ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20230323000469
Original scientific paper

3D Reconstruction of Optical Building Images Based on Improved 3D-R2N2 Algorithm

Qianying ZOU*, Fengyu LIU

Abstract: Three-dimensional reconstruction technology is a key element in the construction of urban geospatial models. Addressing the current shortcomings in reconstruction accuracy, registration results convergence, reconstruction effectiveness, and convergence time of 3D reconstruction algorithms, we propose an optical building object 3D reconstruction method based on an improved 3D-R2N2 algorithm. The method inputs preprocessed optical remote sensing images into a Convolutional Neural Network (CNN) with dense connections for encoding, converting them into a low-dimensional feature matrix and adding a residual connection between every two convolutional layers to enhance network depth. Subsequently, 3D Long Short-Term Memory (3D-LSTM) units are used for transitional connections and cyclic learning. Each unit selectively adjusts or maintains its state, accepting feature vectors computed by the encoder. These data are further passed into a Deep Convolutional Neural Network (DCNN), where each 3D-LSTM hidden unit partially reconstructs output voxels. The DCNN convolutional layer employs an equally sized 3 × 3 × 3 convolutional kernel to process these feature data and decode them, thereby accomplishing the 3D reconstruction of buildings. Simultaneously, a pyramid pooling layer is introduced between the feature extraction module and the fully connected layer to enhance the performance of the algorithm. Experimental results indicate that, compared to the 3D-R2N2 algorithm, the SFM-enhanced AKAZE algorithm, the AISI-BIM algorithm, and the improved PMVS algorithm, the proposed algorithm improves the reconstruction effect by 5.3%, 7.8%, 7.4%, and 1.0% respectively. Furthermore, compared to other algorithms, the proposed algorithm exhibits higher efficiency in terms of registration result convergence and reconstruction time, with faster computational speed. This research contributes to the enhancement of building 3D reconstruction technology, laying a foundation for future research in deep learning applications in the architectural field.

Keywords: 3D-R2N2; 3D reconstruction; computer vision; optical building image; urban geographical space

1 INTRODUCTION

With the development of modern technology, 3D reconstruction technology [1] for buildings has received increasing attention, and 3D model construction has become a key element of urban geographic spatial data frameworks. This paper focuses on the optical image-based 3D reconstruction of buildings using 3D-R2N2. Currently, many scholars both domestically and internationally have conducted research in this area and achieved significant results. Wu et al. [2] were the first to propose 3D Shape-Nets, a voxel-based 3D reconstruction network. However, this network suffered from texture defects, specular reflections, and baseline matching issues. Choy et al. [3] proposed the 3D-R2N2 method, which primarily addresses the problem of object feature matching. However, the accuracy and efficiency of this method are not high. Kanazawa et al. [4] developed a Warp-Net network framework based on convolutional neural networks, which achieved reconstruction quality similar to that of supervised methods. However, the targets reconstructed by this method were distorted. Wu J. et al. [5] combined the Marr-Net model, which is trained end-to-end on real images, but this algorithm has issues of computational complexity and lacks finer geometric shapes. Lu Chuan et al. [6] used an improved 2DPCA-SIFT feature-matching algorithm to match feature points in images of ancient buildings from the Qing Dynasty. They then achieved 3D reconstruction through image sequence fusion. This method has good completeness and accuracy, but the reconstruction efficiency is not high. Chen Jiankun et al. [7] proposed a 3D reconstruction method based on deep neural networks, which achieved good results in SAR images, but the 3D reconstruction effect in optical images is not perfect and cannot fully reflect the continuous structure of the target. Stathopoulou E. K. et al. [8] investigate three of the available commonly used open-source solutions, namely COLMAP, OpenMVG + OpenMVS and AliceVision, evaluating their results under diverse large scale scenarios. Comparisons and critical evaluation on the image orientation and dense point cloud generation algorithms is performed with respect to the completeness and accuracy of the final results. Zhu Pan et al. [9] developed an AISI-BIM 3D reconstruction method, which has high accuracy under AISI network segmentation, and the segmentation boundary is also clearer. However, the reconstruction efficiency of this method is not ideal, and it consumes a lot of time and resources. To improve the accuracy of the deep learning algorithm based on a Multi-View Stereo matching network (MVS-Net) in weak texture scenes, Liu D. et al. [10] proposed a novel 2D-3D CNN with spectral-spatial multi-scale feature fusion for hyperspectral image classification, which consists of two feature extraction streams, a feature fusion module as well as a classification scheme. The authors innovated a classification scheme to lift the classification accuracy. Wang A. et al. [11] showed that the point cloud is dense enough which is reconstructed by the 3d reconstruction algorithm based on regional growth combining CMVS-PMVS and well expressed the practical model of object reconstruction; the reconstruction of objects in remote sensing images has very strong practicability, but this algorithm is suitable for a limited number of building types and is not suitable for all types of buildings.

If solely relying on the 3D-R2N2 algorithm for three-dimensional reconstruction in the model, the following limitations on the reconstruction results may be observed: 1) Accuracy, the outcome of the reconstruction is contingent upon the quantity and quality of the input images. As such, if the number of input images is limited or the image quality is subpar, it may lead to lower accuracy in the generated 3D model. 2) Loss of Detail, this algorithm may encounter issues when dealing with the details of buildings. Particularly for complex architectural structures, such as arches, towers, and intricate decorations, it may not be able to accurately reconstruct these details. 3) Scale Issues, as the scale of buildings is generally large, this algorithm may face certain challenges when handling large-scale 3D models. For instance, it may require a substantial amount of computational resources and time, and improper handling might result in a decline in model quality. 4) Data Sparsity, for certain parts of a building that cannot be directly observed from images, such as the interior or obscured sections of a building, this algorithm may not be able to effectively reconstruct them. 5) Realism, the 3D models reconstructed by this algorithm might not achieve an extremely realistic effect. Especially in terms of details such as lighting and textures, it may not be able to achieve the desired outcome.

