DFNN: Data Fusion Neural Network for Real-scene Reconstruction Model Inpainting of Nature Tree

Hongpeng Wang[1,2,3],Xiao Han[1,2], Zhongzhi Cao[1,2], Yaojing Li[1,2], and Xinwei Chen[3,4*]

Abstract— With the development of 3D reconstruction and multi-modal perceptual data fusion, it is a fundamental problem to inpainting a natural scene reconstruction model based on Generative Adversarial Network. In this paper, we propose Data Fusion Neural Network(DFNN) to solve the problem of inpainting the natural tree reconstruction model, which is reconstructing inpainting. The DFNN includes a generator network and a discriminator network. This model has higher real-time and availability than the existing methods, especially for the real-time task reconstruction of the natural environment. The experiments demonstrate the feasibility efficiency and effectiveness of our proposed method. Finally, the loss values of the trained generator and discriminator are 0.364 and 0.115 respectively, and the trained generator network inpainting the original model can satisfy the requirement for natural scene reconstruction.

I. INTRODUCTION

Nowadays, 3D point cloud data is widely used in reconstruction tasks. In reverse engineering, the point cloud is a point dataset of product appearance surface obtained by measuring instrument. The points in the point cloud contain rich information, including spatial coordinates, color, classification values, intensity values, time, and so on. Real scenes can be restored from high-precision point cloud data, therefore point cloud is a good choice for 3D reconstruction.

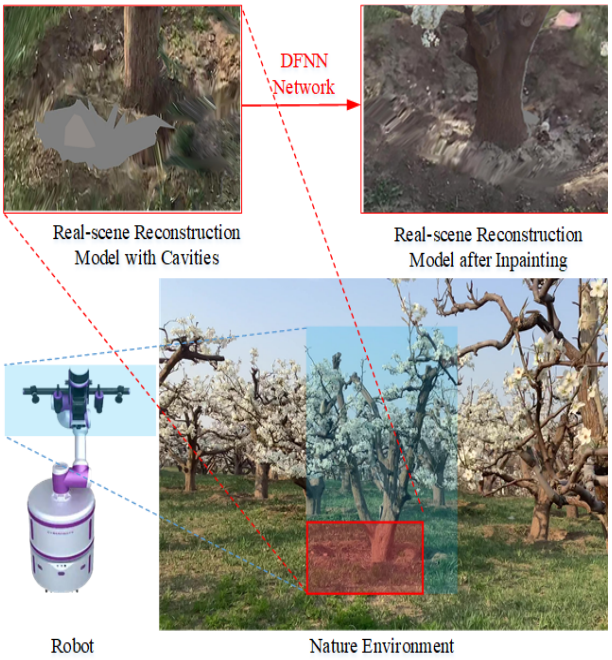

Fig. 1. Real scene reconstruction of natural environment.Top is the robot system that performs scanning and reconstruction tasks.Bottom left corner shows the cavity generated by multi-view reconstruction.Bottom right corner shows the inpainting result by DFNN.

To achieve better performance on 3D reconstruction tasks, it is essential to collect more accurate and complete point cloud data. Limited by the posture of the sensor and the occlusion of obstacles, it is unrealistic to collect complete point cloud data, such as the top of the tree and the puddles on the ground. to get more complete data, one possible approach is to fill in gaps in the incomplete point cloud data. In order to get a better filling effect, the technique of neural networks can be applied.

