Hindawi
Advances in Multimedia
Volume 2024, Article ID 5528497, 12 pages
https://doi.org/10.1155/2024/5528497

Research Article

Design of 3D Environment Combining Digital Image Processing Technology and Convolutional Neural Network

Xiaofei Lu and Shouwang Li

Art College of Xi'an University of Science and Technology, Shaanxi 710054, China

Correspondence should be addressed to Xiaofei Lu; luxf230214@163.com

Received 29 March 2023; Revised 8 October 2023; Accepted 31 October 2023; Published 12 January 2024

Academic Editor: K Abhimanyu Kumar Patro

As virtual reality technology advances, 3D environment design and modeling have garnered increasing attention. Applications in networked virtual environments span urban planning, industrial design, and manufacturing, among other fields. However, existing 3D modeling methods exhibit high reconstruction error precision, limiting their practicality in many domains, particularly environmental design. To enhance 3D reconstruction accuracy, this study proposes a digital image processing technology that combines binocular camera calibration, stereo correction, and a convolutional neural network (CNN) algorithm for optimization and improvement. By employing the refined stereo-matching algorithm, a 3D reconstruction model was developed to augment 3D environment design and reconstruction accuracy while optimizing the 3D reconstruction effect. An experiment using the ShapeNet dataset demonstrated that the evaluation indices—Chamfer distance (CD), Earth mover's distance (EMD), and intersection over union—of the model constructed in this study outperformed those of alternative methods. After incorporating the CNN module in the ablation experiment, CD and EMD increased by an average of 0.1 and 0.06, respectively. This validates that the proposed CNN module effectively enhances point cloud reconstruction accuracy. Upon adding the CNN module, the CD index and EMD index in the dataset increased by an average of 0.34 and 0.54, respectively. These results indicate that the proposed CNN module exhibits strong predictive capabilities for point cloud coordinates. Furthermore, the model demonstrates good generalization performance.

1. Introduction

With the development of Internet technology, image processing technology has become an important means of information technology. People can easily use image processing technology to obtain information, so as to construct different technical models. The improvement of the image processing effect by computer is an important part of information realization. With the increasing demand for information technology in the whole society, image engineering is playing a more and more important role in contemporary science and technology.

With the development of virtual reality technology, 3D environment design and modeling technology have been paid more and more attention. It has been applied in virtual network environments, urban planning, industrial design, manufacturing, and other fields [1]. However, the existing 3D modeling methods have large error accuracy defects. In many fields, especially in environmental design, the practicability is limited to some extent [2]. Moreover, the 3D modeling of environmental design requires a high degree of 3D reduction. This is because the restoration accuracy of reconstruction methods based on a single perspective is limited [3]. With the progress of technology, the 3D modeling method based on double-view multidimensional data has gradually become the mainstream [4]. Under the multidirectional 3D modeling framework, the environment modeling method based on texture mapping can achieve 3D restoration to a certain extent [5]. In order to further improve the modeling accuracy, 3D reconstruction methods based on learn-perception classes have been widely studied.

There are many methods and theories for image-based 3D reconstruction. Among them, structure from motion recovery (SfM) is one of the most widely used classical methods [6]. SfM calculates that the feature points successfully matched between images have 3D information and can be restored to 3D coordinates to form 3D point clouds. However, the feature point information contained in the image is

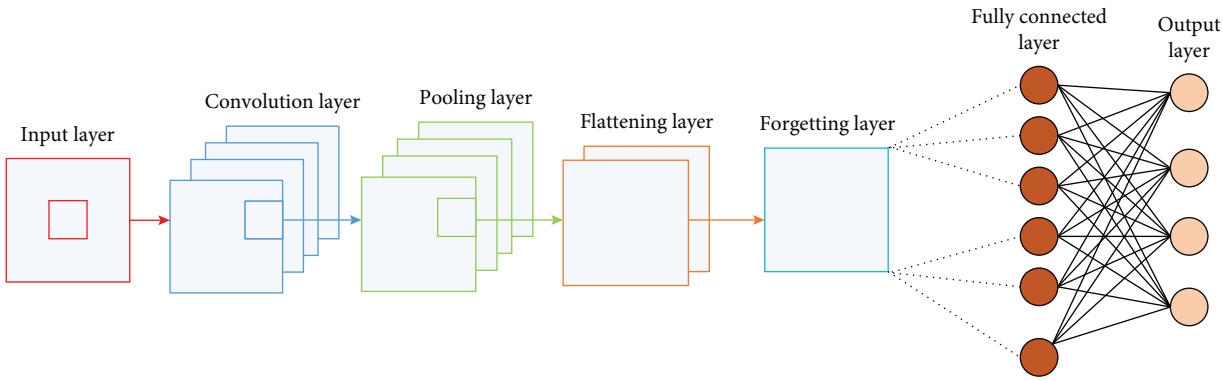

Figure 1: CNN structure diagram.

relatively small [7]. Therefore, the point cloud model calculated by SfM is sparse, and the accuracy of the reconstructed model is low. The multiview stereo (MVS) [8] method can calculate the dense 3D point cloud of the scene from multiple view images of the object. Patch-based MVS [9] takes sparse point cloud reconstructed by SfM as input information. Then, using the image surface neighborhood information iteration, a point cloud expansion strategy is used for point cloud expansion and filtering. Finally, dense point clouds are reconstructed by this method. Wang et al. [10] took the sparse reconstruction model and camera attitude obtained by SfM as input. This method uses depth map fusion to recover dense point clouds. The MVS method based on learning is shown in literature [11]. The depth map fusion method used in literatures [12, 13] is also effective in restoring high-precision dense point clouds in the scene. Literature [14] proposed a 3D model reconstruction method based on point cloud, which achieved better reconstruction accuracy by defining loss functions such as chamfering distance and spatial distance. Literature [15] classifies internal points and external points based on fusion features and proposes a point cloud sampling optimization strategy. The scheme allows for a more detailed reconstruction of the point cloud. In order to effectively restore the occlusion area of the single view of the object, literature [16] combines the 3D encoder–decoder structure with the generative antagonism network. The detailed dimensional structure of the object is reconstructed from a single view, and good experimental results are obtained on the synthesized dataset.

In order to improve the accuracy of 3D object reconstruction with a single view, a fusion of digital image processing technology and convolutional neural network (CNN) algorithm is proposed to optimize and improve CNN. Through the improved stereo-matching algorithm, the 3D reconstruction model was constructed to improve the 3D environment design and reconstruction accuracy and optimize the 3D reconstruction effect. Experiments on the dataset of ShapeNet [17] show that the evaluation indexes of Chamfer distance (CD), Earth mover's distance (EMD), and intersection over union (IoU) in the model experiments constructed in this paper are superior to other traditional methods. The ablation experiment also verifies that the CNN module proposed in this paper can effectively improve the reconstruction accuracy of point clouds, has a good prediction of point cloud coordinates, and the generalization performance of the model presented in this paper is also good.

