OctNetFusion: Learning Depth Fusion from Data

Gernot Riegler[1] Ali Osman Ulusoy[2] Horst Bischof[1] Andreas Geiger[2,3]
[1]Institute for Computer Graphics and Vision, Graz University of Technology
[2]Autonomous Vision Group, MPI for Intelligent Systems Tübingen
[3]Computer Vision and Geometry Group, ETH Zürich
{riegler, bischof}@icg.tugraz.at {osman.ulusoy,andreas.geiger}@tue.mpg.de

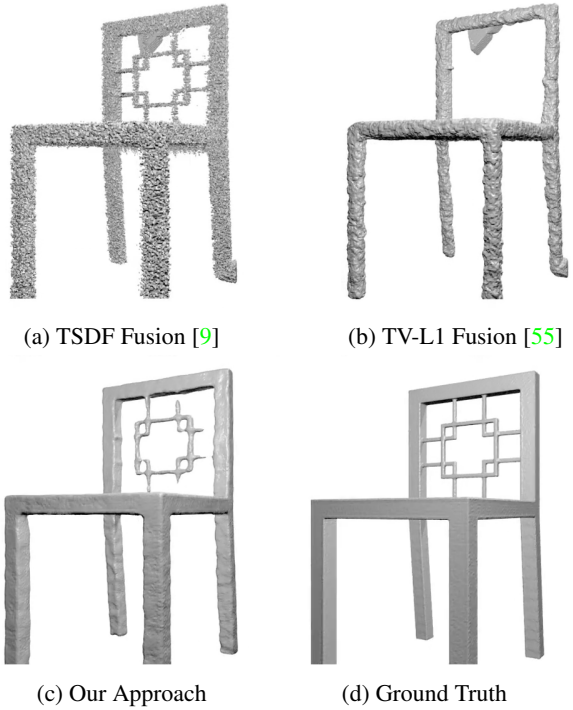
Abstract

In this paper, we present a learning based approach to depth fusion, i.e., dense 3D reconstruction from multiple depth images. The most common approach to depth fusion is based on averaging truncated signed distance functions, which was originally proposed by Curless and Levoy in 1996. While this method is simple and provides great results, it is not able to reconstruct (partially) occluded surfaces and requires a large number frames to filter out sensor noise and outliers. Motivated by the availability of large 3D model repositories and recent advances in deep learning, we present a novel 3D CNN architecture that learns to predict an implicit surface representation from the input depth maps. Our learning based method significantly outperforms the traditional volumetric fusion approach in terms of noise reduction and outlier suppression. By learning the structure of real world 3D objects and scenes, our approach is further able to reconstruct occluded regions and to fill in gaps in the reconstruction. We demonstrate that our learning based approach outperforms both vanilla TSDF fusion as well as TV-L1 fusion on the task of volumetric fusion. Further, we demonstrate state-of-the-art 3D shape completion results.

1. Introduction

Reconstructing accurate and complete 3D surface geometry is a core problem in computer vision. While image based techniques [40, 15, 12, 49, 39] provide compelling results for sufficiently textured surfaces, the introduction of affordable depth sensors, in particular the Microsoft Kinect sensor, allows scanning a wide variety of objects and scenes. This has led to the creation of large databases of real-world 3D content [7, 10, 42, 43], enabling progress in a variety of areas including 3D modeling [28, 54], 3D reconstruction [6, 52], 3D recognition [18, 44] and 3D scene understanding [16, 42].

However, creating complete and accurate 3D models from 2.5D depth images remains a difficult problem. Challenges include sensor noise, quantization artifacts, outliers (e.g., bleeding at object boundaries) and missing data such as occluded surfaces. Some of these challenges can be addressed by integrating depth information from multiple viewpoints in a volumetric representation. In particular, Curless and Levoy demonstrate that averaging truncated signed distance functions (TSDF) allows for a simple yet effective approach to depth fusion [9] which is used in a large number of reconstruction pipelines today [6, 31, 32, 46, 51].



(a) TSDF Fusion [9] (b) TV-L1 Fusion [55]

(c) Our Approach (d) Ground Truth

Figure 1: **Depth fusion results** using four uniformly spaced views around the object. (a) TSDF fusion produces noisy results. (b) TV-L1 fusion smoothes the shape, but also removes thin details. (c) Our approach learns from large 3D repositories and significantly improves fusion results in terms of noise reduction and surface completion.

Despite its popularity, TSDF fusion has two major drawbacks: First, it requires a large number of frames to smooth out sensor noise and outliers. Second, it does not allow for reconstructing occluded regions or completing large holes.

In this paper, we tackle these problems by proposing a learning-based solution to volumetric 3D fusion. By utilizing large datasets of 3D models and high capacity 3D convolutional neural networks (3D CNNs), our approach learns to smooth out sensor noise, deals with outliers and completes missing 3D geometry. To the best of our knowledge, our approach is the first to learn volumetric fusion from noisy depth observations.

3D CNNs [8, 22, 28, 34, 53] are a natural choice for formulating this task as an end-to-end learning problem. However, existing deep learning approaches are limited to small resolutions (typically 32^3 voxel grids) due to the cubic growth in memory requirements. A notable exception is the OctNet approach of Riegler et al. [36] which takes advantage of the sparsity of 3D volumetric models using an octree representation [37], enabling deep learning at resolutions of 256^3 voxels and beyond.

The main limitation of OctNets [36], however, is that the octree representation is derived from the *input* and fixed during learning and inference. While this is sufficient for tasks such as 3D classification or semantic segmentation where the input and the output share the same octree representation, the OctNet framework does not directly apply to tasks where the 3D space partitioning of the output is *unknown* a priori and may be different than that of the input. In particular, for tasks such as depth map fusion and 3D completion the location of the implicit surface is unknown and needs to be inferred from noisy observations.

