Robust Bayesian Scene Reconstruction by Leveraging Retrieval-Augmented Priors

Herbert Wright[1] Weiming Zhi[2] Matthew Johnson-Roberson[2] Tucker Hermans[1,3]

***Abstract*—Constructing 3D representations of object geometry is critical for many downstream robotics tasks, particularly tabletop manipulation problems. These representations must be built from potentially noisy partial observations. In this work, we focus on the problem of reconstructing a multi-object scene from a single RGBD image, generally from a fixed camera in the scene. Traditional scene representation methods generally cannot infer the geometry of unobserved regions of the objects from the image. Attempts have been made to leverage deep learning to train on a dataset of observed objects and representations, and then generalize to new observations. However, this can be brittle to noisy real-world observations and objects not contained in the dataset, and cannot reason about their confidence. We propose BRRP, a reconstruction method that can leverage preexisting mesh datasets to build an informative prior during robust probabilistic reconstruction. In order to make our method more efficient, we introduce the concept of retrieval-augmented prior, where we retrieve relevant components of our prior distribution during inference. The prior is used to estimate the geometry of occluded portions of the in-scene objects. Our method produces a distribution over object shape that can be used for reconstruction or measuring uncertainty. We evaluate our method in both simulated scenes and in the real world. We demonstrate the robustness of our method against deep learning-only approaches while being more accurate than a method without an informative prior.**

I. Introduction

The ability to construct internal representations of its operating environment is key for robot autonomy. These representations need to be particularly fine-grained for robotic manipulation, which often requires closely interacting with and avoiding objects. These interactions make it necessary for robots to develop an understanding of the geometry within their vicinity. Explicit 3D representations of the geometry of the scene are often required for the robust usage of downstream grasping and motion planning algorithms. These representations must be built from observations that are both noisy and, due to occlusion, only contain partial information of the scene. In our case, we focus on the problem of building a 3D representation of multi-object scenes from a single RGBD camera image.

One approach to this problem is to train a neural network to *predict* the full geometry of an object given a partial view. Such approaches are able to use existing mesh datasets to more accurately infer the occluded backside of objects. Unfortunately, these approaches also tend to have a number of problems when used on real-world depth cameras. The presence of unknown objects, significant occlusion, noisy point clouds, or inaccurate segmentation prevents these methods from being reliably deployed in less-structured environments.

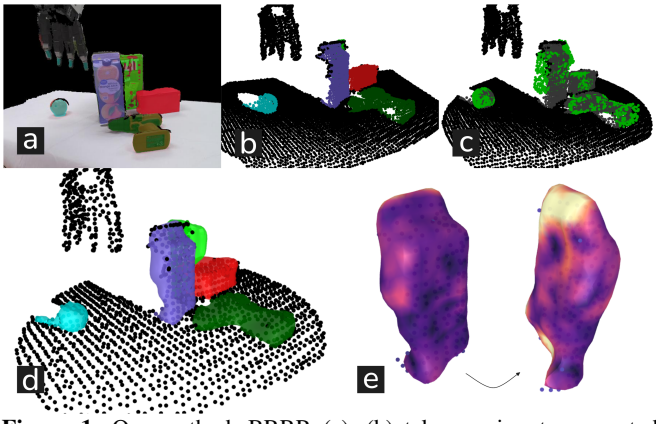

Figure 1: Our method, BRRP, (a), (b) takes an input segmented RGBD image, and (c) *retrieves* objects to use as a prior, which allow it to (d) reconstruct the scene as well as (e) capture principled *uncertainty* about object shape.

Deep learning methods for 3D reconstruction generally lack a sense of uncertainty about the shape of the object. This can be particularly detrimental when an object in the scene is only partially observed and we seek to factor in the geometry of observed and occluded regions. Such uncertainty can enable safer and more robust operation in a range of downstream tasks, such as robot grasping [1], [2], safe motion generation and active learning.

Another common approach to building full 3D representations from a partial view is to build a representation, solely from the observed data without considering prior information from mesh datasets. A common example of this is the Gaussian Process Implicit Surface [3] model. A more recent example is V-PRISM [4], which probabilistically maps tabletop scenes without using prior information. These reconstruction methods are more robust to unknown objects because they do not rely on any training distribution. However, this also means that they cannot reconstruct the unobserved backside of objects. Humans have the remarkable ability to infer the geometry of scenes based on prior experience, we seek to imbue robots with the same capability.

In this work, we introduce a novel Bayesian approach for robustly reconstructing multi-object tabletop scenes by leveraging object-level shape priors. We present **B**ayesian **R**econstruction with **R**etrieval-augmented **P**riors (BRRP). BRRP is resilient to many of the pitfalls of learning-based methods while still being able to leverage an *informative prior* to more accurately reconstruct known objects. To further improve efficiency, we introduce the idea of a retrieval-

[1] University of Utah Robotics Center and Kahlert School of Computing, University of Utah, Salt Lake City, UT, USA
[2] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[3] NVIDIA Corporation, Seattle, WA, USA

augmented prior, where we retrieve relevant components of our prior distribution based on classification results. We begin with an observed RGBD image with corresponding instance segmentations. Then, we compute an identification result that predicts which objects from our database should be retrieved to use as a prior during reconstruction. We use this prior along with a sampled likelihood to infer a posterior distribution over object shapes. Because we solve for a distribution, we can recover principled uncertainty about each object's shape. In practice, we use a pre-existing foundation model to perform the identification and use registration to make our prior pose invariant. An example scene and reconstruction can be seen in Figure 1.

