Multimedia Tools and Applications (2021) 80:463–498
https://doi.org/10.1007/s11042-020-09722-8

Single image 3D object reconstruction based on deep learning: A review

Kui Fu[1] · Jiansheng Peng[1,2] · Qiwen He[1] · Hanxiao Zhang[2]

Received: 1 January 2020 / Revised: 19 August 2020 / Accepted: 25 August 2020 /
Published online: 3 September 2020

Abstract
The reconstruction of 3D object from a single image is an important task in the field of computer vision. In recent years, 3D reconstruction of single image using deep learning technology has achieved remarkable results. Traditional methods to reconstruct 3D object from a single image require prior knowledge and assumptions, and the reconstruction object is limited to a certain category or it is difficult to accomplish a good reconstruction from a real image. Although deep learning can solve these problems well with its own powerful learning ability, it also faces many problems. In this paper, we first discuss the challenges faced by applying the deep learning method to reconstruct 3D objects from a single image. Second, we comprehensively review encoders, decoders and training details used in 3D reconstruction of a single image. Then, the common datasets and evaluation metrics of single image 3D object reconstruction in recent years are introduced. In order to analyze the advantages and disadvantages of different 3D reconstruction methods, a series of experiments are used for comparison. In addition, we simply give some related application examples involving 3D reconstruction of a single image. Finally, we summarize this paper and discuss the future directions.

Keywords Single image 3D reconstruction · Deep learning · Computer vision · 3D shape representation

Kui Fu and Jiansheng Peng contributed equally to this work.

✉ Jiansheng Peng
sheng120410@163.com

[1] School of Physics and Mechanical and Electronic Engineering, Hechi University, Yizhou, Guangxi 546300, China

[2] School of Electrical and Information Engineering, Guangxi University of Science and Technology, Liuzhou, Guangxi 545006, China

1 Introduction

Three-dimensional reconstruction of images is a common topic in computer vision, medical image processing [74, 4] and virtual reality [109]. The main purpose of theory and technology related to computer vision is to obtain information from images or multi-dimensional data to establish artificial intelligence systems. 3D reconstruction of images is one of the main tasks of computer vision, and its purpose is to study the generation of corresponding 3D structures from a single image or multiple images [93, 82]. According to the different reconstruction targets, the 3D reconstruction of images can be divided into 3D scene reconstruction and 3D object reconstruction. A big challenge for single-view 3D scene reconstruction is to predict invisible parts from a single image [38, 108, 100]. Multi-view 3D scene reconstruction [36, 39] and multi-view 3D object reconstruction [18] can integrate the information of multiple images to compensate for the defect of single image prediction uncertainty for the invisible part. Compared with the traditional 3D reconstruction methods of multi views [25, 95] and models [14, 45], deep learning has the ability to process big data. Therefore, there have been many studies in recent years that combine traditional methods with deep learning [114, 149, 27].

Some studies have reviewed the image 3D reconstruction techniques [32, 143]. Ham et al. [32] reviewed the methods of 3D reconstruction of single still image, RGB depth image, multi-perspective of 2D images, and video sequences. Most of the methods reviewed by Ham et al. use traditional 3D reconstruction algorithms, and a few methods use deep learning techniques. Yuniarti et al. [143] briefly reviewed the method of 3D reconstruction of single image or multiple images based on deep learning. This review is different from the review by Ham et al. or Yuniaart et al. We reviewed the single image 3D reconstruction method based on deep learning more comprehensively, including the challenges faced by the method, reconstruction algorithms for different 3D representations, 3D reconstruction training architecture, etc. In this paper, our main research goal is reconstruction of 3D object from a single image. The problem of single-image 3D object reconstruction is similar to that encountered in single-view 3D scene reconstruction. Since a single image loses a lot of information about the three-dimensional object, the reconstruction result is uncertain. Traditional methods for 3D object reconstruction of a single image are often based on prior models [5, 52], or use 2D annotation [49]. These methods are often limited to object reconstruction in a certain category. For example, assuming the illumination is fixed and the shape is restored from the shadow [92, 20, 2]. Assuming surface smoothness, the shape is restored from the texture [68]. Due to the complexity of the real datasets, this class of model that requires assumptions is less effective in real applications.

With the continuous rise of deep learning in recent years, it has become a hot direction to reconstruct 3D objects from a single 2D image by using deep neural networks [147].The 3D object reconstruction method based on deep learning is to train neural networks to learn the mapping relationship between 2D image and 3D object. The main motivation of this paper is to provide a survey of generating 3D shapes from a single image using deep learning in recent years. We focus on analyzing the main challenges and research methods in the 3D reconstruction of a single image. These challenges and methods represent the problems and development trends that deep learning needs to solve in the single image 3D reconstruction in the future. The following sections are arranged as follows: In Sect. 2, this review first discusses the challenges faced by reconstructing 3D objects from a single image based on deep learning. In Sect. 3 and 4, we introduce the encoders and decoders currently used in 3D

reconstruction of a single image, respectively. In Sect. 5, the public training details in many literatures are introduced, including loss function and network training architecture. In Sect. 6, we introduce the datasets and evaluation metrics for 3D reconstruction experiments. In Sect. 7, we conduct multiple comparative experiments to analyze the advantages and disadvantages of different 3D reconstruction methods. In Sect. 8, we introduce the related applications of single image 3D reconstruction. In Sect. 9, we summarize the full text and look forward to the future development trends.

2 Challenges of single image 3D object reconstruction

The single image 3D reconstruction based on deep learning faces multiple challenges, which lead to the development of this direction is still in its infancy. In general, the 3D reconstruction of a single image mainly has the following challenges: (1) shape complexity reconstruction of objects, (2) uncertainty reconstruction of objects, (3) reconstruction of fine-grained objects, (4) memory requirements and calculation time, (5) training datasets, (6) selectivity of 3D shape representations.

2.1 Challenge 1: shape complexity of objects

First of all, the shape complexity of objects is mainly reflected in the differences between the shapes of different classes of objects, which exist in the reconstruction results of individual training for the same class of objects and joint training of different classes of objects. Therefore, a good 3D reconstruction model should have the ability to characterize objects with different complexity. In addition, the model needs to learn various connections between different classes of objects while keep its own uniqueness among the same classes of objects. Secondly, the shape complexity of objects is also reflected in itself. The structure of a simple object can often be represented by combining multiple cuboids. When small parts occupy less of the overall structure, a simple object tends to have a higher reconstruction score. However, when complex objects are small in structure and have fine-grained parts (e.g. gun trigger), reconstruction results tend to be poor. Under this challenge, improving the resolution of reconstructed 3D objects is a relatively straightforward solution.

2.2 Challenge 2: uncertainty of objects

Single image 3D reconstruction is an ill-posed problem [78]. Because a single image loses a lot of three-dimensional information and lacks prior knowledge or assumptions, its reconstruction results are not unique. Therefore, some recent studies have tried to predict the correct shape by some auxiliary means [31]. For humans, we can deduce about invisible 3D shapes from a single RGB image from our own rich experience accumulation. This reflects from the side that 3D reconstruction of a single image based on deep learning requires sufficient datasets for training.

2.3 Challenge 3: reconstruction of fine-grained objects

For most 3D reconstruction models, their goal is to generate 3D objects with fine-grained, not just a rough 3D representation. Different 3D shape representations face different difficulties.

For example, for voxel-based 3D reconstruction methods, most of them face high memory usage and computing costs. For mesh-based methods, most of them are limited by mesh topology. For different 3D reconstruction methods, the corresponding solutions are introduced in Sect. 4. However, they all have problems that need to be solved. Therefore, it is a huge challenge how to reconstruct fine-grained 3D objects from a single image.

2.4 Challenge 4: memory requirements and calculation time

For an excellent single image 3D object reconstruction model, it should have lightweight parameters. In addition, with limited memory requirements, it is not only necessary to reconstruct the correct shape of fine-grained parts from a single image, but also to have good training and inference time. Currently, some solutions to this challenge are introduced in Sect. 4.

2.5 Challenge 5: training datasets

The deep neural network can exert their powerful learning ability in the era of large model and big computing, which is attributed to the existing big data. However, recent studies have shown that the single image 3D object reconstruction based on deep learning actually learns recognition capabilities (search and clustering) [112], and it rarely learns reconstruction capabilities. ShapeNet [7] dataset is a commonly used dataset for 3D object reconstruction. The entire dataset is usually divided into a training set, a validation set and a test set. Because the 3D models in the testing and training sets are highly similar, neural networks may be misled to learn to recognize. In addition, there are great differences between wild datasets and synthetic datasets. For images that have not been seen by neural networks, it may lead to different reconstruction results that chooses different coordinate systems to reconstruct 3D shapes [99]. The image content in the real dataset is complex, such as occlusion, multiple categories of objects, and different lighting. Therefore, it is difficult to accomplish the 3D reconstruction of the object on the real dataset after training on a clean synthetic dataset. Recently, a lot of studies have been tried to render 2D images using texture datasets [16] and background datasets [135]. However, there is still a large difference between the actually rendered 2D image set and the real scene 2D image set. It is difficult for the model to adapt to the dataset in the real scene after training on the rendered dataset. In the end, the 3D shape reconstructed by the model is poor. Therefore, it is a challenging problem how to improve the existing training dataset and make a dataset suitable for 3D reconstruction.

2.6 Challenge 6: selectivity of 3D shape representations

At present, most studies choose different 3D shape representations to accomplish the 3D reconstruction of a single image. The method based on voxel representation can use 3D convolutional neural networks (CNN), and it can reconstruct objects with arbitrary topological structure. However, the huge memory requirements and computation time limit the reconstruction results of most methods to low resolution. Although many improvements have been proposed for this problem [111, 88, 103], the reconstruction results still fail to achieve ultra-high precision reconstruction. Point cloud representation is relatively simple and highly flexible. Because the point cloud is not a regular structure, it cannot adapt well to traditional 3D CNN. The mesh helps to restore the details of the model in the three-dimensional object,

and its representation accuracy is high. Similarly, the mesh is not a regular form of geometric data, so it cannot directly apply 3D CNN. Most of the current mesh-based research adopts the method of deformation from the mesh model. However, this method cannot deal well with objects of unknown topology. Although some methods solve the problem of topology [83, 29], they also introduce some problems. Both parametric and implicit surface representation methods can represent objects with smooth surfaces, and the generated objects have better visual attraction. However, the 3D reconstruction based on parametric surface representation is difficult to adapt to the reconstruction of multi-genus complex structure objects by using the global surface parametrization method [101]. Approximating 3D shapes with multiple locally parameterized surfaces also faces the problem of stitching between surfaces [30]. The decoder based on implicit surface representation needs to predict all points in 3D space, which is time-consuming in the inference stage. Using volume primitives to represent 3D shapes can predict relatively correct 3D structures. Because the volume primitives used to represent 3D shapes are relatively simple, this method can only reconstruct simple 3D structures at present. Surface primitives use multiple planar patches to approximate 3D shapes. Although this method simplifies the 3D representation, it also needs to solve the problem of stitching between the planes. In general, these 3D representations have advantages and disadvantages.

3 2D encoder

The study of 3D reconstruction of 2D images by deep learning has become a popular research in recent years. Deep learning is also known as deep neural networks. The learning capabilities of deep neural networks can be used to accomplish many tasks related to computer vision, such as image classification [118, 44], image segmentation [3, 10, 11], object recognition [144, 126, 106], and image super-resolution [54, 47, 61, 65]. The success of applying deep learning methods in the field of 2D image also promotes the development of 3D reconstruction tasks [91, 19, 87, 48, 97, 13, 24, 77, 1, 151]. Generally, a 3D reconstruction model based on deep learning can represent the input image set as $I = \{I_1, I_2, \ldots, I_n\}$, let the corresponding ground truth 3D shape be Y, and the reconstructed 3D shape can be optimized by Eq. (1):

$$L = \sigma argmin dis(f_\sigma(I), Y) \tag{1}$$

Here, $f_\sigma(\cdot)$ represents the reconstructor, including 2D encoder and 3D decoder, and σ represents the $f(\cdot)$ parameter set. The reconstructor reconstructs a 3D shape from the input image. *dis* is a measure of the distance between the reconstructed shape and the ground truth shape, and it is represented as L when the two achieve a minimum.

In the 2D encoder stage, input images are encoded into a latent space for feature compression. According to the encoding method, it can be divided into encoded images to discrete latent space and encoded images to continuous latent space. The way of encoding input images

Table 1 Different methods to encode images into latent space

Encoding space	Encoding mode	Method
Discrete latent space	Direct encoding	Conv, ResNet, RNN, FC
	Intermediate representation encoding	
Continuous latent space	Direct encoding	VAE

into discrete latent space can be further divided into direct encoding and intermediate representation encoding. Commonly used networks for image encoding to discrete space are standard convolutional (Conv) networks, residual Networks (ResNet), recurrent neural networks (RNN) and fully connected (FC) networks. Encoding an input image into a continuous latent space often uses the encoder part of the variational auto-encoder (VAE) [55]. The comparison results of encoding input images into a latent space are shown in Table 1.