The key contribution of this paper is to lift this restriction. More specifically, we propose a novel 3D CNN architecture termed *OctNetFusion* which takes as input one or more depth images and estimates both the 3D reconstruction and its supporting 3D space partitioning, i.e. the octree structure of the *output*. We apply this architecture to the depth map fusion problem and formulate the task as the prediction of truncated signed distance fields which can be meshed using standard techniques [27].

We evaluate our approach on synthetic and real-world datasets, studying several different input and output representations, including TSDF. Our experiments demonstrate that the proposed method is able to reduce noise and outliers compared to vanilla TSDF fusion [9, 31] while avoiding the shrinking bias of local regularizers such as TV-L1 [55]. Besides, our model learns to complete missing surfaces and fills in holes in the reconstruction. We demonstrate the flexibility of our model by evaluating it on the task of volumetric shape completion from a single view where we obtain improvements wrt. the state-of-the-art [14]. Our code and data will be made available.

2. Related Work

Volumetric Fusion: In their seminal work, Curless and Levoy [9] proposed to integrate range information across viewpoints by averaging truncated signed distance functions. The simplicity of this method has turned it into a universal approach that is used in many 3D reconstruction pipelines. Using the Microsoft Kinect sensor and GPGPU processing, Newcombe et al. [31] showed that real-time 3D modeling is feasible using this approach. Large-scale 3D reconstruction has been achieved using iterative re-meshing [51] and efficient data structures [32, 46]. The problem of calibration and loop-closure detection has been considered in [6, 57]. Due to the simplicity of the averaging approach, however, these methods typically require a large number of input views, are susceptible to outliers in the input and don't allow to predict surfaces in unobserved regions.

Noise reduction can be achieved using variational techniques which integrate local smoothness assumptions [2, 19, 55] into the formulation. However, those methods are typically slow and can not handle missing data. In this paper, we propose an alternative learning based solution which significantly outperforms vanilla TSDF fusion [9, 31] and TV-L1 fusion [55] in terms of reconstruction accuracy.

Ray Consistency: While TSDF fusion does not explicitly consider free space and visibility constraints, ray potentials allow for modeling these constraints in a Markov random field. Ulusoy et al. [49, 48] consider a fully probabilistic model for image based 3D reconstruction. Liu and Cooper [26] formulate the task as MAP inference in a MRF. In contrast to our method, these algorithms do not learn the geometric structure of objects and scene from data. Instead, they rely on simple hand-crafted priors such as spatial smoothness [26], or piecewise planarity [47]. Notably, Savinov et al. combine ray potentials with 3D shape regularizers that are learned from data [38]. However, their regularizer is *local* and relies on a semantic segmentation as input. In this work, we do not consider the semantic class of the reconstructed object or scene and focus on the generic 3D reconstruction problem using a *global* model.

Shape Completion: If exact 3D models are available, missing surfaces can be completed by detecting the objects and fitting 3D models to the observations [3, 21]. In this paper, we assume that such prior knowledge is not available. Instead we directly learn to predict the 3D structure from training data in an end-to-end fashion.

Shape completion from a single RGB-D image has been tackled in [14, 23, 45, 56]. While [23] use a CRF for inference, [14] predict structured outputs using a random forest and [45] use a CNN to jointly estimate voxel occupancy and semantic class labels. In contrast to our approach, these methods reason at the voxel level and therefore do not pro-

vide sub-voxel surface estimates. Furthermore, existing 3D CNNs [45] are limited in terms of resolution. In this paper, we demonstrate a unified approach which allows to reason about missing 3D structures at large resolution while providing sub-voxel surface estimates. In contrast to single-image reconstruction methods, our approach naturally handles an arbitrary number of input views. For the task of 3D shape completion from a single image we obtain results which are superior to those reported by Firman et al. [14].

In very recent work, Dai et al. [11] consider the problem of high-resolution 3D shape completion. Their approach first regresses 32^3 voxel volumes using a 3D CNN, followed by a multi-resolution 3D shape synthesis step using a large database of 3D CAD models [5]. While their object-centric approach is able to reconstruct details of individual objects with known 3D shape, we put our focus on general 3D scenes where such knowledge is not available.

3. Method

This section introduces our OctNetFusion architecture. Our work builds upon the recent work of 3D octree convolutional networks [36, 50]. As our work specifically extends [36], we follow its notation whenever possible. To make this paper self-contained, we first briefly review OctNet [36] in Section 3.1. Then, we present our OctNetFusion approach in Section 3.2 which learns to jointly estimate the output quantity (e.g., signed distance or occupancy) and the space partitioning. Section 3.3 specifies the feature representations considered in our experiments.

3.1. OctNet

The main limitation of conventional 3D CNNs that operate on regular voxel grids is the cubic growth in memory requirements with respect to the voxel resolution. However, 3D data is often sparse in nature [25]. For instance, the surface of an object can be interpreted as a 2D manifold in 3D space. Riegler et al. [36] utilize this observation and define a CNN on the grid-octree data structure of [29]. The data structure itself consists of a grid of shallow octrees with maximum depth $D = 3$, trading off computation and memory. The structure of the shallow octrees can be efficiently encoded as bit strings that allows for rapid retrieval of neighboring cells. One important property of OctNets is that none of the operations (i.e., convolution, pooling, unpooling) changes the grid-octree data structure which is based on the input (e.g., point cloud, voxel grid). This can be seen as a data-adaptive pooling operation which maps the output of each layer back to the grid-octree representation.

We now introduce the basic notation. Consider a dense voxel grid $T \in \mathbb{R}^{8D \times 8H \times 8W}$ where $T_{i,j,k}$ denotes the value at voxel (i, j, k). Further, let O denote a grid-octree data structure that covers the same volume. Given tree depth of $D = 3$ this data structure would contain $D \times H \times W$ shallow octrees, where each shallow octree covers 8^3 voxels. The important difference to the dense representation is that the cells in O can comprise a variable number of voxels.