In this study, an improved version of the 3D-R2N2 algorithm was proposed. We first replaced the CNN module with a convolutional layer that uses densely connected forms. Subsequently, a pyramid pooling layer was inserted between the feature extraction module and the fully connected layer to enhance the model's performance. The experimental results show that the proposed algorithm improves the reconstruction performance by 5.3%, 7.8%, 7.4%, and 1.0%, respectively, compared to the 3D-R2N2 algorithm, the SFM-improved AKAZE algorithm, the AISI-BIM algorithm, and the improved PMVS algorithm. Compared with the 3D-R2N2 algorithm, the SFM-improved AKAZE algorithm, and the AISI-BIM algorithm, the proposed algorithm has higher efficiency and faster computing speed in terms of algorithm registration result convergence and reconstruction time.

2 BASIC PRINCIPLE OF 3D-R2N2 ALGORITHM

2.1 Overall Framework

This study focuses on the research of heterogeneous buildings. The overall process is shown in Fig. 1.

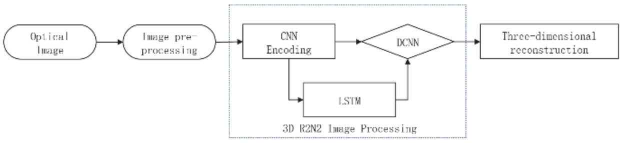


Figure 1 Overall framework

Preprocessed optical remote sensing images are input to CNN [12] for encoding to obtain a low-dimensional feature matrix. To enhance the role of deep neural networks, a residual connection is added between every two convolutional layers in CNN, and 3D-LSTM is used for transitional connection, while conducting recurrent learning [13]. The 3D network structure in 3D-LSTM is formed by arranging each unit of 3D-LSTM, because 3D-LSTM will selectively adjust or maintain the state of each unit, forming a three-dimensional grid structure unit while also accepting a feature vector, which is the result calculated by the encoder, and finally transmitting the data of these feature vectors to DCNN. Finally, each 3D-LSTM hidden unit reconstructs a part of the output voxel, and DCNN convolutional layer takes a 3 × 3 × 3 equal-sized convolution kernel to process various pixel data with building image feature information, and then decodes it to obtain the three-dimensional reconstruction of the building.

2.2 Image Preprocessing

2.2.1 Image Denoising

Due to the presence of noise and blur in optical images [14], direct 3D reconstruction from optical images can lead to inaccurate results. Therefore, preprocessing of optical images is necessary. In this study, a total variation (*TV*) model was used to denoise the images. The principle is to use physical noise natural harmonics to explain the inherent physical regularity of noise images, which is easy to accurately reflect the authenticity and natural characteristics of real noise images from the inherent physical noise of real natural noise images and natural harmonic images, as shown in Eq. (1).

$$f_0(x,y) = f(x,y) + s(x,y) \tag{1}$$

where, *f* represents the original high-frequency image of the noise-free simulation, f_0 represents the clear noisy image contaminated by high-frequency noise, *s* has a zero-mean property, and (*x*, *y*) denotes the pixel position in the image. To eliminate noise, the study uses total variation (*TV*) minimization. The image denoising problem can be formulated as the following minimization problem, as shown in Eq. (2).

$$\min TV(f) = \int_\Omega \sqrt{|\nabla f|^2}\, d_x d_y = \int_\Omega \sqrt{f_x^2 + f_y^2}\, d_x d_y \tag{2}$$

The satisfied constraint conditions are shown as Eq. (3) and Eq. (4).

$$\int_\Omega f d_x d_y = \int_\Omega f_0 d_x d_y \tag{3}$$

$$\frac{1}{|\Omega|} \int_\Omega (f - f_0)^2 d_x d_y = \sigma^2 \tag{4}$$

where, σ^2 represents the variance of noise, δ represents the domain of the target region, Ω represents the domain of the image, $(x,y) \in \Omega$, and the pixel point $(x,y) \in \delta$. The above equation is the system data fidelity term, which can mainly preserve the distortion characteristics of the original image and greatly reduce the distortion degree of the system image noise. The derived equation is shown as Eq. (5).

$$-\nabla \cdot \left(\frac{\nabla f}{|\nabla f|} \right) + \lambda (f - f_0) \tag{5}$$

The equation is the parameter regularization value variation term, where λ is a parameter of regular integers, which plays an important role in balancing noise reduction and noise smoothing in the image.

In this research, the Total Variation model (*TV*) is applied to denoise the optical images of anisotropic buildings, effectively extracting the authenticity and natural noise characteristics of the images. This approach significantly enhances the accuracy of the 3D reconstruction results.

2.2.2 Image Enhancement

The optical image was enhanced using the histogram equalization method [15]. The input optical image was transformed into a histogram image and then grayscale.

The method extended the grayscale range of the specific comparison to the entire range of all grayscale regions, achieving non-uniform extension stretching of the input image. Then, the pixels in the input image were redistributed, as shown in Eq. (6).

$$p\left(k\right)=\frac{m_k}{m}\ k=0,1,2\ldots\ldots L-1 \tag{6}$$

where, $L-1=255$, k denotes an integer value that marks the number of object pixels of the given image in a certain grayscale, m is the total number of pixels in the image, and $p(k)$ represents the frequency.