3.1 Images to discrete latent space

In this way, the encoder encodes the input images into a low-dimensional latent layer vector. The decoder then maps the latent layer vector to a 3D shape. Encoding the image into discrete a latent space can be roughly divided into two ways. The first way, 2D convolutional neural networks directly encode input images into a fixed low-dimensional latent vector. The second way, the input images are first encoded to generate an intermediate representation (e.g. 2.5D representation), and then the intermediate representation is encoded similar to the first way.

3.1.1 Direct encoding

For most 3D object reconstruction methods, they directly encode the input image into a lower dimensional discrete latent space. The encoding diagram is shown in Fig. 1.

Choy et al. [15] proposed a shallow network and a deep residual network. A shallow network uses a standard convolutional neural network to encode the input image to low dimensional feature. The deep residual network uses a shortcut connection to improve the standard convolutional neural network. Similarly, Shin et al. [99] used an encoder with residual units. Moreover, there is work with recurrent 2D encoders [138].

Subsequently, Girdhar et al. [28] introduced a TL-embedding network. At the bottom of the T-network, the input image is encoded into a 64D embedding space by using 5 standard convolution layers. At the top of the T-network, an input $20 \times 20 \times 20$ voxel grid is encoded into a 64D embedding space by 3D auto-encoder and the output voxel grid of the same size is decoded. Moreover, there are many studies that use standard convolutional networks to encode

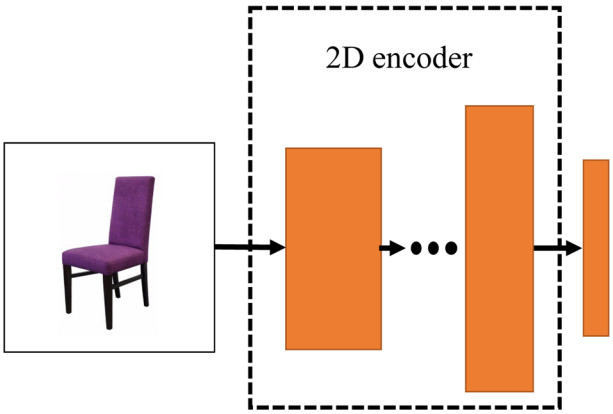

Fig. 1 Image directly encoded into a discrete latent space. In this method, the encoder is used to directly encode the input image into a latent vector of fixed length

directly input images into a discrete hidden space. [137, 86, 115, 136, 122]. In addition to using convolutional neural networks, some studies also use fully connected networks [26].

3.1.2 Intermediate representation encoding

Many studies first tried to generate an intermediate representation of the input image through a 2D encoder-decoder network. Then, the intermediate representation is encoded as a latent vector using a 2D encoder, and the basic encoding diagram is shown in Fig. 2. Wu et al. [129] proposed MarrNet. They first used ResNet-18 [34] to encode a 256×256 RGB image into multiple feature maps. Then, the corresponding intermediate representation (depth maps, surface normal, and silhouette images) is output through a decoder. Whereafter, the intermediate representation is encoded into a 200-dimensional vector. Finally, this vector outputs a $128 \times 128 \times 128$ voxel grid through a decoder. In addition, there are some similar studies [110, 130, 148].

3.2 Images to continuous latent space

Unlike encoding images into a discrete latent space, encoding images into a continuous latent space pays more attention to learning the probability density function in the latent feature space, and the basic encoding diagram is shown in Fig. 3. The encoder of the VAE observes samples from the target distribution and produces a vector of means μ and variances σ parameterizing a set of Gaussians, which are sampled to produce a latent vector. To be able to optimize network parameters using a backpropagation technique, the network needs to use a reparameterization trick that randomly samples ε from a unit Gaussian.

Wu et al. [128] used an encoder in VAE to encode the input image to a latent representation vector, which is then fed into a 3D generative adversarial network (3D GAN) to accomplish the 3D volumetric reconstruction of a single image. In addition, there are also some studies using VAE to encode images into a continuous latent space [102, 71].

4 3D decoder

The 2D encoder based on neural networks learns to encode an input image into a latent vector from a large amount of data. The latent vector is then converted into three-dimensional data by

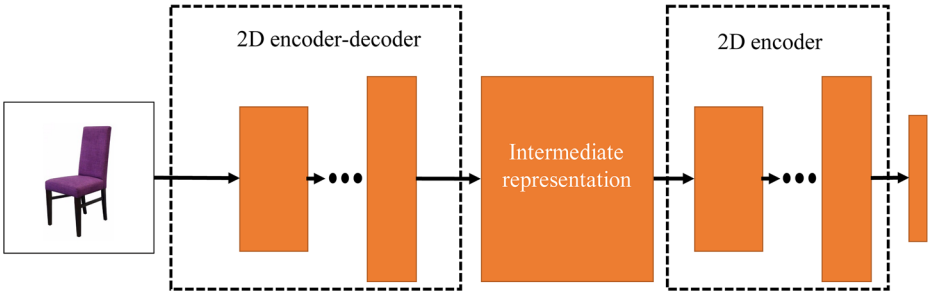

Fig. 2 Intermediate representation encoding. This method first predicts an intermediate representation from an input RGB image through an encoder-decoder network. Then, an encoder is used to encode the intermediate representation into a discrete latent space

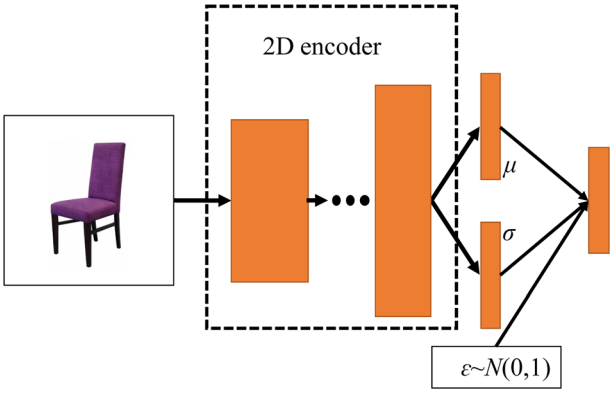

Fig. 3 Image directly encoded into a continuous latent space

a 3D decoder. In order to generate 3D shapes from the input image, the entire network needs to combine low-level image features with higher-level part arrangement knowledge. Most 3D object reconstruction methods based on a single image choose to use low-level image features for reasoning. However, these methods lack an understanding of object structure or structure relationship at the expression level. They have several output representations: voxel grid, point cloud, mesh, parametric surface, and implicit surface. In addition, some works attempt to understand higher-level representations of object structures, and they see 3D objects as a collection of two primitives (volumetric primitives or surface primitives). Further work attempts to understand the symmetrical relationship between parts at a higher level [76]. Three-dimensional decoder classification is shown in Table 2.

The following content reviews the 3D decoders based on different 3D representations. In order to better show the difference between 3D decoders based on different representations, we review them separately according to their expression level.

4.1 Low-level expression decoding

In low-level expressions, the voxel grid, point cloud, and mesh expressed in discrete forms are more studied, and the parametric surface and implicit surface in continuous form are relatively less studied.

4.1.1 Voxel-based representation decoding

3D decoding based on voxel representation can be divided into dense voxel decoding, intermediate representation voxel decoding, sparse voxel decoding and other decoding.

Table 2 Different levels of 3D decoder categories

Expression level	Reasoning clue	Representation form
Low level	Color, texture, shadow, edge	Voxel grid, point cloud, mesh, parametric, surface, implicit surface
Higher level	Part structure, structural relationship	Volumetric primitives, surface primitives

Dense voxel decoding With the development of deep learning research, deep learning models based on CAD databases have been proposed for 3D modeling of a single image. Wu et al. [127] began to propose a 3D ShapeNets model, which used a deep convolutional belief network to learn the joint distribution of all 3D voxels in a data-driven manner. This work is one of the earlier models for 3D shape expression using voxel form. Although the reconstruction results are rough, the experimental results show that this is a good start. Choy et al. [15] first feed the latent layer vectors into an intermediate module (3d long short-term memory), and then a 3D shape is generated by a decoder with a residual network (see Fig. 4c). Similarly, Yang et al. [141] introduced an attentional aggregation module (AttSets) between the latent layer vector and the decoder. These two methods can use the intermediate module to accomplish the 3D reconstruction of single-view image or multi-view images. In addition, Yang et al. [138] used a recurrent 3D decoder to decode the latent layer units to generate a 3D volumetric grid.

Different from the above methods, there are also some studies to decode directly the latent layer vector into a 3D shape [28, 137, 86, 115, 136, 26, 128, 102]. These methods use similar decoder architectures (see Fig. 4b).

Intermediate representation voxel decoding In recent years, many studies have added an intermediate representation (2.5D sketch) between the 2D image and the 3D shape prediction. Compared to predicting 3D shape directly from a single 2D image, this method is easier to express 3D objects. Wu et al. [129] proposed MarrNet, which first estimated 2.5D sketches (depth, normal maps and silhouette) of an input RGB image. Subsequently, a 3D encoder-decoder is used to estimate a 3D shape from the 2.5D sketch represented in the middle. Similarly, Sun et al. [110] and Wu et al. [130] sequentially estimate the 2.5D representation and 3D shape from the input RGB image. Different from the method of directly estimating the 3D shape from 2.5D, Zhang et al. [148] decomposed the 2.5D to 3D shape process into two stages of partial 3D completion and full 3D completion. They processed the depth map in turn by a partial spherical map and an inpainted spherical map to represent the full surface of the object. Finally, the voxel reconstruction network combined the back projection of both the

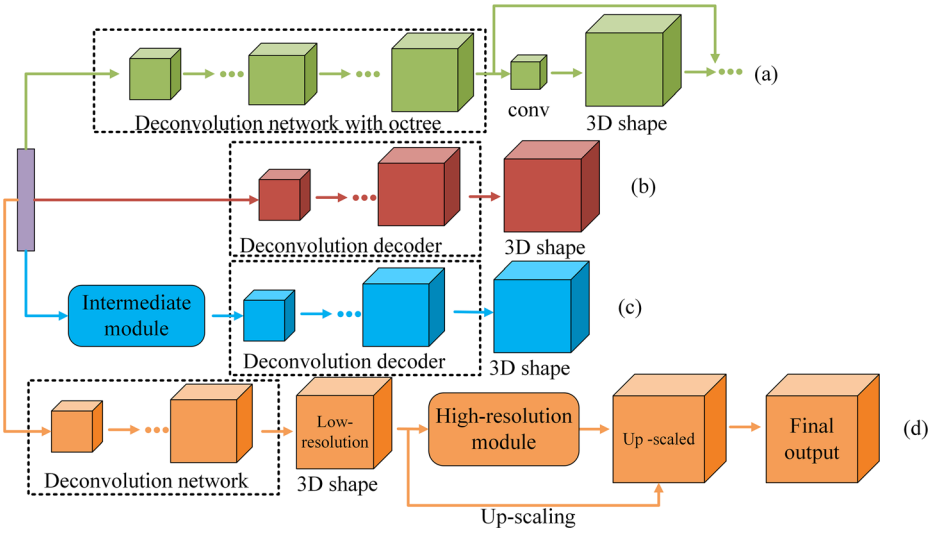

Fig. 4 Four voxel decoder architectures

depth map and the inpainted spherical map to output a 3D shape. Experimental results show that the network can also obtain results that are closer to ground truth on untrained classes. The resolution of 3D object reconstructed by these methods can reach $128 \times 128 \times 128$, and the reconstruction results also have more details. However, there is still a big gap compared with the appearance of real 3D models.

Sparse voxel decoding In 3D space, the representation of 3D shape is closely related to the surface resolution of the reconstructed object. In recent years, a sparse voxel representation method for octree has been proposed [89, 119, 90]. Coarse-resolution voxel prediction can be used for most of the objects in the space that are mostly empty and fully occupied. The mixed part needs to be further subdivided. Voxel sparse representation using the octree method can make the reconstruction object resolution reach $512 \times 512 \times 512$. Tatarchenko et al. [111] proposed octree generating networks (OGN) (see Fig. 4a). The entire network starts from a certain layer, and the convolutional network is placed on the octree to run until the resolution of the output meets the set conditions. Compared with the dense voxel decoding method, OGN can represent a higher resolution 3D output in a limited memory space. However, when the resolution increases to a certain value, the network is difficult to adapt to big data training. In this case, the performance of the model will decrease. Similarly, Häne et al. [33] proposed a hierarchical surface prediction (HSP) network.

