

Research on 3D Object Reconstruction Method based on Deep Learning

Xiaoyang Liu

Department of Mathematical science, Tiangong University, Tianjin, China

1911650218@Tiangong.edu.cn

Abstract. 3D reconstruction is a classic task in the field of computer graphics. More and more researchers try to replicate the success of deep learning in 2D image processing tasks to 3D reconstruction tasks, so 3D reconstruction related research based on deep learning has gradually become a research hotspot. Compared with the traditional 3D reconstruction methods that require precision acquisition equipment and strict calibration of image information, the 3D reconstruction method based on deep learning completes the matching of 2D images to 3D models through deep neural networks, and can reconstruct 3D models of various categories of objects from RGB images obtained by ordinary acquisition equipment in a large number and quickly. This paper introduces the state of the art of 3D voxel reconstruction, 3D points cloud reconstruction and 3D mesh reconstruction, respectively. According to the different representation methods of 3D objects, the 3D object reconstruction methods based on deep learning are classified and reviewed, the characteristics and shortcomings of existing methods are analyzed, and three important research trends are summarized.

Keywords: 3D Reconstruction; Theory of Multi-view Geometry; Point Cloud; Voxel; Grid.

1. Introduction

3D object reconstruction from limited input is an important research topic in computer vision and computer graphics. In order to facilitate computer processing, 3D objects are usually represented by some ways, including point clouds, voxels, hidden fields, and grids. The methods of 3D object reconstruction can be divided into multi-view geometry theory and deep learning. The method based on multi-view geometry theory generally needs many different views to reconstruct the overall structure of the object, but it does not need a large amount of data for training, and the preliminary result of reconstruction is generally pointing cloud. In the actual environment, the matching between views will be seriously affected by illumination and other factors, which will lead to inaccurate matching [1].

Although the traditional 3D reconstruction method has been widely used in production and life, the traditional machine learning has inevitable shortcomings due to the restriction of learning methods and learning equipment. For example, the traditional 3D reconstruction method needs more human resources to supervise. When facing multiple shape modification and generation tasks at the same time, it is difficult to accurately identify the geometric and topological structure differences of object shapes. When there are fine geometric details, the accuracy of 3D reconstruction is not high enough because global operation cannot be realized. It is difficult to model the structure number of 3D shapes due to the inability to output the relationship of individual parts. With the rapid development of deep learning methods, some researchers have combined it with traditional 3D reconstruction and achieved good results. Deep learning has brought profound changes and new challenges to 3D reconstruction at multiple levels, such as data processing, geometric inference, structural reasoning, and semantic understanding [1].

This paper mainly reviews the 3D object reconstruction methods based on deep learning, describes the main representation forms of 3D objects, and discusses the 3D object reconstruction methods based on multi-view geometry theory and deep learning. Because the 3D reconstruction method based on multi-view geometry theory lacks the ability of inference, and cannot solve the problem of 3D object structure completion with less input information, and the deep learning method has good

learning and inference ability, so the 3D reconstruction method based on deep learning will become the focus of future development.

2. The Representation of 3D Objects

The representation of 3D objects mainly includes point cloud, voxel, hidden field, grid and so on. Point cloud is composed of points uniformly distributed on the surface of 3D objects, which is the most direct and convenient way to represent 3D objects. It can be obtained by scanning the object with lidar or sampling on the existing 3D model, so it is also the easiest way to obtain 3D representation. However, the point cloud is irregular and discrete, and there is a problem of missing adjacent topological relationships between points. Therefore, it is necessary to transform the point cloud into an explicit grid representation in the post-processing step. In order to obtain the 3D voxel representation of a 3D object, the space in which the 3D object is located needs to be rasterized, and the smallest unit obtained by rasterization is called a voxel. A voxel is valid if it falls inside the object or intersects its surface and is usually assigned a value of 1, and a value of 0 if it falls outside the object. 3D voxel representations are usually divided into dense voxels and sparse voxels. Sparse voxel representation cannot describe 3D objects in detail, while dense voxel representation requires a large amount of storage resources. The latent field is an implicit representation of a 3D object, indicating whether any point in space is inside or outside the object, or the distance from the object's surface.

The hidden field can also be discretized into non-binary voxels, and the Marching Cubes algorithm is used to generate the grid structure. The latent field can generally be learned and represented by neural networks. The mesh representation of a 3D object can be regarded as an explicit graph representation. Compared with the other three representations, the mesh representation is the most realistic, which is also beneficial to the rendering of 3D objects. Figure 1 illustrates the common representation of 3D objects. From the perspective of representation, 3D object reconstruction is to realize the process from 2D representation to 3D representation, or from one kind of 3D representation to other 3D representation [2].