We conduct experiments on BRRP in both procedurally generated scenes and in the real world. We quantitatively show that BRRP results in accurate reconstructions and that BRRP is robust to unknown objects in the generated scenes. We qualitatively show that BRRP is robust to noisy real world scenes collected from an RGBD camera, and is able to capture uncertainty within the reconstructions.

Our three primary contributions can be summarized as follows:

1) The formulation of retrieval-augmented priors for Bayesian inference.
2) A formulation for a prior over Hilbert maps with pose and scale invariance
3) A novel, robust *Bayesian* scene reconstruction method that utilizes prior information from existing mesh datasets

Our paper has the following organization. We overview related works in Section II. In Section III we cover the necessary mathematical preliminaries for our method. In Section IV, we introduce the concept of retrieval-augmented priors. Our method, BRRP, is introduced in Section V. Then, we perform experiments in Section VI before a conclusion in Section VII

II. Related Works

3D Representations. There are many different ways of representing 3D geometry of a scene. In the mapping literature, techniques such as truncated signed distance functions [5] build voxelized representations of an environment. Hilbert maps [6], on the other hand, are a continuous occupancy map that takes the form of a linear function over some hinge point feature space. Hilbert map representations have also been extended to Bayesian Hilbert maps of various forms [7]–[9]. Neural implicit functions have also been used to represent continuous 3D geometry [10]–[12]. Other representations are built using differentiable rendering and combining multiple views, especially for novel-view synthesis. A Neural radiance field (NeRF) [13] is an example of this, in which a neural network maps 3D position to density and color. 3D Gaussian splatting [14] tackles a similar but with a set of Gaussians instead of a neural network. Foundation models have also been developed for this task in [15] and applied to robotics [16]. Other representation primitives have been studied, including super quadrics [17].

3D Reconstruction with Deep Learning. Many methods have been proposed as ways to leverage deep learning to reconstruct scenes or objects. While some methods aim to predict object shape from RGB data only [18]–[21], we instead focus on using depth measurements during reconstruction. DeepSDF [10] is a method to reconstruct an object by running inference-time optimization to recover a latent code for a neural implicit function. In the context of robotics, [22] extends DeepSDF to have uncertainty-awareness. Other work, such as occupancy networks [11] or PointSDF [23] try to directly predict such a latent code without inference-time optimization. Deep learning has also been leveraged to learn kernels, which are used to construct a continuous signed distance function [24], [25]. Language is also used during reconstruction in [26] and [27]. Another work uses a voxel-to-voxel variational autoencoder conditioned on bounding boxes [28]. While these works typically focus on single object scenes, other work focuses on reconstruction scenes with multiple occluding objects. For example, [29] learns a reconstruction from a voxel representation with different channels to account for occlusion. Another method uses silhouettes to refine the initial reconstructions [30]. In practice, these deep learning approaches can struggle to reconstruct noisy scenes with multiple, highly occluded, unknown objects on real world depth cameras. Inaccurate segmentation can also be a problem for many of these methods as well.

3D Reconstruction without Deep Learning. There are also many approaches to perform probabilistic 3D reconstruction without deep learning. Some of such methods for reconstruction use informative prior information by assuming fixed classes of objects, such as 3DP3 [31]. Other methods use an uninformative prior, such as Gaussian process implicit surfaces (GPIS) [3]. While there is an extension of GPIS to a slightly more informative prior [32], the only priors that can be enforced are specifically spherical, ellipsoidal, cylidrical, or planar priors. In our work, we derive our prior from pre-existing mesh datasets. V-PRISM [4] is another method that probabilistically *maps* the scene using a multi-class framing.

Using Reconstructions in Manipulation. 3D reconstruction methods have seen extensive use in manipulation. In [23], PointSDF provides collision constraints during grasping. PointSDF is also utilized in [33], where tactile sensors are used along with the reconstruction during grasping. A learning-based voxel representation is used for grasping in [34]. Neural shape completion is also used during the anthropomorphic grasping pipeline proposed in [35]. GPIS is also a common representation for manipulation applications. Some recent work has utilized the uncertainty from GPIS representations during grasp selection [1], [2]. We believe BRRP provides principled uncertainty measurements that can similarly be utilized in downstream manipulation tasks.

III. Background

A. Hilbert Maps

A Hilbert map [6] is a *continuous* occupancy map. It represents the environment by a continuous function that is defined by a linear function of a fixed feature transform.

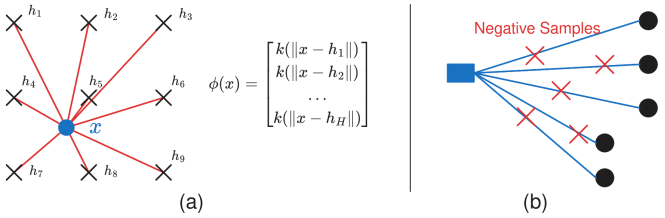

Figure 2: (a) A *hinge point* feature transform induced by a set of hinge points is used by Hilbert maps [6]; (b) these maps are built by first sampling *negative samples* along the unoccupied portions of the camera ray.