Deep learning has made a breakthrough in the 3D model generation or shape completion.3DGAN learns the mapping from 2D image to 3D model by using random potential vector as input. The network uses an encoder to map voxelized 3D objects to probabilistic potential space, then gener-

This work is supported in part by the National Natural Science Foundation of China (Grants Nos.61973173),Technology Reasearch and Development Program of Tianjin (Grant No. 18ZXZNGX00340).And Open Fund Project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (No. MJUKF-IPIC201904)

[1]College of Artificial Intelligence, Nankai University, Tianjin, 300353

[2]Institute of Intelligence Technology and Robotic Systems Shenzhen Research Institute of Nankai University, Shenzhen, China

[3]Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University), 350108

[4]IStrong Technology Co., LTD

*Corresponding author:

Xinwei Chen(chen_xinwei@126.com)

ates a generator of confrontation network (GAN) to predict complete volume objects according to potential features. Although the 3DGAN is comparable or even better than the traditional methods, due to the disorder of the point cloud, the abovementioned methods often convert the 3D model into a more regular format which results in a huge amount of data to be processed and may cause quantization loss or projection loss.

The Deep Learning method does not need to measure the size of a 3D model and design a CAD model for it, but due to the disorder and sparsity of point cloud, the deep learning method for 3D model shape completion usually converts the training data into voxel representation, so it is easy to generate redundant voxel data, resulting in a huge amount of data to be processed, which leads to the increase of calculation. Our contributions are summarized as follows:

- We propose a new architecture for natural environment reconstruction inpainting, named Data Fusion Neural Network, which includes discriminator and generator with T-Net.
- We design the new reconstruction inpainting loss function which combines the loss of GAN and chamfers distance combination of the point cloud shape. And adjust the total loss function through the alpha parameter. Through experiments, the optimal alpha value is obtained.
- We compare the proposed method with state-of-the-arts approaches on several datasets in terms of evaluation metrics to demonstrate its effectiveness.

II. RALATED WORK

Real scene reconstruction inpainting with a multi-modal sensing system in the natural environment is the task of using neural network technology with the idea of generative Adversaria to fix the cavity generated by image single-modal reconstruction. Such a task mainly relates to two research fields that are intimately connected:3D Reconstruction with GAN and Point Cloud with Neural Network.

A. 3D Reconstruction with GAN

In the development of semantic image processing using deep learning algorithm, the generative adversarial networks(GAN)[1] proposed by Goodflow's team in 2014 is a landmark research achievement.

Conditional generative adversarial networks (CGAN)[2] takes labeled data as input, which enables CGAN generators to learn more efficiently. Deep convolutional generative adversarial networks (DCGAN)[3] is introduced to solve the instability of GAN training. Based on the results of CGAN and DCGAN, the author [4] and others introduced tag auxiliary information, which greatly improved the convergence speed of the network and the clarity of the generated image. The auxiliary classifier generative adverse network[5] (ACGAN) improves CGAN and gives the discriminator the function of image classification.

LSGAN [8](least-squares GAN) replaces GAN's loss function with the least square loss function, which improves the stability of the model. Wasserstein GAN [9] optimizes the loss function of GAN and enhances its stability. Gulrajani [10] and others improved Wasserstein GAN to accelerate the convergence speed of the network and eliminate the problems of gradient disappearance and gradient explosion. Pix2pix [11] Based on CGAN puts forward a new idea for image conversion. Pix2pix uses two datasets. Pix2pix generates an approximation image through the mapping between training datasets A and B. Since then, Wang T C etc.[12] and Regmi K ,etc [13] have continued the Generic Advantageous research of pix2pix. Perceptual adversarial network (PAN) also proposes a general idea, which can map the input image to the desired target image. DR-GAN [6] and TP-GAN [7] network models proposed by Tran et al are used to generate attitude transformation images.

B. Point Cloud with Neural Network

Qi et al proposed Pointnet [17] and PointNet++ [18] Networks, which directly use 3D point cloud data as the input of the network to complete the segmentation and recognition of the point cloud model, thus proving that deep learning can solve the problems of disorder and sparsity. Achlioptas et al put forward the point cloud generation confrontation network and autoencoder network, which take the point cloud data as the input, learn the potential feature representation of a 3D point cloud, and remap these potential features to generate a 3D point cloud. Although autoencoder network can learn the point features of 3D point cloud data, it still has some shortcomings, that is, it only accepts clean and complete point sets for training.

