

Neural Object Descriptors for Multi-View Shape Reconstruction

Edgar Sucar, Kentaro Wada, and Andrew Davison

Dyson Robotics Lab
Imperial College London, UK
`{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.

Keywords: generative, object, reconstruction, render, SLAM, multi-view

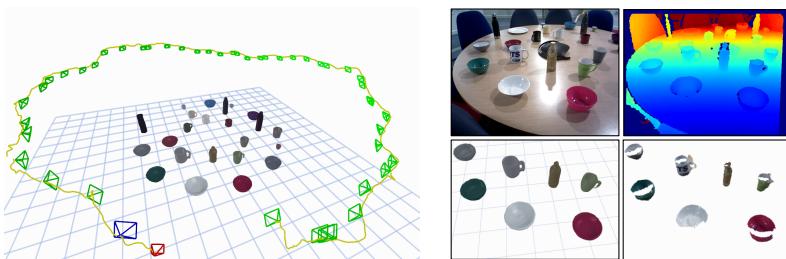


Fig. 1. Left: Compact and optimisable shape models are used to build an object-level SLAM system that can handle a real world cluttered table top scene with varied object shapes from different classes. **Right:** The use of class-level priors allows us to get accurate and complete object reconstructions (bottom-left) even from a single image in contrast to the partial reconstruction from TSDF fusion (bottom-right).

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, and especially of the objects available for interaction. The desired properties of object representations include: (i) compactness, enabling principled and efficient optimisation; (ii) consideration of semantic priors, so that complete models can be built from partial observation; and (iii) the ability to 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 in multi-object scenes.

A generative object model allows for optimisation and for principled probabilistic integration of multiple image measurements. This is in contrast to discriminative approaches which learn a function (commonly a neural network) for mapping images of objects into their full 3D shape. Single image reconstruction approaches such as [11,29,25,26,5,30] lack the ability to integrate multiple observations in a principled way. For example, DeepSLAM++ [7] directly averages 3D shapes predicted by Pix3D [22], while 3D-R2N2 [2] uses a recurrent network to update shapes from new images.

Our work comes at a time when many authors are building accurate and flexible generative models for objects [28,16,15,17]. However, there is a lack of progress in using these models in real world shape reconstruction from images. Existing methods which use generative models for inferring object shape [3,4,27,12,33,13] are normally constrained to single observations or single classes. An emphasis of this paper is to design object models that work robustly in the real world. The use of rendering-based optimisation combined with class-level priors enables multi-view reconstruction of objects from different categories, even with noisy depth measurements. We prove this through a robotic manipulation application, an augmented reality demo, and the use of learned models in a full jointly-optimised object SLAM system.

We take advantage of the high regularity in object shapes from the same class and train a class-conditioned Variational Auto Encoder (VAE) on aligned volumetric models of several classes (Section 2). This allows us to represent objects through a compact code, and to model class shape prior information through the decoder network which is used to recover the shape from the object code. To be able to use a generative method in inference we need a measurement function. In our case we use depth images with object masks as measurements. We introduce a novel probabilistic and differentiable rendering engine for transforming object volumes into depth images with associated uncertainty (Section 3). This allows us to optimise an object shape represented by a compact code against one or more depth images to obtain a full object shape reconstruction (Section 4). The combination of the object VAE with the probabilistic rendering engine allows us to tackle the important question of how to integrate several measurements for shape inference using semantic priors.

When it comes to building models of scenes with many objects and from multiple observations, our optimisable compact object models can serve as the landmarks in an object-based SLAM map. Object-based SLAM has been previously attempted, but our work fills a gap between systems which make separate reconstructions of every identified object [14,23], which are general with respect

to arbitrary classes but do not take account of shape priors and can only reconstruct directly observed surfaces, and those which populate maps with instances of object CAD models [19], 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. We use our object models to build the first jointly optimisable object-level SLAM system, which uses the same measurement function for camera tracking as well as for joint optimisation of object poses and shapes, and camera poses (Section 5).