Let $\Omega[i, j, k]$ denote the smallest grid-octree cell that contains the voxel at (i, j, k). $\Omega[i, j, k]$ can be interpreted as the set of voxel indices, whose data is pooled to a single value as described above. Furthermore, $|\Omega[i, j, k]|$ denotes the number of voxels comprised by this cell. If the cell is at the finest resolution of the tree, we have $|\Omega[i, j, k]| = 1$, i.e., the cell is equal to the voxel in $T_{i,j,k}$. In contrast, if the complete shallow octree consists of only a single leaf cell, then $|\Omega[i, j, k]| = 512$ as all 8^3 voxels are pooled.

Given this basic notation, the authors of [36] show how the convolution, pooling and unpooling operation can be efficiently implemented on this data structure. We refer to [36] for further details.

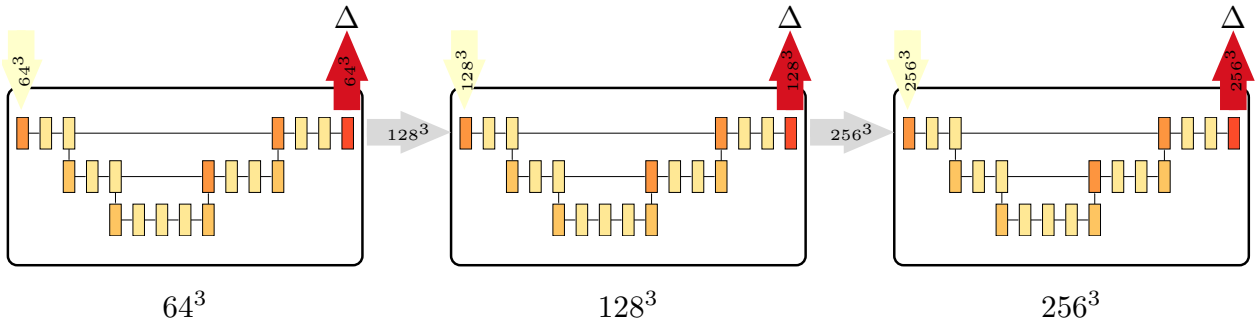
3.2. OctNetFusion

The main drawback of OctNets [36] is that the octree structure of the input and output, i.e. the partitioning of the 3D space, has to be known a priori. This is a reasonable assumption for tasks like 3D point cloud labeling (e.g., semantic segmentation) where the input and the output octree structures are the same. However, for tasks where the output geometry is different from the input geometry, e.g., in volumetric fusion or shape completion, the grid-octree data structure has to be adapted during inference.

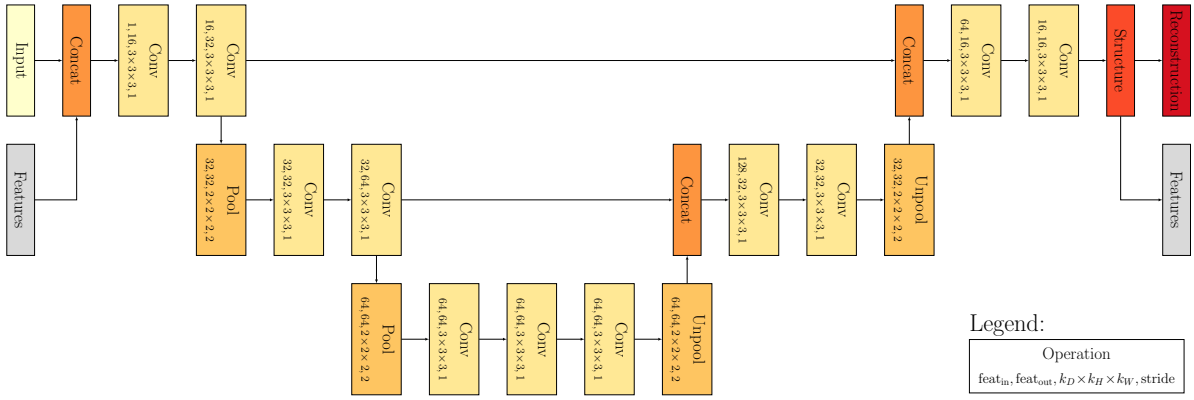
We now present our OctNetFusion architecture, illustrated in Fig. 2, which allows to learn the grid-octree structure along with the 3D task in a principled manner.

Network Architecture: Our overall network architecture is illustrated in Fig. 2a. We represent the voxelized input and output using the grid-octree structure described in Section 3.1. The input to the network is a feature volume (e.g., TSDF), calculated from a single or multiple depth maps, see Section 3.3 for details. The output may encode a TSDF or a binary occupancy map, depending on the application.

As the 3D input to our method can be sparse and incomplete, we refrain from using the classical U-shaped architecture as common for 2D-to-2D prediction tasks [1, 13]. Instead, we propose a coarse-to-fine network with encoder-decoder modules, structure manipulation modules and a loss defined at every pyramid level. More specifically, we create a 3D scale pyramid where the number of voxels along each dimension increases by a factor of two between pyramid levels. At each level, we process the input using an encoder-decoder module which enlarges the receptive field and captures contextual information. We pass the resulting features to a structure manipulation module which computes the output at the respective resolution, increases the resolution and updates the structure of the network for further processing. We propagate features to successively finer resolutions until we have reached the final target resolution.



(a) **OctNetFusion Architecture**



(b) **Encoder-Decoder Module**

Figure 2: **OctNetFusion Architecture.** (a) Overall structure of our coarse-to-fine network. (b) Each encoder-decoder module increases the receptive field size and adds contextual information. The structure module (orange) is detailed in Fig. 3.