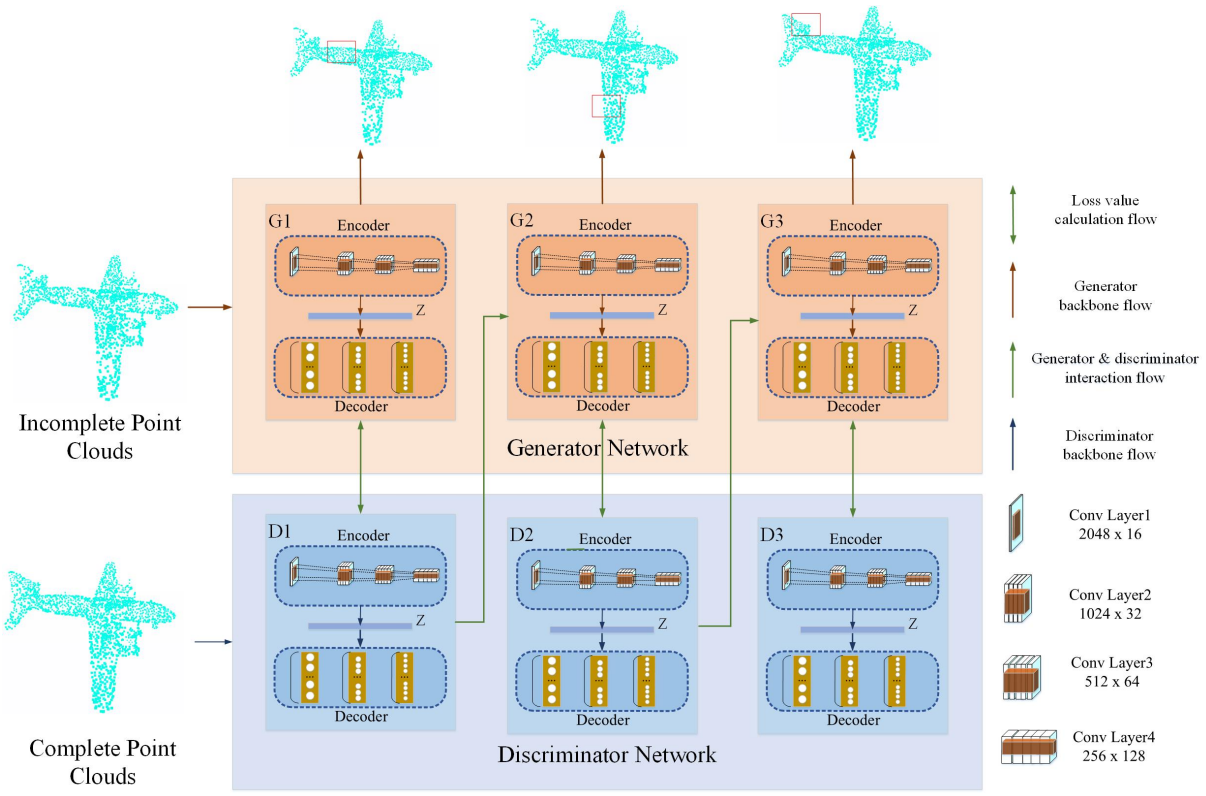

Fig. 2. The architecture of DFNN.The network structure mainly includes generator (G) and discriminator (D). Through the confrontation training between G and D networks,Final G that can meet the requirements is obtained

III. METHODOLOGY

In 3D reconstruction, due to the limitation of camera pose and the occlusion of obstacles, the collected point cloud data are missing in some positions, and the geometric shape and semantic information are incomplete. Therefore, the completion of the 3D point cloud has become an important research direction in the field of computer vision. In the following three sections, the research work of this paper is described in detail. They are network architecture design, loss function and weight parameter update.