To demonstrate the capacity of our system we choose 4 classes of common table-top items: ‘mugs’, ‘bowls’, ‘bottles’, and ‘cans’. We construct a synthetic dataset and prove that our models achieve comparable surface reconstruction quality to purely data driven TSDF fusion, while reaching full object reconstruction from much fewer observations, at a minimum even only one view (Section 6). We qualitatively show that our SLAM system can handle real world cluttered scenes with varied object shapes as shown on Figure 1. Furthermore we show that the completeness and accuracy of our object reconstructions enable robotic tasks such as packing objects into a tight space 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:

- Multi-view object shape reconstruction enabled by a novel differentiable and probabilistic depth rendering engine combined with multi-class object descriptors represented with a VAE network.
- The first object-level SLAM capable of jointly optimising full object shapes and poses together with camera trajectory from real world images.
- The integration into a practical robot system that can do useful manipulation tasks with varied object shapes from different categories due to the high quality surface reconstructions.

2 Class-Level Object Shape Descriptors

Objects of the same semantic class exhibit strong regularities in shape under common pose alignment. We make three key observations: (i) Given two objects of the same class, there is a pose alignment between them that allows for a smooth surface deformation between the two objects; (ii) This pose alignment is common among all instances of the same class, which defines a class-specific coordinate frame; (iii) If we select two random objects of a certain class and smoothly deform one into the other, there will be other object instances of the same class which are similar to the intermediate deformations.

We leverage these characteristics to construct a class specific smooth latent space which allows us to represent the shape of an instance with a small number of parameters. This is motivated by the fact that the space of valid inter-class surface deformations is much smaller than the space of all possible deformations; there are high correlations between the surface points in a valid deformation.

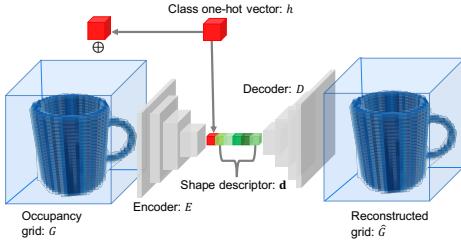


Fig. 2. Occupancy Variational Autoencoder: The class one hot vector h is concatenated channel-wise to each occupancy voxel in the input occupancy grid G . The input is compressed into shape descriptor \mathbf{d} by encoder network E . The shape descriptor and the class-one hot vector are concatenated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

Rather than manually designing a parameterised shape model for a class of objects, we propose instead to learn the latent space by training a single Class-Conditional Variational Autoencoder neural network.

2.1 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. We store occupancy values to enable probabilistic rendering and inference.

The 3D models used 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 [9] 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 [21,24]. 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. We use G to denote the 3D voxel occupancy grid, h the class one-hot vector, $E()$ the network’s encoder function, $\mathbf{d} = E(G, h)$ the encoded shape descriptor, $D()$ the decoder function, and $\hat{G} = D(\mathbf{d}, h)$ the reconstructed voxel occupancy grid.

3 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 $\hat{\delta}_\mu, \hat{\delta}_{var} = R(G, T_{GC})$. Our rendering engine differs from existing occupancy volume [25,32], mesh [8], and SDF [13] renderers by providing the capacity to render depth uncertainty and by having a higher receptive field either in the depth image or along each back-projected ray.

When designing our render function, we wish for it to satisfy three important requirements: for it to be differentiable so that it can be used for optimisation, for it to be probabilistic so that it can be used in a principled way in inference, and for it 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 backprojected 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 will consider the depth at pixel $[u, v]$ as a random variable $\mathbf{D}[u, v]$. 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 shows the relation between 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 Now, we can 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.

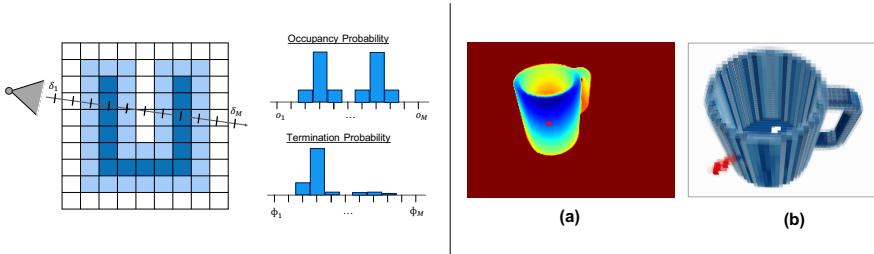


Fig. 3. Pixel rendering: Each pixel is back projected into a ray from which uniform depth samples δ_i are taken. For each depth sample the occupancy probability o_i is obtained from the voxel grid by trilinear interpolation. After this, the probability ϕ_i that the ray terminates at each depth sample is calculated. The figure shows how later depth samples have low termination probability. **(a):** Rendering example of a $32 \times 32 \times 32$ mug occupancy grid. **(b):** the derivative of the red highlighted pixel with respect to occupancy values is shown by a red color map. Increasing the occupancy value of the red voxels would decrease the depth of the highlighted pixel.