The cumulative histogram is calculated as shown in Eq. (7).

$$P_k=\sum_{i=0}^{k}\frac{n_i}{m}=\sum_{i=0}^{k}p\left(i\right)\left(k=0,1,\ldots\ldots L-1\right) \tag{7}$$

where, n_i is the number of pixels corresponding to the i gray level, and P_k represents the sum of frequencies obtained from the gray level i in [0, $L-1$]. Then, P_k is extended by rounding, as shown in Eq. (8).

$$P_k=\text{int}\left[\left(L-1\right)P_k+0.5\right] \tag{8}$$

The mapping relationship is as follows: $k\rightarrow P_k$, where k represents the integer value labeled by the number of object pixels in the given image, and P_k represents the sum of frequencies obtained for the gray level i in [0, $L-1$].

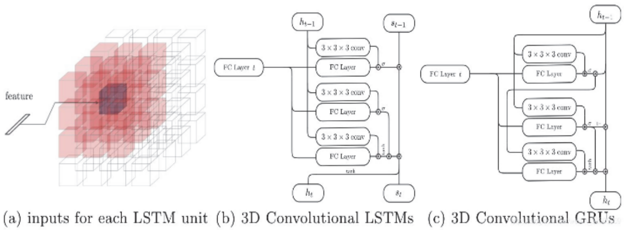
In this study, the histogram equalization stretching method was used to enhance the contrast and details of the optical anisotropic building images. This method improves the visual effects and readability of the images, aiding in more accurate execution of image encoding, feature extraction, and 3D reconstruction. By enhancing the less discernible details in dark or bright areas, the accuracy and quality of the reconstruction are significantly improved.

2.3 Build 3D-R2N2 Network

Voxel refers mainly to the probability distribution of three-dimensional objects represented as three-dimensional binary variables under the action of deep convolutional belief networks [16]. The obtained probabilities are input into a depth map and the data is continuously predicted and filled using Gibbs sampling [17] to reconstruct the shape voxel of the building target. To improve the effect of building reconstruction, this study uses the voxel-based 3D-R2N2 algorithm for 3D reconstruction. The combination of voxel and the 3D-R2N2 algorithm makes the 3D reconstruction result more accurate and solves problems such as texture defects and wide baseline feature matching.

Firstly, random sampling is performed on preprocessed images, and the sampled images are extracted and encoded with different 2D-CNN standards, including feedforward CNN and deep residual variation. To match the number of channels after convolution, this study uses 1 × 1 convolution for residual connections. Next, the output of the encoder is unfolded and fed into a fully connected layer, which maps the output to a 1024-dimensional feature vector, i.e., a low-dimensional feature vector.

The obtained low-dimensional feature vector is inputted into 3D-LSTM units, which are arranged in a 3D grid structure. In the 3D grid, there are $N\times N\times N$ 3D-LSTM units, where N is the spatial resolution of the 3D-LSTM grid.



(a) inputs for each LSTM unit (b) 3D Convolutional LSTMs (c) 3D Convolutional GRUs

Figure 2 LSTM specific structure

As shown in Fig. 2, each 3D-LSTM unit receives a feature vector of the image after encoding and the hidden state of the previous unit. The index (i, j, k) has an independent hidden state h_t, and $\left(i,j,k\right)\in R^{N_h}$. The output gate f_t, input gate i_t, and storage unit s_t control the 3D-LSTM grid according to Eq. (9), Eq. (10), Eq. (11), and Eq. (12).

$$\sigma\left(W_fT\left(x_t\right)+U_f*h_{t-1}+b_f\right) \tag{9}$$

$$i_t=\sigma\left(W_fT\left(x_t\right)+U_f*h_{t-1}+b_i\right) \tag{10}$$

$$s_t=f_t\odot s_{t-1}+i_t\tanh\left(W_sT\left(x_t\right)+U_s*h_{t-1}+b_s\right) \tag{11}$$

$$h_t=\tanh\left(s_t\right) \tag{12}$$

Represents convolutional operation. We set $n=4$ in our study. A notable characteristic of our study is the absence of an output gate f_t, as we only output the final result. To reduce the number of parameters, we removed the redundant output gate, which enables the 3D-LSTM unit to handle inconsistencies between the reconstruction region and the true model, allowing each unit to focus on learning a part of the voxel space rather than participating in the entire space reconstruction. This structure gives the network a locality, allowing it to selectively update predictions of previously occluded parts of the object. The specific portion of the reconstructed final output is calculated and input into the decoder.

After inputting the image sequence $x_1,x_2,\cdots,x_T$, the 3D-LSTM transfers the hidden state features to the decoder, which uses 3D convolutions, nonlinearity, and 3D transposed convolution to increase the resolution of the hidden state until the target output resolution is achieved. As shown in Fig. 3, similar to the encoder, this study designs a decoding network, consisting of 5 convolutions and a deep residual (with 4 residual connections). After the final activation before reaching the target output resolution in the last layer, voxel-wise-softmax [18] is applied to the final activation, as shown in Eq. (13).

$$V \in R^{N_{\text{vox}} \times N_{\text{vox}} \times N_{\text{vox}} \times 2} \tag{13}$$

The voxel cross-entropy loss function is used to calculate the probability of each voxel. X represents the input image sequence, $y(i,j,k)$ represents the actual ground truth, and $p(i,j,k)$ represents the predicted probability of the voxel.

