Stereo Occupancy Networks

Ethan Chun
Massachusetts Institute of Technology
elchun@mit.edu

Sarah Zhang
Massachusetts Institute of Technology
sjzhang@mit.edu

Abstract

Current algorithms for three-dimensional implicit neural representations often use single images or point clouds as input data. Subsequently, the noise artifacts or lack of information from these data can degrade the robustness of their implicit representations. In this study, we propose an alternative approach that utilizes stereo input images with an occupancy network to create an implicit neural representation of household objects. We train and evaluate our method on a subset of the ShapeNet dataset, with a focus on reconstructing common objects such as mugs, bowls, and bottles. We find that stereo occupancy networks can outperform traditional monocular occupancy networks in reconstructing some objects. However, the average volumetric intersection over union scores (IoU) of our stereo occupancy network were 40% lower than those of a similar monocular occupancy network. We suspect that more training is needed reach the full potential of stereo vision. Our findings contribute to the field of robotic perception and manipulation, paving the way for improved methods for representing and understanding three-dimensional objects in real-world settings.

1. Introduction

One of the key challenges in robotics is to perceive and manipulate objects in the real world with accuracy and precision [26]. To tackle this challenge, researchers have been exploring the use of implicit neural representations to represent three-dimensional information for robotic manipulation tasks [22]. These representations provide several advantages over traditional object representations, including infinite resolution, low data consumption, and the ability to represent more nuanced information about perceived three-dimensional geometry. However, current implicit neural representations rely on single images, which provide limited information, or point clouds, which may be noisy and unreliable.

To address these limitations, we aim to investigate whether stereo input images can be directly mapped to occupancy values to create an implicit neural representation of household objects. By using deep neural networks to generate occupancy values directly from stereo input images, the intermediate step of constructing a depth map or point cloud can be eliminated. We hypothesize that this, in turn, may provide improvements in reconstruction accuracy over conventional single-image occupancy networks.

More specifically, we propose a method to create a three-dimensional cost volume from stereo image pairs, which is then used to condition an occupancy network [15] to reconstruct mugs, bowls, and bottles. We demonstrate the efficacy of our method in comparison to a conventional image-based occupancy network through both qualitative analysis and quantitative analysis using averaged volumetric IoU scores.

2. Related Work

2.1. Implicit Neural Representations

Previous work has demonstrated the capabilities of fully connected networks as continuous and memory-efficient implicit representations for capturing objects [1, 8, 9, 17, 18] and scenes [3, 11, 13, 23, 24]. Unlike conventional signal representations, which are discrete, implicit neural representations parameterize an object as a continuous function. This allows for a more flexible and expressive representation, capable of grasping the fine-grained details and variations in shape and scene structures.

Occupancy Networks. Instead of relying on voxelized representations at fixed resolutions, which tend to have a substantial memory footprint, occupancy networks learn a three-dimensional occupancy function using a neural network. This function maps any point in three-dimensional space to an occupancy value, conveying the probability that that point is occupied by a portion of the represented object. This representation encodes a description of the three-dimensional output at infinite resolution, which leads to precise three-dimensional reconstruction from various input types, including single images, noisy point clouds, and coarse discrete voxel grids [15]. Absent from these input