Other decoding In addition to using intermediate representation and octree method to generate high-resolution 3D objects, there are also methods to treat generation of 3D shape as 2D prediction [88, 103, 98]. Richter et al. [88] proposed a 2D encoding method of 3D geometry after thinking about 2D prediction instead of 3D object reconstruction. In order to express low resolution 3D shape more effectively, they developed a method to predict the entire voxel tube from each pixel of the reference view. In addition, the fusion of 6 nested depth maps is used to extend the generated 3D object resolution. Smith et al. [103] first reconstructed a rough 3D shape through a low-resolution 3D encoder-decoder. Then, six orthogonal depth maps with high resolution are restored by 3D super-resolution network. Finally, the high-resolution depth maps are used to carve up-sampled rough 3D shapes to accomplish high-resolution 3D shape (see Fig. 4d). In addition, Shen et al. [98] proposed a fourier-based 3D reconstruction method to predict slices in the frequency domain to reconstruct 3D shapes from 2D space.

4.1.2 Point cloud-based representation decoding

Inspired by the study of point sets in 2D space [21, 8, 84, 56], some studies started using 3D shape generation networks based on point cloud representations [60, 139, 105, 69, 75, 62, 63, 113, 125]. 3D decoders based on point cloud representation can be divided into sparse point cloud decoding and dense point cloud decoding. Due to the irregular structure of the point cloud, it cannot be well adapted to 3D convolutional neural network. Generally, point cloud-based decoders are mainly composed of fully connected networks, which predict point clouds from latent layer vectors (see the middle of Fig. 5).

Sparse point cloud decoding Fan et al. [23] used a point set generation network to generate a 3D point cloud from a single image. As an enlightening study, this method demonstrates the

powerful expression ability of point clouds. Jiang et al. [46] also adopted a similar method, but they introduced a geometric adversarial loss (GAL) for improvement. Unlike an architecture that generates a 3D point cloud from a single image, Zeng et al. [145] first generated an intermediate representation depth map from a single image. Then, the depth map is sequentially generated into a local point cloud and a full point cloud.

The 3D shape of the generated sparse point cloud is similar to the ground truth as a whole. However, it is difficult to express the surface details of the object due to the small number of generated points.

Dense point cloud decoding Insafutdinov et al. [40] predicted 3D shape and pose by two fully connected layer decoders, respectively. They used the projection module to generate 2D projection maps from predicted camera poses with real images for training. In addition to the method of one-step prediction of dense point clouds from single images, there are also methods of stepwise increasing the reconstruction resolution. Based on the idea of point cloud up-sampling [142], Mandikal et al. [70] proposed DensePCR to generate high-resolution 3D point clouds. Specifically, this is a deep pyramid network that continuously predicts higher resolution 3D point clouds in a hierarchical manner (see the bottom of Fig. 5). First, the network outputs a sparse point cloud of 1024 points. Then the sparse point cloud passes through two dense networks to increase the resolution by 16 times, and finally forms a 3D dense point cloud. The dense network first aggregates global and local features, and then estimates the dense point cloud after adjusting around the coordinate grid.

In addition to the decoders using the fully connected network described above, there is also decoder using 2D convolutional neural networks. Lin et al. [66] proposed a method for generating dense point clouds. They used a structure generator with 2D convolution to predict 3D structures at multiple different viewpoints, and transformed them into the canonical coordinates, and finally fused into a dense point cloud (see the top of Fig. 5). These studies

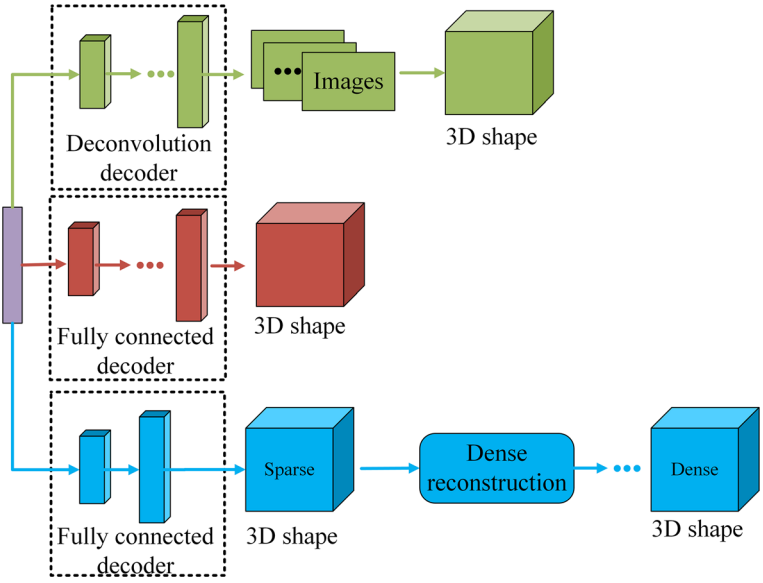

Fig. 5 Three point cloud decoder architectures

have produced high-resolution dense point clouds with high accuracy, but the problem of edge point artifacts needs to be further addressed.

4.1.3 Mesh-based representation decoding

Three-dimensional shapes are essentially represented using vertices and faces, which can be represented by multiple meshes. From traditional two-dimensional to three-dimensional, since most of the geometric information is contained on the surface boundary, using meshes to represent three-dimensional objects is a more accurate modeling method. The 3D shape decoding method based on mesh representation can be divided into 0-genus mesh decoding, multi-genus mesh decoding, and arbitrary genus mesh decoding.

0-genus mesh decoding One of the biggest problems with 0-genus mesh reconstruction is that it is limited to the topology of the initial mesh model. This type of work assumes that the mesh is a 0-genus initial mesh form. Because the mesh is an irregular structure, it cannot be well adapt to convolutional neural networks. Kato et al. [51] used a fully connected network to predict vertex coordinates from a latent space and deform a predefined mesh model (see the top of Fig. 6).In addition, there are other similar methods [50, 67, 12]. These methods use similar decoders, but they differ in optimizing 3D shapes.

In addition to decoders for fully connected networks, there are also methods using graph-based convolutional neural networks [58, 120, 94, 6, 104]. Wang et al. [120] proposed Pixel2mesh to predict 3D mesh shape from a single color picture. They use a decoder based on graph convolutional networks [94, 6]. The decoder uses the perceptual features extracted by the image feature network to deform gradually an initial ellipsoid to produce the correct solid geometry (see the bottom of Fig. 6). Smith et al. [104] introduced an adaptive mesh reconstruction method. They used an extended zero-neighbor graph convolutional networks (0N-GCN) to deform a predefined mesh model, and finally increased the local complexity of the reconstructed mesh through an adaptive face splitting.

Multi-genus mesh decoding The initial mesh of this type of method can select a multi-genus mesh model to deform. Jack et al. [41] first learned multiple templates and then inferred appropriate

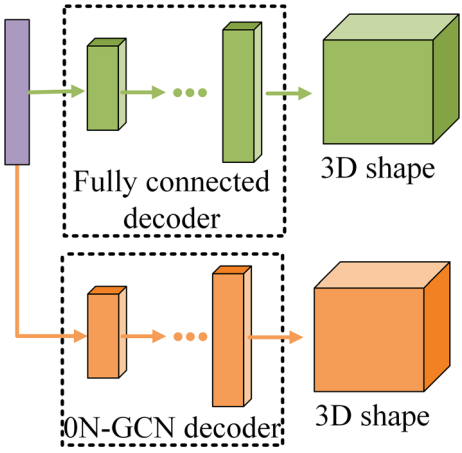

Fig. 6 Two mesh decoder architectures

templates from a single image for deformation. The reconstructed shape is closer to the ground truth most of the time, but it is also limited by the choice of deformed template topology. To solve this problem, Wang et al. [123] used a 3D deformation network (3DN) to deform from the source model. In this way, any existing high-quality mesh model can be transformed into an object model. In addition, similar work has been done by Pontes et al. [83]. This kind of work can accomplish the reconstruction of the multi-genus mesh, but it needs to solve the deformation problem when the object 3D shape genus is inconsistent with the initial mesh.

Arbitrary genus mesh decoding Different from the above two types of deformation methods from the existing initial mesh, a method of 2D perception combined with 3D reconstruction has recently been proposed. In real scenes, 2D images usually have multiple objects. In order to be able to perform 3D shape prediction from end-to-end, Gkioxari et al. [29] combined object detection with 3D reconstruction on the basis of Mask R-CNN [35]. Different from previous mesh reconstruction methods, they first used Mask R-CNN to estimate the bounding box, category label and 2D segmentation mask from a single RGB image. The model then predicted a rough voxel model. Finally, it is converted to a mesh, and the fine branch optimization of the mesh is used to realize the fine prediction of arbitrary geometric structures. Their work solved the reconstruction problem of arbitrary topological mesh, but another problem was how to extract accurately the mask of object.

4.1.4 Parametric surface-based decoding

Sinha et al. [101] studied the problem of generating non-rigid and rigid 3D shaped surfaces using parametric representations. Their 3D shape generation work was developed for category-specific shape surface generated. This is a method that uses a deep neural network to generate a 3D shape surface, but the limitation is that the generated 3D shapes are all surfaces with 0-genus. Then, Grogux et al. [30] proposed a novel structure, named AtlasNet. This structure is also expressed using parametric surfaces. The difference is that they use the method of local surface approximation, which is to map a set of rectangular meshes onto a three-dimensionally shaped surface. This method solved the problem of surface reconstruction topology limitation, and it can reconstruct high-precision 3D object surfaces. However, this method needs to solve how to stitch multiple meshes tightly together.

4.1.5 Implicit surface-based decoding

The 3D object reconstruction method based on the mesh representation can also be called a method based on the explicit surface. Recently, some new studies have focused on 3D object reconstruction from implicit surface networks. These studies use implicit decoders to decode latent feature vectors. Using implicit decoders can generate 3D shapes with superior visual quality.

Indicator function decoding This type of work classifies points in three-dimensional space as the inside or outside of a shape, which is a binary classification problem. Chen et al. [9] proposed using implicit field to generate 3D shapes. The shape encoder takes the extracted feature vector and point coordinates as input, and then the state value of the point inside or outside the shape is returned through an implicit decoder. Similarly, Mescheder et al. [72] proposed occupancy networks. Occupancy networks learn continuous decision boundaries,

which is similar to a binary classifier. Finally, the network estimates the occupancy probability of each point in the 3D shape between 0 and 1. Since the decoder needs to decode the coordinates of each input point, it has a longer training time and inference time.

Signed distance function (SDF) decoding The indicator function can be regarded as a special case of the signed distance function, considering only the sign of the SDF values [79]. Park et al. [79] introduced DeepSDF, which is a continuous signed distance function. For a point $p = (x, y, z) \in R^3$ in space, the signed distance function maps the point to $s \in R$. The absolute value of s represents the distance from the point to the nearest surface, and its symbol indicates the point on the surface of the object inside or outside (The schematic diagram of the decoder is shown at the top of Fig. 7). This method can represent complex shapes, but the efficiency of the method is limited by the potential vector optimization in the reasoning stage. Similarly, Wang et al. [124] introduced a deep implicit surface network (DISN) that predicts a symbolic distance function from a 2D image to represent a 3D surface. Given the predicted camera parameters, the points are projected onto a 2D plane to collect multi-scale features. Finally, DISN combines local features, global features and point features to decode the signed distance function values. This method can currently generate a topological structure that is almost consistent with the corresponding object of the input image on the synthetic datasets.

Level sets decoding Michalkiewicz et al. [73] used level sets method to generate continuous 3D object surfaces. Since level sets are defined on regular grids, a 3D deconvolution network can be used to decode the latent layer vector to generate a 3D shape (as shown in the bottom of Fig. 7). This method of level sets representation can reconstruct a natural 3D shape. Under the same architecture, the 3D shape represented by the level set can show more details than the voxel representation.

