few representative examples of failures in figure 7. Although these shapes match the provided shadow, they are incorrect because they either lack physical stability or produce objects that are unlikely to be found in the natural visual world. These failures suggest that further progress on this problem is possible by integrating more extensive priors about objects and physics.

5 Conclusions

This paper shows that generative models are a promising mechanism to explain shadows in images. Our experiments show that jointly searching the latent space of a generative model and parameters for the light position and object pose allows us to reconstruct 3D objects from just an observation of the shadow. We believe tightly integrating empirical priors about objects with models of image formation will be an excellent avenue for resolving occlusions in scenes.

6 ETHICS STATEMENT

This paper is primarily based off ShapeNet (a publicly released dataset) and Blender (a publicly available software). In addition, all the code, models, and data will be released. As authors of the paper, we have carefully reviewed and will adhere to the code of ethics provided at `https://iclr.cc/public/CodeOfEthics`.

7 REPREDUCIBILITY STATEMENT

We will release all code, models, and data. To create $G(\mathbf{z})$, we use the unconditional 3D generative model from Mescheder et al. (2019), which is trained to produce an occupancy network with a 128-dimensional latent vector. The generative model is trained separately on four categories of the ShapeNet dataset, as in Mescheder et al. (2019). When the location of the illumination source is unknown, we sample a 3-dimensional coordinate $\mathbf{c}$ from the surface of the northern hemisphere above the ground plane with a fixed radius of 3. When the SE(3) transformation for the object pose is unknown, we sample a 4-dimensional quaternion ϕ to parameterize the rotation matrix. A non-zero rotation for "pitch" and "roll" are physically implausible given a level ground plane and the assumption of upright object, so we constrain them to be zero during optimization. To optimize the full model, we use spherical gradient descent to optimize $\mathbf{z}$ (and optionally $\mathbf{c}$ and ϕ) for up to 300 steps. We use a step size of 1.0 for known light and pose experiments and 0.01 for unknown light and pose experiments.

To accomplish the differentiable shadow rendering π, we evenly sample 128 points along each light ray emitted from the illumination source, then evaluate them for occupancy. In the case of occlusion from other objects as well as self-occlusion, we calculate the segmentation mask of all objects in the scene, and disable gradients coming from light rays intersecting with these masks.

REFERENCES

Sameer Agarwal, Noah Snavely, Steven M Seitz, and Richard Szeliski. Bundle adjustment in the large. In *European conference on computer vision*, pp. 29–42. Springer, 2010.

Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, pp. 1–11, 2011.

Jean-Yves Bouguet and Pietro Perona. 3d photography using shadows in dual-space geometry. *International Journal of Computer Vision*, 35(2):129–149, 1999.

Adrian Broadhurst, Tom W Drummond, and Roberto Cipolla. A probabilistic framework for space carving. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, volume 1, pp. 388–393. IEEE, 2001.

Andrew Brock, Theodore Lim, J. M. Ritchie, and Nick Weston. Generative and Discriminative Voxel Modeling with Convolutional Neural Networks. *arXiv:1608.04236 [cs, stat]*, August 2016. URL `http://arxiv.org/abs/1608.04236`. arXiv: 1608.04236.

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

Jeremy S De Bonet and Paul Viola. Poxels: Probabilistic voxelized volume reconstruction. In *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 418–425. Citeseer, 1999.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image Super-Resolution Using Deep Convolutional Networks. *arXiv:1501.00092 [cs]*, July 2015. URL `http://arxiv.org/abs/1501.00092`. arXiv: 1501.00092.

Haoqiang Fan, Hao Su, and Leonidas Guibas. A Point Set Generation Network for 3D Object Reconstruction from a Single Image. *arXiv:1612.00603 [cs]*, December 2016. URL `http://arxiv.org/abs/1612.00603`. arXiv: 1612.00603.

Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Gipuma: Massively parallel multi-view stereo reconstruction. *Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e. V*, 25(361-369):2, 2016.

Shubham Goel, Angjoo Kanazawa, , and Jitendra Malik. Shape and viewpoints without keypoints. In *ECCV*, 2020.

Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 216–224, 2018.

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

Derek Hoiem, Alexei A Efros, and Martial Hebert. Automatic photo pop-up. In *ACM SIGGRAPH 2005 Papers*, pp. 577–584. 2005.

Youichi Horry, Ken-Ichi Anjyo, and Kiyoshi Arai. Tour into the picture: using a spidery mesh interface to make animation from a single image. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pp. 225–232, 1997.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018.

Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Computer Vision and Pattern Regognition (CVPR)*, 2019.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1646–1654, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.182. URL `http://ieeexplore.ieee.org/document/7780551/`.

Alexander Krull, Eric Brachmann, Frank Michel, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Learning analysis-by-synthesis for 6d pose estimation in rgb-d images. In *Proceedings of the IEEE international conference on computer vision*, pp. 954–962, 2015.

Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *European Conference on Computer Vision*, pp. 677–693. Springer, 2020.

Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7708–7717, 2019.

Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 2437–2445, 2020.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4460–4470, 2019.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pp. 405–421. Springer, 2020.

Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3504–3515, 2020.

Gregory Ongie, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, and Rebecca Willett. Deep Learning Techniques for Inverse Problems in Imaging. *arXiv:2005.06001 [cs, eess, stat]*, May 2020. URL http://arxiv.org/abs/2005.06001. arXiv: 2005.06001.