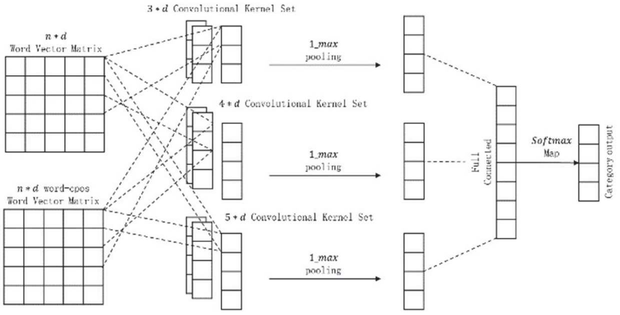


Figure 3 Decoding network

Let the final output at each voxel (i,j,k) follow a Bernoulli distribution $[1-p(i,j,k)$, $p(i,j,k)]$, where the dependency on the input $X=\{Xt\}, t\in\{1,...,t\}$ is omitted. Let the corresponding ground truth occupancy be $y(i,j,k)\in\{0,1\}$, as shown in Eq. (14).

$$L(X,y)=\sum_{i,j,k} y_{(i,j,k)}\log\left(p_{(i,j,k)}\right)+ \left(1-y_{(i,j,k)}\right)\log\left(1-p_{(i,j,k)}\right) \tag{14}$$

This study adopts Intersection over Union (*IoU*) as the evaluation metric, which is larger the better, as shown in Eq. (15).

$$IoU=\sum_{i,j,k}\left[I\left(p_{(i,j,k)}>t\right)I\left(y_{(i,j,k)}\right)\right]/ \sum_{i,j,k}\left[\left(I\left(p_{(i,j,k)}>t\right)+I\left(y_{(i,j,k)}\right)\right)\right] \tag{15}$$

where, *I*(*x*) is the indicator function and *t* is the threshold. If the probability is greater than this threshold, then the corresponding voxel exists.

2.4 Improved 3D-R2N2 Model

This study mainly focuses on improving the CNN module in 3D-R2N2. The linearly connected convolutional layers used in the original 3D-R2N2 CNN are replaced with convolutional layers using dense connections. In addition, a pyramid pooling layer is added between the CNN feature extraction module and the fully connected layer. Compared to the old model, the new model is faster and more stable during training, with a faster decline in cross-entropy loss function, and the reconstructed 3D models are more accurate.

2.4.1 Compact Module

The first structural module in 3D-R2N2, Encoder, uses a CNN to extract and encode image features, which includes 12 convolutional layers and 5 residual connections. Although the convolutional part is aided by 5 residual connections, the feature extraction ability of the entire Encoder is still limited. In this paper, while retaining the advantages of 3D-R2N2 as much as possible, we improve and optimize the model's structure and algorithm. The specific improvements are as follows:

The original CNN structure in the Encoder module was modified by replacing the conventional connection in the convolutional layer with densely connected (DC) connections [19]. This modification not only allows training of a deeper network without increasing the training difficulty, but also enhances the feature extraction ability of the Encoder while making the network as a whole more easily trainable. The content of the network's recurrent learning is the feature, and the more features obtained, the more accurate the reconstructed 3D model will be.

The main focus of this study is to add two dense modules to the CNN structure. The CNN structure before and after the improvement is shown in Fig. 4 and Fig. 5, respectively.

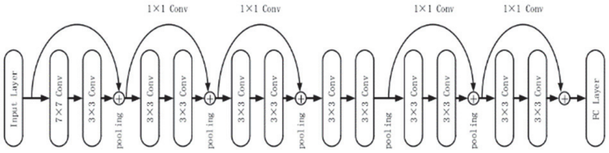


Figure 4 CNN module of 3D-R2N2 before improvement

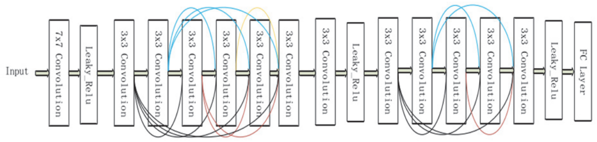


Figure 5 CNN module of improved 3D-R2N2

The traditional convolutional layers with regular links in CNN were replaced by convolutional layers with dense links. From the improved Fig. 5, it can be seen that the improved network added two modules, each containing 6 and 5, 3 × 3 convolutional layers, respectively. In these convolutional layers, each layer is connected to all subsequent layers. The layers after the improved convolutional layers can obtain feature information extracted from previous layers, and each layer input can be represented as x_n, as shown in Eq. (16).

$$x_n = Y_n\left(\left[x, x_1 \cdots x_{n-1}\right]\right) \tag{16}$$

where, Y_n represents the transportation method, and $\left[x_0, x_1 \cdots x_{n-1}\right]$ represents the set of extracted feature information. Each convolutional layer contains all the feature information extracted from the previous layers, so each layer requires very few feature maps, unlike other networks that have a large width. At the same time, the transfer of features and gradients is more efficient under this dense connection scheme, making training the improved network easier.

In the improved densely connected network, there is also a parameter growth rate *v*, which controls the number of feature extractions in the network. This allows for more efficient use of the channel number, making the CNN network more efficient and improving encoding efficiency.

2.4.2 Feature Classification Module

The conventional convolutional neural network [20] typically follows the feature extraction module with multiple fully connected layers, which output the classification results. The improvements made in this study are as follows:

Inserting a spatial pyramid pooling (SPP) [21] layer between the feature extraction module and the fully connected layer can solve the problem of varying input image sizes. The SPP layer, as shown in Fig. 6, extracts multi-scale features and improves classification performance by fusing these features with the feature classification module.

As shown in Fig. 6, the principle of the SPP method is to process from the input layer to the output layer of the network. Taking the feature map group with a depth of 256 as an example, the operation process of the SPP layer is described, where the height and width of the feature map group are not fixed. In the SPP layer, the leftmost blue box divides each channel of the feature map group into 16 parts, and the feature map group is divided into 16×256 parts in this way. The green box and the right purple box are similar, and they divide the feature map group into 4 × 256 parts and 1 × 256 parts, respectively. After pooling operation is performed on each part, the feature map group is transformed into a 256 × 21-dimensional matrix, which is then flattened into a feature vector of length 5376 before entering the fully connected layer. This ensures that regardless of the dimension of the input data, the data will be unified to the same dimension after passing through the SPP layer before entering the fully connected layer, solving the problem of inconsistent sizes of input feature maps caused by saliency segmentation and cropping in sonar image feature extraction module.

