A review of 3D reconstruction methods based on deep learning

Yuanchun Wang

School of Computer and Artificial Intelligence, Wuhan University of Technology, Wuhan 430063, China

wangyuanchun@whut.edu.cn

Abstract. In computer vision, an important research area is three-dimensional reconstruction. Using computer technology to reconstruct three-dimensional models of objects has become an indispensable part of in-depth research in many fields. This thesis presents the development process of 3D reconstruction methods that use deep learning. Compared with traditional methods, the 3D reconstruction method based on deep learning has more flexible input and output and higher efficiency. This thesis classifies the methods by the type of 3D model representation and discusses different frameworks for 3D reconstruction based on deep learning. With the introduction of the method NeRF (Neural Radiance Field), the three-dimensional reconstruction work based on deep learning has got a great development. NeRF can achieve good results in a very short period of time in the face of various complex scenes. With the continuous improvement of NeRF by researchers, this method has achieved more amazing results. Finally, the existing problems in the field of 3D reconstruction, the causes of problems and possible solutions are analyzed. Finally, the future development trend and direction of this field are hypothesized and discussed.

Keywords: Three-Dimensional Reconstruction, Deep Learning, Computer Vision, Neural Radiance Field.

1. Introduction

Three-dimensional reconstruction is a process that abstracts all kinds of three-dimensional information of objects from two-dimensional images, such as the shape, material, lighting, etc., of an object obtained from a group of photos, and then the real object can be rendered in the virtual world through this information. 3D reconstruction is one of a most important research directions in computer vision. Using computer to reconstruct 3D models of objects has become an indispensable part of in-depth research in many fields. In the medical field, the use of three-dimensional models to diagnose physical conditions; In the field of history and culture, cultural relics are reconstructed in three dimensions for scientific research and tourists to visit. In addition, 3D reconstruction technology has important application prospects in the fields of game development, industrial design, aerospace and navigation.

The origin of three-dimensional reconstruction technology dates back to 1963, when Roberts [1] suggested that computer vision could be used to extract three-dimensional information from two-dimensional images. Since then, three-dimensional reconstruction methods have been developed very rapidly in various fields. Especially with the rise of artificial intelligence technology in the past 10 years, 3D reconstruction work has been further developed based on deep learning methods.

There are many kinds of representation of 3D model, such as voxel model, grid model, point cloud model, occupation field, surface light field, signed distance field and so on. Similarly, for the three-way reconstruction method of deep learning, we can also divide it according to the representation method of the final model, and the narrative order is arranged by time, which is successively voxel, point cloud, grid, and neural radiance field.

2. 3D reconstruction algorithm based on deep learning

2.1. 3D reconstruction based on voxels

A voxel model is a 3D image reconstruction method that uses small cubes to represent the shape and color of an object. It is one of the earliest representation that was used for in deep learning 3D reconstruction. Depending on how many images are used as input, voxel-based 3D reconstruction can be classified into two categories: single-image and multi-image input.

For 3D reconstruction of a single image as input, the 3D model is usually reconstructed using a network of encoder decoder structures. For example, in 2016, Choy et al. [2] designed a 3D LSTM network based on LSTM to process the encoded information of a single image. In 2019, Yang et al. [3] used GAN to improve the reconstruction network of voxel model, but the depth map needed to be input at the same time, which increased the difficulty of obtaining input information.

The decoder of the 3D reconstruction network based on the voxel model is usually composed of 3D convolution, and the required memory and calculation requirements are equal to the voxel model’s resolution, which usually requires a larger memory, so the resolution of the reconstructed voxel model is low. In order to improve the voxel resolution, in 2017, Tatarchenko et al. [4] used 3DCNN to output the feature map, decode the feature map into an octree, and then gradually refine the low-resolution structure to high resolution.

For 3D reconstruction with multiple images as input, early methods used recurrent neural networks to fuse image features. For example, Choy et al. [2] processed multiple images in sequence, used 3D-CNN decoding to reconstruct voxel models, and carried out 3D reconstruction of multiple images based on image feature fusion. In 2019, Xie et al. [5] used VGG network to encode different views respectively, 3D convolutional decoding to obtain corresponding rough models, and then used context attention module for feature fusion to obtain the final voxel model.

2.2. Three-dimensional reconstruction based on point cloud

Point cloud is a different method from voxel model for 3D image reconstruction. It uses points with 3D coordinates, colors, and other attributes to represent the surface of an object. This method can produce smoother shapes and use less memory for the 3D reconstruction network. Point cloud based 3D reconstruction of a single image usually uses encoder-decoder structure network to reconstruct the point cloud model.

