A Review of Deep Learning Techniques for 3D Reconstruction of 2D Images

Anny Yuniarti
Department of Informatics, Faculty of Information Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
anny@if.its.ac.id

Nanik Suciati
Department of Informatics, Faculty of Information Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
nanik@if.its.ac.id

***Abstract*—Deep learning techniques have attracted many researchers in computer vision field to solve computer vision problems such as image segmentation and object recognition. This success also led to the implementation of deep learning techniques in 3D reconstruction. 3D reconstruction itself is a classical problem in computer vision that has been approached by many techniques. However, deep learning techniques for 3D reconstruction are still in the early phase, whereas the opportunity of the techniques is large. Hence, to improve the performance of such approaches, it is important to study the research and applications of the approaches through literature review. This paper reviews deep learning-based methods in 3D reconstruction from single or multiple images. The research scope includes single or multiple image sources but excludes RGB-D type input. Several methods and their significance are discussed, also some challenges and research opportunities are proposed for further research directions.**

Keywords—computer vision, 3D reconstruction, machine learning, deep learning, end-to-end approach.

I. INTRODUCTION

Three-dimensional (3D) reconstruction of 2D images is the process of producing 3D representations of an object, given only a single or multiple images of the object. Recently, several approaches have been proposed using learning-based method for 3D reconstruction of general images. However, such approaches are still in the early stage, meanwhile the potential of the learning-based approaches is large. Therefore, in order to enhance such approaches, it is important to study the research and application of the approaches through literature review.

3D reconstruction from 2D images have been discussed in several review papers. Reyneke et al. [1] discussed several methods of 3D reconstruction from 2D radiographs (X-ray, DXA, fluoroscopic, ultrasound). The reconstruction is accomplished using a deformable model, which encodes prior knowledge and assumptions about the typical 3D appearance of a structure. These models can then be manipulated using a set of parameters. Reyneke et al. [1] suggested to use X-ray images rather than CT and MRI machines since the 3D imaging technologies may not possible, and the risk of cancer due to ionizing radiation can be reduced. Paturkar, Gupta and Bailey [2] reviewed limitations and challenges of 3D reconstruction techniques in agricultural applications. Ma and Liu [3] reviewed 3D reconstruction techniques in civil engineering and their applications. Besides summarizing the techniques and their applications in civil engineering, Ma and Liu [3] also proposing future research directions in the field. All of the three review papers emphasize on one particular applications of 3D reconstruction, ie. medical, agricultural, and civil engineering. Neither of them discussed about deep learning-based 3D reconstruction.

The problem of 3D reconstruction has been addressed in the past using various traditional computer vision and machine learning techniques. However, the deep learning techniques has shown better performance over traditional computer vision and machine learning techniques, as discussed by Garcia-Garcia et al. [4] in the topic of image segmentation. In addition, we believe that this is the first work that reviews deep learning for 3D reconstruction.

The remainder of this paper focuses on the process, algorithms and methods of learning-based 3D reconstruction techniques. Specifically, Section 2 discusses about the terminology and some background concepts. Section 3 discusses about methodology used for searching literature and for defining the research scope, also explains about datasets and methods. Section 4 discusses evaluation metrics and summary. Finally, we conclude this paper in Section 5.

II. TERMINOLOGY AND BACKGROUND CONCEPTS

A. 3D Reconstruction of 2D Images

3D reconstruction of 2D images aims to infer a 3D structure of an object from single or multiple image sources. Based on the 3D model representation, there are several approaches of 3D reconstruction. Following are types of the 3D representation. First, 3D structure of an object is represented as primitive assembly of simpler 3D objects. In the approach that uses this representation type, a more complex 3D object is defined as an assembly of simpler 3D objects, e.g., cube, cylinder, etc. Second, voxel representation. The main limitations of voxel representation are inefficient and less scalable. However, voxel representation is still mainly used in medical image analysis, e.g. representation of CT-scan data. Third, octree representation, a tree-like structure but more efficient than voxel since only represents surface of an object more detail without representing the object's interior. Fourth, point cloud representation, a more popular 3D representation that a collection of points in 3D space are used to define surface of an object. A point cloud can be defined as a uniform structure that allows simple manipulation and is easier to learn [5]. Last, mesh (surface patch) representation, a collection of nodes and connectivity information between nodes. 3D reconstruction methods are affected by the 3D representation discussed above.