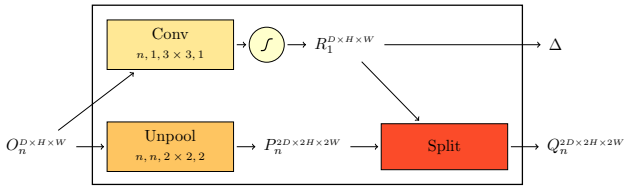


Figure 3: **Structure Module**. The structure manipulation module doubles the spatial resolution of the feature maps. A loss at the end of each pyramid level (Δ) measures the quality of the reconstruction at the respective resolution.

We will now describe the encoder-decoder module and the structure module in detail.

Encoder-Decoder Module: The encoder-decoder module is illustrated in Fig. 2b. It combines convolution layers with pooling and unpooling layers similar to the segmentation network used in [36]. All convolutional layers are followed by a ReLU non-linearity [30]. The convolution layer before each pooling operation doubles the number of feature maps while the convolution layer after each unpooling operation halves the number of features. Pooling operations reduce spatial information but increase the level of context captured in the features. The result of the unpooling operation is concatenated with the corresponding high-resolution features from the encoder path to combine high-resolution information with low-resolution contextual cues.

Structure Module: As discussed above, the unpooling operation of the original OctNet architecture [36] has one major drawback: the octree structure must be known in advance to determine which voxels shall be split. While for 3D point labeling the structure can be split according to the input, the final output structure is unknown for tasks like volumetric fusion or completion. Naïvely splitting all voxels would eliminate the advantage of the data-adaptive representation and limit the output resolution to small volumes.

Consequently, we introduce a *structure module* after each encoder-decoder module which determines for each voxel if it shall be split (i.e., close to the surface) or not (i.e., far from the surface). Our structure module is illustrated in Fig. 3. The main idea is to add a split mask to the standard unpooling operation that indicates which of the octree cells should be further subdivided. This splitting mask is then used to subdivide the unpooled grid-octree structure.

More formally, let us consider an input grid-octree structure O with n feature channels and $D \times H \times W$ shallow octrees as illustrated in Fig. 3. After the unpooling operation we obtain a structure P that consists of $2D \times 2H \times 2W$ shallow octrees where each octree cell comprises eight-times the number of voxels, i.e., $|\Omega_P[2i, 2j, 2k]| = 8|\Omega_O[i, j, k]|$.

To determine the new octree structure, we additionally predict a reconstruction R at the resolution of O using a single convolution followed by a sigmoid non-linearity or a 1×1 convolution depending on the desired output (occupancy or TSDF, respectively). A reconstruction loss (Δ) ensures that the predicted reconstruction is close to the ground truth reconstruction at each resolution of the scale pyramid (for learning the network we provide the ground truth reconstruction at each resolution as described in Section 4).

We define the split mask implicitly by the surface of the reconstruction R. The surface is defined by the gradients of R when predicting occupancies or by the zero-levelset of R in the case of TSDF regression. Given the surface, we split all voxels within distance τ from the surface. For TSDF regression, τ equals the truncation threshold. For occupancy classification, τ is a flexible parameter which can be tuned to trade reconstruction accuracy vs. memory usage. The output of this split operation finally yields the high resolution structure Q which serves as input to the next level.

3.3. Input / Output Encoding

This section describes the input and output encodings for our method. An ablation study, analyzing the individual encodings is provided in Section 4.

3.3.1 Input Encoding

The input to our method are one or more 2.5D depth maps. We now discuss several ways to project this information into 3D voxel space which, represented using grid-octree structures, forms the input to the OctNetFusion architecture described above. The traditional volumetric fusion approach [9] calculates the weighted average TSDF with respect to all depth maps independently for every voxel where the distance to the surface is measured along the line of sight to the sensor. While providing for a simple one-dimensional signal at each voxel, this encoding does not capture all information due to the averaging operation. Thus, we also explore higher dimensional input encodings which might better retain the information present in the sensor recordings. We now formalize all input encodings used during our experiments, starting with the simplest one.

Occupancy Fusion (1D): The first, and simplest encoding fuses information at the occupancy level. Let $d_v^{(i)}$ be the depth of the center of voxel v wrt. camera i. Further, let $d_c^{(i)}$ be the value of the depth map when projecting voxel v onto the image plane of camera i. Denoting the signed distance between the two depth values $\delta_v^{(i)} = d_c^{(i)} - d_v^{(i)}$, we define the occupancy of each voxel $o(v)$ as

$$o(v) = \begin{cases} 1 & \exists i : \delta_v^{(i)} \leq s \wedge \nexists i : \delta_v^{(i)} > s \\ 0 & \text{else} \end{cases} \quad (1)$$

where s is the size of voxel v. The interpretation is as follows: If there exists any depth map in which voxel v is observed as free space the voxel is marked as free, otherwise it is marked as occupied. While simple, this input encoding is susceptible to outliers in the depth maps and doesn't encode uncertainty. Furthermore, the input distance values are not preserved as only occupancy information is encoded.

TSDF Fusion (1D): Our second input encoding is the result of traditional TSDF fusion as described in [9, 31]. More specifically, we project the center of each voxel into every depth map, calculate the TSD value using a truncation threshold τ (corresponding to the size of four voxels in all our experiments), and average the result over all input views. While various weight profiles have been proposed [46], we found that the simple constant profile proposed by Newcombe et al. [31] performs well. This input representation is simple and preserves distances, but it does not encode uncertainty and thus makes it harder to resolve conflicts.

TDF + Occupancy Fusion (2D): The TSDF encoding can also be split into two channels: one channel that encodes the truncated unsigned distance to the surface (TDF) and one that encodes occupancy. Note that, if $t(v)$ is the TDF of voxel v, and $o(v)$ its occupancy, then $-t(v) \cdot o(v)$ is equivalent to the truncated signed distance function of v.

