# Multi-view 3D Object Reconstruction and Uncertainty Modelling with Neural Shape Prior

Ziwei Liao and Steven L. Waslander
Robotics Institute & Institute for Aerospace Study
University of Toronto
ziwei.liao@mail.utoronto.ca, steven.waslander@utoronto.ca

## Abstract

*3D object reconstruction is important for semantic scene understanding. It is challenging to reconstruct detailed 3D shapes from monocular images directly due to a lack of depth information, occlusion and noise. Most current methods generate deterministic object models without any awareness of the uncertainty of the reconstruction. We tackle this problem by leveraging a neural object representation which learns an object shape distribution from large dataset of 3d object models and maps it into a latent space. We propose a method to model uncertainty as part of the representation and define an uncertainty-aware encoder which generates latent codes with uncertainty directly from individual input images. Further, we propose a method to propagate the uncertainty in the latent code to SDF values and generate a 3d object mesh with local uncertainty for each mesh component. Finally, we propose an incremental fusion method under a Bayesian framework to fuse the latent codes from multi-view observations. We evaluate the system in both synthetic and real datasets to demonstrate the effectiveness of uncertainty-based fusion to improve 3D object reconstruction accuracy.*

## 1. Introduction

Identifying and modelling 3D objects in the scene is an important step towards semantic scene understanding [29]. Accurate object representations are key elements for down stream tasks such as object detection, segmentation, tracking, manipulation and dynamic change detection. However, reconstructing detailed 3D object shapes from limited image data remains challenging [5, 9, 50, 49, 13, 45, 12, 37, 24]. Man-made objects have highly variable shapes. Monocular observations can be degraded by issues such as occlusion, noise, truncation, lack of depth measurements, which makes the reconstruction task a ill-posed problem. Some form of prior knowledge of object shape is needed,

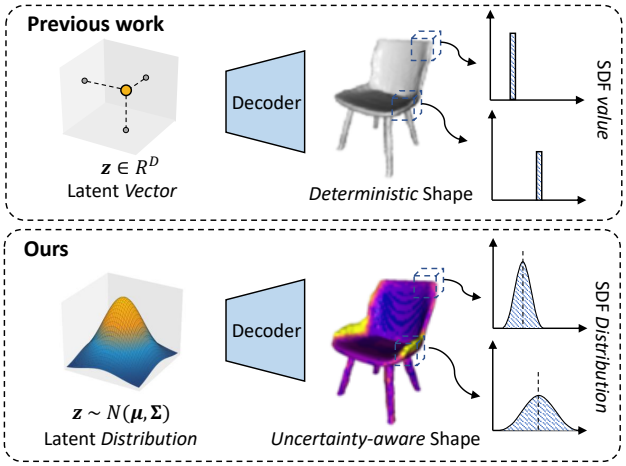


Figure 1. The proposed uncertainty-aware 3D object representation. While the previous work decodes a deterministic shape from a latent vector, Ours decodes a shape with uncertainty from a latent distribution. We can reconstruct 3D objects with uncertainty attached to each part from monocular or multi-view images.

and has been shown to significantly improve reconstruction performance from single [24] and multiple viewpoints [37].

This paper aims to propose a robust 3D object reconstruction method based on monocular images. To address the problems mentioned above, we concentrate on efficiently and robustly combining object priors through fusion of multi-view observations. In particular, we propose an uncertainty-aware fusion framework that refines 3D object representations as new viewpoints are observed, and exploits large object shape datasets to maximize prior knowledge of object geometries.

Recently, neural implicit representations [28] have presented remarkable achievements in multiple areas, including object reconstruction [37], object SLAM [47, 43], and scene reconstruction [42, 56, 55]. Neural object representations, e.g., DeepSDF [34], OccupancyNet [27], show the ability for interpolation, partial completion of 3d shapes and reconstruction from point cloud or monocular images

[27, 37]. The representation can be trained on object shape datasets to learn a prior distribution of a specific class of objects.

We take a step further to couple uncertainty into the neural object representation as in Fig. 1. Uncertainty modelling is highly critical for AI safety and robotics. For example, autonomous driving vehicles need to make safe and reliable decisions based on incomplete or noisy data. The uncertainty accumulated in the perception system can be propagated to the down stream tasks to help applications like localization, tracking, motion planning, and make system maximize the usage of the multi-view observations with robustness to corrupted observations. We discuss more uncertainty applications in the Supplementary Materials.

Uncertainty modelling in deep learning is a widely studied area. Methods including Bayesian Neural Networks [26], sampling approaches (e.g., MC Dropout [11], Ensembles [21]) and direct methods [18], have been proposed and used in real applications such as object detection [10] and semantic segmentation [17]. As far as we know, this paper is the first time to estimate uncertainty for neural object representation from monocular images.

To couple uncertainty, we propose a framework that can propagate uncertainty from image space, to latent space, and finally to 3D object shape, as in Fig. 2. Specifically, we propose a way to teach the encoder to produce a code uncertainty that leads to the right model uncertainties from single images. Then, we propose a method to propagate the uncertainty through the decoder to the SDF and onto the mesh. We design a two-stage training strategy following the previous work [37]. First, we train the decoder to learn a latent space. Then, holding the decoder fixed, we force the encoder to output the correct code uncertainty. This strategy makes the encoder and decoder loosely coupled, and stores the uncertainty in the latent space, which can in theory generalize to different types of decoders. We summarize our contributions below:

- We propose a 3d object modelling approach that relies on an implicit neural representation and provides both a 3D object reconstruction and an uncertainty measure for each object.
- We propose an image encoder with direct uncertainty modelling to estimate latent codes with uncertainty from a single image.
- We propose an incremental fusion method that relies on Bayesian inference to fuse multi-view observations in the latent space to improve reconstruction accuracy and reduce spatial uncertainty.
- We evaluate the system in both synthetic and real datasets, demonstrating the benefit of fusing object models produced from different views through Bayesian inference on the encoded representation.

