Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Review:

Deep 3D reconstruction: methods, data, and challenges*

Caixia LIU†1, Dehui KONG1, Shaofan WANG†‡1, Zhiyong WANG2, Jinghua LI1, Baocai YIN1

1*Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Beijing Institute of Artificial Intelligence,*
Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
2*Multimedia Laboratory, School of Computer Science, University of Sydney, Sydney NSW 2006, Australia*
†E-mail: lcxxib@emails.bjut.edu.cn; wangshaofan@bjut.edu.cn

Received Feb. 11, 2020; Revision accepted Mar. 23, 2020; Crosschecked Dec. 29, 2020

Abstract: Three-dimensional (3D) reconstruction of shapes is an important research topic in the fields of computer vision, computer graphics, pattern recognition, and virtual reality. Existing 3D reconstruction methods usually suffer from two bottlenecks: (1) they involve multiple manually designed states which can lead to cumulative errors, but can hardly learn semantic features of 3D shapes automatically; (2) they depend heavily on the content and quality of images, as well as precisely calibrated cameras. As a result, it is difficult to improve the reconstruction accuracy of those methods. 3D reconstruction methods based on deep learning overcome both of these bottlenecks by automatically learning semantic features of 3D shapes from low-quality images using deep networks. However, while these methods have various architectures, in-depth analysis and comparisons of them are unavailable so far. We present a comprehensive survey of 3D reconstruction methods based on deep learning. First, based on different deep learning model architectures, we divide 3D reconstruction methods based on deep learning into four types, recurrent neural network, deep autoencoder, generative adversarial network, and convolutional neural network based methods, and analyze the corresponding methodologies carefully. Second, we investigate four representative databases that are commonly used by the above methods in detail. Third, we give a comprehensive comparison of 3D reconstruction methods based on deep learning, which consists of the results of different methods with respect to the same database, the results of each method with respect to different databases, and the robustness of each method with respect to the number of views. Finally, we discuss future development of 3D reconstruction methods based on deep learning.

Key words: Deep learning models; Three-dimensional reconstruction; Recurrent neural network; Deep autoencoder; Generative adversarial network; Convolutional neural network
https://doi.org/10.1631/FITEE.2000068 **CLC number:** TP391

1 Introduction

Vision-based three-dimensional (3D) reconstruction refers to the computing process and technology that recovers 3D information (such as geometric shape and texture) of objects from images acquired by a camera. Using this technique, 3D shapes with accurate information and photorealism are reconstructed. These 3D shapes can provide functions of scene visualization and virtual roaming, and also meet the high requirements of data archiving, measurement, and analysis, which consequently leads to wide digital applications with respect to ancient buildings, museums, urban planning, medical research, aerospace, shipbuilding, justice, archaeology, industrial measurement, ade-commerce (Sun YY, 2011; Chen et al., 2015; Udayan et al., 2015), and other fields.

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 61772049, 61632006, 61876012, U19B2039, and 61906011) and the Beijing Natural Science Foundation of China (No. 4202003)

ORCID: Caixia LIU, https://orcid.org/0000-0002-1802-8197; Dehui KONG, https://orcid.org/0000-0001-7722-7172; Shaofan WANG, https://orcid.org/0000-0002-3045-624X; Zhiyong WANG, https://orcid.org/0000-0002-8043-0312; Jinghua LI, https://orcid.org/0000-0002-5583-8260; Baocai YIN, https://orcid.org/0000-0002-8125-4648

Most traditional reconstruction methods, such as structure from motion (SFM) and simultaneous localization and mapping (SLAM) (Furukawa and Ponce, 2006; Goesele et al., 2007), require a large number of views and the assumption that features can be matched across views. Although multi-view stereo (Wu ZR et al., 2015; Gwak et al., 2017) and space carving (Wu JJ et al., 2016b) have shown good performance in 3D reconstruction from images, they require precisely calibrated cameras and high-quality images, which restricts them from being widely applied in practice. Moreover, traditional 3D reconstruction methods consist of several manually designed steps, including image preprocessing, point cloud computing, data fusion, and texture mapping (Fig. 1). These steps lead to cumulative errors and inaccurate semantic features of 3D shapes, and therefore seriously affect 3D reconstruction quality. In addition, it is difficult for traditional methods to reconstruct invisible parts when 3D shapes are partially occluded or missing, which makes 3D reconstruction tasks challenging.

In recent years, the rapid development of deep learning models and the distribution of a large number of 3D shapes have provided ideas for traditional 3D reconstruction. 3D reconstruction methods based on deep learning (i.e., deep 3D reconstruction methods) can avoid manually designed algorithms for extracting features and complicated camera calibration. What is more important, deep 3D reconstruction methods can learn 3D shapes, explore both the common and specific characteristics among different 3D shapes by training deep networks, and transfer the knowledge to testing data for predicting the corresponding 3D shapes. In other words, these methods can fully learn visible parts of 3D shapes and reconstruct occluded parts of 3D shapes by training objective functions, which makes up for the inherent defect of traditional vision-based 3D reconstruction and improves the reconstruction accuracy. Therefore, deep 3D reconstruction methods have been widely studied, and promising results are achieved.

Han XF et al. (2019) and Laga (2019) proposed state-of-the-art surveys on 3D reconstruction. In particular, Han XF et al. (2019) first divided 3D reconstruction methods according to the decomposed architectures that contain the encoding stage from different latent spaces and the decoding stage from different shape representations, and then analyzed them by different training mechanisms. Laga (2019) reviewed related topics and addressed works that use deep learning techniques to estimate 3D depth from one or multiple images. We provide a comprehensive survey of deep 3D reconstruction methods and review architectures and characteristics of different reconstruction methods. Compared with previous surveys (Han XF et al., 2019; Laga, 2019), the main contributions and differences in this survey are summarized as follows:

1. Progressively hierarchical classification. We divide deep 3D reconstruction methods into four types based on the model's whole architecture, and divide each type of method based on decoding method, data priors, and shape representations. In addition, we introduce the methodology of each method.

2. Detailed introduction of four representative databases. We describe four databases that are commonly used for 3D reconstruction, introduce them based on many aspects, such as the type of 3D shape, the number of 3D shape categories, the number of 3D shapes, and images, and provide the download links.

3. Comprehensive comparison of reconstruction

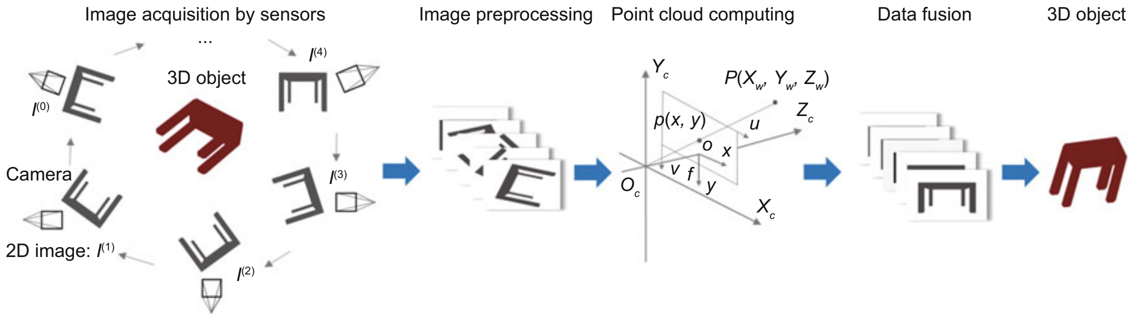


Fig. 1 Traditional three-dimensional (3D) reconstruction based on stereo vision

methods. We provide a comparison of different methods with respect to common databases, a comparison of each method with respect to different databases, and a comparison of the robustness of each method with respect to the number of views.

2 Methodologies of deep 3D reconstruction methods

Applying deep learning models to 3D reconstruction brings new ideas to traditional vision-based 3D reconstruction methods. In this section, we first analyze deep learning models and introduce their applications to 3D reconstruction. Then, we sort existing deep 3D reconstruction methods into four types and progressively introduce the methodology of each reconstruction type.

Fig. 2 shows the development history of deep learning models. These models have achieved great success in many pattern recognition tasks for many reasons: a large amount of training data, various deep networks whose architectures mimic the neural connections in human brains, the rise of graphics processing unit (GPU) computing, and novel techniques (dropout and batch normalization) that help improve training and generalization of the networks. Specifically, prior knowledge, such as common and specific characters among different categories of data, can be explored from training data through multiple single-layer nonlinear networks and then transferred to testing data by subtle models (e.g., classifiers and regressors) for various applications. This makes deep 3D reconstruction methods outperform other methods.

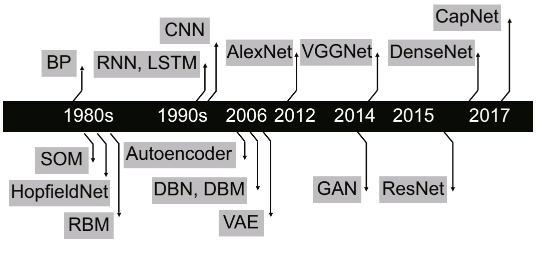


Fig. 2 Development history of deep learning models

In general, deep learning models consist of an encoder and a decoder whose parameters are learned from a training set. The trained encoder maps images to a latent vector in an implicit feature space, and then the trained decoder maps the latent vector to an output space so that the predicted results correspond to the ground-truth targets as close as possible (Zeiler et al., 2010). 3D reconstruction methods have witnessed significant progress through the use of deep learning models, which include the back-propagation neural network, recurrent neural network (RNN), long short-term memory (LSTM), convolutional neural network (CNN), and deep generative models. In particular, deep generative models, which include the restricted Boltzmann machine, deep belief network, deep Boltzmann machine, deep autoencoder (DAE), variational autoencoder (VAE), and generative adversarial networks (GANs), have attracted more and more attention (see Fig. 3 for a classification of deep generative models). The main idea of 3D reconstruction based on these models is as follows: the encoder maps input images to a latent vector space and the decoder maps the latent vector to a 3D shape space, which minimizes the distance between the generated 3D shapes and ground-truth 3D shapes. Deep learning models have a strong ability to learn complex 3D or higher-dimensional data distribution and capture latent features of 3D shapes by various deep networks when applied to 3D reconstruction. As a result, deep 3D reconstruction methods outperform vision-based 3D reconstruction methods.

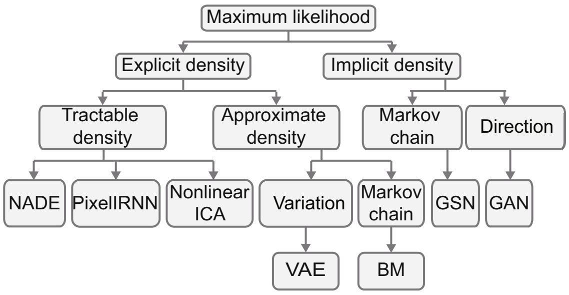


Fig. 3 Taxonomy of generative models (Goodfellow, 2016)

According to network architectures, deep 3D reconstruction methods are divided roughly into RNN-, DAE-, GAN-, and CNN-based methods. In summary, RNN-based methods have a memory function that can save and transfer the features of previous inputs to current inputs; DAE-based methods exhibit a strong ability to expand and infer latent information from images; GAN-based methods have a strong ability to learn both the similarities and differences between 3D shapes through adversarial discriminators; CNN-based methods have a strong ability to learn features of 3D shapes through convolution and pooling operations. We will introduce these methods in detail.

