# On Image to 3D Volume Construction for E-Commerce Applications

Mohamed Abdelfattah
*German University in Cairo*
muhamad.abdelfattah@guc.edu.eg

Mostafa Alaa
*German University in Cairo*
mostafa.talaat@guc.edu.eg

Mayar Mohamed
*German University in Cairo*
mayar.mohamed@guc.edu.eg

Sama El Baroudy
*German University in Cairo*
sama.elbaroudy@guc.edu.eg

Rania Reda
*C.E.O Augmania GmbH*
rania@augmania.com

Mohammed A.-M. Salem
*German University in Cairo*
mohammed.salem@guc.edu.eg

Slim Abdennadher
*German University in Cairo*
slim.abdennadher@guc.edu.eg

***Abstract*—Ever since the corona pandemic started back in 2020 and the world stopped. However, ecommerce did not. In fact, one of the very few sectors that did rise during the covid-19 pandemic was ecommerce. Being able to recover the 3D shape of an object from a single or multiple images is a difficult problem that has been attracting a lot of attention lately. Specifically after most shops have had to shut down to prevent the spread of covid-19. A lot of existing solutions are available online but unfortunately, a global solution that deals with any object in any image does not exist yet. In this paper we try to discuss those solutions and how deep learning could be used to bridge the gap between the research and the industry. Finally we discuss the open challenges that are in this problem space and opportunities for future work.**

***Index Terms*—photogrammetry, deep learning, voxels, mesh, 3D computer vision**

## I. Introduction

Ecommerce thrived in 2020 because of store closures and fears of contracting the coronavirus in public. As a direct result of that, online sales went up 32.4%. Naturally, the demand for a platform that allows the customer to be able to visualize the final product before purchase, has risen. Here, Augmented Reality (AR) applications play an essential role to visualize the products that are sold efficiently.

Deducing the accurate and precise 3D shape of an object from a single or multiple 2D image is an important problem in 3D modelling, object recognition and medical diagnosis. Finding a complete approach to that problem that does not require complex camera calibration or depends on triangulation techniques via depth information has proven to be a quite challenging problem. Albeit the fact that using traditional methods that utilize triangulation techniques and complex camera calibration can produce 3D reconstructions that are considered to be of consistent and of satisfactory quality. Being capable of utilizing those techniques for the infinite amount of objects that exist in our world is not practical, and is infeasible in most situations.

This is why recovering the 3D shape of an object from single or multiple images with deep neural networks has been attracting an increasing amount attention in the past couple of years. Specifically that ever since deep learning has emerged as a field, an immense amount of benchmark datasets have been made available to the public so as to make use of those deep learning techniques.

This paper is organized as following. In section II a survey is conducted on the available running platforms that enable augmented reality for business to customer sales. Section III introduces the state of the art techniques used to create 3D volumes from single or multiple 2D images. The techniques are categorized as voxel-based methods and mesh-based methods.

## II. Overview on AR for E-Commerce Applications

### A. Houzz

Houzz is an augmented reality app for furniture and house accessories. It allows users to search for and to buy products online. It aids users in seeing the 3D model being viewed in the room in a high quality as shown in Figure 1 (left).

### B. IKEA Place

IKEA Place is an augmented reality application for home decor and furniture from Ikea. It allows users to search and to place the 3D model in their home with a high quality mesh and texture as shown in Figure 1 (right).

### C. Augment

Augment is a 3D Augmented Reality app that creates augmented images of their products. It helps visualize 3D models in Augmented Reality, integrated in real time and in their actual size and environment as shown in Figure 2 (left)

### D. AR in Google Search

The AR in the google search app allows us to place 3D digital objects right in our own space directly from Search or from websites on Chrome. It provides a great sense of context and scale as shown in Figure 2 (right)

