

NodeSLAM: Neural Object Descriptors for Multi-View Shape Reconstruction

Edgar Sucar Kentaro Wada Andrew Davison
 Dyson Robotics Laboratory, Imperial College London
 {e.sucar18, k.wada18, a.davison}@imperial.ac.uk

Abstract

The choice of scene representation is crucial in both the shape inference algorithms it requires and the smart applications it enables. We present efficient and optimisable multi-class learned object descriptors together with a novel probabilistic and differential rendering engine, for principled full object shape inference from one or more RGB-D images. Our framework allows for accurate and robust 3D object reconstruction which enables multiple applications including robot grasping and placing, augmented reality, and the first object-level SLAM system capable of optimising object poses and shapes jointly with camera trajectory.

1. Introduction

To enable advanced AI applications, computer vision algorithms must build useful persistent 3D representations of the scenes they observe from one or more views, especially of the objects available for interaction. Ideal object representations are: (i) efficient, for principled and fast optimisation, (ii) robust to noisy measurements and variable uncertainties, and (iii) incremental, with the ability to grow and improve with new measurements. In this paper, we study the use of generative class-level object models and argue that they provide the right representation for principled and practical shape inference as shown in the experiments. We use a novel probabilistic rendering measurement function which enables efficient and robust optimisation of object shape with respect to depth images, and build from this an object-level SLAM system capable of incrementally mapping multi-object scenes.

There have been two main approaches for 3D shape reconstruction from images. Classical reconstruction techniques infer geometry by minimizing the discrepancy between a reconstructed 3D model and observed data through a measurement function [10, 19, 36]. These methods are flexible and general, but they can only reconstruct directly observed parts of a scene and are limited in accuracy when observations are weak or noisy. On the other hand, dis-

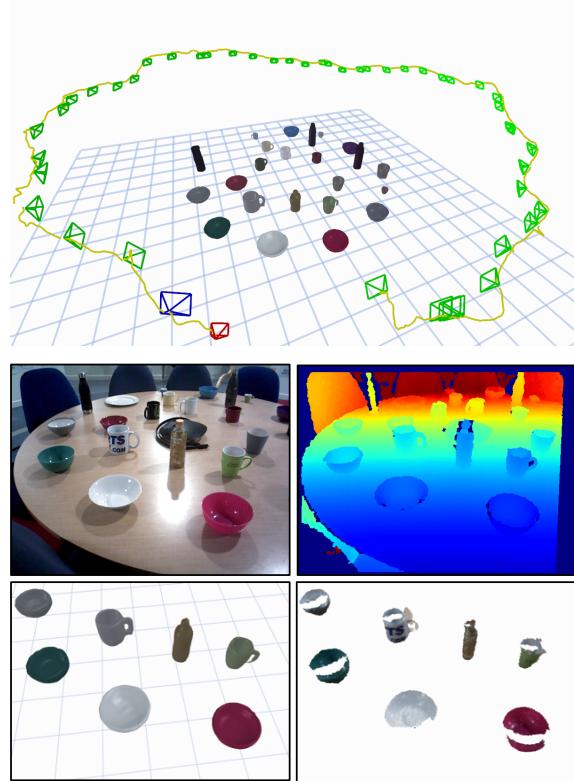


Figure 1. **Top:** Compact, optimisable shape models used in an object-level SLAM system which maps a real world cluttered table top scene with varied object shapes from different classes. **Bottom:** Class-level priors allow accurate and complete object reconstruction (bottom-left) even from a single image in contrast to partial reconstruction from TSDF fusion (bottom-right).

criminative methods learn to map image measurements to 3D shape, such as through a feed-forward neural network [7, 15, 37, 33, 34, 39]. These methods take advantage of regularities in data for robustness but have trouble in generalisation and lack the ability to integrate multiple measurements in a principled way.

Our work sits between these two approaches. We capture regularities in data though a volumetric 3D generative model represented though a class conditioned Variational Auto Encoder (VAE), trained on a collection of CAD mod-

els, allowing us to represent object shape through a compact code. We then use the generative model for shape inference though iterative optimisation of the latent code with respect to any number of depth image measurements.

To use a generative method for inference we need a rendering function to transform 3D volumes into measurements; in our case depth images with object segmentation. The design of this function will influence optimisation speed and convergence success. Two important design considerations are (1) receptive field, the size of 3D region which influences each rendered pixel, and (2) uncertainty modeling, the confidence of each rendered pixel depth. We introduce a novel probabilistic volumetric rendering function based on these two design principles, improving the state of the art in volumetric rendering.

In scenes with many objects, our optimisable compact object models can serve as the landmarks in a SLAM system, where we use the same measurement function for camera tracking, object poses and shape optimisation. We quantitatively show that joint optimisation leads to more robust tracking and reconstruction, with comparable surface reconstruction to the data driven Fusion++ [19], while reaching full object reconstruction from far fewer observations.

An emphasis of this paper is to design object models that work robustly in the real world. We demonstrate the robustness of our proposed rendering function through qualitative demonstrations of our object-level SLAM on real world image sequences from a cluttered table-top scene obtained with a noisy depth camera, and on an augmented reality demo. Furthermore we integrate our efficient shape inference method into a real time robotic system, and show that the completeness and accuracy of our object reconstructions enable robotic tasks such as packing objects into a tight box or sorting objects by shape size. We encourage readers to watch the associated **video** which supports our submission.