2.1 RNN-based 3D reconstruction methods

RNN is a deep model that specializes in processing sequence data. Each RNN neuron has two inputs, current input information and previously generated memory information, which results in the RNN retaining the sequence-dependent type. As shown in Fig. 4, the RNN unfolds in time series, with the target memory given by

$$\begin{cases} \boldsymbol{s}_t = f(\boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{W}\boldsymbol{s}_{t-1} + \boldsymbol{b}), \\ \boldsymbol{o}_t = g(\boldsymbol{V}\boldsymbol{s}_t), \end{cases}$$

where $\boldsymbol{U}$ is the weight matrix of the input $\boldsymbol{x}_t$, $\boldsymbol{W}$ is the weight matrix of $\boldsymbol{s}_{t-1}$, $\boldsymbol{V}$ is the weight matrix of the output, $\boldsymbol{s}_t$ is a memory that captures information at time t, $\boldsymbol{b}$ is the bias, $f(\cdot)$ and $g(\cdot)$ are activation functions, and $\boldsymbol{o}_t$ is the output. It can be concluded that RNN performs the same task for each element in a sequence, and that the output depends on both the input and the memory. It is difficult to establish feature correspondences between views with traditional 3D reconstruction methods due to local appearance changes, self-occlusion, or a lack of texture for 3D shapes (Fitzgibbon and Zisserman, 1998; Lhuillier and Quan, 2005; Agarwal et al., 2009; Engel et al., 2014). In contrast, RNN-based methods can effectively avoid the problem by selectively concatenating memory cell information that corresponds to views when a single view (or multiple views) is (are) fed into networks.

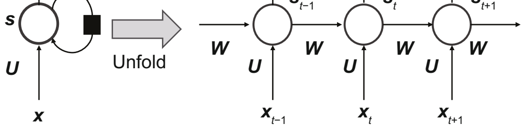


Fig. 4 Structure of a recurrent neural network

According to the decoding process, RNN-based methods can be divided roughly into two types: direct decoding based RNN reconstruction methods and indirect decoding based RNN reconstruction methods. In general, the first type of method generates 3D shapes by direct deconvolution of the latent vector obtained by an encoder, while the second type of method generates 3D shapes by first obtaining the parts of 3D shapes one-by-one with deconvolution of the latent vector and then combining those parts.

2.1.1 Direct decoding based RNN reconstruction methods

Choy et al. (2016) proposed a 3D recurrent neural network (3D-R2N2) which uses the power of LSTM (Hochreiter and Schmidhuber, 1997; Sundermeyer et al., 2012; Sutskever et al., 2014) to retain previous observations and incrementally refined reconstruction results as more observations become available. Specifically, 3D-R2N2 takes one or more images of a 3D shape from arbitrary views as the input, encodes each input image as a latent vector using 2D-CNN, selectively updates cell states of the latent vector or retains the cell states by closing input gates using 3D-LSTM, decodes the cell states of the LSTM units, and generates a 3D shape using 3D-DCNN. 3D-R2N2 takes full advantage of information from different views, and allows adaptive and consistent learning of the appropriate representation of 3D shapes, leading to a faithful mapping between images and 3D shapes from a large amount of synthetic data. 3D-R2N2 requires neither annotations of images nor class labels of 3D shapes during training and testing. Moreover, 3D-R2N2 is superior to SFM/SLAM reconstruction methods.

2.1.2 Indirect decoding based RNN reconstruction methods

Different from the first type of method, indirect decoding based methods usually combine geometric primitives generated by deep models to reconstruct 3D shapes. The motivation comes mainly from the fact that 3D shapes are often composed of multiple primitives which are regarded as a structured and abstract representation of the 3D world. Zou et al. (2017) proposed a 3D primitive recurrent neural network (3D-PRNN) to predicte primitive sequences in shape-centered coordinates. Specifically, 3D-PRNN first sends a latent vector of a depth image to a recurrent generator consisting of LSTMs and mixture density networks, iteratively predicts a set of primitives and features, and combines the primitives to reconstruct 3D shapes according to the predefined contexts. 3D-PRNN achieves comparable accuracy and high efficiency with fewer parameters than voxel-based reconstruction methods (Hu and Zhu, 2015; Huang et al., 2015; Bansal et al., 2016) during training, because it retains the information of a previous primitive to generate the current primitive. It is worth noting that 3D-PRNN uses the Gaussian field and energy minimization (Zhu et al., 1997) to acquire ground-truth primitives of 3D shapes, which is challenging for 3D reconstruction. However, 3D-PRNN has less freedom in the representation of 3D shapes.

We conclude that RNN-based methods are suitable for 3D reconstruction from input samples with a dependent or complementary relationship because they can transfer useful information learned from previous input to current input. In other words, these methods are conducive to the reconstruction of 3D shapes with commonalities.

2.2 DAE-based 3D reconstruction methods

DAE is a type of deep models that consist of two parts, an encoder represented by the function $\boldsymbol{z} = f(\boldsymbol{X})$ and a decoder $\boldsymbol{X}' = g(\boldsymbol{z})$ that represents the reconstruction with some constraints to ensure the closeness between $\boldsymbol{X}'$ and $\boldsymbol{X}$ (Fig. 5). It is difficult for traditional 3D reconstruction methods to learn features of 3D shapes from images automatically. In contrast, DAE-based methods can consider which parts of 3D shapes need to be reconstructed first, learn useful characteristics of 3D shapes, and mine underlying semantics of 3D shapes.

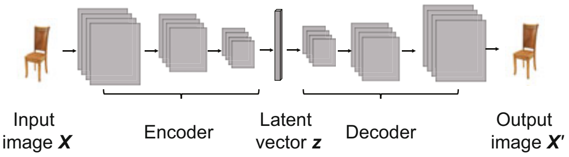


Fig. 5 Structure of deep autoencoder

DAE-based methods are divided into two types, database matching based DAE reconstruction methods and deconvolution decoding based DAE reconstruction methods. In general, the first type of method reconstructs 3D shapes through prior databases based on the idea that similar 3D shapes share similar structural characteristics, while the second type of method predicts 3D shapes by deconvolution layers after obtaining the latent vector from input images.

2.2.1 Database matching based DAE reconstruction methods

In view of the noise, low resolution, occlusion, and missing depth information in general images, Kong et al. (2017) reconstructed 3D shapes from images using a local dense correspondence (LDC) graph, which is a directed graph whose nodes are 3D shapes and edges are the dense correspondences between 3D shapes. Specifically, Kong et al. (2017) first created an LDC graph using the non-rigid iterative closest point algorithm, quickly selected the “closest” 3D shape from the graph by orthogonal matching tracking, and finally combined the similar 3D shapes to fit the final 3D shape sparsely and linearly. Pontes et al. (2017) improved the means of choosing the “closest” 3D shape from the graph using the idea of 3D-2D registration, i.e., to make the selected 3D shape best fit the contours of input images. Pontes et al. (2018) proposed a learning-based 3D reconstruction architecture called Image2Mesh, which further improves the means of selecting the “closest” 3D shape. Specifically, Image2Mesh uses a convolutional autoencoder to extract a latent vector from images, trains a multi-label classifier to classify the vector according to an index of similar 3D shapes, and selects a 3D shape whose latent vector is the closest to the input images. Compared to previous works, Image2Mesh relies on neither contours nor image landmarks, and effectively reconstructs 3D shapes while retaining important geometric features.

Nan et al. (2012), Shao et al. (2012), Li YY et al. (2015), and Shi et al. (2016) proposed 3D reconstruction methods that are similar to the above works. Their common idea is that given a partial shape input, DAE encodes the input to obtain latent features representing 3D shapes. It then sets these features as the inputs to a classifier that attempts to retrieve the same or the most likely 3D shapes and aligns them with the scan, and finally combines these 3D shapes linearly with the combination coefficients given by deep models. However, these methods suffer from two issues: (1) they explicitly assume that databases contain the same or highly similar shapes as reconstructed 3D shapes; (2) they show less robustness and generalization (i.e., they are less effective if novel 3D shapes or categories are included in the testing phase).

2.2.2 Deconvolution decoding based DAE reconstruction methods

Inspired by the idea that a good latent vector can generate a 3D shape and should also be able to be predicted from images, Girdhar et al. (2016) proposed a TL-embedding network which consists of T network for training and L network for testing. Specifically, the T network sends images and ground-truth 3D shapes to two encoders for predicting embedded latent vectors and sends the latent vectors of images into a decoder for reconstructing 3D shapes. The L network sends images to the corresponding trained encoder and reconstructs 3D shapes using the trained decoder. The TL-embedding network adopts the latent vector of ground-truth 3D shapes to constrain those images on the basis of common reconstruction losses. As a result, the TL-embedding network outperforms the baselines on each object of CAD models from ShapeNet (Chang et al., 2015), and generalizes well to real images from IKEA (Lim et al., 2014).

Lin et al. (2018) proposed a 3D generation method that effectively reconstructs 3D shapes with dense point clouds from images. Specifically, the method predictes 3D shapes by fitting multi-view depth images encoded by latent vectors and optimizing objective functions with 2D projection of 3D shapes. This method has great advantages in predicting similarity and density of 3D shapes. Lun et al. (2017) proposed 3D point cloud generation (3D-PCG), which predicts 3D point clouds from 2D sketches. Specifically, 3D-PCG first uses an encoder to convert 2D sketches into a compact representation of 3D shapes. Then a decoder converts the representation into multi-view depth images and normal maps, and 3D-PCG merges the images and maps to obtain 3D point clouds. 3D-PCG offers several advantages over voxel-based reconstruction methods, such as higher-resolution 3D outputs and better 3D topology retention.

As an improvement of DAE, VAE (Gregor et al., 2015) is an encoder-decoder network for learning complex distributions. The encoder of VAE observes samples from the target distribution and produces a set of Gaussian mean and variance vectors, which are sampled to produce a latent vector; the VAE decoder attempts to reproduce the original samples from the latent vector. Nash and Williams (2017) introduced the shape variation autoencoder (ShapeVAE), which is a deep generative model based on VAE for 3D reconstruction. Specifically, ShapeVAE learns low-dimensional shape embedding using a deep probabilistic autoencoder, extracts samples from previously embedded distributions, and obtains novel shapes with point orientations using a decoder. ShapeVAE can capture semantically meaningful shapes and generate reasonable shapes.

Rezende et al. (2016) proposed a conditional generative model that learns 3D shapes from volumetric data or images. Different from previous DAE-based reconstruction methods, context cues are added to the latent vector by an encoder, where the context cues can be category labels, or more views from different cameras. The method shows how to reconstruct 3D shapes without any use of ground-truth 3D labels, and demonstrates the feasibility of learning to infer 3D representation of the world in a purely unsupervised manner.

