Shape Reconstruction using Differentiable Projections and Deep Priors

Matheus Gadelha Rui Wang Subhransu Maji
University of Massachusetts, Amherst
{mgadelha, ruiwang, smaji}@cs.umass.edu

Abstract

We investigate the problem of reconstructing shapes from noisy and incomplete projections in the presence of viewpoint uncertainties. The problem is cast as an optimization over the shape given measurements obtained by a projection operator and a prior. We present differentiable projection operators for a number of reconstruction problems which when combined with the deep image prior or shape prior allows efficient inference through gradient descent. We apply our method on a variety of reconstruction problems, such as tomographic reconstruction from a few samples, visual hull reconstruction incorporating view uncertainties, and 3D shape reconstruction from noisy depth maps. Experimental results show that our approach is effective for such shape reconstruction problems, without requiring any task-specific training.

1. Introduction

Consider the problem of reconstructing a 3D shape from silhouettes. The classic visual hull algorithm that intersects the visible volumes from each viewpoint is easy to implement but is sensitive to errors in viewpoint estimation and silhouette noise. A Bayesian approach for this problem would be to add appropriate priors over the shape and viewpoint estimates and perform posterior inference. This is challenging for two reasons. First, the search space of 3D shape is large since there is no compact shape basis to search over for general shapes. Second, Bayesian inference is typically expensive for high-dimensional data.

To this end we present *differentiable projection operators* $\mathcal{T}$ and *deep shape priors* for which Bayesian inference can be performed via stochastic gradient descent and their variants [23]. While many priors exist, of interest is the "deep shape prior" of Ulyanov *et al*. [21] which showed that the space of natural images $\boldsymbol{x}$ can represented as a parametric family $f_{\boldsymbol{\theta}}(\eta)$ where f is a convolutional network, θ its parameters, and η is a fixed input. Their work showed that search over natural images can be replaced by a search over the parameters of the network $\boldsymbol{\theta}$, which can be efficiently done via gradient descent.

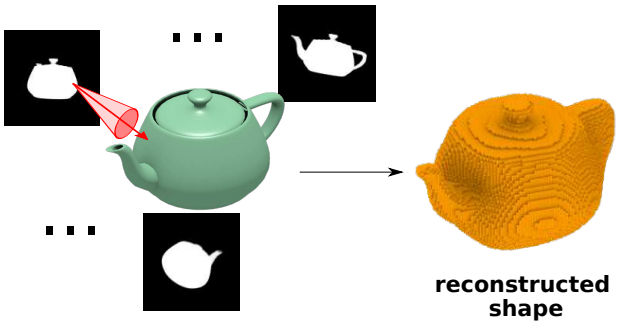

Figure 1: **Shape reconstruction from binary images with uncertain viewpoints.** We propose to use deep networks together with differentiable projection operators for shape reconstruction. Our approach leverages the shape prior induced by neural networks to reconstruct shapes from projections without any learning procedure. Additionally, our approach can use differentiable operators to reconstruct shapes under noisy projection measurements, like perturbed viewpoint information.

Our work takes this idea further. First, we endow the deep image prior with 3D convolutions resulting in a deep shape prior. Second, we incorporate differentiable projection operators $\mathcal{T}$ that model projection measurements, such as silhouettes, given projection parameters ϕ such as viewpoints. Thus inferring a shape $\boldsymbol{x}$ given noisy projection measurements $\boldsymbol{y}$ reduces to the following optimization over network parameters $\boldsymbol{\theta}$ and projection parameters ϕ:

$$\min_{\phi, \boldsymbol{\theta} \in \mathbb{R}^D} E\left(\boldsymbol{y}, \mathcal{T}(f_{\boldsymbol{\theta}}(\boldsymbol{\eta}), \phi)\right) + P(\phi), \tag{1}$$

where $P(\phi)$ is a prior over projection parameters, which is often a simple function. We show that for a number of shape construction problems such as tomographic reconstruction, shape from silhouettes or depth maps, it is possible to construct projection operators using existing neural network building blocks that are differentiable with respect to both the input and projection parameters. Thus the objective can be minimized using "backpropagation" machinery, which is generally much faster than Bayesian inference using Markov Chain Monte Carlo (MCMC) techniques.

Apart from choosing the network architecture and the projection operator, the approach does not require any task-specific training. Nevertheless, it yields compelling results for tomographic reconstruction in the low sampling regime, where it outperforms a state-of-the-art approach based on iterative BM3D [13]. Our work also shows that the deep image prior generalized to 3D volumes is effective at modeling 3D shapes. In problems such as visual hull reconstructions, or reconstruction from depth maps, we can accurately estimate the 3D shape of an object from only a few views, even when there are uncertainties in the view estimates, or when depth maps are corrupted by noise. The reconstruction results are significantly better than handcrafted priors. These tasks are illustrated in Figures 3-9.