To summarise, the key contributions of our paper are: (i) A novel volumetric probabilistic rendering function which enables robust and efficient multi-view shape optimisation. (ii) The first object-level SLAM capable of jointly optimising full object shapes and poses together with camera trajectory from real world images. (iii) The integration into a real-time robotic system that can achieve useful manipulation tasks with varied object shapes from different categories due to complete high quality surface reconstructions.

2. Related Work

There are two categories of methods for full object shape inference from images, discriminative and generative approaches. Discriminative single image reconstruction approaches such as [7, 15, 37, 33, 34, 39] lack the ability to integrate multiple observations in a principled way. For example, DeepSLAM++ [9] directly averages 3D shapes predicted by Pix3D [30], while 3D-R2N2 [2] uses a recurrent

network to update shapes from new images.

Generative methods for shape reconstruction were first developed using linear models such as PCA [3, 6, 35, 16, 42] which constrained them to simple shapes. Our work comes at a time when many authors are building accurate and flexible generative models for objects using neural networks [38, 24, 21, 25] and starting to use them in shape inference from images [17, 11]. However existing methods for generative full shape inference have not yet shown to work in incremental real-world multi-object settings in real time.

To allow for robust inference, critical in real world settings, and efficient optimisation, necessary for real time application, we introduce a novel differential probabilistic rendering formulation. Existing differential volumetric rendering techniques used for shape inference such as [18, 23, 11] have a local receptive field, which means each rendered pixel only depends on a single 3D point sample; this causes optimisation to be slow and easily stuck in local minima because of local gradient flow. We take inspiration from traditional volumetric rendering in graphics [12] to define a probabilistic rendering formulation, which gathers occupancy probabilities along samples of a back-projected ray for each pixel to render a depth image. Additionally, we build a Gaussian pyramid after rendering to increase the spatial receptive field. Our probabilistic formulation allows us to measure the uncertainty of each rendered pixel which improves optimisation by weighting residuals.

One of the targets of visual SLAM research has been a continuous improvement in the richness of scene representations, from sparse SLAM systems which reconstruct point clouds, [4, 5] via dense surface representations [22, 36] to semantically labelled dense maps [20, 40]. Dense maps of whole scenes are very expensive to store, and difficult to optimise probabilistically from multiple views. A sensible route towards representations which are both efficient and semantically meaningful is to focus representation resources on the most important elements of a scene, objects.

Dense object-based SLAM has been previously attempted, but our work fills a gap between systems which make separate reconstructions of every identified object [19, 31], which are general with respect to arbitrary classes but can only reconstruct directly observed surfaces, and those which populate maps with instances of object CAD models [27], which are efficient but limited to precisely known objects. Our approach can efficiently reconstruct whole objects which are recognised at the class level, and cope with a good range of shape variation within each class.

3. Class-Level Object Shape Descriptors

Objects of the same semantic class exhibit strong regularities in shape under a common pose alignment. We leverage this to construct the class specific smooth latent space, which allows us to represent the shape of an instance with

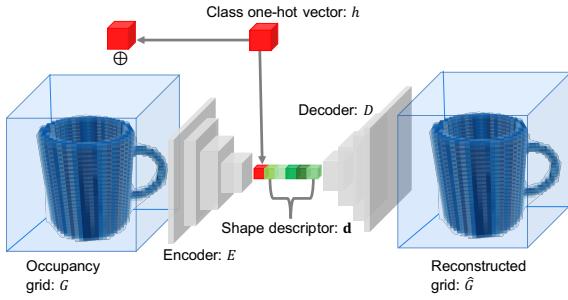


Figure 2. **Occupancy Variational Autoencoder:** The class one hot vector h is concatenated channel-wise to each occupancy voxel in the input occupancy grid G . The input is compressed into shape descriptor d by encoder network E . The shape descriptor and the class-one hot vector are concatenated and passed through decoder network D to obtain occupancy reconstruction \hat{G} .

a small number of parameters by training a single Class-Conditional Variational Autoencoder neural network.

3.1. Network Design

3D object shapes are represented by voxel occupancy grids of dimension $32 \times 32 \times 32$, with each voxel storing a continuous occupancy probability between 0 and 1. A voxel grid was chosen to enable shapes of arbitrary topology.

The 3D models were obtained from the ShapeNet database [1], which comes with annotated model alignment. The occupancy grids were obtained by converting the model meshes into a high resolution binary occupancy grid, and then down-sampling by average pooling.

A single 3D CNN Variational Autoencoder [13] was trained on objects from 4 classes: ‘mug’, ‘bowl’, ‘bottle’, and ‘can’, common table-top items. The encoder is conditioned on the class by concatenating the class one-hot vector as an extra channel to each occupancy voxel in the input, while the decoder is conditioned by concatenating the class one-hot vector to the encoded shape descriptor, similar to [29, 32]. A KL-divergence loss is used in the latent shape space, while a binary-crossentropy loss is used for reconstruction. We choose a latent shape variable of size 16. The 3D CNN encoder has 5 convolutional layers with kernel size 4 and stride 2, each layer doubles the channel size except the first one which increases it to 16. The decoder mirrors the encoder using deconvolutions.

The elements of our VAE network are shown in Figure 2.