It is concluded that deconvolution decoding based DAE reconstruction methods are more popular than database matching based DAE reconstruction methods. This is because the former is an end-to-end architecture, in which it is easy for objective functions to feed errors back and update network parameters, while the latter requires similar and high-quality 3D shapes in databases, which are limited in many practical applications.

2.3 GAN-based 3D reconstruction methods

Goodfellow et al. (2014) proposed GAN, which consists mainly of two parts, a generator and a discriminator (Fig. 6). The generator is used mainly to learn distribution of real images, making the generated images more realistic and fooling the discriminator. The discriminator needs to make judgment on (real or fake) generated images. Through training, two networks finally achieve a dynamic equilibrium: the generated image is close to the real image distribution and the discriminator does not recognize the real or fake images. It is difficult for traditional 3D reconstruction methods to reconstruct invisible parts when 3D shapes are partially occluded

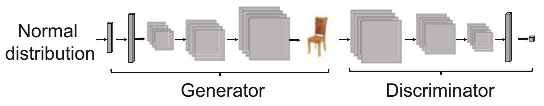


Fig. 6 Structure of a generative adversarial network

or missing. In contrast, GAN-based methods incorporate the adversarial discriminator into the process of generation modeling, which implicitly learns rich similarities and differences of 3D shapes and thus helps infer occluded or missing parts.

Based on whether the input information has priors from 3D shapes, GAN-based methods are divided into two types, single GAN based reconstruction methods and fusion GAN based reconstruction methods. The first type of method reconstructs 3D shapes from random noise, while the second type of method reconstructs 3D shapes from priors, such as images and incomplete shapes.

2.3.1 Single GAN based reconstruction methods

Motivated by the applications of GAN on image processing (Denton et al., 2015; Radford et al., 2015; Li C and Wand, 2016; Wang XL and Gupta, 2016), Wu JJ et al. (2016a) proposed 3D-GAN, which uses volume convolutions based GAN to generate 3D shapes from a latent space. Specifically, the generator of 3D-GAN maps a 200-dimensional latent vector randomly sampled from the probability latent space to a $64 \times 64 \times 64$ cube, which represents a 3D shape in the voxel space, and the discriminator of 3D-GAN outputs a confidence value to determine if 3D shapes are real or fake. 3D-GAN has three advantages: first, using adversarial standards rather than traditional heuristics, it enables the generator to implicitly capture details of 3D shapes and thus produce high-quality 3D shapes; second, the generator establishes the mapping from low-dimensional probability spaces to 3D-shape spaces, so that 3D shapes sampled without reference images can be explored; third, the discriminator provides a powerful 3D shape descriptor that has a wide range of applications in 3D recognition without supervision.

Although 3D-GAN is proficient in generating high-quality 3D shapes for each category, it is difficult to train multi-category data whose distributions are quite different. 3D-GAN uses Kullback-Leibler divergence to minimize the difference between real data and the generated one. This makes the gradient of objective functions disappear when the distributions of real data and the generated one do not overlap, and thus causes unstable training and low reconstruction accuracy.

Wu JJ et al. (2016a) proposed 3D-IWGAN, which stably trains a generator and a discriminator in tandem for predicting realistic 3D shapes from random noise. 3D-IWGAN uses WGAN-GP; that is, it incorporates Wasserstein GAN (Gulrajani et al., 2017), which uses Wasserstein distance to measure the distance between real data and the generated one by weight clipping (i.e., limiting the weights of the network within a compact space) and the gradient penalty (GP), as the discriminator loss. WGAN-GP can avoid a few issues: modeling weakening, gradient explosion, and disappearance. As a result, 3D-IWGAN can perform stable training for multi-category data.

It is concluded that GAN-based 3D reconstruction methods can generate realistic samples of 3D shapes from complex distributions, yet they may suffer from instability during training and fail to generate target 3D shapes from images when using original GAN loss. In addition, it is noted that the dimension of input random noise is difficult to determine; that is, high-dimensional noise needs a large amount of training time, while low-dimensional noise fails to guarantee the generalization of complete and accurate 3D shapes.

2.3.2 Fusion GAN based reconstruction methods

The input of GAN-based methods is the random noise, which generates a type of 3D shape that approximates only ground-truth 3D shapes but fails to reconstruct the corresponding 3D shapes from images. Alternatively, GAN is combined with other deep models to add prior knowledge of 3D shapes from images and then jointly generate target 3D shapes.

As a popular fusion GAN model, DAE-GAN (Wang LJ and Fang, 2017; Yang et al., 2018, 2019) combines the effective coding of DAE with the generation capability of GAN to generate samples based on prior data. The working principle is to use a single network as both a generator and a decoder (i.e., the decoding network of DAE acts as the generation network of GAN), and combine two loss functions during the training process. Yang et al. (2018) proposed 3D-RecGAN, which combines DAE and conditional GAN to reconstruct 3D shapes. 3D-RecGAN requires only the 2.5D voxel grid representation of a depth image as the input and can generate a complete 3D voxel grid with the resolution of $64 \times 64 \times 64$ by filling the occlusion/missing regions. 3D-RecGAN achieves the best reconstruction effect on the chair image in the large synthetic dataset. On this basis, Yang et al. (2019) proposed 3D-RecGAN++ as an improvement of 3D-RecGAN. 3D-RecGAN++ improves mainly the resolution of 3D shapes by adding two upsampling layers in the decoder part, and generates 3D shapes of the $256 \times 256 \times 256$ resolution, while the intersection-over-union (IoU) score drops 0.033. In addition, Wang LJ and Fang (2017) proposed U3DRec based on DAE-GAN, which reconstructs 3D shapes from images in an unsupervised way (that is, images have neither manual annotation nor corresponding 3D shapes). Specifically, U3DRec first embeds images and the synthesized images from similar 3D shapes into a shared latent vector and then converts the latent vector into a 3D shape space. U3DRec designs the reconstruction loss ℓ_1 norm between the reconstructed 3D shapes and similar 3D shapes.

Similar to DAE-GAN, VAE-GAN (Wu JJ et al., 2016a; Smith and Meger, 2017) fuses the advantages of VAE into GAN. The working principle is to use the decoder of VAE as the generator of GAN and combine their loss functions during training. Wu JJ et al. (2016a) proposed 3D-VAE-GAN, which learns a set of parameters (mean and variance) by an encoder, achieves a latent vector sampled from the parameters, generates 3D shapes from the latent vector by a decoder/generator, and sends the generated 3D shapes and ground-truth 3D shapes to a discriminator. 3D-VAE-GAN outperforms previously state-of-the-art models in voxel-level prediction and other baseline methods. Smith and Meger (2017) proposed 3D-VAE-IWGAN, which improves mainly the training of the discriminator in 3D-VAE-GAN by penalizing the gradient of weights to stabilize the training process.

RNN-GAN has been proposed by introducing the memory function of RNN into GAN. Gwak et al. (2017) proposed McRecon, which combines RNN and GAN to reconstruct 3D shapes. Specifically, McRecon encodes multi-view images as a latent vector by an RNN network, decodes the latent vector as a 3D shape, and sends the generated 3D shapes and similar 3D shapes to a discriminator. McRecon has two remarkable advantages: for adversarial loss, it compensates for the missing information in 3D shapes by learning shape distribution of similar 3D shapes; for generative loss, it renders the generated 3D shapes into 2D contours using a perspective ray-tracing pooling layer and uses real 2D contours as mask supervision.

Compared with previous methods, fusion GAN based reconstruction methods perform better because of the advantage of GAN and other deep models, but take longer for convergence due to the instability of GAN during training. In general, the architectures of fusion GAN are worth exploring in the future.

2.4 CNN-based 3D reconstruction methods

CNN is the most widely used deep learning model and is successfully applied to image classification and signal reconstruction, because it can learn the characteristics of images at various levels through convolution and pooling operations. Traditional 3D reconstruction methods involve manually designed algorithms, which have difficulty in completely learning local and global features, surface and abstract features, and geometric and structural features of 3D shapes given input images. In contrast, CNN-based methods can adopt convolution and pooling operations to mine the adaptive, discriminative, and semantic features of 3D shapes, and thus improve reconstruction accuracy.

It should be noted that all methods in Sections 2.1–2.3 can be regarded as CNN-based methods because they use CNN to extract features by convolution and pooling operations. However, other methods, e.g., Wu JJ et al. (2016b)'s and Wang NY et al. (2018)'s approaches, do not belong to any category in Sections 2.1–2.3. To reflect the differences among all CNN-based methods, we divide them into four types, skeleton-based methods, voxel-based methods, point cloud based methods, and mesh-based methods, according to their representations of 3D shapes.

2.4.1 Skeleton-based CNN reconstruction methods

Skeletons composed of the connections between pairwise keypoints are popularly used to represent 3D shapes due to their robustness to the variation of 3D shapes. Wu JJ et al. (2016b) proposed a 3D Interpreter network (3D-INN), which first uses a multi-scale CNN to estimate 2D keypoint heatmaps from images and uses four fully connected layers to infer the skeletons of 3D shapes from the heatmaps. 3D-INN updates the skeletons of 3D shapes by making their projected 2D keypoints closer to ground-truth

2D marker images. In addition, Akhter and Black (2015) used skeletons to represent different 3D human body poses for 3D reconstruction from images.

2.4.2 Voxel-based CNN reconstruction methods

Häne et al. (2017) proposed hierarchical surface prediction (HSP), which first maps images to a latent vector using a convolutional encoder and predictes high-resolution voxel grids of 3D shapes from the latent vector using a convolutional decoder. Because of the subtle design of the decoding process (cropping the feature block part centered around the child node's octant and upsampling the feature map to the new feature block with higher spatial resolution), HSP seems promising in its ability to capture 3D shape surfaces. Choy et al. (2016) and Yang et al. (2018, 2019) predicted higher-resolution voxel grids of 3D shapes than those of HSP.

2.4.3 Point cloud based CNN reconstruction methods

Fan et al. (2017) proposed a point cloud output network (PointOutNet), which first maps images and random vectors to a latent vector and generates point clouds of 3D shapes from the latent vector. Specifically, PointOutNet combines a fully connected branch for showing good performance at describing intricate structures and a deconvolution branch for large smooth 3D shape surfaces. PointOutNet also integrates the hourglass model (Newell et al., 2016) for learning global and local 3D shape information well. PointOutNet infers mainly the point positions in a 3D frame determined by input images and view positions. However, Lun et al. (2017) reconstructed point clouds of 3D shapes by consolidating multi-view depth and normal maps obtained from a deep encoder-decoder network.

2.4.4 Mesh-based CNN reconstruction methods

Wang NY et al. (2018) proposed Pixel2Mesh, which reconstructs meshes of 3D shapes by progressively deforming an ellipsoid and using perceptual features extracted from input images. Specifically, Pixel2Mesh represents 3D mesh with a graph-based convolutional network (GCN), where the vertices and edges of the mesh are directly represented as nodes and connections in a graph, respectively. So, Pixel2Mesh enables features to be exchanged across neighbor nodes by convolutional layers and eventually regresses the 3D location for each vertex by forward propagation. Pixel2Mesh also designs a projection layer that incorporates features of input images into 3D geometry represented by GCN. In addition, Henderson and Ferrari (2019) generated 3D mesh samples by learning the pose of 3D meshes from latent codes. Liu et al. (2019) proposed a 3D mesh reconstruction method, which first outputs the displacement vectors from input images by a mesh generator and obtains the reconstructed 3D mesh from the vectors by a template model.