## 2. Related Work

### 2.1. 3D Object Representations and Reconstruction

Common 3D object representations include meshes [12], voxels [24], octrees [54], TSDFs [34] and point clouds [9], which are all flexible representations but require heavy storage and computation. Each element in the representation is discrete and independent, thus it remains difficult to reconstruct detailed shape from partial observations. There are also compact representations using geometric primitives such as cuboids and quadrics [53, 31, 23], which are significantly more computation efficient but only provide limited information for localization and insufficient information for collision detection and manipulation.

Detection and reconstruction methods have been proposed that generate dense reconstructions in the form of 3D cuboids [19] or meshes [12] directly from single images, without reliance on prior knowledge of object shapes. To make these systems more robust, researchers have also proposed fusing multiple observations from different viewpoints [30], and coupling semantic information as priors into object reconstruction pipelines [38, 47]. Multi-view reconstruction methods [48] generate 3D models from multiple frame, are also called structure from motion (SfM). Geometric priors have also been widely used for different objects and scenarios, e.g., shape prior [47, 43], size prior [53, 33, 23]. PointFlow [52] uses normalization flow to learn a prior distribution of point clouds and reconstructs shapes from partial points. It remains an open question to design an object representation with prior knowledge for detailed shape reconstruction from images and yet can generalize to many objects of different shapes.

### 2.2. Neural Implicit Representation

Recently, neural implicit representations have attracted wide attention in image and scene rendering [28, 32], voice encoding [41], 3D objects [34, 27] and scenes representations [42]. For object representation, DeepSDF [34] proposes to use a neural network to approximate a continual signed distance function for modelling both known and unknown objects, which are captured via interpolation and completion of partial observation. Similar ideas are used in object reconstruction [37, 8], object-level SLAM [43, 47] and multiple object tracking [22]. For example, Duggal et al. [8] reconstruct cars from single-frame lidar points and optionally an image, but without uncertainty quantification and multi-view fusion. Besides object-level details, researchers have further proved the effectiveness of neural implicit representations in representing large scenes, e.g., NeRF [28], visual SLAM [42, 56, 55] and scene reconstruc-

tion [44].

Implicit representation in 3D currently presents many open problems to address, such as effective neural architectures, multi-view fusion methods and uncertainty representations. As described above, this work focuses on identifying an effective fusion method and providing accurate uncertainty measures for downstream tasks.

### 2.3. Uncertainty Modelling in Deep Learning

Modelling uncertainty in deep learning inference has been well studied in the area of Bayesian Neural Networks [26]. Common uncertainty modelling techniques include sampling methods such as MC Dropout and Deep Ensembles, Error Propagation and Direct Modelling [10]. MC Dropout [11] and Deep Ensembles [21] need to run the network multiple times to produce samples from which to estimate uncertainty. Directly Modelling [18] can output uncertainty from a single forward pass and is much more efficient, so we use it to estimate the uncertainty in our work. Error Propagation [36] can also be run efficiently at inference time but requires complex modification of network layers which can affect network performance adversely, so we leave it as future work.

Direct modelling faces the problem of inaccurate and uncalibrated uncertainty in classification and regression [20]. Several methods are proposed to evaluate the output calibration, including calibration plot [14], and proper scoring rules [15] such as Energy Score and Negative Log Likelihood. A recalibration method [14] has been proposed to rectify the calibration via temperature scaling. We will give a detailed analysis with proper scoring rules, and a calibration plot for our uncertainty output.

Very limited work exists for considering uncertainty in neural implicit representation. Researchers [6, 7, 35] have investigated learning a distribution of different topology shapes by changing a low dimensional hyperspace and can model the correspondences between shapes. Deng et al. [6] models the correspondence uncertainty inside a shape category, instead of the reconstruction uncertainty from image observations, e.g., occlusion and ambiguity. It is also non-trivial to propagate the uncertainty for multi-view fusion. Ours aims to derive from Bayesian formulation and output well-calibrated uncertainty. Most related to ours is [39] which models uncertainty in the color and density output of a scene-level neural representation. However, we concentrate on the problem of 3D object reconstruction and multi-view fusion. As far as we know, we are the first to estimate uncertainty for neural object representation from monocular images.

## 3. Methods

### 3.1. Framework Overview

The system framework is shown in Fig 2. The inputs are monocular image sequences of an object taken from different viewpoints. For each input image, the system outputs a reconstructed 3D object shape with uncertainty. The system can fuse multi-view observations in an uncertainty-aware way to incrementally update the shape.

The system consists of an uncertainty-aware neural object representation, and an uncertainty-aware Image Encoder. The neural object representation learns an object shape prior in a latent code space. It has a decoder to generate Signed Distance Function (SDF) values conditioned on each latent code. Then, the Marching Cubes algorithm [25] is used to generate a mesh from the SDF values, with uncertainty represented as an isotropic variance attached to its vertices.

The uncertainty-aware Image Encoder takes in monocular images and outputs latent codes with uncertainty. In this work, we consider a diagonal covariance matrix for all the dimensions of the latent space. When there are multiple images, the multi-view fusion module fuses each output through a Bayesian update rule to estimate both the mean and covariance of the latent code. We now proceed with a more detailed formulation of our approach.

### 3.2. Uncertainty-aware Neural Object Model

Building on DeepSDF [34], we propose to expand the current decoder-based neural object representation to model uncertainty. It is worth mentioning that the proposed uncertainty modelling and fusion method is generalizable to other similar neural representations with limited modification.