Fig 1. Representation of 3D objects [4]

3. Reconstruction Methods based on Deep Learning

3.1 3D Object Reconstruction based on Point Cloud

3D point cloud is a set of points that are sampled on the surface of an object according to a certain distribution, and the shape information of the object surface is represented by the set of these points. The goal of 3D point cloud reconstruction is to reconstruct the position coordinate matrix of the 3D point cloud from the input image. Unlike regular 3D voxels, 3D points clouds have irregular structures, so existing deep neural networks cannot be simply used to operate on 3D point clouds, especially convolutional neural networks. However, the storage form of 3D point cloud in computer is a two-dimensional matrix with the horizontal axis as the spatial coordinates and the vertical axis as the sampling points. This matrix form representation can be used for convolution operation. Compared with 3D voxels, the storage form of 3D point cloud has the advantages of small consumption of computing resources and memory resources, and can represent finer shape information and surface details through the increase of the number of sampling points [5].

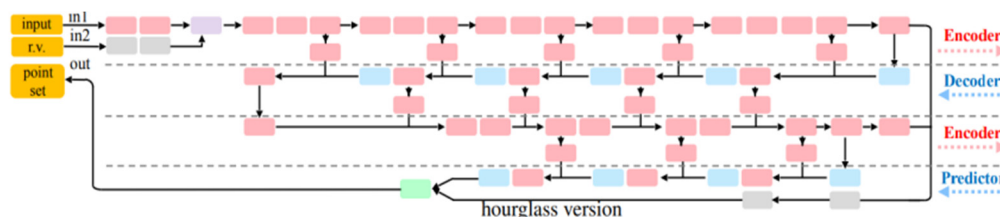


Fig 2. Network structure of PSGN [6].

Instead of directly generating voxel and mesh representations of 3D structures, Fan is the first to propose reconstruction methods from single views to 3D point clouds as shown in the figure 2. In addition, in order to solve the ill-posed problem of single-view 3D reconstruction, this method uses a variational autoencoder to generate possible 3D shapes. Yang proposed Folding Net, in which the grid points of two-dimensional unit element are transformed by two-stage multilayer perceptron to generate point clouds. However, the results of Folding Net are easily affected by the adjacent topological relationship of two-dimensional unit element grid points. Pang proposed Tearing Net, which can eliminate the influence of the topological relationship of the initial mesh points on the generated point cloud by clipping the mesh faces of two-dimensional unit elements. Adversarial loss has also been introduced into 3D object reconstruction in order to be able to generate 3D point clouds well. Yu first applied transformer to the completion problem of 3D point clouds and achieved very good results.

Compared with voxel representation, point cloud representation of 3D shape requires less storage resources. In contrast to the grid representation, do not consider the relationship between points; In contrast to the hidden-domain field representation, the step of subsequent explicit mesh generation is not required. Therefore, the point cloud representation of 3D shape is more convenient, and there are many application scenarios, such as autonomous driving. However, the point cloud representation of 3D objects is relatively rough, and the true adjacent relationship between points is difficult to determine. In the case of not enough points, the detail information of the object is also difficult to be expressed [6].

3.2 Voxel-based 3D Object Reconstruction

3D voxel is a 3D model representation that discretizes the surface and internal entities of 3D objects into voxel blocks and fills them into the 3D voxel grid. In simple terms, if a cell of the discretized grid is on the surface or inside of the original object, the cell is filled with voxel blocks, and if the cell is not on the interior or inside of the original object, the cell is left blank. The higher the degree of discretization, that is, the more voxel blocks, the finer and more accurate the corresponding 3D model can be represented. 3D voxel reconstruction is to reconstruct a voxel grid from an input image such that the 3D shape represented by the voxel grid is as close as possible to the real 3D shape of the corresponding object in the input image. 3D voxel reconstruction is the first 3D reconstruction to apply deep neural networks. The reason is that both the representation of 3D voxels and the representation of 2D rasterized images are regular and uniformly distributed. Therefore, in many existing deep neural network architectures specifically designed for 2D image analysis, the 2D pixel matrix can be simply replaced by a 3D voxel grid. Then, 3D convolution and pooling operations are used to process the 3D voxel data.