Histogram (10D): While the previous two encodings captures surface distance and uncertainty, they do not maintain the multi-modal nature of fused depth measurements. To capture this information, we propose a histogram-based representation. In particular, we encode all distance values for each voxel using a 10D histogram with 5 bins for negative and 5 bins for positive distance values. The first and the last bin of the histogram capture distance values beyond the truncation limit, while the bins in between collect non-truncated distance values. To allow sub-voxel surface estimation, we choose the histogram size such that a minimum of 2 bins are allocated per voxel. Furthermore, we populate the histogram in a smooth fashion by distributing the vote of each observation linearly between the two closest bins, e.g., we assign half of the mass to both neighboring bins if the prediction is located at their boundary.

3.3.2 Output Encoding and Loss

Finally, we describe the output encodings and the loss we use for the volumetric fusion and the volumetric completion tasks we consider in the experimental evaluation.

Volumetric Fusion: For volumetric fusion, we choose the TSDF as output representation using an appropriate truncation threshold τ. Note that in contrast to binary occupancy, TSDF outputs allow for predicting implicit surfaces with sub-voxel precision. We regress the TSDF values at each resolution (i.e., within the structure module) using a 1×1 convolution layer and use the ℓ_1 loss for training.

Volumetric Completion: For volumetric completion from a single view, we use a binary occupancy representation to match the setup of the baselines as closely as possible. Following common practice, we leverage the binary cross entropy loss for training the network.

4. Evaluation

In this section, we present our experiments and evaluations. In Section 4.1 we consider the task of volumetric fusion from multiple depth images and in Section 4.2 we compare our approach to a state-of-the-art baseline on the task of volumetric completion from a single depth image.

4.1. Volumetric Fusion

In this section we consider the volumetric fusion task. We evaluate our OctNetFusion approach on the synthetic ModelNet40 dataset of Wu et al. [54] as well as on real Kinect object scans that are generated using the depth image dataset by Choi et al. [7].

Unfortunately, the ground truth TSDF can not be calculated directly from the 3D models in ModelNet as the meshes are not watertight, i.e., they contain holes and cracks. Moreover, the meshes typically do not have consistently oriented normals. Instead, we obtain the ground truth TSDF by densely sampling views around the object, rendering the input 3D model from all views and running traditional volumetric fusion on all generated (noise-free) depth maps. We found that 80 views cover all object surfaces and hence allow for computing highly accurate TSDF ground truth. For each of the categories we used 200 models for training and 20 for testing from the provided train/test split, respectively.

Besides the synthetic ModelNet dataset, we also evaluated our approach on real data from a Kinect RGB-D sensor. In particular, we use the 10 videos captured by Choi et al. [7] which include a diverse set of objects such as chairs, tables, trash containers, plants, signs, etc. Unfortunately, the dataset does not include ground-truth 3D models or camera poses. We thus estimated the camera poses using Kintinuous [51] and visually inspect all models to remove those for which the pose tracking failed. Similar to the ModelNet experiments, we leverage TSDF fusion to obtain reference 3D models. However, for this dataset we leverage 1000 viewpoints for each object to average the effect of noise. This is possible as the dataset has been carefully collected with many slow camera motion and many redundant views. At test time we provide only a small fraction of views (10-20) to each algorithm to simulate challenging real-world conditions. Example reference models produced by this procedure are shown in Fig. 4. We augment the dataset by generating 20 different view configurations per scene by selecting different and disjoint subsets of view points at random.

	VolFus [9]	TV-L1 [55]	Occ	TDF + Occ	TSDF	TSDF Hist
64^3	4.136	3.899	2.095	1.987	**1.710**	1.715
128^3	2.058	1.690	0.955	0.961	0.838	**0.736**
256^3	1.020	0.778	0.410	0.408	0.383	**0.337**

Table 1: **Evaluation of Input Encodings** (MAD in mm)

	VolFus [9]	TV-L1 [55]	Ours
1 view	14.919	14.529	**1.927**
2 views	3.929	3.537	**0.616**
4 views	1.020	0.778	**0.337**
6 views	0.842	0.644	**0.360**

Table 2: **Number of Input Views** (MAD in mm)

We train each stage of the network with a constant learning rate of 10^{-4} and Adam [24] as optimizer. We train the first stage for 50 epochs, and the next two stages for 25 epochs each. We initialize the first stage according to the randomization scheme proposed in [20], and initialize the weights of the other stages with those of the previous stage.

Input Encoding: We first investigate the impact of the input encodings discussed in Section 3.3.1 on the quality of the output. Towards this goal, we scaled the ModelNet objects to fit into a cube of $3 \times 3 \times 3$ meters and rendered the depth maps onto 4 equally spaced views sampled from a sphere. To simulate real data, we added depth dependent Gaussian noise to the inputs as proposed in [33].

Our results are shown in Table 1. We compare the traditional volumetric fusion approach of Curless et al. [9] ("VolFus") and the variational approach of Zach et al. [55] ("TV-L1") to our method using the input encodings described in Section 3.3.1. The parameters of the baselines have been chosen based on cross-validation. We evaluate our results in terms mean absolute distance (MAD) which we compute over all voxels in the scene. Each row shows results at a particular resolution, ranging from 64^3 to 256^3 voxels.

First, we observe that our model outperforms the traditional fusion approach [9, 55] as well as TV-L1 fusion [55] by a large margin. Improvements are particularly pronounced at high resolutions which demonstrates that our learning based approach is able to refine and complete geometric details which can not be recovered using existing techniques. Furthermore, we observe that the TSDF histogram based encoding yields the best results. We thus use this input encoding for all experiments that follow.