2. State of the Art

2.1. Structure and Principle of CNNs. CNN is a representative algorithm in deep learning [18]. The algorithm is a deep feed-forward neural network with local connection and weight sharing. CNN continuously extracts features through multiple convolution kernels to realize image classification and natural language processing. CNN consists of an input layer, convolution layer, pooling layer, flattening layer, forgetting layer, and fully connected (FC) layer. Its structure is shown in Figure 1.

The convolution layer mainly realizes feature extraction of data. The convolution kernel in the convolutional layer slides on the input data one by one and carries out the dot product operation with the data at each position, and the output is the feature graph. The convolution operation can be expressed as shown in Formula (1):

$$C_x = f\ (g \times I_{xx+b-1} + h). \tag{1}$$

In the above formula, g represents weight and h represents bias.

The pooling layer replaces the network output in the region by using the region's overall characteristics. This can achieve the purpose of reducing network parameters and reducing the amount of calculation, so as to avoid the overfitting problem.

The flattening layer is the realization of 2D data 1D.

The forgetting layer is to temporarily hide some weight values by setting parameters to alleviate the occurrence of overfitting. This can achieve the regularization effect to a certain extent.

The FC layer completes the classification task. Output the data, get the classification result, and use the Sigmoid function to output the classification probability value. The function formula is shown in Formula (2):

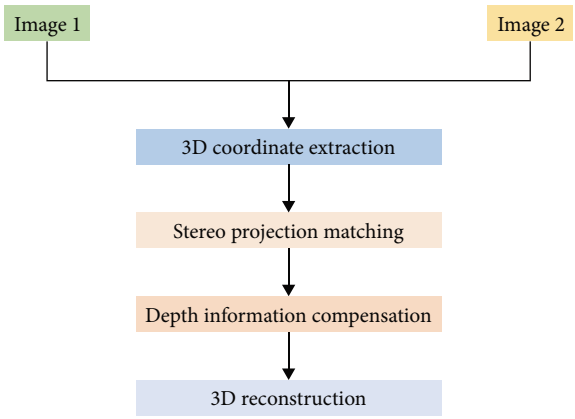

Figure 2: Principle of bidirectional 3D imaging method.

$$a(s) = \frac{1}{1 + e^{-s}}. \tag{2}$$

In Formula (2), s represents the output of the upper layer of the model.

2.2. Digital Image Processing Technology. Digital image processing technology is widely used for the practical needs of environment design. Among them, stereo imaging technology is developing rapidly. This paper studies the principle of 3D environment design based on stereo imaging technology. Digital image processing technology can effectively model 3D scenes and improve the authenticity of environmental design. See Figure 2 for the specific method principle.

The 3D coordinates of scenes in different coordinate systems can be extracted by triangular projection. On this basis, this paper uses the stereo projection-matching algorithm to coordinate the pixel points of the 3D scene. Considering the 3D reconstruction modeling using 2D images will have stereo distortion. Based on traditional 3D modeling, this paper can make stereo compensation for the extracted image depth information and finally realize the reconstruction of a highly restored 3D scene. The schematic diagram of the nonparallel bidirectional stereoscopic imaging 3D modeling method is shown in Figure 3.

In Figure 3, U is projected sterically in two coordinate systems, O_1 and O_2. Its projection points in the projection plane are, respectively, U_1 and U_2. The observed coordinates of U_1 and U_2 in the coordinate system with the origin of O_1 and O_2 are $U_1(i_1, j_1)$ and $U_2(i_2, j_2)$, respectively. Let I_n represent the true coordinates of U. Use I_1 and I_r to represent the coordinates of U_1 and U_2, respectively, in the observed coordinate system, then the corresponding relationship can be obtained, as shown in Formula (3):

$$\begin{cases} I_1 = z_1 I_n + n_1 \\ I_r = z_r I_n + n_r \end{cases}, \tag{3}$$

where z_1, z_r, n_1, and n_r are the parameters of the stereoscopic projection transformation between the two observed coordinate systems and the real 3D coordinate systems. Transform

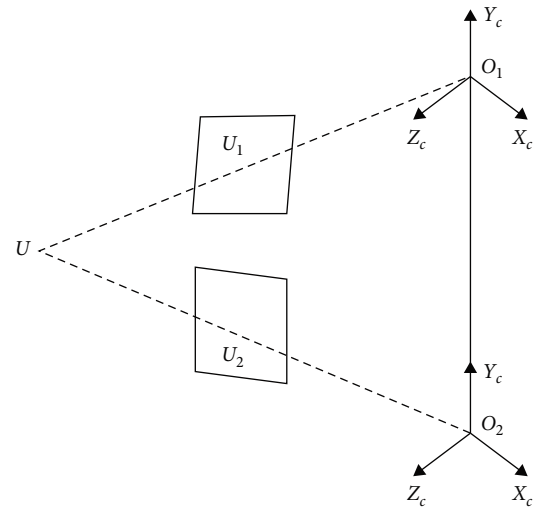

Figure 3: Nonparallel bidirectional stereoscopic imaging 3D modeling diagram.

Formula (3), as shown in Formula (4).

$$I_r = \boldsymbol{K} I_1 + \boldsymbol{T}, \tag{4}$$

where $\boldsymbol{K}$ and $\boldsymbol{T}$ are stereoscopic projection transformation parameter matrices. It is defined as shown in Formula (5).

$$\begin{cases} \boldsymbol{K} = z_r z_1^{-1} \\ \boldsymbol{T} = n_r - Z n_1 \end{cases}. \tag{5}$$

The stereo projection transformation parameters are different at different points. 3D matching is a nonlinear optimization to determine the optimal stereo projection transformation parameter matrix.

3. Methodology

The algorithm in this paper is a combination of digital image processing technology and an improved CNN algorithm. The stereo-matching algorithm model can be constructed by this algorithm.

In the experiment part, the reconstruction accuracy of the data is measured, and the 3D reconstruction effect of different models is analyzed. Through the analysis and verification of the model, the model with higher reconstruction accuracy and better 3D reconstruction effect is selected. Through this model, the precision of 3D reconstruction can be improved, so as to achieve the purpose of optimizing 3D environment design.