The skeleton representation can capture geometric changes of articulated shapes and preserve the structural properties that we are interested in, but skeleton-based methods predict only abstract 3D shapes. Currently, most existing works resort to the voxel grid representation, which is a regular structure and easily fits into deep model architectures; however, voxel-based methods lead to imbalanced tradeoffs between sampling resolution of 3D shapes and network efficiency, and obscure natural invariance of 3D shapes under rigid motions. The point cloud representation is a simple and uniform structure that is easy to learn and allows simple manipulation during geometric transformation and deformation of 3D shapes; however, point cloud based methods may lose important surface details of 3D shapes. In contrast, the mesh representation is lightweight, capable of modelling shape details, and easily deformed for animation, so mesh-based methods are more desirable for many real applications.

3 Commonly used 3D geometric databases

In this section, we first summarize several public 3D databases (Table 1) by providing the data type, number of categories, number of samples, and download links. Then we present detailed discussions on four of the most commonly used databases (Table 2), ShapeNet, ModelNet, PASCAL3D+, and IKEA.

3.1 ShapeNet database

ShapeNet is an annotated, large-scale 3D shape database created by Princeton, Stanford, and TTIC researchers (Chang et al., 2015). The raw 3D shapes for ShapeNet come from public online repositories or existing research datasets.

Table 1 Public 3D geometric databases

Database	Data type	Number of categories	3D shape		Image		
			Number	Type	Number	Type	Camera
ShapeNet (Chang et al., 2015)	Synthetic	55/270[§]	51 300/12 000[§]	Mesh	–	Rendered	✓
ModelNet (Wu ZR et al., 2015)	Synthetic	662	127 915	Mesh	–	Rendered	
PASCAL3D+ (Xiang et al., 2014)	Real-world, indoor, outdoor	12	36 000	Mesh	30 899	Real-world	
IKEA (Lim et al., 2014)	Real-world, indoor	7	219	Mesh	759	Real-world	
Pix3D (Sun XY et al., 2018)	Real-world, indoor	9	418	Mesh	16 913	Real-world	
ObjectNet3D (Xiang et al., 2016)	Real-world, indoor	100	44 147	Mesh	90 127	Real-world	✓
NYUdv2 (Silberman et al., 2012)	Real-world, indoor	894	35 064	Mesh	1449	Real-world	✓

[§]: Red for ShapeNetCore and blue for ShapeNetSem. ShapeNet: https://www.shapenet.org; ModelNet: http://modelnet.cs.princeton.edu; PASCAL3D+: http://cvgl.stanford.edu/projects/objectnet3d; IKEA: http://ikea.csail.mit.edu; Pix3D: http://pix3d.csail.mit.edu; ObjectNet3D: http://cvgl.stanford.edu/projects/objectnet3d; NYUdv2: https://cs.nyu.edu/∼silberman/datasets. "–" denotes that there are no images in the database, but images can be obtained by rendering 3D shapes from the database. References to color refer to the online version of this table

Table 2 Deep learning models used on 3D geometric databases

Method	Model	Literature	ShapeNet	ModelNet	PASCAL3D+	IKEA	ObjectNet3D	NYUdv2	Pix3D	Note
RNN	3D-R2N2	Choy et al. (2016)	✓	–	✓	–	–	–	–	Supervised
RNN	3D-PRNN	Zou et al. (2017)	–	✓	–	–	–	✓	–	Supervised
DAE	TL	Girdhar et al. (2016)	✓	–	–	✓	–	–	–	Supervised
DAE	Image2Mesh	Pontes et al. (2018)	✓	–	✓	–	–	–	–	Supervised
DAE	3D-PCG	Lun et al. (2017)	✓	–	–	–	–	–	–	Supervised
DAE	SketchModeling	Nash and Williams (2017)	✓	–	–	–	–	–	–	Supervised
DAE	PointOutNet	Fan et al. (2017)	✓	–	–	–	–	–	–	Supervised
DAE	Pix2Vox	Xie et al. (2019)	✓	–	✓	–	–	–	–	Supervised
DAE	Base-AttSets	Yang et al. (2020)	✓	–	–	–	–	–	–	Supervised
DAE	HRShapeCompletion	Han XG et al. (2017)	✓	–	–	–	–	–	–	Supervised
DAE	Object-completion	Varley et al. (2017)	✓	–	–	–	–	–	–	Supervised
DAE	MarrNet	Wu JJ et al. (2017)	–	–	–	–	–	–	✓	Supervised
DAE	HSP	Häne et al. (2017)	✓	–	–	–	–	–	–	Supervised
GAN	3D-GAN	Wu JJ et al. (2016a)	–	✓	–	–	–	–	–	Unsupervised
GAN	3D-IWGAN	Smith and Meger (2017)	–	✓	–	–	–	–	–	Unsupervised
DAE-GAN	3D-RecGAN++	Yang et al. (2019)	✓	–	–	–	–	–	–	Supervised
DAE-GAN	3D-RecGAN	Yang et al. (2018)	–	✓	–	–	–	–	–	Supervised
DAE-GAN	3D-VAE-GAN	Wu JJ et al. (2016a)	–	–	–	✓	–	–	–	Supervised
DAE-GAN	3D-VAE-IWGAN	Smith and Meger (2017)	–	✓	–	✓	–	–	–	Supervised
DAE-GAN	U3DRec	Wang LJ and Fang (2017)	–	–	✓	–	–	–	–	Unsupervised
RNN-GAN	McRecon	Gwak et al. (2017)	✓	–	–	–	✓	–	–	Weakly supervised
CNN	3D-INN	Wu JJ et al. (2016b)	–	–	✓	✓	–	–	–	Supervised
CNN	Pixel2Mesh	Wang NY et al. (2018)	✓	–	–	–	–	–	–	Supervised

ShapeNet contains models that span a multitude of semantic categories and provides extensive sets of annotations for every model and correspondences between models. The annotations are mainly language-related annotations (category and descriptions of 3D shapes), geometric annotations (rigid alignments, parts and keypoints, and symmetry and shape size), functional annotations (functional parts and affordances), and physical annotations (surface material and weight), which make ShapeNet uniquely valuable. Fig. 7 shows this dense network of interlinked attributes of shapes.

ShapeNet provides a view of data in a hierarchical categorization according to WordNet synsets (Fig. 8). It consists of two subsets, ShapeNetCore and ShapeNetSem. ShapeNetCore is composed of clean 3D shapes, manually validated category labels, and alignment annotations. ShapeNetCore covers 55 common shape categories with approximately 5300 unique 3D shapes. ShapeNetSem is a smaller, more intensive subset with 12 000 models spread over a broader set of 270 categories. In addition to manually validated category labels and consistent alignment, the models from ShapeNetSem are annotated with real-world dimensions, material composition at the category level, volume, and weight.

ShapeNetCore is a widely used subset in the literature; its shapes are shown in Fig. 9. Its v1 release

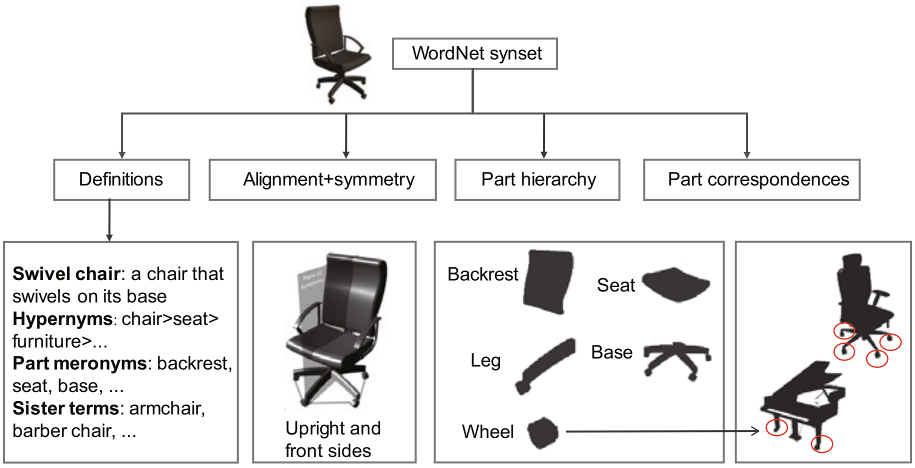
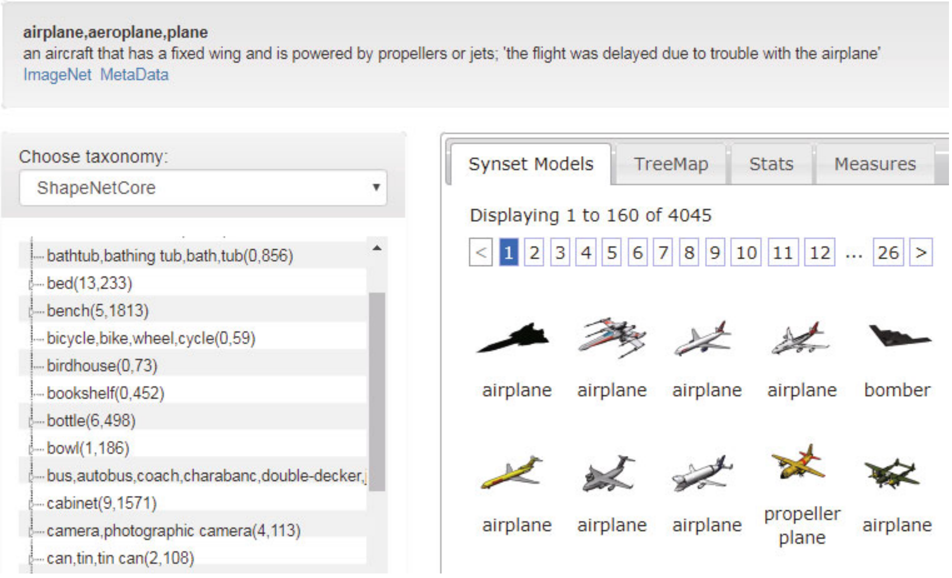

WordNet synset
Definitions
Alignment+symmetry
Part hierarchy
Part correspondences
Swivel chair: a chair that swivels on its base
Hypernyms: chair>seat> furniture>...
Part meronyms: backrest, seat, base, ...
Sister terms: armchair, barber chair, ...
Upright and front sides
Backrest
Seat
Leg
Base
Wheel

Fig. 7 ShapeNet annotations illustrated for chair (Chang et al., 2015)



airplane,aeroplane,plane
an aircraft that has a fixed wing and is powered by propellers or jets; 'the flight was delayed due to trouble with the airplane'
ImageNet MetaData

Choose taxonomy: ShapeNetCore

- bathtub,bathing tub,bath,tub(0,856)
- bed(13,233)
- bench(5,1813)
- bicycle,bike,wheel,cycle(0,59)
- birdhouse(0,73)
- bookshelf(0,452)
- bottle(6,498)
- bowl(1,186)
- bus,autobus,coach,charabanc,double-decker,
- cabinet(9,1571)
- camera,photographic camera(4,113)
- can,tin,tin can(2,108)

Displaying 1 to 160 of 4045

Fig. 8 Screenshot of the online ShapeNet taxonomy view

is of 30.3 GB of size and contains mainly zip and csv files. Each zip file name is an eight-digit, zero-padded string by the offset in WordNet synset. Within each synset zip file, there is a set of annotated 3D shapes, each of which contains mainly a geometric structure (obj file), material component (mtl file), and texture images (jpg files). The csv files are the interpretation and description of 3D shapes in the zip files.