2. Related work

In this section we briefly summarize techniques for solving inverse problems for image and volumetric reconstruction of the form:

$$\min_{\boldsymbol{x}\in\mathcal{X}} P(\boldsymbol{x}) + E(\boldsymbol{y}, \mathcal{T}(\boldsymbol{x})). \tag{2}$$

The data term E and the projection operator $\mathcal{T}$ are application specific, but there is considerable flexibility on modeling the prior term P. These include smoothness priors such as total variation (TV) [17] and L_0 gradients [25], Gaussian mixture models over patches [29], denoising autoencoders [22]. The deep image prior [21] represents images as the output convolutional network with random parameters from a fixed (random) input. The authors showed that outputs of networks consisting of several convolutional and pooling layers, followed by several deconvolutional layers with few or no skip connections in between tend to generate natural images. Recently, an extension to the deep image prior shows that it is asymptotically equivalent to a Gaussian Process [5]. This suggests a Bayesian approach to the problem: conducting posterior inference through Langevin dynamics avoids the need for early stopping and improves results for denoising and inpainting tasks. The deep image prior is also related to procedural priors such as bilateral filtering [20], non-local means [3], or block matching 3D (BM3D) [7]. These models use non-local self-similarity of patches in images to collectively denoise them.

For complex projection operators $\mathcal{T}$ involving noisy and incomplete measurements $\boldsymbol{y}$, applying procedural priors is non-trivial. Suppose $\boldsymbol{y}$ and $\boldsymbol{z}$ denote the observed and unobserved projection measurements corrupted by noise: $(\boldsymbol{y}, \boldsymbol{z}) = T(\boldsymbol{x}) + \delta$. For example $\boldsymbol{y}$ could denote the subset of frequencies in the Fourier transform, or projections of data in a compressed sensing application. Maggioni *et al*. [13] proposed the following iterative scheme:

1. Estimate $\boldsymbol{x}$ by inverting the measurement $\boldsymbol{x}^{(k)} = \mathcal{T}^{-1}(\boldsymbol{y}, \boldsymbol{z}^{(k)})$ starting from $\boldsymbol{z}^{(1)} = 0$.
2. Denoise $\boldsymbol{x}^{(k)}$ using BM3D to obtain $\boldsymbol{x}^{(k+1)}$.
3. Re-estimate $(., \boldsymbol{z}^{(k+1)}) = \mathcal{T}(\boldsymbol{x}^{(k+1)}) + \delta^{(k)}$. Note that only the unobserved part of projection is estimated keeping $\boldsymbol{y}$ fixed across iterations.

The iterative BM3D can be applied to problems where the support of $\mathcal{Y}$ is small. This procedure is related to the alternating direction method of multipliers (ADMM) [2] which has been applied for solving linear inverse problems of the form: $\min_{\boldsymbol{x}} ||\boldsymbol{y} - A\boldsymbol{x}||_2^2 + \lambda P(\boldsymbol{x})$. ADMM solves the augmented Lagrangian $\mathcal{L}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{u})$:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{u}) = ||\boldsymbol{y} - A\boldsymbol{z}||_2^2 + \lambda P(\boldsymbol{x}) + \frac{\rho}{2}||\boldsymbol{x} - \boldsymbol{z} + \boldsymbol{u}||_2^2$$

over auxiliary variables $\boldsymbol{z}$ and $\boldsymbol{u}$ for $\rho > 0$ by alternatively optimizing $\boldsymbol{x}$, $\boldsymbol{z}$, and $\boldsymbol{u}$ as:

$$\begin{aligned}
\boldsymbol{x}^{(k+1)} &\leftarrow \operatorname*{argmin}_{\boldsymbol{x}} \lambda P(\boldsymbol{x}) + \frac{\rho}{2}||\boldsymbol{x} - \boldsymbol{z}^{(k)} + \boldsymbol{u}^{(k)}||_2^2 \\
\boldsymbol{z}^{(k+1)} &\leftarrow \operatorname*{argmin}_{\boldsymbol{z}} ||\boldsymbol{y} - A\boldsymbol{z}||_2^2 + \frac{\rho}{2}||\boldsymbol{x}^{(k+1)} - \boldsymbol{z} + \boldsymbol{u}^{(k)}||_2^2 \\
\boldsymbol{u}^{(k+1)} &\leftarrow \boldsymbol{x}^{(k+1)} - \boldsymbol{z}^{(k+1)} + \boldsymbol{u}^{(k)}
\end{aligned}$$

The optimization decouples the reconstruction and the prior. The first involves inference with an image prior and squared-loss term. The second objective is quadratic in $\mathbf{z}$ can be solved with conjugate gradient decent. The decoupling allows use of explicit or implicit priors, as well as learned proximal projection operators [4,26] $\textbf{proj}(\boldsymbol{z} - \boldsymbol{u}, \rho)$ that map a vector $\boldsymbol{z} - \boldsymbol{u}$ to $\boldsymbol{x}$ in the manifold of natural images within a distance ρ from it, similar to a denoising autoencoder, to solve the inverse problem.

Finally, a class of approaches directly learn the inverse mapping $\mathcal{G} : \mathcal{Y} \rightarrow \mathcal{X}$ using rich parametric models such as a neural network in a fully-supervised manner. These models amortize inference during training and enable efficient inference given noisy measurements. Such models have been successfully applied for various inverse problems such as super resolution [8], denoising [24], colorization [12, 28], and estimating depth and normals from images [9]. However a disadvantage is the architecture and parameters of the model are likely to be specific to the noise and projection operators, which require separate training for each task.

Closely related to this work, recent approaches have employed geometric transformations on deep features to generate novel views of a 3D object [14, 19]. In contrast to our approach, those techniques do not explicitly define the projection operators – they are parameterized by a deep neural network. As a consequence, the inferred representation does not directly correspond to a 3D shape, but to a higher lever representation learned by the model.

