Proceedings of 9th IEEE International Conference
on CYBER Technology in Automation, Control, and Intelligent Systems
July 29- August 2, 2019, Suzhou, China

Object Reconstruction with Deep Learning: A Survey

Zishu Gao, En Li, Guodong Yang, Zhe Wang, Yunong Tian, Zize Liang, Rui Guo, Shengchuan Li

Abstract— **Object reconstruction is one of the most crucial branches of computer vision. With the development of deep learning, many tasks have achieved remarkable improvements in computer vision. 3D reconstruction with deep learning also has attracted much attention in recent years. Deep learning methods based on CNN-based and GAN-based architectures have been adopted for 3D object prediction. In addition, researchers utilize different inputs such as RGB and depth images to achieve prediction based on different problem. In this paper, we provide a detailed overview of recent advances in 3D object reconstruction. The reviewed approaches are categorized into three groups depending on the input modality: RGB-based, depth-based and other-input-based. Particularly, we introduce the various methods and indirectly classify the shape representation. As a survey, we discuss the strong and weak points of exciting approaches.**

I. Introduction

The ability to infer the accurate 3D model of an object is a fundamental problem for many potential applications, such as robot-object-manipulation [1], 3D printing and 3D acquisition [2], AR applications [3], object deformation [4] and semantic understanding. Object reconstruction focuses on acquiring 3D shape representation of an object from single or multiple images. Traditional methods for object reconstruction tends to use Structure-from-motion (SFM) and Simultaneous Localization and Mapping (SLAM), which achieve shape prediction depending on the correspondence of geometric features. However, these approaches always perform poorly in some cases, such as having little brightness information, having views with wide baseline and suffering from cumulative error. In recent years, despite of the impressive performance on diversity of computer vision tasks using deep learning, 3D object reconstruction has also been improved by deep neural networks.

In this paper, a comprehensive survey of object reconstruction using deep learning methods is presented. Regular methods are favored by deep auto-encoder architectures [5-19]. In addition, many researchers try to utilize generative adversarial network to learn latent space to infer 3D model shapes, such as GANs [20-23, 35, 36], Variational Autoencoders (VAEs) [26]. Since multiple-view object reconstruction can be viewed sequence tasks, it can be solved by recurrent neural network [24].

It is important to talk about the input modality while focusing on deep learning techniques. There are three important input modalities: RGB, depth and binary image. RGB images are the dominant input in most reconstruction tasks because of its abundant information [5-9, 11-13, 15-20, 32-34]. Along with the popularity of consumer-level RGB-D cameras, depth view is now widely available for many applications. So various methods appear using depth view as input [10,14, 21, 22]. Furthermore, [23] renders the images as binary images and takes the binary images into neural network.

Shape representation is also one of most important property of objects. Most extant approaches advocate voxel representation [5, 18-24, 27-30, 41, 42]. A few methods employ the unordered point cloud representation [7, 8, 15, 37, 44]. Moreover, other shape representations are also be explored such as mesh [6, 17], sketch [31], Octree [11, 12], Diserete Cosine Transform [38], multi-view maps [13, 39] and a set of cuboid part primitives [40].

As Figure 1 shown, our overview classifies three broad categories of approaches based on the input modality. The categories consists of RGB-based, depth-based and other-input-based. In RGB-based and depth-based category, two or three sub-divisions are further identified, namely CNN-based, GAN-based and RNN-based. Particularly, we identify the shape representation indirectly. Table 1 shows the comparison between different methods.

II. RGB-based Reconstruction with Deep Learning

In this section, we discuss various approaches for 3D object reconstruction with RGB inputs. RGB-based volume prediction is one crucial method because of rich information, including color, shape and texture. Based on modeling architecture, these techniques can be divided into three

* Zishu Gao is with The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation Chinese Academy of Sciences and The University of Chinese Academy of Sciences, Beijing, China. (e-mail: gaozishu2016@ia.ac.cn)

En Li is with The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation Chinese Academy of Sciences, Beijing, China. (corresponding author, phone: 86-10-82544783; e-mail: en.li@ia.ac.cn).