**3D object shape modelling with a neural network**. A neural network $f_\theta$ can be trained as a function to map any 3D coordinate, $\mathbf{X} = [x, y, z] \in \mathbb{R}^3$, to its SDF value of $s \in \mathbb{R}$:

$$s = f_\theta(\mathbf{X}) \tag{1}$$

where $\theta$ are the network parameters. Given a 3D grid of SDF values, the Marching Cubes algorithm can then generate a mesh. We can model a 3D shape with each parameter $\theta$. To model a specific class of objects, e.g. chairs or tables, we make the network conditional on a $D$-dimensional latent code, $\mathbf{z} \in \mathbb{R}^D$:

$$s = f_\theta(\mathbf{X}, \mathbf{z}), \mathbf{z} \in \mathbb{R}^D \tag{2}$$

By varying $\mathbf{z}$, the SDF function will also change, as well as the 3D reconstruction it produces. In this manner, a single decoder network can be trained to express the SDF representations of multiple semantically and geometrically similar objects, based on a latent code associated with each training object instance.

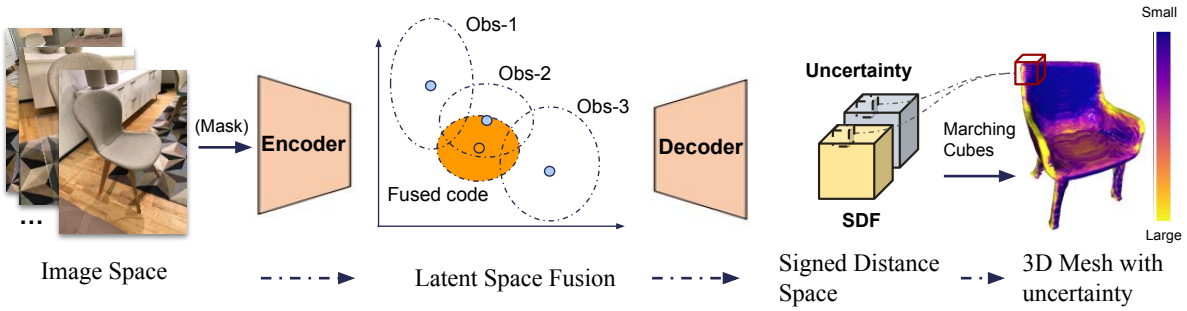


Figure 2. Proposed System Framework. It consists of an uncertainty-aware image Encoder and a pre-trained decoder. We fuse multi-view observations in the latent space under a Bayesian framework. The decoder takes the fused latent space encoding and generates SDF values and associated uncertainties. Finally, the Marching Cubes algorithm is used to generate a mesh from the SDF values with uncertainty at each vertex. We visualize the relative uncertainty values with a color bar inside each models in this paper.

**Modelling uncertainty into 3D object shape.** In Eq. 2, the code $\mathbf{z}$ is deterministic. To model uncertainty, we model the $D$-dimensional latent code $\mathbf{z}$ as a probabilistic variable obeying a multivariate Gaussian distribution $\mathbf{z} \sim \mathcal{N}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. To simplify the problem, we assume each dimension of $\mathbf{z}$ is independent, which leads to a diagonal covariance matrix $\boldsymbol{\Sigma}$. We will train a neural network to output the mean and variance for each dimension of $\mathbf{z}$.

We also model the SDF value at $\mathbf{X}$ as a random variable, $s \sim \mathcal{N}(\mu_s, \sigma_s^2)$. According to Eq. 2, we can propagate the code uncertainty in $\mathbf{z}$ to the SDF value through the decoder network. Since the neural network $f_\theta$ is nonlinear, we can not directly solve for $\sigma_s^2$, and must employ some form or approximation to propagate the uncertainty from code input to SDF output.

**Uncertainty propagation through neural network**. We use Monto Carlo Sampling [16] to propagate the uncertainty through the nonlinear network. First, we sample $M$ codes $Z = \{\mathbf{z}_m\}_{m=1}^M$ from the code distribution $\mathbf{z} \sim \mathcal{N}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For a 3D point $\mathbf{X}$, to get the variance $\sigma_s^2$ of its SDF, we pass each code $\mathbf{z}_m \in Z$ through Eq. 2 to get $s_m$. We then calculate the sample variance [2] from the $M$ SDF values:

$$\sigma_s^2 = \frac{1}{M-1} \sum_{m=1}^{M} (s_m - s_\mu)^2 \tag{3}$$

where $s_\mu = \frac{1}{M} \sum s_m$ is the sample mean. We then calculate the SDF uncertainty for each of the vertices of the mesh generated using Marching Cubes. Now we can use the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of the latent code distribution $\mathcal{N}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to represent a 3D object shape and its uncertainty. The remaining question is how to estimate the mean $\boldsymbol{\mu}$ *and* the covariance $\boldsymbol{\Sigma}$ from input images.

### 3.3. Uncertainty-aware Image Encoder

We propose training a simple encoder network $f_\beta$ to map an RGB image $\mathbf{m} \in \mathbb{R}^{H \times M \times 3}$ with height $H$ and width $W$ to a D-dimensional latent code $\mathbf{z}$ with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$. Since we assume each code dimension is independent, the covariance matrix is diagonal and can be represented as $\boldsymbol{\Sigma} = diag(\boldsymbol{\sigma}^2)$, where $\boldsymbol{\sigma} \in \mathbb{R}^D$.

$$\boldsymbol{\mu}, \boldsymbol{\sigma} = f_\beta(\mathbf{m}) \tag{4}$$

We use the Direct Modelling [18] approach to output uncertainty, which is well-established and does not add computational complexity. We leave the comparison of other uncertainty modelling methods as future work. The Encoder consists of a feature backbone, ResNet-50, and an output layer for the mean and variance. The architecture is straight forward and we concentrate on the choice of proper losses [15] to generate calibrated and accurate uncertainty. We consider two common losses, Negative Log-Likelihood loss (NLL) and Energy Score. We conduct extensive experiments to explore the effectiveness compared with the baseline model trained without uncertainty. We will briefly introduce the two losses below. Their advantages and applications in object detection have been discussed in [15].