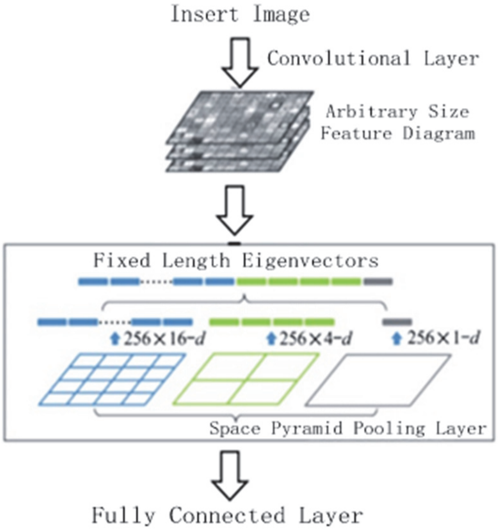

Insert Image
Convolutional Layer
Arbitrary Size Feature Diagram
Fixed Length Eigenvectors
256×16-d
256×4-d
256×1-d
Space Pyramid Pooling Layer
Fully Connected Layer

Figure 6 Pyramid pooling layer (SPP)

In addition, unlike regular pooling layers, spatial pyramid pooling (SPP) uses multiple sampling scales. The partitioning of SPP can be flexibly set according to the application requirements. For example, as shown in Fig. 6, a 3-level pyramid is used, with sampling windows of size 4 × 4, 2 × 2, and 1 × 1, respectively. By down-sampling the feature map with multiple window sizes and strides, a set of new feature maps with different scales is obtained. These feature maps are then concatenated to form a new vector that contains both high-level semantic and spatial information. By inputting this feature vector to a fully connected layer, the accuracy of target recognition can be improved.

3 EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Experimental Environment

The experimental environment of this study was Windows 10 operating system, with 32 GB memory, and the training was performed using a Ge Force RTX 3090 SUPRIM X 24G graphics card. The training framework used was TensorFlow.

After data preprocessing, a total of 100 000 optical images were used in the comparative experiments in this study. The dataset consists of 20 different categories, among which 12 were selected as the training samples and 8 were used as the testing samples for this research.

3.2 Evaluation Metrics

In this study, the *IoU* value [22], i.e., the accuracy of 3D reconstruction, was used as the evaluation criterion for experiments 3.3 and 3.6. The evaluation was mainly based on the cross-overlap ratio of the 3D reconstruction result and the true model. The ratio of the intersection and union between the measured target volume and the true volume was measured. Therefore, when dealing with other *FFr* representations based on surface reconstruction, voxelization of the reconstructed and true models is required first, as shown in Eq. (17).

$$IoU = \frac{V^{\text{pred}} \cap V^{\text{gt}}}{V^{\text{pred}} \cup V^{\text{gt}}} = \frac{\sum_i \left\{ I\left(V_i^{\text{pred}}\right) \cdot I\left(V_i^{\text{gt}}\right) \right\}}{\sum_i \left\{ I\left(V_i^{\text{pred}}\right) + I\left(V_i^{\text{gt}}\right) - I\left(V_i^{\text{pred}}\right) \cdot I\left(V_i^{\text{gt}}\right) \right\}} \tag{17}$$

where, $I(\cdot)$ denotes the indicator function, V_i^{pred} represents the *i* voxel of the predicted model, V_i^{gt} represents the *i* voxel of the ground truth model. The higher the *IoU* value, the better the reconstruction accuracy.

The reconstruction results of the target object may vary with different network structures, and the improved Encoder network structure may also have differences in the reconstruction results compared with the original Encoder. In this experiment, the proposed algorithm improved the Encoder module by adding Inception modules while preserving the residual convolutional network. After the optical images were segmented to a certain extent, in order to solve the problem of gradient vanishing, residual connections in Resnet were used to extract more accurate target image features.

The evaluation criterion for experiment 3.4 is the convergence of the matching results, which is calculated using the method shown in Eq. (18).

$$R_1 = \lim_{z \to \infty} sup\left(a^{2z} + b^{2z}\right)^{\frac{1}{2z}} \tag{18}$$

R_1 represents the convergence factor, and $\left(a^{2z} + b^{2z}\right)^{\frac{1}{2z}}$ represents the vector matrix. The closer R_1 is to 0, the better the convergence of the algorithm, and vice versa.

3.3 Comparison of the Same Image Reconstruction Accuracy of Different Algorithms

To evaluate the 3D reconstruction accuracy of the proposed algorithm, this experiment compared it with the 3D-R2N2 algorithm, SFM-improved AKAZE algorithm, AISI-BIM algorithm, and improved PMVS algorithm on the same type of optical images, and calculated the corresponding IoU values for different algorithms, as shown in Tab. 1.

As shown in Tab. 1, the proposed algorithm has higher reconstruction accuracy (*IoU*) than other algorithms in different types of buildings. Compared with the 3D-R2N2 algorithm, the proposed algorithm has an increase of 5.3% in *IoU*, an increase of 7.8% in IoU compared with the SFM-improved AKAZE algorithm, an increase of 7.9% in *IoU* compared with the AISI-BIM algorithm, and an increase of 1.0% in IoU compared with the improved PMVS algorithm. Although the proposed algorithm has not shown a significant improvement in reconstruction accuracy compared with the improved PMVS algorithm, the experiment shows that the improved PMVS algorithm has certain limitations in applicable building types, as it lacks data for buildings in Guting, adobe houses, and ancient town buildings.