A. Network Architecture Design

In this section, a network model based on DFNN is proposed to inpainting the shape hole in natural environment reconstruction. The generation network model is elaborated. Inspired by the research work of PointNet, the encoder-decoder network model is used in the point cloud shape completion network, and the complete point cloud shape is generated according to the input incomplete point cloud shape. After the generator replenishes the point cloud, compared with the point cloud model with labels, the loss function from the discriminator and chamfered distance combination of the point cloud shape is used as the overall loss function. Different alpha values are set to test the influence of different weights on the final loss. The three indicators of accuracy, completion and F1-score are tested, which verify the effectiveness of the algorithm.

The 3D point cloud shape completion countermeasure neural network based on encoder-decoder includes the point cloud shape completion part, namely the generator part and the recognizer part. The network structure of the recognizer is shown in Fig 2. The point cloud shape completion network part is based on the encoder-decoder network, including the encoder and decoder parts. The encoder consists of one layer of ectopic convolution layer with a core size of 1, step size of 1 and four layers of ectopic convolution layer with a core size of 3, step size of 1, which is used to learn to compress the features of point cloud data into potential space vectors. In order to accelerate convergence and avoid overfitting, a RELU activation layer and a batch normalization layer are added after each

convolution layer, and maximum pooling is used to extract high-dimensional features in the last four layers.

The input of the encoder is a $2048*3$ point cloud matrix. However, due to the difference between the shape of the point cloud data and the complete point cloud data, there is a slight deviation to the coordinate system of the trained point cloud model. When the missing point cloud dataset is directly input, the convergence speed of the network model is slow. Inspired by the point net alignment method, this paper proposes a T-Net structure to predict the affine transformation matrix and act on the secondary coordinates of the input point cloud. Compared with the common encoder-decoder network model, the network model with T-Net structure has better performance. The decoder consists of three fully connected layers, each layer consists of 512 neurons, 256 neurons and 6144 neurons. The goal of the designed neural network is to map the potential space vector generated by the encoder into a $2048*3$ complete point cloud, and enter the discriminator together with the sampled complete point cloud to judge whether the generated complete point cloud model is true or false.

B. Loss Function

The proposed DFNN model has two loss functions, similar to 3D-ED-GAN, which are LGAN and LCH. The reconstruction loss comes from the chamfered distance, which is a differentiable function and is more efficient than the permutation invariant measure of an unordered point set. A distance transform for the image is often used in shaped-based object detection. The distance transformation is to solve the distance from each point to the nearest feature point. Therefore, the reconstruction loss is defined by chamfered distance. Chamfered distance is used to calculate the square distance between each point in the S1 subset and the nearest neighbor in the S2 subset. The specific formula is demonstration as (1) and (2).

$$\begin{aligned} L_{CH} &= D_{CH}(S_1, S_2) \\ &= \sum_{F_1} \min \|x-y\|_2^2 + \sum_{F_2} \min \|x-y\|_2^2 \end{aligned} \tag{1}$$

$$F_1 : x \in S_1, y \in S_2 \tag{2}$$

$$F_2 : x \in S_2, y \in S_1 \tag{3}$$

$$\text{Loss} = L_{GAN} + \alpha L_{CH} \tag{4}$$

C. Weight parameter update

The weight update based on the DFNN model inpainting proposed in this section adopts the ADAM optimizer, and the weight update solution is shown in algorithm 1. ADAM combines the advantages of Adagrad and the momentum gradient descent algorithm. It can adapt to sparse gradients and alleviate the problem of gradient oscillation. ADAM combines the momentum term with the adaptive learning rate and avoids the impact of oscillation and improper parameter initialization in network training.