**NLL loss**. The NLL loss can be viewed as a standard $L_2$ loss weighted by uncertainty. Considering a batch of outputs $\{(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)\}_{i=1}^N$ directly from the encoder with $N$ data samples, and the ground-truth codes $\{\mathbf{z}_i\}_{i=1}^N$, NLL can be written as:

$$\text{NLL} = \frac{1}{2N} \sum_{i=1}^{N} (\boldsymbol{\mu}_i - \mathbf{z}_i)^\intercal \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\mu}_i - \mathbf{z}_i) + \log \det \boldsymbol{\Sigma}_i \tag{5}$$

where $\boldsymbol{\Sigma}_i = diag(\boldsymbol{\sigma}_i^2) \in \mathbb{R}^{D \times D}$ and $\boldsymbol{\sigma}_i \in \mathbb{R}^D$. The first term pushes down the error, where the variance, $\boldsymbol{\Sigma}_i$, acts to reduce the weight of samples in high uncertainty areas. The second, regularization term avoids uncertainty from growing too large.

**Energy Score**. Energy Score (ES) can be generalized to any distribution. It concentrates on optimizing the result of

high uncertainty data samples to improve performance during training. For computation efficiency, we use a Monte-Carlo approximation version [15], which is represented as:

$$\mathrm{ES} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{M}\sum_{m=1}^{M}\|\mathbf{z}_{i,m}-\mathbf{z}_i\| - \frac{1}{2(M-1)}\sum_{m=1}^{M-1}\|\mathbf{z}_{i,m}-\mathbf{z}_{i,m+1}\|\right) \tag{6}$$

where $\mathbf{z}_{i,m}$ is the $m^{th}$ i.i.d sample from $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. We take $M = 1000$ with very little computational overhead.

### 3.4. Multi-view Bayesian Fusion in Latent Space

**Bayesian Fusion**. Consider $N$ input images and the corresponding outputs $\{(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)\}_{i=1}^{N}$ from the encoder. Since we assume each of the $D$ code dimensions is independent, the covariance matrix is diagonal $\boldsymbol{\Sigma}_i = diag(\boldsymbol{\sigma}_i^2)$. We follow Gaussian Inference [1] to get a fused latent code $\mathbf{z}$. It follows the Gaussian distribution $\mathbf{z} \sim \mathcal{N}_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where:

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\sum_{i=1}^{N}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i, \quad \boldsymbol{\Sigma} = \left(\sum_{i=1}^{N}\boldsymbol{\Sigma}_i^{-1}\right)^{-1} \tag{7}$$

Then, we can use Monto Carlo Sampling to propagate $\mathbf{z}$ through the decoder to get the mean and variance of the SDF value for each 3D point as described in Sec. 3.2.

**Outlier rejection**. When facing extreme situations such as highly occluded objects, experimentation revealed that performance improves by treating them as outliers and filter them out of the fusion process, instead of incorporating them with high uncertainty. We define a modified inference strategy, "Bayesian-$N$", which only selects the $N$ observations with the lowest uncertainty for Bayesian fusion. When $N = 1$, we simply select the lowest uncertainty viewpoint. When $N = N_{max}$, we use all available measurements without rejection, referred to as "Bayesian" by omitting $N$.

## 4. Experiments

### 4.1. Implementation and Training Details

Our system consists of an encoder and a decoder. For the decoder, we follow the implementation and training of DeepSDF [34] on ShapeNet [3]. For the encoder, we use ResNet-50 pretrained on ImageNet as the feature backbone, modify the output layer to the dimensions of the code $N$, and further add $K$ dimensions for the uncertainty. We take $N = K = 64$ in the experiments.

We need monocular images and ground-truth latent codes to train the encoder. We use the images from ShapNetRendering dataset [5] which contains rendered images of 24 different views from the CAD models in ShapeNet [3]. After training the decoder, we get optimized latent codes for each CAD models, and we use them as the ground-truth latent codes for the training and evaluation of the Encoder.

For training the Encoder, we use the same dataset split as FroDO [37]. We augmented the training data with random resize and horizontal flip, and random background clip from SUN dataset [51]. We set a learning rate of 0.1, a batchsize

| Methods | **Shape** | **IoU** ↑ | **EMD** ↓ | **CD** ↓ |
|---|---|---|---|---|
| 3D-R2N2 [5] | Voxel | 0.136 | 0.211 | 0.239 |
| PSGN [9] | Points | N/A | 0.216 | 0.2 |
| 3D-VAE-GAN [50] | Voxel | 0.171 | 0.176 | 0.182 |
| DRC [46] | Voxel | 0.265 | 0.144 | 0.16 |
| MarrNet [49] | Voxel | 0.231 | 0.136 | 0.144 |
| AtlasNet [13] | Mesh | N/A | 0.128 | 0.125 |
| Sun et al. [45] | Voxel | 0.282 | 0.118 | 0.119 |
| FroDO [37] | Mesh | 0.302 | 0.112 | 0.103 |
| FroDO* w/ GT Mask | Mesh | 0.319 | 0.107 | 0.109 |
| FroDO* w/ Seg Mask | Mesh | 0.285 | 0.120 | 0.121 |
| FroDO* w/o Mask | Mesh | 0.257 | 0.127 | 0.123 |
| Ours w/ GT Mask | Mesh | **0.335** | **0.102** | **0.102** |
| Ours w/ Seg Mask | Mesh | 0.293 | 0.116 | 0.116 |
| Ours w/o Mask | Mesh | 0.268 | 0.122 | 0.118 |

Table 1. Single-view reconstruction of the chairs category on the Pix3D dataset. Metrics include Intersection of Union (IoU), Earth Moved Distance (EMD) and Chamfer Distance (CD). GT: Groundtruth. Seg: Semantic Segmentation algorithm. * Our own implementation.