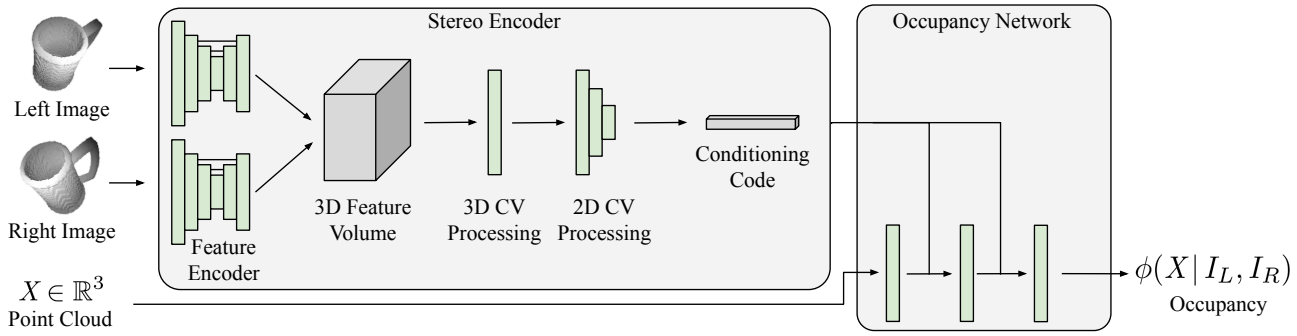

Figure 1. **Network Architecture**. Our network takes in a stereo image pair and a set of sample points. The stereo images are encoded with a dilated ResNet, converted to a 3D cost volume, then processed with a series of 3D and 2D convolutions into a single conditioning code. This conditioning code and the original sample points are fed into an occupancy network, which uses a series of linear layers to predict whether each point in the sample points is occupied by the object represented by the conditioning code.

types is stereo imagery, a gap we intend to fill.

Convolutional Occupancy Networks. While traditional implicit models have shown promising results, their simple fully-connected network architecture limits their ability to perform structured reasoning. As a consequence, the surface reconstructions produced by these models tend to appear overly smooth. In contrast, Convolutional Occupancy Networks combine convolutional encoders with implicit occupancy decoders to capture both local and global information, resulting in equivariant and scalable implicit representations [19].

2.2. Stereo Vision and Depth Estimation

RGB-D cameras have significantly improved the reconstruction of objects due to their ability to capture depth information in addition to color information. However, changes in illumination can lead to reduced depth accuracy and insufficient depth information [5]. Consequently, researchers have introduced methods to estimate depth directly from RGB images. These approaches seeks to overcome the constraints imposed by various lighting conditions, fostering more reliable and versatile depth reconstruction.

Monocular Depth Estimation. The pursuit of estimating depth from a single image, known as monocular depth estimation, has gained significant attention due to the success of deep neural networks in computer vision [7, 10, 30]. However, the inherent limitations arising from the absence of direct depth information in a single image cannot provide accurate and reliable depth estimation in complex and unfamiliar real-world environments [2, 25]. Therefore, alternative approaches like stereo depth estimation, which leverage multiple images for inference, have been used to better reason about the scene geometry [6, 12, 20].

Stereo Depth Estimation. Unlike depth cameras and sensors, stereo images estimate depth or disparity by taking bidirectional disparities and feature correspondences between the two views into account [28]. Building off a model proposed by Kendal et al. [12], the Toyota Research Institute has developed a passive stereo depth system specifically designed for human environments, capable of generating dense and accurate point clouds [21]. The learned stereo model excels in producing high-resolution depth maps even in challenging scenarios, such as on dark, texture-less, thin, reflective, and specular objects and surfaces.

3. Methods

3.1. Architecture

Our proposed model draws inspiration from the influential works of Mescheder et al. on occupancy networks [15] and Shankar et al. on learned stereo depth [21]. To this end, we use an encoder-decoder architecture, where the encoder converts stereo image pairs into a latent conditioning code and the decoder then predicts object occupancy using this code. This architecture aims to harness the strengths of both occupancy networks and learned stereo depth techniques to enhance reconstruction capabilities. We provide a visual representation of this architecture in Figure 1.

Encoder. Our encoder consists of a dilated ResNet [27] which extract features from both the left and right images. These features are merged into a cross-correlation cost volume [14] to create a four dimensional representation of the input data. Next, we run three-dimensional and two-dimensional convolutions on this cost volume, finally down sampling to a conditioning code of size 512. This approach enables us to leverage the advantages of stereo image pairs to produce a more accurate conditioning code for our occupancy network.