3.1. Stereo-Matching Algorithm Based on Improved CNN. Deformable CNN is a deep learning model for image processing that adaptively adjusts the shape of the convolution kernel to better capture nonlinear features in images. By introducing deformable convolution, the algorithm is able to more accurately capture the subtle differences in the surface of an object in a stereoscopic image, which improves

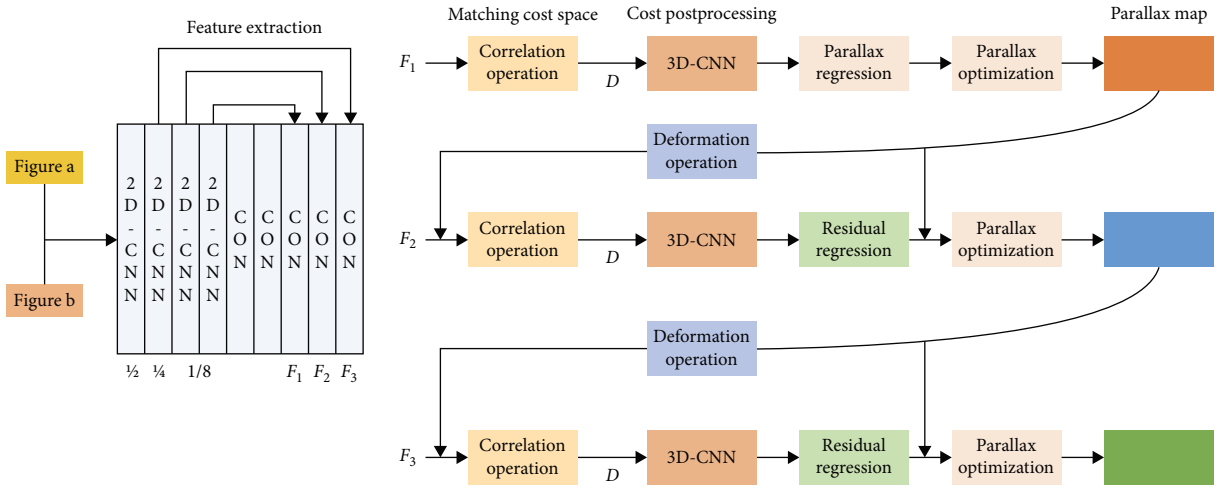

FIGURE 4: Design of the stereo-matching algorithm.

the accuracy and detail representation of point cloud reconstruction.

The stereo-matching algorithm based on deformable convolution is composed of feature extraction, matching cost space, cost postprocessing, parallax/residual regression, and parallax optimization modules. The design structure of the stereo-matching algorithm is shown in Figure 4.

The feature extraction module is an encoder–decoder that introduces a 2D deformable convolution hourglass in the encoding stage. The matching cost space is constructed by the associated operation of DispNetC to form the 3D cost space. In the cost postprocessing module, the 3D deformable convolution of the residual structure is used to regularize the matching cost space. The parallax regression module adopts the softargmin method proposed by GC-Net. Its expression is shown in Formula (6):

$$\widehat{d} = \sum_{d=0}^{D_{\max}^{-1}} d \times \sigma(-c_d), \tag{6}$$

where $\widehat{d}$ represents the predicted parallax value. d represents the parallax value of the candidate. $D_{\max}$ represents the maximum candidate parallax. σ indicates the softmax function. c_d indicates the matched generation value.

The parallax optimization module is a spatial propagation network [19]. The network can extract the similarity matrix of the image and optimize the predicted parallax value.

The algorithm is divided into three stages to get a parallax map with different precision.

In the first stage, the feature extraction module extracted feature map F1 with a resolution of 1/16. Therefore, the candidate parallax value ranges from 0 to 1/16 $D_{\max}$. After parallax regression and optimization, it is necessary to obtain the parallax map of the first stage by up-sampling operation and multiplying by 16 times.

In the second stage, the range of candidate residual d is set to −2–2. According to the parallax map from Stage 1, the new feature map is warped on the right feature map F2 at 1/8 resolution. Then, the matching cost space is formed with the left feature map. The residuals of regression are added to the parallax map of Stage 1. Then, the parallax map is optimized to get the parallax map of the second stage.

The third stage is the same as the second stage.

3.2. Deformable Convolution. An ordinary convolution consists of two steps. The process is shown below:

(1) A regular grid R is used for sampling on input feature graph i.

(2) The sampling value is multiplied by the weight m and summed. For example, $R = \{(-1, 0),\ldots,(0, 1), (1, 1)\}$ represents a 3×3 grid with expansion rate of 1. For each position u_0 on the output feature graph y, the expression is shown in Formula (7):

$$j(u_0) = \sum_{u_t \in R} m(u_t) \cdot i(u_0 + u_t), \tag{7}$$

where u_t represents every position belonging to R. In the deformable convolution, R has an offset $\{\Delta u_t | t = 1, \cdots T = |R|\}$. Transform Formula (7) into Formula (8):

$$j(u_0) = \sum_{u_t \in R} m(u_t) \cdot i(u_0 + u_t + \Delta u_t). \tag{8}$$

Now, the sampling is $u_t + \Delta u_t$ at the regular and offset position. Because Δu_t is a decimal, Formula (8) needs to be implemented by linear interpolation. Its expression is shown in Formula (9):

$$i(u) = \sum_{v} A(v, u) \cdot i(v). \tag{9}$$

In the above formula, u represents any position. In Formula (8), $u = u_0 + u_t + \Delta u_t$. v represents each integer position in the feature graph i. $A(*)$ has two dimensions and can

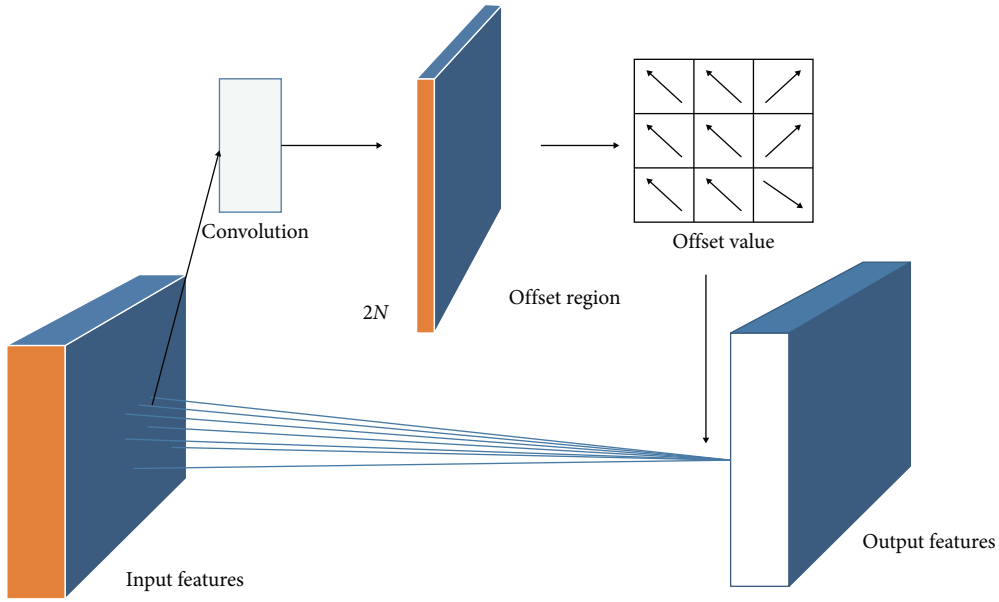

Figure 5: 3 × 3 2D deformable convolution.