Guodong Yang is with The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation Chinese Academy of Sciences, Beijing, China. (e-mail: guodong.yang@ia.ac.cn).

Zhe Wang is with The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation Chinese Academy of Sciences and The University of Chinese Academy of Sciences, Beijing, China. (e-mail: wangzhe2016@ia.ac.cn)

Yunong Tian is with The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation Chinese Academy of Sciences and The University of Chinese Academy of Sciences, Beijing, China. (e-mail: tianyunong2016@ia.ac.cn)

Zize Liang is with The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation Chinese Academy of Sciences, Beijing, China. (e-mail: zize.liang@ia.ac.cn).

Rui Guo is with State Grid Shandong Electric Power Company, Jinan, China. (e-mail: guoruihit@qq.com).

Shengchuan Li is with State Grid Liaoning Electric Power Company Limited, Shenyang, China. (e-mail: lnlsc@163.com).

978-1-7281-0770-7/19/$31.00 © 2019 IEEE

categories: CNN-based approach, GAN-based approach, RNN-based approach.

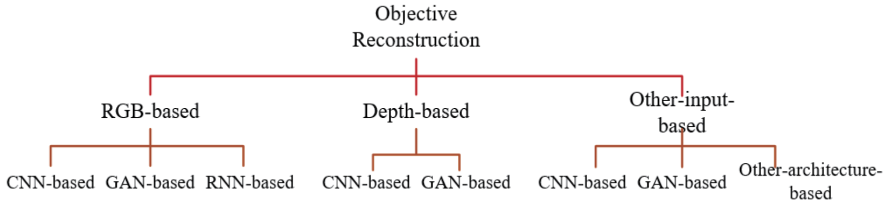

Figure 1. Categorisation of the methods for object reconstruction using deep learning

TABLE I. METHODS COMPARISON AMONG DIFFERENT WORK USING DEEP LEARNING.

	Reference	Method	Shape
RGB	[5]	Autoencoder+Perspective transfirmer	Voxel
	[6]	ConvNet(Autoencoder)	Point cloud
	[7]	MVPNet(Autoencoder)	Point cloud
	[8]	PointOut Network(Autoencoder)	Voxel or mesh
	[9]	CNN+projection layer	Skeleton
	[11]	CNN+decoder	Octree
	[12]	OctNet(convoilution+	Octree
	[13]	2DConvNets+	Point cloud
	[15]	Autoencoder+segmentation	Point cloud
	[16]	Autoencoder+embedding matching	Point cloud
	[18]	TL-embedding	Voxel
	[20]	3D-GAN	Voxel
	[24]	LSTM+CNN	Voxel
	[25]	Context-conditional general model+3D-2D projection	Voxel or mesh
	[26]	Stochastic Gradient VB	Voxel
	[28]	CNN+ray consistency	Voxel
	[29]	TL-embedding+VAE+GAN	Mesh
	[30]	RNN+GAN	Point cloud
	[31]	Autoencoder+reprojection consistency	Mesh
	[34]	Autoencoder+ Spatial Transformer Networks	Voxel
	[37]	CNN+Grid Deformation Unit	Voxel
	[40]	Autoencoder	Voxel
	[41]	Autoencoder + inverse Discrete Cosine Transform	Voxel
	[45]	Conv_gru+view planning	Voxel
Depth	[21]	GAN+U-net+upsampling	Voxel
	[22]	GAN+U-net	Voxel
	[27]	Matching+deformation	Mesh
	[38]	LSTM+ a Mixture Density Network (MDN)	Primitive
	[46]	CNN+projection	Point cloud
Binary image	[23]	Projective GAN	Voxel
sketch	[32]	Encoder + multi-view decoder	Mesh

A. *CNN-based approach*

For this group of approaches, there are mainly four methods to encode RGB information. First, many researches apply 2D and 3D convolutional layers to extract features and predict a volumetric representation. Yan et al [5] introduces auto-encoder network to infer a volume. As shown in Figure 2, it feeds single image into 2D encoder to extract spatial information, and leverages 3D convolutional layers with 3D kernels to build a volume generator. Furthermore, a plane loss function based on perspective transformation is utilized to project the 3D volume to 2D silhouette, which make it possible for the network to train without 3D ground truth data. The advantage of this method comes from its ability of learning an end-to-end network by providing 2D observations only which is time saving. [47] introduces view planning model to decide which view will be fed into network at each step based on [5]. In addition, [10] also uses Next-best-view for the same purpose as [47]. Compared to this method, [47] makes a better performance on reconstruction accuracy. [25] also adds 3D-2D projection mechanism to general network with learnable parameters, but [5] is parameter-free.