In terms of input type and number of images, there are several variations of the 3D reconstruction problem, i.e, 3D reconstruction from single image (shape-from-X), multiple images (stereo), or 3D reconstruction from RGB-D data. This paper only considers the first and second type of problem, i.e. 3D reconstruction from single or multiple images.

Recovering 3D shape from single image is an ill-posed problem in computer vision because many 3D shapes may result in the same set of images. Meanwhile 3D reconstruction

978-1-7281-2133-8/19/$31.00 ©2019 IEEE

from multiple images, usually called stereo, may be performed by traditional or learning-based methods. Traditional methods for uncalibrated photometric stereo problem include image registration using *keypoint* or landmark detection or using optical flow to achieve pixel-wise correspondence between several images. Then, at each pixel location, the surface normal and the surface albedo is estimated using the image registration results. However, there are some limitations using this traditional method, for example, it is difficult to obtain dense and accurate correspondence at pixel level.

An example of 3D reconstruction from a single image is illustrated in Fig. 1. In Fig. 1, the first column is a sample input image, created by rendering the 3D ShapeNet dataset, the second column is the corresponding 3D model as the ground truth data, columns 3-7 are the 3D reconstruction results using several methods discussed in [6]. In this example, the reconstructed 3D model is represented using voxel. A slightly variation of 3D reconstruction from a single image is the 3D bas relief reconstruction from an image as in [7]. The sample input and output of the 3D bas relief reconstruction is as shown in Fig. 2. In the example, the color input image is converted into a grayscale image, then blurred, and the edges are extracted followed by edge thickening. The reconstructed 3D relief is created using high carving in the example shown in Fig. 2.

In addition, an example of 3D reconstruction from multiple images is as shown in Fig. 3.

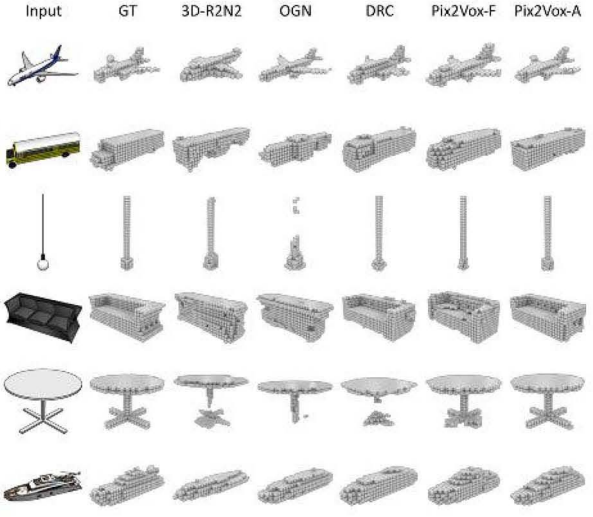

Fig. 1. Some illustrations of 3D reconstruction from a single image [6].

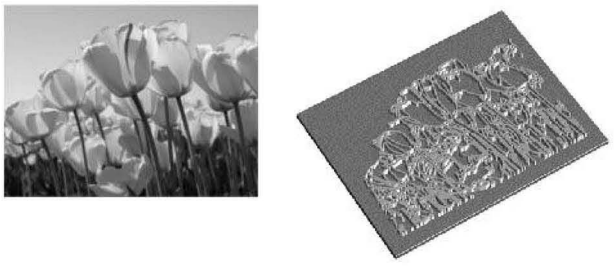

Fig. 2. An example of 3D bas relief reconstruction from a tulip image [7].

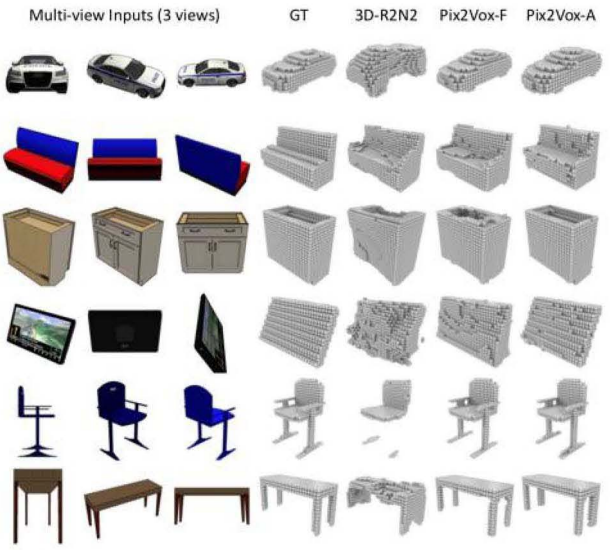

Fig. 3. Some illustrations of 3D reconstruction from multiple images [6].