4. Probabilistic Rendering

Rendering is the process of projecting a 3D model into image space. Given the pose of the grid with respect to the camera T_{CG} , we wish to render a depth image. We denote the rendered depth image as $\hat{\delta}_\mu$ with uncertainty

$\hat{\delta}_{var}$, and the rendering function $R()$, such that $\hat{\delta}_\mu, \hat{\delta}_{var} = R(G, T_{GC})$. When designing our render function, we wish for it to satisfy three important requirements: to be differentiable and probabilistic so that it can be used for principled inference, and to have a wide receptive field so that its gradients behave properly during optimisation. These features make a robust function that can handle real world noisy measurements such as depth images.

We now describe the algorithm for obtaining the depth value for pixel (u, v) :

Point sampling. Sample M points uniformly along back-projected ray r , in depth range $[\hat{\delta}_{min}, \hat{\delta}_{max}]$. Each sampled depth $\hat{\delta}_i = \hat{\delta}_{min} + \frac{i}{M}(\hat{\delta}_{max} - \hat{\delta}_{min})$ and position in the camera frame $s_i^C = \hat{\delta}_i r$. Each sampled point is transformed into the voxel grid coordinate frame as $s_i^G = T_{GC}s_i^C$.

Occupancy interpolation. Obtain occupancy probability $o_i = Tril(s_i^O, G)$, for point s_i^O from the occupancy grid, using trilinear interpolation from its 8 neighbouring voxels.

Termination probability. We denote the depth at pixel $[u, v]$ by $\mathbf{D}[u, v]$. Now we can calculate $p(\mathbf{D}[u, v] = \hat{\delta}_i)$ (that is, the termination probability at depth $\hat{\delta}_i$) as:

$$\phi_i = p(\mathbf{D}[u, v] = \hat{\delta}_i) = o_i \prod_{j=1}^{i-1} (1 - o_j) . \quad (1)$$

Figure 3 relates occupancy and termination probabilities.

Escape probability Now we define the escape probability (the probability that the ray doesn’t intersect the object) as:

$$\phi_{M+1} = p(\mathbf{D}[u, v] > \hat{\delta}_{max}) = \prod_{j=1}^M (1 - o_j) , \quad (2)$$

$\{\phi_i\}$ forms a discrete probability distribution.

Aggregation We obtain the rendered depth at pixel $[u, v]$ as the expected value of the random variable $\mathbf{D}[u, v]$:

$$\hat{\delta}_\mu[u, v] = \mathbb{E}[\mathbf{D}[u, v]] = \sum_{i=1}^{M+1} \phi_i \hat{\delta}_i . \quad (3)$$

d_{M+1} , the depth associated to the escape probability is set to $1.1d_{max}$ for practical reasons.

Uncertainty Depth uncertainty is calculated as:

$$\hat{\delta}_{var}[u, v] = Var[\mathbf{D}[u, v]] = \sum_{i=1}^{M+1} \phi_i (\hat{\delta}_i - \hat{\delta}_\mu)^2 . \quad (4)$$

Mask Note that we can render a segmentation mask as:

$$m[u, v] = 1 - \phi_{M+1} \quad (5)$$

For multi-object rendering we combine all the renders by taking the minimum depth at each pixel, to deal with cases when objects occlude each other:

$$\begin{aligned} \hat{\delta}_\mu[u, v] &= R(\{\hat{G}_i\}, \{T_G^i\}, T_C)[u, v] \\ &= \min\{\hat{\delta}_\mu^1[u, v], \dots, \hat{\delta}_\mu^N[u, v]\} . \end{aligned} \quad (6)$$

Figure 3 shows the relation between rendered depth and occupancy probabilities. Additionally, we apply Gaussian blur down-sampling to the resulting rendered image at different pyramid levels (4 levels with 1 pixel standard deviation each) to perform coarse to fine optimisation, this increases the spatial receptive field in the higher levels of the pyramid because each rendered pixel is associated to several back projected rays.

5. Object Shape and Pose Inference

Given a depth image from an object of a known class, we wish to infer the full shape and pose of the object. We assume we have a segmentation mask and classification of the object, which in our case is obtained with Mask-RCNN [8]. To formulate our inference method, we integrate the object shape models developed on Section 3 with a measurement function, the probabilistic render algorithm outlined in Section 4. We will now describe the inference algorithm for a single object observation setup, and this will be extended to multiple objects and multiple observations in the SLAM system described in Section 6.

5.1. Shape and Pose Optimisation

An object’s pose T_{CG} is represented as a 9-DoF homogeneous transform with \mathcal{R} , t , and S the rotation, translation and scale of the object with respect to the camera.

The shape of the object is represented with latent code \mathbf{d} , which is decoded into full occupancy grid \hat{G} using the

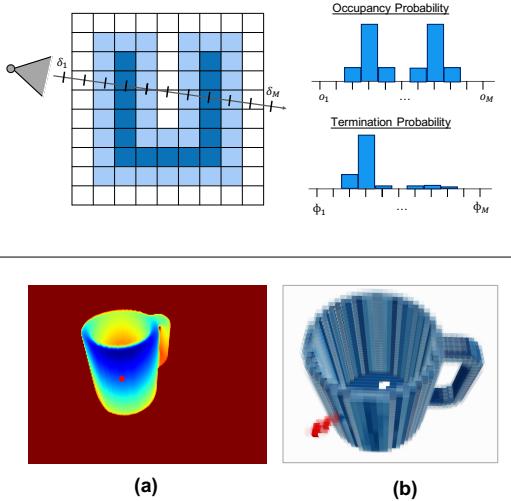


Figure 3. Pixel rendering: Each pixel is back-projected into a ray from which uniform depth samples δ_i are taken. Occupancy probability o_i is obtained from the voxel grid by trilinear interpolation, and termination probability ϕ_i is calculated. **(a):** A $32 \times 32 \times 32$ mug occupancy grid. **(b):** The derivative of the highlighted red pixel with respect to occupancy values is shown in red.

decoder described in Section 3.