Uncertainty The uncertainty of the depth can be calculated as:

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

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:

$$\begin{aligned} \hat{\delta}_\mu[u, v] &= R(\{\hat{G}_i\}, \{T_G^i\}, T_C)[u, v] \\ &= \min\{\hat{\delta}_\mu^1[u, v], \dots, \hat{\delta}_\mu^N[u, v]\}. \end{aligned} \quad (5)$$

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 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. This rendering formulation meets our initially established requirements.

4 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 [6]. To formulate our inference method, we integrate the object shape models developed on Section 2 with a measurement function, the probabilistic render algorithm outlined in Section 3. 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 5.

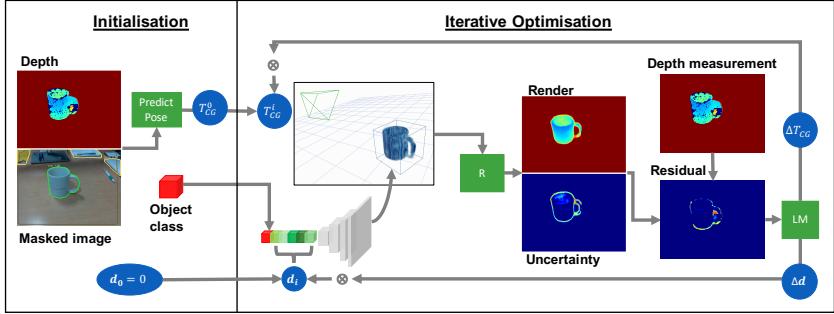


Fig. 4. Initialisation: Initial object pose T_{CG}^0 is estimated from a depth image and masked RGB image; object class is inferred from RGB only. The shape descriptor \mathbf{d} is set to 0, representing the mean class shape. **Optimisation:** The shape descriptor is decoded into a full voxel grid, 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.

4.1 Shape and Pose Optimisation

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

$$T_{CG} = \begin{bmatrix} S\mathcal{R} & t \\ 0 & 1 \end{bmatrix}, S = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & s_z \end{bmatrix}. \quad (6)$$

The shape of the object is represented with latent code \mathbf{d} , which is decoded into full occupancy grid \hat{G} using the decoder described in Section 2.

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 LevenbergMarquardt algorithm for estimation:

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

We perform coarse to fine optimisation at different pyramid levels obtained by Gaussian blurring. A structural prior is added to the optimisation loss enforcing the bottom surface of the object to be in contact with the supporting plane. This is done by rendering an image from a virtual camera under the object. The mesh for the surface is recovered from the occupancy grid by marching cubes. Figure 4 illustrates the single object shape and pose inference pipeline.

4.2 Variable Initialisation

Second order optimisation methods such as LevenbergMarquardt require a good initialisation. The object’s translation and scale are initialised using the back-projected 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 [20] backbone pre-trained on ImageNet [18].

When training our Class-Conditional Variational Autoencoder we assume the latent shape space has a canonical Gaussian distribution. This means that



Fig. 5. Shape descriptor influence: We show the derivative of the rendering of a decoded voxel grid with respect to 8 entries of the shape descriptor. Red represents negative derivatives and blue positive. Color transparency shows derivative intensity. We observe that the descriptor’s influence is distributed throughout the object, with each code entry clustering on certain regions.

if take the zero shape descriptor $\mathbf{d} = 0$, conditioned on a given class h , the decoder will output the mean shape for that class. Given the detected class, we initialise the shape descriptor as 0 when starting optimisation. We can interpret the optimisation as iteratively deforming the mean shape of the given class into the resulting shape that best describes our observations while keeping the shape consistent with the class (0 prior on descriptor). Figure 5 illustrates how changes in the shape descriptor are reflected in the shape of the object.

5 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.