3. Method

Our approach for Bayesian inference will be to optimize the objective in Equation 1 using Stochastic Gradient De-

scent (SGD). This corresponds to a Maximum Likelihood Estimate (MLE), or Maximum A-Posteriori (MAP) estimate if priors over parameters $\boldsymbol{\theta}$ are added. Although more sophisticated schemes for SGD based posterior sampling exist [5, 23], we find that SGD works reasonably well for the problems we consider.

Solving reconstruction problems with SGD requires formulating differentiable projection operators and differentiable priors over the shapes. We use the deep image prior for image-based reconstruction tasks, and a 3D convolutional version for shape reconstruction tasks. In earlier work the deep image prior was used to solve a number of reconstruction problems with linear measurements [21]. For example in denoising the projection operator is the identity transformation, while in inpainting the projection operator is a mask indicating which pixels are present and absent. In this section, we present three differentiable projection operators that can be combined with deep neural networks for reconstructing shapes from partial and noisy observations.

3.1. Radon Projection ($\mathcal{T}_R$)

In [15], Radon proposed the utilization of the inverse of an integral transform to reconstruct images from a CT scan. The forward version of this transform is known as Radon transform R and can be described by the following:

$$R(\phi, r) = \int_L s(x, y)dl,\ L = \{(x, y)|x\sin\phi - y\cos\phi = r\} \tag{3}$$

where s represents a density function, ϕ is the angle of projection, and this transform represents data obtained as the output of a CT scan. Let $T_\psi(s)$ be an operator that rotates s by ψ degrees, i.e. $T_\psi(s)(x, y) = s(x\cos\psi - y\sin\psi, x\sin\psi + y\cos\psi)$. Plugging this in Equation (3) we have:

$$\begin{aligned} R(\phi, r) &= \int_L T_\psi(s)(x, y)dl \\ L &= \{(x, y)|x\sin(\phi + \psi) - y\cos(\phi + \psi) = r\} \end{aligned}$$

Taking $\psi = -\phi$:

$$R(\phi, r) = \int_L T_{-\phi}(s)(x, y)dl, \quad L = \{(x, y)|y = r\} \tag{4}$$

$$R(\phi, r) = \int_{\mathbb{R}} T_{-\phi}(s)(x, r)dx \tag{5}$$

In practice, s is represented by image and $T_{-\phi}(s)$ is computed by rotating a regular grid and resampling the image as described in [11]. Specifically, let $I_{i,j}^{(\phi)}$ be the value of the pixel i, j in the image formed by s rotated by $-\phi$ degrees, the discrete version of the Radon transform is:

$$R(\phi, r) = \sum_{i=1}^{S} I_{i,r}^{(\phi)}, \tag{6}$$

where S is the size of the image. Notice that the result of the Radon transform R is also an image (called sinogram and is parametrized by ϕ and r) as can be seen in Figure 3. Finally, our operator $\mathcal{T}_R$ receives an image I of size $S \times S$, a set of values ϕ representing the projection angles and outputs an image of size $S \times |\phi|$. The process is differentiable and can be implemented as a sum over one dimension of multiple rotated images.

3.2. Silhouette Projection ($\mathcal{T}_S$)

Shape reconstruction from silhouettes consists in the following problem: given a set of silhouette images of the same object from different views, estimate the 3D shape of the object. Silhouette projection can be formulated as a differentiable operator $\mathcal{T}_S(V, \phi)$. To do so, we represent 3D shape as a voxel grid V, and the projection $\mathcal{T}_S(V, \phi)$ generates a silhouette of the shape V captured from a view ϕ. The formulation of $\mathcal{T}_S$ follows [10]. Specifically, let $V : \mathbb{Z}^3 \to [0, 1] \in \mathbb{R}$ be the voxel grid, representing the occupancy value at a given integer 3D coordinate $c = (i, j, k)$. The rotated version of the voxel grid $V(c)$ is defined as $V_\phi(c) = \Phi(V, T_\phi(c))$, where $T_\phi(c)$ is the coordinate obtained by rotating c around the origin according to ϕ and $\Phi(V, c)$ is a procedure that samples a value of V in a position c – trilinear or nearest neighbor sampling.

The next step consists in performing the projection to create an image from the rotated voxel grid. This is done by applying the projection operator $P(V)_{i,j} = 1 - e^{-\tau\sum_k V(i,j,k)}$. The intuition behind this operator is similar to the idea of the Radon transform: compute a line integral of the occupancy function V along each line of sight (assuming othographic projection), with the difference that here we apply an exponential falloff to create a smooth and differentiable function. The smoothness can be controlled by the parameter τ: bigger values result in binary images. If there all voxels along the line of sight are empty, the projection results in a value of 0; as the number of non-empty voxels increases, the value approaches 1. Combined with the rotated version of the voxel grid, we define our final projection operator as: $\mathcal{T}_S(V, \phi)_{i,j} = 1 - e^{-\tau\sum_k V_\phi(i,j,k)}$ where i, j is the pixel coordinate of the resulting image.

3.3. Depth Image Projection ($\mathcal{T}_D$)

Given a 3D shape represented as a voxel occupancy grid V and a view ϕ, the depth image captures the distance values from the viewpoint to the visible points on the shape. This is useful in practical applications as depth images are frequently captured by LiDAR and similar depth sensors. Here, we demonstrate that the depth projection operator can be built upon the silhouette projection operator. To do so, we first define a visibility function $A(V, \phi, c)$ that describes whether a given voxel c inside the grid V is visible, when

seen from a view ϕ:

$$A(V, \phi, i, j, k) = \exp\left\{ -\tau \sum_{l=1}^{k} V_\phi(i, j, l) \right\} \quad (7)$$

Intuitively, this is the complement of the silhouette projection, the difference is that we are incrementally accumulating the occupancy (from the first voxel on the line of sight) as we traverse the voxel grid, instead of summing all voxels on entire the line of sight. If voxels on the path from the first to the current voxel are all empty, the value of A is 1 (indicating the current voxel is 'visible' to the view ϕ). If there is at least one non-empty voxel on the path, the value of A will be close to 0 (indicating this voxel is not visible).