B. Deep Learning Techniques

The main goal of deep learning techniques usually includes classifying objects by predicting an input, or if there are many solutions then it will provide a ranked list of possible solutions.

Currently, there are several widely known architectures that have made significant contributions to the computer vision field, e.g. AlexNet, VGG-16, GoogLeNet, and ResNet. AlexNet [8] is the first deep Convolutional Neural Network (CNN) that won ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012 with an accuracy of 84.6%, compared to the traditional techniques as the second winner with an accuracy of 73.8%. AlexNet consists of five convolutional layers, max-pooling, Rectified Linear Units (ReLUs) as non-linearities, three fully-connected layers, and dropout. Fig. 4 shows the AlexNet architecture [8].

Meanwhile, University of Oxford's Visual Geometry Group (VGG) created VGG model called VGG-16 submitted to the ILSVRC in 2013 and achieved 92.7% accuracy [9]. Fig. 5 illustrates the VGG-16 architecture. VGG-16 increases the performance of the model by employing smaller receptive fields in its first layers.

GoogLeNet was developed by Szegedy et al. [10], won the ILSVRC in 2014 with an accuracy of 93.3%. The architecture of GoogLeNet consists of 22 layers and inception modules, computed in parallel, consisting of a NiN layer, a pool operation, and convolution layers. All have 1x1 convolution operations to reduce dimensionality. Fig. 6 shows the inception module from the GoogLeNet architecture [10].

ResNet [11] by Microsoft won the ILSVRC in 2016 with 96.4% accuracy. The ResNet architecture is well-known because of its 152 layers and its residual blocks. The residual blocks introduced identity skip connections such that layers can copy their inputs to the next layer. Fig. 7 shows the residual block in a ResNet [11].

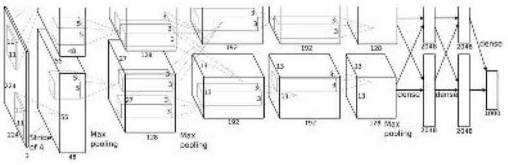

Fig. 4. AlexNet Architecture [8].

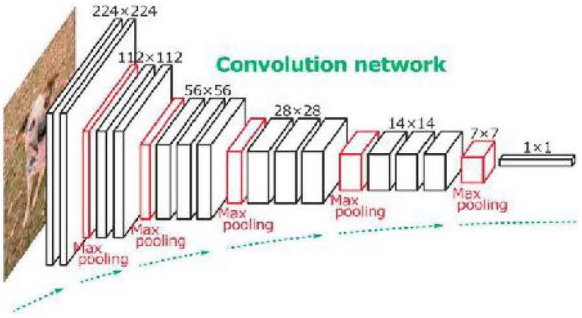

Fig. 5. VGG-16 Architecture [9].

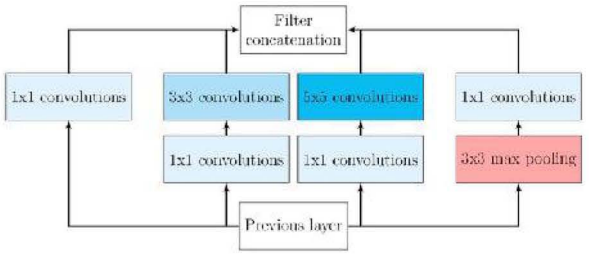

Fig. 6. Inception module from the GoogLeNet architecture [10].

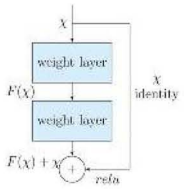

Fig. 7. The residual block in ResNet architecture [11].

III. Methodology

A. Literature Search Conditions

The research databases used in this review include those databases accessible via ScienceDirect website produced by Elsevier, the world's leading source for scientific, technical and medical research. Keywords used in this research consist of '3D reconstruction review', '"3D reconstruction" AND "learning"', '"3D" AND "deep learning"'. As the results of using the keywords mainly are not associated with deep learning-based 3D reconstruction methods, the results are then cleaned through reading the abstract to exclude those that are not related to learning-based methods. Based on this condition, there are eight journal or conference articles. In addition to the ScienceDirect, the ACM Digital Library (ACM DL) is also used as the source database of this research. There are ten articles resulted from the search conditions above in the ACM DL. Lastly, there are 37 related articles indexed in the IEEE Xplore website. In total, there are 55 research papers studied in this review. Several papers and its description are shown in Table 1.