We wish to find the pose and shape parameters that best explain our depth measurement δ . We consider the rendering \mathbf{D} of the object as Gaussian distributed, with mean $\hat{\delta}_\mu$ and variance $\hat{\delta}_{var}$ calculated through the render function:

$$\begin{aligned} \hat{\delta}_\mu, \hat{\delta}_{var} &= R(\hat{G}, T_{GC}) \\ &= R(D(\mathbf{d}, h), T_{GC}), \end{aligned} \quad (7)$$

with h the class one-hot vector of the detected object.

When training the latent shape space a Gaussian prior distribution is assumed on the shape descriptor. With this assumption and by taking $\hat{\delta}_{var}$ as constant, our MAP objective takes the form of least squares problem. We apply the Levenberg–Marquardt algorithm for estimation:

$$\begin{aligned} &\min_{\mathbf{d}, T_{CG}} -\log(p(\delta|\mathbf{d}, T_{CG})p(\mathbf{d})) \\ &= \min_{\mathbf{d}, T_{CG}} (L_{render}(\mathbf{d}, T_{CG}) + L_{prior}(\mathbf{d})) \\ &= \min_{\mathbf{d}, T_{CG}} \left(\sum_{u,v} \frac{(\delta[u, v] - \hat{\delta}_\mu[u, v])^2}{\hat{\delta}_{var}[u, v]} + \sum_i \mathbf{d}_i^2 \right). \end{aligned} \quad (8)$$

A structural prior is added to the optimisation loss to force the bottom on the object to be in contact with the supporting plane. We render an image from a virtual camera under the object and recover the surface mesh from the occupancy grid by marching cubes. Figure 4 illustrates the single object shape and pose inference pipeline.

5.2. Variable Initialisation

Second order optimisation methods such as Levenberg–Marquardt require a good initialisation. The object’s translation and scale are initialised using the backprojected point cloud from the masked depth image. The first is set to the centroid of the point cloud, while the latter is recovered from the centroid’s distance to the point cloud boundary.

Our model classes (‘mug’, ‘bowl’, ‘bottle’, and ‘can’) are often found in a vertical orientation in a horizontal surface. For this reason we detect the horizontal surface using the point cloud from the depth image and initialise the object’s orientation to be parallel to the surface normal. One class of objects, ‘mug’, is not symmetric around the vertical axis. To initialise the rotation of this class along the vertical axis we train a CNN. The network takes as input the cropped object from the RGB image, and outputs a single rotation angle along the vertical object axis. We train the network in simulation using realistic renders from a physically-based rendering engine, Pyrender, randomising object’s material, lighting positions and colors. The network has a VGG-11 [28] backbone pre-trained on ImageNet [26].

The shape descriptor is initialised to $\mathbf{d} = 0$, which gives the mean class shape under the Gaussian prior of a VAE.

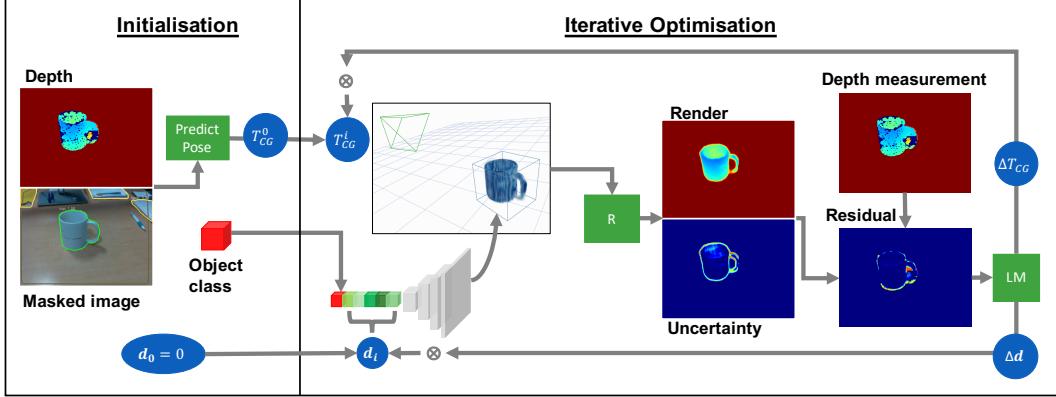


Figure 4. **Initialisation:** Initial object pose T_{CG}^0 is estimated from a depth image and masked RGB image; object class is inferred from RGB only. The shape descriptor d is set to 0, representing the mean class shape. **Optimisation:** The shape descriptor is decoded into a full voxel grid, which is used with the pose to render an object depth map. The least squares residual between this and depth is used update the shape descriptor and object pose iteratively with the Levenberg-Marquardt algorithm.

Optimisation iteratively deforms the mean shape to best fit our observations. Figure 5 illustrates how changes in the shape descriptor alter the shape of the object.