Algorithm 1 Parameter Update Algorithm

Input: step size $stride = 1$, decay rate $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate $\alpha = 0.001$, size of point cloud pre_size, weight parameter θ_d, θ_g

Output: the optimal value of weight parameter θ_d, θ_g

1: sample m data $\{x^1, x^2, ..., x^m\}$ from $ShapeNet$;
2: sample m noise $\{z^1, z^2, ..., z^m\}$ from P_{data};
3: for $index = 0 \rightarrow pre\_size - 1$ do
4: $min\_dis[index] \leftarrow min\_dis[index - stride]$;
5: $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$;
6: $V \leftarrow \frac{1}{m}\sum_{i=1}^{m} \log D(G(x^i)) + \frac{1}{m}\sum_{i=1}^{m}(1\text{-}\log D(x^i)) + L_{ch}$;
7: $\theta_d \leftarrow \theta_d + \eta \nabla V(\theta_d)$;
8: sample m noise data $\{z^1, z^2, ..., z^m\}$ from $P_{example}$;
9: sample m conditions that satisfy L_{ch} from database $ShapeNet$;
10: $V \leftarrow \frac{1}{m}\sum_{i=1}^{m} \log D(G(x^i))$;
11: $\theta_g \leftarrow \theta_g + \eta \nabla V(\theta_g)$;
12: Or $index \leftarrow index + 1$
13: end for

IV. EXPERIMENTS AND RESULTS

A. Multi Sensor System and Dataset

The experimental equipment used in this paper is an innovative multi-modal perception fusion system. It includes the vision system, which is composed of lidar (radium intelligent c−16b) and stereo binocular vision with a large baseline (FLIR industrial network camera), and the close-range subsystem, which is composed of a depth camera (RealSense D435i) and stereo binocular vision with

small baseline. Among them, all sensor devices can be flexibly moved, installed and disassembled. It is convenient for different scanning tasks.

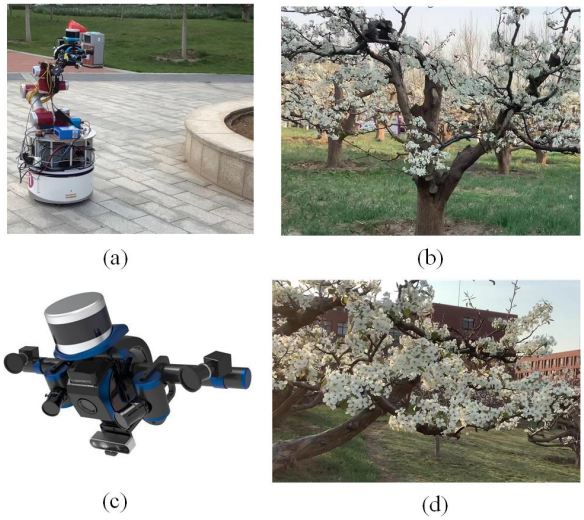

Fig. 3. multi-modal perception fusion system and datasets.(a) is the Mobile robot scanner.(c) is the multi sensor sensing fusion system.(c) and (d) represente the image dataset part of pear orchard.

B. 3D Reconstruction of Cavity Filling Experiment

In the reconstruction effect optimization part, the experimental processor used in this paper is 16GB RAM, Intel (R) Core (TM) $i7-8700k$ CPU, GTX1070. The test environment is Ubuntu LTS 16.04, and the programming language and framework are Python and PyTorch.

ShapeNet dataset and 3D reconstruction point cloud data of a pear orchard were used in the experimental dataset. 1000 groups of point cloud data were randomly extracted from the point clouds of pear orchard in the experiment. First, the ShapeNet dataset is used to pre-train the network, and then the point cloud data of the pear orchard after 3D reconstruction is used for secondary training. Two RTX2080ti graphics cards are used to speed up the training, and PyTorch is used as the deep learning framework.