Typically this feature transform is induced by a set of *hinge points*, $\{\mathbf{h}_1, ..., \mathbf{h}_H\} \subset \mathbb{R}^3$ and a translation-invariant kernel $k(d)$. The transform is then defined as:

$$\phi(\mathbf{x}) = [1, k(\mathbf{x} - \mathbf{h}_1), ..., k(\mathbf{x} - \mathbf{h}_H)]^\top.$$

Typically, a Gaussian kernel is used and hinge points are placed in an evenly-spaced grid. An occupancy map can then be defined by a single weight vector, $\mathbf{w} \in \mathbb{R}^{H+1}$, as such:

$$m(\mathbf{x}) = \sigma(\mathbf{w}^\top \phi(\mathbf{x})).$$

To recover the weight vector corresponding to a given depth observation, *negative sampling* is performed along the unoccupied portions of the depth rays. These negative samples are assigned a label of unoccupied and the points at the end of the ray are labeled as occupied. Then, stochastic gradient descent (SGD) is performed on the binary cross entropy (BCE) of the negative samples and terminal points of the ray. The binary cross entropy measures the *likelihood* of the samples, and is defined as:

$$\text{BCE}(y, \mathbf{w}^\top \phi(\mathbf{x})) = \begin{cases} -\ln\left[\sigma\left(\mathbf{w}^\top \phi(\mathbf{x})\right)\right], & y = 1 \\ -\ln\left[1 - \sigma\left(\mathbf{w}^\top \phi(\mathbf{x})\right)\right], & y = 0 \end{cases} \tag{1}$$

Figure 2 shows an illustration of both the hinge point feature transform and the negative samples used during Hilbert map construction.

Hilbert maps have previously been extended to Bayesian Hilbert maps, where a distribution over weights is modeled as a multivariate Gaussian [7]–[9]. There is also a multiclass variant that defines a weight *matrix* with Gaussian rows described in [4]. In this work, we adopt the Hilbert map representation, but model the distribution over weights as a collection of particles. This allows our method the capability to capture irregular, non-Gaussian posterior distributions.

B. Stein Variational Gradient Descent

Stein Variational Gradient Descent (SVGD) [36] is an algorithm for variational inference that closely resembles gradient descent. The general problem of variational inference is to find a distribution $q^* \in \mathcal{Q}$ that is close to some target distribution p. Usually, this takes the form of an optimization problem over the Kullback-Leibler (KL) divergence:

$$q^* = \arg\min_{q \in \mathcal{Q}} \mathbb{KL}(q \| p).$$

SVGD aims to iteratively transform q in descent directions of the KL divergence in a d-dimensional reproducing kernel Hilbert space (RKHS), $\mathcal{H}^d$. Because this Hilbert space is a space of functions, a descent direction requires deriving the *functional gradient* of our KL divergence objective.

***Theorem** 1:* From [36]. Let $T(\mathbf{x}) = \mathbf{x} + f(\mathbf{x})$, where $f \in \mathcal{H}^d$ and $q_{[T]}$ is the density of random variable $\mathbf{z} = T(\mathbf{x})$ when $\mathbf{x} \sim q$. Then

$$\nabla_f \mathbb{KL}(q_{[T]} \| p)|_{f=0} = -g^*_{q,p},$$

where $g^*_{q,p} = \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})}[k(\mathbf{x}, \cdot)\nabla_{\mathbf{x}} \ln p(\mathbf{x}) + \nabla_{\mathbf{x}} k(\mathbf{x}, \cdot)]$.

In SVGD, q is approximated by a set of particles $\mathbf{x}_1^{(0)}, ..., \mathbf{x}_P^{(0)} \sim q(\mathbf{x})$. This can be used to approximate the gradient in Theorem 1 with $\hat{g}^*$:

$$\hat{g}^*(\mathbf{x}) = \frac{1}{P} \sum_{i=1}^{P} k(\mathbf{x}_i, \mathbf{x}) \nabla_{\mathbf{x}_i} \ln p(\mathbf{x}_i) + \nabla_{\mathbf{x}_i} k(\mathbf{x}, \cdot). \tag{2}$$

The particles can then be iteratively updated according to $\hat{g}^*$ in Equation (2) with:

$$\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} + \epsilon \hat{g}^*(\mathbf{x}_i^{(t)})$$

The result of these iterations is that the set of particles converges to an approximation of the target distribution p. Importantly, Equation (2) only relies on the gradient of the log of p, which means we can perform variational inference to an unnormalized distribution. Such unnormalized distributions are commonplace in many Bayesian inference problems, including ours.

IV. Retrieval-Augmented Priors

Retrieval-augmented generation [37] was originally introduced in the context of improving language generation. The work has served as inspiration for an approach to affordance-prediction in [38]. In our case, we draw inspiration from retrieval-augmented generation, but we use the retrieved results to improve efficiency in certain *explicit* formulations for prior density functions during Bayesian inference.

To motivate retrieval-augmented priors, consider the problem of Bayesian inference with a mixture model acting as the prior distribution. Given some data, we would like to infer a posterior distribution over hypotheses. If we have a mixture model as a prior distribution, then:

$$P(H|D) \propto P(D|H) \sum_{c=1}^{C} P(H|c). \tag{3}$$

If our prior distribution has a lot of components, it may be inefficient to fully evaluate. This could be a serious problem for algorithms like SVGD, which requires iteratively computing the gradient of both the likelihood and prior. Inspired by [37], the insight behind retrieval-augmented priors is to determine which subset of the prior distribution components to retrieve and use given some detection result R. Conditioning on this detection result, we have a new posterior distribution, $P(H|D, R)$. Making an independence assumption,

$$P(H|D, R) \propto P(D|H) \cdot \mathbb{E}_{c \sim P(c|R)}[P(H|c)].$$

Comparing to Equation (3), the expectation now replaces the true prior. Then, we can use a top-k approximation for the expectation:

$$P(H|D, R) \propto P(D|H) \sum_{c \in \text{topk}} P(H|c) P(c|R) \tag{4}$$

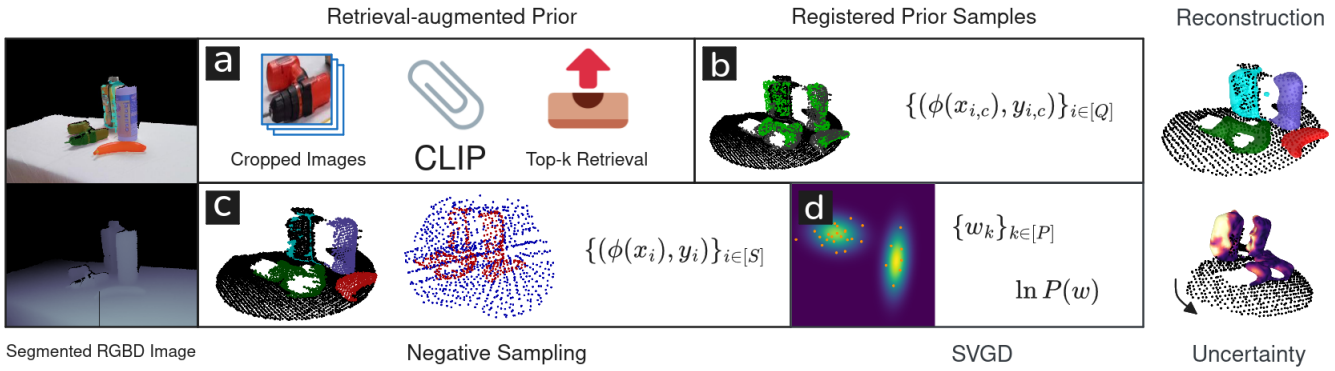

Figure 3: Overview of BRRP method. We begin with a segmented RGBD image and (a) feed cropped images of each segment into CLIP to get object probabilities (Section V-B) Then, we retrieve and (b) register the the top-k objects in the prior. This gives us a set of registered prior samples (Section V-A). We also (c) compute negative samples based on the observed segmented point cloud (Section V-C). Finally, (d) we run SVGD optimization to recover a posterior distribution over Hilbert map weights (Section V-D). We can use this distribution to both reconstruct the scene as well as measure uncertainty.

This means that we only need to evaluate a subset of the prior distribution components.

V. The BRRP Method

Our method takes a single RGBD image and produces reconstructions for each object in the scene. We treat the problem as a Bayesian inference problem over an observation described by negative samples. We incorporate prior information on the shape of the object by leveraging retrieval-augmented priors introduced in Section IV. We use CLIP [39] to determine which objects to retrieve and define our object-specific priors by a registered set of pre-computed samples from the stored mesh. We then use SVGD to optimize for a set of samples over map weights. We can generate predicted reconstructions by taking the expected occupancy over our weights for a given location. Figure 3 shows a visual overview of our method.

In Section V-A, we explain how we leverage pre-existing mesh assets to create a prior that is robust to different poses and scales. Then, in Section V-B, we explain how we utilize the retrieval-augmented priors paradigm to retrieve relevant objects in the prior. The specific negative sampling is explained in Section V-C. Then we give the specific SVGD objective used in Section V-D.

A. Negative Samples as Reconstruction Priors

We want to leverage existing mesh assets as our priors during Bayesian reconstruction. We define our prior as a mixture over different objects, $c_1, ..., c_C$. Because there is not a direct way to convert a mesh into a Hilbert map, we instead *sample* points $\tilde{\mathbf{x}}_{c,1}, ..., \tilde{\mathbf{x}}_{c,Q} \in \mathbb{R}^3$ around each object c's mesh. We refer to these samples as the *prior samples*. We give them labels $\tilde{y}_{c,1}, ..., \tilde{y}_{c,1} \in \{-1, 1\}$ determined by whether they are outside or inside the mesh. Then we simply define our prior using this data combined with a Gaussian prior over weight norm:

$$P(\mathbf{w}|c) := P(\{\tilde{y}_{c,i}\}|\{\tilde{\mathbf{x}}_{c,i}\}, \mathbf{w})P(\mathbf{w}) \tag{5}$$

$$\propto \exp(\lambda\|w\|^2) \prod_{i=1}^{Q} \exp\left(\text{BCE}(\tilde{y}_{c,i}, \mathbf{w}^\top \phi(\tilde{\mathbf{x}}_{c,i}))\right), \tag{6}$$

where BCE is the same as in Equation (1).

In order to enforce pose-invariance, we first register a small stored point cloud of the object to the observed points and then transform the prior samples to this reference frame. In practice we use RANSAC [40] and the FPFH features from [41] to perform registration. In order to also have scale invariance, we do a linear scan over 10 different scales and select the the scale that resulted in the most inlier pairs from the registration.

B. Retrieval-Augmented Priors for Hilbert Maps

Because it would be inefficient to register all meshes that are part of the prior mixture model, we propose using the retrieval-augmented prior approach introduced in Section IV. In order to determine which objects to use, we need to compute $P(c|R)$ from Equation (4). In our case, we use CLIP [39] as a zero-shot classifier for our different objects. For each object in our prior, we store a small textual description of the object. These descriptions are then used as classes for CLIP to classify each segmented object. In order to make sure CLIP knows which object we are targeting, we crop the RGB image to fit the predicted segmentation of each object and feed the cropped images as input into CLIP.

Once we have the probability of each object, we retrieve and register the stored point clouds of the top-k objects. After registration, we retrieve the prior samples corresponding to these objects to define our prior according to Equation (5).