3.2 ModelNet database

ModelNet (Wu ZR et al., 2015) is a 3D shape database provided by the Princeton ModelNet Project for researchers in computer vision, computer graphics, robotics, and cognitive science. To build the database, the organization compiles a list of the most common shape categories in the world using statistical information obtained from the extensive Scene UNderstanding (SUN) database (Xiao et al., 2010). One can query each shape category using the shape vocabulary, collect 3D shapes of shape

Fig. 9 Common 3D shapes for reconstruction in ShapeNetCore (Chang et al., 2015)

classes using the online search engine, and manually determine whether each model belongs to a specified category on Amazon Mechanical Turk using the internally designed quality control tool.

3D shapes in ModelNet are shown in Fig. 10. ModelNet provides two versions, ModelNet10 and ModelNet40. ModelNet10 includes 10 popular categories of 3D shapes for training networks in deep learning projects, and each category has multiple 3D shape styles; ModelNet40 contains 40 categories of 3D shapes, and shares similar properties with ModelNet10. 3D shapes in ModelNet are indoor and clean, with manually aligned orientation, and stored as off files. Unlike ShapeNet, the data in ModelNet consists of training data and testing data.

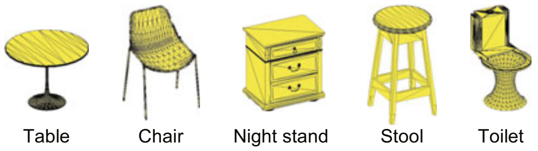


Fig. 10 Common 3D shapes for reconstruction in ModelNet (Wu ZR et al., 2015)

3.3 PASCAL3D+ database

PASCAL3D+ (Xiang et al., 2014) was originally a 3D shape detection and pose recognition dataset created by the Stanford University Computational Vision and Geometry Lab (CVGL). PASCAL3D+ augments 12 rigid categories of images and corresponding 3D shapes from PASCAL VOC (Everingham et al., 2015) with 3D annotations. Many researchers often use PASCAL3D+ to verify the robustness of 3D reconstruction methods by comparing it with other databases.

3D shapes in PASCAL3D+ are shown in Fig. 11. Currently, PASCAL3D+ provides two versions, Release 1.0 (1 GB size) and Release 1.1 (7.5 GB size). The 3D shapes in each category are collected from Google 3D Warehouse when they represent intraclass variations of a particular category. The annotations for each 3D shape include related 3D shapes, 2D markers, and 3D continuous poses. The images in PASCAL3D+ are more variable than those in the existing 3D datasets.

However, when PASCAL3D+ is used for benchmark 3D reconstruction, it is not appropriate to train 3D shapes in PASCAL3D+, because the same set of 3D shapes is used to annotate the test set. Therefore, using 3D shapes in training and testing is biased for 3D reconstruction. In addition, the images in PASCAL3D+ are derived from real-world scenes; that is, the images have complex backgrounds, which may affect the reconstruction accuracy of target 3D shapes.

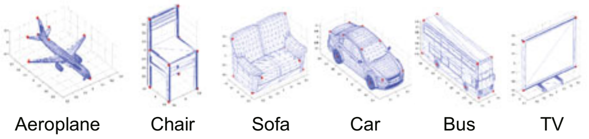


Fig. 11 Common 3D shapes for reconstruction in PASCAL3D+ (Xiang et al., 2014)

3.4 IKEA database

IKEA was originally created for developing and evaluating fine pose estimation based on 3D shapes. IKEA includes images and corresponding 3D shapes from indoor scenes, which are collected from Google 3D Warehouse and Flickr, respectively. Similar to PASCAL3D+, IKEA is often used to verify the robustness of 3D reconstruction methods.

3D shapes in IKEA are shown in Fig. 12. IKEA contains approximately 759 images and 219 3D shapes. All images are annotated with available models (approximately 90 different models). In addition, the images are divided into two different partitions, IKEA shape and IKEA room. Each 3D shape contains mainly geometry structure (obj file), material component (mtl file), and view images (png files). IKEA images have complex backgrounds, which may affect the extraction of pertinent information, thereby reducing reconstruction accuracy of target 3D shapes.

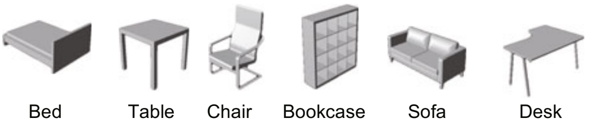


Fig. 12 Common 3D shapes for reconstruction in IKEA (Lim et al., 2014)

4 Performance comparison of deep 3D reconstruction methods

To judge whether deep 3D reconstruction methods are effective in predicting 3D shapes, a common means is to compare their results on the same database, using the same evaluation function. At present, the evaluation functions for 3D reconstruction include mainly the IoU, average precision (AP), chamfer distance (CD), Earth mover's distance (EMD), and cross entropy (CE), where IoU and AP are the two most common functions. According to the experimental results provided by the literature, we list quantitative results of different 3D reconstruction methods on four databases in Tables 3–6. In this section, we will analyze and compare the above deep 3D reconstruction methods.

4.1 Comparison of reconstruction methods based on different deep models

As can be seen from Table 3, most of the methods mentioned in Section 2 have reconstruction results on ShapeNet. Therefore, we highlight representative reconstruction methods on this database, analyze their characteristics in detail, and compare their performances.

RNN-based reconstruction methods obtain relatively low IoU scores which are about 0.567 on average in Table 3. Specifically, 3D-R2N2 and HRShapeCompletion (Han XG et al., 2017) adopt LSTM to learn a set of view images for reconstructing 3D shapes. LSTM is a special structure of RNN and consists of three control units (called cells), input gate, output gate, and forget gate. As features of view images enter networks, each cell will judge whether the features remain or are forgotten, so RNN is often used to fuse the features of multi-view images. However, RNN can memorize only a certain number of view images. Once too many view images are input, the reconstruction accuracy cannot be improved and the computation burden is increased. Moreover, the input order of view images affects the reconstruction effect; that is, inconsistent 3D shapes are estimated from the same image set with different orders. The reconstruction accuracy of McRecon is much lower than those of other methods, because it uses inexpensive 2D silhouettes and approximate views as weak supervision instead of complete 3D shapes.

DAE-based reconstruction methods perform better than RNN-based reconstruction methods except Image2Mesh and HSP (Table 3). This is because DAE-based methods process the features of previous input and current input independently, which may reduce the interference of irrelevant features from previous input for current input. In particular, Image2Mesh produces a lower IoU score than other DAE methods (Fan et al., 2017; Varley et al., 2017; Yang et al., 2018, 2020; Xie et al., 2019) on ShapeNet. This is because Image2Mesh fits limited 3D shapes from databases to reconstruct 3D shapes from images. However, PointOutNet, Pix2Vox (Xie et al., 2019), Object-completion (Varley et al., 2017), Base-AttSets, and 3D-RecAE (Yang et al., 2018) employ a decoder to reconstruct 3D shapes from the latent features instead of fitting 3D shapes, and they are end-to-end architectures whose parameters are updated by loss functions. HSP produces worse results than other DAE methods, but it can predict 3D voxel grids with a high resolution of 256 × 256 × 256.

Fusion GAN based reconstruction methods with supervision achieve relatively high IoU scores than other methods (Table 3). It is also found that fusion GAN based methods perform the best on other

Table 3 Quantitative results of different methods on ShapeNet

Method	Model	Literature	IoU					3D	Resolution	Note
			Bench	Car	Chair	Table	Mean	representation		
RNN	3D-R2N2	Choy et al. (2016)	0.421	0.798	0.466	0.513	0.550	Voxel grid	32^3	Supervised
	HRShapeCompletion	Han XG et al. (2017)	0.611	–	0.524	0.615	0.583	Voxel grid	32^3	Supervised
DAE	HSP	Häne et al. (2017)	–	0.698	0.361	–	0.530	Voxel grid	256^3	Supervised
	Image2Mesh	Pontes et al. (2018)	–	0.664	0.403	–	0.534	Mesh	–	Supervised
	PointOutNet	Fan et al. (2017)	0.550	0.831	0.544	0.606	0.633	Point cloud	32^3	Supervised
	Pix2Vox	Xie et al. (2019)	0.613	0.806	0.599	0.613	0.658	Voxel grid	32^3	Supervised
	Base-AttSets	Yang et al. (2020)	0.569	0.848	0.571	0.597	0.646	Voxel grid	32^3	Supervised
	Object-completion	Varley et al. (2017)	0.653	–	0.619	0.678	0.650	Voxel grid	32^3	Supervised
	3D-RecAE	Yang et al. (2018)	0.800	–	0.790	0.808	0.799	Voxel grid	32^3	Supervised
	3D-RecAE	Yang et al. (2018)	0.733	–	0.736	0.759	0.742	Voxel grid	64^3	Supervised
DAE-GAN	3D-RecGAN++	Yang et al. (2019)	0.806	–	0.793	0.821	0.807	Voxel grid	32^3	Supervised
DAE-GAN	3D-RecGAN++	Yang et al. (2019)	0.745	–	0.741	0.772	0.753	Voxel grid	64^3	Supervised
RNN-GAN	McRecon	Gwak et al. (2017)	0.295	0.562	0.350	0.353	0.390	Voxel grid	32^3	Weakly supervised

databases from Tables 4–6. In particular, according to Table 3, 3D-RecGAN++ achieves the highest mean IoU score of 0.807 among all methods, and can reconstruct 3D shapes with a resolution of 64 × 64 × 64. However, other methods predict only 3D shapes with lower resolution. The success of 3D-RecGAN++ benefits from combining the generative capabilities of DAE and GAN. McRecon performs worse than 3D-RecGAN++, because 3D-RecGAN++ takes a real 3D shape and a fake 3D shape as the inputs of a discriminator, while McRecon takes a real 2D silhouette and a fake 2D occupancy map rendered by a reconstructed 3D shape as the inputs of a discriminator. Although both discriminators can distinguish whether the estimated 3D shapes are plausible or not, McRecon observes 3D shapes using 2D views and it is a weakly supervised method.

CNN-based reconstruction methods learn features of 3D shapes using convolution operations instead of traditional manually designed algorithms. From Tables 3–6, it is found that most deep 3D reconstruction methods use voxel grids to represent 3D shapes, because voxel grids are regular structures and easily fit into deep model architectures. From Table 7, it can be seen that Pixel2Mesh performs the best by achieving the lowest CD and EMD, which benefits from the mesh representation of 3D shapes provided by GCN. We analyze that point clouds (Fan et al., 2017) and voxel grids (Choy et al., 2016) tend to lose important surface details of 3D shapes; however, meshes are capable of modeling shape details and are easy to deform for animation.