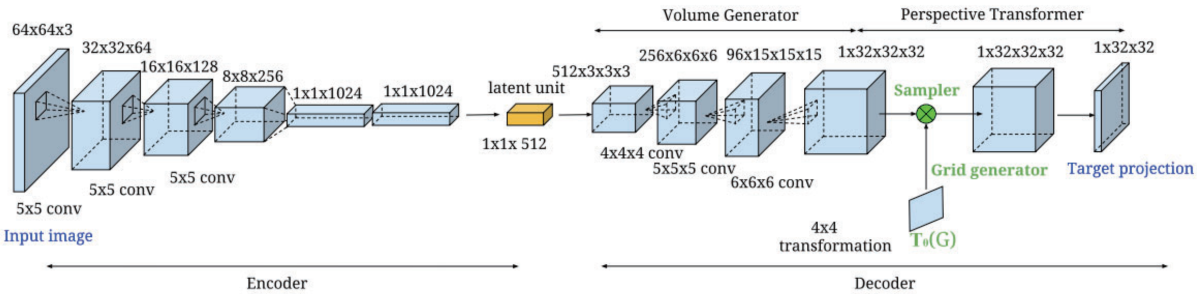

Figure 2. Auto-encoder network. Figure from Yan et al[5]

Tatarchenko et al[6] proposes a ConvNet network predicting the depth maps and unseen views of the object from a random RGB image. Then, the network aggregates multiple depth maps together to produce a full point cloud and turns it to be a mesh representation. This work has an elegant architecture network which is simpler than usual networks. The simple network first achieves reconstruction utilizing natural images without non-homogeneous background. This method shows the efficiency of reconstruction using rich features and desired viewpoints. But for real images, it is still hard to get explicit inferring since the training data cannot provide enough variations appearing in natural images.

Wang et al[7] exploits multi-view point regression network to generate a set of view-dependent point clouds by computing 3D coordinates and visibilities of points from an arbitrary image, then the various point clouds form a 3D surface of an object. Figure 3 shows that they incorporate camera parameters into the auto-encoder network (MVPNet network) and decode them into multiple point clouds representation. These point clouds being embedded in 2D rigid makes it easy to feed into CNN-based architecture. It is also worth to noting that the geometric loss integrating variance over 3D surfaces rather than 2D silhouette improves the reconstruction performance.

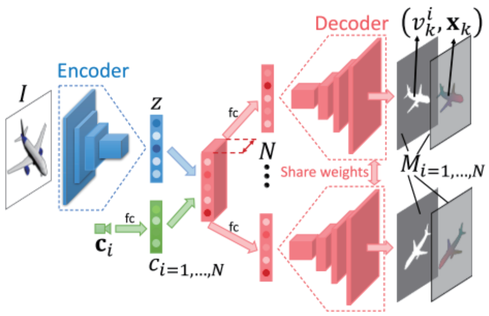

Figure 3. MVPNet network. Figure from Wang et al[7]

The above methods all fuse the RGB input images together into volumetric or mesh representation, which is efficient to represent the continuous 3D shapes. However, they also need to tackle the problem of obscure invariance of 3D reconstructions under geometric transformations. Fan et al [8] explores a simpler point set representation with generative network from a single RGB image. There is the auto-encoder network (PointOut Network) in this work. The encoder feeds the input image and random vector into a latent space. The Predictor has two branches, one is fully-connected branch, which is used to capture intricate structures, and the other is deconvolution branch, which generates smooth 3D surfaces by exploiting spatial continuity. The experiments show that fully-connected branch is better at predicting detailed components of an object, and deconvolution branch performs good on reconstruction of the overall shapes. This network has many excellent characteristics such as simple and easy to learn.