In 2017, Fan et al. [6] used full connection and deconvolution as decoders after image coding, and used chamfer distance and EMD (the earth mover’s distance) as indicators of loss function to reconstruct three-dimensional model in the form of point cloud. In 2019, Mandikal et al. [7] used full connection as a decoder, established loss function based on EMD to reconstruct sparse point cloud, then used MLP to extract point cloud features, and used chamfer distance as an indicator of loss function to conduct intensive reconstruction of the initial sparse point cloud to (obtain the point cloud model of objects.

In 2018, Jiang et al. [8] jointly generated adversarial losses and multi-view consistent losses, and reconstructed point cloud models using GAN networks. Other researchers combine different loss functions to design networks reconstructed by point cloud models of single images. For example, Mandikal et al. [7] use point cloud auto encoders to learn the potential space of three-dimensional point clouds. The image encoder maps the 2D image to the potential space in a probabilistic way, extrapolates multiple 3D reconstruction models, and combines matching loss and diversity loss reconstruction point cloud models.

2.3. Three-dimensional reconstruction based on grid

Grid model is the most common form of three-dimensional model representation. Compared with voxel model and point cloud model, grid model can represent the surface shape of objects more completely. Therefore, grid-based three-dimensional reconstruction work is the most extensive, which can also be divided into single image input and multiple image input. Single image input can be divided into four types based on multi-stage network, template, implicit function and basis differentiable rendering, while multi-image input can be divided into three types based on image feature fusion, shape feature fusion and graph convolutional neural network.

The 3D reconstruction method based on multi-stage network was first proposed by Groueix et al. [9] in 2018. Resnet [10] was used as the encoder of the image, then MLP was used for decoding, the two-dimensional points were mapped to three-dimensional points, the point cloud model was reconstructed using the point's bevel loss as the loss function, and the grid model was reconstructed using the Poisson reconstruction algorithm.

In order to reconstruct a more refined three-dimensional model in the form of grid, Wang et al. [11] applied a graph convolutional neural network in 2008 to build the grid model from the image features that the encoder extracted. The next year, Tang et al. [12] developed the skeleton bridging network by combining different methods.

In order to reduce the memory during training and further improve the reconstruction effect, some researches reconstruct the three-dimensional model in the form of grid by learning the implicit function of the reconstruction target. After the neural network constructs an implicit function to represent the 3D shape, an extraction algorithm is used to extract information from the learned 3D representation and reconstruct the 3D grid model. For example, Wang et al. [13] estimated the camera attitude and projected it in 2019, then used MLP to construct a symbolic distance function, used the symbolic distance function to represent the object shape implicitly, and reconstructed the grid model.

In 2019, Chen et al. [14] designed a microrenderable framework to render the initially reconstructed 3D model into a two-dimensional image through differentiable rendering, construct a two-dimensional image loss with the input image, and reconstruct a grid model with color texture by estimating shape, lighting and texture.

The input of grid-based 3D reconstruction of multiple images is usually multiple images with known camera parameters. By combining the corresponding camera parameters of each image in multiple views, the corresponding relationship between images can be obtained, thus improving the effect of 3D model reconstruction. For example, Wen et al. [15] used graph convolutional neural networks to iteratively refine the rough model in 2019. In 2021, Bautista et al. [16] generated feature maps by U-shaped network encoder, then connected the feature maps according to camera parameters, generated feature points through MLP, and then used a method similar to space occupancy network to predict space occupancy and reconstruct three-dimensional models through implicit functions. In the same year, Shrestha et al. [17] estimated the voxel model of the object, used the voxel model to render the depth map, compared the rendered depth map with the depth map of multi-view stereo estimation, further refined the three-dimensional model by using the contrast features, and finally obtained the three-dimensional model in the form of grid.

2.4. Three-dimensional reconstruction based on neural radiance field

The concept of surface light field [18] was first proposed in 2000. It is a technology used to capture the location information and direction information of light in space, which can present the appearance and geometry of a three-dimensional model, so as to realize the effect of rendering three-dimensional models from different perspectives.

In 2020, NeRF [19] or neural radiation field was proposed firstly by Mildenhall et al., which can be seen as an implicit representation of surface light field, which does not require explicit storage of light information, but uses neural networks to encode and decode the color and volume density of each point in space. It uses volume rendering with implicit neural scene representation for new perspective synthesis via MLP.

NeRF [19] is a technology that can create highly realistic views of complex scenes from different angles. It has received a lot of interest from researchers in this field. In some improved methods of NeRF [19], they only need a few pictures to start training, and even can complete the training of a scene in a few seconds and render an excellent effect quickly.