TABLE I
Analysis of Loss Index

Value of α	Accuracy	Completeness	FI Score
$\alpha = 0$	63.2	55.2	59.2
$\alpha = 0.7$	72.5	64.5	68.5
$\alpha = 0.8$	77.3	67.2	72.3
$\alpha = 0.9$	80.2	70.3	75.3

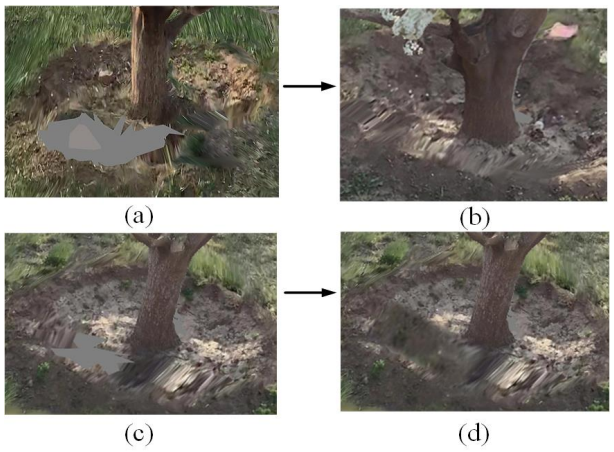

Fig. 4. The natural scene inapinting displayed based on DFNN (a) and (c) show origin reconstruction. (c) and (d) show the result of inpainting.

In this section, the evaluation metrics of point cloud shape completion are similar to accuracy and recall. Accuracy, completeness and F1-score are defined as follows.

- Accuracy. The minimum distance between each point and the nearest point in the tag point cloud is calculated. If the minimum distance is less than the threshold value of 0.03, it will be counted as the correct matching point, otherwise, it is not.
- Completeness is the number of point clouds correctly matched with P_{comp} within the threshold of 0.03 in P_{gt}.
- F1-score is the average value of accuracy and completeness, which is used to comprehensively evaluate the effect of point cloud completion. It is close to 1, indicating that the better the experimental effect is.

TABLE II
Real scene inpainting Result on ShapeNet Dataset

Methods	RMSE(m)	Time(h)
Raw GAN	6.73	3.7
Latent GAN	4.94	3.4
Ours	5.31	2.1

The super parameters in the experiment are set as follows. To speed up the training, the random gradient descent algorithm of batch processing is used in the training, and the batch size is 64. The whole model uses Adam optimizer $\beta_1 = 0.5$, $\beta_2 = 0.999$. In the training process, the recognizer learns faster than the generator.

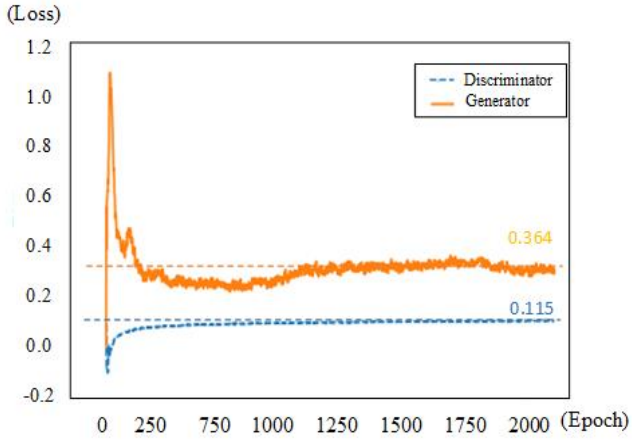

Fig. 5. DFNN training data visualization. The horizontal axis represents the training epoch times; The vertical axis represents the loss value (blue points in the figure represents the loss value of the discriminator; Orange points represents the loss value of the generator).

Therefore, the encoder-decoder network is trained separately with only reconstruction loss, and the learning rate is set to $5e$-5, and 20 iterations are trained separately. Then the recognizer and the encoder-decoder network are trained together 100 times, and the learning rate of the encoder-decoder is set to $1e$-4, and the learning rate of the recognizer is set to $1e$-5. When training together, if the accuracy of each generation's recognizer no longer changes significantly (fluctuates between one percent), only the recognizer is updated. To observe the different α in the experiment, four different parameters of 0.0, 0.7,0.8, 0.9 are used.