5.1 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 [6]. For each detected object instance, we try to associate it with one of the objects already reconstructed in the map. 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. If a mask from the previous frame is associated with a reconstructed object, the current matched mask is also associated with that object.

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

If a segmentation is not matched with any existing objects we initialise a new object by inferring its pose and shape as described in Section 4.

5.2 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 5.1, 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 3:

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

Now we can compute the loss between render and depth measurement as:

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

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 LevenbergMarquardt optimisation algorithm.

5.3 Sliding-Window Joint Optimisation

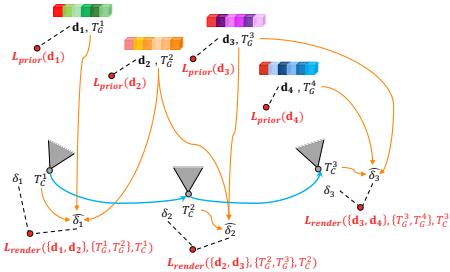


Fig. 6. Joint optimisation graph: Graph of shape descriptors and camera poses which is jointly optimised in a sliding window manner. Render and prior factors connect the different variables. Render factor compares object shape renders with depth measurements. Prior factors constrain how much each object shape can deviate from the mean shape of its class.

We have shown how to reconstruct objects from a single observation, and how to track the position of the camera by 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 [10].

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 object is different enough than from the keyframe when the object was initialised. For this we take the maximum viewpoint angle difference for all objects and if it is above a threshold of 13 degrees we set the frame as a keyframe.

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.

When performing joint optimisation the pose of the oldest keyframe is fixed. It is also important to highlight that the poses of all frames and objects are represented relative to their parent keyframe, so after each bundle optimisation their world pose is automatically updated.

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 $\hat{G}_i = D(\mathbf{d}_i, h_i)$.

Now for each render we compute a render 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 has a graph representation of the joint optimisation problem. Our final joint loss, optimised using Levenberg-Marquardt, is:

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

6 Experimental Results

6.1 Experimental setup:

For evaluating shape reconstruction quality and tracking accuracy 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 [30] which are not used during training of the shape model.

For each scene a random trajectory is generated by sampling random camera positions in spherical coordinates, with random look at points in the volume bounded by the table. The camera positions are sorted and interpolated to guarantee a smooth motion. Image and depth renders are obtained from the trajectory with PyBullet library, which uses a different rendering engine than the one used for training the pose prediction network.

Metrics: For shape reconstruction evaluation we use two metrics [16], *shape completion* and *reconstruction accuracy*. Given the meshes for the object reconstruction and ground truth CAD model, we first sample a point cloud of 2000 points from each. For each point from the CAD model we compute the distance to the closest point in the reconstruction, and if the distance is smaller than 1cm the point is considered successful. *Shape completion* is defined as the proportion of successful points ranging from 0 to 1.

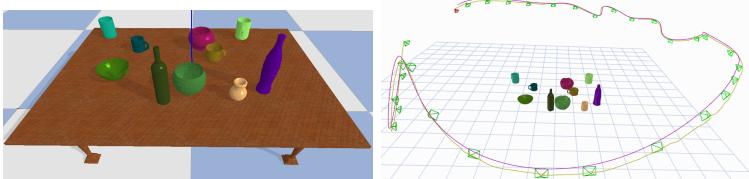


Fig. 7. Synthetic scene example along with reconstruction and camera trajectory. Ground truth trajectory is shown in purple and tracked one in yellow. Inserted keyframes during SLAM are shown with green frustum.

For each point in the reconstruction we compute the distance (in mm) to the closest point in the CAD model, and average to get *reconstruction accuracy*.

6.2 Evaluation

We compare our proposed method with open-source TSDF fusion [31]. For every frame in the sequence we fuse the masked depth map into a TSDF volume for each object, with voxel size 3mm, which is approximately the same as the used by our method. Only in this experiment ground truth poses are used, to decouple tracking accuracy and reconstruction quality evaluations. Gaussian noise is added to the depth image and camera poses (2mm, 1mm, 0.1 degrees standard deviation for depth, translation and orientation, respectively).

For all objects visible at each keyframe we evaluate shape completion and accuracy. Results are accumulated for each class from the 5 simulated sequences. Figure 9 shows how shape completion evolves with respect to the number of times each object is updated. This graph clearly demonstrates the advantage of using class-based priors for object shape reconstruction. With our proposed 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 to reconstruct occluded or unseen parts of an object for manipulation and in augmented reality for realistic simulations, as shown in Figure 8.