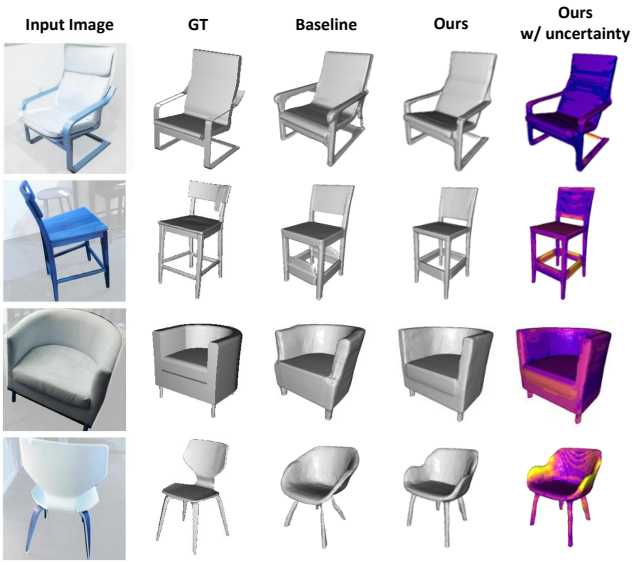


Figure 3. Qualitative results of single-view reconstruction on chairs of Pix3D dataset. The models are trained on ShapeNet dataset. Ours has fewer artifacts than the baseline FroDO*, and can further output uncertainty for each object part. Our uncertainty highlights the areas with errors, indicating information for downstream tasks.

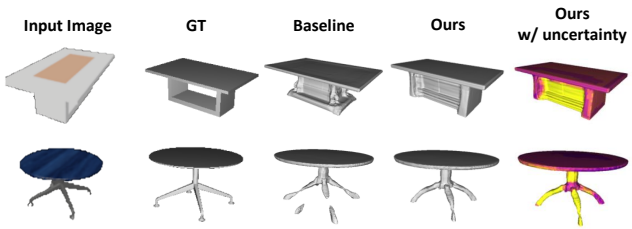


Figure 4. Qualitative results of single-view reconstruction of tables on ShapeNet dataset. The models are both trained and tested on ShapeNet dataset. Compared with the baseline Frodo*, Ours further outputs uncertainty and has fewer artifacts.

of 64, and a random seed of 1000. We use a polynomial learning rate scheduler, and trained for 50 epochs.

To verify that our model can generalize to different categories, we test on both chairs and tables categories on ShapeNet dataset. To verify the performance in real scenarios, after training on the sythetic Shapenet dataset, we directly evaluate on the Pix3D dataset without finetunning.

### 4.2. Metrics and Baselines

**Metrics**. For the reconstruction, we calculate the metrics of Intersection over Union (IoU), Chamfer Distance (CD), and Earth Moved Distance (EMD) on the voxelized mesh with a resolution of $32^3$ following [45]. For the uncertainty, we use Negative Log Likelihood (NLL) and Energy Score (ES), which evaluate the error of the regression, and the calibration and sharpness of the estimated uncertainty.

**Baselines**. We train our model with Energy Score and denote it as Ours. We also compare the choice of the two training losses, Energy Score and NLL in the ablation study.

FroDO [37] is a baseline closest to ours with an encoder trained with L2 loss and a DeepSDF decoder for reconstruction but without uncertainty. It averagely fuses multiple latent codes to get the final reconstruction. It also supports pose estimation and optimization with both shape and pose together. Since pose estimation is out of the scope of the paper, we compare with the Encoder parts to investigate the effectiveness of uncertainty. Note that the results on Pix3D dataset of the origin paper do not use the pose module so it is a fair comparason. Since FroDO is not open-sourced, to fully investigate the performance, we implemented it by ourselves and denote it as FroDO*. We compare our implemented version with origin published version on Pix3D dataset. We also fully compare our models with other published state-of-the-art models for the reconstruction accuracy on Pix3D dataset.

### 4.3. Single-view reconstruction

We show the results on the chairs category of Pix3D dataset in Fig. 3 and Table 1. When using the origin RGB image as input (see Ours w/o Mask), Ours outperforms the baselines PSGN, DRC, AtlasNet and our implementation of Frodo, but still has a gap with the published state-of-the-art performance from Frodo. We trained the Encoder on the synthetic dataset ShapeNet and inferred on the real Pix3D dataset. The domain gap of the real texture and background limits the performance of the encoder, which is only a vanilla ResNet originally designed for classification.