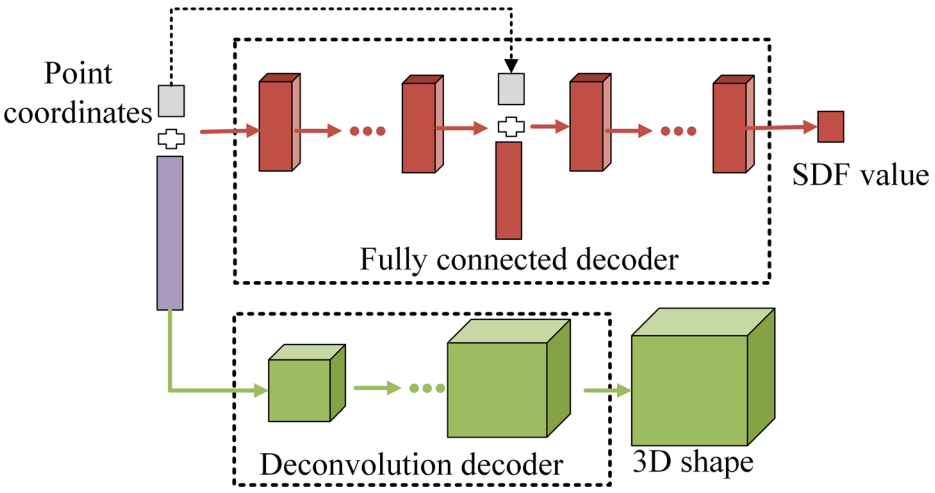

Fig. 7 Two implicit surface decoder architectures

4.2 Higher-level 3d object representation

Different from predicting the shape of 3D objects from low-level features, a decoder based on primitives predicted the parts of an object, and then assembled them into a 3D shape. The 3D shape representation can be regarded as a set of several simple primitives, which can be divided into volumetric primitives and surface primitives. Simple primitives can be defined as geometric elements with fixed types: planes, cuboids, and spheres.

4.2.1 Volumetric primitives decoding

Abstracting a 3D shape into multiple simple entities is a representation of volumetric primitives. Tulsiani et al. [116] proposed a learning framework to abstract complex shapes. The primitive decoder predicts the primitive combination of the different parts and outputs the final shape (see top of Fig. 8). Zou et al. [152] proposed 3D-PRNN, which is a generative recurrent neural network. In detail, the network decodes feature vectors and predicts primitive sequences through a recurrent generator composed of long short-term memory (LSTM) and mixture density network (MDN).

In addition to the decoder using recurrent neural networks above, Niu et al. [76] used a primitive decoder based on a recursive neural network. They proposed to recover a 3D shape structure from a single RGB image, which is a structure represented by cuboids and parts relationships, including connectivity and symmetry. They encoded a low-level mask feature map and an original RGB image separately, and then fed into a decoder to decode the cuboid hierarchical structure after fusion.

Most of the time, this type of method can accurately estimate the true topological structure of an object. However, due to the use of a single volumetric primitive representation, it is difficult to represent correctly complex objects. In addition, these methods cannot reconstruct fine-grained objects.

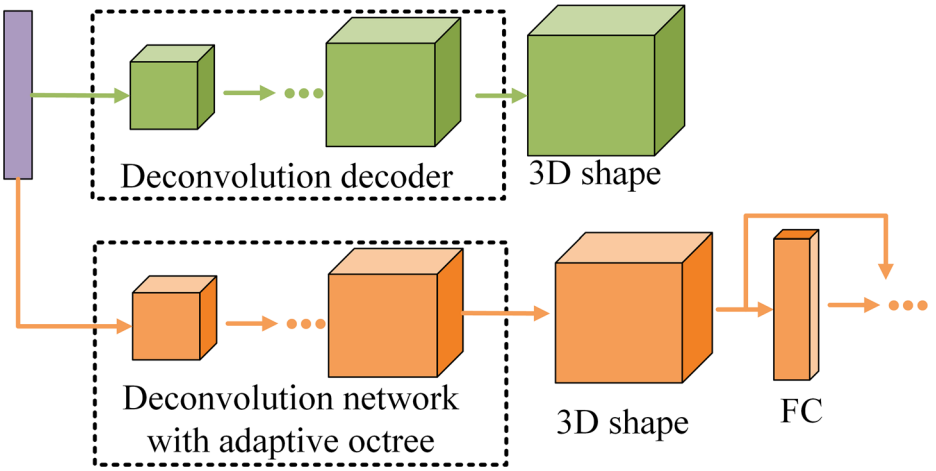

Fig. 8 Two primitive decoder architectures

4.2.2 Surface primitives decoding

Using surface primitives to approximate the surface of three-dimensional object is a simplified method of three-dimensional representation. Wang et al. [121] proposed an adaptive octree-based convolutional neural network (Adaptive O-CNN) for 3D shape decoding. The decoder of Adaptive O-CNN outputs a patch-guided adaptive octree from a latent code (see bottom of Fig. 8). The decoder predicts the patch approximation status for each octant: empty surface, good approximate surface, and poor approximate surface. The surface of the third poor patch is further divided until the max depth is reached. This method greatly reduces network memory requirement, and it can produce high-precision 3D objects. However, this method requires further stitching of the gaps between the individual patches. In addition, the planar patch cannot be very close to an object with a curved feature (e.g. the car wheel).

5 Training details

The above content summarizes the encoders and decoders for single-image 3D object reconstruction. The following sections introduce the use of the loss function and training architecture in these work.

5.1 Loss function

Next, we introduce some loss functions commonly used in the 3D reconstruction network training process.

L_1 loss The loss function can be written as:

$$L_1 = ||Pred_x - GT_y||_1 \tag{2}$$

Here, GT_Y represents a 3D shape of the ground truth, and $Pred_x$ represents a predicted 3D shape.

L_2 loss The loss function is given by:

$$L_2 = ||Pred_x - GT_y||_2^2 \tag{3}$$

where, the meanings of GT_Y and $Pred_x$ correspond to the GT_Y and the $Pred_x$ in Eq. (2). It should be noted that we distinguish between L_2 loss and Mean Squared Error (*MSE*) loss. The standard *MSE* loss is equivalent to averaging the overall L_2 loss.

Binary cross-entropy (BCE) loss The loss function is defined as:

$$L_{BCE} = -\frac{1}{N}\sum_{n=1}^{N}[p_n log q_n + (1-p_n) log(1-q_n)] \tag{4}$$

Here, N represents the total number of voxels in the 3D volume (or the total number of pixels in the 2D image), p_n is the ground truth probability (1 or 0) of the filled voxels (pixels), and q_n is the prediction probability.

Generative adversarial networks loss The loss function of the generative adversarial network includes the loss of the generative network and the discriminative network. Here we introduce the 3D-VAE-GAN adversarial loss function applied by Wu et al. [128], and the goal of the loss function is to maximize/minimize the binary cross-entropy:

$$L_{3D-GAN} = logD(x) + log(1 - D(G(E(I)))) \tag{5}$$

Here, I represents an input image, E represents an image encoder, G represents a generator (also called a decoder), and D represents a discriminator.

Earth mover's distance (EMD) EMD solves the allocation problem, expressed as the minimum cost when one point set S_1 is transformed into another point set S_2, $|S_1| = |S_2|$:

$$d_{EMD}(S_1, S_2) = \min_{\sigma:S_1 \to S_2} \sum_{x \epsilon S_1} \|x - \sigma(x)\|_2 \tag{6}$$

where, $\sigma: S_1 \to S_2$ is a bijection.

Chamfer distance (CD) For each point in point set S_1, the CD finds the nearest neighbor point in point S_2 to calculate the squared distance sum:

$$d_{CD}(S_1, S_2) = \sum_{x \epsilon S_1} \min_{y \epsilon S_2} \|x - y\|_2^2 + \sum_{y \epsilon S_2} \min_{x \epsilon S_1} \|x - y\|_2^2 \tag{7}$$

5.2 Training architectures

According to the existing training architectures, we divide them into three categories: typical training architecture, adversarial training architecture, and hybrid training architecture.

5.2.1 Typical training architecture

A typical training architecture involves optimization of 3D shapes, 2D projections of 3D shapes, and latent layer vectors. For a generated 3D shape, the optimization method can use 3D supervision (see Fig. 9a). This type of training architecture uses a loss function to train the neural network so that the generated 3D data (e.g. volume) is as similar as possible to ground truth 3D data [111, 15, 23, 70, 120, 123, 73]. Some of these studies use voxel-wise binary cross-entropy as the loss function [111, 15], and some use point set loss [23, 70, 123]. Different from the above optimization of the output 3D shape alone, some studies combined with the latent layer vectors for joint optimization [28, 104]. Girdhar et al. [28] encoded the input image and ground truth 3D volumetric grid into the embedding space for joint optimization, respectively (see Fig. 9c). Smith et al. [104] adopted a similar idea. The difference is that they encode the predicted mesh and the ground truth mesh, separately. Both Girdhar [28] and Smith et al. [104] use mean square error to optimize the embedding space.

In addition to using 3D supervision, 2D supervision has also been used (see Fig. 9b). Generally, a 2D supervision method uses a projection-like operation to obtain multiple projection images from various viewpoints of the generated 3D shape for optimization [137, 129, 40, 66, 51, 12]. Some of these studies use L_2 loss [137], a combination of binary cross-entropy loss function and L_1 loss function [66], or standard *MSE* [40]. In addition, 3D supervision combined with 2D supervision can also be adopted [103, 29, 138, 145].

5.2.2 Adversarial training architecture

Typical 3D reconstruction networks need to design corresponding loss functions for training neural networks. The adversarial training architecture uses adversarial mechanisms to train neural networks. This mechanism avoids the need to design complex loss functions in neural networks to optimize the generated 3D shapes. Recently, the generative adversarial networks (GAN) has shown a strong ability in image generation [85, 59]. Inspired by this, Wu et al. [128] extended the 2D generative adversarial networks to a 3D GAN. They used the encoder of VAE [55] to encode the image into a latent vector and feed it into 3D-GAN for 3D reconstruction tasks (see Fig. 10). Similarly, Smith et al. [102] used a variant GAN to accomplish the 3D reconstruction from a single image. Although the GAN method has achieved good 3D reconstruction results, it is still necessary to solve the problem of self-training stability.

In addition to the above 3D supervision method, there is also 2D supervision method. Gadelha et al. [26] used 3D shape projection images of different viewpoints for adversarial training.

5.2.3 Hybrid training architecture

For the existing typical training architecture and adversarial network training architecture, a natural idea is to use both together [130, 46, 50]. Under 3D supervision, Wu et al. [130] combined the typical training architecture and the adversarial training architecture. The difference is that Kato et al. [50] applied two training architectures under 2D supervision. In

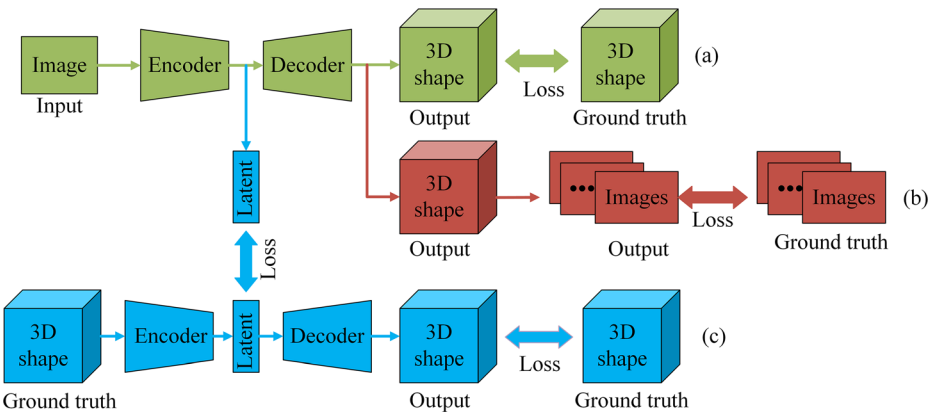

Fig. 9 Typical training architectures

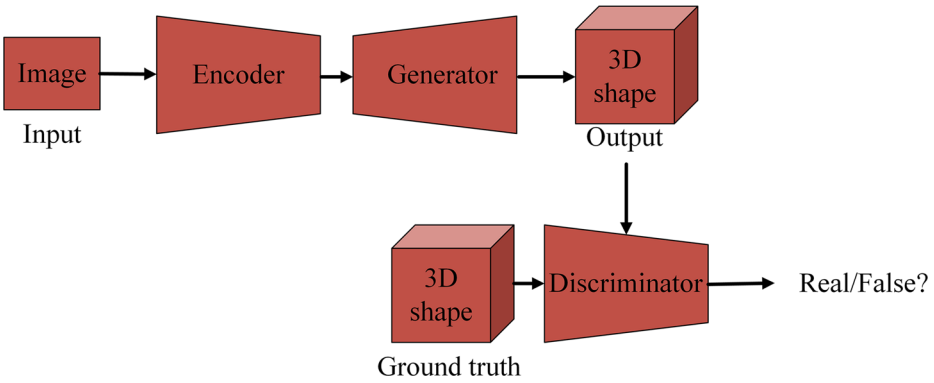

Fig. 10 An adversarial training architecture

addition to applying 3D supervision or 2D supervision alone, Jiang et al. [46] adopted 2D supervision combined with 3D supervision.