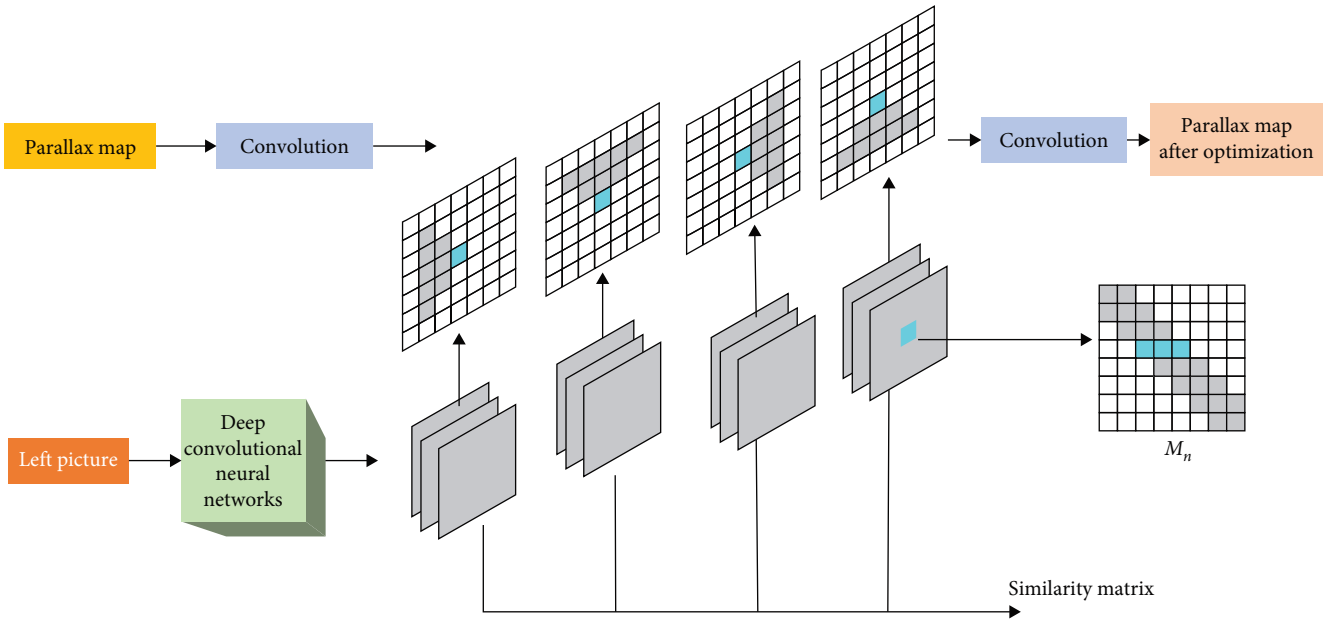

Figure 6: Spatial communication network structure.

be divided into two 1D cores. Its expression is shown in Formula (10).

$$A(v,u) = g(v_i,u_i)\cdot g\left(v_j,u_j\right), \tag{10}$$

where $g(v_i,u_i) = \max(0,1-|v_i-u_i|)$.

Figure 5 shows a 2D deformable convolution with a convolution kernel size of 3 × 3. The offset value is obtained by adding a layer of convolution to the same feature graph. The size and expansion rate of the convolution kernel are similar to the current deformable convolution kernel. $2N$ is the number of channels in the convolution, corresponding to N 2D offsets. 3D deformable convolution is a generalization of 2D deformable convolution. The principle is the same as in two dimensions, but one dimension is added to the dimension of the convolution.

3.3. Space Propagation. The spatial propagation network structure is shown in Figure 6, a parallax map used to optimize regression. It mainly consisted of a differentiable linear propagation module and a deep CNN model that learned the similarity matrix. Linear propagation of spatial propagation network is to scan the matrix row by row or column by row in four fixed directions. The four fixed directions are left to

right, right to left, top to bottom, and bottom to top. The following is mainly introduced from left to right direction, and other directions are the same principle.

First, assume two 2D images, I and B, both of size $t \times t$, where I is the image before spatial propagation. B is the image after space propagation. i_t and b_t are their respective tth columns. They are both $t \times 1$ in size. Linear propagation is performed from left to right in two adjacent columns using the $t \times t$ linear transformation matrix M_n. Its expression is shown in Formula (11):

$$b_n = (X - D_n)i_n + M_n b_{n-1}, n \in [2, t], \tag{11}$$

where M denotes the $t \times t$ identity matrix. The initial condition is $b_1 = i_1$. $D_n(x, x)$ is the diagonal matrix. The xth entry is the sum of row x in $\boldsymbol{M}_n$. Its expression is shown in Formula (12):

$$\boldsymbol{D}_n(x, x) = \sum_{y=1, y \neq x}^{t} \boldsymbol{M}_n(x, y). \tag{12}$$

Therefore, the matrix $B(\in B, n \in [1, t])$ is updated recursively by column. For each column, b_n is the preceding column b_{n-1} multiplied by the matrix $\boldsymbol{M}_n$ and combined with x_n, which is linear.

When the recursion is complete, the matrix expression of Formula (11) is shown in Formula (13):

$$\boldsymbol{H}_q = \begin{bmatrix} \boldsymbol{I} & 0 & \cdots & \cdots & 0 \\ \boldsymbol{M}_2 & \lambda_2 & 0 & \cdots & \cdots \\ \boldsymbol{M}_3\boldsymbol{M}_2 & \boldsymbol{M}_3\lambda_2 & \lambda_3 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \lambda_t \end{bmatrix} \boldsymbol{X}_q = \boldsymbol{G}\boldsymbol{X}_q, \tag{13}$$

where $\boldsymbol{G}$ represents a triangular transformation matrix under $T \times T$ $(T = t^2)$. $\boldsymbol{H}_q = [b_1^N, \cdots, b_t^N]^N, \boldsymbol{X}_q = [i^N, \cdots, i_t^N]^N$. The dimension is $T \times 1$. The parameter is $\{\lambda_n, M_n, D_n, X\}$, $n \in [2, t]$ and the size is $t \times t$, $\lambda_n = X - D_n$.

The deep CNN module is mainly used to output the similarity matrix A, and then linear propagation is carried out to obtain $\boldsymbol{H}_q$. The algorithm mainly uses deep CNN and linear propagation modules to learn $\boldsymbol{H}$ from the left image to guide the optimization of the regression parallax map.

3.4. Loss Function. In order to predict the position of a point cloud, EMD, CD, symmetric loss, and an equidistant prior loss are used as loss functions for model training. The specific definition is as follows:

(1) EMD

EMD is defined as the minimum sum of the distances between elements u in the set and all elements in the set S_{an}. Its expression is shown in Formula (14).

$$L_{\text{EMD}}(S_1, S_{\text{an}}) = \min_{\sigma S_1 \to S_{an}} \sum_{u \in S_1} \| u - \sigma(u) \|_2, \tag{14}$$

where S_1 stands for reconstructed point cloud, and S_{an} stands for ground truth (GT) true point cloud. σ is the bijective relation.