Subsequently, many researchers carried out innovations on the basis of NeRF [19], and also made good progress. For example, in December 2020, PixelNeRF [20] was proposed, this method is capable of accurately predicting the 3D representation of multiple objects from one or several images, without requiring any additional supervision or optimization. Moreover, it exhibits robustness in handling larger viewpoint changes and more complex scenes.

FastNeRF [21] was proposed. It uses a graphics-inspired decomposition method that can compactly cache the depth radiation map of each position in the space and efficiently query the map using the ray direction, thereby rendering high-fidelity realistic images on high-end consumer GPUs at a speed of 200 Hz, which is 3000 times faster than the basic NeRF [19].

Mip-NeRF [22] used cone tracing to introduce integrated position coding instead of ray-tracing for standard NeRF [19] volume rendering. It mainly solves the problem of aliasing and blurring of NERF [19] and improves the rendering effect.

RapNeRF [23] was proposed to improve the NeRF effect on view extrapolation without requiring any additional supervision or optimization.

PointNeRF [24] was proposed in volume rendering, which to use feature point clouds as an intermediate step. Pre-trained 3DCNN can create depth and surface probabilities from the cost volume in the training view with generating a dense point cloud. It optimizes point clouds through pruning and growth mechanisms to improve the speed and quality of reconstruction, and can handle more complex scenes.

NeRFusion [25] used the 2D image features that CNN extracted to extract 3D cost quantity. It then used a sparse 3DCNN to process the 3D cost quantity into local feature quantity. NeRFusion can reconstruct a variety of small-scale object and large-scale indoor scenes with higher speed and better effect, surpassing NeRF [19] and other methods at that time.

AutoRF [26] focused on new perspective synthesis of objects that without any background. A 3D object detection algorithm and panoramic segmentation are applied to a 2D multi-view image in AutoRF to obtain 3D bounding boxes and object masks. The bounding boxes are used to set up standardized object coordinate spaces for each object. Then, volume rendering is performed for each object in its coordinate space. It can synthesize realistic new perspective images from a single image and can handle objects with complex shapes and textures.

2.5. Classification summary of 3D reconstruction methods

The voxel-based 3D reconstruction method is to discretize the three-dimensional space into a series of voxels, and use voxels to characterize the three-dimensional model. The voxels can be compared to the pixels in the two-dimensional space. Therefore, such methods can model complex three-dimensional geometric shapes. The advantage is very simple and easy to understand, but the voxel processing requires a large amount of calculation.

The three-dimensional reconstruction method based on point cloud uses point cloud to represent the three-dimensional shape. Point cloud is a collection of points in the coordinate system, including three-dimensional coordinates, colors, classification values and other information. This kind of method usually uses the encoder-decoder structure network to decode the image features into point cloud coordinates or features, so as to reconstruct the point cloud model. The advantage of this method is that it can flexibly represent any shape, but there are also problems such as sparseness and discontinuity.

The mesh-based 3D reconstruction method uses the mesh to represent the 3D model. The mesh is a polyhedral structure composed of multiple triangles. This kind of method usually uses the encoder-decoder structure network to decode the image features into mesh vertices or facets to reconstruct the mesh model. The advantage of this method is that it can accurately represent the surface of the object, but there are also problems such as complex topology and difficult optimization.

The three-dimensional reconstruction method based on the neural radiation field is to use the neural network to perform three-dimensional reconstruction of the radiation field. The neural radiation field is a method of implicitly representing complex scenes. It uses a multi-layer perceptron (MLP) to map each position and direction in the scene to color and density. The three-dimensional reconstruction method based on NeRF [19] uses NeRF [19] to represent the three-dimensional shape, and uses the volume rendering equation to synthesize a new perspective image. This method can generate high-fidelity, high-resolution, continuous and differentiable images. The advantage is that high-quality results can be output through a small number of images, but there are also problems such as large amount of calculation and weak generalization ability.

3. Problems and prospects

This chapter mainly discusses the existing problems of 3D reconstruction in four aspects: reconstruction accuracy, model performance, generalization performance and multi-modal processing, and gives possible solutions. Finally, the future development of 3D reconstruction is forecasted.

3.1. Reconstruction accuracy problem

3D reconstruction techniques based on deep learning usually need to represent 3D shapes as a discrete or low-dimensional data structure, such as voxels, point clouds, grids, etc. These representations have certain limitations, such as voxel will lead to memory consumption and resolution loss, point clouds will ignore surface information and topology, and grids will increase complexity and instability. Existing methods have the problem of insufficient accuracy, or high-precision reconstruction results need to rely on a large number of input data and training data, and in most cases, we can not get a large number of images for training or generation. Therefore, how to improve the precision and quality of 3D reconstruction is a key problem. Some solutions include: Using multi-scale or hierarchical network structures to improve resolution and detail Using implicit functions or grid-free representations to avoid issues of discretization and parameterization. Use multi-modal or multi-view inputs to increase the amount of information and constraints.