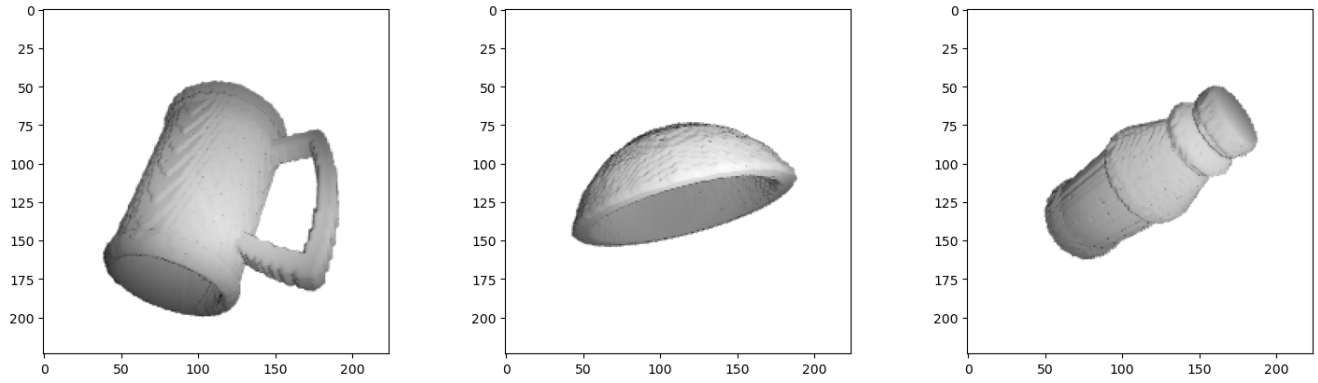

Figure 2. **Dataset**. We use a subset of the ShapeNet dataset consisting of mugs, bowls, and bottles. These objects are commonly found in the household, lending themselves to research in household robotics. We anticipate that our neural implicit representations may be particularly relevant in this field of robotics.

Decoder. Our occupancy network decoder follows the architecture of Mascheder et al., taking in a sample point cloud, $X \in \mathbb{R}^3$, and the conditioning code. A series of linear layers are used to predict whether each sample point is occupied by a portion of the object encoded by the conditioning code. This system can be used to reconstruct objects directly [15] or as a way to create neural fields for use in other applications [22]. We use the former approach.

3.2. Dataset

To train and evaluate the model, we utilize a subset of the ShapeNet dataset consisting of 114 mugs, 174 bowls, and 352 bottles. Some examples of these objects are shown in Figure 2. ShapeNet is a popular, richly-annotated, large-scale dataset of three-dimensional shapes [4], which has been used to evaluate implicit neural representations in previous works [22, 29]. We choose this specific selection of objects for their prevalence in household environments, a popular application for robotic manipulation systems.

Data Preprocessing. To evaluate our network, we create a set of data points consisting of four values: a left image, a right image, a set of sample points, and the occupancies for these sample points. The left and right image correspond to a stereo image pair showing the same object from slightly different positions. The sample points are used to evaluate the network's reconstruction ability.

We render a set of 1,000 stereo image pairs for each object in our dataset, each pair similar to those shown in Figure 3. Using Trimesh and Pyrender, we model two cameras in simulation. Each camera is a distance $r = 1$ meter from the object and a distance $\theta = \pi/16$ from a center line, as shown in Figure 4. For each image pair, the object is set at a random orientation in 3D space ($R \in \text{SO}(3)$). This provides us with a total of 640,000 data points.

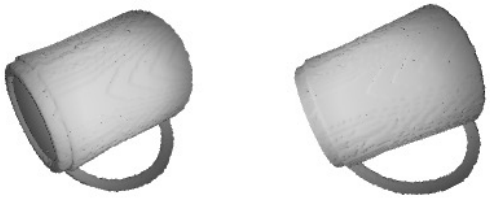

Figure 3. **Stereo Image Pairs**. For each of our mugs, bowls, and bottles, we generate a stereo image pair consisting of a left and right image. Pictured above is an example stereo pair of a mug.

To generate our sample points, we sample $n = 2,048$ points in the bounding box of each object, then extract ground truth occupancy values for these points. We transform the points to have the same pose as the photographed object.

Finally, we split the dataset into train, test, and validation data with 80% of the data in train, 10% in validation, and 10% in test.

3.3. Training and Inference

We use Binary Cross Entropy on the predicted occupancy of our sample points to calculate the network loss. This provides a convenient means to enforce the accuracy of our occupancy predictions. We train for a total of 40,000 iterations, validating every 4,000 iterations on both loss and volumetric IoU.

3.4. Baselines

We compare our proposed method to an occupancy network with a single pre-trained ResNet-18 encoder in place of our stereo encoder [15]. This network only receives the left image from the stereo pair and must perform the same reconstruction task. Similar to our stereo system, we train this model for 40,000 iterations.