(2) CD

The CD is used to measure the distance between two sets of point clouds. Formally defined as Formula (15):

$$\begin{aligned} L_{\text{CD}}(S_1, S_{\text{an}}) = & \sum_{i_1 \in S_1} \min_{i_2 \in S_{\text{sn}}} \| i_1 - i_2 \|_2^2 \\ & + \sum_{i_2 \in S_s} \min_{i_1 \in S_1} \| i_2 - i_1 \|_2^2. \end{aligned} \tag{15}$$

The first term represents the sum of the minimum distances from any point in S_1 to S_{an}, and the second term represents the sum of the minimum distances from any point in S_{an} to S_1.

(3) Equidistant prior loss

Let S_1 be the reconstructed point cloud and s be any point in S_1. $S_x(S_i^x, S_j^x, S_k^x)$ is the xth adjacent point to s. After Gaussian filtering, the position of s changes accordingly. Take x coordinate as an example, as shown in Formulae (16) and (17).

$$s_i^{'} = \sum_x f(s_i^x) \times s_i^x, \tag{16}$$

$$f(s_i^x) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(s_i^x - s_i)^2}{2\sigma^2} \right). \tag{17}$$

Equidistant prior losses are defined as shown in Formula (18):

$$L_{\text{iso}} = L_{\text{CD}}(S', S_1), \tag{18}$$

where S_1 is the initial point cloud, and S' is the point cloud after Gaussian filtering. The introduction of equidistant prior loss function can make adjacent points close to each other.

(4) Symmetric loss

In order to maintain the symmetry of the point cloud model in the deformation process, the symmetric loss function of the point cloud is introduced, and the expression is shown in Formula (19).

$$L_{\text{sym}} = L_{\text{CD}}(M(S_1), S_{\text{an}}). \tag{19}$$

In the above formula, $M(S_1)$ is the specular reflection transformation.

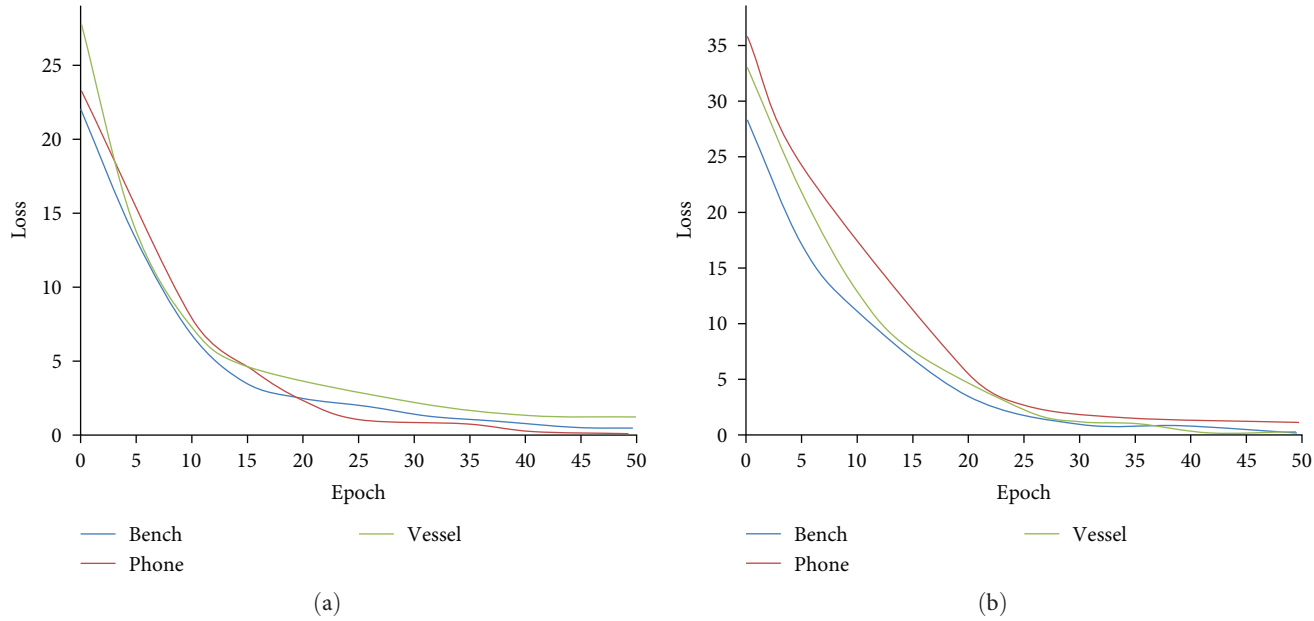

Figure 7: Convergence curve of training loss function: (a) 3D-matching algorithm training process loss function convergence curve; (b) figure convergence curve of the loss function in the convolution training process.

4. Result Analysis and Discussion

4.1. Experimental Setup. In all experiments, the model inputs are RGB color images, and the output is a 3D point cloud with 2,048 vertices. Meanwhile, in order to train the graph-convolutional network end-to-end, the Ad-am optimizer is used in the experiments, and the learning rate is initialized to 5×10^{-5}. The number of iterations of the model is 50 epochs, and the batch size is 32. All the experiments are implemented on NVIDIA GeForce GTX1080Ti GPUs using the open-source machine learning framework Pytorch.4.2 Experimental data and evaluation criteria

In order to evaluate the reconstruction performance of the proposed algorithm, ShapeNet synthetic dataset, ModelNet, and dataset and Pix3D [20, 21] real scene dataset were used for experiments. ShapeNet has a total of 51,300 3D models in 13 model categories. The ModelNet dataset contains about 17,210 3D models in about 50 different categories. The partially occluded or truncated data is excluded, and the training set and test set are randomly divided according to the ratio of 4 : 1. The same Pix3D dataset is used to do the preprocessing, with the background of the mask information to remove useless background and moved to the center of the object, will eventually image zooming or cut to 224 by 224 as the input image. In this paper, IoU, CD, and EMD were used as indicators to measure experimental results. IoU represents the intersection ratio between the 3D voxel shape of the network reconstruction and the shape of the true solid element. Here, the same voxel generation method as literature [14] is adopted. CD and EMD represent the difference between two point clouds. Here, the GT point cloud is sampled to generate a point cloud model with a number of vertices of 2,048, and the reconstructed point cloud is compared with the reconstructed point cloud in this paper.

4.2. Experimental Data and Evaluation Criteria. Verify the robustness of the loss function design strategy proposed in this paper, as shown in Figure 7. Figure 7(a) shows the comparison of the effects of loss function on different training sets. By comparison, it can be seen that on the three different training sets, the loss function of the training set generally keeps a downward trend during the training. The loss function of the training set decreases rapidly in the first 25 times of the epoch and tends to be stable after the 40th time. It can be seen that the method in this paper has high robustness. Further, Figure 7(b) shows the convergence of the loss function in the point cloud deformation process of the CNN. It can be seen from Figure 7(b) that the CNN has a good convergence result in the deformation stage, indicating that the model has a good 3D reconstruction effect.