6 Datasets and evaluation metrics

It is uncertain to reconstruct 3D objects from a single image, and it requires estimation of the occluded object structure. Currently, several datasets for single image reconstruction are published. Table 3 provides a brief summary of these datasets: ShapeNet [7], ModelNet [127], PASCAL 3D+ [133], IKEA [64], Pix3D [110].

6.1 ShapeNet

As a subset of ShapeNet datasets, ShapeNetCore contains 55 common object categories and has approximately 51,300 unique 3D models. In addition, the dataset was manually verified category labels and alignment annotations. At present, this dataset is the most commonly used single image 3D reconstruction synthetic dataset, and its 3D models have no corresponding 2D images. Figure 11 shows some examples of the dataset rendering pictures and ground truth CAD models.

6.2 ModelNet

ModelNet currently contains three classes of subsets: ModelNet10, ModelNet40, and Aligned 40-Class ModelNet. The dataset covers 662 object categories and has approximately 127,915 CAD models. ModelNet10 and Aligned 40-Class ModelNet manually align the orientation of the CAD model. This dataset is also a synthetic dataset, and its

Table 3 Popular datasets for 3D Reconstruction

DataSet	Data Type	Classes	Models /Images	Website
ShapeNet [7]	Synthetic	55	51,300/-	https://www.shapenet.org/
ModelNet [127]	Synthetic	662	127,915/-	http://modelnet.cs.princeton.edu/
PASCAL3D+ [133]	Real	12	3000+/-	http://cvgl.stanford.edu/projects/pascal3D.html
IKEA [64]	Real	6	219/759	http://ikea.csail.mit.edu/
Pix3D [110]	Real	9	395/10,069	http://pix3D.csail.mit.edu/

model is divided into a training set and a test set. Figure 12 shows some examples of this dataset.

6.3 PASCAL3D+

The two datasets introduced above are synthetic datasets, and the 3D models of these datasets do not have corresponding 2D images. Therefore, 2D images need to be rendered from 3D models before training neural networks. PASCAL3D + is a dataset enhanced by PASCAL VOC 2012 [22] using 3D annotation. The dataset contains 12 rigid categories, with an average of more than 3,000 object instances per category. However, it is relatively rough alignment between the 2D image and the 3D model. The 3D models of the dataset have corresponding indoor or outdoor pictures. Because the 2D pictures of this dataset have multiple objects, occlusion, and truncation, it is a challenging dataset. Figure 13 shows some examples of this dataset.

Fig. 11 Partial rendering pictures and CAD models on ShapeNet

Fig. 12 Some CAD models on ModleNet

6.4 IKEA

IKEA is a dataset of indoor 3D models. The dataset consists of six categories, about 759 pictures and 219 3D models. All images in the dataset are accurately annotated using available models (90 different models). The dataset consists of two parts: IKEAobject and IKEAroom. IKEAobject is a simple scene that contains larger objects. IKEAroom is the opposite. Figure 14 shows some examples of the dataset.

6.5 Pix3D

The datasets described above have respective limitations. First, the synthetic datasets lack corresponding real pictures. Then, for PASCAL3D, its CAD models have the corresponding real images, but there are lack of accurate alignments between the CAD models and the images. Finally, accurate alignment between 2D-3D can be guaranteed, but the dataset (IKEA) contains fewer objects. To make up for these shortcomings, Pix3D was extended on the IKEA dataset. Pix3D is a pixel-level alignment dataset between 3D shapes and their silhouettes. The dataset contains 9 object classes with 395 3D shapes and 10,069 pictures. Each 3D shape has a corresponding real image in different environments, but the environment of this dataset is limited to indoors. Figure 15 shows some examples of this dataset.

6.6 Other datasets

ObjectNet3D [134] has 100 categories and 90,127 images. Each picture contains multiple objects, a total of 201,888 objects and 44,147 3D shapes. However, the 2D silhouettes of the images in this dataset are not precisely aligned with the 3D objects. In addition, there is an Online Products dataset [107] for qualitative evaluation of model expression capabilities. The dataset contains 22,634 classes with a total of 120,053 real pictures.

Fig. 13 Partial PASCAL3D + pictures and CAD models

6.7 Evaluation metrics

We will introduce some evaluation metrics used in 3D reconstruction. CD and EMD metrics in Sect. 5.1 have been introduced. Different 3D shape representations can be converted to each other after processing to use different evaluation metrics.

Intersection-over-Union (IoU) This IoU calculates the intersection and union ratio between a predicted object volume and a ground truth object volume. If the calculated IoU is larger, it means that the reconstructed 3D object is better. The IoU can be written as below:

$$\mathrm{IoU} = \frac{\sum_{i,j,k}\left[\mathrm{I}(x_{(i,j,k)} > t)\mathrm{I}\left(y_{(i,j,k)}\right)\right]}{\sum_{i,j,k}\left[\mathrm{I}(x_{(i,j,k)} > t) + \mathrm{I}\left(y_{(i,j,k)}\right)\right]} \tag{8}$$

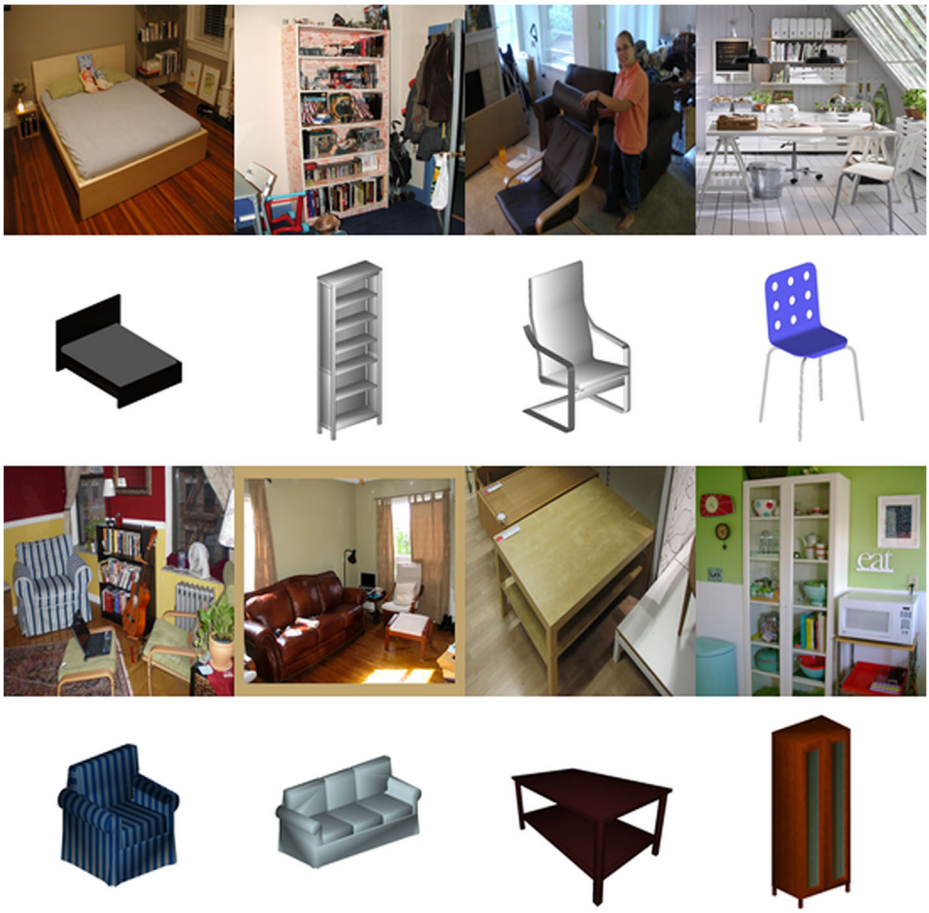

Fig. 14 Some IKEA pictures and CAD models

Here, $x(i, j, k) \in [0, 1]$, $y(i, j, k) \in [0, 1]$. They represent the occupancy values at each voxel (i, j, k), respectively. $I(\cdot)$ represents the indicator function, and t represents the voxelization threshold, which is generally taken as 0.40 or 0.45.

F-score Y represents the ground truth and X represents the reconstructed point set. For a reconstructed point $x \in X$, the accuracy rate of X at any distance threshold d is defined as:

$$P(d) = \frac{100}{|X|} \sum_{x \epsilon X} \left[\min_{y \epsilon Y} \|x - y\| \right. \tag{9}$$

Where $[\cdot]$ is the Iverson bracket. $P(d)$ represents a percentage. Similarly, the recall rate is defined as:

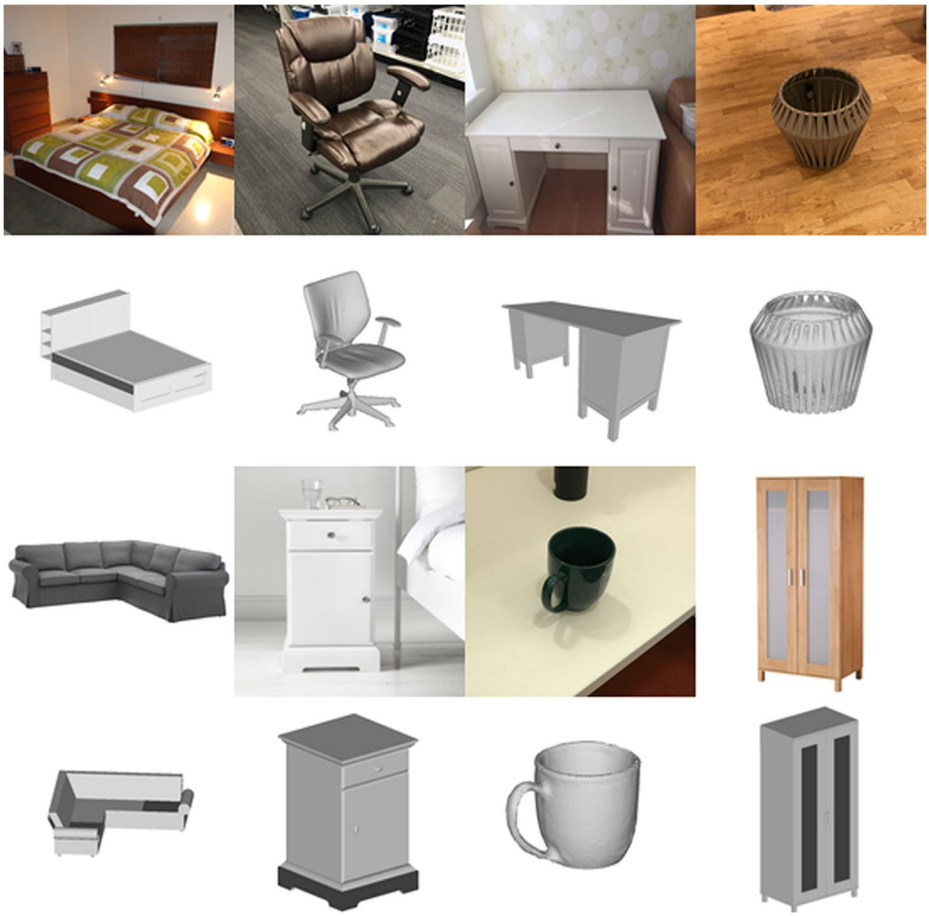

Fig. 15 Some Pix3D pictures and CAD models

$$R(d) = \frac{100}{|Y|} \sum_{y \epsilon Y} [\min_{x \epsilon X} \|y - x\| \tag{10}$$

F-score can be expressed as a combination of accuracy and recall:

$$F(\tau) = \frac{2P(\tau)R(\tau)}{P(\tau) + R(\tau)} \tag{11}$$

The F-score is the harmonic mean of the precision and recall rate at a given threshold τ.

7 Comparison

In the following sections, we compare the advantages and disadvantages of different methods through several aspects.

7.1 Comparison of reconstruction results of different datasets

In the first comparison experiment, we select 13 categories of synthetic dataset ShapeNet and compare it with the real dataset Pix3D. "*" indicates that the result was reported by Xie et al. [136]. It can be seen from Tables 4 and 5 that the reconstruction results on the synthetic dataset are better than the reconstruction results on the real dataset. As discussed in Sect. 2.5, most models are difficult to adapt to complex and varied scenarios due to the large differences between the synthetic and real datasets.

7.2 Comparison under different training architectures

We compare the performance of various 3D reconstruction networks under different training architectures using 2D supervision, 3D supervision, or a combination of them, as shown in Table 5. The performance of various networks in a typical training architecture is similar, and the highest MIoU score obtained under different supervision has little difference. In the method using only 2D supervision, Kato et al. [50] achieved higher performance using a typical training architecture combined with a generative adversarial approach. In addition, Jiang et al. [46] used a hybrid training architecture to increase the benchmark score from 0.560 to 0.712 under 2D and 3D supervision. In general, the 3D reconstruction network under various supervision methods can produce competitive results. On the other hand, a hybrid training architecture approach can achieve better results.