6. Object-Level SLAM System

We have developed class level shape models and a measurement function that allows us to infer object shape and pose from a single RGB-D image. From stream of images we want to incrementally build a map of all the objects in a scene while simultaneously tracking the position of the camera. For this, we will show how to use the render module for camera tracking, and for joint optimisation of camera poses, object shapes, and object poses with respect to multiple image measurements. This will allow us to construct a full, incremental, jointly optimisable object-level SLAM system with sliding keyframe window optimisation.

6.1. Data association and Object Initialisation

For each incoming image, we first segment and detect the classes of all objects in the image using Mask-RCNN [8]. For each detected object instance, we try to associate it with one of the objects already reconstructed in the map.



Figure 5. **Shape descriptor influence:** the derivative of a rendered decoded voxel grid with respect to 8 entries of the shape descriptor.

This is done in a two stage process:

Previous frame matching: We match the masks in the image with masks from the previous frame. Two segmentations are considered a match if their IoU is above 0.2.

Object mask rendering: If a mask is not matched in stage 1, we try to match it directly with map objects by rendering their masks and computing IoU overlaps.

If a segmentation is not matched with any existing objects we initialise a new object as in Section 5.

6.2. Camera Tracking

We wish to track the camera pose T_C^j for the latest depth measurement δ_j . Once we have performed association between segmentation masks and reconstructed objects as described in Section 6.1, we have a list of matched object descriptors $\{\mathbf{d}_1, \dots, \mathbf{d}_N\}$. We initialise our estimate for T_C^j as the tracked pose of the previous frame T_C^{j-1} , and render the matched objects as described in Section 4:

$$\hat{\delta}_\mu, \hat{\delta}_{var} = R(\{\hat{G}_i\}, \{T_G^i\}, T_C^j). \quad (9)$$

The loss between render and measured depth is:

$$L_{render}(\{\mathbf{d}_i\}, \{T_G^i\}, T_C^j) = \sum_{u,v} \frac{(\delta_j[u, v] - \hat{\delta}_\mu[u, v])^2}{\hat{\delta}_{var}[u, v]}. \quad (10)$$

Notice that this is the same loss used when inferring object pose and shape, but now we assume that the map (the object shapes and poses) is fixed and we want to estimate the camera pose T_C^j . As before, we use the iterative Levenberg–Marquardt optimisation algorithm.

6.3. Sliding-Window Joint Optimisation

We have shown how to reconstruct objects from a single observation, and how to track the position of the camera by

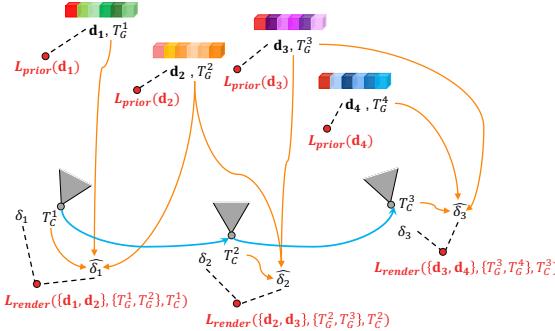


Figure 6. **Optimisation graph**, showing all jointly-optimised variables. Render and prior factors connect the different variables. A render factor compares object shape renders with depth measurements. Prior factors constrain how much each object shape can deviate from the mean shape of its class.

assuming the map is fixed. This will however lead to the accumulation of errors, causing motion drift. Integrating new viewpoint observations for an object is also desirable, to improve its shape reconstruction. To tackle these two challenges, we wish to jointly optimise a bundle of camera poses, object poses, and object shapes. Doing this with all frames is however computationally infeasible, so we jointly optimise the variables associated to a select group of frames, called keyframes, in a sliding window manner, following the philosophy introduced by PTAM [14].

Keyframe criteria: There are two criteria for selecting a frame as a keyframe. If an object was initialised in the frame then it is selected as a keyframe, or second if the frame viewpoint for any of the existing objects is larger than 13 degrees from the frame in which the object was initialised.

Bundle Optimisation: Each time that a frame is selected as a keyframe we jointly optimise the variables associated with a bundle of N keyframes. In particular we select a window of 3 keyframes, the new keyframe and its two closest keyframes, with the previously defined distance.

To formulate the joint optimisation loss, consider, T_C^1 , T_C^2 , and T_C^3 , the poses of the keyframes in the optimisation window; T_C^1 is held fixed. Now suppose $\{\mathbf{d}_i\}$ is the set of shape descriptors for the objects observed by the three keyframes. Then we can render a depth image and uncertainty for each keyframe as:

$$\hat{\delta}_\mu^j, \hat{\delta}_{var}^j = R(\{\hat{G}_i\}, \{T_G^i\}, T_C^j), \quad (11)$$

with $\hat{G}_i = D(\mathbf{d}_i, h_i)$. For each render we compute a loss with the respective depth measurement, L_{render}^j as in Equation 10, and a prior loss, L_{prior}^i on all codes as in Equation 8. Figure 6 illustrates the joint optimisation problem.

Our final loss, optimised using Levenberg-Marquardt, is:

$$L_{joint}(\{\mathbf{d}_i\}, \{T_G^i\}, \{T_C^j\}) = \sum_j L_{render}(\{\mathbf{d}_i\}, \{T_G^i\}, T_C^j) + \sum_i L_{prior}(\mathbf{d}_i). \quad (12)$$

Timings: Rendering a single object takes 7ms, full object reconstruction takes approximately 1.5 seconds. Camera tracking is 7fps and joint optimisation takes 2 seconds.