C. Negative Sampling

We adopt the negative sampling method introduced in [4]. The negative sampling method makes the assumption that all objects are lying on or above a planar surface. We

begin by labeling the points segmented to each object as occupied for that object. Next, we perform stratified sampling along each camera ray near each object to recover a set of negatively sampled points, labeled as unoccupied. Then, we use RANSAC over points not segmented to any object to recover the flat surface all objects are resting on. This plane is used to randomly sample points in a sphere underneath each object that are near the object. We also label these points as occupied. Finally, we use grid subsampling from [42] to reduce the number of points and increase uniformity of sampled points. We refer to these points and labels as *observed samples* and denote them as $\{\mathbf{x}_i\}_{i\in[S]}, \{y_i\}_{i\in[S]}$. The entire negative sampling process can be easily parallelized for efficient computation.

D. SVGD Reconstruction

Once we have retrieved our prior samples and computed our observed samples, we can perform optimization-based reconstruction with SVGD. Given both sets of samples and our prior definition from Equation (5), we have the following posterior distribution:

$$P(\{y_{c,i}\}|\{\mathbf{x}_{c,i}\}, \mathbf{w})P(\mathbf{w}) \sum_{c\in\text{topk}} P(c|R)P(\{\tilde{y}_{c,i}\}|\{\tilde{\mathbf{x}}_{c,i}\}, \mathbf{w}),$$

taking the log and applying Equation (6) gives us the following objective:

$$\sum_{i=1}^{S} \text{BCE}(y_i, \mathbf{w}^\top \phi(\mathbf{x}_i)) + \sum_{c\in\text{topk}} P(c|R) \ln \left[\sum_{i=1}^{Q} \exp\left(\text{BCE}(\tilde{y}_{c,i}, \mathbf{w}^\top \phi(\tilde{\mathbf{x}}_{c,i}))\right)\right] + \lambda\|\mathbf{w}\|^2 + \text{const.}$$

In practice, we introduce multipliers to each term as hyperparameters during optimization, drop the constant, and use means instead of sums, creating the following objective:

$$\frac{\lambda_3}{S} \sum_{i=1}^{S} \text{BCE}(y_i, \mathbf{w}^\top \phi(\mathbf{x}_i)) \tag{7}$$

$$+ \frac{\lambda_2}{K} \sum_{c\in\text{topk}} P(c|R) \ln \left[\frac{1}{Q} \sum_{i=1}^{Q} \exp\left(\text{BCE}(\tilde{y}_{c,i}, \mathbf{w}^\top \phi(\tilde{\mathbf{x}}_{c,i}))\right)\right] \tag{8}$$

$$+ \lambda_3\|\mathbf{w}\|^2, \tag{9}$$

where K is the number of objects retrieved for the prior. This objective is used as the log of the target distribution, $\ln P(\mathbf{w})$, in Equation (2), where we also adopt the original median kernel suggested in [36]. We also opt to use SVGD in a stochastic manner, where both the observed samples and query samples are mini-batched.

From a non-probabilistic standpoint, one can interpret Equation (7) as the likelihood of the observed data, Equation (8) as the object shape prior, and Equation (9) as a regularization term.

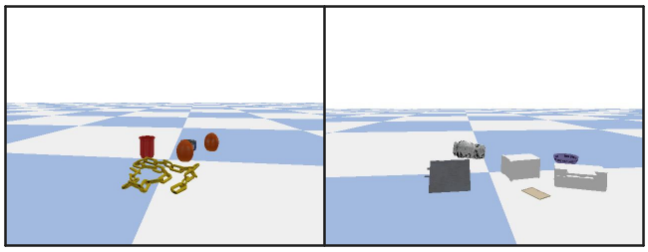

Figure 4: Sample images of procedurally generated scenes used to evaluate BRRP. **Left:** a YCB scene. **Right:** a ShapeNet scene.

Method	ShapeNet Scenes	YCB Scenes	Objaverse Scenes
V-PRISM [4]	0.3092	0.5003	0.4640
PointSDF [23]	**0.3600**	0.4601	0.3471
BRRP (ours)	0.3124	**0.5277**	**0.4809**

Table I: Intersection over union (IoU) on procedurally generated scenes from three different mesh datasets. BRRP uses a YCB prior and PointSDF is trained on ShapeNet scenes.

VI. Experiments

In this section, we aim to experimentally validate the following claims: (1) BRRP is more *robust* than deep learning methods; (2) BRRP is more *accurate* than methods that use uninformative priors; (3) BRRP can capture principled uncertainty. We begin by providing details on the experiments such as the baselines and metrics in Section VI-A, then we introduce results and analyses in Section VI-B

A. Experimental Details

BRRP Implementation: We use a set of 50 objects from the YCB dataset [43] to act as the prior for our experiments with BRRP. We implement the method in PyTorch and run the method on an NVIDIA RTX GeForce 2070 GPU.

Baselines: We compare our work against two main baselines, V-PRISM [4] and a version of PointSDF [23] that predicts occupancy and is trained on ShapeNet [44] scenes. V-PRISM is a probabilistic mapping method that uses an uninformative prior. This means that it is robust to novel objects, but doesn't accurately reconstruct object backsides. We refer to this baseline as **V-PRISM**. In contrast, PointSDF is a learning-based method. This means it can leverage prior information from mesh datasets to accurately reconstruct the backside, but can suffer in performance under significant distributional shift. We refer to this baseline as **PointSDF**. When computing reconstruction meshes with PointSDF, a level set of $\tau = 0.3$ is used.

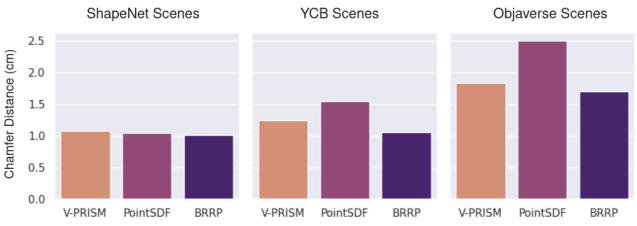

Figure 5: Chamfer distances (lower is better) for various methods across the procedurally generated scenes. Values are reported in centimeters. BRRP has the lowest chamfer distance on each dataset.