Yan proposed an unsupervised 3D voxel reconstruction method, which could input a 2D image and output the projection of its 3D structure from a certain view, and then predict the 3D voxel structure of the object by adding the constraint of equal input and output. Tulsiani proposed a similar approach to achieve the reconstruction of 3D voxels using multi-view constraints. Yang proposed a semi-supervised single-view 3D voxel reconstruction method. After adding limited pose information to the input image, the accuracy of 3D shape reconstruction was further improved. In general, it is difficult to obtain good object structure due to the uncertainty of shape prior information in single-view 3D reconstruction. Therefore, scholars try to unify single-view and multi-view 3D

reconstruction into a method framework. Kar proposed a multi-view stereo vision machine to improve the ability of the method to reconstruct object details. Choy processes the view sequence using a 3D recurrent neural network to obtain the voxels of the 3D object as shown in the figure 3 [6].

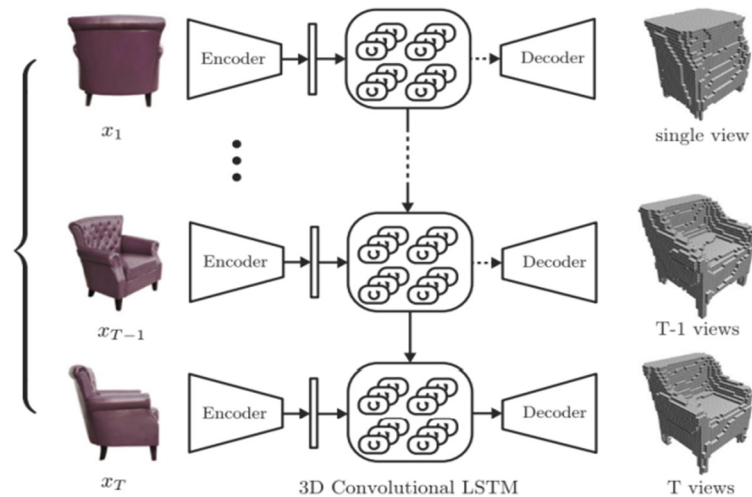


Fig 3. Network structure of 3d-r 2n 2 [7].

Häne proposed a hierarchical surface reconstruction method, which can obtain higher resolution voxel representation. In order to achieve high precision 3D object reconstruction with limited voxels, researchers have introduced new techniques. Wu used adversarial generative networks to add shape prior information to the voxel generation process of 3D objects, which well solved the ill-posed problem of single-view 3D voxel reconstruction. Ma proposed the separated channel spatio-temporal convolutional network, and introduced the attention mechanism into the reconstruction of 3D voxels for the first time. Yang proposed a 3D voxel reconstruction method based on memory mechanism. Firstly, the features of 3D voxels are pre-stored. Then, the voxel features similar to the input image were found and input into the long short-term memory network to obtain the fused features. Finally, this feature is used to decode the 3D voxel structure of the object. This memory-based method makes full use of the existing voxel information of 3D objects to optimize the reconstruction, but when the resolution of 3D voxels increases, this method also faces the problem of huge storage resource consumption.

3.3 Mesh-based 3D Object Reconstruction

Three-dimensional grid is based on three-dimensional point cloud, and every three specific and adjacent points are connected to form triangular surfaces. All triangular surfaces together form a closed surface model. However, three-dimensional grid reconstruction is not optimized based on three-dimensional point cloud reconstruction, because the reconstructed three-dimensional point cloud is disordered. If the disordered points are arranged in order and the connections are formed into triangular surfaces, there will be many defects such as staggered or missing triangular surfaces due to the imprecise correspondence between points. At present, the 3D mesh reconstruction generally adopts the deformation-based method. The vertex displacement is carried out on the preset geometric prototype, such as the two-dimensional plane or the three-dimensional sphere, to obtain the final reconstruction result. The advantage of deform-based method is that the geometric prototype already has topological relations between vertices, so it can keep the topological structure invariance when the vertices are shifted, so as to prevent the crossing and missing of triangular surfaces. However, the flexibility in deformation is given up because the topological structure needs to be maintained in the deformation process. As a result, the expression ability of the final reconstructed 3D mesh is still weaker than that of the 3D point cloud [8].