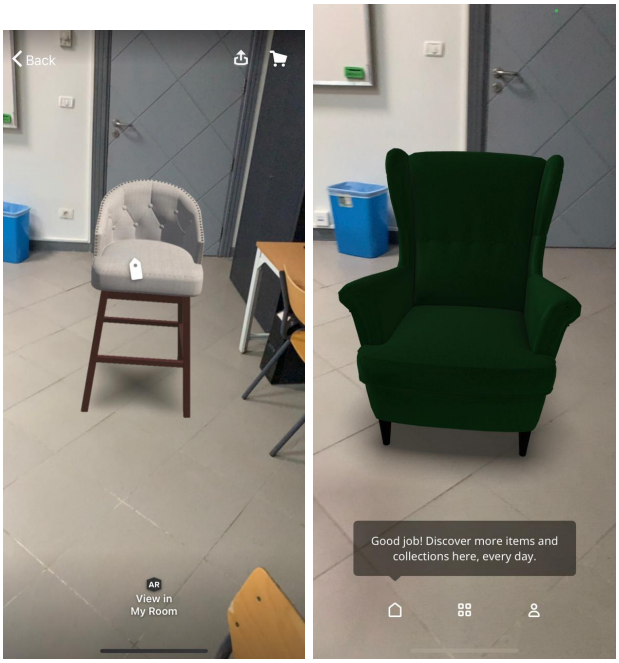


Fig. 1. Samples of Houzz (left) and IKEA Place (right) applications.

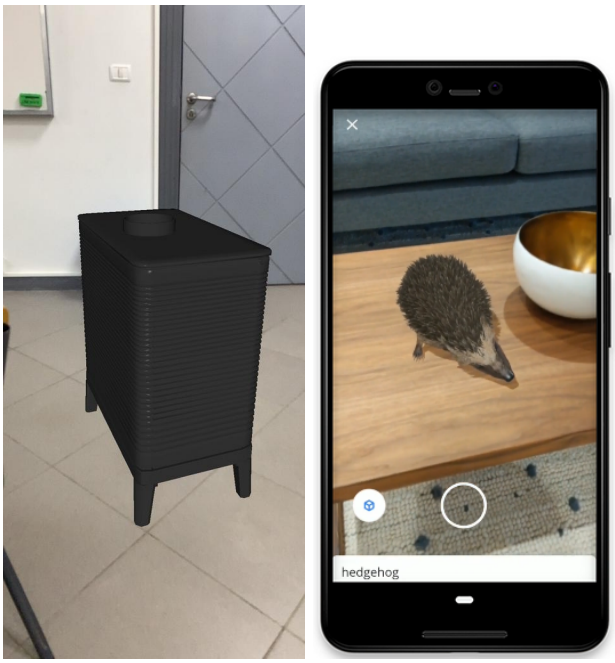


Fig. 2. Samples of Augment (left) and AR in Google Search (right).

## III. Image to Volume Construction Techniques

### A. Material Preparation

Due to the computational complexity of the problem at hand, we have chosen to consider a cloud computing service in order to be able to store, process our data in a decentralized fashion that could output results within a limited time frame.

The Amazon Web service (aws) was chosen because it is so much bigger than the other competitors and provides an edge when it comes to storage and ease of access. This is exceptionally important in our use-case due to the enormous size of our datasets. The aws instance of choice was EC2-3.8Large

### B. Benchmark Datasets

In an endeavour to answer our research question, two benchmark datasets were obtained by downloading and unpacking them into our aws instance. The first dataset, ShapeNet Core [1], which is a dataset of synthetic images that is made up of 51,300 unique 3D models that are categorized into 12 different categories and their corresponding 2D images. The second dataset, Pix3D [2], which is a dataset of realistic images that is made up of 4000 2D Images along with their corresponding masks and 3D models.

Both benchmark datasets were downloaded and unpacked into the S3 aws instance that was prepared for the purpose of this research. The stack of required python libraries was installed onto the aws machine to aid in loading and processing both datasets.

Towards bridging the gap between the realistic (Pix3D) and synthetic images (ShapeNet), an aggregate of both was created and 10% of that aggregated dataset was left out for testing. The remaining 90% was used for training and validation.

### C. Deep Learning Techniques

Recently, numerous deep learning techniques have been employed to recover the 3D shape of an object from a single or multiple 2D images. Most of those techniques employed recurrent neural networks in order to extract multiple feature maps from the input sequence of images and incrementally reconstruct the 3D shape of the object contained within that sequence.

Unfortunately, such methods are incapable of robustly estimating the 3D shape of an object due to the shortcomings of recurrent neural networks. Firstly, due to permutation variance (Vinyalset al., 2016) [3], recurrent neural networks are unable to incessantly produce the same output when given the same sequence of 2D images with different orders. Secondly, due to the long-term memory loss in recurrent neural networks (Pascanu et al., 2013) [4], important features

of preliminary 2D images in the sequence are usually obliterated. Thirdly, due to the sequential processing nature of recurrent neural networks (Hwang and Sung, 2015) [5] training proves to require an immense amount of time.