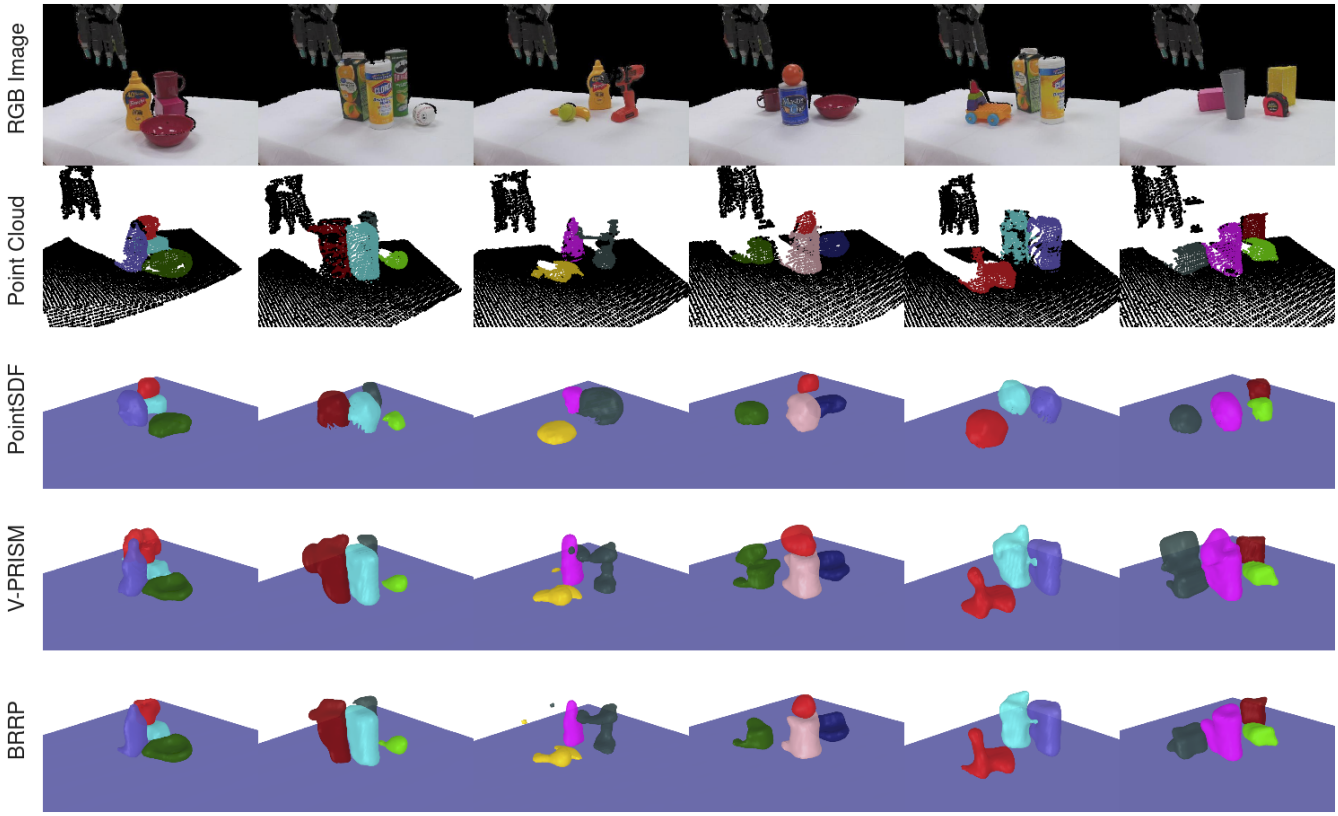

Figure 6: Qualitative comparison of BRRP and our baselines. PointSDF tends to predict a spherical shape for many non-spherical objects. V-PRISM can sometimes predict occupancy in portions of the scene that are not occupied. Our method is more robust and can more accurately reconstruct the scenes.

Procedurally Generated Scenes: We use the generated scenes from [4] to evaluate our method. These scenes are constructed with objects from ShapeNet [44], YCB [43], and Objaverse [45] datasets. There are 100 multi-object scenes for each mesh dataset. Each scene contains up to 10 objects. Some meshes in the Objaverse and ShapeNet scenes did not have correctly rendered textures and were instead rendered as plain white objects. Figure 4 contains two example images of these procedurally generated scenes. We also conduct an experiment on robustness where we perturb instance segmentation of the ShapeNet scenes by 2 pixels and evaluate reconstructions.

We evaluate performance on the procedurally generated scenes with two metrics: intersection over union (IoU) and chamfer distance. We refer readers to [4] for further explanation of these metrics.

Real World Scenes: In order to showcase robustness to real-world noise, we evaluate on real world scenes collected with a Kinect depth camera. In order to obtain instance segmentations, we use Grounded SAM [46] along with some depth filters. We evaluate on these real world scenes qualitatively with images of scene reconstruction and visualizing surface uncertainty.

B. Results

In Table I, we display the IoU results from procedurally generated scenes. The chamfer distances for the procedurally generated scenes is shown in Figure 5. The qualitative reconstructions on real world scenes can be seen in Figure 6.

Insight 1: *BRRP is more accurate than a method with an uninformative prior.*

As showcased in Table I, BRRP outperforms V-PRISM on each set of procedurally generated scenes. It has the highest IoU improvement from V-PRISM on the YCB scenes. A similar pattern can be seen in the chamfer distances in Figure 5, where BRRP consistently outperforms V-PRISM, with the biggest gap of 0.19 (cm) occuring on the YCB scenes. This makes sense because BRRP uses a subset of the YCB objects as its prior.