3.2. Model performance issues

The volume of existing models is still too large to be mounted on terminals with weak computing performance, which means that the use scenario of 3D reconstruction technology will be limited. For example, in the field of AR/VR, such virtual reality terminals usually cannot obtain the computing power conditions of platforms such as PCS, so they cannot be mounted on models that require more computing power or cause the performance of models to decline. For example, the training time is too long, the reasoning speed is too slow, and the memory usage is too large. Therefore, how to improve the model performance is an important problem. Some solutions include: using lightweight or compact network structures to reduce parameters and computation; Using techniques such as knowledge distillation or network pruning to compress model size; Use a parallelized or distributed computing framework to speed up training, reasoning, etc.

3.3. Generalization performance problems

At present, 3D reconstruction relying on deep learning technology often relies heavily on data sets, but in practical applications, there may be great differences between the input image and the training data, such as viewing Angle, illumination, occlusion, noise, etc. When switching domains does not migrate well, a model needs to be retrained for each scenario, which results in reduced model generalization performance. Therefore, how to improve the model generalization performance is a challenging problem. Some solutions include: using techniques such as data enhancement or adversarial training to increase data diversity and robustness; Use techniques such as self-supervised or unsupervised to reduce the need for annotated data; Use techniques such as meta-learning or transfer learning to adapt to new fields or tasks, etc.

3.4. Multimodal Processing Problems

3D reconstruction techniques based on deep learning usually use only a single input mode, such as RGB images or depth images, to reconstruct 3D shapes. However, in practical applications, there may be many different input modes, such as RGBA image, video, voice, text, etc., and there may be complementary or conflicting information between these modes, such as color, texture, motion, semantics, etc. Therefore, how to deal with multi-modal input effectively is a problem worth studying. Some solutions include: using a multi-task or multi-target network architecture to handle multiple modes simultaneously; Use attention mechanisms or fusion strategies to enhance or select relevant modes; Use techniques such as conditional generation or adversarial generation to generate different 3D shapes according to different modes.

3.5. Future Outlook

The future development of 3D reconstruction may be carried out from the following aspects:

1) Combining deep learning methods with traditional methods, such as combining neural fields with traditional vision methods.

2) Further exploration of multiple modes. We look forward to seeing work that can do 3D reconstruction between image and image, text and image, video and image. Existing studies such as CLIP-NeRF[27] provide some ideas for our research.

3) Exploration of low-level semantics. At present, the exploration of low-level semantics in the field of 3D reconstruction is not complete, such as denoising, image recovery and other aspects need to be further developed.

4) Combine with generative artificial intelligence. With the rise of generative artificial intelligence models in recent years, 3D reconstruction technology can rely on AIGC technology to quickly build virtual worlds.

4. Conclusion

This thesis mainly introduces the three-dimensional reconstruction methods based on deep learning, including voxel, point cloud, grid, neural radiation field and other representation methods, as well as the deep learning framework based on these representation methods. Among them, NeRF[19] technology has made important progress in the field of 3D reconstruction, which can complete the 3D reconstruction of complex scenes in a very short time. It can be seen that compared with the traditional 3D reconstruction method, the 3D reconstruction method based on deep learning has more flexible input and output, and higher efficiency. It is believed that in the future, 3D reconstruction technology will combine the advantages of deep learning methods and traditional methods to further explore and develop in order to achieve more efficient and accurate 3D reconstruction. It can be more widely used in multi-modal input, low-level semantics, generative artificial intelligence and other more scenarios.

References

[1] Roberts L G. Machine Perception of Three-Dimensional Solids [Ph.D.dissertation],Massachusetts Institute of Technology,USA,1963

[2] Choy C B, Xu D and Gwak J, 2016. Choy et al.(2016): A Unified Approach for Single and Multi-view 3D Object Reconstruction//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands: Springer: 628-644. [DOI: 10.1007/ 978- 3-319-46484 - 8_38]

[3] Yang B, Rosa S and Markham A, 2019. Dense 3D Object Reconstruction from a Single Depth View. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(12): 2820-2834. [DOI:10.1109/TPAMI.2018.2868195]

[4] Tatarchenko M, Dosovitskiy A and Brox T, 2017. Octree Generating Networks: Efficient Convolutional Architectures for HighResolution 3D Outputs// Proceedings of the IEEE International Conference on Computer Vision. Honolulu, USA: IEEE: 2088- 2096. [DOI:10.1109/ICCV.2017.230]