Based on the above analysis, it is concluded that different methods have their own advantages and disadvantages for 3D reconstruction, but fusion GAN methods are more conducive because of the inherent architectures, and the mesh representation of 3D shapes is more suitable for 3D reconstruction. In addition, it is worth noting that deep 3D reconstruction methods often use popular CNNs, e.g., VGGNet (Simonyan and Zisserman, 2015) and

Table 4 Quantitative results of different methods on ModelNet

Method	Model	Literature	IoU						3D	Resolution	Note
			Chair	Table	Night stand	Stool	Toilet	Mean	representation		
RNN	3D-PRNN	Zou et al. (2017)	0.245	0.188	0.204	–	–	0.212	Point cloud	30^3	Supervised
DAE	Object-completion	Varley et al. (2017)	0.564	–	–	0.273	0.503	0.447	Voxel grid	40^3	Supervised
DAE	3D-RecAE	Yang et al. (2018)	0.633	–	–	0.488	0.520	0.547	Voxel grid	64^3	Supervised
DAE-GAN	3D-RecGAN	Yang et al. (2018)	0.661	–	–	0.501	0.569	0.577	Voxel grid	64^3	Supervised

Table 5 Quantitative results of different methods on PASCAL3D+

Method	Model	Literature	IoU							3D	Resolution	Note
			Aeroplane	Bus	Car	Chair	Sofa	TV	Mean	representation		
RNN	3D-R2N2	Choy et al. (2016)	0.544	0.816	0.699	0.280	0.332	0.574	0.541	Voxel grid	32^3	Supervised
DAE	Image2Mesh	Pontes et al. (2018)	0.366	0.280	0.371	0.236	0.207	–	0.292	Mesh	–	Supervised
DAE	Pix2Vox	Xie et al. (2019)	0.690	0.760	0.657	0.593	0.634	0.694	0.671	Voxel grid	32^3	Supervised
DAE-GAN	U3DRec	Wang LJ and Fang (2017)	–	–	0.634	0.241	0.450	0.247	0.393	Voxel grid	32^3	Unsupervised

Table 6 Quantitative results of different methods on IKEA

Method	Model	Literature	AP							3D	Resolution	Note
			Bed	Bookcase	Chair	Desk	Sofa	Table	Mean	representation		
DAE	TL	Girdhar et al. (2016)	56.3	30.2	32.9	25.8	71.7	23.3	40.03	Voxel grid	20^3	Supervised
DAE-GAN	3D-VAE-GAN	Wu JJ et al. (2016a)	63.2	46.3	47.2	40.7	78.8	42.3	53.08	Voxel grid	20^3	Supervised
DAE-GAN	3D-VAE-IWGAN	Smith and Meger (2017)	77.7	51.8	56.2	49.8	82.0	52.6	61.68	Voxel grid	32^3	Supervised
RNN-GAN	McRecon	Gwak et al. (2017)	–	–	32.0	28.6	55.7	29.0	36.33	Voxel grid	32^3	Weakly supervised

Table 7 Quantitative results of different 3D representations for CNN-based reconstruction methods on ShapeNet

Method	Literature	CD					EMD					3D
		Bench	Car	Chair	Table	Mean	Bench	Car	Chair	Table	Mean	representation
3D-R2N2	Choy et al. (2016)	1.891	0.845	1.432	1.116	1.321	1.136	1.670	1.466	1.641	1.478	Voxel grid
Pixel2Mesh	Wang NY et al. (2018)	0.624	0.268	0.610	0.498	0.500	0.965	1.297	1.399	1.480	1.285	Mesh
PointOutNet	Fan et al. (2017)	0.629	0.333	0.645	0.517	0.531	1.113	1.747	1.946	2.121	1.732	Point cloud

AlexNet (Krizhevsky et al., 2012), to learn 3D shape features. For example, Niu et al. (2018) initialized a network using the trained parameters of VGGNet and then re-trained it using suitable loss functions to accurately reconstruct 3D shapes from images.

4.2 Performance comparison of reconstruction methods on different databases

As can be seen from Tables 3–6, different reconstruction methods use different databases because deep learning models have their own advantages when dealing with a specific database. We will analyze four commonly used databases by comparing the reconstruction methods in detail in this subsection.

For synthetic databases, ShapeNet is more widely used than ModelNet. According to the results of the chair image (which is the only common shape in these two databases) from Tables 3 and 4, ShapeNet obtains higher IoU than ModelNet under the same model 3D-RecAE. We argue that this is because ShapeNet has more annotations and representations for each 3D shape than ModelNet, and the additional information helps deep learning models learn more prior knowledge and mine more feature cues for complete recovery of 3D shapes during that reconstruction.

For real-world databases, the reconstruction on PASCAL3D+ and IKEA does not work well. This may because images provided by the databases are real-world images; that is, the images have cluttered backgrounds which cause interference and noise and affect accurate reconstruction of target 3D shapes, thus resulting in lower precision. Moreover, PASCAL3D+ and IKEA contain a relatively small number of categories, which may cause poorer generalization capabilities than the previous two databases.

The above analysis demonstrates that the differences in database quality, quantity, and additional information can directly affect the performance of 3D reconstruction methods. In addition, there are two findings: (1) many deep 3D reconstruction methods train networks by synthetic databases and verify their robustness and generalization on real-world databases; (2) many deep 3D reconstruction methods use AP to evaluate the performance on IKEA.

4.3 Impact of the number of views on 3D reconstruction methods

As can be seen from Table 8, whether supervised or unsupervised, the number of input views affects 3D reconstruction results. When the number of views increases, the IoU score increases, which means that the reconstructed 3D shapes are closer to ground-truth 3D shapes. In essence, the more views there are, the more comprehensive are the features of representing 3D shapes. In this subsection, we analyze 3D reconstruction methods by input views in detail.

According to different views of input images, 3D reconstruction tasks can be divided into single-view reconstruction and multi-view reconstruction. The main idea of single-view reconstruction is to first encode an input image as a latent vector, and then decode the latent vector to recover a 3D shape. Currently, single-view reconstruction has two types of architectures reconstructing 3D shapes from single-view images (Girdhar et al., 2016; Smith and Meger, 2017; Yang et al., 2018) and reconstructing 3D shapes from multi-view images predicted from single-view images (Lin et al., 2018). The difference between these two types of methods lies in the decoding process. The first type of method decodes

Table 8 Quantitative results with respect to the number of views on ShapeNet

Method	Model	Literature	Number of views	IoU					3D representation	Resolution	Note
				Chair	Car	Table	Bench	Mean			
RNN	3D-R2N2	Choy et al. (2016)	1	0.466	0.798	0.513	0.421	0.550	Voxel grid	32^3	Supervised
			2	0.515	0.821	0.550	0.484	0.593			
			3	0.533	0.829	0.564	0.502	0.607			
			4	0.541	0.833	0.573	0.516	0.616			
			5	0.550	0.836	0.580	0.527	0.623			
RNN-GAN	McRecon	Gwak et al. (2017)	1	0.350	0.562	0.353	0.295	0.390	Voxel grid	32^3	Weakly supervised
			5	0.437	0.614	0.420	0.401	0.468			

the latent vector to reconstruct a 3D shape by 3D deconvolution layers which take up a large computer memory. The second type of method makes multiple copies of the latent vector, and then decodes these vectors to reconstruct a 3D shape by 2D deconvolution layers. The second type of method uses inexpensive 2D depth images from several views as optimization targets instead of 3D shapes, and uses 2D deconvolution accordingly instead of 3D deconvolution, which largely reduces computational cost. It is difficult to predict complete 3D shapes using single-view reconstruction because of incomplete sampling information. However, this problem is offset by the continuous improvement in network performance and the completeness of prior knowledge acquired by network training. In addition, the convenience of single-view reconstruction in applications further enhances its availability.

Multi-view reconstruction provides a more precise result than single-view reconstruction; however, it suffers from two issues: (1) it requires images of 3D shapes from as many views as possible; (2) it requires a subtle fusion scheme of multi-view images, which needs both multi-view feature descriptors and cross-view feature matching. With the development of hardware technology and the reduction of hardware cost, the former problem has been increasingly ignored. Currently, the advantage of reconstruction accuracy brought by abundant sampling information has become increasingly prominent for multi-view reconstruction.

Based on the above analysis, it is concluded that single-view reconstruction and multi-view reconstruction have their own strengths and weaknesses, so researchers can choose network designs according to their needs.

5 Future development of 3D reconstruction

Although deep 3D reconstruction methods have made breakthrough progress, further studies are necessary. In this section, we summarize the possible future development of deep 3D reconstruction methods from three aspects: supervised learning versus unsupervised learning models, matrix neural networks versus non-matrix neural networks, and low-resolution versus high-resolution outputs.

5.1 Supervised learning versus unsupervised learning models

We live in a 3D world, but our observations are usually in a form of 2D projections captured by eyes or cameras. A key goal of computer vision is to restore 3D shapes from these 2D observations. A lot of attempts have been made to infer 3D shapes from images. Currently, supervised 3D reconstruction methods have witnessed significant progress. However, these methods require large-scale annotation of 2D/3D data. Actually, complete 3D shapes are difficult to obtain and are often unknown in reality, resulting in a severe imbalance between the available number of images and the corresponding 3D shapes. Therefore, supervised reconstruction methods cannot achieve satisfactory results because of the lack of data during training.

A few researchers have studied weakly supervised 3D reconstruction. Rezende et al. (2016) minimized the loss between projected images of predicted 3D shapes and input images for reconstructing 3D shapes from images. Yan et al. (2016) introduced a 2D silhouette loss function based on perspective transformations for reconstructing 3D shapes from images. Gwak et al. (2017) used a 2D contour loss defined by a perspective ray tracing pooling layer for reconstructing 3D shapes from images, which has better generalization ability than the one Yan et al. (2016) designed. These methods prove the feasibility of inferring 3D representation from images in a weakly unsupervised way without ground-truth 3D shapes. However, Wang LJ and Fang (2017) proposed U3DRec, which is a purely unsupervised network. Specifically, U3DRec learns features of synthesized images from 3D shapes similar to ground-truth 3D shapes, and uses the similar features to predict target 3D shapes. U3DRec optimizes the distance between predicted 3D shapes and similar 3D shapes, and produces results close to those of the supervised method (Choy et al., 2016) on the vehicle and chair images from PASCAL3D+.

At present, most deep 3D reconstruction methods are still supervised. Through the above analysis, it can be seen that the weakly supervised and unsupervised 3D reconstructions are feasible. Although they currently have lower performance than supervised methods, this is a direction worth studying for practical applications.

5.2 Matrix neural networks versus non-matrix neural networks

Deep learning models have achieved significant breakthroughs in the hierarchical representation of images. CNN has become an alternative method for learning the representation of 2D planar images. However, CNN exhibits the invariance of translation and rotation for image representation, which leads to the position information loss of image features and hence affects the 3D reconstruction effect. A few researchers proposed non-matrix networks, such as group equivariant convolutional neural network (G-CNN) and GCN, which are expected to be applied to 3D reconstruction for solving the problem of feature information loss.