7.3 Comparison of memory requirements and computing time

Table 6 shows the memory requirements and iteration time for some methods. The results with "#" were reported by Xie et al. [136]. The results with "*" were reported by Park et al. [79]. For voxel-based reconstruction methods on regular grids [15, 136], their reconstruction resolution is limited by memory size and computation time. Tatarchenko et al. [111] used an octree representation to significantly reduce the memory requirements during high-resolution 3D reconstruction. However, the iterative time of this method at 256^3 resolution is 40 times longer than that at 32^3 resolution. In addition, Shen et al. [98] used frequency-domain slicing to reconstruct 3D shapes from 2D space. Compared with OGN [111], this method costs more in memory, but it reduces the iteration time. In terms of model inference time, the time required by Park et al. [79] using the implicit decoder method is about 30 times that of the high-resolution 3D voxel reconstruction method [111]. The long inference time of the model is the most important problem to be solved by this type of method.

Table 4 Comparison of synthetic and real datasets. MIoU is calculated at 32^3 resolution

Method/Year	Decoder	Datasets	MIoU
Choy et al. [15]/2016	Voxel	ShapeNet	0.560
		Pix3D	0.136*
Xie et al. [136]/2019	Voxel	ShapeNet	0.661
		Pix3D	0.288
Sun et al. [110]/2018	Voxel	Pix3D	0.282
Wu et al. [130]/2018	Voxel	Pix3D	0.284

Table 5 Comparison of various supervision methods under different training architectures. The MIOU of all experimental results calculates 13 common categories of shapenet at 32^3

Method/ Year	Training	Network description	Supervision	MIoU
Choy et al. [15]/2016	Typical	Encoder + LSTM + Decoder	3D	0.560
Tatarchenko et al. [111]/2017	Typical	Encoder + Octree Decoder	3D	0.596
Fan et al. [23]/2017	Typical	Encoder+(FC + Deconv) decoder	3D	0.640
Yang et al. [141]/2019	Typical	Encoder + AttSets + Decoder	3D	0.642
Xie et al. [136]/2019	Typical	Encoder + Decoder + Refiner	3D	0.661
Kato et al. [51]/2018	Typical	Encoder + FC decoder + Renderer	2D	0.602
Kato et al. [50]/2019	Hybrid	Encoder + 2D decoder + Renderer + Discriminator	2D	0.655
Liu et al. [67]/2019	Typical	Encoder + FC decoder + Renderer	2D	0.646
Yang et al. [138]/2018	Typical	Recurrent (Encoder + Decoder)	2D + 3D	0.600
Shen et al. [98]/2019	Typical	Encoder + 2D decoder + Fourier transform	2D + 3D	0.605
Jiang et al. [46]/2018	Hybrid	Encoder-decoder hourglass + FC + GAL	2D + 3D	0.712
Zeng et al. [145]/2018	Typical	Image to depth + Sparse + Dense	2D + 3D	0.648

7.4 Comparison of different decoders

In this section, we give the results of single image reconstruction of 13 categories on the ShapeNet dataset for different 3D shape decoders, as shown in Table 7. At low resolution 32^3, the point cloud-based output representation method achieved the highest MIoU score. Jiang et al. [46] used an hourglass version of the point-set generative network [23], and combined geometric loss with conditional adversarial loss to optimize 3D point clouds. In the end, they raised the baseline from 0.560 to 0.712. Compared with the MIoU metric, recent a research result suggest that the F-score is used as a metric to measure the accuracy of 3D objects [112]. Gkioxari et al. [29] first predicted a rough voxelized object from an input image, and then converted it to a mesh and refined it. This hybrid reconstruction method increased the benchmark score from 59.72–75.83%.

In recent years, decoders based on different 3D shape representations have attempted to reconstruct fine-grained objects. At present, 3D shapes generated by decoders based on meshes or implicit surfaces have better visual appeal and physical structure. For most mesh-based decoders, the reconstructed 3D shape topology is limited by the initial deformed mesh model [51, 50, 67]. The network based on the implicit surface decoder

Table 6 Comparison of memory consumption and iteration time calculations of different methods. Batch size 1. Res.: Resolution, Mem.: Memory, Iter.: Iteration, Inf.: Inference

Method/ Year	Res.	GPU/number	Mem.(GB)	Iter.(ms)	Inf.(ms)
Choy et al. [15]/2016	32^3	GTX 1080Ti/1	1.37	385.8	73.35#
Xie et al. [136]/2019	32^3	GTX 1080Ti/1	2.66	81.91	9.90#
Tatarchenko et al. [111]/2017	32^3	TitanX Maxwell/1	0.29	16	37.90#
	256^3		0.54	640	320*
Groueix et al. [30]/2018	25 patches	Nvidia/-	0.172	-	320*
Park et al. [79]/2019	-	Nvidia/-	0.0074	-	9720
Shen et al. [98]/2019	256^3	1080 Maxwell/1	1.93	470	-

learns a continuous volume field to represent the surface of an object. This method can generate realistic high-resolution 3D shapes.

8 3D reconstruction applications

At present, there are few researches on single image 3D reconstruction based on deep learning. Here, we briefly introduce 3D face reconstruction and 3D human shape and pose estimation.

8.1 Reconstruction and recognition of 3D face

The goal of 3D face reconstruction is to reconstruct a 3D face from a single 2D face image or multiple 2D face images. 3D face reconstruction technology has a wide range of extended application examples, such as 3D face recognition. Deep learning-based single image 3D face reconstruction has been partially studied using voxel-based expression methods [42, 96]. Jackson et al. [42] used voxel-based methods to reconstruct the corresponding 3D face from a single 2D face image. This method reconstructs a 3D face shape from a single face image in an end-to-end manner, and can process images with different facial poses and expressions. Sharma et al. [96] designed a 3D face recognition system using voxel-based 3D reconstruction methods, and achieved state-of-the-art results in terms of computation time and recognition accuracy.

8.2 Shape and pose estimation of 3D human

The shape and pose of 3D human estimation technology have a wide range of application scenarios in real life. 3D human shape reconstruction technology can be used for virtual fitting

Table 7 Comparison of various supervision methods under different training architectures. MIOU and F-scores are calculated in 13 common classes of shapenet. The voxel resolution and threshold τ are 32^3 and 10^{-4}, respectively

Method/ Year	Decoder	Network description	MIoU	F(τ)%
Choy et al. [15]/2016	Voxel	Encoder(ResNet) + LSTM + Decoder(ResNet)	0.560	-
Yang et al. [138]/2018	Voxel	Recurrent (Encoder + Decoder)	0.600	-
Yang et al. [141]/2019	Voxel	Encoder + AttSets + Decoder	0.642	-
Xie et al. [136]/2019	Voxel	Encoder + Decoder + Refiner	0.661	-
Tatarchenko et al. [111]/2017	Voxel	Encoder + Octree Decoder	0.596	-
Richter et al. [88]/2018	Voxel	Encoder + 2D decoder + Depth fusion	0.641	-
Smith et al. [103]/2018	Voxel	Encoder + Decoder + Depth carving	0.610	66.39
Shen et al. [98]/2019	Voxel	Encoder + 2D decoder + Fourier transform	0.605	-
Fan et al. [23]/2017	Point	Encoder+(FC + Deconv) decoder	0.640	-
Jiang et al. [46]/2018	Point	Encoder-decoder hourglass + FC + GAL	0.712	-
Zeng et al. [145]/2018	Point	Image to depth + Sparse + Dense	0.648	-
Kato et al. [51]/2018	Mesh	Encoder + FC decoder + Mesh renderer	0.602	-
Kato et al. [50]/2019	Mesh	Encoder + 2D decoder + Discriminator	0.655	-
Liu et al. [67]/2019	Mesh	Encoder + FC decoder + Renderer	0.646	-
Mescheder et al. [30]/2019	Implicit	Encoder(ResNet-18) + FC decoder(ResNet)	0.571	-
Wang et al. [120]/2018	Mesh	Encoder + GCN decoder	-	59.72
Smith et al. [104]/2019	Mesh	Encoder + 0N-GCN decoder	-	67.37
Gkioxari et al. [29]/2019	Mesh	Image to voxel to mesh + Mesh refiners	-	75.83

and special effects production. At present, many studies have used voxel-based or mesh-based methods to reconstruct 3D human shapes from a single image [117, 43, 57, 150]. Moreover, 3D human pose estimation can be used for motion capture and motion recognition. Motion capture technology can be widely used in the production of film and television animation. Motion recognition technology can estimate human motion behavior, which can be applied to unmanned driving scenarios. At present, some studies have used voxel-based and mesh-based methods to study the 3D human pose estimation of a single image [80, 81].

9 Conclusions and future directions

In this paper, we have reviewed the 3D object reconstruction of a single image using deep learning methods in recent years. This review focuses on three-dimensional decoding methods for different 3D shape representations. In terms of the voxel-based and point cloud-based 3D reconstruction methods proposed earlier, they increased the reconstruction resolution to generate more detailed 3D shapes. Due to memory limitations, some voxel-based methods use sparse voxel representation or transfer high-resolution 3D reconstruction parts to 2D space. Experimental results show that these methods can reconstruct high-resolution 3D objects with better visual quality. However, it should be noted that the high-resolution 3D shape reconstruction results in these methods depend on the reconstructed low-resolution 3D volume shapes. Currently, mesh-based and implicit surface-based methods can generate 3D shapes with higher visual quality. Most of these methods use low-level features for 3D shape inference, and occasionally incorrect 3D reconstruction results appear in the reconstruction results. Conversely, primitive methods that use higher-level features are more able to reconstruct correct 3D shapes. There is no absolute standard for the use of 3D shape representation for the reconstruction of 3D objects from a single image. Different 3D representations have advantages and disadvantages. In addition, the training details of the 3D reconstruction network need to be further explored. At present, the hybrid training architecture has greater advantages. The deep learning-based method can reconstruct a better 3D shape by inputting a single picture, which is far beyond the traditional methods. However, most of the methods lack the universality of traditional methods and are difficult to adapt to different reconstruction scenarios.

Based on the review of this paper, we propose several possible research directions in the future:

(1) Combination of 3D expression at different levels. Low-level 3D representations can use pixels, silhouettes and other information for 3D object reconstruction. However, most current research methods lack an understanding of part hierarchy or part relationships. Although the evaluation score is high, the comparison with the real object structure may be another structure. The higher-level expression can relatively accurately capture the structure of an object, which guarantees the accuracy of reconstructing the structure of the object in a certain extent. Therefore, we suggest to include a high-level understanding of low-level 3D representations in the future [53].

(2) Establish lifelike or real scene training datasets. At present, there are many methods for 3D object reconstruction of a single image based on deep learning. Most studies use clean backgrounds or rendered backgrounds, but they still differ significantly from real datasets. Therefore, these methods perform poorly on wild datasets. At present, since the number of 2D images and 3D models contained in the real scene dataset is too small, it is

difficult to use the advantages of deep learning methods for training. This problem can be solved by building a large-scale training dataset in real scenes.

(3) Combining traditional 3D reconstruction with deep learning. Deep learning-based methods can learn prior information through neural networks. However, the current learning-based method for 3D reconstruction of a single image is limited to the dataset and lacks generalization capabilities. Similarly, the traditional 3D reconstruction method needs to add various prior conditions, which is a limitation of the traditional method. However, traditional methods are universal. Therefore, future work can try to use deep learning combined with traditional 3D reconstruction algorithms to accomplish 3D reconstruction from a single image.

(4) Combining 2D object detection with 3D reconstruction [29, 140]. In recent years, 2D object detection based on deep learning has achieved remarkable results in various fields. Currently, most studies on 3D shape reconstruction use synthetic datasets. However, the real scene dataset has a complex environment, there are multiple objects, and the shape complexity of each object may vary greatly. The conventional method is to predict the 3D shape after cropping the input image, which increases the workload and gets poor reconstruction results. The method combined with 3D object reconstruction after extracting the object mask can better deal with complex environment, complex shape reconstruction and multi-object reconstruction problems. In addition, objects can also be divided into multiple parts for reconstruction through semantic segmentation [37].

(5) 3D shape representation based on GAN or combined GAN. Currently, 3D reconstruction models based on GAN networks are applied to various 3D representations. Although the mechanism of GAN itself introduces noise or causes unstable training problems, this method still shows good potential. In addition, combining GAN as part of the model for learning shape or view prior has been applied in different 3D representation methods, and it has achieved good results. Therefore, in the future, GAN can be considered as an auxiliary method for 3D reconstruction.