V. CONCLUSION

In this paper, a novel DFNN based repair method is proposed to fill the holes in image reconstruction, which includes discriminator and generator with T-Net. Then, we design the new reconstruction inpainting loss function which combines the loss of GAN and chamfers distance combination of the point cloud shape. On the one hand, the cavities comes from the failure of the reconstruction model generation caused by the failure of the image data itself to match the feature points; on the other hand, the hole is caused by the failure of the ground mobile robot system to obtain the data at the top of the tree. The proposed algorithm is tested on the public dataset, and the effectiveness and the advanced nature of the algorithm are verified. The algorithm is applied to the reconstruction experiment of a pear garden, and the practicability of the algorithm is verified.

References

[1] Goodfellow I J, et al. "Generative Adversarial Networks," Comput. Res. Repository, 2014.

[2] Ghamisi P and Yokoya N. "Img2dsm: Height Simulation from Single Imagery Using Conditional Generative Adversarial Net," IEEE Geosci. Remote Sens. Lett., vol. 15, no. 5, pp. 794-798, 2018.

[3] Radford A, Metz L and Chintala S. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," in Int. Conf. Learn. Represent., Poster, 2016.

[4] He Z, Liu H, Wang Y, et al. "Generative Adversarial Networks-based Semi-supervised Learning for Hyperspectral Image Classification," Remote Sens., vol. 9, no. 10, pp. 1042, 2017.

[5] Odena A, Olah C, Shlens J. "Conditional Image Synthesis with Auxiliary Classifier GANs," in Int. Conf. Mach. Learn., 2015, pp. 2642-2651.

[6] Tran L, Yin X and Liu X. "Disentangled Representation Learning GAN for Pose-invariant Face Recognition," in Proc. IEEE Conf. Comput. Vision Pattern Recognit., 2017, pp. 1415-1424.

[7] Huang R, Li T, et al. "Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis," in Proc. IEEE Int. Conf. Comput. Vision, 2017, pp. 2439-2448.

[8] Arjovsky M, Chintala S and Bottou L. "Wasserstein GAN," in Proc. IEEE Int. Conf. Mach. Learn., 2017, pp. 214-223.

[9] Gulrajani I, Ahmed F, Arjovsky M, et al. "Improved Training of Wasserstein GANs," in Conf. Neural Inf. Process. Syst., 2017, pp. 5767-5777.

[10] Mao X, Li Q, Xie H, et al. "Least Squares Generative Adversarial Networks," in Proc. IEEE Int. Conf. Comput. Vision, 2017, pp. 2794-2802.

[11] Isola P, Zhu J Y, Zhou T, et al. "Image-to-image Translation with Conditional Adversarial Networks," in Proc. IEEE Conf. Comput. Vision Pattern Recognit., 2017, pp. 1125-1134.

[12] Wang T C, Liu M Y, Zhu J Y, et al. "High-resolution Image Synthesis and Semantic Manipulation with Conditional GANs," in Proc. IEEE Conf. Comput. Vision Pattern Recognit., 2018, pp. 8798-8807.

[13] Regmi K, Borji A. "Cross-view Image Synthesis Using Conditional GANs," in Proc. IEEE Conf. Comput. Vision Pattern Recognit., 2018, pp. 3501-3510.

[14] Quan T M, Nguyen-Duc T, et al. "Compressed Sensing MRI Reconstruction Using a Generative Adversarial Network with a Cyclic Loss," IEEE Trans. Med. Imaging, vol. 37, no. 6, pp. 1488-1497, 2018.

[15] Wu J, Zhang C, Xue T, et al. "Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-adversarial Modeling," in Conf. Neural Inf. Process. Syst., 2016, pp. 82-90.

[16] Yang B,et al."3D Object Reconstruction from a Single Depth View with Adversarial Learning," in Proc. IEEE Int. Conf. Comput. Vision, 2017, pp. 679-688.

[17] Qi C R, Su H, Mo K and Guibas L J. "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in Proc. IEEE Conf. Comput. Vision Pattern Recognit., 2017, pp. 652-660.

[18] Qi C R, Su H, et al. "Pointnet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," in Conf. Neural Inf. Process. Syst., 2017, pp. 5099-5108.