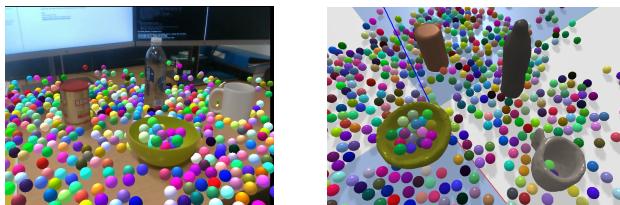


Fig. 8. Few-shot augmented reality: Complete and watertight meshes can be obtained from few images due to the learned shape priors. This are then loaded into a physics engine to perform realistic augmented reality demonstrations. Clear occlusion boundaries demonstrate the quality of the reconstruction.

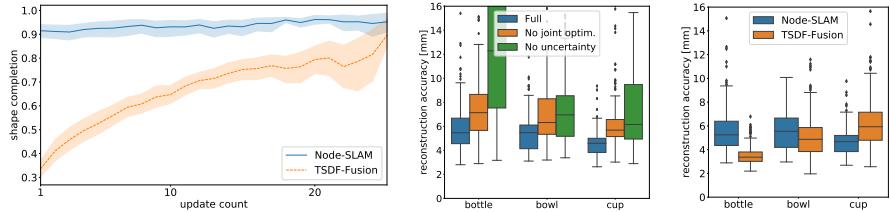


Fig. 9. Left: Graph with object surface completion comparison between NodeSLAM and TSDF fusion, with respect to number of times an object is updated. We can see how NodeSLAM jumps very fast to an almost full completion while TSDF fusion requires many more observations. **Middle:** Box plots of surface reconstruction accuracy from our ablation study ran on 5 scenes with 10 objects in each. These results highlight the importance of multi-view optimisation and uncertainty rendering to get high quality reconstructions. **Right:** The same metric shown on the left but comparing the proposed system with TSDF fusion, using ground truth tracking. The results show comparable reconstruction accuracy between methods.

Figure 9 displays the shape accuracy of Node-SLAM compared with TSDF fusion. Purely data driven TSDF fusion is considered a gold standard for surface reconstruction quality, as demonstrated by its use in a wide variety of applications such as augmented reality. We observe comparable surface reconstruction quality of close to 5mm. Our method performs better on the ‘mug’ class, possibly because it handles noise on thin structures better.

6.3 Ablation Study

We have evaluated shape reconstruction accuracy and tracking *absolute pose error* on 3 different versions of our system. We compare the full version of our SLAM system with a version without sliding window joint optimisation, and a version without uncertainty rendering, the main novelty in our rendering module. Figure 9 shows the importance on these features for shape reconstruction quality, with decreases in performance from 2 up to 7 mm. Table 1 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 joint optimisation or rendering uncertainty leads to significantly lower accuracy on most trajectories.

Table 1. Ablation study for tracking accuracy on 5 different scenes, highlighting the importance of using a principled 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

6.4 Robot Manipulation Application

We have developed a manipulation application which uses our object reconstruction system. We demonstrate two tasks: object packing and object sorting. 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 on top of each other 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, robot kinematics are used for camera tracking.

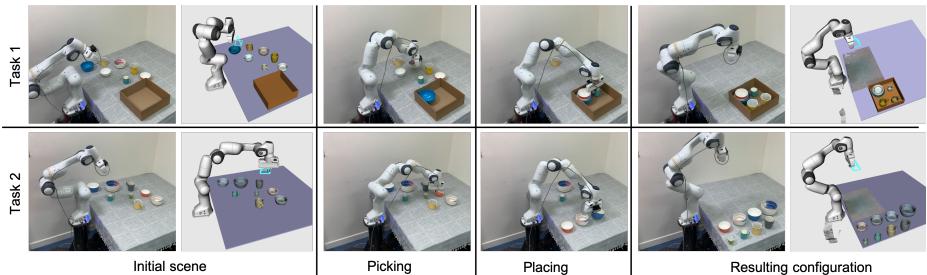


Fig. 10. Robotic demonstration of packing (task 1) and sorting (task 2) of objects.