(6) Reconstruction of rigid and non-rigid shapes. The real world is a combination of rigid and non-rigid objects. Most 3D reconstruction of single images based on deep learning focuses on rigid or non-rigid shape reconstruction [69]. However, the object that needs to be dealt with in specific applications is not fixed in a certain category, such as driverless cars, movie production. In this case, a more generalized model is needed that can handle the reconstruction of rigid and non-rigid objects [117]. In addition, under such conditions, it is a challenging task to optimize the deep learning algorithm to meet the requirements of real-time operation.

(7) Applied to 3D data exchange. Recent studies have shown that deep learning technology is very helpful for solving some of the problems in the manufacturing industry [17]. With the continuous growth of product development, collaborative manufacturing has become a trend. However, different companies or departments will choose different CAD systems for their own reasons. This leads to inconsistencies between the source model and the target model when data is converted between different CAD systems. Therefore, it is necessary to solve the problem of 3D data exchange between heterogeneous CAD systems [146, 131, 132]. The 3D data exchange of heterogeneous systems also involves how to reconstruct 3D models across different CAD systems, which has high potential value for real-world application scenarios.

Three-dimensional reconstruction of images is an important research problem in the field of computer vision. Compared with the previous traditional methods, 3D reconstruction of images using deep learning is a new method. Although there are still many problems with the method of deep learning, this method has achieved significant research results. It is believed that with the further study of deep learning related knowledge, the method of image 3D reconstruction can be improved.

Acknowledgements The authors are highly thankful to the Development Research Center of Guangxi Relatively Sparse-populated Minorities (ID: GXRKJSZ201901), to the Natural Science Foundation of Guangxi Province (NO.2018GXNSFAA281164),This research was financially supported by the project of outstanding thousand young teachers' training in higher education institutions of Guangxi, Guangxi Colleges and Universities Key Laboratory Breeding Base of System Control and Information Processing.

References

1. Alldieck T, Magnor M, Bhatnagar BL, Theobalt C, Pons-Moll G (2019) Learning to reconstruct people in clothing from a single RGB camera. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1175–1186
2. Atick JJ, Griffin PA, Redlich AN (1996) Statistical approach to shape from shading: reconstruction of three-dimensional face surfaces from single two-dimensional images. Neural Comput 8(6):1321–1340
3. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495
4. Baka N, Kaptein BL, Bruijne MD, Walsum TV, Giphart WJ, Lelieveldt BPF (2011) 2D-3D shape reconstruction of the distal femur from stereo x-ray imaging using statistical shape models. Med Image Anal 15(6):840–850
5. Blanz V, Vetter T (1999) A morphable model for the synthesis of 3D faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, pp 187–194
6. Bronstein MM, Bruna J, Lecun Y, Szlam A, Vandergheynst P (2017) Geometric deep learning: going beyond euclidean data. IEEE Signal Process Mag 34(4):18–42
7. Chang AX, Funkhouser T, Guibas L et al (2015) Shapenet: an information-rich 3D model repository. arXiv preprint arXiv:1512.03012
8. Charles RQ, Su H, Mo K, Guibas LJ (2017) Point net: deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 77–85
9. Chen Z, Zhang H (2019) Learning implicit fields for generative shape modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5939–5948
10. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille A (2017) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
11. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille A (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision, pp 801–818
12. Chen W, Ling H, Gao J, Smith E, Lehtinen J et al (2019) Learning to predict 3D objects with an interpolation-based differentiable renderer. In: Proceedings of the Advances in Neural Information Processing Systems, pp 9605–9616
13. Chinaev N, Chigorin A, Laptev I (2018) Mobileface: 3D face reconstruction with efficient CNN regression. In: Proceedings of the European Conference on Computer Vision, pp 15–30
14. Choi J, Medioni G, Lin Y, Silva L, Regina O, Pamplona M, Faltemier TC (2010) 3D face reconstruction using a single or multiple views. In: Proceedings of the International Conference on Pattern Recognition, pp 3959–3962
15. Choy CB, Xu D, Gwak J, Chen K, Savarese S (2016) 3D-r2n2: a unified approach for single and multi-view 3D object reconstruction. In: Proceedings of the European Conference on Computer Vision, pp 628–644
16. Cimpoi M, Maji S, Kokkinos I, Mohamed S, Vedaldi A (2014) Describing textures in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3606–3613
17. Dekhtiar J, Durupt A, Bricogne M, Eynard B, Rowson H, Kiritsis D (2018) Deep learning for big data applications in CAD and PLM–research review, opportunities and case study. Comput Ind 100:227–243

18. Dou P, Kakadiaris IA (2018) Multi-view 3D face reconstruction with deep recurrent neural networks. Image Vis Comput 80:80–91
19. Dou P, Shah K, Kakadiaris IA (2017) End-to-end 3D face reconstruction with deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5908–5917
20. Dovgard R, Basri R (2004) Statistical symmetric shape from shading for 3D structure recovery of faces. In: Proceedings of the European Conference on Computer Vision, pp 99–113
21. Eckart B, Kim K, Troccoli A, Kelly A, Kautz J (2016) Accelerated generative models for 3D point cloud data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5496–5505
22. Everingham M, Gool LV, Williams CKI, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338
23. Fan H, Su H, Guibas L (2017) A point set generated network for 3D object reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 605–613
24. Feng Y, Wu F, Shao X, Wang Y, Zhou X (2018) Joint 3D face reconstruction and dense alignment with position map regression network. In: Proceedings of the European Conference on Computer Vision, pp 534–551
25. Furukawa Y, Curless B, Seitz SM, Szeliski R (2010) Towards internet-scale multi-view stereo. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp 1434-1441
26. Gadelha M, Maji S, Wang R (2017) 3D shape induction from 2D views of multiple objects. In: Proceedings of the International Conference on 3D Vision, pp 402–411
27. Genova K, Cole F, Maschinot A, Sarna A, Vlasic D, Freeman WT (2018) Unsupervised training for 3D morphable model regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 8377–8386
28. Girdhar R, Fouhe DF, Rodriguez M, Gupta A (2016) Learning a predictable and generative vector representation for objects. In: Proceedings of the European Conference on Computer Vision, pp 484–499
29. Gkioxari G, Malik J, Johnson J (2019) Mesh r-cnn. arXiv preprint arXiv:1906.02739
30. Groueix T, Fisher M, Kim VG, Russell BC, Aubry M (2018) A papier-mâché approach to learning 3D surface generated. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 216–224
31. Gwak JY, Choy CB, Chandraker M, Garg A, Savarese S (2017) Weakly supervised 3D reconstruction with adversarial constraint. In: Proceedings of the International Conference on 3D Vision, pp 263–272
32. Ham H, Wesley J, Hendra H (2019) Computer vision based 3D reconstruction: a review. Int J Electr Comput Eng 9(4):2394–2402
33. Häne C, Tulsiani S, Malik J (2017) Hierarchical surface prediction for 3D object reconstruction. In: Proceedings of International Conference on 3D Vision, pp 76–84
34. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778
35. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2961–2969
36. Hepp B, Nießner M, Hilliges O (2018) Plan3D: viewpoint and trajectory optimization for aerial multi-view stereo reconstruction. ACM Trans Graphics 38(1):1–17
37. Huang Q, Wang H, Koltun V (2015) Single-view reconstruction via joint analysis of image and shape collections. ACM Trans Graph 34(4):1–10
38. Huang S, Qi S, Zhu Y, Xiao Y, Xu Y, Zhu SC (2018) Holistic 3D scene parsing and reconstruction from a single rgb image. In: Proceedings of the European Conference on Computer Vision, pp 187–203
39. Huang PH, Matzen K, Kopf J, Ahuja N, HuangJB (2018) Deepmvs: learning multi-view stereopsis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2821–2830
40. Insafutdinov E, Dosovitskiy A (2018) Unsupervised learning of shape and pose with differentiable point clouds. In: Proceedings of the Advances in Neural Information Processing Systems, pp 2802–2812
41. Jack D, Pontes JK, Sridharan S et al (2018) Learning free-form deformations for 3D object reconstruction. In: Proceedings of the Asian Conference on Computer Vision, pp 317–333
42. Jackson AS, Bulat A, Argyriou V, Tzimiropoulos G (2017) Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1031–1039
43. Jackson AS, Manafas C, Tzimiropoulos G (2018) 3D human body reconstruction from a single image via volumetric regression. In: Proceedings of the European Conference on Computer Vision, pp 64–77
44. Jeon Y, Kim J (2017) Active convolution: learning the shape of convolution for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4201–4209

45. Jiang L, Zhang J, Deng B, Li H, Liu L (2018) 3D face reconstruction with geometry details from a single image. IEEE Trans Image Process 27(10):4756–4770
46. Jiang L, Shi S, Qi X, Jia J (2018) Gal: geometric adversarial loss for single-view 3D-object reconstruction. In: Proceedings of the European Conference on Computer Vision, pp 802–816
47. Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the European Conference on Computer Vision, pp 694–711
48. Kanazawa A, Tulsiani S, Efros AA, Malik J (2018) Learning category-specific mesh reconstruction from image collections. In: Proceedings of the European Conference on Computer Vision, pp 371–386
49. Kar A, Tulsiani S, Carreira J, Malik J (2015) Category-specific object reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1966–1974
50. Kato H, Harada T (2019) Learning view priors for single-view 3D reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 9778–9787
51. Kato H, Ushiku Y, Harada T (2018) Neural 3D mesh renderer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3907–3916
52. Kemelmacher-Shlizerman I (2013) Internet based morphable model. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3256–3263
53. Khan SH, Guo Y, Hayat M, Barnes N (2019) Unsupervised primitive discovery for improved 3D generative modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 9739–9748
54. Kim J, Lee JK, Lee KM (2016) Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1646–1654
55. Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114
56. Klokov R, Lempitsky V (2017) Escape from cells: deep kd-networks for the recognition of 3D point cloud models. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2380–7504
57. Kolotouros N, Pavlakos G, Daniilidis K (2019) Convolutional mesh regression for single-image human shape reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4501–4510
58. Kulon D, Wang H, Güler RA, Bronstein M, Zafeifiou S (2019) Single image 3D hand reconstruction with mesh convolutions. arXiv preprint arXiv:1905.01326
59. Larsen ABL, Sønderby SK, Larochelle H, Winther O (2015) Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300
60. Le T, Duan Y (2018) Pointgrid: a deep network for 3D shape understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 9204–9214
61. Ledig C, Theis L, Huszár F et al (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp 4681–4690
62. Li CL, Zaheer M, Zhang Y, Poczos B, Salakhutdinov R (2018) Point cloud gan. arXiv preprint arXiv: 1810.05795
63. Li K, Pham T, Zhan H, Reid I (2018) Efficient dense point cloud object reconstruction using deformation vector fields. In: Proceedings of the European Conference on Computer Vision, pp 497–513
64. Lim JJ, Pirsiavash H, Torralba A (2013) Parsing ikea objects: Fine pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2992–2999
65. Lim B, Son S, Kim H, Nah S, Lee KM (2017) Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 136–144
66. Lin CH, Kong C, Lucey S (2018) Learning efficient point cloud generated for dense 3D object reconstruction. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, pp 7114–7121
67. Liu S, Li T, Chen W, Li H (2019) Soft rasterizer: a differentiable renderer for image-based 3D reasoning. arXiv preprint arXiv:1904.01786
68. Loh AM, Hartley RI (2005) Shape from non-homogeneous, non-stationary, anisotropic, perspective texture. In: Proceedings of the 2005 British Machine Vision Conference, pp 5:69–78
69. Lun Z, Gadelha M, Kalogerakis E, Maji S, Wang R (2017) 3D shape reconstruction from sketches via multi-view convolutional networks. In: Proceedings of the International Conference on 3D Vision, pp 67–77
70. Mandikal P, Radhakrishnan VB (2019) Dense 3D point cloud reconstruction using a deep pyramid network. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pp 1052–1060
71. Mandikal P, Murthy N, Agarwal M, Babu RV (2018) 3D-lmnet: latent embedding matching for accurate and diverse 3D point cloud reconstruction from a single image. arXiv preprint arXiv:1807.07796