4.3. Quantitative Comparison of Experimental Results. In order to quantitatively analyze the differences between the proposed method and other methods, Tables 1 and 2 show the comparison of reconstruction accuracy in the ShapeNet dataset and ModelNet dataset. The evaluation index was scaled 100 times and compared with the methods of literatures [14, 22, 23]. In terms of CD evaluation indexes, the method in this paper achieves higher reconstruction accuracy in 13 categories, such as airplanes. Similarly, in terms of EMD evaluation indexes, the method in this paper is superior to other methods in all categories. The average reconstruction accuracy of CD and EMD is higher than that of other methods.

Further, we compared the differences between the proposed method and literatures [22, 23] in different categories of IoU. As can be seen from Table 3, the IoU of this paper's method is higher in eight categories, such as airplane and literature [22], and is higher in sofa and speaker.

Table 1: CD and EMD evaluation indicators on ShapeNet dataset.

Item	CD				EMD			
	Literature [14]	Literature [22]	Literature [23]	Ours	Literature [14]	Literature [22]	Literature [23]	Ours
Airplane	3.76	3.32	3.34	**2.37**	6.41	4.75	3.82	**2.65**
Bench	4.61	4.54	4.65	**3.38**	5.91	5.04	4.32	**3.43**
Cabinet	6.97	6.14	6.12	**4.47**	6.06	6.37	4.96	**4.33**
Car	5.24	4.57	4.36	**3.31**	4.14	4.83	3.65	**2.92**
Chair	6.41	6.40	6.54	**3.43**	9.71	8.03	6.47	**3.52**
Lamp	6.33	7.12	6.60	**4.88**	16.23	15.91	8.54	**6.17**
Monitor	6.13	6.42	6.46	**4.52**	7.61	7.24	5.94	**4.13**
Phone	4.55	4.62	4.24	**3.45**	5.13	5.41	3.82	**3.45**
Rifle	2.93	2.74	2.83	**2.32**	8.45	6.17	4.22	**3.71**
Sofa	7.02	5.84	5.93	**4.16**	7.45	5.61	5.04	**3.85**
Speaker	8.72	8.16	8.44	**5.62**	8.71	9.16	7.34	**5.06**
Table	6.02	6.08	6.25	**4.04**	8.41	7.84	6.03	**4.35**
Vessel	4.38	4.44	4.53	**3.51**	6.24	5.75	4.93	**3.86**
Mean	5.60	5.43	5.44	**3.83**	7.71	7.06	5.32	**3.94**

The bold data represent a comparison of the data obtained by the method used in this article compared to other methods.

Table 2: CD, EMD evaluation indicators on ModelNet dataset.

Item	CD				EMD			
	Literature [14]	Literature [22]	Literature [23]	Ours	Literature [14]	Literature [22]	Literature [23]	Ours
Airplane	2.65	2.21	2.23	**1.26**	5.3	3.64	2.71	**1.54**
Bench	3.5	3.43	3.54	**2.27**	4.8	3.93	3.21	**2.32**
Cabinet	5.86	5.03	5.01	**3.36**	4.95	5.26	3.85	**3.22**
Car	4.13	3.46	3.25	**2.2**	3.03	3.72	2.54	**1.81**
Chair	5.3	5.29	5.43	**2.32**	8.6	6.92	5.36	**2.41**
Lamp	5.22	6.01	5.49	**3.77**	15.12	14.8	7.43	**5.06**
Monitor	5.02	5.31	5.35	**3.41**	6.5	6.13	4.83	**3.02**
Phone	3.44	3.51	3.13	**2.34**	4.02	4.3	2.71	**2.34**
Rifle	1.82	1.63	1.72	**1.21**	7.34	5.06	3.11	**2.6**
Sofa	5.91	4.73	4.82	**3.05**	6.34	4.5	3.93	**2.74**
Speaker	7.61	7.05	7.33	**4.51**	7.6	8.05	6.23	**3.95**
Table	4.91	4.97	5.14	**2.93**	7.3	6.73	4.92	**3.24**
Vessel	3.27	3.33	3.42	**2.4**	5.13	4.64	3.82	**2.75**
Mean	4.49	4.32	4.33	**2.72**	6.6	5.95	4.21	**2.83**

The bold data represent a comparison of the data obtained by the method used in this article compared to other methods.

Literature [23] achieved the best performance in the car and phone categories under 5-view reconstruction. Overall, on the ShapeNet dataset, the average IoU of the proposed method is improved by 9.16% over the literature [23] in five views and 7.63% over the literature [22]. On ModelNet dataset, the average IoU of the proposed method is improved by 11.11% over literature [23] at five views and 9.22% over literature [22].

4.4. Comparison of Ablation Data

(1) CNN module ablation experiment comparison

In this paper, the CNN module is used to adjust the 3D reconstructed point cloud model of the stereo-matching algorithm. In order to verify the effectiveness of this method, the CNN module is replaced by a common FC layer, and the model is trained and tested. CD and EMD are used to measure the quality of the generated point cloud, and the test results are shown in Table 4.

As can be seen from Table 4, after the CNN module is added, CD and EMD have a certain improvement in most datasets. CD and EMD schemes only showed slight declines in some datasets. CD increases by 0.1 on average, and EMD increases by 0.07 on average. For the CD indicator, the chair dataset was increased by 0.34. For EMD indicators, the monitor dataset is increased by 0.44. It can be seen that the introduction of the CNN module can effectively improve the accuracy of point cloud reconstruction.

The performance of the stereo-matching algorithm is verified by experiments. Evaluation indicators were trained and tested on bench, monitor, and phone datasets. As shown in Table 5, after the CNN module is added, the evaluation indexes of different datasets are improved. CD index

Table 3: IoU evaluation indicators.

Item	ShapeNet dataset				ModelNet dataset			
	Literature [22]	Literature [23]		Ours	Literature [22]	Literature [23]		Ours
		3 views	5 views			3 views	5 views	
Airplane	0.605	0.551	0.562	0.721	0.494	0.44	0.451	0.61
Bench	0.553	0.504	0.531	0.764	0.442	0.393	0.42	0.653
Cabinet	0.775	0.761	0.776	0.73	0.664	0.65	0.665	0.619
Car	0.832	0.836	0.843	0.654	0.721	0.725	0.732	0.543
Chair	0.541	0.533	0.554	0.682	0.43	0.422	0.443	0.571
Lamp	0.462	0.414	0.427	0.634	0.351	0.303	0.316	0.523
Monitor	0.552	0.547	0.563	0.665	0.441	0.436	0.452	0.554
Phone	0.751	0.736	0.754	0.752	0.64	0.625	0.643	0.641
Rifle	0.607	0.593	0.604	0.714	0.496	0.482	0.493	0.603
Sofa	0.714	0.692	0.703	0.636	0.603	0.581	0.592	0.525
Speaker	0.743	0.714	0.725	0.747	0.632	0.603	0.614	0.636
Table	0.601	0.563	0.58	0.652	0.49	0.452	0.469	0.541
Vessel	0.615	0.602	0.613	0.635	0.504	0.491	0.502	0.524
Mean	0.642	0.619	0.633	0.691	0.531	0.508	0.522	0.58

Table 4: Evaluation indicators of CNN ablation experiments CD.