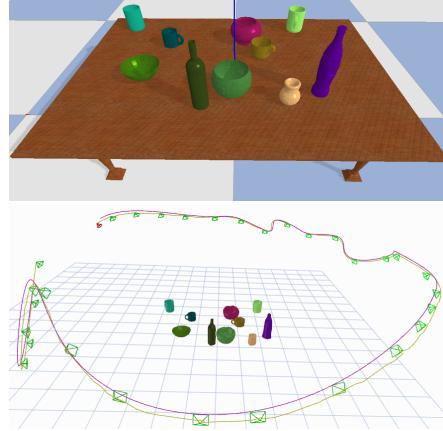


Figure 7. Synthetic scene example along with reconstruction and camera trajectory. Ground truth trajectory is shown in purple and tracked one in yellow, keyframes with green frustum.

7. Experimental Results

7.1. Metrics

For shape reconstruction evaluation we use three metrics: chamfer- L_1 distance and accuracy as defined in [21] and completeness (with 1cm threshold) as defined in [17]. We sample 20000 points on both reconstruction and ground truth CAD model meshes.

7.2. Rendering Evaluation

In this evaluation test the optimisation performance of our rendering formulation. We perform object shape and pose optimisation on all the objects of the ‘mug’ category in the ShapeNet dataset. For each instance we generate three random views of the object. Initial object pose is predicted from the first view. We perform 30 optimisation iterations for 1, 2, and 3 views. Table 1 shows median shape accuracy, completion, and chamfer distance after optimisation. We compare our full system with versions without uncertainty, without Gaussian pyramid, and with a loss only between the rendered and Mask-RCNN segmentation masks. We compare with the state of the art volumetric differential rendering component in paper Differential Volumetric Rendering (DVR) [23] with our shape representation.

Table 1. Shape reconstruction results for 1, 2, and 3 views. We do an ablation study of our method and compare with DVR [23].

| | Full | No Unc. | No Gauss. | [23] | Mask |
|---------------------|--------|---------|-----------|--------|--------|
| 1 view | | | | | |
| accuracy [mm] | 4.459 | 4.998 | 4.701 | 8.967 | 15.806 |
| chamfer- L_1 [mm] | 4.439 | 4.844 | 4.928 | 11.896 | 18.386 |
| completion [1cm] | 93.492 | 91.857 | 90.812 | 43.075 | 30.212 |
| 2 views | | | | | |
| accuracy [mm] | 3.752 | 4.270 | 4.237 | 8.408 | 4.709 |
| chamfer- L_1 [mm] | 3.854 | 4.185 | 4.723 | 11.325 | 4.438 |
| completion [1cm] | 95.72 | 94.627 | 90.752 | 43.342 | 93.73 |
| 3 views | | | | | |
| accuracy [mm] | 3.484 | 4.158 | 3.827 | 8.277 | 4.620 |
| chamfer- L_1 [mm] | 3.648 | 4.010 | 4.281 | 10.913 | 4.210 |
| completion [1cm] | 96.065 | 95.165 | 93 | 44.815 | 95.44 |

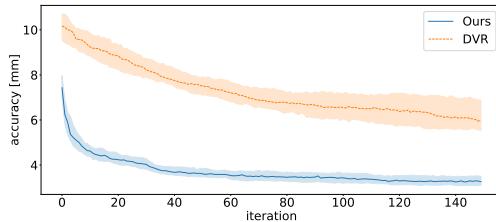


Figure 8. Median reconstruction accuracy (95% confidence) across 150 optimisation iterations of all ‘mug’ objects instances comparing our proposed renderer with [23].

We observe that additional views improve shape reconstruction, more drastically in the mask optimisation because of the scale ambiguity in a single image. We also see that both the uncertainty and Gaussian pyramid are necessary for more accurate and complete shape reconstructions. Our method significantly improves on DVR, which is both less precise and has much lower shape completion, because of its local receptive field.

To further illustrate the comparison, we plot in Figure 8 median reconstruction accuracy across 150 optimisation iterations with all object instances against our proposed method. The plot illustrates the much faster convergence of our method and its ability to reach a lower accuracy error.

7.3. SLAM evaluation

In this evaluation we evaluate our full SLAM system and how it generalises to new object instances. We create a synthetic dataset. Random object CAD models are spawned on top of a table model with random positions and vertical orientation. Five scenes are created with 10 different objects on each from three classes: ‘mug’, ‘bowl’, and ‘bottle’. The models are obtained from the ModelNet40 dataset [39] which are not used during training of the shape model.

For each scene a random trajectory is generated by sampling and interpolating random camera positions and look at points in the volume bounded by the table. Image and depth renders are obtained from the trajectory with PyBul-

let render, which is different rendering engine than the one used for training pose prediction.

7.3.1 Fusion++ comparison

We compare our proposed method with a custom implementation of Fusion++ [19] using open-source TSDF fusion [41] for each object volume. In this experiment ground truth poses are used to decouple tracking accuracy and reconstruction quality. Gaussian noise is added to the depth image and camera poses (2mm, 1mm, 0.1° standard deviation for depth, translation and orientation, respectively).