Cohen and Welling (2016) introduced G-CNN, which uses G-convolution layers to extract image features. It is proved that image features learned by G-CNN on CIFAR10 and rotating MNIST are more accurate and more reliable than those captured by CNN. This is because G-CNN has a higher degree of weight sharing, which enhances the expressive power of a network without increasing the number of parameters. In addition, G-CNN is equivariant to translations and transformations of feature maps, which enables the network to preserve the position information of image features. Therefore, G-CNN is conducive to reconstructing 3D shapes from images.

Cohen et al. (2018) proposed a spherical CNN, which is used to deal with the classification problem of spherical signals. It is demonstrated that spherical CNN produces better computational efficiency, numerical accuracy, and effectiveness than CNN. This is because the spherical cross-correlation in spherical CNN satisfies a generalized Fourier theorem, which computes the network efficiently using a generalized fast Fourier transform algorithm. Therefore, it is reasonably expected that a spherical CNN can capture the details of spherical 3D shapes well, and be applied to 3D reconstruction from images.

Inspired by the work of GCN (Bruna et al., 2013), Kipf and Welling (2017) proposed a deep model based on GCN to classify graph data, and achieved good accuracy. This demonstrates that GCN can learn features of graph data. Wang NY et al. (2018) represented a 3D shape as a graph interconnecting part and used GCN to capture different levels of attributes of a 3D shape for reconstructing 3D meshes from images. This method ensures visually appealing and physically accurate 3D geometry information, and has better performance than CNN-based reconstruction methods (Choy et al., 2016; Fan et al., 2017).

Therefore, many reconstruction methods reconstruct 3D shapes with important surface detail lost, which contributes to a common challenge of these methods. The reason is that these methods are insufficiently powerful to represent the features of 3D shapes. A few researchers have proposed novel networks for learning complex 3D data. Therefore, to improve the reconstruction accuracy, the application of these novel networks to 3D reconstruction is worth exploring.

5.3 Low-resolution versus high-resolution outputs

A common limitation of the works of Calakli and Taubin (2011), Choy et al. (2016), Girdhar et al. (2016), Rezende et al. (2016), Yan et al. (2016), and Gadelha et al. (2017) is that they can predict only coarse voxel grids (e.g., $32 \times 32 \times 32$ resolution) of 3D shapes. It is observed that predicting high-resolution 3D shapes becomes computationally infeasible. However, a few researchers have devoted to high-resolution reconstruction of 3D shapes from images.

Inspired by the sparse convolutional networks (Graham, 2014, 2015), Tatarchenko et al. (2017) proposed octree generating networks (OGNs), which store sparse non-trivial feature sets rather than dense feature maps for efficient 3D shape analysis, thus reducing computational cost. In particular, OGN gradually refines an estimated rough low-resolution structure to a desired high-resolution one and predicts only a sparse set of spatial locations at each level. The OGN represents its volumetric output as an octree. The representation is significantly more efficient than a dense voxel grid and allows generating volumes as large as $512 \times 512 \times 512$ voxels on a modern GPU in a single forward pass. In addition, Dai et al. (2017), Häne et al. (2017), and Cao et al. (2018) predicted high-resolution meshes in a progressive, coarse-to-fine manner.

Currently, most deep 3D reconstruction methods can predict only low-resolution voxel grids of 3D shapes. However, as we all know, the need for high-resolution shapes is becoming more and more urgent in real applications. Therefore, generating high-resolution 3D shapes with surface details is worth studying.

6 Conclusions

Through the comparison and analysis of the deep 3D reconstruction methods on databases, the following conclusions can be drawn:

1. More knowledge should be mined from existing data by deep networks to improve 3D reconstruction accuracy, which stems from three phenomena: (1) deep 3D reconstruction methods trained on datasets with large capacity and many labels perform better than the ones trained on datasets with small capacity and few labels; (2) supervised deep 3D reconstruction methods obtain higher accuracy and generate higher-quality 3D shapes than unsupervised or weakly supervised reconstruction methods; (3) multi-view deep 3D reconstruction methods learn the mapping between images and 3D shapes better than single-view reconstruction methods.

2. Generative models are more conducive to reconstruction than non-generative models because of inherent architectures. Specifically, fusion GAN based reconstruction methods perform the best for 3D reconstruction, and representing 3D shapes with meshes is desirable for many real applications in the future.

3. The research trend of 3D reconstruction is diversification with regard to network architectures and applications. The reason is twofold: for one thing, the research with emphasis on feature extraction, encoding, and decoding can improve performance of networks and reconstruction accuracy; for the other, different application backgrounds give rise to the needs of diverse 3D reconstruction, resulting in a variety of network architectures.

In this study, we attempt to reveal the essence of deep 3D reconstruction methods, classify them, and compare their performances from the methodological sense, 3D shape representation, and databases. This work will help researchers better understand and improve the methods.

Contributors

Caixia LIU designed the research outline and collected the data. Caixia LIU and Shaofan WANG drafted the manuscript. Dehui KONG and Zhiyong WANG helped organize the manuscript. Caixia LIU, Shaofan WANG, Jinghua LI, and Baocai YIN revised and finalized the paper.

Compliance with ethics guidelines

Caixia LIU, Dehui KONG, Shaofan WANG, Zhiyong WANG, Jinghua LI, and Baocai YIN declare that they have no conflict of interest.

References

Agarwal S, Snavely N, Simon I, et al., 2009. Building Rome in a day. IEEE 12th Int Conf on Computer Vision, p.72-79. https://doi.org/10.1109/ICCV.2009.5459148

Akhter I, Black MJ, 2015. Pose-conditioned joint angle limits for 3D human pose reconstruction. IEEE Conf on Computer Vision and Pattern Recognition, p.1446-1455. https://doi.org/10.1109/CVPR.2015.7298751

Bansal A, Russell B, Gupta A, 2016. Marr revisited: 2D-3D alignment via surface normal prediction. IEEE Conf on Computer Vision and Pattern Recognition, p.5965-5974. https://doi.org/10.1109/CVPR.2016.642

Bruna J, Zaremba W, Szlam A, et al., 2013. Spectral networks and locally connected networks on graphs. Int Conf on Learning Representations, p.1-14.

Calakli F, Taubin G, 2011. SSD: smooth signed distance surface reconstruction. *Comput Graph Forum*, 30(7):1993-2002. https://doi.org/10.1111/j.1467-8659.2011.02058.x

Cao YP, Liu ZN, Kuang ZF, et al., 2018. Learning to reconstruct high-quality 3D shapes with cascaded fully convolutional networks. Proc 15th European Conf on Computer Vision, p.616-633. https://doi.org/10.1007/978-3-030-01240-3_38

Chang AX, Funkhouser T, Guibas L, et al., 2015. ShapeNet: an information-rich 3D model repository. https://arxiv.org/abs/1512.03012

Chen K, Lai YK, Hu SM, 2015. 3D indoor scene modeling from RGB-D data: a survey. *Comput Vis Media*, 1(4): 267-278. https://doi.org/10.1007/s41095-015-0029-x

Choy CB, Xu DF, Gwak J, et al., 2016. 3D-R2N2: a unified approach for single and multi-view 3D object reconstruction. Proc 14th European Conf on Computer Vision, p.628-644. https://doi.org/10.1007/978-3-319-46484-8_38

Cohen TS, Welling M, 2016. Group equivariant convolutional networks. Proc 33rd Int Conf on Machine Learning, p.2990-2999.

Cohen TS, Geiger M, Köhler J, et al., 2018. Spherical CNNs. Int Conf on Learning Representations, p.1-15.

Dai A, Qi CR, Nießner M, 2017. Shape completion using 3D-encoder-predictor CNNs and shape synthesis. IEEE Conf on Computer Vision and Pattern Recognition, p.6545-6554. https://doi.org/10.1109/CVPR.2017.693

Denton E, Chintala S, Szlam A, et al., 2015. Deep generative image models using a Laplacian pyramid of adversarial networks. Proc 28th Int Conf on Neural Information Processing Systems, p.1486-1494.

Engel J, Schöps T, Cremers D, 2014. LSD-SLAM: large-scale direct monocular SLAM. Proc 13th European Conf on Computer Vision, p.834-849. https://doi.org/10.1007/978-3-319-10605-2_54

Everingham M, Eslami SMA, van Gool L, et al., 2015. The PASCAL visual object classes challenge: a retrospective. *Int J Comput Vis*, 111(1):98-136. https://doi.org/10.1007/s11263-014-0733-5

Fan HQ, Su H, Guibas L, 2017. A point set generation network for 3D object reconstruction from a single image. IEEE Conf on Computer Vision and Pattern Recognition, p.2463-2471. https://doi.org/10.1109/CVPR.2017.264

Fitzgibbon A, Zisserman A, 1998. Automatic 3D model acquisition and generation of new images from video sequences. Proc 9th European Signal Processing Conf, p.129-140.

Furukawa Y, Ponce J, 2006. Carved visual hulls for image-based modeling. Proc 9th European Conf on Computer Vision, p.564-577. https://doi.org/10.1007/11744023_44

Gadelha M, Maji S, Wang R, 2017. 3D shape induction from 2D views of multiple objects. Int Conf on 3D Vision, p.402-411. https://doi.org/10.1109/3DV.2017.00053

Girdhar R, Fouhey DF, Rodriguez M, et al., 2016. Learning a predictable and generative vector representation for objects. Proc 14th European Conf on Computer Vision, p.484-499. https://doi.org/10.1007/978-3-319-46466-4_29

Goesele M, Snavely N, Curless B, et al., 2007. Multi-view stereo for community photo collections. IEEE 11th Int Conf on Computer Vision, p.1-8. https://doi.org/10.1109/ICCV.2007.4408933

Goodfellow I, 2016. NIPS tutorial: generative adversarial networks. https://arxiv.org/abs/1701.00160

Goodfellow IJ, Pouget-Abadie J, Mirza M, et al., 2014. Generative adversarial nets. Proc 27th Int Conf on Neural Information Processing Systems, p.2672-2680.

Graham B, 2014. Spatially-sparse convolutional neural networks. https://arxiv.org/abs/1409.6070v1

Graham B, 2015. Sparse 3D convolutional neural networks. Proc British Machine Vision Conf, p.150.1-150.9. https://doi.org/10.5244/C.29.150

Gregor K, Danihelka I, Graves A, et al., 2015. DRAW: a recurrent neural network for image generation. Proc 32nd Int Conf on Machine Learning, p.1462-1471.

Gulrajani I, Ahmed F, Arjovsky M, et al., 2017. Improved training of Wasserstein GANs. Advances in Neural Information Processing Systems, p.5767-5777.