Item	CD		Deviation	EMD		Deviation
	FC	CNN		FC	CNN	
Airplane	2.42	2.32	0.22	2.67	2.63	0.04
Bench	3.38	3.35	0.03	3.56	3.42	0.14
Cabinet	4.52	4.44	0.08	4.15	4.31	−0.16
Car	3.43	3.31	0.12	3.03	2.9	0.13
Chair	3.76	3.42	0.34	3.68	3.53	0.15
Lamp	4.75	4.91	−0.16	5.76	6.11	−0.35
Monitor	4.53	4.51	0.02	4.58	4.14	0.44
Phone	3.74	3.46	0.27	3.51	3.48	0.03
Rifle	2.45	2.33	0.12	3.98	3.71	0.27
Sofa	4.18	4.13	0.05	4.07	3.83	0.24
Speaker	5.82	5.65	0.17	4.85	5.02	−0.17
Table	3.96	4.07	−0.13	4.46	4.35	0.11
Vessel	3.67	3.52	0.15	3.92	3.84	0.08
Mean	3.9	3.8	0.1	4.01	3.94	0.07

Table 5: Evaluation indicators of ELAS ablation experiments.

Item	CD		Deviation	EMD		Deviation
	FC	CNN		FC	CNN	
Bench	3.73	3.39	0.34	4.12	3.47	0.65
Monitor	4.75	4.51	0.24	4.94	4.17	0.77
Phone	3.96	3.45	0.51	3.65	3.48	0.17
Mean	4.14	3.78	0.36	4.23	3.71	0.53

increased by 0.36 on average, and the EMD index increased by 0.53 on average. It is proved that the CNN module has a good prediction for point cloud coordinates.

(2) Loss function ablation experiment comparison

In order to verify the effectiveness of the loss function adopted in this paper, different combinations of loss functions are selected, and the model is retrained. Based on bench, rifle, and vessel datasets, the test results are shown in Table 6. It can be seen from Table 5 that after all loss

Table 6: CD comparison of loss function ablation experiments item.

Item	Remove isometric prior loss	Remove symmetry loss	All loss
Bench	3.3632	3.3597	3.3447
Rifle	2.3502	2.3466	2.3339
Vessel	3.5753	3.5464	3.5275
Mean	3.0962	3.0842	3.0687

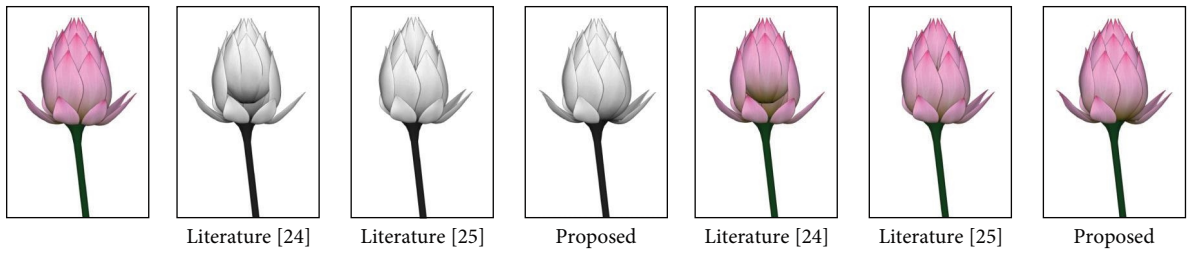

Figure 8: Comparison of 3D reconstruction results: (a) original image; (b) the modeled image; (c) image after texture map.

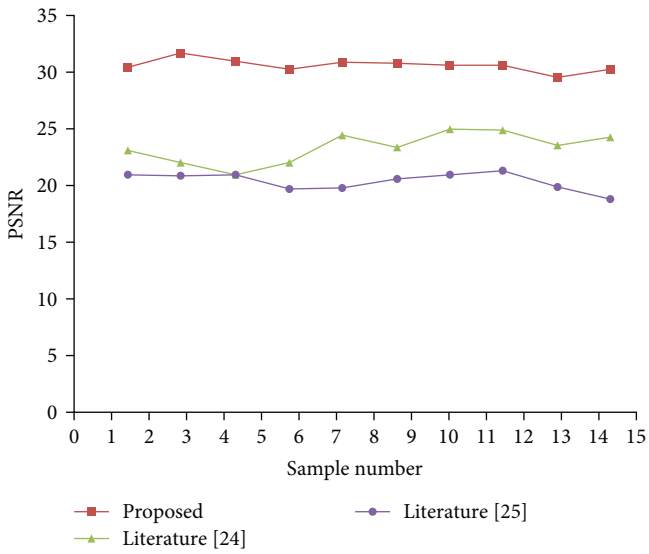

Figure 9: Analysis result of image distortion degree.

functions are adopted, CD performs better than the other two strategies and is effective for different datasets, improving the generalization performance of the model.

4.5. Comparison of 3D Modeling. In order to test the effectiveness of the algorithm, the reduction degree of this paper and different algorithm models is compared, which is shown in Figure 8. In this paper, the lotus flower is chosen as the experiment in the reconstruction of the natural environment. The algorithm in this paper, literatures [24, 25] are used to reconstruct the 3D model of the same lotus flower in the collected sample data. The model effect after reconstruction is shown in Figure 8(b).

According to Figure 8(c), by comparing the image models reconstructed by the three algorithms, we can see that the model reconstructed by the proposed algorithm is clearer. The distortion degree of both rod diameter part and petal part is small. After texture mapping, the image restoration degree is higher, and the feature point recognition is more accurate.

In order to verify the distortion degree of reconstructed images, PSNR values of red dog images were compared by the above three methods. The comparison results are shown in Figure 9. The image with a higher PSNR value has a lower distortion degree, which proves that the image restoration quality is higher.

5. Conclusion

In this study, we combine binocular camera calibration and stereo correction of digital image processing technology with a CNN to optimize and improve the 3D reconstruction method, constructing a 3D reconstruction model using a stereo-matching algorithm. In the experimental portion, we measure the reconstruction accuracy of the data and analyze the 3D reconstruction effects of different models. Experiments demonstrate that the proposed method achieves higher reconstruction accuracy in 13 categories, such as airplanes. Regarding EMD evaluation indices, the proposed method outperforms other methods in all categories. In terms of average reconstruction accuracy, the proposed algorithm yields better CD and EMD results compared to other methods. The proposed algorithm also demonstrates good performance in terms of average IoU. After incorporating the CNN module in the ablation experiment, CD and EMD increased by an average of 0.1 and 0.06, respectively. This validates that the proposed CNN module effectively enhances point cloud reconstruction accuracy. Upon adding the CNN module, the CD index and EMD index in the dataset increased by an average of 0.34 and 0.54, respectively, indicating that the proposed CNN module has strong

predictive capabilities for point cloud coordinates. Furthermore, the model demonstrates good generalization performance.