Wu et al[9] applies CNN to build a 3D interpreter network(3D-INN), an end-to-end network which infers a skeleton representation from a single annotated image. The overall network mainly has three parts. The keypoint estimator maps the input image to the heatmaps of 2D keypoints, which is a latent unit connecting real and synthetic data. The 3D interpreter predicts 3D skeleton and viewpoint parameters from the heatmaps of 2D keypoints. This method has the breakthrough that it use annotated image instead of 3D structural models as supervision, which is more adaptable and productive when reconstruct real image lack of 3D object annotations.

B. GAN-based approach

Besides the commonly used CNN-based methods for 3D object reconstruction, there are some GAN-based architectures that are adopted for this work. Wu et al [20] proposes 3D Generative Adversarial Network (3D-GAN) for 3D object generation from images. As shown on Figure 4, the vanilla network is composed of two parts, one is generator which maps the low-dimensional latent space to a final 3D voxel occupancy map, the other one is adversarial discriminator that aims to distinguish whether the predicted 3D shapes are real or not. The overall network is used 3D convolutional layers with 3D filters, and the discriminator mostly mirrors the generator. This architecture has many benefits compared to auto-encoder based approaches. First, the generator is able to extract object information and infer implicit volume due to the utilization of adversarial criterion. In addition, the discriminator helps to learn without supervision, which has promising application on not only reconstruction but also object recognition.

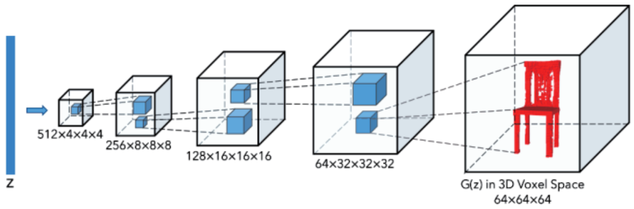

Figure 4. The generator in 3D-GAN.The discriminator mostly mirrors the generator. Figure from Wu et al[20].

C. *RNN-based approach*

Recurrent network are well developed for sequence-based problems. Inspired of this, Choy et al [24] designs a unified approach using LSTM for single or multi-view object reconstruction. The architecture of 3D Recurrent Reconstruction Neural Network (3D-R2N2) consists of an encoder, a 3D-LSTM (Long Short-Term Memory) and a decoder. The images are fed into the encoder to capture compressed features. The 3D-LSTM takes in the features captured by encoder and retains previous observations. Input gates and forget gates in 3D-LSTM help to selectively update hidden representations. The decoder establishes a mapping from the hidden states to 3D volume prediction. The method give an impressive performance on dealing with images which have insufficient texture or wide baseline viewpoints. Since the method can take a sequence of images into the network and remember previous observations using memory cells in 3D-LSTM, it overcomes the challenge of object self-occlusion. [43] uses 3D-R2N2 as baseline network and replaces 3D deconvolutional decoder with inverse discrete cosine transform (IDCT) decoder. This work gets high resolution volume with 128^3 compared with 32^3 in [24]. For high resolution, [21] reconstructs 3d shape by volumes of at 256^3 using 3D-RecGAN++, which will be introduced in detail later.

III. Depth-based Reconstruction with Deep Learning

Along with the development of the inexpensive 2.5D depth sensors, object reconstruction have gained remarkable success using 2.5D depth views as input with deep learning. We introduce different methods for 3D object shape prediction with depth-view inputs. Based on network architecture, these approaches can be divided into two categories: CNN-based approach and GAN-based approach.

A. *CNN-based approach*

For CNN-based method, Wu et al[10] proposes a 3D-ShapeNets. They consider the object shape as a probability distribution of binary variables on a 3D voxel grid. As Figure 5 shown, Convolutional Deep Belief Network (CDBN) is leveraged to learn the joint distribution of 3D volume. Given an arbitrary depth map of an object, 3D-ShapeNets maps it to a 3D volume representation. In addition, it also achieves recognizing object category and predicting the next best view in the case of uncertain initial recognition. It is the state-of-the-art reconstruction method in recent years.