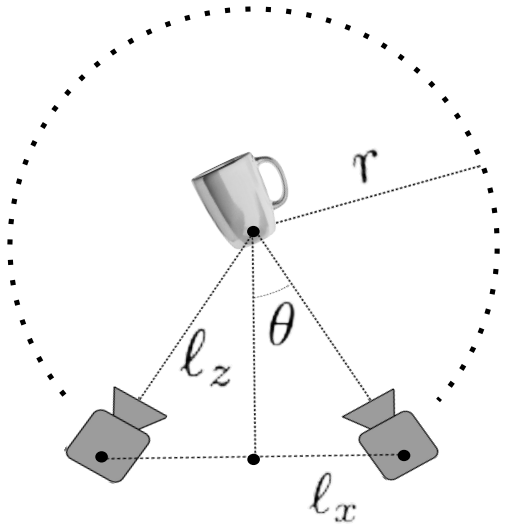

Figure 4. **Rendering Stereo Images**. To render our stereo image pairs, we position an object at the origin of our scene. Then, we place two cameras a distance $r = 1$ meter away from the center. These cameras are placed at an angle $\theta = \pi/16$ on either side of the mid-line.

3.5. Evaluation Metrics

We evaluate our method and baselines by performing reconstruction tests on mugs, bowls, and bottles. We measure reconstruction performance using the volumetric intersection over union (IoU) metric, which is defined as the quotient of the volume of the two meshes' intersection and the volume of their union, where higher IoU values indicate better reconstruction results. More formally, we are concerned with Equation 1, where $\mathcal{M}_{pred}$ and $\mathcal{M}_{GT}$ denote the set of all points that are inside or on the predicted and ground truth mesh, respectively.

$$\text{IoU}(\mathcal{M}_{pred}, \mathcal{M}_{GT}) = \frac{|\mathcal{M}_{pred} \cap \mathcal{M}_{GT}|}{|\mathcal{M}_{pred} \cup \mathcal{M}_{GT}|}. \quad (1)$$

This metric has been commonly used to evaluate object detection and three-dimensional reconstruction models [15, 16], making it suitable for our application.

4. Results

Quantitative Evaluation. Our first step in evaluation is to understand the quantitative accuracy of the model's reconstruction abilities. As shown in Table 1, we find that the monocular occupancy network produces a higher volumetric IoU than the stereo system. This translates to the monocular system producing *more* accurate reconstruction of objects. This is somewhat expected, as the limited training resources available to us meant that neither model could be fully trained. Since the monocular system already uses a pre-trained encoder, these weights are likely closer to optimal, leading to higher reconstruction accuracy for a correspondingly low training time.

Volumetric IoU	
Monocular	**Stereo**
0.1896	0.1153

Table 1. **Volumetric IoU**. We measure reconstruction performance of mugs, bowls, and bottles using volumetric IoU for monocular and stereo reconstructions. A higher IoU score corresponds to a more accurate model.

Qualitative Evaluation. To gain insight into the performance of each system on individual object categories, we also inspected the visual quality of the reconstructions produced. As shown in Figure 5, our reconstructions highlight the differential performance of stereo and monocular reconstruction in capturing specific object features.

Specifically, we observed that stereo reconstruction outperforms monocular reconstruction in capturing the neck of the bottle. The additional depth information obtained from stereo image pairs allows for a more accurate representation of the bottle's geometry, resulting in a more faithful reconstruction of the neck region. While the performance of the stereo system on bottles highlights some performance increases, we find that for bowls, the results of each network were extremely similar, both systems producing similar reconstructions of the relatively simple geometry. However, it is important to note that stereo reconstruction did not demonstrate a similar advantage in capturing the handle of the mug. In this case, monocular reconstruction produced a clearer reconstruction of the mug's handle while the stereo system completely omitted the handle.

5. Discussion

One key limitation we faced was the restricted computational resources, which limited the training of our stereo model to only 40,000 steps. As a result, the stereo model was undertrained and most likely could not reach its full potential. To fully unleash the capabilities of the stereo model, a longer training period with several hundred thousand steps would be needed, which exceeded our available compute resources. Consequently, we recognize the need for more extensive training time to capture the subtleties between stereo and monocular reconstructions. With additional training resources and an extended training period, we can thoroughly assess the benefits and drawbacks of stereo reconstruction compared to its monocular counterpart.