72. Mescheder L, Oechsle M, Niemeyer M, Nowozin S, Geiger A (2019) Occupancy networks: learning 3D reconstruction in function space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4460–4470
73. Michalkiewicz M, Pontes JK, Jack D, Baktashmotlagh M, Eriksson A (2019) Deep level sets: implicit surface representations for 3D Shape inference. arXiv preprint arXiv:1901.06802
74. Montefusco LB, Lazzaro D, Papi S, Guerrini C (2010) A fast compressed sensing approach to 3D MR image reconstruction. IEEE Trans Med Imaging 30(5):1064–1075
75. Navaneet KL, Mandikal P, Agarwal M, Babu RV (2019) CAPNet: continuous approximation projection for 3D point cloud reconstruction using 2d supervision. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 33:8819–8826
76. Niu C, Li J, Xu K (2018) Im2struct: recovering 3D shape structure from a single RGB image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4521–4529
77. Omran M, Lassner C, Pons-Moll G, Gehler P, Schiele B (2018) Neural body fitting: unifying deep learning and model based human pose and shape estimation. In: Proceedings of the International Conference on 3D Vision, pp 484–494
78. Oswald MR, Töppe E, Nieuwenhuis C, Cremers D (2013) A review of geometry recovery from a single image focusing on curved object reconstruction. Innovations for Shape Analysis, pp 343–378
79. Park JJ, Florence P, Straub J, Newcombe R, Lovegrove S (2019) Deepsdf: learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 165–174
80. Pavlakos G, Zhou X, Derpanis KG, Daniilidis K (2017) Coarse-to-fine volumetric prediction for single-image 3D human pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7025–7034
81. Pavlakos G, Zhu L, Zhou X, Daniilidis K (2018) Learning to estimate 3D human pose and shape from a single color image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 459–468
82. Pollefeys M, Koch R, Vergauwen M, Gool LV (2000) Automated reconstruction of 3D scenes from sequences of images. ISPRS J Photogramm Remote Sens 55(4):251–267
83. Pontes JK, Kong C, Sridharan S, Lucey S, Eriksson A, Fookes C (2018) Image2mesh: a learning framework for single image 3D reconstruction. In: Proceedings of the Asian Conference on Computer Vision, pp 365–381
84. Qi CR, Yi L, Su H, Guibas LJ (2017) Pointnet++: deep hierarchical feature learning on point sets in a metric space. In: Proceedings of the Advances in Neural Information Processing Systems, pp 5099–5108
85. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv: 1511.06434
86. Rezende DJ, Eslami SMA, Mohamed S, Battaglia P, Jaderberg M, Heess N (2016) Unsupervised learning of 3D structure from images. In: Proceedings of the Advances in Neural Information Processing Systems, pp 4996–5004
87. Richardson E, SelaLUN M, Or-EI R, Kimmel R (2017) Learning detailed face reconstruction from a single image. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 1259 – 126
88. Richter SR, Roth S (2018) Matryoshka networks: predicting 3D geometry via nested shape layers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1936–1944
89. Riegler G, Ulusoy AO, Geiger A (2017) Octnet: learning deep 3D representations at high resolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3577–3586
90. Riegler G, Ulusoy AO, Bischof H, Geiger A (2017) Octnetfusion: learning depth fusion from data. In: Proceedings of the International Conference on 3D Vision, pp 57–66
91. Rock J, Gupta T, Thorsen J, Gwak JY, Shin D, Hoiem D (2015) Completing 3D object shape from one depth image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2484–2493
92. Samaras D, Metaxas D, Fua P, Leclerc YG (2000) Variable albedo surface reconstruction from stereo and shape from shading. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1:480–487
93. Saxena A, Sun M, Ng AY (2008) Make3D: learning 3D scene structure from a single still image. IEEE Trans Pattern Anal Mach Intell 31(5):824–840
94. Scarselli F, Gori M, Tsoi AC (2009) The graph neural network model. IEEE Trans Neural Netw 20(1):61–80
95. Schönberger JL, Zheng E, Frahm JM, Pollefeys M (2016) Pixelwise view selection for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision, pp 501–518
96. Sharma S, Kumar V (2020) Voxel-based 3D face reconstruction and its application to face recognition using sequential deep learning. Multimedia Tools and Applications 1–28

97. Sharma A, Grau O, Fritz M (2016) Vconv-dae: deep volumetric shape learning without object labels. In: Proceedings of the European Conference on Computer Vision, pp 236–250
98. Shen W, Jia Y, Wu Y (2019) 3D Shape reconstruction from images in the frequency domain. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4471–4479
99. Shin D, Fowlkes CC, Hoiem D (2018) Pixels, voxels, and views: a study of shape representations for single view 3D object shape prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3061–3069
100. Shin D, Ren Z, Sudderth EB, Fowlkes CC (2019) Multi-layer depth and epipolar feature transformers for 3D scene reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 39–43
101. Sinha A, Unmesh A, Huang Q, Ramani K (2017) Surfnet: generating 3D shape surfaces using deep residual networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6040–6049
102. Smith E, Meger D (2017) Improved adversarial systems for 3D object generated and reconstruction. arXiv preprint arXiv:1707.09557
103. Smith E, Fujimoto S, Meger D (2018) Multi-view silhouette and depth decomposition for high resolution 3D object representation. In: Proceedings of the Advances in Neural Information Processing Systems, pp 6479–6489
104. Smith EJ, Fujimoto S, Romero A, Meger D (2019) GEOMetrics: exploiting geometric structure for graph-encoded objects. arXiv preprint arXiv:1901.11461
105. Soltani AA, Huang H, Wu J, Kulkarni TD, Tenenbaum JB (2017) Synthesizing 3D shapes via modeling multi-view depth maps and silhouettes with deep generative networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1511–1519
106. Song S, Xiao J (2014) Sliding shapes for 3D object detection in depth images. In: Proceedings of the European Conference on Computer Vision, pp 634–651
107. Song HO, Xiang Y, Jegelka S, Savarese S (2016) Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4004–4012
108. Song S, Yu F, Zeng A, Chang AX, Savva M, Funkhouser T (2017) Semantic scene completion from a single depth image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 190–198
109. Sra M, Garrido-Jurado S, Schmandt C, Maes P (2016) Procedurally generated virtual reality from 3D reconstructed physical space. In: Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology, pp 191–200
110. Sun X, Wu J, Zhang X et al (2018) Pix3D: dataset and methods for single-image 3D shape modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2974–2983
111. Tatarchenko M, Dosovitskiy A, Brox T (2017) Octree generating networks: efficient convolutional architectures for high-resolution 3D outputs. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2088–2096
112. Tatarchenko M, Richter SR, Ranftl R, Li Z, Koltun V, Brox T (2019) What do single-view 3D reconstruction networks learn?. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3405–3414
113. Tchapmi LP, Kosaraju V, Rezatofighi H, Reid I, Savarese S (2019) TopNet: structural point cloud decoder. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 383–392
114. Tran L, Liu X (2018) Nonlinear 3D face morphable model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7346–7355
115. Tulsiani S, Zhou T, Efros AA, Malik J (2017) Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2626–2634
116. Tulsiani S, Su H, Guibas LJ, Efros A, Malik J (2017) Learning shape abstractions by assembling volumetric primitives. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2635–2643
117. Varol G, Ceylan D, Russell B et al (2018) Bodynet: volumetric inference of 3D human body shapes. In: Proceedings of the European Conference on Computer Vision, pp 20–36
118. Wang F, Jiang MQ, Qian C et al (2017) Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3156–3164
119. Wang PS, Liu Y, Guo YX, Sun CY, Tong X (2017) O-cnn: octree-based convolutional neural networks for 3D shape analysis. ACM Trans Graph 36(4):72–81

120. Wang N, Zhang Y, Li Z, Fu Y, Liu W, Jiang YG (2018) Pixel2mesh: generating 3D mesh models from single rgb images. In: Proceedings of the European Conference on Computer Vision, pp 55–71
121. Wang PS, Sun CY, Liu Y, Tong X (2018) Adaptive o-cnn: a patch-based deep representation of 3D shapes. ACM Trans Graph 37(6):1–11
122. Wang H, Yang J, Liang W, Tong X (2019) Deep single-view 3D object reconstruction with visual hull embedding. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 33:8941–8948
123. Wang W, Ceylan D, Mech R, Neumann U (2019) 3DN: 3D deformation network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1038–1046
124. Wang WY, Xu Q, Ceylan D, Mech R, Neumann U (2019) Disn: deep implicit Surface network for high-quality single-view 3D reconstruction. arXiv preprint arXiv:1905.10711
125. Wei Y, Liu S, Zhao W, Lu J (2019) Conditional single-view shape generated for multi-view stereo reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 9651–9660
126. Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: Proceedings of the European Conference on Computer Vision, pp 499–515
127. Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X, Xiao J (2015) 3D shapenets: a deep representation for volumetric shapes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1912–1920
128. Wu J, Zhang C, Xue T, Freeman B, Tenenbaum J (2016) Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In: Proceedings of the Advances in Neural Information Processing systems, pp 82–90
129. Wu J, Wang Y, Xue T, Sun X, Freeman B, Tenenbaum J (2017) Marrnet: 3D shape reconstruction via 2.5D sketches. In: Proceedings of the Advances in Neural Information Processing Systems, pp 8–15
130. Wu J, Zhang C, Zhang X, Zhang Z, Freeman WT, Tenenbaum JB (2018) Learning shape priors for single-view 3D completion and reconstruction. In: Proceedings of the European Conference on Computer Vision, pp 673–691
131. Wu Y, He F, Zhang D, Li X (2018) Service-oriented feature-based data exchange for cloud-based design and manufacturing. IEEE Trans Serv Comput 11(2):341–353
132. Wu Y, He F, Yang Y (2020) A grid-based secure product data exchange for cloud-based collaborative design. Int J Coop Inf Syst 29(01n02):2040006
133. Xiang Y, Mottaghi R, Savarese S (2014) Beyond pascal: a benchmark for 3D object detection in the wild. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pp 75–82
134. Xiang Y, Kim W, Chen W et al (2016) Objectnet3D: a large scale database for 3D object recognition. In: Proceedings of the European Conference on Computer Vision, pp 160–176
135. Xiao J, Ehinger KA, Hays J, Torralba A, Oliva A (2016) Sun database: exploring a large collection of scene categories. Int J Comput Vis 119(1):3–22
136. Xie H, Yao H, Sun X, Zhou S, Zhang S (2019) Pix2Vox: context-aware 3D reconstruction from single and multi-view images. arXiv preprint arXiv:1901.11153
137. Yan X, Yang J, Yumer E, Guo Y, Lee H (2016) Perspective transformer nets: learning single-view 3D object reconstruction without 3D supervision. In: Proceedings of the Advances in Neural Information Processing Systems, pp 1696–1704
138. Yang X, Wang Y, Wang Y et al (2018) Active object reconstruction using a guided view planner. arXiv preprint arXiv:1805.03081
139. Yang Y, Feng C, Shen Y, Tian D (2018) Foldingnet: point cloud auto-encoder via deep grid deformation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 206–215
140. Yang B, Lai Z, Lu X et al (2018) Learning 3D scene semantics and structure from a single depth image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 309–312
141. Yang B, Wang S, Markham A, Trigoni N (2020) Robust attentional aggregation of deep feature sets for multi-view 3D reconstruction. Int J Comput Vis 128(1):53–73
142. Yu L, Li X, Fu CW, Cohen-Or D, Heng PA (2018) Pu-net: point cloud upsampling network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2790–2799
143. Yuniarti A, Suciati N (2019) A Review of Deep Learning Techniques for 3D Reconstruction of 2D Images. In: Proceedings of the 2019 12th International Conference on Information & Communication Technology and System, pp 327–331
144. Zeng N, Zhang H, Song B, Liu W, Li Y, Dobaie AM (2018) Facial expression recognition via learning deep sparse autoencoders. Neurocomputing 273:643–649
145. Zeng W, Karaoglu S, Gevers T (2018) Inferring Point Clouds from Single Monocular Images by Depth Intermediation. arXiv preprint arXiv:1812.01402

146. Zhang D, He F, Han S, Li X (2016) Quantitative optimization of interoperability during feature-based data exchange. Integr Comput Aided Eng 23(1):31–50
147. Zhang J, Li K, Liang Y, Li N (2017) Learning 3D faces from 2D images via stacked contractive autoencoder. Neurocomputing 257:67–78
148. Zhang X, Zhang Z, Zhang C, Tenenbaum J, Freeman B, Wu J (2018) Learning to reconstruct shapes from unseen classes. In: Proceedings of the Advances in Neural Information Processing Systems, pp 2257–2268
149. Zhao R, Wang Y, Benitez-Quiroz CF, Liu Y, Martinez M (2016) Fast and precise face alignment and 3D shape reconstruction from a single 2D image. In: Proceedings of the European Conference on Computer Vision, pp 590–603
150. Zheng Z, Yu T, Wei Y, Dai Q, Liu Y (2019) Deephuman: 3D human reconstruction from a single image. In: Proceedings of the IEEE International Conference on Computer Vision, pp 7739–7749
151. Zhu H, Zuo X, Wang S, Cao X, Yang R (2019) Detailed human shape estimation from a single image by hierarchical mesh deformation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4491–4500
152. Zou C, Yumer E, Yang J, Ceylan D, Hoiem D (2017) 3D-prnn: generating shape primitives with recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp 900–909