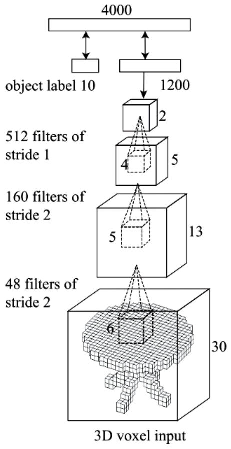

Figure 5. 3D-ShapeNets model architecture. Figure from Wu et al[10]

B. *GAN-based approach*

Yang et al[22] designs a network combining auto-encoder and conditional GAN network (3D-RecGAN) to reconstruct a 3D volumetric shape from a single random 2.5D depth view. [21] makes some improvements on high resolution shape prediction based on [22]. The generator is composed of a skip-connected auto-encoder and an up-sampling module. The encoder in the generator converts the depth view into compressed latent representation and the decoder maps the latent space to the plausible 3D volume. The skip-connection between the encoder and decoder is devoted to remember high frequency information. The conditional discriminator is utilized to classify real and synthetic object. The method has the advantage over other methods is that it generate a high resolution of 256^3 volumetric representation by recovering the missing regions.

IV. Other-input-based Reconstruction with Deep Learning

There are some other architectures have been proposed besides the commonly used RGB-based and Depth-based inputs for object reconstruction.

2D binary images are fed into projective generative adversarial networks (PrGANs)[23]. The PrGAN architecture is shown as Figure 6, which consists of generator, projection module and discriminator. 3D shape generator establishes the map from binary images to 3D volume. Given a viewpoint, the 3D volume is rendered by projection module to generate an image. The discriminator classifies whether the input image is real. The crucial strength is that it achieves making 3D reconstruction in an unsupervised manner. Compared to 3D-GAN [20], PrGANs gets better reconstruction results on complex objects such as airplanes.

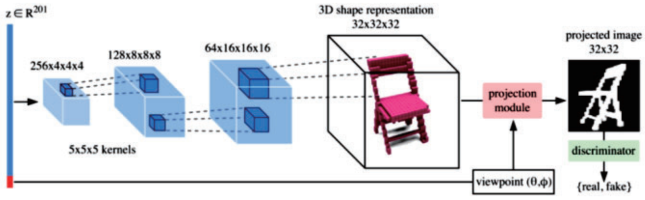

Figure 6. The PrGAN architecture. Figure from Wu et al[23]

V. Conclusion

In this paper, we introduce an extensive survey of object reconstruction using deep learning. We divide available approaches into three categories based on modality: RGB-based, depth-based and other-input-based. Every category is classified according to the deep learning method used. Their advantages and limitations are overviewed in each category. In addition, we group the methods based on shape representation indirectly. From the reviewed methods, we can see that volumetric representation is now dominant due to its flexibility during convolutional operations, but it faces the shortcoming of computational complexity for dense sampling. Point cloud and mesh prediction could be obtained through good mapping and they are trivial to optimization. However, the unordered property can lead to sparse shape prediction. Octree make it possible to predict high-resolution model by reducing the memory. Based on the insights drawn from the overview, we hope this survey will provide valuable viewpoints for the object reconstruction and encourage new idea in the future.

Acknowledgment

This work was supported by the National Key R&D Program of China (2018YFB1307400) and the National Natural Science Foundation of China (61873267, U1713224).

References

[1] J. Varley, C. DeChant, A. Richardson, J. Ruales and P. Allen, "Shape completion enabled robotic grasping," *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Vancouver, BC, 2017, pp. 2442-2447.

[2] S.Choi, Q.Y. Zhou, S. Miller, V. Koltun: A large dataset of object scans. arXiv preprint arXiv:1602.02481

[3] A. Sharma, O. Grau, and M. Fritz. VConv-DAE : Deep Volumetric Shape Learning Without Object Labels. *In European Conference on Computer Vision* , Springer, Cham, 2016, pp. 236-250.