	$\sigma = 0.00$			$\sigma = 0.01$			$\sigma = 0.02$			$\sigma = 0.03$		
	VolFus [9]	TV-L1 [55]	Ours	VolFus [9]	TV-L1 [55]	Ours	VolFus [9]	TV-L1 [55]	Ours	VolFus [9]	TV-L1 [55]	Ours
64^3	3.020	3.272	**1.647**	3.439	3.454	**1.487**	4.136	3.899	**1.715**	4.852	4.413	**1.938**
128^3	1.330	1.396	**0.744**	1.647	1.543	**0.676**	2.058	1.690	**0.736**	2.420	1.850	**0.804**
256^3	0.621	0.637	**0.319**	0.819	0.697	**0.321**	1.020	0.778	**0.337**	1.188	0.858	**0.402**

Table 3: **Evaluation wrt. Input Noise** (MAD in mm)

		VolFus [9]	TV-L1 [55]	Seen	Unseen
all	64^3	4.136	3.899	1.715	**1.686**
	128^3	2.058	1.690	**0.736**	0.799
	256^3	1.020	0.778	**0.337**	0.358
airplane	64^3	0.668	0.583	**0.419**	0.470
	128^3	0.324	0.297	**0.174**	0.192
	256^3	0.157	0.111	**0.076**	**0.076**
desk	64^3	5.122	4.767	**1.954**	2.000
	128^3	2.540	2.165	**0.777**	0.898
	256^3	1.260	0.987	**0.334**	0.383

Table 4: **Seen vs. Unseen Categories** (MAD in mm)

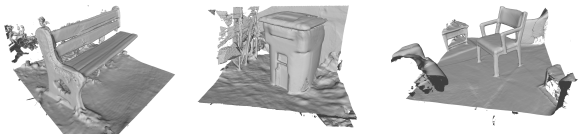


Figure 4: **Examples from Kinect Scan Dataset**

Number of Views: Next, we evaluate the performance of our network when varying the number of input views from one to six on the ModelNet dataset. All experiments are conducted at a resolution of 256^3 voxels. Our results are shown in Table 2. Again, our approach outperforms both baselines in all categories. As expected, performance increases with the number of viewpoints. The largest difference between the baseline and our approach is visible for the experiment with only one input view. While no fusion is performed in this case, this demonstrates that our learned model is effective in completing missing geometry. When considering four input views, our approach reduces errors by a factor of 2 to 3 wrt. TSDF fusion and TV-L1 fusion.

Noise on Input: In our next experiment we evaluate the impact of noise in the depth maps on the reconstruction. Table 3 summarizes our results. We observe that our method is faithful wrt. the increase of input noise. The MAD increases from 0.274 mm for no noise to 0.374 mm for severe noise ($\sigma = 0.03$). In contrast, for TSDF fusion the MAD increases by more than 0.5 mm and for TV-L1 fusion by more than 0.2 mm.

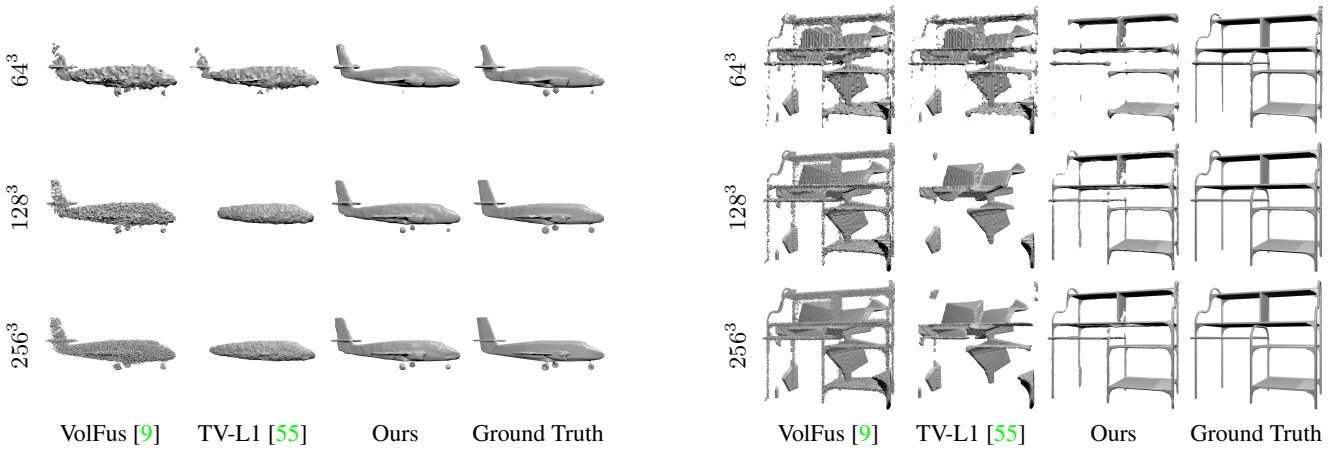
Generalization on Unseen Categories: Most existing approaches that leverage deep learning for 3D reconstruction train a model *specific* for a particular object class [4, 8, 17, 41, 35] which typically does not generalize well to other classes or scenes with varying backgrounds as present in the scans of Choi et al. [7]. In contrast, here we are interested in 3D reconstruction of *general* scenes. We therefore analyze how our model behaves on shapes that where not seen during training. In Table 4 we trained a network on only 8 categories out of 10 and use the two unseen ones for testing ("Unseen"). We compare these results to the case where we train on all 10 categories ("Seen"). Note that in neither case training shapes are part of the test set, but shapes from the same category are used or ignored. While we can observe a slight decrease in performance for the unseen categories, it is still far better than simple TSDF fusion, or TV-L1 fusion.

	views=10			views=20		
	VolFus [9]	TV-L1 [55]	Ours	VolFus [9]	TV-L1 [55]	Ours
64^3	103.855	25.976	**22.540**	72.631	22.081	**18.422**
128^3	58.802	12.839	**11.827**	41.631	11.924	**9.637**
256^3	31.707	5.372	**4.806**	22.555	5.195	**4.110**

Table 5: **Fusion of Kinect Object Scans** (MAD in mm)

Qualitative Results on ModelNet: We show qualitative results on ModelNet in Fig. 5 using the TSDF encoding and 4 views. The same TSDF truncation threshold has been used for traditional fusion, our OctNetFusion approach and the ground truth generation process. While the baseline approach is not able to resolve conflicting TSDF information from different viewpoints, our approach learns to produce a smooth and accurate 3D model from highly noisy input.