Now that we have the visibility value of each voxel, the depth value of a pixel in the projected image is simply the line integral of A along the line of sight: $D(i, j) = \sum_k A(V, \phi, i, j, k)$. This accumulates the number of voxels along the entire line of sight that are visible, therefore it gives the depth value. Refer to Figure 2 for illustrations.

While using this operator along with a neural network, we found that it works better if we apply an exponential decay. Thus, we can define the depth projection operator $\mathcal{T}_D$ as follows:

$$\mathcal{T}_D(V, \phi)_{i,j} = 1 - \exp\left\{ -\sum_{k} A(V, \phi, i, j, k) \right\} \quad (8)$$

This smoothly maps the depth value to the range between [0,1]. Specifically, it maps a depth value of 0 to 0, and infinity to 1, while still remaining a differentiable operator.

4. Experiments

This section presents the results of applying our shape projection operators along with deep shape priors for three reconstruction tasks.

Network Architecture. In the volumetric reconstruction experiments (i.e. reconstructing 3D shapes from silhouette images and depth images respectively), the network architecture is a fully convolutional UNet [16] where the encoder has 5 layers with 8, 16, 32, 64 and 128 filters. The decoder is a mirrored version of the encoder and skip connections are applied just in the 2 innermost layers. The upsampling is done through bilinear/trilinear interpolation followed by a convolution. All convolutions have filter size 3 and are followed by batch normalization and ReLU activation function. The input to the network is a tensor of the same size as the output, and its values sampled from $\mathcal{N}(0, 1)$. In all the experiments, we used Adam optimizer with learning rate=10^{-2}.

For the image reconstruction (i.e. tomography) we doubled the number of filters in each layer keeping the rest of the network architecture identical to account for higher spatial frequency of the underlying signal. The only other difference between the network that produces images and the one that produces voxel grids is that the convolutional operations are performed in 2D instead of 3D. Even though the network can be used to generate data of any size (since it is fully convolutional), in our experiments we set our image resolutions to 256×256 and voxel resolution to 128^3.

4.1. Tomography Reconstruction

In tomographic reconstruction our goal is to invert the sinograms as described in Section 3. With deep image prior the reconstruction involves solving the following optimization problem:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^D} ||R - \mathcal{T}_R(f_{\boldsymbol{\theta}}(\boldsymbol{\eta}))||_1, \quad (9)$$

where f is our neural network described above, $\boldsymbol{\eta}$ is its noise input, and R is the input sinogram (which may have low angular sampling rate and/or be corrupted by noise). To test the ability of our algorithm when handling challenging input, we use a low angular sample rate ($n = 30$) and simulate noisy sinograms by adding a Gaussian noise of $\sigma = 1$. Figure 3 shows the reconstruction results of the Shepp-Logan phantom image [18] and two separate slices of a sample from the BrainWeb database [6]. These images have been commonly used to evaluate CT reconstruction algorithms. For each reconstruction we compute the structured similarity (SSIM) index and PSNR values with respect to the groundtruth image (higher is better).

The standard solution for tomography is Filtered Back Projection (FBP): it inverts the Radon transform using the Fourier slice theorem. When angular sampling rate is low, the reconstruction using FBP turns out to have severe aliasing artifacts as seen in Figure 3 third column. The TV prior significantly improves the reconstructions for all three images. The iterative BM3D approach [13] described in Section 2 was run for 100 iterations. We noticed that the PSNR values converged after 100 iterations with the largest gains in PSNR in the first 20 iterations. Note that running BM3D on the FBP reconstruction corresponds to one iteration of this approach. For the deep prior we obtain results by running 2000 gradient steps. Compared to iterative BM3D, the deep prior produces reconstructions with significantly better SSIM values and comparable or better PSNR values (last two columns in Figure 3). The relatively poor performance of BM3D may be because the aliasing noise in CT reconstructions tends to be more structured and less like natural image noise when compared to the noise observed in image denoising applications. It takes many iterations for the iterative BM3D algorithm to get rid of the artifacts produced by the inverse radon transform but this causes smoothing of the underlying structures leading to lower SSIM scores.

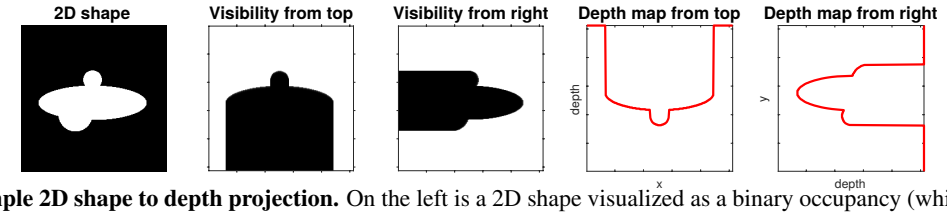

Figure 2: **A example 2D shape to depth projection.** On the left is a 2D shape visualized as a binary occupancy (white is occupied). The visibility map for each pixel from the top and right views are shown next – a pixel is white (value=1) if it is visible. The depth maps are obtained by summing the visibility maps along the vertical and horizontal directions for the top and right views receptively.