[4] Z. Wang et al., "DEFO-NET: Learning Body Deformation Using Generative Adversarial Networks," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, QLD, 2018, pp. 2440-2447.

[5] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. *In Advances in Neural Information Processing Systems*, pages 1696–1704, 2016.

[6] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Multi-view 3d models from single images with a convolutional network. *In European Conference on Computer Vision*, Springer, 2016, pp. 322–337.

[7] J. Wang, B. Sun, and Y. Lu, MVPNet: Multi-View Point Regression Networks for 3D Object Reconstruction from A Single Image. arXiv preprint arXiv:1811.09410.

[8] H. Fan, H. Su, and L. J. Guibas, A point set generation network for 3d object reconstruction from a single image. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 605-613.

[9] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single Image 3D Interpreter Network. *In European Conference on Computer Vision (ECCV)*, 2016, pp. 365-382

[10] Zhirong Wu et al., "3D ShapeNets: A deep representation for volumetric shapes," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 1912-1920.

[11] M. Tatarchenko, A. Dosovitskiy and T. Brox, "Octree Generating Networks: Efficient Convolutional Architectures for High-resolution 3D Outputs," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017, pp. 2107-2115.

[12] G. Riegler, A. O. Ulusoy and A. Geiger, "OctNet: Learning Deep 3D Representations at High Resolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 6620-6629.

[13] C. H. Lin, C. Kong, and S. Lucey, Learning efficient point cloud generation for dense 3D object reconstruction. *In Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[14] D. Shin, C. C. Fowlkes and D. Hoiem, "Pixels, Voxels, and Views: A Study of Shape Representations for Single View 3D Object Shape Prediction," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 3061-3069.

[15] K. Li, R. Garg, M. Cai, and I. Reid, Optimizable Object Reconstruction from a Single View. arXiv preprint arXiv:1811.11921.

[16] P. Mandikal, N. Murthy, M. Agarwal, and R. V. Babu, 3D-LMNet: Latent Embedding Matching for Accurate and Diverse 3D Point Cloud Reconstruction from a Single Image. arXiv preprint arXiv:1807.07796.

[17] A. Kar, S. Tulsiani, J. Carreira and J. Malik, "Category-specific object reconstruction from a single image," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 1966-1974.

[18] R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta, Learning a predictable and generative vector representation for objects. *In European Conference on Computer Vision (ECCV)*, Springer, Cham, 2016, pp. 484-499.

[19] A. Dai, C. R. Qi and M. Nießner, "Shape Completion Using 3D-Encoder-Predictor CNNs and Shape Synthesis," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 6545-6554.

[20] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum, Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *In Advances in neural information processing systems*, 2016, pp. 82-90.

[21] B. Yang, S. Rosa, A. Markham, N. Trigoni and H. Wen, "Dense 3D Object Reconstruction from a Single Depth View," *in IEEE Transactions on Pattern Analysis and Machine Intelligence.*

[22] B. Yang, H. Wen, S. Wang, R. Clark, A. Markham and N. Trigoni, "3D Object Reconstruction from a Single Depth View with Adversarial Learning," *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Venice, 2017, pp. 679-688.

[23] M. Gadelha, S. Maji and R. Wang, 3D Shape Induction from 2D Views of Multiple Objects, *2017 International Conference on 3D Vision (3DV)*, Qingdao, 2017, pp. 402-411.

[24] 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 European Conference on Computer Vision (ECCV), Springer, Cham, 2016, pp. 628-644.

[25] D. J. Rezende, S. A. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess, Unsupervised learning of 3d structure from images. *In Advances in Neural Information Processing Systems* , 2016, pp. 4996-5004.

[26] D. P. Kingma and M. Welling, Auto-Encoding Variational Bayes, arXiv preprint arXiv:1312.6114.

[27] J. Rock, T. Gupta, J. Thorsen, J. Gwak, D. Shin and D. Hoiem, "Completing 3D object shape from one depth image," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 2484-2493..