B. Scope Definition

By analyzing research publications discussed in the previous section, the frequency of publications in terms of its type is shown in Table 2. Several publications use RGB-D images as the input. In this study, we later omit this publication type since we focus on RGB images as the input to the 3D reconstruction system. In terms of 3D objects, there are three types of specific 3D objects in the research scope, i.e. generic objects (synthetic or real), face, and MRI.

C. Datasets

This section describes the most popular large-scale datasets that are being used for 3D reconstruction. Table 3 shows the popular 3D reconstruction datasets.

D. Methods

Deep learning methods have main benefits in terms of its capabilities to learn features of a problem in an end-to-end fashion. Thus, the manual effort of selecting contributing features or some fine-tunings that need human expertise can be minimized [4]. However, one of the main concerns in 3D reconstruction of images is the format used for representing the 3D model. Following are methods used for 3D reconstruction based on the 3D representations.

1) Point Cloud

The forerunner of deep learning techniques used point cloud representation is PointNet [25], followed by PointNet++ [26]. Originally, they were used for point cloud classification, but a number of works have sprung up in this domain ranging from point cloud processing to reconstruction. The first approach in 3D reconstruction is the Point Set Generation (PSG) Network by Fan, Su, and Guibas [5]. The approach attempted to generate point cloud coordinates straight from the input without hand-crafted feature extraction. However, there is a condition that the ground-truth 3D object for a 2D input image may be ambiguous. During training phase, the Earth Mover's Distance (EMD) was used to measure loss. The network architecture has two prediction branches, one is flexible enough to capture complicated structures and the other ensures geometric continuity.

TABLE I. Selected Publications and Its Description

Ref No.	Application Types	Object Type	3D Representation	Dataset
[12]	Identification	Face	Parametric Model	3DFace (authors')
[13]	Generic	CAD	Triangular Mesh	ShapeNet
[14]	Generic	CAD	Voxel	ShapeNet, ObjectNet3D
[15]	Generic	CAD	Voxel Block Octree	ShapeNet-Core
[16]	Generic	CAD	Point Cloud	ShapeNet
[17]	Generic	CAD	Polygon Mesh	ShapeNet-Core
[18]	Generic	CAD	Voxel	Pascal3D+, Kitti
[19]	Generic	CAD	Point Cloud	ShapeNet, ObjectNet3D
[20]	Facial recognition, facial analysis, facial animation	Face	Parametric model	FRGC2, BU-3DFE, UHDB31

TABLE II. THE NUMBER OF PUBLICATIONS BASED ON ITS TYPE FROM 2016 TO 2019

Type of publications	2016	2017	2018	2019
Journal articles	1	2	8	10
Conference/workshops/symposium	3	12	16	2
Poster	0	0	1	0

TABLE III. THE MOST POPULAR DATASETS USED IN 3D RECONSTRUCTION RESEARCH

Dataset Name and Reference	Year	#categories	#images	#3D shapes	3D annotation type
ShapeNetCore [21]	2016	55	NA	51,300	2D-3D alignment
3DFace [12]	2019	NA	80,000	187,860	NA
ObjectNet3D [22]	2016	100	90,127	44,161	2D-3D alignment
Pascal3D+ [23]	2014	12	30,899	79	2D-3D alignment
KITTI [24]	2012	2	14,999	N/A	3D point

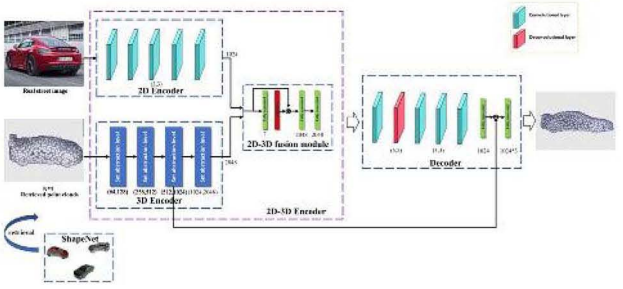

Fig. 8. The RealPoint3D architecture [19].

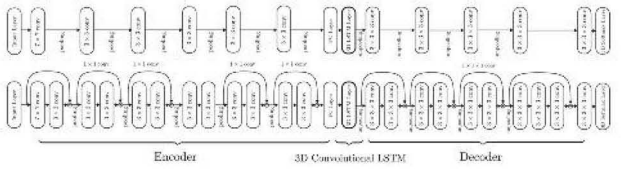

Fig. 9. The 3D-R2N2 architecture [27].