We further use Mask2Former, an off-the-shelf semantic segmentation method [4], to filter the background (see Ours w/ Seg Mask), and notice an obvious improvement on the reconstruction accuracy. When using groundtruth masks to filter all the background (see Ours w/ GT Mask), ours achieves an IoU of 0.335 and outperforms all the baselines, which demonstrates the accuracy upper bound of our design of an image encoder and a shape decoder. It demonstrates that the decoder learns a powerful category-level

| | Views | Methods | IoU | % |
|---|---|---|---|---|
| FroDO* | 1 | Single-view | 0.3225 | 0 |
| | 10 | Average | 0.3456 | 0 |
| Ours | 1 | Single-view | 0.3373 | 4.6 |
| | 10 | Average | 0.3750 | 8.5 |
| | | Bayesian-1 | 0.3719 | 7.6 |
| | | Bayesian-2 | 0.3828 | 10.8 |
| | | Bayesian-3 | 0.3874 | 12.1 |
| | | Bayesian-4 | **0.3902** | **12.9** |
| | | Bayesian-5 | 0.3867 | 11.9 |
| | | Bayesian-6 | 0.3867 | 11.9 |
| | | Bayesian-7 | 0.3836 | 11.0 |
| | | Bayesian-8 | 0.3828 | 10.8 |
| | | Bayesian-9 | 0.3822 | 10.6 |
| | | Bayesian(-10) | 0.3816 | 10.4 |

Table 2. Multi-view fusion IOU performance on the chair category of the Pix3D-MV dataset. Ours with uncertainty can higher IOU than the deterministic baseline. % denotes percent improvement over baseline single/multiview.

| | Min Scale | | | | |
|---|---|---|---|---|---|
| **Methods** | **1.0** | **0.8** | **0.4** | **0.2** | **0.1** |
| FroDO* | 0.346 | 0.343 | 0.338 | 0.323 | 0.318 |
| NLL - Bayesian | 0.327 | 0.326 | 0.318 | 0.302 | 0.301 |
| NLL - Bayesian-4 | 0.346 | 0.349 | 0.335 | 0.330 | 0.341 |
| Ours - Bayesian | 0.382 | 0.362 | 0.345 | 0.332 | 0.322 |
| Ours - Bayesian-4 | **0.390** | **0.381** | **0.374** | **0.369** | **0.365** |

Table 3. Multi-view reconstruction (IoU) on the Pix3D-MV chair set when the input images are cropped to a randomly selected area between [Min Scale, 1.0]. As Min Scale decreases, the fusion task becomes more difficult. Ours with uncertainty gets better robustness in difficult tasks.

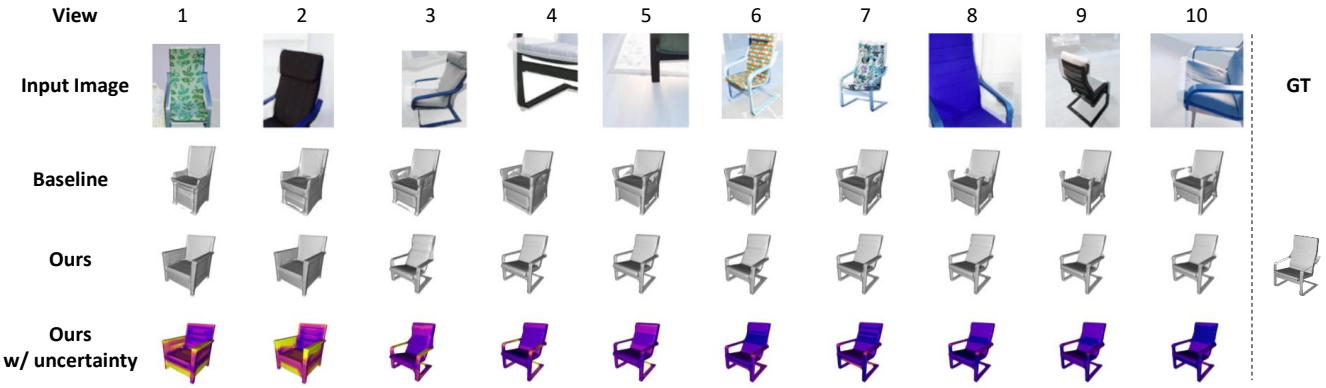


Figure 5. Qualitative results of multi-view fusion on the Pix3D dataset. After fusing 1 to 10 corrupted observation images, Ours reduces the shape uncertainty based on Bayesian Fusion and outputs a more accurate final reconstruction compared with the baseline FroDO*. Note that our method can fuse observations without knowing the camera poses and can work even when the objects have different textures.

prior distribution of the objects shape even from the synthetic dataset, and can transfer to real objects.

Further training and fine-tuning with background augmentation, e.g., randomly sampled texture images, has been shown to make the network adapt to varied backgrounds [40]. However, the evaluation of the background invariance is out of the scope of this paper.

We concentrate on the performance improvement brought by the uncertainty both in single-view and multi-view scenarios. We implemented the Encoder in FroDO [37] which outputs deterministic latent code as baseline, and keep the same decoder. We show the performance when using different masks (w/o Mask, w/ Seg Mask and w/GT mask). Ours gets better performance for all the mask types, thanks to the introduction of uncertainty during training which makes the network more robust to the domain gap so that it does better on difficult reconstruction tasks.

Our implementation (FroDO w/o Mask) has a moderate gap compared with the published result of FroDO. However, when using GT mask (FroDO w/ GT Mask), it achieves higher performance than the published result, which indicates the validity of the accuracy upper bound. Despite extensive experiments in data augmentation and training methods, we were unable to recreate the published performance of Frodo, and the authors of that work have not released their code. In the following experiments, we highlight the effectiveness of uncertainty during multi-view fusion by comparing our work to our own implementation of FroDO*.

### 4.4. Multi-view Reconstruction

The Pix3D dataset contains real images and groundtruth CAD models but has no splits for instances and their multi-view observations. To evaluate the multi-view performance, we group the images of the chair category into separate instances according to their GT models, and keep 10 views as one instance, which results in a multi-view dataset with totally 1490 images from 149 instances. We denote this multi-view dataset as Pix3D-MV which is a subset of the original Pix3D dataset. We show the results of multi-view fusion on Pix3D-MV chair set in Fig. 5 and Table 2. We consider the following methods as multi-view fusion baselines: *Average* equally fuses each estimated latent code; *Bayesian* fuses with uncertainty according to Bayesian Fusion in Equation 7; *Bayesian-$K$* keeps the top-$K$ observations with the lowest uncertainty evaluated by taking the trace of the covariance matrix, and then fuses with *Bayesian*.