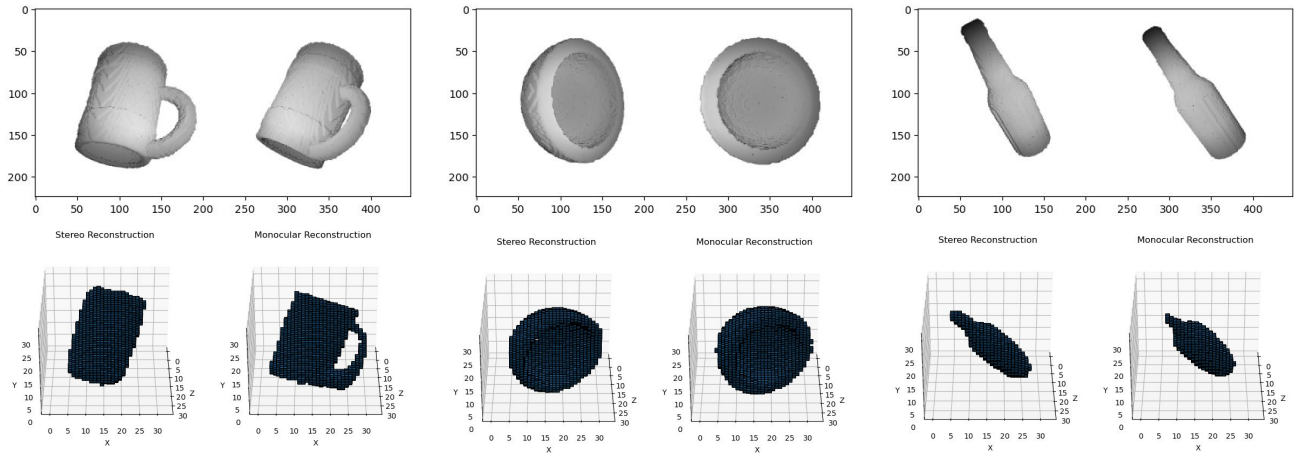

Figure 5. **Monocular and Stereo Reconstructions**. We compare the reconstruction ability of stereo and monocular occupancy networks for mugs, bowls, and bottles. Note that in some instances, such as with the bottle, the stereo system shows greater ability to reconstruct fine details like the neck of the bottle. In other instances, as with the bowl, the reconstructions from both networks appear extremely similar. However, in some cases, such as with the mug, we find that the monocular system can instead produce finer reconstructions.

Despite these shortcomings, the ability of our stereo system to produce reconstructions at all indicates that that occupancy network architecture is a highly adaptable framework, lending itself to a wide range of improvements and extensions. Following this prototype, future work may explore a variety of encoder architectures: variations of stereo systems, transformers, or even audio or text-based systems.

6. Conclusion and Future Research

In this paper, we proposed a novel extension of the occupancy network architecture. While our study focused on the reconstruction of mugs, bowls, and bottles as a representative example, the generalizability of our findings to other object categories and complex scenes remains an open question. Further investigation is necessary to evaluate the performance of stereo reconstruction across a wider range of objects and scenes, such as packages in warehouse environments.

By incorporating a stereo encoder to gather more information about an object, our system demonstrates a promising approach for generating high-quality implicit neural representations of commonly found household objects. We note that our network relies solely on color images, making it particularly well-suited for the reflective and textureless surfaces encountered in household environments. By focusing on this aspect, our work paves the way for robots to better comprehend and interact with household objects in real-world settings, thereby opening up new possibilities for home assistance, maintenance, and the integration of robots into our daily lives.

Author Contributions

The contributions of each author are described as follows: Ethan Chun contributed to the implementation of the data processing pipeline. Sarah Zhang contributed to the implementation of the model architecture. Both team members were fully involved in the writing and editing of the report, generation of figures, and qualitative and quantitative evaluation of results.