Jhony K Pontes, Chen Kong, Sridha Sridharan, Simon Lucey, Anders Eriksson, and Clinton Fookes. Image2mesh: A learning framework for single image 3d reconstruction. In *Asian Conference on Computer Vision*, pp. 365–381. Springer, 2018.

Kaustubh Sadekar, Ashish Tiwari, and Shanmuganathan Raman. Shadow art revisited: a differentiable rendering based approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 29–37, 2022.

Silvio Savarese, Marco Andreetto, Holly Rushmeier, Fausto Bernardini, and Pietro Perona. 3D Reconstruction by Shadow Carving: Theory and Practical Evaluation. *International Journal of Computer Vision*, 71(3):305–336, March 2007. ISSN 0920-5691. doi: 10.1007/s11263-006-8323-9. URL https://doi.org/10.1007/s11263-006-8323-9.

Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pp. 501–518. Springer, 2016.

Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(2):151–173, 1999.

Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, pp. 519–528. IEEE, 2006.

Steven A Shafer and Takeo Kanade. Using shadows in finding surface orientations. *Computer Vision, Graphics, and Image Processing*, 22(1):145–176, April 1983. ISSN 0734-189X. doi: 10.1016/0734-189X(83)90099-3. URL https://www.sciencedirect.com/science/article/pii/0734189X83900993.

Daeyun Shin, Charless C. Fowlkes, and Derek Hoiem. Pixels, Voxels, and Views: A Study of Shape Representations for Single View 3D Object Shape Prediction. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3061–3069, Salt Lake City, UT, June 2018. IEEE. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00323. URL https://ieeexplore.ieee.org/document/8578421/.

Kushagra Tiwary, Tzofi Klinghoffer, and Ramesh Raskar. Towards learning neural representations from shadows. *arXiv preprint arXiv:2203.15946*, 2022.

A.J. Troccoli and P.K. Allen. A Shadow Based Method for Image to Model Registration. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pp. 169–169, June 2004. doi: 10.1109/CVPR.2004.289.

David Waltz. Understanding Line Drawings of Scenes with Shadows. In *The Psychology of Computer Vision*, pp. pages. McGraw-Hill, 1975.

Tianyu Wang, Xiaowei Hu, Qiong Wang, Pheng-Ann Heng, and Chi-Wing Fu. Instance Shadow Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1877–1886, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.00195. URL https://ieeexplore.ieee.org/document/9157490/.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688. Citeseer, 2011.

Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1–10, 2020.

Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8843–8852, 2021.

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

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

A APPENDIX

A.1 DIFFERENTIABLE RENDERING OF SHADOWS

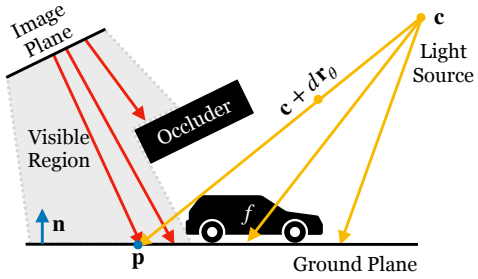

Figure 10: **Differentiable rendering of shadows.** A point $\mathbf{p}$ on the ground plane will be a shadow if the corresponding ray from the light source intersects with the volume f. We calculate whether $\mathbf{p}$ is a shadow by finding the intersecting ray $\mathbf{r}_\theta$, and max pooling f_Ω along the ray.

A.2 KAGEMUSHA: A DATASET OF SHADOWS

Kagemusha[1] is a dataset of 3D objects and their shadows. The dataset contains four common objects of the ShapeNet dataset Chang et al. (2015). For each 3D object, we sample a random point light source location from the northern hemisphere with a radius of 3 and a camera location sampled from the northern hemisphere with a radius of 2, to create a scene with shadow. We then compute the segmentation mask and shadow mask of the object. The dataset uses the same train/validation/test split as the original ShapeNet dataset Chang et al. (2015).

A.3 BASELINE METHODS

(i) Regression: An alternative approach to the same problem is to train a regression model to map images of shadows $\mathbf{s}$ to 3D volumes Ω. We modified the occupancy network from Mescheder et al. (2019) to perform this task, which is a widely adopted and highly competitive model for single-view 3D reconstruction. During training and inference, we replace the input RGB image with a shadow image. We also loaded the occupancy decoder pre-trained on ShapeNetChang et al. (2015) and supervised further training with 3D ground truth. *(ii) Nearest Neighbor:* We experiment with a nearest neighbor approach. For each shadow image $\mathbf{s}$ in the test set, we search in the training set for the object whose shadow $\mathbf{s}'$ minimizes $||\mathbf{s} - \mathbf{s}'||_2$ and use the corresponding 3D object as prediction. *(iii) Random:* We compute chance by selecting a random 3D object from the training set. *(iv) Full Image:* For analysis purposes, we also compare against an off-the-shelf 3D reconstruction method Mescheder et al. (2019) that is able to see the entire image (not just the shadows). Since this approach has more information than our method, we do not expect to outperform it. However, this comparison allows us to quantify the amount of information lost when we only operate with shadows.

[1]Named after Akira Kurosawa's movie, which translates to "shadow warrior".