We can see this improvement qualitatively in Figure 6. While BRRP and V-PRISM are comparable for most objects, there exist certain objects that V-PRISM predicts to occupy a large portion of space that the object *doesn't* occupy. BRRP is able to more accurately reconstruct these objects. The clearest example of this is the dark green object in the right-most scene in Figure 6.

Both of these quantitative and qualitative results suggest BRRP is generally more accurate than V-PRISM. It is the most accurate when evaluated on objects in its prior distribution.

Insight 2: *BRRP is more robust than a deep learning method.*

While PointSDF outperformed BRRP on the ShapeNet scenes on IoU (Table I), BRRP had a lower Chamfer distance

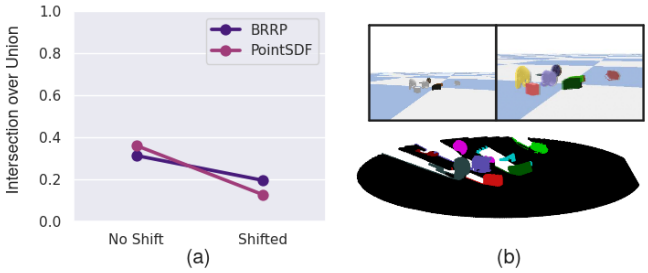

Figure 7: (a) IoU of BRRP and PointSDF on ShapeNet scenes with and without shifted segmentations. Our method is more robust to segmentation shifts. (b) An example of a scene and the corresponding point cloud with shifted segmentation.

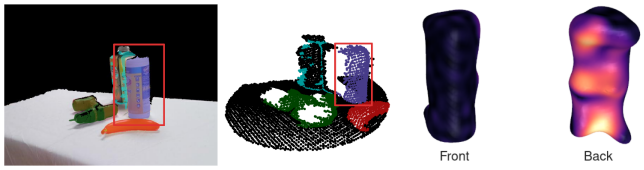

Figure 8: Visualization of a cylindrical Clorox container surface uncertainty from BRRP. Lighter areas correspond to higher uncertainty about the shape. We observe that the occluded back-side of the container has high uncertainty.

(Figure 5). BRRP also performed better on mesh datasets that PointSDF was not trained on. In Table I, we can see that on Objaverse scenes, where the objects were novel to both methods, BRRP performed better than PointSDF. When measuring chamfer distance, BRRP outperformed PointSDF on all datasets as shown in Figure 5. These results suggests BRRP is more robust to different object distributions than PointSDF.

Next, we evaluate robustness to slightly incorrect instance segmentations. We take the procedurally generated ShapeNet scenes and shift the segmentation over by 2 pixels. In Figure 7, we compare the IoU of BRRP and PointSDF on the scenes with and without the shift. Our method performs better on the shifted scenes compared to PointSDF.

On the real world scenes in Figure 6, BRRP is qualitatively more robust than PointSDF. PointSDF struggles with the noise associated with real world scenes as well as the novel objects. It tends to predict a spherical object on many non-spherical objects in the real world scenes. Our method on the other hand, is better able to reconstruct these scenes, including the objects that are out-of-distribution for its prior.

These results suggest BRRP is more robust than PointSDF. Even though by one metric PointSDF outperforms BRRP on ShapeNet scenes, when the scenes are perturbed BRRP performs better.

Insight 3: *BRRP can capture principled uncertainty about object shape.*

Figure 8 shows a qualitative example of uncertainty from BRRP. We measure the uncertainty by taking the variance of logits over weight particles. Our method predicts the highest uncertainty in areas of the surface that are occluded. This suggests that we can utilize surface uncertainty from BRRP in a similar way to how GPIS surface uncertainty is utilized in many grasping applications.

VII. Conclusion

We introduced the concept of retrieval-augmented priors, where we retrieve relevant components of a prior distribution during Bayesian inference. We also introduced a novel Bayesian method for scene reconstruction that uses *informative priors*. Our method, BRRP, leveraged existing mesh datasets to build its prior and probabilistically reconstructed the scene leveraging these priors in a retrieval-augmented manner. We experimentally showed our method was more robust than a deep learning method as well as more accurate than a mapping method. Finally, we showed a qualitative example of recovering principled uncertainty from our method.

References

[1] S. Chen, J. Bohg, and C. K. Liu, "Springgrasp: An optimization pipeline for robust and compliant dexterous pre-grasp synthesis," *arXiv preprint arXiv:2404.13532*, 2024.

[2] C. de Farias, B. Tamadazte, M. Adjigble, R. Stolkin, and N. Marturi, "Task-informed grasping of partially observed objects," *IEEE Robotics and Automation Letters*, 2024.

[3] S. Dragiev, M. Toussaint, and M. Gienger, "Gaussian process implicit surfaces for shape estimation and grasping," in *2011 IEEE International Conference on Robotics and Automation*, pp. 2845–2850, IEEE, 2011.

[4] H. Wright, W. Zhi, M. Johnson-Roberson, and T. Hermans, "V-prism: Probabilistic mapping of unknown tabletop scenes," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1078–1085, 2024.

[5] H. Oleynikova, A. Millane, Z. Taylor, E. Galceran, J. Nieto, and R. Siegwart, "Signed distance fields: A natural representation for both mapping and planning," in *RSS 2016 workshop: geometry and beyond-representations, physics, and scene understanding for robotics*, University of Michigan, 2016.