The IoU values corresponding to (b), (c), (d), (e), and (f) in Fig. 7 are 0.671, 0.650, 0.657, 0.712, and 0.729, respectively, when specific building images are selected. As higher IoU values indicate higher accuracy, it can be seen from Fig. 7 that the proposed algorithm is more complete in 3D reconstruction details.

Table 1 IoU comparison of 3D-R2N2, SFM-improved AKAZE, AISI-BIM, improved PMVS and our proposed method for different objects

Category	3D-R2N2	SFM-improved AKAZE	AISI-BIM	Improved PMVS	Algorithm in this paper
Guting	0.511	0.494	0.512	-	0.552
Workshop	0.731	0.653	0.621	0.701	0.821
Adobe house	0.523	0.511	0.510	-	0.564
Bell tower	0.642	0.649	0.643	0.645	0.658
Skyscraper	0.668	0.696	0.701	0.660	0.730
Ancient town buildings	0.734	0.659	0.655	-	0.806
Mean	0.635	0.610	0.607	0.668	0.688

In the table, "-" represents missing data.

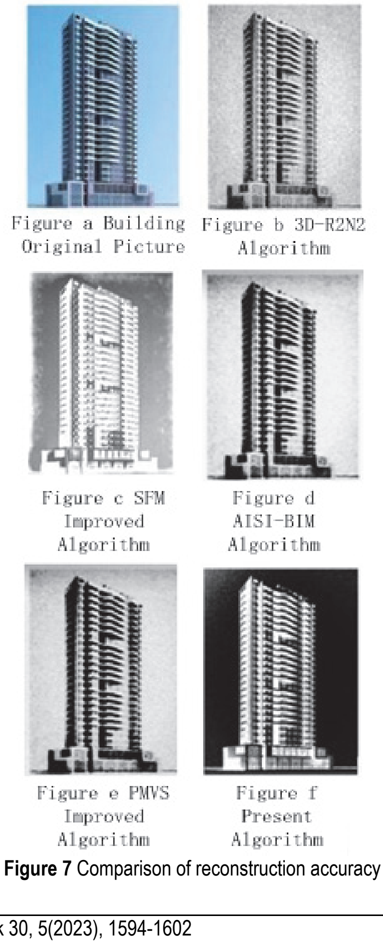


Figure a Building Original Picture

Figure b 3D-R2N2 Algorithm

Figure c SFM Improved Algorithm

Figure d AISI-BIM Algorithm

Figure e PMVS Improved Algorithm

Figure f Present Algorithm

Figure 7 Comparison of reconstruction accuracy

3.4 Comparison of Convergence of Registration Results for Different Algorithms on the Same Dataset

Convergence is one of the methods to test the stability of an algorithm. In this experiment, convergence tests of registration results were conducted on 3D-R2N2, SFM-improved AKAZE, AISI-BIM algorithm, improved PMVS algorithm, and the proposed algorithm using the same dataset to evaluate the stability of different algorithms. The results are shown in Tab. 2.

Table 2 Convergence comparison of registration results

Optical image data scale (number of images)	Convergence of registration results				
	3D-R2N2	SFM-improved AKAZE	AISI-BIM	Improved PMVS	Algorithm in this paper
4000	0.00639	0.00560	0.00512	0.00554	0.00501
17311	0.00382	0.00233	0.00380	0.00312	0.00212
37780	0.00378	0.00230	0.00370	0.00320	0.00209

According to Tab. 2, it can be observed that the convergence of the proposed algorithm is stronger than other algorithms in optical image datasets of 4000, 17311, and 37780 images, respectively. When performing 3D reconstruction on three different datasets, the proposed algorithm is compared with 3D-R2N2 algorithm, SFM-improved AKAZE algorithm, AISI-BIM algorithm, and improved PMVS algorithm. The results show that the proposed algorithm can effectively improve the convergence of algorithm registration results by 1.6%, 3.2%, 2.0%, and 2.6%.

3.5 Comparison of the Same Image Reconstruction Effect and Convergence Time of Different Algorithms

The qualitative analysis method was used in this experiment to evaluate the 3D building reconstruction effect of the proposed algorithm. A teaching building with various architectural features, such as a clock tower and multi-story buildings, was chosen as the experimental object. The 3D-R2N2 algorithm, SFM-improved AKAZE algorithm, AISI-BIM algorithm, improved PMVS algorithm, and the proposed algorithm were used in sequence for 3D reconstruction of the building images, and the results are shown in Figs. 8 to 12. As can be seen from Figs. 8 to 12, the 3D reconstruction effect obtained using the proposed algorithm has clearer outlines and better reconstruction details compared to other algorithms.