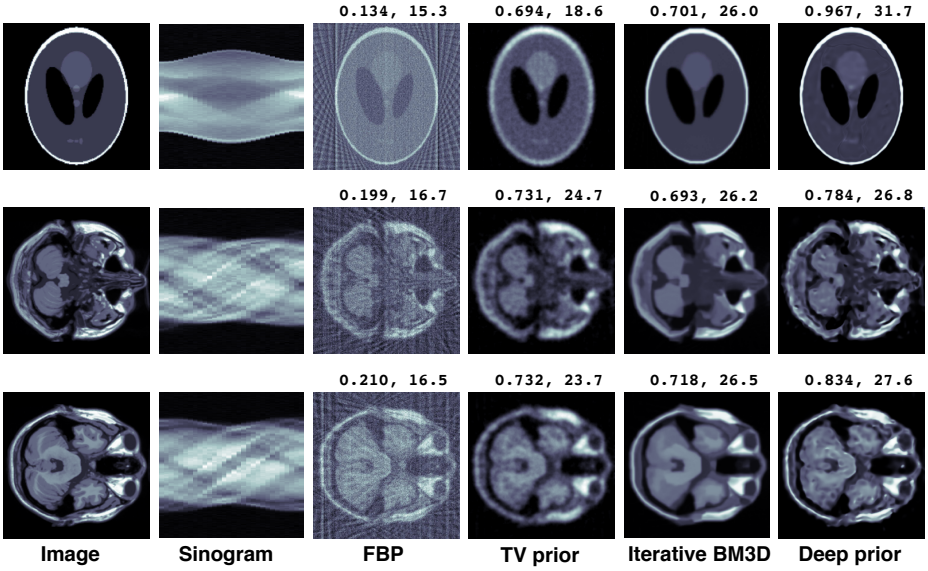

Figure 3: **Tomographic reconstruction results** from sinograms (radon transforms) sampled with $n = 30$ angles and noise ($\sigma = 1$). The sinogram is rescaled to the image size with nearest neighbor interpolation for visibility. From left to right in each row is the noise free image, the noisy sinogram, reconstruction with the filtered backprojection (FBP), TV prior, BM3D, and deep image prior. The SSIM and PSNR are shown for each approach on top of the corresponding figure. Our approach outperforms the other learning free baselines by a significant margin. *Zoom in for details.*

4.2. Shape-from-Silhouette 3D Reconstruction

For 3D shape reconstruction from silhouette images, we employ the 3D convolutional neural network described as before to generate a voxel grid V where each voxel represents an occupancy value. The output of the network is then passed to the projection operator $\mathcal{T}_S$ along with a view direction ϕ. Given a set of N viewpoints $\boldsymbol{\phi} = \{\phi_1, \phi_2, ..., \phi_n\}$ and its associated images I_{ϕ_i}, our problem is described by the following optimization:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^D} \sum_{i=1}^{N} ||I_{\phi_i} - \mathcal{T}_S(f_{\boldsymbol{\theta}}(\boldsymbol{\eta}), \phi_i)||_1, \tag{10}$$

where f is our neural network and $\boldsymbol{\eta}$ its noise input. We solve this minimization using gradient descent and then use $f_{\boldsymbol{\theta}}(\boldsymbol{\eta})$ to generate our final reconstruction. The results can be seen in Figure 4. Even with a small number of silhouette images, our method is able to reconstruct reasonable 3D shapes. The viewpoints for this example are chosen by evenly rotating the object along the horizontal axis (e.g. with 4 views, each view is 90 degrees apart; with 8 views, each is 45 degrees apart and so on). A baseline approach for this problem is space carving, which takes the intersection of all the projected views to generate the occupancy grid. We show a qualitative comparison with space carving in Figure 5. Space carving provides reasonable reconstructions for most of the shapes, but some of the objects contain artifacts like creases or even missing parts. On the other hand, the deep shape prior tends to create overly smooth shapes, which sometimes means removing some parts of the object (chairs in Figure 5) or adding content where should exist a sharp boundary (lamp in Figure 5).

View uncertainties. In the previous formulation, we assume that the set viewpoints $\boldsymbol{\phi}$ corresponds exactly to the observed views. However, a more realistic scenario is to assume that we are given a set of noisy viewpoint measurements. In this case, besides estimating the parameters of the

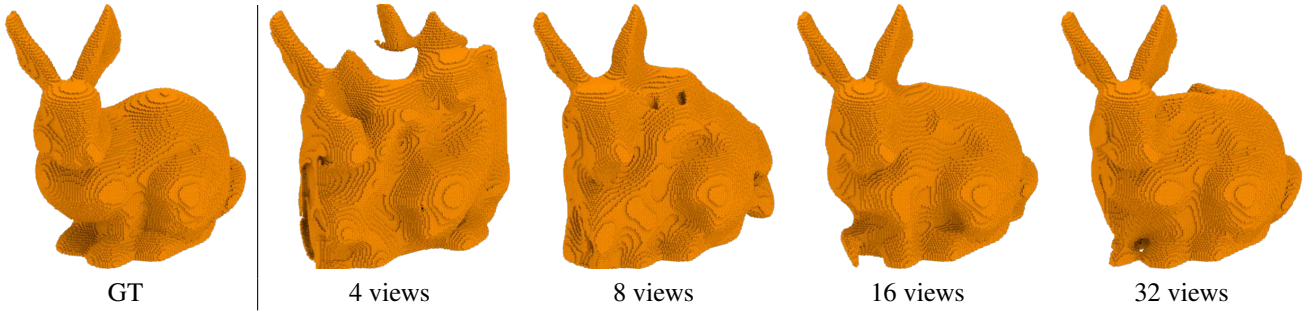

Figure 4: **Effect of the number of views in the reconstruction from silhouettes.** 3D shape reconstructed from silhouette images of the same object. Even without having access to enough 3D information, our method is still capable of generating plausible shapes.