In an endeavour to overcome these shortcomings, DeepMVS (Huang et al., 2018) [6] exploit max pooling layers in order to find an effective method for aggregating information across a set of unordered images. In addition, Ray-Net (Paschalidou et al., 2019) [7] employ average pooling layers in order to collectively aggregate the deep features extracted from the same position in a 2D image to recover the 3D structure. Most recently, AttSets (Yang et al., 2020) [8] utilized an attention based aggregation component that automatically predicts a matrix of weights, which are used as attention scores for input features. Despite showing promising results, the former method proves to be be quite challenging for images with noisy backgrounds and may not be robust in reconstructing objects in real world scenarios.

According to (Xian-Feng Ha et al., 2021) [9], most 3D reconstruction techniques encode the final output in a 3D volumetric discrete representation, which is called a voxel grid. The finer the discretization is, the more accurate the representation will be of the object embedded within the 2D image. The main advantage of using volumetric grids is that many of the currently existing deep learning architectures that have been designed for processing 2D images can be easily extended to 3D data by replacing the 2D pixels array with their respective 3D representations and then processing the grid using 3D convolution. On the other hand, the second most common final output is a 3D surface mesh, which is a representation of a surface that consists of vertices, edges and faces. In order to retrieve the actual 3D surface mesh from a volumetric representation, a post processing step, e.g. marching cubes [10] is used.

The main drawback of working directly with 3D surface meshes is that common representations such as meshes or point clouds are not regularly structured and thus, they are not readily handled by deep learning techniques, especially those that use CNNs. Despite that, 3D surface meshes are quite favorable from an endpoint stand of view because volumetric representation-based methods are computationally very wasteful since information is rich only on or near the surfaces of 3D shapes.

The comparative summary of the generic 3D object reconstruction techniques, conducted by (Xian-Feng Ha et al., 2021) [9] provides a comprehensive overview of state of the art techniques for both, voxel-based and surface mesh techniques, respectively. Upon further analysis of the state of the art techniques, we concluded that for the sake of our project, one technique will be investigated from each, voxel-based and surface mesh techniques, respectively. In our decision making process, many properties and factors were taking into consideration the time constraint, the complexity of the problem space and the exhaustive amount of computational resources required. Finally, a settlement was made to explore Pix2Vox [11] for voxel-based techniques and DeepMesh [12] for mesh-based techniques, respectively.

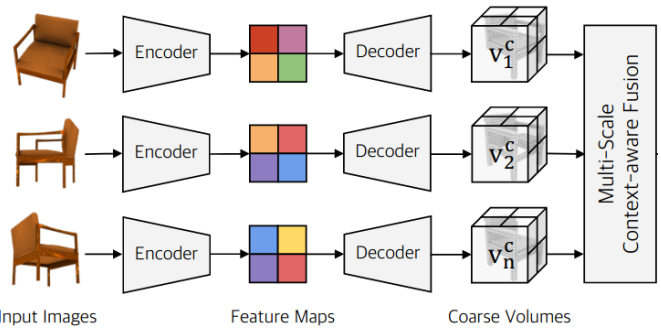


Fig. 3. Overview of the encoder, decoder, and context-aware fusion module.

### D. Voxel-based Deep Learning Techniques

Pix2Vox++ [11] provides a novel framework for single view and multi view 3D reconstruction. It consists of four modules: encoder, decoder, multi-scale context-aware fusion module, and refiner. The encoder computes a set of features for the decoder to recover the 3D shape of the object whereas the decoder is responsible for transforming the information encapsulated in the 2D feature maps into 3D volumes as shown in figure 3.

There are two encoder architectures, Pix2Vox++/Fast and Pix2Vox++/Accurate, as shown in Figure 4. The former involves much fewer parameters and lower computational complexity. The latter has more parameters, which can reconstruct more accurate3D shapes but has higher computational complexity. On the other hand, the decoder produces output encoded in two different resolutions. Low resolution reconstructions are of size 32x32x32 and are produced by 5 transposed convolutional layers whereas high resolution reconstructions are of size 64x64x64 and are produced by 6 transposed convolutional layers, respectively.