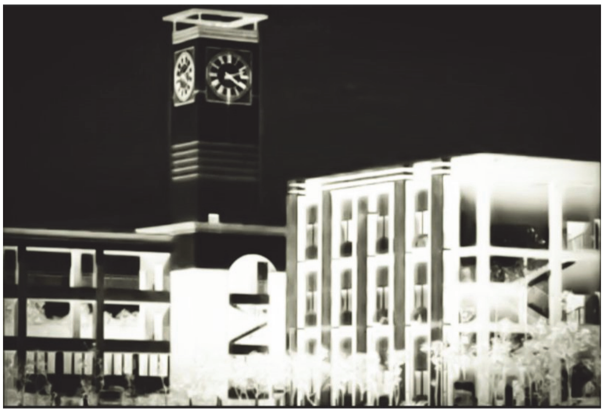


Figure 8 3D-R2N2 algorithm

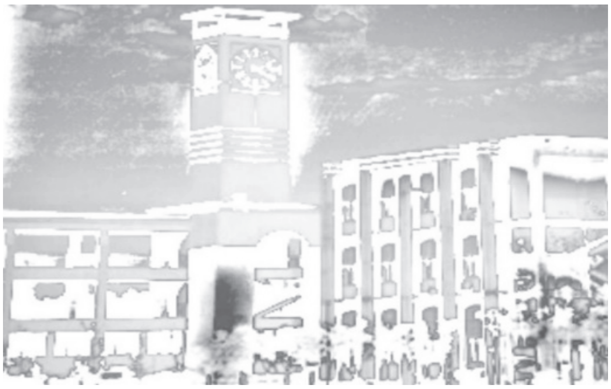


Figure 9 SFM-improved AKAZE

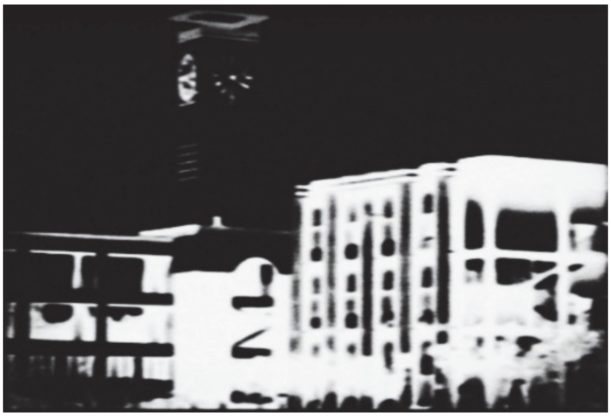


Figure 10 AISI-BIM algorithm

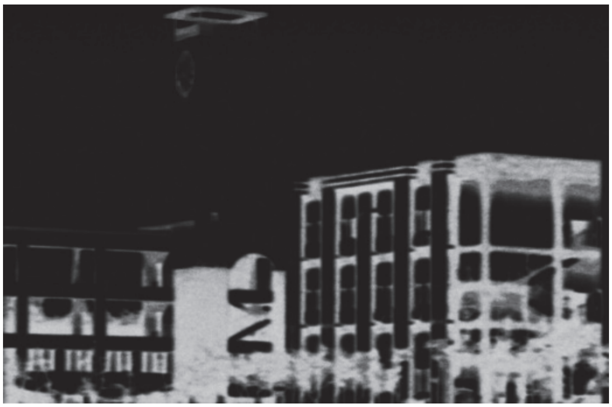


Figure 11 Improved PMVS algorithm

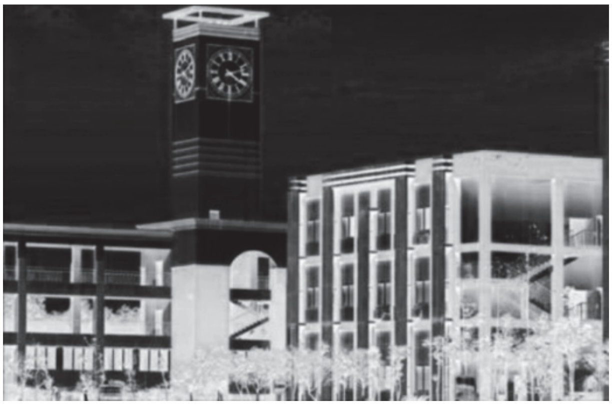


Figure 12 Algorithm in this study

At the same time, the experiment conducted a comparative study on the convergence time of optical images of different data sizes, as shown in Tab. 3.

Table 3 Convergence time of different image data scales

Optical image data scale (number of images)	Convergence time /s				
	3D-R2N2	SFM-improved AKAZE	AISI-BIM	improved PMVS	Algorithm in this paper
4000	2.001	2.588	2.501	1.986	1.351
17311	10.188	12.222	7.088	8.892	6.082
37780	28.478	33.511	25.125	22.345	20.193
66700	79.934	100.441	70.450	80.886	60.432

According to Tab. 3, the convergence time of the five algorithms is directly proportional to the data scale, that is, the larger the data scale, the longer the convergence time. Among the algorithms compared in the same data scale, the convergence time of the 3D-R2N2 algorithm, SFM-improved AKAZE algorithm, AISI-BIM algorithm, and improved PMVS algorithm were all higher than that of this study algorithm. The proposed algorithm reduces the convergence time by an average of 8.1%, 15.1%, 4.2%, and 6.5%, respectively, compared to other algorithms.

3.6 Evaluation Metrics

The ablation experiment is one of the methods used to demonstrate the effectiveness of the improved algorithm, as shown in Tab. 4. DC represents the improved dense connection module in CNN, and SPP represents the improved feature classification module.

The experimental results show that when the DC module is applied to improve 3D-R2N2, the reconstruction accuracy increases by 11.1% compared to 3D-R2N2; when the SPP module is applied to improve 3D-R2N2, the reconstruction accuracy increases by 11.39%; when both DC and SPP modules are applied to improve 3D-R2N2, i.e., the proposed algorithm in this study, the reconstruction accuracy is increased by 17.81% compared to 3D-R2N2. The results demonstrate that the proposed algorithm has a significant advantage in reconstruction accuracy.

Table 4 Performance comparison results of 3D-R2N2 algorithm before and after improvement

Algorithm	*IoU*
3D-R2N2	0.6735
3D-R2N2 fusion with DC	0.7845
3D-R2N2 fusion with SPP	0.7874
Algorithm in this paper	0.8516

4 CONCLUSION

This study is based on the 3D-R2N2 algorithm for three-dimensional reconstruction of optical building images. The principle of the 3D-R2N2 algorithm was analyzed, and the algorithm was improved and compared with traditional 3D reconstruction algorithms. The experimental results show that the improved 3D-R2N2 algorithm has significant advantages over traditional 3D reconstruction algorithms in terms of reconstruction accuracy, algorithm registration result convergence, reconstruction effect, and convergence time. The efficacy of the algorithm itself was verified through ablation experiments. Furthermore, the results of this study can be applied not only to optical building images but also to other types of complex planar stereoscopic optical images for three-dimensional reconstruction. However, there are still many aspects of this study that need improvement, such as the poor performance of the algorithm on curved surface and irregular edge buildings, which are the future research directions and goals.