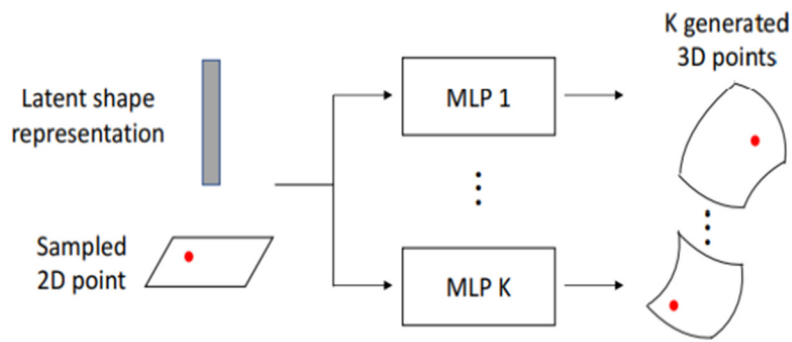


Fig 4. Overview of AtlasNet [11]

Table 1. The summary of 3D representation reconstruction [13]

Method	Number of views	Data category	Supervision mode	Deep learning training network	Output 3D representation	Data set
3D ShapeNets	Multiview	2D,3D,5D	Supervised	Deep Belief Networks	Voxel	ModelNet
TL	Multiview	2D	Supervised	autoencoder	Voxel	IKEA
3D-GAN	Multiview	2D,3D	Supervised	VAE-GAN	Voxel	SUN
MarrNet	Single view	2D,3D	Supervised	TL,3D-VAE+Encoder-decoder	Voxel	ShapeNet, PASCAL
SurfaceNet	Multiview	2D,3D	Supervised	Encoder-decoder	Voxel	DTU
DeepMVS	Multiview	2D	Supervised	MVS+CNN	Voxel	ETH3D
MVSNet	Multiview	2D	Supervised	MVS+CNN	Voxel	DTU
PointMVSNet	Multiview	2D	Supervised	MVS+CNN	Point cloud	DTU
IM-NET	Single view	2D,3D	Supervised	VAE+GAN	Implicit surface +Point cloud	ShapeNet, HSP
PF-Net		3D	Supervised	MLP+GAN	Point cloud	ShapeNet
Foldingnet	Multiview	2D	Unsupervised	Encoder-decoder ;	Point cloud	ShapeNet
AtlasNet	Single view	2D	Supervised	MLP	Grid	ShapeNet
Pixel2mesh	Single view	2D	Supervised	Figure CNN	Grid	ShapeNet
Pix2Vox	Multiview	2D	Supervised	Encoder-decoder	Voxel	ShapeNet
IF-Nets		3D	Supervised	Encoder-decoder	Implicit surface	3DBodyTex
3D-R2N2	Multiview	2D	Supervised	LSTM,Encoder-decoder	Voxel	PASCAL VOC
VoxNet	Multiview	2D	Supervised		Encoder-decoder	ModelNet
MVCNN	Multiview	2D	Supervised	Encoder-decoder	Voxel	ModelNet
OGN	Multiview	2D,3D	Supervised	Octree generating network	Voxel	ShapeNet

If the resolution of 3D voxels is increased, the voxel representation of 3D shape will consume a large amount of storage resources and processing time, and the point cloud and hidden field representation will eventually be transformed into grid structure to represent the surface structure of the object. The mesh representation of 3D shape can not only combine with texture to represent the surface structure of real objects, but also greatly reduce the demand for storage resources and speed up the processing of 3D shape. Therefore, grid-based 3D object reconstruction has become a hot research topic.

Kar is the first learning-based method to make progress in single-view 3D reconstruction, but the output 3D mesh structure is relatively coarse. Kato proposed a neural network renderer from visual images to 3D meshes (Neural 3D Mesh Renderer). Starting from a spherical mesh with 642 points, the final mesh tends to the structure of the object by changing the position of each point on the mesh through constant iterations. In the iterative process, the generated mesh is rendered to the image and compared with the real rendering image, so that the rendered image is as consistent as possible with the real image. Wang proposed an iterative model named Pixel2Mesh, which starts from a spherical mesh with 156 points, uses graph convolutional network and perceptual feature pooling technology to calculate the new features and positions of each point, and uses the de-pooling technology to continuously increase the number of vertices in the mesh, and achieves good reconstruction results. Wang also proposed a network similar to 3D deformation, which can start from any network and generate a target mesh through continuous deformation.

The topological structure of most objects is different from the original mesh. In this case, the detail information of 3D object reconstruction based on mesh may be seriously wrong. Therefore, it is worth studying how to make the topology of mesh adaptive in 3D reconstruction based on mesh [9].