Gwak J, Choy CB, Chandraker M, et al., 2017. Weakly supervised 3D reconstruction with adversarial constraint. Int Conf on 3D Vision, p.263-272. https://doi.org/10.1109/3DV.2017.00038

Han XF, Laga H, Bennamoun M, 2019. Image-based 3D object reconstruction: state-of-the-art and trends in the deep learning era. *IEEE Trans Patt Anal Mach Intell*, 43(5):1578-1604. https://doi.org/10.1109/TPAMI.2019.2954885

Han XG, Li Z, Huang HB, et al., 2017. High-resolution shape completion using deep neural networks for global structure and local geometry inference. IEEE Int Conf on Computer Vision, p.85-93. https://doi.org/10.1109/ICCV.2017.19

Häne C, Tulsiani S, Malik J, 2017. Hierarchical surface prediction for 3D object reconstruction. Int Conf on 3D Vision, p.412-420. https://doi.org/10.1109/3DV.2017.00054

Henderson P, Ferrari V, 2019. Learning single-image 3D reconstruction by generative modelling of shape, pose and shading. *Int J Comput Vis*, 128:835-854. https://doi.org/10.1007/s11263-019-01219-8

Hochreiter S, Schmidhuber J, 1997. Long short-term memory. *Neur Comput*, 9(8):1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu WZ, Zhu SC, 2015. Learning 3D object templates by quantizing geometry and appearance spaces. *IEEE Trans Patt Anal Mach Intell*, 37(6):1190-1205. https://doi.org/10.1109/TPAMI.2014.2362141

Huang QX, Wang H, Koltun V, 2015. Single-view reconstruction via joint analysis of image and shape collections. *ACM Trans Graph*, 34(4):87. https://doi.org/10.1145/2766890

Kipf TN, Welling M, 2017. Semi-supervised classification with graph convolutional networks. Int Conf on Learning Representations, p.1-13.

Kong C, Lin CH, Lucey S, 2017. Using locally corresponding CAD models for dense 3D reconstructions from a single image. IEEE Conf on Computer Vision and Pattern Recognition, p.5603-5611. https://doi.org/10.1109/CVPR.2017.594

Krizhevsky A, Sutskever I, Hinton GE, 2012. ImageNet classification with deep convolutional neural networks. Proc 25th Int Conf on Neural Information Processing Systems, p.1-9.

Laga H, 2019. A survey on deep learning architectures for image-based depth reconstruction. https://arxiv.org/abs/1906.06113

Lhuillier M, Quan L, 2005. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE Trans Patt Anal Mach Intell*, 27(3):418-433. https://doi.org/10.1109/TPAMI.2005.44

Li C, Wand M, 2016. Precomputed real-time texture synthesis with Markovian generative adversarial networks. Proc 14th European Conf on Computer Vision, p.702-716. https://doi.org/10.1007/978-3-319-46487-9_43

Li YY, Dai A, Guibas L, et al., 2015. Database-assisted object retrieval for real-time 3D reconstruction. *Comput Graph Forum*, 34(2):435-446. https://doi.org/10.1111/cgf.12573

Lim JJ, Pirsiavash H, Torralba A, 2014. Parsing IKEA objects: fine pose estimation. IEEE Int Conf on Computer Vision, p.2992-2999. https://doi.org/10.1109/ICCV.2013.372

Lin CH, Kong C, Lucey S, 2018. Learning efficient point cloud generation for dense 3D object reconstruction. AAAI Conf on Artificial Intelligence, p.7114-7121.

Liu SC, Chen WK, Li TY, et al., 2019. Soft rasterizer: differentiable rendering for unsupervised single-view mesh reconstruction. https://arxiv.org/abs/1901.05567v1

Lun ZL, Gadelha M, Kalogerakis E, et al., 2017. 3D shape reconstruction from sketches via multi-view convolutional networks. Int Conf on 3D Vision, p.67-77. https://doi.org/10.1109/3DV.2017.00018

Nan LL, Xie K, Sharf A, 2012. A search-classify approach for cluttered indoor scene understanding. *ACM Trans Graph*, 31(6):137.1-137.10. https://doi.org/10.1145/2366145.2366156

Nash C, Williams CKI, 2017. The shape variational autoencoder: a deep generative model of part-segmented 3D objects. *Comput Graph Forum*, 36(5):1-12. https://doi.org/10.1111/cgf.13240

Newell A, Yang KY, Deng J, 2016. Stacked hourglass networks for human pose estimation. Proc 14th European Conf on Computer Vision, p.483-499. https://doi.org/10.1007/978-3-319-46484-8_29

Niu CJ, Li J, Xu K, 2018. Im2Struct: recovering 3D shape structure from a single RGB image. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1-9. https://doi.org/10.1109/CVPR.2018.00475

Pontes JK, Kong C, Eriksson A, et al., 2017. Compact model representation for 3D reconstruction. Int Conf on 3D Vision, p.88-96. https://doi.org/10.1109/3DV.2017.00020

Pontes JK, Kong C, Sridharan S, et al., 2018. Image2Mesh: a learning framework for single image 3D reconstruction. Proc 14th Asian Conf on Computer Vision, p.365-381. https://doi.org/10.1007/978-3-030-20887-5_23

Radford A, Metz L, Chintala S, 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. Int Conf on Learning Representations, p.1-16.

Rezende DJ, Eslami SMA, Mohamed S, et al., 2016. Unsupervised learning of 3D structure from images. Proc 30th Conf on Neural Information Processing Systems, p.4997-5005.

Shao TJ, Xu WW, Zhou K, et al., 2012. An interactive approach to semantic modeling of indoor scenes with an RGBD camera. *ACM Trans Graph*, 31(6):136. https://doi.org/10.1145/2366145.2366155

Shi YF, Long PX, Xu K, et al., 2016. Data-driven contextual modeling for 3D scene understanding. *Comput Graph*, 55:55-67. https://doi.org/10.1016/j.cag.2015.11.003

Silberman N, Hoiem D, Kohli P, et al., 2012. Indoor segmentation and support inference from RGBD images. Proc 12th European Conf on Computer Vision, p.746-760. https://doi.org/10.1007/978-3-642-33715-4_54

Simonyan K, Zisserman A, 2015. Very deep convolutional networks for large-scale image recognitions. Int Conf on Learning Representations, p.1-14.

Smith EJ, Meger D, 2017. Improved adversarial systems for 3D object generation and reconstruction. Proc 1st Annual Conf on Robot Learning, p.87-96.

Sun XY, Wu JJ, Zhang XM, et al., 2018. Pix3D: dataset and methods for single-image 3D shape modeling. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2974-2983. https://doi.org/10.1109/CVPR.2018.00314

Sun YY, 2011. A survey of 3D reconstruction based on single image. *J North China Univ Technol*, 23(1):9-13 (in Chinese). https://doi.org/10.3969/j.issn.1001-5477.2011.01.002

Sundermeyer M, Schlüter R, Ney H, 2012. LSTM neural networks for language modeling. https://core.ac.uk/display/22066040

Sutskever I, Vinyals O, Le Q, 2014. Sequence to sequence learning with neural networks. Proc 27th Int Conf on Neural Information Processing Systems, p.3104-3112.

Tatarchenko M, Dosovitskiy A, Brox T, 2017. Octree generating networks: efficient convolutional architectures for high-resolution 3D outputs. IEEE Int Conf on Computer Vision, p.2107-2115. https://doi.org/10.1109/ICCV.2017.230

Udayan JD, Kim H, Kim JI, 2015. An image-based approach to the reconstruction of ancient architectures by extracting and arranging 3D spatial components. *Front Inform Technol Electron Eng*, 16(1):12-27. https://doi.org/10.1631/FITEE.1400141

Varley J, DeChant C, Richardson A, et al., 2017. Shape completion enabled robotic grasping. IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.2442-2447. https://doi.org/10.1109/IROS.2017.8206060

Wang LJ, Fang Y, 2017. Unsupervised 3D reconstruction from a single image via adversarial learning. https://arxiv.org/abs/1711.09312

Wang NY, Zhang YD, Li ZW, et al., 2018. Pixel2Mesh: generating 3D mesh models from single RGB images. Proc 15th European Conf on Computer Vision, p.55-71. https://doi.org/10.1007/978-3-030-01252-6_4

Wang XL, Gupta A, 2016. Generative image modeling using style and structure adversarial networks. Proc 14th European Conf on Computer Vision, p.318-335. https://doi.org/10.1007/978-3-319-46493-0_20

Wu JJ, Zhang CK, Xue TF, et al., 2016a. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. Advances in Neural Information Processing Systems, p.82-90.

Wu JJ, Xue TF, Lim JJ, et al., 2016b. Single image 3D interpreter network. Proc 14th European Conf on Computer Vision, p.365-382. https://doi.org/10.1007/978-3-319-46466-4_22

Wu JJ, Wang YF, Xue TF, et al., 2017. MarrNet: 3D shape reconstruction via 2.5D sketches. Advances in Neural Information Processing Systems, p.540-550.

Wu ZR, Song SR, Khosla A, et al., 2015. 3D ShapeNets: a deep representation for volumetric shapes. IEEE Conf on Computer Vision and Pattern Recognition, p.1912-1920. https://doi.org/10.1109/CVPR.2015.7298801

Xiang Y, Mottaghi R, Savarese S, 2014. Beyond PASCAL: a benchmark for 3D object detection in the wild. IEEE Winter Conf on Applications of Computer Vision, p.75-82. https://doi.org/10.1109/WACV.2014.6836101

Xiang Y, Kim W, Chen W, et al., 2016. ObjectNet3D: a large scale database for 3D object recognition. Proc 14th European Conf on Computer Vision, p.160-176. https://doi.org/10.1007/978-3-319-46484-8_10

Xiao JX, Hays J, Ehinger KA, et al., 2010. SUN database: large-scale scene recognition from abbey to zoo. IEEE Computer Society Conf on Computer Vision and Pattern Recognition, p.3485-3492. https://doi.org/10.1109/CVPR.2010.5539970

Xie HZ, Yao HX, Sun XS, et al., 2019. Pix2Vox: context-aware 3D reconstruction from single and multi-view images. IEEE/CVF Int Conf on Computer Vision, p.1-9. https://doi.org/10.1109/ICCV.2019.00278

Yan XC, Yang JM, Yumer E, et al., 2016. Perspective transformer nets: learning single-view 3D object reconstruction without 3D supervision. Advances in Neural Information Processing Systems, p.1696-1704.

Yang B, Wen HK, Wang S, et al., 2018. 3D object reconstruction from a single depth view with adversarial learning. IEEE Int Conf on Computer Vision Workshop, p.679-688. https://doi.org/10.1109/ICCVW.2017.86

Yang B, Rosa S, Markham A, et al., 2019. 3D object dense reconstruction from a single depth view. *IEEE Trans Patt Anal Mach Intell*, 41(12):2820-2834. https://doi.org/10.1109/TPAMI.2018.2868195

Yang B, Wang S, Markham A, et al., 2020. Robust attentional aggregation of deep feature sets for multi-view 3D reconstruction. *Int J Comput Vis*, 128:53-73. https://doi.org/10.1007/s11263-019-01217-w

Zeiler MD, Krishnan D, Taylor GW, et al., 2010. Deconvolutional networks. IEEE Computer Society Conf on Computer Vision and Pattern Recognition, p.2528-2535. https://doi.org/10.1109/CVPR.2010.5539957

Zhu CY, Byrd RH, Lu PH, et al., 1997. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans Math Softw*, 23(4):550-560. https://doi.org/10.1145/279232.279236

Zou CH, Yumer E, Yang JM, et al., 2017. 3D-PRNN: generating shape primitives with recurrent neural networks. IEEE Int Conf on Computer Vision, p.900-909. https://doi.org/10.1109/ICCV.2017.103