The reconstruction qualities of a part of an object that are visible from one viewpoint are much higher than those of invisible parts. Thus, towards meeting that end, the multi-scale context-aware fusion module was developed in order to adaptively select high-quality reconstruction for each part from different coarse 3D volumes and fuse them to produce a 3D volume of the entire object. In addition, the multi-scale context-aware fusion module is used to preserve details in shallower convolutional layers because even though deeper convolutional layers have larger receptive fields, they may lose details of the object.

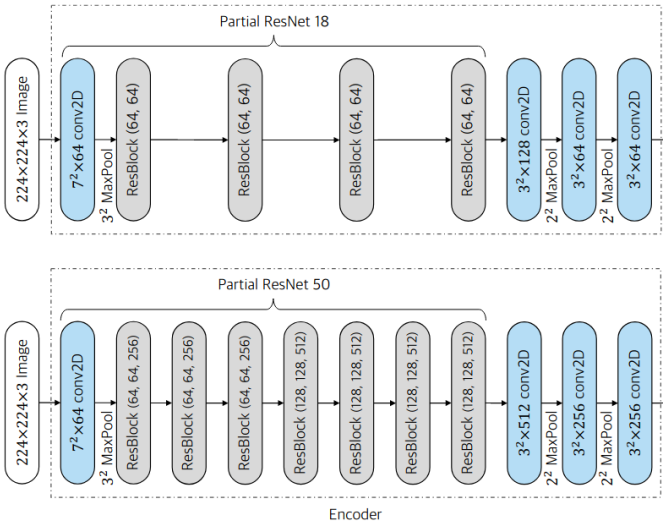


Fig. 4. Overview of the fast (top) and accurate (bottom) encoder architectures.

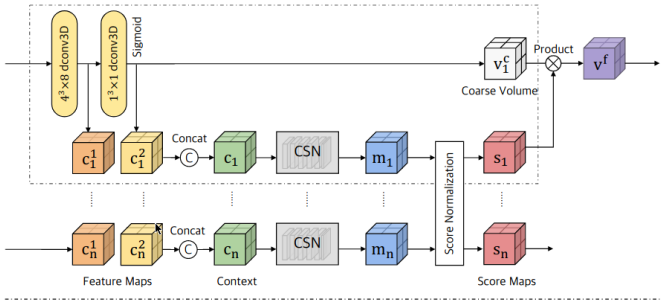


Fig. 5. Overview of the context-aware fusion module.

### E. Mesh-based Deep Learning Techniques

DeepMesh [12], which is shown in figure 6, strives to generate a 3D mesh from a single template spherical mesh by dynamically modifying the topology of that template mesh by face pruning. In the process a trade-off between the deformation flexibility of the template sphere and the target output meshing quality, is achieved. The DeepMesh encoder-decoder network is composed of an encoder, 2 main sub-networks/modules for topology modification and a final module that refines the boundary conditions of the obtained surface mesh.

The encoder, which is inherently made up of ResNet-18, is used for shape generation by taking as input a 2D image and extracting a 1024-dimensional feature vector x. On the other hand, the decoder contains three successive modules. Each of the first two modules consist of a mesh deformation module and a topology modification module, and the last module comprises a single boundary refinement module. The three modules allow modification of the coordinates and connectivity of the vertices on the predefined template mesh.

Each mesh deformation module, shown in figure 7,

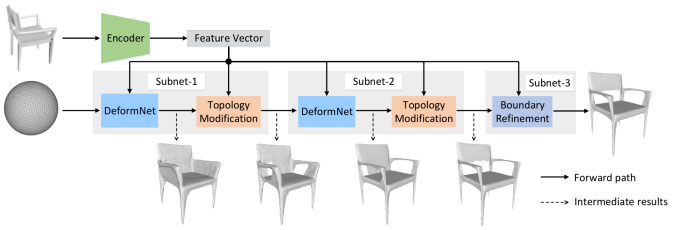


Fig. 6. DeepMesh Architecture.