[5] Xie H, Yao H and Sun X, 2019. Pix2Vox: Context-Aware 3D Reconstruction From Single and Multi-View Images//Proceedings of the International Conference on Computer Vision. Seoul, Korea (South): IEEE: 2690-2698. [DOI:10.1109/ICCV. 2019.00278]

[6] Fan H, Su H and Guibas L J, 2017. A Point Set Generation Network for 3D Object Reconstruction From a Single Image//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 605-613. [DOI:10.1109/CVPR.2017.264]

[7] Mandikal P, Navaneet K L and Agarwal M, 2019. 3D-LMNet: Latent Embedding Matching for Accurate and Diverse 3D Point Cloud Reconstruction from a Single Image//Proceedings of British machine vision conference. Newcastle, UK: 662-674

[8] Jiang L, Shi S and Qi X, 2018. GAL: Geometric Adversarial Loss for Single-View 3D-Object Reconstruction//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer: 802-816. [DOI:10.1007/978-3-030- 01237- 3_49]

[9] Groueix T, Fisher M and Kim V G, 2018. A Papier-Mâché Approach to Learning 3D Surface Generation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 216-224. [DOI:10.1109/CVPR. 2018. 00030]

[10] He K , Zhang X , Ren S ,et al.Deep Residual Learning for Image Recognition[J].IEEE, 2016.DOI:10.1109/CVPR.2016.90.

[11] Wang N, Zhang Y and Li Z, 2018. Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer: 52- 67. [DOI:10.1007/978-3-030-01252-6_4]

[12] Tang J, Han X and Pan J, 2019. A Skeleton-Bridged Deep Learning Approach for Generating Meshes of Complex Topologies From Single RGB Images//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach,USA: IEEE: 4541-4550. [DOI:10.1109/CVPR.2019.00467]

[13] Wang W, Xu Q and Ceylan D, 2019. DISN: deep implicit surface network for high-quality single-view 3D reconstruction// Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates Inc.: 492-502

[14] Chen W, Ling H and Gao J, 2019. Learning to Predict 3D Objects with an Interpolation-based Differentiable Renderer//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, USA: Curran Associates, Inc.: 9609-9619.

[15] Wen C, Zhang Y and Li Z, 2019. Pixel2Mesh++: Multi-View 3D Mesh Generation via Deformation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1042-1051. [DOI:10.1109/ICCV.2019.00113]

[16] Bautista M A, Talbott W and Zhai S, 2021. On the Generalization of Learning-Based 3D Reconstruction//Proceedings of the Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2180-2189. [DOI:10.1109/WACV48630.2021. 00223]

[17] Shrestha R, Fan Z and Su Q, 2021. MeshMVS: Multi-View Stereo Guided Mesh Reconstruction//International Conference on 3D Vision. London, UK: IEEE: 1290-1300. [DOI: 10.1109/3DV53792. 2021. 00136]

[18] Wood D N ,Azuma D I ,Aldinger K , et al. Surface light fields for 3D photography[C]// SIGGRAPH conference. 2000.

[19] Mildenhall, B. , Srinivasan, P. P. , Tancik, M. , Barron, J. T. , Ramamoorthi, R. , & Ng, R. . (2020). NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis.

[20] Alex Yu, Vickie Ye, Matthew Tancik, Angjoo Kanazawa, pixelNeRF: Neural Radiance Fields From One or Few Images, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4578-4587.

[21] Stephan J. Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, Julien Valentin, FastNeRF: High-Fidelity Neural Rendering at 200FPS, in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14346-14355

[22] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. MartinBrualla, and P. P. Srinivasan, Mip-nerf: A multiscale representation for anti-aliasing neural radiance fifields, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 5855–5864.

[23] J. Zhang, Y. Zhang, H. Fu, X. Zhou, B. Cai, J. Huang, R. Jia, B. Zhao, and X. Tang, Ray priors through reprojection: Improving neural radiance fifields for novel view extrapolation,in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18 376–18 386.

[24] Q. Xu, Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, and U. Neumann,Point-nerf: Point-based neural radiance fifields, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5438–5448.

[25] X. Zhang, S. Bi, K. Sunkavalli, H. Su, and Z. Xu, Nerfusion: Fusing radiance fifields for large-scale scene reconstruction, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5449–5458.

[26] N. Muller, A. Simonelli, L. Porzi, S. R. Bulo, M. Nießner, and P. Kontschieder, Autorf: Learning 3d object radiance fifields from single view observations, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3971– 3980.

[27] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, Jing Liao,CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields,Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3835-3844