Compared with the deterministic baseline, Ours with uncertainty achieves an IoU of 0.3816 vs. 0.3456 with a margin of 10.4%. When using Bayesian-4 to filter the outliers and keep the first 4 observations, ours can further boost up to an IoU of 0.3902 with a margin of 12.9% compared with the baseline. The experiment demonstrates that uncertainty can effectively help the system to trust the observations that contain more valid information, and make the system more robust to outliers in the multi-view observations.

We further push the limit of the robustness brought by the uncertainty during multi-view fusion in Table 3. In real applications like robots, the input images are heavily corrupted because of occlusions, errors in segmentation or sensor noise. We simulate challenging situations by randomly cropping images into a specific size range [$c$, 1.0], so that only part of the origin image is kept. By changing the value of the min scale, $c$, we vary the difficulty of the experiments. As is visible in Table 3, when the min scale decreases, the task becomes more difficult. In the most difficult task, where images can be cropped to only 10% of the origin images, the deterministic FroDO model suffers from the occlusions obviously and decreases to an IoU of only 0.318 while Ours w/ uncertainty remains robust to the cropping and maintains an IoU of 0.366, with an improvement of 15.1%. The experiments prove the effectiveness of using

| Methods | Views | Chairs | | | Tables | | |
|---|---|---|---|---|---|---|---|
| | | Easy | Mid | Hard | Easy | Mid | Hard |
| FroDO* | 1 | **0.388** | 0.341 | 0.326 | 0.403 | 0.321 | 0.300 |
| Ours | | 0.383 | **0.343** | 0.326 | **0.410** | 0.321 | 0.300 |
| FroDO* | 10 | **0.410** | 0.375 | 0.359 | 0.446 | 0.376 | 0.350 |
| Ours | | 0.400 | **0.391** | **0.385** | 0.446 | **0.399** | **0.378** |

Table 4. Single- and multi-view reconstruction results of the Chairs and Tables on the ShapeNet dataset. Ours shows obvious improvements when the task becomes hard for multi-view fusion, showing the robustness to corruption brought by uncertainty.

uncertainty in multi-view fusion to select valid information from a group of corrupted input images.

### 4.5. Ablation Study

**Loss function.** We compare two options for uncertainty training loss: NLL (NLL) and ES (Ours) in Table 3. Even though training with NLL can improve the performance in difficult tasks, it presents lower accuracy in general than when training with ES.

**Selection of $K$ in Bayesian.** With uncertainty, we can detect outliers and take active actions to deal with them. As in Table 2, when decreasing $K$, the performance increases since the outlier codes are filtered out. The highest IOU performance of 0.39 is achieved with $k = 4$, which has an improvement of 12.9% compared with the baseline. When further decreasing the number, the system has too few observations to fuse and the performance begins to drop. An interesting finding is that, with 1 best views we get better performance than the *Average Equal* of FroDO. In real applications, we have the option of adjusting the parameter $K$ to better suit the data.

### 4.6. Evaluation on ShapeNet

Our network architecture, including the uncertainty framework, is not specifically designed for any categories. If the training data is available, we can support the new categories. We show more experiments results on the Chairs and Tables categories on ShapeNet dataset in Table 4. We also show the results of Tables in Fig. 4. We train each categories separately. During inference, we consider two tasks, *Easy* for taking origin rendered images in Shapenet, and *Mid*/*Hard* for randomly croping the images into a range of [$c$,1.0] ($c = 0.1$ for *Mid* and $c = 0.01$ for *Hard*). For the multi-view fusion method, we use *Bayesian Fusion* for Ours and *Average* for FroDO*. In the Easy task, we get comparable but slightly lower IOU performance on Chairs, but higher IoU performance on Tables. Since our model, with the same architecture, requires a part of the model capacity to regress uncertainty. It is notable that uncertainty is not very helpful in easy tasks where each input image contains enough information for reconstruction. In the Mid and Hard task, the effectiveness of uncertain becomes more pronounced, as Ours gets higher multi-view accuracy than the baseline. Especially on Hard task, we got an IoU of 0.385 v.s. 0.359 on Chairs, and 0.378 v.s. 0.350 on Tables. This result demonstrates that uncertainty can robustly find and fuse the valid data from a set of input data of varying quality.

## 5. Conclusion

We propose an uncertainty-aware 3D object reconstruction framework that can take in both monocular and multi-view images. Based on the neural shape models, we introduce a method to model and estimate uncertainty in latent space and a method to propagate uncertainty into 3D object space, so that we can output 3D object shape with uncertainty awareness. Our proposed method can be trained on a purely synthetic dataset and directly evaluated on real datasets. It achieves higher reconstruction performance than deterministic models, and in particular demonstrates better robustness and accuracy in multi-view fusion when the input image sequences are corrupted.

In future work, we plan to scale up to multi-classes objects reconstruction and uncertainty estimation. Also, it will be interesting to leverage the uncertainty-aware shape model for down stream tasks related to objects such as detection, segmentation, tracking, and object-level SLAM.

## Acknowledgement

The authors would like to thank Jordan Hu for the fruitful discussion on uncertainty.

## References

[1] Timothy D Barfoot. *State estimation for robotics*. Cambridge University Press, 2017.

[2] Tony F Chan, Gene H Golub, and Randall J LeVeque. Algorithms for computing the sample variance: Analysis and recommendations. *The American Statistician*, 37(3):242–247, 1983.