Kinect Object Scans: To evaluate the performance of our algorithm on real data, we use the dataset of Kinect object scans published by Choi et al. [7]. For this experiment we vary the number of input views from ten to twenty as the scenes are larger and the camera trajectories are less regular than in the synthetic ModelNet experiments. Our results are shown in Table 5. While overall errors are larger compared to the ModelNet experiments, we again outperform the traditional fusion approach at all resolutions and for all number of input views. Notably, the relative difference between the two methods increases with finer resolution which demonstrates the impact of learned representations for reconstructing details. In contrast to the experiments on ModelNet, the TV-L1 baseline [55] reduces errors more substantially when compared to vanilla volumetric fusion [9], yet our method is consistently more accurate.



(a) Airplane (b) Desk

Figure 5: **Qualitative Results on ModelNet** using 4 equally spaced input viewpoints

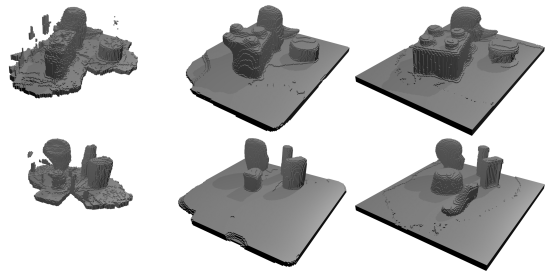
Runtime: Given its simplicity, Vanilla TSDF fusion [9] is the fastest method, using just a few milliseconds on a GPGPU. In contrast, TV-L1 fusion [55] is computationally more expensive. We ran all experiments using 700 iterations of the optimization algorithm. For an output resolution of 64^3 TV-L1 needs 0.58 seconds on average and for an output resolution of 256^3 it needs 24.66 seconds on average. In comparison, our proposed OctNetFusion CNN requires 0.005 seconds on average for an output resolution of 64^3 and 10.1 seconds on average for an output resolution of 256^3. All numbers were obtained on a NVidia K80 GPGPU.

4.2. Volumetric Completion

In this section, we provide a comparison to Firman's Voxlets approach et al. [14] on the task of volumetric shape completion from a single image. For this experiment, we use the dataset and metrics proposed by [14] and modify our model to predict binary occupancy maps instead of real-valued TSDFs. Our results are shown in Table 6. Qualitative results for three different scenes are visualized in Fig. 6. As evidenced by our results, our approach improves upon Voxlets [14] as well as the method of Zheng et al. [56] in terms of intersection-over-union (IoU) of the occupied space, precision and recall. Unfortunately, even after communication with the authors we were not able to reproduce their results due to post-publication changes in their dataset.

Method	IoU	Precision	Recall
Zheng et al. [56]*	0.528	0.773	0.630
Voxlets et al. [14]*	0.585	0.793	0.658
Voxlets et al. [14]	0.550	0.734	0.705
Ours	**0.650**	**0.834**	**0.756**

Table 6: **Volumetric Completion on Tabletop Dataset.** The numbers marked with asterisk (*) are taken from [14] while the others have been recomputed by us and verified by the authors of [14]. Unfortunately, even in a joint effort with the authors, we were not able to reproduce the original results due to irreversible post-publication changes in their dataset, thus we provide both results in this table.



(a) Firman [14] (b) Ours (c) Ground Truth

Figure 6: **Volumetric Completion Results**

5. Conclusion

We propose OctNetFusion, a deep 3D convolutional neural network that is capable of fusing depth information from different viewpoints to produce accurate and complete 3D reconstructions. Our experiments demonstrate the advantages of our learning-based approach over the traditional fusion baseline and show that our method generalizes to novel object categories, producing compelling results on both synthetic and real Kinect data. While in this paper we have focused on the problem of fusing depth maps, an interesting direction for future work is to extend our model to 3D reconstruction from RGB images where 3D representations are learned jointly with 2D image representations in an end-to-end fashion.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv.org*, 1511.00561, 2015. 3

[2] M. Blaha, C. Vogel, A. Richard, J. D. Wegner, T. Pock, and K. Schindler. Large-scale semantic 3d reconstruction: An adaptive multi-resolution model for multi-class volumetric labeling. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[3] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother. Learning 6d object pose estimation using 3d object coordinates. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 2

[4] A. Brock, T. Lim, J. M. Ritchie, and N. Weston. Generative and discriminative voxel modeling with convolutional neural networks. *arXiv.org*, 1608.04236, 2016. 7

[5] A. X. Chang, T. A. Funkhouser, L. J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. Shapenet: An information-rich 3d model repository. *arXiv.org*, 1512.03012, 2015. 3

[6] S. Choi, Q. Zhou, and V. Koltun. Robust reconstruction of indoor scenes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2

[7] S. Choi, Q. Zhou, S. Miller, and V. Koltun. A large dataset of object scans. *arXiv.org*, 1602.02481, 2016. 1, 6, 7

[8] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2016. 2, 7

[9] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *ACM Trans. on Graphics (SIGGRAPH)*, 1996. 1, 2, 5, 6, 7, 8