Currently, the method that attained state-of-the-art accuracy in 3D reconstruction from a single image that uses volumetric and point set generation methods is the RealPoint3D by Zhang et al. [19]. With a given 2D image, the approach generated point cloud data by projecting the pixel information into a 3D space, then computed a Chamfer distance and generated a projection loss between the generated point cloud data and the real point cloud data. To be more specific, in addition to the 2D image, there is an additional input to the network, i.e. the most similar point cloud from the ShapeNet dataset searched by some image similarity measurement. The network itself consists of an encoding part, a 2D-3D fusion module, and a decoding part. The encoding part extracts 2D features from the input image and 3D features from the input point cloud data. The 2D-3D fusion module generates both image and spatial features from the previous step. Lastly, the decoding part generates the object's predicted 3D point clouds. The RealPoint3D architecture is illustrated in Fig. 8.

2) Voxel

One significant method of 3D reconstruction that used voxel representation is 3D-R2N2 by Choy et al. [27]. 3D-R2N2 learns a mapping from images to their underlying 3D shapes from the ShapeNet dataset. The output of the network is in the form of a 3D occupancy grid. Before that, Gwak et al. [14] proposed weakly supervised reconstruction with adversarial constraint. However, in the Gwak et al. approach, we need to provide camera viewpoints as the input.

The 3D-R2N2 is an extension of the standard LSTM framework and can selectively update hidden representations by controlling input gates and forget gates [27]. Fig. 9 shows the 3D-R2N2 architecture. The loss function of the network is the sum of voxel-wise cross entropy [27].

3) Mesh

A polygon mesh can be used for representing a 3D model. A mesh consists of a list of vertices, edges, and surface normal of each face. The problem of 3D reconstruction from images using mesh representation may be compromised using additional input, i.e. a pre-defined mesh, as in [13] or [28]. To be more specific, it is a problem of learning a parametrization that transforms vertices in the input mesh where edges and surface normal of faces are conserved.

One of the recent methods in using mesh representation for 3D reconstruction is proposed by Pan et al. [28]. They proposed Residual MeshNet (ResMeshNet) that consists of stacked blocks of multi-layer perceptrons (MLPs). The initial block extracts shape features. The consecutive blocks receive a mesh (vertices coordinates combined with the shape features) and produce its deformed version. The output of each block is a set of meshes that approximate the target surface.

IV. DISCUSSION

A. Evaluation Metrics

1) Point Cloud

There are two distance metrics that can be used for point cloud representation: The Chamfer distance and the EMD. They are differentiable and can be used as the loss function [5].

2) Voxel

There are two metrics that can be used for voxel representation, i.e. the voxel Intersection-over-Union (IoU) between a 3D voxel reconstruction and its ground-truth, and the cross-entropy loss [27]. Higher IoU values but lower loss values indicate better results.

3) Mesh

Mesh connectivity can be evaluated by "Metro" criteria, which is an average Euclidean distance between two meshes. This metric was used in [13]. ResMeshNet [28] was trained using Chamfer distance (CD) based objective, that encourages the produced meshes to be consistent with the ground-truth meshes.

B. Summary

Voxel representation is costly in term of memory and is limited to coarser resolutions. This can be improved by using octree representation. Another approach to overcome this is by learning to encode and decode a 3D point representation. But this approach has no surface connectivity embedded. Mesh representation has surface connectivity embedded. However, there is a problem of combinatorial optimization in this case, using assumptions that vertices are sampled from the object's surface. Moreover, given an underlying variation, there may exist more than one way to build meshes.

Based on current works, the mesh-based reconstruction, ResMeshNet, has better performance compared to point cloud approach like the PSG or another mesh-based approach, i.e. the AtlasNet. The metrics used in the comparison was the mean-CD, where the values are 6.09 x 10^{-3}, 3.42 x 10^{-3}, and 3.23 x 10^{-3} for PSG, AtlasNet, and ResMeshNet respectively.

V. Conclusion

This paper is the first attempt to review the literature which focuses on 3D reconstruction using deep learning techniques. Compared to existing papers that review 3D reconstruction methods, this paper is dedicated to the rising of deep learning, covering the most advanced and state-of-the-art work in this field. Finally, we can conclude that 3D reconstruction has been approached with many deep learning techniques successfully, but still can be further enhanced due to the better solution would be very beneficial for many real-world applications. Moreover, the results of current implementation of deep learning in 3D reconstruction have shown good performance. Therefore, a further examination and many researches are required in this field.