Despite the significant 3D reconstruction accuracy improvement achieved by the proposed method, however, there are some limitations of the method and areas that need to be further explored. For example, (1) the CNN may be sensitive to input variations such as lighting conditions, object orientation, and occlusion. There is a need to further investigate the robustness of the method to these variables. Techniques to improve the robustness of the method to noise, uncertainty, and occlusion will be further explored in the future to enhance its performance in real-world scenarios. (2) The paper provides an overview of stereo-matching algorithms based on deformable CNNs, but the complexity and computational cost of the algorithms are not discussed in detail. It is necessary to elaborate on the practical feasibility of the method in real-time or resource-limited situations. Carry out case studies in specific application scenarios. In the future, some real-world scenarios, such as industrial automation, robot navigation, urban planning, and industrial design, will be selected for practical applications, and the performance of the algorithm will be tested in these scenarios.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] J. Abich IV, J. Parker, J. S. Murphy, and M. Eudy, "A review of the evidence for training effectiveness with virtual reality technology," *Virtual Reality*, vol. 25, pp. 919–933, 2021.

[2] F. Okura, "3D modeling and reconstruction of plants and trees: a cross-cutting review across computer graphics, vision, and plant phenotyping," *Breeding Science*, vol. 72, no. 1, pp. 31–47, 2022.

[3] L. Gong, W. Wang, T. Wang, and C. Liu, "Robotic harvesting of the occluded fruits with a precise shape and position reconstruction approach," *Journal of Field Robotics*, vol. 39, no. 1, pp. 69–84, 2022.

[4] V. C. Anadebe, P. C. Nnaji, O. D. Onukwuli et al., "Multidimensional insight into the corrosion inhibition of salbutamol drug molecule on mild steel in oilfield acidizing fluid: experimental and computer aided modeling approach," *Journal of Molecular Liquids*, vol. 349, Article ID 118482, 2022.

[5] X. Fan, B. Zhou, and H. H. Wang, "Urban landscape ecological design and stereo vision based on 3D mesh simplification algorithm and artificial intelligence," *Neural Processing Letters*, vol. 53, pp. 2421–2437, 2021.

[6] S. Jiang, C. Jiang, and W. Jiang, "Efficient structure from motion for large-scale UAV images: a review and a comparison of SfM tools," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 167, pp. 230–251, 2020.

[7] Y. Wang and Y. Chen, "Non-destructive measurement of three-dimensional plants based on point cloud," *Plants*, vol. 9, no. 5, Article ID 571, 2020.

[8] Y. Peng, Z. Wu, G. Cao et al., "Three-dimensional reconstruction of wear particles by multi-view contour fitting and dense point-cloud interpolation," *Measurement*, vol. 181, Article ID 109638, 2021.

[9] R. Zhang, X. Yi, H. Li et al., "Multiresolution patch-based dense reconstruction integrating multiview images and laser point cloud," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 153–159, 2022.

[10] Q. Wang, X. Zhou, B. Hariharan, and N. Snavely, "Learning feature descriptors using camera pose supervision," in *Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science*, vol. 12346, pp. 757–774, Springer International Publishing, Glasgow, UK, 2020.

[11] R. Chen, S. Han, J. Xu, and H. Su, "Visibility-aware point-based multi-view stereo network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3695–3708, 2020.

[12] S. Song, K. G. Truong, D. Kim, and S. Jo, "Prior depth-based multi-view stereo network for online 3D model reconstruction," *Pattern Recognition*, vol. 136, Article ID 109198, 2023.

[13] E. R. Hillesund, L. R. Sagedal, E. Bere, and N. C. Øverby, "Family meal participation is associated with dietary intake among 12-month-olds in Southern Norway," *BMC Pediatrics*, vol. 21, Article ID 128, 2021.

[14] S. V. Brant Pinheiro, V. B. de Freitas, G. V. de Castro et al., "Acute post-streptococcal glomerulonephritis in children: a comprehensive review," *Current Medicinal Chemistry*, vol. 29, no. 34, pp. 5543–5559, 2022.

[15] R. Li, X. Li, P.-A. Heng, and C.-W. Fu, "PointAugment: an auto-augmentation framework for point cloud classification," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6377–6386, Seattle, WA, USA, 2020.

[16] T. Ma, P. Kuang, and W. Tian, "An improved recurrent neural networks for 3D object reconstruction," *Applied Intelligence*, vol. 50, pp. 905–923, 2020.

[17] F. Nammour, U. Akhaury, J. N. Girard et al., "ShapeNet: shape constraint for galaxy image deconvolution," *Astronomy & Astrophysics*, vol. 663, Article ID A69, 2022.

[18] P. Wang, E. Fan, and P. Wang, "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning," *Pattern Recognition Letters*, vol. 141, pp. 61–67, 2021.

[19] Y. Ji, Z. Kang, and X. Liu, "The data filtering based multiple-stage Levenberg–Marquardt algorithm for Hammerstein nonlinear systems," *International Journal of Robust and Nonlinear Control*, vol. 31, no. 15, pp. 7007–7025, 2021.

[20] Y. Jiang, Y. Li, S. Zou, H. Zhang, and Y. Bai, "Hyperspectral image classification with spatial consistence using fully convolutional spatial propagation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 10425–10437, 2021.

[21] D. Liu, G. Han, P. Liu et al., "A novel 2D–3D CNN with spectral-spatial multi-scale feature fusion for hyperspectral image classification," *Remote Sensing*, vol. 13, no. 22, Article ID 4621, 2021.

[22] B. Li, Y. Zhang, B. Zhao, and H. Shao, "3D-ReConstnet: a single-view 3D-object point cloud reconstruction network," *IEEE Access*, vol. 8, pp. 83782–83790, a single-view 3d-object point cloud reconstruction, 2020.

[23] H. Chen and Y. Zuo, "3D-ARNet: an accurate 3D point cloud reconstruction network from a single-image," *Multimedia Tools and Applications*, pp. 1–14, 2022.

[24] X. Wang, Y. Guo, Z. Yang, and J. Zhang, "Prior-guided multi-view 3D head reconstruction," *IEEE Transactions on Multimedia*, vol. 24, pp. 4028–4040, 2021.

[25] C. Chen, D. Gong, H. Wang, Z. Li, and K. Y. K. Wong, "Learning spatial attention for face super-resolution," *IEEE Transactions on Image Processing*, vol. 30, pp. 1219–1231, 2020.