Acknowledgements

We express sincere appreciation to Professors Bill Freeman, Vincent Sitzmann, and Mina Konaković Luković for introducing us to this research area and providing valuable insights into stereo vision, as well as three-dimensional scene representation and reconstruction. We also express gratitude to Christine Casatelli and Sai Bangaru for providing invaluable feedback on the proposal of the paper. Furthermore, we would like to express our thanks to the 6.8301 TAs for their unwavering support and assistance throughout the semester.

References

[1] Matan Atzmon and Yaron Lipman. SAL: sign agnostic learning of shapes from raw data. *CoRR*, abs/1911.10414, 2019. 1

[2] Amlaan Bhoi. Monocular depth estimation: A survey. *CoRR*, abs/1901.09402, 2019. 2

[3] Rohan Chabra, Jan Eric Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard A. Newcombe. Deep local shapes: Learning local SDF priors for detailed 3d reconstruction. *CoRR*, abs/2003.10983, 2020. 1

[4] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015. 3

[5] Yan Chen, Jimmy S. J. Ren, Xuanye Cheng, Keyuan Qian, and Jinwei Gu. Very power efficient neural time-of-flight. *CoRR*, abs/1812.08125, 2018. 2

[6] Yilun Du, Cameron Smith, Ayush Tewari, and Vincent Sitzmann. Learning to render novel views from wide-baseline stereo pairs, 2023. 2

[7] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue, 2016. 2

[8] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas A. Funkhouser. Deep structured implicit functions. *CoRR*, abs/1912.06126, 2019. 1

[9] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T. Freeman, and Thomas A. Funkhouser. Learning shape templates with structured implicit functions. *CoRR*, abs/1904.06447, 2019. 1

[10] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *CoRR*, abs/1609.03677, 2016. 2

[11] Chiyu "Max" Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Niessner, and Thomas Funkhouser. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[12] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. *CoRR*, abs/1703.04309, 2017. 2

[13] Amit P. S. Kohli, Vincent Sitzmann, and Gordon Wetzstein. Inferring semantic information with 3d neural scene representations. *CoRR*, abs/2003.12673, 2020. 1

[14] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. *CoRR*, abs/1512.02134, 2015. 2

[15] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 4

[16] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Supplementary material for occupancy networks: Learning 3d reconstruction in function space. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[17] Mateusz Michalkiewicz, Jhony K. Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[18] Jeong Joon Park, Peter R. Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *CoRR*, abs/1901.05103, 2019. 1

[19] Songyou Peng, Michael Niemeyer, Lars M. Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. *CoRR*, abs/2003.04618, 2020. 2

[20] Sudeep Pillai, Srikumar Ramalingam, and John J. Leonard. High-performance and tunable stereo reconstruction, 2016. 2

[21] Krishna Shankar, Mark Tjersland, Jeremy Ma, Kevin Stone, and Max Bajracharya. A learned stereo depth system for robotic manipulation in homes. *IEEE Robotics and Automation Letters*, 7:2305–2312, 2022. 2

[22] Anthony Simeonov, Yilun Du, Andrea Tagliasacchi, Joshua B. Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, and Vincent Sitzmann. Neural descriptor fields: Se(3)-equivariant object representations for manipulation. *International Conference on Robotics and Automation (ICRA)*, 2022. 1, 3

[23] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Proc. NeurIPS*, 2020. 1

[24] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *CoRR*, abs/1906.01618, 2019. 1

[25] Nikolai Smolyanskiy, Alexey Kamenev, and Stan Birchfield. On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach. *CoRR*, abs/1803.09719, 2018. 2

[26] Yu Sun, Joe Falco, Máximo A. Roa, and Berk Çalli. Research challenges and progress in robotic grasping and manipulation competitions. *CoRR*, abs/2108.01483, 2021. 1

[27] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1451–1460, 2018. 2

[28] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, Xiaojun Tong, and Wenxiu Sun. Toward 3d object reconstruction from stereo images. *Neurocomputing*, 463:444–453, 2021. 2

[29] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomír Mech, and Ulrich Neumann. DISN: deep implicit surface network for high-quality single-view 3d reconstruction. *CoRR*, abs/1905.10711, 2019. 3

[30] Chaoqiang Zhao, Qiyu Sun, Chongzhen Zhang, Yang Tang, and Feng Qian. Monocular depth estimation based on deep learning: An overview. *CoRR*, abs/2003.06620, 2020. 2