We evaluate shape completion and accuracy; results are accumulated for each class from the 5 simulated sequences. Figure 11 shows how mean shape completion evolves with respect to frame number. This graph demonstrates the advantage of class-based priors for object shape reconstruction. With our method we see a jump to almost full completion, while TSDF fusion slowly completes the object with each new fused depth map. Fast shape completion without the need for exhaustive 360 degree scanning is important in robotic applications and in augmented reality, as shown in Figure 9. Figure 11 displays the median shape accuracy of Node-SLAM compared with TSDF fusion. We observe comparable surface reconstruction quality of close to 5mm.

7.3.2 Ablation Study

We evaluate shape reconstruction accuracy and tracking *absolute pose error* on 3 different versions of our system. We compare our full SLAM system (with camera tracking) with a version without sliding window joint optimisation, and a version without uncertainty rendering. Figure 11 shows the importance of these features for shape reconstruction quality, with decreases in performance from 2 up to 7 mm. Table 2 shows mean *absolute pose error* for each version of our system for all 5 trajectories. These results prove that the precise shape reconstructions from objects provide enough information for accurate camera tracking with mean errors between 1 and 2 cm. It also shows how tracking without



Figure 9. **Few-shot augmented reality:** Complete and watertight meshes can be obtained from few images due to the learned shape priors. This are then loaded into a physics engine to perform realistic augmented reality demonstrations.

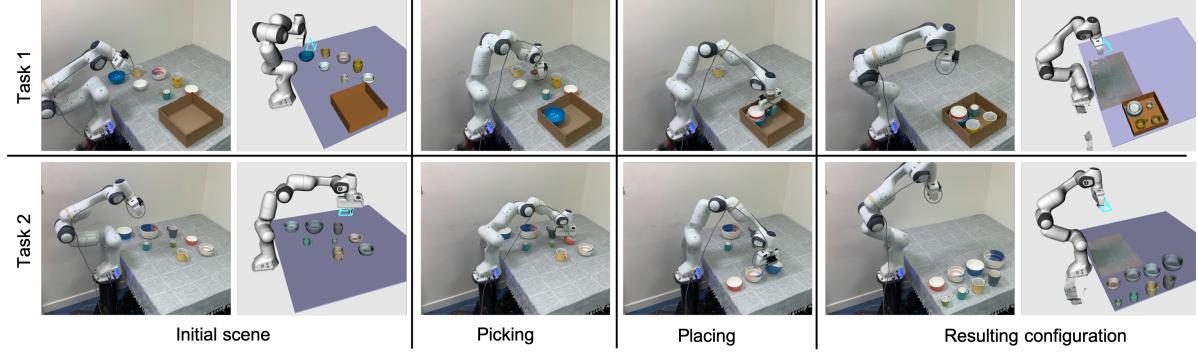


Figure 10. Robotic demonstration of packing (task 1) and sorting (task 2) of objects.

Table 2. Ablation study for tracking accuracy on 5 scenes, highlighting the importance of a joint optimisation with uncertainty.

| Absolute Pose Error [cm] | Scene 1 | Scene 2 | Scene 3 | Scene 4 | Scene 5 |
|--------------------------|---------|---------|---------|---------|---------|
| NodeSLAM | 1.73 | 1 | 0.81 | 1.24 | 1.15 |
| NodeSLAM no joint optim. | 8.6 | 10.17 | 0.7 | 2.14 | 1.25 |
| NodeSLAM no uncertainty | 4.37 | 3.41 | 0.88 | 3.05 | 6.99 |

joint optimisation or uncertainty leads to significantly lower accuracy on most trajectories.

7.4. Robot Manipulation Application

We have developed a manipulation application which uses our object reconstruction system. We demonstrate two tasks: object packing and object sorting; see Figure 10 and the attached video. A rapid pre-defined motion is first used to gather a small number of RGB-D views which our system uses to estimate the pose and shape of the objects laid out randomly on a table. Heuristics are used for grasp point selection and a placing motion based on the class and pose of the object and the shape of the reconstructed mesh. All the reconstructed objects are then sorted based on height and radius. For the packing task all the scanned objects are placed in a tight box, with bowls stacked in decreasing size order and all mugs placed inside the box with centers and orientations aligned. In the sorting task all objects are placed in a line in ascending size. In this robot application only, robot kinematics are used for camera tracking.

8. Conclusions

Our generative multi-class object models allow for principled and robust full shape inference. We have shown practical use in a jointly optimisable object SLAM system as well as in two robotic manipulation demonstrations and an augmented reality demo. We believe that this proves that

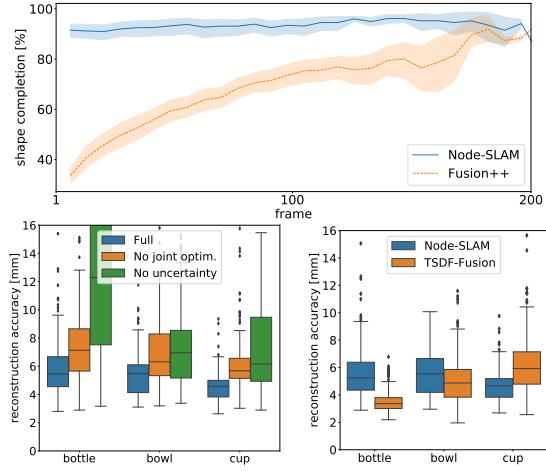


Figure 11. **Top:** Graph with mean object surface completion (95% confidence) comparison between NodeSLAM and TSDF fusion, with respect to the number of times an object is updated. **Bottom left:** Box plots of median surface reconstruction accuracy from our ablation study on 5 scenes with 10 objects in each. **Bottom right:** The same metric but comparing our system with Fusion++.

decomposing a scene into full object entities is a powerful idea for robust mapping and smart interaction. Not all object classes will be well represented by the single code object VAE we used in this paper, and in near future work we plan to investigate alternative coding schemes such as methods which can decompose a complicated object into parts, and methods which can be trained by self-supervision.