References

[1] C. J. F. Reyneke, M. Lüthi, V. Burdin, T. S. Douglas, T. Vetter, and T. E. M. Mutsvangwa, "Review of 2-D/3-D Reconstruction Using Statistical Shape and Intensity Models and X-Ray Image Synthesis: Toward a Unified Framework," *IEEE Rev. Biomed. Eng.*, vol. 12, pp. 269–286, 2019.

[2] A. Paturkar, G. Sen Gupta, and D. G. Bailey, "Overview of image-based 3D vision systems for agricultural applications," *2017 Int. Conf. Image Vis. Comput. New Zeal.*, pp. 1–6, 2017.

[3] Z. Ma and S. Liu, "A review of 3D reconstruction techniques in civil engineering and their applications," *Adv. Eng. Informatics*, vol. 37, pp. 163–174, 2018.

[4] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Appl. Soft Comput.*, vol. 70, pp. 41–65, 2018.

[5] H. Fan, H. Su, and L. J. Guibas, "A Point Set Generation Network for 3D Object Reconstruction from a Single Image," *CoRR*, vol. abs/1612.0, 2016.

[6] H. Xie, H. Yao, X. Sun, S. Zhou, S. Zhang, and X. Tong, "Pix2Vox: Context-aware 3D Reconstruction from Single and Multi-view Images," *arXiv Prepr. arXiv1901.11153*, 2019.

[7] N. Suciati *et al.*, "Converting Image into Bas Reliefs Using Image Processing Techniques," *J. Phys. Conf. Ser.*, vol. 1196, p. 12037, Mar. 2019.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[9] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *CoRR*, vol. abs/1409.1, 2014.

[10] C. Szegedy *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] Y. Guo, j. zhang, J. Cai, B. Jiang, and J. Zheng, "CNN-Based Real-Time Dense Face Reconstruction with Inverse-Rendered Photo-Realistic Face Images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 6, pp. 1294–1307, Jun. 2019.

[13] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry, "A Papier-Mache Approach to Learning 3D Surface Generation," *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 216–224, 2018.

[14] J. Gwak, C. B. Choy, M. Chandraker, A. Garg, and S. Savarese, "Weakly supervised 3d reconstruction with adversarial constraint," in *Proceedings - 2017 International Conference on 3D Vision, 3DV 2017*, 2018, pp. 263–272.

[15] C. Hane, S. Tulsiani, and J. Malik, "Hierarchical surface prediction for 3D object reconstruction," in *Proceedings - 2017 International Conference on 3D Vision, 3DV 2017*, 2018.

[16] L. Jiang, S. Shi, X. Qi, and J. Jia, "GAL: Geometric adversarial loss for single-view 3D-object reconstruction," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11212 LNCS.

[17] H. Kato, Y. Ushiku, and T. Harada, "Neural 3D Mesh Renderer," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[18] A. Kundu, Y. Li, and J. M. Rehg, "3D-RCNN: Instance-Level 3D Object Reconstruction via Render-and-Compare," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3559–3568.

[19] Y. Zhang, Z. Liu, T. Liu, B. Peng, and X. Li, "RealPoint3D: An Efficient Generation Network for 3D Object Reconstruction From a Single Image," *IEEE Access*, vol. 7, pp. 57539–57549, 2019.

[20] P. Dou and I. A. Kakadiaris, "Multi-view 3D face reconstruction with deep recurrent neural networks," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 483–492.

[21] A. X. Chang *et al.*, "ShapeNet: An Information-Rich 3D Model Repository," *CoRR*, vol. abs/1512.0, 2015.

[22] Y. Xiang *et al.*, "ObjectNet3D: A Large Scale Database for 3D Object Recognition," in *European Conference Computer Vision (ECCV)*, 2016.

[23] Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond PASCAL: A Benchmark for 3D Object Detection in the Wild," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

[24] A. Geiger, "Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361.

[25] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-January.

[26] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," *CoRR*, vol. abs/1706.0, 2017.

[27] 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," *CoRR*, vol. abs/1604.0, 2016.

[28] J. Pan, J. Li, X. Han, and K. Jia, "Residual MeshNet: Learning to Deform Meshes for Single-View 3D Reconstruction," in *2018 International Conference on 3D Vision (3DV)*, 2018, pp. 719–727.