Groueix proposed AtlasNet in order to address topological changes in mesh reconstruction, as shown in the figure 4. The network consists of several multilayer perceptron's, and each multilayer perceptron is responsible for generating a part of the point cloud structure of the 3D shape. The experimental results show that AtlasNet achieves very good point cloud and mesh reconstruction results [10]. However, the generated region corresponding to the multilayer perceptron in Atlas-Net has uncertainty, and there is discontinuity between the generated mesh blocks. We summarize the 3D representation reconstruction method in the following table 1, which includes the training process, the output 3D representation, and the dataset.

4. Future Directions

The powerful representation learning ability and geometric reasoning ability of 3D deep learning bring substantial promotion to 3D reconstruction or restoration based on single-view images or incomplete geometric data. At present, there are roughly two mainstream methods:

First, discriminative models based on geometric reasoning: training an end-to-end neural network to directly map input images or geometric data to the target 3D model.

The second is to train a deep generative model to learn the shape space of 3D objects, and then embed the input image or geometric data into the shape space based on metric learning. Finally, the 3D model is decoded from the embedding vector, to realize the 3D reconstruction of the input.

At present, there are many public 3D datasets in the world, covering single objects and indoor and outdoor scenes, but most of them are created by foreign teams. The domestic contribution in 3D datasets needs to be strengthened, and the ability of 3D deep learning will be limited in the case of insufficient training data. Therefore, constructing sufficient and accurate 3D datasets is a very important task [12].

5. Conclusion

According to the different representation methods of 3D objects, this paper classifies and summarizes the 3D object reconstruction methods based on deep learning, analyzes the characteristics and shortcomings of existing methods, and summarizes three important research trends. This survey surveys the progress in reconstructing 3D shapes of generic objects using deep learning. In the general trend, the field of 3D reconstruction is gradually moving towards commercialization and practicality. The requirements for real-time reconstruction, reconstruction quality, and realism of motion and rendering are becoming higher and higher. At the same time, the 3D reconstruction of human body in the simple indoor environment gradually transitioned to the 3D reconstruction in the complex outdoor environment. The equipment used is gradually simplified, from multiple expensive cameras

to monocular cameras, and then to consumer-grade monocular cameras and even mobile cameras. At the same time, the reconstruction goal develops from single target to multiple targets. Transfer learning has been successful in 2D image representation, but how to apply these techniques to 3D domain with less data structure is still unclear. This should motivate further research on new tasks designed specifically for 3D data. Therefore, a practical direction for future research is to combine the techniques of deep learning and transfer learning to improve the universality of the latter.

References

- [1] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.
- [2] William E. Lorensen, Harvey E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. *Acm sig-graph computer graphics*, 1987, 21(4), 1-11.
- [3] YU Qian, LU Jinxiao, BAI Jiquan, Fan Honghui. Review of three-dimensional object reconstruction methods based on deep learning. *Journal of Jiangsu Institute of Technology*, 2022, 28(04):31-41.
- [4] He Xiao, Zhu Mingrui, et al. BiTGAN: bilateral generative adversarial networks for Chinese ink wash painting style transfer. *Science China Information Sciences*, 2022,66(1): 239-248.
- [5] Mahad Fairuz Safwan, et al. Learning Pyramidal Feature Hierarchy for 3D Reconstruction. *IEICE Transactions on Information and Systems*, 2022, E105. D (2).
- [6] Yang G D, Cui Y, Belongie S, et al. Learning single view 3D reconstruction with limited pose supervision, *European Conference on Computer Vision*, 2018, 2018:90-105.
- [7] Peng S Y, Niemeyer M, Mescheder L, et al. Convolutional occupancy networks, *European Conference on Computer Vision*, 2020: 523-540.
- [8] Liu S L, Guo H X, Pan H, et al. Deep implicit moving least- squares functions for 3D reconstruction, *2021 Conference on Computer Vision and Pattern Recognition*, 2021:1788-1797.
- [9] Wang P S, Liu Y, Tong X. Deep octree-based CNNs with output-guided skip connections for 3D shape and scene completion, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020:1074-1081.
- [10] Yang J Y, Alv J M, Liu M M. Self-supervised learning of depth inference for multi-view stereo. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021:7522-7530.
- [11] Park J J, Florence P, Straub J, et al. DeepSDF: learning continuous signed distance functions for shape representation. *2019 Conference on Computer Vision and Pattern Recognition*, 2019:165- 174.
- [12] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 2017, 60 (6):84-90.
- [13] LI Mingyang, Chen Wei, Wang Shanshan, et al. Review of 3D reconstruction methods for visual deep learning. *Computer Science and Exploration*, 2022, 1-26.