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

[4] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021.

[5] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016.

[6] Yu Deng, Jiaolong Yang, and Xin Tong. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *Proceedings of the IEEE/CVF Conference on Com-*

*puter Vision and Pattern Recognition*, pages 10286–10296, 2021.

[7] Shivam Duggal and Deepak Pathak. Topologically-aware deformation fields for single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1536–1546, 2022.

[8] Shivam Duggal, Zihao Wang, Wei-Chiu Ma, Sivabalan Manivasagam, Justin Liang, Shenlong Wang, and Raquel Urtasun. Mending neural implicit modeling for 3d vehicle reconstruction in the wild. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1900–1909, 2022.

[9] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017.

[10] Di Feng, Ali Harakeh, Steven L Waslander, and Klaus Dietmayer. A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[11] Yarin Gal et al. Uncertainty in deep learning. 2016.

[12] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9785–9795, 2019.

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

[14] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[15] Ali Harakeh and Steven L Waslander. Estimating and evaluating regression predictive uncertainty in deep object detectors. *arXiv preprint arXiv:2101.05036*, 2021.

[16] W Keith Hastings. *Monte Carlo sampling methods using Markov chains and their applications*. Oxford University Press, 1970.

[17] Po-Yu Huang, Wan-Ting Hsu, Chun-Yueh Chiu, Ting-Fan Wu, and Min Sun. Efficient uncertainty estimation for semantic segmentation in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 520–535, 2018.

[18] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[19] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018.

[20] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR, 2018.

[21] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[22] Kejie Li, Hamid Rezatofighi, and Ian Reid. Moltr: Multiple object localization, tracking and reconstruction from monocular rgb videos. *IEEE Robotics and Automation Letters*, 6(2):3341–3348, 2021.

[23] Ziwei Liao, Yutong Hu, Jiadong Zhang, Xianyu Qi, Xiaoyu Zhang, and Wei Wang. SO-SLAM: Semantic object slam with scale proportional and symmetrical texture constraints. *IEEE Robotics and Automation Letters*, 7(2):4008–4015, 2022.

[24] Feng Liu and Xiaoming Liu. Voxel-based 3d detection and reconstruction of multiple objects from a single image. *Advances in Neural Information Processing Systems*, 34:2413–2426, 2021.

[25] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987.

[26] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[27] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019.

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

[29] Muzammal Naseer, Salman Khan, and Fatih Porikli. Indoor scene understanding in 2.5/3d for autonomous agents: A survey. *IEEE access*, 7:1859–1887, 2018.

[30] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. Ieee, 2011.

[31] Lachlan Nicholson, Michael Milford, and Niko Sünderhauf. Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam. *IEEE Robotics and Automation Letters*, 4(1):1–8, 2018.

[32] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021.

[33] Kyel Ok, Katherine Liu, Kris Frey, Jonathan P How, and Nicholas Roy. Robust object-based slam for high-speed autonomous navigation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 669–675. IEEE, 2019.

[34] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019.

[35] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.

[36] Janis Postels, Francesco Ferroni, Huseyin Coskun, Nassir Navab, and Federico Tombari. Sampling-free epistemic uncertainty estimation using approximated variance propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2931–2940, 2019.

[37] Martin Runz, Kejie Li, Meng Tang, Lingni Ma, Chen Kong, Tanner Schmidt, Ian Reid, Lourdes Agapito, Julian Straub, Steven Lovegrove, et al. Frodo: From detections to 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14720–14729, 2020.

[38] Renato F Salas-Moreno, Richard A Newcombe, Hauke Strasdat, Paul HJ Kelly, and Andrew J Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1352–1359, 2013.

[39] Jianxiong Shen, Antonio Agudo, Francesc Moreno-Noguer, and Adria Ruiz. Conditional-flow nerf: Accurate 3d modelling with reliable uncertainty quantification. *arXiv preprint arXiv:2203.10192*, 2022.

[40] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[41] Vincent Sitzmann, Julien NP Martel, Alexander W Bergman, David B Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *arXiv preprint arXiv:2006.09661*, 2020.

[42] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021.

[43] Edgar Sucar, Kentaro Wada, and Andrew Davison. Nodeslam: Neural object descriptors for multi-view shape reconstruction. In *2020 International Conference on 3D Vision (3DV)*, pages 949–958. IEEE, 2020.

[44] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021.

[45] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2974–2983, 2018.

[46] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2626–2634, 2017.

[47] Jingwen Wang, Martin Rünz, and Lourdes Agapito. Dsp-slam: Object oriented slam with deep shape priors. In *2021 International Conference on 3D Vision (3DV)*, pages 1362–1371. IEEE, 2021.

[48] Thomas Whelan, Stefan Leutenegger, R Salas-Moreno, Ben Glocker, and Andrew Davison. Elasticfusion: Dense slam without a pose graph. In *Robotics: Science and Systems*, 2015.

[49] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. *Advances in neural information processing systems*, 30, 2017.

[50] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016.

[51] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[52] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4541–4550, 2019.

[53] Shichao Yang and Sebastian Scherer. Cubeslam: Monocular 3-d object slam. *IEEE Transactions on Robotics*, 35(4):925–938, 2019.

[54] Ming Zeng, Fukai Zhao, Jiaxiang Zheng, and Xinguo Liu. Octree-based fusion for realtime 3d reconstruction. *Graphical Models*, 75(3):126–136, 2013.

[55] Zihan Zhu, Songyou Peng, Viktor Larsson, Zhaopeng Cui, Martin R Oswald, Andreas Geiger, and Marc Pollefeys. Nicer-slam: Neural implicit scene encoding for rgb slam. *arXiv preprint arXiv:2302.03594*, 2023.

[56] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022.