[6] F. Ramos and L. Ott, "Hilbert maps: Scalable continuous occupancy mapping with stochastic gradient descent," *The International Journal of Robotics Research*, vol. 35, no. 14, pp. 1717–1730, 2016.

[7] R. Senanayake and F. Ramos, "Bayesian hilbert maps for dynamic continuous occupancy mapping," in *Conference on Robot Learning*, pp. 458–471, PMLR, 2017.

[8] R. Senanayake, A. Tompkins, and F. Ramos, "Automorphing kernels for nonstationarity in mapping unstructured environments.," in *CoRL*, pp. 443–455, 2018.

[9] W. Zhi, L. Ott, R. Senanayake, and F. Ramos, "Continuous occupancy map fusion with fast bayesian hilbert maps," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 4111–4117, IEEE, 2019.

[10] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.

[11] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. 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.

[12] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in neural information processing systems*, vol. 33, pp. 7462–7473, 2020.

[13] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[14] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, 2023.

[15] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[16] W. Zhi, H. Tang, T. Zhang, and M. Johnson-Roberson, "Simultaneous geometry and pose estimation of held objects via 3d foundation models," *IEEE Robotics and Automation Letters*, 2024.

[17] W. Liu, Y. Wu, S. Ruan, and G. S. Chirikjian, "Robust and accurate superquadric recovery: A probabilistic approach," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2676–2685, 2022.

[18] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan, "Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models," *arXiv preprint arXiv:2404.07191*, 2024.

[19] H. Jun and A. Nichol, "Shap-e: Generating conditional 3d implicit functions," *arXiv preprint arXiv:2305.02463*, 2023.

[20] M. Liu, C. Xu, H. Jin, L. Chen, M. Varma T, Z. Xu, and H. Su, "One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[21] F. Engelmann, K. Rematas, B. Leibe, and V. Ferrari, "From points to multi-object 3d reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4588–4597, 2021.

[22] Z. Liao, J. Yang, J. Qian, A. P. Schoellig, and S. L. Waslander, "Uncertainty-aware 3d object-level mapping with deep shape priors," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4082–4089, IEEE, 2024.

[23] M. Van der Merwe, Q. Lu, B. Sundaralingam, M. Matak, and T. Hermans, "Learning continuous 3d reconstructions for geometrically aware grasping," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11516–11522, IEEE, 2020.

[24] F. Williams, Z. Gojcic, S. Khamis, D. Zorin, J. Bruna, S. Fidler, and O. Litany, "Neural fields as learnable kernels for 3d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18500–18510, 2022.

[25] J. Huang, Z. Gojcic, M. Atzmon, O. Litany, S. Fidler, and F. Williams, "Neural kernel surface reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4369–4379, 2023.

[26] Y. Kasten, O. Rahamim, and G. Chechik, "Point cloud completion with pretrained text-to-image diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[27] Y.-C. Cheng, H.-Y. Lee, S. Tulyakov, A. G. Schwing, and L.-Y. Gui, "Sdfusion: Multimodal 3d shape completion, reconstruction, and generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4456–4465, 2023.

[28] B. Saund and D. Berenson, "Diverse plausible shape completions from ambiguous depth images," in *Conference on Robot Learning*, pp. 1802–1813, PMLR, 2021.

[29] W. Agnew, C. Xie, A. Walsman, O. Murad, Y. Wang, P. Domingos, and S. Srinivasa, "Amodal 3d reconstruction for robotic manipulation via stability and connectivity," in *Conference on Robot Learning*, pp. 1498–1508, PMLR, 2021.

[30] L. Li, S. Khan, and N. Barnes, "Silhouette-assisted 3d object instance reconstruction from a cluttered scene," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[31] N. Gothoskar, M. Cusumano-Towner, B. Zinberg, M. Ghavamizadeh, F. Pollok, A. Garrett, J. Tenenbaum, D. Gutfreund, and V. Mansinghka, "3dp3: 3d scene perception via probabilistic programming," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9600–9612, 2021.

[32] W. Martens, Y. Poffet, P. R. Soria, R. Fitch, and S. Sukkarieh, "Geometric priors for gaussian process implicit surfaces," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 373–380, 2016.

[33] M. Matak and T. Hermans, "Planning visual-tactile precision grasps via complementary use of vision and touch," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 768–775, 2022.

[34] J. Lundell, F. Verdoja, and V. Kyrki, "Robust grasp planning over uncertain shape completions," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1526–1532, IEEE, 2019.

[35] D. Hidalgo-Carvajal, H. Chen, G. C. Bettelani, J. Jung, M. Zavaglia, L. Busse, A. Naceri, S. Leutenegger, and S. Haddadin, "Anthropomorphic grasping with neural object shape completion," *IEEE Robotics and Automation Letters*, 2023.

[36] Q. Liu and D. Wang, "Stein variational gradient descent: A general purpose bayesian inference algorithm," *Advances in neural information processing systems*, vol. 29, 2016.

[37] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[38] Y. Kuang, J. Ye, H. Geng, J. Mao, C. Deng, L. Guibas, H. Wang, and Y. Wang, "Ram: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation," *arXiv preprint arXiv:2407.04689*, 2024.

[39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[40] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[41] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *2009 IEEE international conference on robotics and automation*, pp. 3212–3217, IEEE, 2009.

[42] H. Thomas, *Learning new representations for 3D point cloud semantic segmentation*. PhD thesis, Université Paris sciences et lettres, 2019.

[43] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *2015 international conference on advanced robotics (ICAR)*, pp. 510–517, IEEE, 2015.

[44] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An Information-Rich 3D Model Repository," Tech. Rep. arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

[45] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3d objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023.

[46] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.