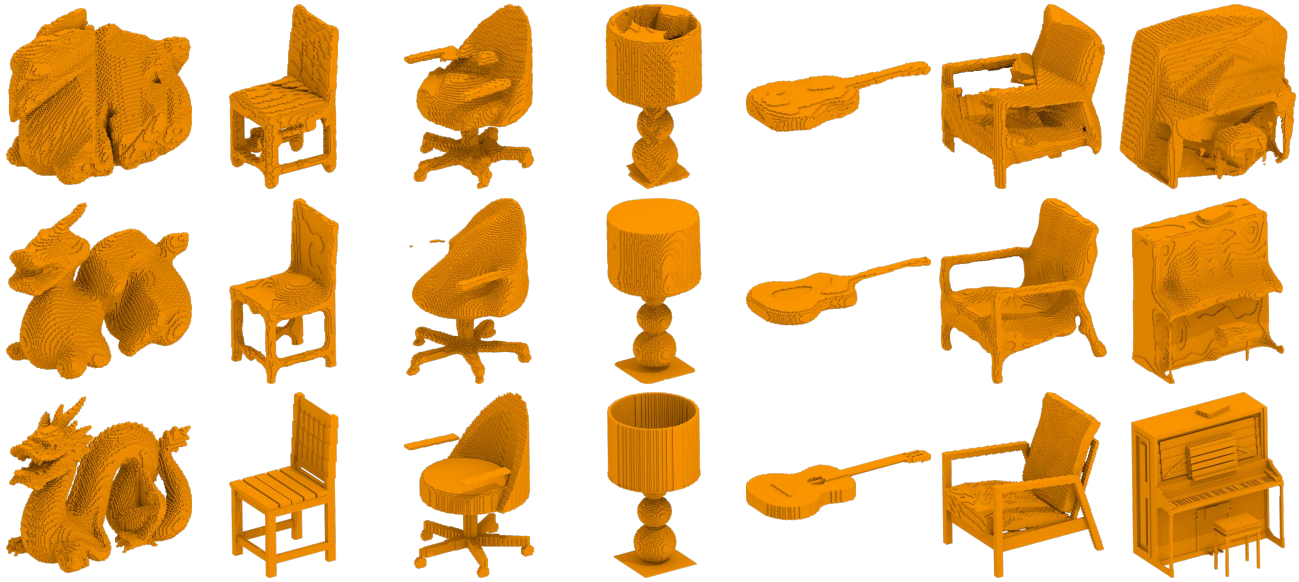

Figure 5: **Reconstruction from silhouettes without viewpoint noise.** 3D shapes reconstructed from 8 silhouette images of the same object. Viewing angles were sampled uniformly at random. Top row using space carving baseline, middle row using the deep image prior, bottom row is ground-truth.

network predicting the shape, we are also looking to estimate viewpoints $\hat{\boldsymbol{\phi}}$. We assume that the noisy viewpoints $\boldsymbol{\phi}$ are sampled independently from $VonMises(\boldsymbol{\phi}^*, \kappa)$, where κ is the dispersion of the VonMises distribution with mean $\boldsymbol{\phi}^*$ (the ground-truth viewpoints). This leads to adding an extra term to Equation 10 and also optimizing over the predicted viewpoints $\hat{\boldsymbol{\phi}}$:

$$\min_{\hat{\boldsymbol{\phi}},\boldsymbol{\theta}} \sum_{i=1}^{N} ||I_{\boldsymbol{\phi}_i^*} - \mathcal{T}_S(f_{\boldsymbol{\theta}}(\boldsymbol{\eta}), \hat{\boldsymbol{\phi}}_i)||_1 + \lambda \cos(\hat{\boldsymbol{\phi}}_i - \boldsymbol{\phi}_i), \quad (11)$$

where λ is the weight of the viewpoint regularization term. We use $\lambda = 0.1$ in our experiments. Notice that our projection operator is fully differentiable with respect to the viewpoint parameters and can be easily implemented using automatic differentiation packages.

Evaluation To evaluate our approach we selected twelve meshes from standard benchmarks. Three of them are well know 3D shapes (Stanford bunny, dragon and Utah teapot) while the others were selected from 9 different categories of the ModelNet40 dataset [27]. We voxelize those shapes filling their interior to generate binary occupancy grids of resolution 108^3. Those voxel grids will correspond to our ground-truth data. Our network generates 128^3 occupancy grids, but we use data in a smaller resolution to zero-pad the volume and avoid artifacts in the boundaries. Next, we randomly sample 8 viewpoints and render a binary image I_{ϕ_i} from each sampled view. Since we want to evaluate the ability of the methods to reconstruct the 3D shape while dealing with view uncertainty, we sample views $\hat{\boldsymbol{\phi}}$ from $VonMises(\boldsymbol{\phi}, \kappa)$ and associate them with the corresponding binary images. We use $\kappa = 100$ for all the experiments.