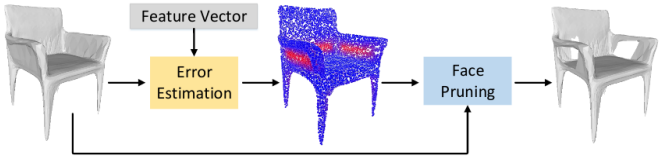


Fig. 7. Topology Modification Network.

consists of a single multilayer perceptron which performs the affine transformation on each vertex of the template mesh and generates the vertex displacements. The multi-layer perceptron aids in predicting the offsets instead of directly regressing the coordinates, to enable more accurate learning of fine geometric details with even less training time. Coherently, the topology modification network updates the topological structure of the reconstructed mesh by pruning the faces which deviate significantly from the ground truth.

## IV. Results

With regards to Pix2Vox++, their main contribution was the Things3D dataset which unfortunately was not available for testing. As such we used the aggregated dataset that we created by merging Pix3D and Shapenet. Using both pretrained models, the fast and the accurate, failed to produce any adequately meaning output. In fact, even after pre-processing the input images, we can see that the generated model fails to represent the object's shape accurately. At best, the model would be able to generate an object with similar overall shape. But even then, it seems to always have missing parts as well as lots of noise. We also saw that, in most cases, the model generates either a very dense or a very sparse model as shown in figure 8.

Upon further examination of the results we noticed that generally speaking, results that were produced as an output from ShapeNet as input produced an overall better 3D reconstruction. This, our main goal behind the addition of these pre-processing techniques was to convert the Pix3D data into something that resembles ShapeNet. Meaning that, we never intended to apply any of these techniques on ShapeNet nor Synthetic data in general.

In an endeavour to quantifying our results, we used the Intersection over Union (IoU) metric and even though we

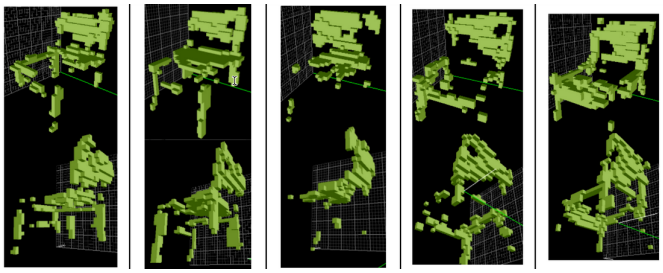

Fig. 8. Sample of Pix2vox++ reconstructed model.

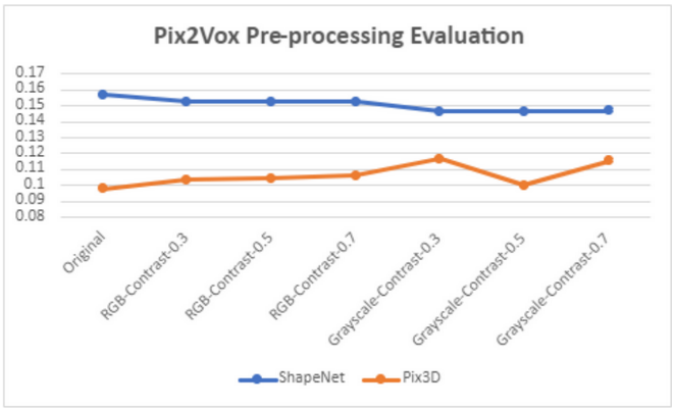


Fig. 9. IoU values VS pre-processing.

managed to achieve an average improvement of 19% higher than the results without the preprocessing, the actual values of the IOU are way below the acceptable range. This is not only true for the Pix3D dataset but also for the ShapeNet one as shown in figure 9.

With regards to DeepMesh, from a qualitative perspective, the approach struggles to reconstruct shapes with complex topologies when the topology modification module is not applied. However, it has outperformed the other approaches in terms of visual quality. Nonetheless, even the state-of-the-art approach is not yet ready for production and further research should be conducted. A sample of the output is shown in figure 10.

Quantitatively, DeepMesh adopts the widely used Chamfer Distance [13] and Earth Mover's Distance [14] to quantitatively evaluate the results. Both metrics are computed between the ground truth point cloud and 10,000 points uniformly sampled from the generated mesh. Although this method can generate visually appealing meshes with smooth surfaces and complex topologies, it still has the inherent drawback of producing open surfaces due to the face pruning operations. To avoid the above mentioned drawbacks, we densely sampled the surface and reconstruct the mesh from the obtained point cloud. The latter approach enhanced the results, however, missed surfaces can not be repainted nor validated.