7 Conclusions

We have developed generative multi-class object models which allow for principled and robust full shape inference. We demonstrated their 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 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. In the longer term, we wish to investigate how to allow for object models to represent arbitrary scenes, which will need the development of more general shape priors and the ability to learn directly from images without the need for 3D object models.

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. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
2. Choy, C., Xu, D., Gwak, J., Chen, K., Savarese, S.: 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. Dame, A., Prisacariu, V.A., Ren, C.Y., Reid, I.: Dense reconstruction using 3d object shape priors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
4. Engelmann, F., Stückler, J., Leibe, B.: Samp: shape and motion priors for 4d vehicle reconstruction. In: Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV) (2017)
5. Gkioxari, G., Malik, J., Johnson, J.: Mesh r-cnn (2019)
6. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the International Conference on Computer Vision (ICCV) (2017)
7. Hu, L., Xu, W., Huang, K., Kneip, L.: Deep-slam++: Object-level rgbd slam based on class-specific deep shape priors. arXiv preprint arXiv:1907.09691 (2019)
8. Kato, H., Ushiku, Y., Harada, T.: Neural 3D mesh renderer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3907–3916 (2018)
9. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: Proceedings of the International Conference on Learning Representations (ICLR) (2014)
10. Klein, G., Murray, D.W.: Parallel Tracking and Mapping for Small AR Workspaces. In: Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR) (2007)
11. Kundu, A., Li, Y., Rehg, J.M.: 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)
12. Li, K., Garg, R., Cai, M., Reid, I.: Single-view object shape reconstruction using deep shape prior and silhouette (2019)
13. Liu, S., Zhang, Y., Peng, S., Shi, B., Pollefeys, M., Cui, Z.: Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. arXiv preprint arXiv:1911.13225 (2019)
14. McCormac, J., Clark, R., Bloesch, M., Davison, A.J., Leutenegger, S.: Fusion++: volumetric object-level slam. In: Proceedings of the International Conference on 3D Vision (3DV) (2018)
15. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
16. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation (2019)

17. Paschalidou, D., Ulusoy, A.O., Geiger, A.: Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
18. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015)
19. Salas-Moreno, R.F., Newcombe, R.A., Strasdat, H., Kelly, P.H.J., Davison, A.J.: SLAM++: Simultaneous Localisation and Mapping at the Level of Objects. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013), <http://dx.doi.org/10.1109/CVPR.2013.178>
20. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015)
21. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Neural Information Processing Systems (NIPS) (2015)
22. Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J.B., Freeman, W.T.: Pix3d: Dataset and methods for single-image 3d shape modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
23. Sünderhauf, N., Pham, T.T., Latif, Y., Milford, M., Reid, I.: Meaningful maps with object-oriented semantic mapping. In: Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS) (2017)
24. Tan, Q., Gao, L., Lai, Y.K., Xia, S.: Variational autoencoders for deforming 3d mesh models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
25. Tulsiani, S., Zhou, T., Efros, A.A., Malik, J.: 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)
26. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 52–67 (2018)
27. Wang, R., Yang, N., Stueckler, J., Cremer, D.: Directshape: Photometric alignment of shape priors for visual vehicle pose and shape estimation. arXiv preprint arXiv:1904.10097 (2019)
28. Wu, J., Zhang, C., Xue, T., Freeman, W., Tenenbaum, J.: Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: Neural Information Processing Systems (NIPS) (2016)
29. Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, B., Tenenbaum, J.: Marrnet: 3d shape reconstruction via 2.5 d sketches. In: Neural Information Processing Systems (NIPS) (2017)
30. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3D ShapeNets: A Deep Representation for Volumetric Shapes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
31. Zhou, Q.Y., Park, J., Koltun, V.: Open3D: A modern library for 3D data processing. arXiv preprint arXiv:1801.09847 (2018)
32. Zhu, J.Y., Zhang, Z., Zhang, C., Wu, J., Torralba, A., Tenenbaum, J., Freeman, B.: Visual object networks: image generation with disentangled 3d representations. In: Neural Information Processing Systems (NIPS) (2018)

33. Zhu, R., Wang, C., Lin, C.H., Wang, Z., Lucey, S.: Object-centric photometric bundle adjustment with deep shape prior. In: Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV) (2018)