Acknowledgements

Research presented in this paper has been supported by Dyson Technology Ltd. We thank Michael Bloesch, Shuaifeng Zhi, and Joseph Ortiz for fruitful discussions.

References

- [1] 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. *arXiv preprint arXiv:1512.03012*, 2015.
- [2] C. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
- [3] A. Dame, V. A. Prisacariu, C. Y. Ren, and I. Reid. Dense reconstruction using 3d object shape priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [4] A. J. Davison, N. D. Molton, I. Reid, and O. Stasse. MonoSLAM: Real-Time Single Camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(6):1052–1067, 2007.
- [5] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2017.
- [6] F. Engelmann, J. Stückler, and B. Leibe. Samp: shape and motion priors for 4d vehicle reconstruction. In *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV)*, 2017.
- [7] G. Gkioxari, J. Malik, and J. Johnson. Mesh r-cnn. 2019.
- [8] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.
- [9] L. Hu, W. Xu, K. Huang, and L. Kneip. Deep-slam++: Object-level rgbd slam based on class-specific deep shape priors. *arXiv preprint arXiv:1907.09691*, 2019.
- [10] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. A. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. J. Davison, and A. Fitzgibbon. KinectFusion: Real-Time 3D Reconstruction and Interaction Using a Moving Depth Camera. In *Proceedings of ACM Symposium on User Interface Software and Technolog (UIST)*, 2011.
- [11] Y. Jiang, D. Ji, Z. Han, and M. Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [12] J. T. Kajiya and B. P. Von Herzen. Ray tracing volume densities. 1984.
- [13] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.
- [14] G. Klein and D. W. Murray. Parallel Tracking and Mapping for Small AR Workspaces. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 2007.
- [15] A. Kundu, Y. Li, and J. M. Rehg. 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [16] K. Li, R. Garg, M. Cai, and I. Reid. Single-view object shape reconstruction using deep shape prior and silhouette. 2019.
- [17] K. Li, M. Rünz, M. Tang, L. Ma, C. Kong, T. Schmidt, I. Reid, L. Agapito, J. Straub, S. Lovegrove, et al. Frodo: From detections to 3d objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [18] S. Liu, Y. Zhang, S. Peng, B. Shi, M. Pollefeys, and Z. Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [19] J. McCormac, R. Clark, M. Bloesch, A. J. Davison, and S. Leutenegger. Fusion++:volumetric object-level slam. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2018.
- [20] J. McCormac, A. Handa, A. J. Davison, and S. Leutenegger. SemanticFusion: Dense 3D semantic mapping with convolutional neural networks. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017.
- [21] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [22] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011.
- [23] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [24] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. 2019.
- [25] D. Paschalidou, A. O. Ulusoy, and A. Geiger. Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [26] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
- [27] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. J. Kelly, and A. J. Davison. SLAM++: Simultaneous Localisation and Mapping at the Level of Objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [28] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.
- [29] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Neural Information Processing Systems (NIPS)*, 2015.

- [30] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, and W. T. Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [31] N. Sünderhauf, T. T. Pham, Y. Latif, M. Milford, and I. Reid. Meaningful maps with object-oriented semantic mapping. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, 2017.
- [32] Q. Tan, L. Gao, Y.-K. Lai, and S. Xia. Variational autoencoders for deforming 3d mesh models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [33] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [34] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang. Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67, 2018.
- [35] R. Wang, N. Yang, J. Stueckler, and D. Cremers. Directshape: Photometric alignment of shape priors for visual vehicle pose and shape estimation. *arXiv preprint arXiv:1904.10097*, 2019.
- [36] T. Whelan, S. Leutenegger, R. F. Salas-Moreno, B. Glocker, and A. J. Davison. ElasticFusion: Dense SLAM without a pose graph. In *Proceedings of Robotics: Science and Systems (RSS)*, 2015.
- [37] J. Wu, Y. Wang, T. Xue, X. Sun, B. Freeman, and J. Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. In *Neural Information Processing Systems (NIPS)*, 2017.
- [38] J. Wu, C. Zhang, T. Xue, W. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Neural Information Processing Systems (NIPS)*, 2016.
- [39] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3D ShapeNets: A Deep Representation for Volumetric Shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [40] Y. Xiang and D. Fox. DA-RNN: Semantic mapping with data associated recurrent neural networks. *arXiv preprint arXiv:1703.03098*, 2017.
- [41] Q.-Y. Zhou, J. Park, and V. Koltun. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847*, 2018.
- [42] R. Zhu, C. Wang, C.-H. Lin, Z. Wang, and S. Lucey. Object-centric photometric bundle adjustment with deep shape prior. In *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV)*, 2018.