Fig. 10. 3D reconstructed output of DeepMesh.

| Original | | | |
|---|---|---|---|
| | All | ShapeNet | Pix3D |
| **max** | 0.147705 | 0.057967 | 0.147705 |
| **min** | 0.000299 | 0.000299 | 0.045639 |
| **avg** | 0.004986 | 0.002727 | 0.085299 |
| **Re-trained ResNet18** | | | |
| | All | ShapeNet | Pix3D |
| **max** | 0.087806 | 0.087806 | 0.023806 |
| **min** | 0.000441 | 0.000441 | 0.000700 |
| **avg** | 0.004658 | 0.004685 | 0.003683 |
| **Re-trained ResNet34** | | | |
| | All | ShapeNet | Pix3D |
| **max** | 0.053511 | 0.053511 | 0.019728 |
| **min** | 0.000431 | 0.000431 | 0.000956 |
| **avg** | 0.005073 | 0.005102 | 0.004050 |
| **Re-trained ResNet101** | | | |
| | All | ShapeNet | Pix3D |
| **max** | 0.087377 | 0.087377 | 0.035920 |
| **min** | 0.000355 | 0.000355 | 0.001322 |
| **avg** | 0.004816 | 0.004824 | 0.004560 |

Fig. 11. Max,min, and avg chamfer distance DeepMesh.

In an attempt to push the results of DeepMesh further, we experimented how the results would change if we tried to change the main encoder architecture to Resnet-34 or Resnet-101 instead of Resnet-18. The best average chamfer distance was acquired over the aggregated dataset when the original ResNet-18 module was used for the encoder. However, similar to Pix2Vox++, DeepMesh performs well on synthetic data but is not yet ready for production when exposed to real-world data. It overfits the synthetic data and is unable to scale well in real-world scenarios.

## V. Conclusion

Several deep learning techniques have been employed to recover the 3D shape of an object from a single or multiple 2D images. Pix2Vox++ and DeepMesh have been chosen to be investigated as they represent the state of the art techniques for voxel-based and mesh-based techniques respectively. Pix2Vox++ performs poorly on realistic image datasets but well enough on synthetic images. DeepMesh performs well on synthetic datasets but is not yet ready for production when exposed to real-world data. It overfits the synthetic data and is unable to scale well in real-world scenarios.

## Acknowledgment

This work is funded by the Borg-Elarab Innovation Cluster.

## References

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

[2] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, and W. T. Freeman, "Pix3d: Dataset and methods for single-image 3d shape modeling," 2018.

[3] O. Vinyals, S. Bengio, and M. Kudlur, "Order matters: Sequence to sequence for sets," 2016.

[4] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," 2013.

[5] K. Hwang and W. Sung, "Single stream parallelization of generalized lstm-like rnns on a gpu," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2015. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2015.7178129

[6] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "Deepmvs: Learning multi-view stereopsis," 2018.

[7] D. Paschalidou, A. O. Ulusoy, C. Schmitt, L. van Gool, and A. Geiger, "Raynet: Learning volumetric 3d reconstruction with ray potentials," 2019.

[8] B. Yang, S. Wang, A. Markham, and N. Trigoni, "Robust attentional aggregation of deep feature sets for multi-view 3d reconstruction," *International Journal of Computer Vision*, vol. 128, no. 1, p. 53–73, Aug 2019. [Online]. Available: http://dx.doi.org/10.1007/s11263-019-01217-w

[9] X.-F. Han, H. Laga, and M. Bennamoun, "Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, p. 1578–1604, May 2021. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2019.2954885

[10] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *COMPUTER GRAPHICS*, vol. 21, no. 4, pp. 163–169, 1987.

[11] H. Xie, H. Yao, S. Zhang, S. Zhou, and W. Sun, "Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images," *International Journal of Computer Vision*, vol. 128, 12 2020.

[12] J. Pan, X. Han, W. Chen, J. Tang, and K. Jia, "Deep mesh reconstruction from single rgb images via topology modification networks," 2019.

[13] A. Hajdu, L. Hajdu, and R. Tijdeman, "Approximations of the euclidean distance by chamfer distances," 2012.

[14] M. van Kreveld, F. Staals, A. Vaxman, and J. Vermeulen, "Approximating the earth mover's distance between sets of geometric objects," 2021.