Figure 6: **Shape-from-silhouette reconstruction using captured images.** For this glass object, we photographed 4 views, with 45° angle apart, against a uniform background color. We then applied background-color removal and converted each image to binary silhouette image. The first reconstructed model is the result using our deep prior, whereas the second is the result using the space carving baseline.

	plane	bunny	car	desk	dragon	guitar	lamp	piano	plant	sofa	table	teapot	mean
Ours	0.35	**0.88**	**0.72**	**0.81**	0.59	0.64	**0.62**	**0.86**	**0.79**	**0.78**	**0.82**	**0.84**	**0.72**
Carving	0.49	0.77	0.59	0.41	0.55	0.51	0.26	0.64	0.58	0.51	0.44	0.83	0.55
Carving*	**0.51**	0.85	**0.72**	0.50	**0.62**	**0.71**	0.28	0.60	0.61	0.57	0.55	0.81	0.61

Table 1: **3D reconstruction from silhouettes with uncertain viewpoints.** Intersection over union of the reconstructed occupancy from 12 different shapes. We randomly sample viewpoints to generate 8 binary images for each shape. Those viewpoints are slightly perturbed before being used by the methods, except for the last (Carving*) which corresponds to using space carving without noisy viewpoints. Our approach significantly outperforms the space carving baseline in all scenarios.

In other words, even though an image I_ϕ was rendered from a viewpoint ϕ, we assign a slightly perturbed viewpoint $\hat{\phi}$ to this image . Finally, we use the binary images I_ϕ and the perturbed viewpoints $\hat{\phi}$ to reconstruct the 3D shape by minimizing the objective described in Equation 11. This is done through 500 steps of gradient descent. We compare our approach with a space carving baseline and report the intersection over union of the estimated occupancy grids in Table 1.

Our method outperforms vanilla space carving even when the viewpoints given are not perturbed, which demonstrates the robustness of our method to viewpoint perturbations. Figure 7 shows a qualitative comparison of the reconstructed shapes. Our approach reconstructs the shapes with high fidelity, preserving details and thin structures. On the other hand, space carving ends up reconstructing objects with missing parts and and rough structures as we can observe in Figure 7.

Reconstructions using captured images. We have also evaluated our method using images captured from a camera. Results are presented in Figure 6. The subject is a glass object, for which we photographed 4 views evenly spaced with 45° horizontal rotation angle apart from each other, against a uniform background color. We then use [1] to remove background and convert each image to a binary silhouette image. We compare results using our method with standard visual hull (i.e. space carving). As can be observed, our method leads to smooth reconstructions and the resulting objects look more natural. In contrast, the visual hull results contain artifacts and sharp transitions around changing views, which would require significantly more number of views to eliminate.

4.3. Shape-from-Depth Images 3D Reconstruction

The setup for 3D reconstruction from depth images is the same for the binary images except for the use of the projection $\mathcal{T}_D$ instead of $\mathcal{T}_S$. All the input depth images have their range scaled to be in $[0, 1]$ using the exponential map in Equation (8). We analyzed the ability of the method to reconstruct 3D shapes from depth images perturbed by different levels of Gaussian noise while using 4 views. Results can be seen in Figure 8. Additionally, we analyzed the reconstruction quality while varying the number of views. Results are presented in Figure 9. For these experiments, we kept the noise level very high ($\sigma = 0.1$). We notice that even when dealing with very noisy projections, our method is able to reconstruct high quality shapes if enough views are given.

5. Conclusion

We showed that by combining the deep image or volumetric prior with differentiable projection operators, signals