[10] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[11] A. Dai, C. R. Qi, and M. Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[12] A. Delaunoy and M. Pollefeys. Photometric bundle adjustment for dense multi-view 3d modeling. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1

[13] A. Dosovitskiy, P. Fischer, E. Ilg, P. Haeusser, C. Hazirbas, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2015. 3

[14] M. Firman, O. Mac Aodha, S. Julier, and G. J. Brostow. Structured prediction of unobserved voxels from a single depth image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3, 8

[15] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 32(8):1362–1376, 2010. 1

[16] A. Geiger and C. Wang. Joint 3d object and layout inference from a single rgb-d image. In *Proc. of the German Conference on Pattern Recognition (GCPR)*, 2015. 1

[17] R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta. Learning a predictable and generative vector representation for objects. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2016. 7

[18] S. Gupta, P. A. Arbeláez, R. B. Girshick, and J. Malik. Aligning 3D models to RGB-D images of cluttered scenes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[19] C. Haene, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3D scene reconstruction and class segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[20] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2015. 6

[21] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. R. Bradski, K. Konolige, and N. Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Proc. of the Asian Conf. on Computer Vision (ACCV)*, 2012. 2

[22] J. Huang and S. You. Point cloud labeling using 3d convolutional neural network. In *Proc. of the International Conf. on Pattern Recognition (ICPR)*, 2016. 2

[23] B.-S. Kim, P. Kohli, and S. Savarese. 3D scene understanding by Voxel-CRF. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2013. 2

[24] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2015. 6

[25] Y. Li, S. Pirk, H. Su, C. R. Qi, and L. J. Guibas. FPNN: field probing neural networks for 3d data. In *Advances in Neural Information Processing Systems (NIPS)*, 2016. 3

[26] S. Liu and D. Cooper. Statistical inverse ray tracing for image-based 3d modeling. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 36(10):2074–2088, 2014. 2

[27] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *ACM Trans. on Graphics (SIGGRAPH)*, 1987. 2

[28] D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *Proc. IEEE International Conf. on Intelligent Robots and Systems (IROS)*, 2015. 1, 2

[29] A. Miller, V. Jain, and J. L. Mundy. Real-time rendering and dynamic updating of 3-d volumetric data. In *Proc. of the Workshop on General Purpose Processing on Graphics Processing Units (GPGPU)*, page 8, 2011. 3

[30] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. of the International Conf. on Machine learning (ICML)*, 2010. 4

[31] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and

A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Proc. of the International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011. 1, 2, 5

[32] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3d reconstruction at scale using voxel hashing. In *ACM Trans. on Graphics (SIGGRAPH)*, 2013. 1, 2

[33] J. Park, H. Kim, Y. Tai, M. S. Brown, and I. Kweon. High quality depth map upsampling for 3d-tof cameras. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2011. 6

[34] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[35] D. J. Rezende, S. M. A. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess. Unsupervised learning of 3d structure from images. *arXiv.org*, 1607.00662, 2016. 7

[36] G. Riegler, A. O. Ulusoy, and A. Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3, 4

[37] H. Samet. *Foundations of Multidimensional and Metric Data Structures (The Morgan Kaufmann Series in Computer Graphics and Geometric Modeling)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005. 2

[38] N. Savinov, L. Ladicky, C. Hne, and M. Pollefeys. Discrete optimization of ray potentials for semantic 3d reconstruction. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[39] J. L. Schnberger and J.-M. Frahm. Structure-from-motion revisited. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[40] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006. 1

[41] A. Sharma, O. Grau, and M. Fritz. Vconv-dae: Deep volumetric shape learning without object labels. *arXiv.org*, 1604.03755, 2016. 7

[42] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGB-D images. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2012. 1

[43] S. Song, S. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[44] S. Song and J. Xiao. Sliding shapes for 3D object detection in depth images. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 1

[45] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3

[46] F. Steinbrucker, C. Kerl, and D. Cremers. Large-scale multi-resolution surface reconstruction from rgb-d sequences. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2013. 1, 2, 5

[47] A. O. Ulusoy, M. Black, and A. Geiger. Patches, planes and probabilities: A non-local prior for volumetric 3d reconstruction. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[48] A. O. Ulusoy, M. Black, and A. Geiger. Semantic multi-view stereo: Jointly estimating objects and voxels. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[49] A. O. Ulusoy, A. Geiger, and M. J. Black. Towards probabilistic volumetric reconstruction using ray potentials. In *Proc. of the International Conf. on 3D Vision (3DV)*, 2015. 1, 2

[50] P.-S. Wang, Y. Liu, Y.-X. Guo, C.-Y. Sun, and X. Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. In *ACM Trans. on Graphics (SIGGRAPH)*, 2017. 3

[51] T. Whelan, M. Kaess, M. F. andH. Johannsson, J. Leonard, and J. McDonald. Kintinuous: Spatially extended KinectFusion. In *RSSWORK*, 2012. 1, 2, 6

[52] T. Whelan, S. Leutenegger, R. F. Salas-Moreno, B. Glocker, and A. J. Davison. Elasticfusion: Dense SLAM without A pose graph. In *Proc. Robotics: Science and Systems (RSS)*, 2015. 1

[53] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3d interpreter network. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2016. 2

[54] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 6

[55] C. Zach, T. Pock, and H. Bischof. A globally optimal algorithm for robust tv-l1 range image integration. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2007. 1, 2, 6, 7, 8

[56] B. Zheng, Y. Zhao, J. C. Yu, K. Ikeuchi, and S.-C. Zhu. Beyond point clouds: Scene understanding by reasoning geometry and physics. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2, 8

[57] Q.-Y. Zhou, S. Miller, and V. Koltun. Elastic fragments for dense scene reconstruction. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2013. 2