[28] S. Tulsiani, T. Zhou, A. A. Efros and J. Malik, "Multi-view Supervision for Single-View Reconstruction via Differentiable Ray Consistency," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 209-217.

[29] R. Zhu, H. K. Galoogahi, C. Wang and S. Lucey, "Rethinking Reprojection: Closing the Loop for Pose-Aware Shape Reconstruction from a Single Image," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017, pp. 57-65.

[30] J. Gwak, C. B. Choy, M. Chandraker, A. Garg and S. Savarese, "Weakly Supervised 3D Reconstruction with Adversarial Constraint," *2017 International Conference on 3D Vision (3DV)*, Qingdao, 2017, pp. 263-272.

[31] J. Wu, Y. Wang, T. Xue, X. Sun, B. Freeman, and J. Tenen- baum. Marrnet: 3d shape reconstruction via 2.5 d sketches. *In Advances in neural information processing systems*, 2017, pp.540–550.

[32] Z. Lun, M. Gadelha, E. Kalogerakis, S. Maji and R. Wang, "3D Shape Reconstruction from Sketches via Multi-view Convolutional Networks," *2017 International Conference on 3D Vision (3DV)*, Qingdao, 2017, pp. 67-77.

[33] X. Di, R. Dahyot, and M. Prasad, Deep Shape from a Low Number of Silhouettes, *In European Conference on Computer Vision (ECCV)*, Springer, Cham, 2016, pp. 251-265.

[34] E. Grant, P. Kohli, and M. V. Gerven, Deep Disentangled Representations for Volumetric Reconstruction, *In European Conference on Computer Vision (ECCV)* ,ECCVWorkshops, Springer, Cham, 2016, pp. 266-279.

[35] H. Huang, E. Kalogerakis, and B. Marlin, “Analysis and synthe- sis of 3D shape families via deep-learned generative models of surfaces,” *Computer Graphics Forum*, vol. 34, no. 5, pp. 25–38, 2015.

[36] P. Achlioptas, O. Diamanti, I. Mitliagkas, L. Guibas, Representation learning and adversarial generation of 3d point clouds, arXiv preprint arXiv:1707.02392, 2017

[37] K. Li, T. Pham, H. Zhan, and I. Reid. Efficient dense point cloud object reconstruction using deformation vector fields. *In Proceedings ofthe European Conference on Computer Vision (ECCV)*, 2018, pp. 497–513.

[38] C. Zou, E. Yumer, J. Yang, D. Ceylan and D. Hoiem, "3D-PRNN: Generating Shape Primitives with Recurrent Neural Networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017, pp. 900-909.

[39] J. Yang, S.E. Reed, M.H. Yang et al. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis, *Advances in Neural Information Processing Systems,* 2015, pp.1099-1107.

[40] C. Häne, S. Tulsiani and J. Malik, "Hierarchical Surface Prediction for 3D Object Reconstruction," *2017 International Conference on 3D Vision (3DV), Qingdao*, 2017, pp. 412-420.

[41] A. Johnston, R. Garg, G. Carneiro and I. Reid, "Scaling CNNs for High Resolution Volumetric Reconstruction from a Single Image," *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Venice, 2017, pp. 930-939.

[42] H. Kato, Y. Ushiku and T. Harada, "Neural 3D Mesh Renderer," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 3907-3916.

[43] Y. Yang, C. Feng, Y. Shen, and D. Tian. Foldingnet: In- terpretable unsupervised learning on 3d point clouds. arXiv preprint arXiv:1712.07262.

[44] X. Yan, J. Yang, E. Yumer et al, Learning volumetric 3d object reconstruction from single-view with projective transformations, *Neural Information Processing Systems*, 2016.

[45] X. Yang, Y. Wang, Y. Wang, B. Yin, Q. Zhang, X. Wei, and H. Fu, Active object reconstruction using a guided view planner. arXiv preprint arXiv:1805.03081.

[46] E. Insafutdinov, and A. Dosovitskiy, Unsupervised learning of shape and pose with differentiable point clouds. *In Advances in Neural Information Processing Systems*, 2018, pp. 2807-2817.