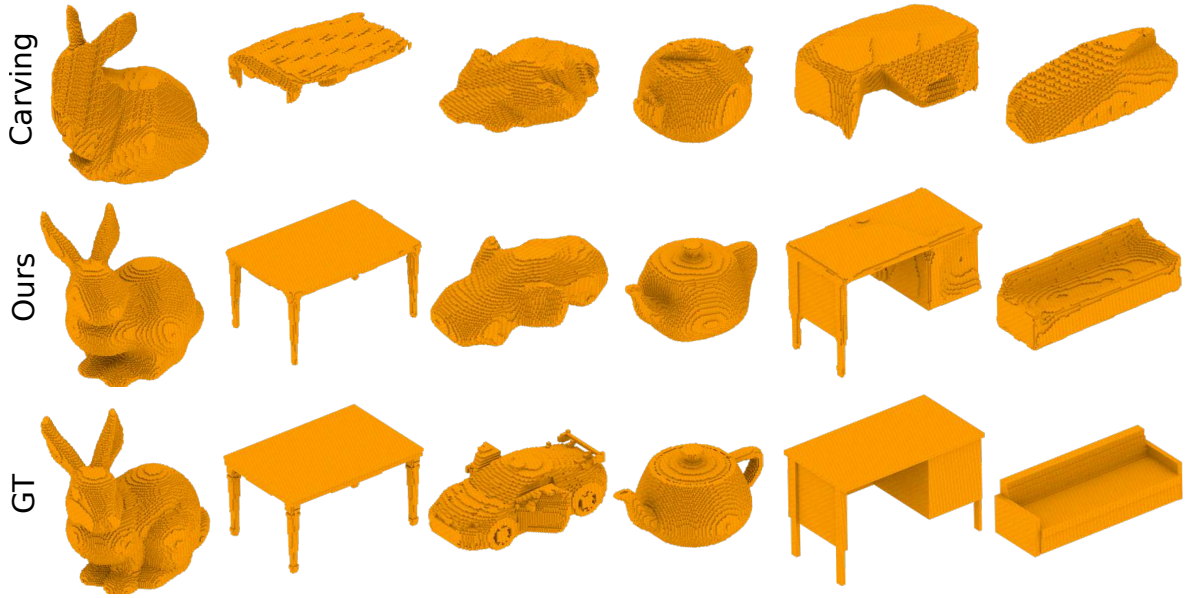

Figure 7: **Shape-from-silhouette reconstruction with perturbed viewpoints.** Results for the space carving baseline in the first row, our method in the second row, ground-truth shapes in the third row. Our results are computed minimizing Equation 11 through 500 gradient descent steps. Our method is capable of updating the initial viewpoint parameters and is capable to recover from imprecise viewpoint assignment. The space carving baseline is not robust to viewpoint perturbations which means it ends up carving the wrong regions of the volume, leading to poor reconstructions and eliminating thin object structures.

Figure 8: **Effect of noise in the reconstruction.** 3D shape reconstructed from 4 noisy depth images of the same object. The variance of the Gaussian noise increases from left to right. Shape prior can reconstruct high quality shapes even with considerable amount of noise.

Figure 9: **Effect of the number of views in the reconstruction from depth images.** 3D shape reconstructed from very noisy ($\sigma = 0.1$) depth images of the same object. On the left, example of the input depth images. If provided enough views, our method is able to reconstruct high quality shapes even from highly noisy inputs.

can be reconstructed from a few noisy projection measurements using stochastic gradient descent. The approach is learning free and can be used as a generic prior. Nevertheless, with a relatively simple network architecture our approach outperformed several handcrafted and procedural priors for image based and volumetric reconstruction tasks. Although we presented results for tomography and for shape reconstruction from silhouettes and depth maps, the approach can be used whenever the rendering or measurement process is differentiable. These include problems such as estimating shape from shading and geometry from multiple shaded images.

A potential issue is the use of volumetric representations for shapes which incurs high memory requirements and longer running times. A possible line of research is to investigate shape priors for more compact 3D representations like point clouds or multi-view. Combining deep priors with work on differentiable computer graphics pipelines opens up the possibility of applying this approach for solving inverse problems in many applications.

Acknowledgments. This work is supported in part by NSF grants IIS-1749833 and IIS-1423082. Our experiments were performed in the UMass GPU cluster obtained under the Collaborative Fund managed by the Massachusetts Technology Collaborative.

References

[1] https://clippingmagic.com/. 7

[2] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning*, 3(1):1–122, 2011. 2

[3] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *Computer Vision and Pattern Recognition (CVPR)*, 2005. 2

[4] JH Rick Chang, Chun-Liang Li, Barnabas Poczos, BVK Vijaya Kumar, and Aswin C Sankaranarayanan. One network to solve them allsolving linear inverse problems using deep projection models. *arXiv preprint*, 2017. 2

[5] Zezhou Cheng, Matheus Gadelha, Subhransu Maji, and Daniel Sheldon. A Bayesian Perspective on the Deep Image Prior. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[6] Chris A Cocosco, Vasken Kollokian, Remi K-S Kwan, G Bruce Pike, and Alan C Evans. Brainweb: Online interface to a 3D MRI simulated brain database. In *NeuroImage*, 1997. 4

[7] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007. 2

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision (ECCV)*, 2014. 2

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems (NIPS)*, 2014. 2

[10] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3D Shape Induction from 2D Views of Multiple Objects. *International Conference on 3D Vision (3DV)*, 2017. 3

[11] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 3

[12] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[13] Matteo Maggioni, Vladimir Katkovnik, Karen Egiazarian, and Alessandro Foi. Nonlocal transform-domain filter for volumetric data denoising and reconstruction. *IEEE transactions on image processing*, 22(1):119–133, 2013. 2, 4

[14] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. HoloGAN: Unsupervised learning of 3D representations from natural images. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[15] J. Radon. On the determination of functions from their integral values along certain manifolds. *IEEE Transactions on Medical Imaging*, 5(4):170–176, Dec 1986. 3

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[17] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992. 2

[18] Lawrence A Shepp and Benjamin F Logan. The Fourier reconstruction of a head section. *IEEE Transactions on nuclear science*, 21(3):21–43, 1974. 4

[19] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. DeepVoxels: Learning Persistent 3D Feature Embeddings. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[20] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *International Conference on Computer Vision (ICCV)*, 1998. 2

[21] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Deep Image Prior. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3

[22] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010. 2

[23] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning (ICML)*, 2011. 1, 3

[24] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In *Advances in neural information processing systems (NIPS)*, 2012. 2

[25] Li Xu, Cewu Lu, Yi Xu, and Jiaya Jia. Image smoothing via L0 gradient minimization. *ACM Transactions on Graphics (TOG)*, 30(6):174, 2011. 2

[26] Yan Yang, Jian Sun, Huibin Li, and Zongben Xu. Deep ADMM-Net for Compressive Sensing MRI. In *Advances in Neural Information Processing Systems (NIPS)*. 2016. 2

[27] A. Khosla F. Yu L. Zhang X. Tang J. Xiao Z. Wu, S. Song. 3D ShapeNets: A Deep Representation for Volumetric Shapes. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 6

[28] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[29] Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *International Conference on Computer Vision (ICCV)*, 2011. 2