Acknowledgments

Funding project: 2022 Sichuan Province Science and Technology Innovation Project in the field of housing and urban-rural construction, "Research and Application of Intelligent Detection of Underground Space Leakage Based on Computer Vision" (NO: SCJSKJ2022-05).

5 REFERENCES

[1] Wang, G., Qin, X., Han, D. et al. (2021). Study on seepage and deformation characteristics of coal microstructure by 3D reconstruction of CT images at high temperatures. *International Journal of Mining Science and Technology*, *31*(2), 175-185. https://doi.org/10.1016/j.ijmst.2020.11.003

[2] Wu, Z., Song, S., Khosla, A. et al. (2015). 3d shape nets: A dee prepresentation for volumetric shapes. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1912-1920.

[3] Choy, C. B., Xu, D., Gwak, J. Y. et al. (2016). 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction. *European Conference on Computer Vision*, 628-644. https://doi.org/10.1007/978-3-319-46484-8_38

[4] Kanazawa, A., Jacobs, D. W., & Chandraker, M. (2016). WarpNet: Weakly Supervised Matching for Single-View Reconstruction. *Proceedings of the IEEEConference on Computer Vision and Pattern Recognition*, 3253-3261. https://doi.org/10.1109/CVPR.2016.354

[5] Wu, J., Wang, Y., Xue, T. et al. (2017). MarrNet: 3D Shape Reconstruction via 2.5D Sketches. *Advances in neural information processing systems*, 540-550.

[6] Lu, C., & Su, J. (2021). Algorithm of 3D Virtual Reconstruction of Ancient Buildings in Qing Dynasty Based on Image Sequence. *Security and Communication Networks*, 8388480. https://doi.org/10.1155/2021/8388480

[7] Nguyen, H., Wang, Y., & Wang, Z. (2020). Single-shot 3D shape reconstruction using structured lightand deep convolutional neural networks. *Sensors*, *20*(13), 3718. https://doi.org/10.3390/s20133718

[8] Stathopoulou, E. K., Welponer, M., & Remondino, F. (2019). Open-source image-based 3D reconstruction pipelines: Review, comparison and evaluation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLII-2/W17*, 331-338. https://doi.org/10.5194/isprs-archives-XLII-2-W17-331-2019

[9] Zhu, P. & Si, J. Y. (2020). Research on BIM 3D reconstruction method based on AISI network. *Journal of Graphics*, *41*(05), 839-846.

[10] Liu, D., Han, G., Liu, P. et al. (2021). A Novel 2D-3D CNN with Spectral-Spatial Multi-Scale Feature Fusion for Hyperspectral Image Classification. *Remote Sensing*, *13*(22), 4621. https://doi.org/10.3390/rs13224621

[11] Wang, A., An, N., Zhao, Y. et al. (2016). 3D Reconstruction of Remote Sensing Image Using Region Growing Combining with CMVS-PMVS. *International Journal of Multimedia and Ubiquitous Engineering*, *11*(8), 29-36. https://doi.org/10.14257/ijmue.2016.11.8.03

[12] Mzoughi, H., Njeh, I., Wali, A. et al. (2020). Deep multi-scale 3D convolutional neural network (CNN) for MRI gliomas brain tumor classification. *Journal of Digital Imaging*, *33*, 903-915. https://doi.org/10.1007/s10278-020-00347-9

[13] Wang, Y., Jiang, L., Yang, M. H. et al. (2019). Eidetic 3d lstm: A model for video prediction and beyond. *International conference on learning representations.*

[14] Yang, X., Zhao, J., Wei, Z. et al. (2022). SAR-to-optical image translation based on improved CGAN. *Pattern Recognition*, *121*, 108208. https://doi.org/10.1016/j.patcog.2021.108208

[15] Pour, A. M., Seyedarabi, H., Jahromi, S. H. A. et al. (2020). Automatic detection and monitoring of diabetic retinopathy using efficient convolutional neural networks and contrast limited adaptive histogram equalization. *IEEE Access*, *8*, 136668-136673. https://doi.org/10.1109/ACCESS.2020.3005044

[16] Cai, W., Liu, D., Ning, X. et al. (2021). Voxel-based three-view hybrid parallel network for 3D object classification. *Displays*, *69*, 102076. https://doi.org/10.1016/j.displa.2021.102076

[17] Ludwig, M., Nothias, L. F., & Dührkop, K. et al. (2020). Database-independent molecular formula annotation using Gibbs sampling through ZODIAC. *Nature Machine Intelligence*, *2*(10), 629-641. https://doi.org/10.1038/s42256-020-00234-6

[18] Liu, M., Li, F., Yan, H. et al. (2020). A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. *Neuroimage*, *208*, 116459. https://doi.org/10.1016/j.neuroimage.2019.116459

[19] Zhang, X., Xu, H., Mo, H. et al. (2021). Dcnas: Densely connected neural architecture search for semantic image segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13956-13967. https://doi.org/10.1109/CVPR46437.2021.01374

[20] Schwendicke, F., Golla, T., Dreher, M. et al. (2019). Convolutional neural networks for dental image diagnostics: A scoping review. *Journal of dentistry*, *91*, 103226. https://doi.org/10.1016/j.jdent.2019.103226

[21] Huang, Z., Wang, J., Fu, X. et al. (2020). DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object detection. *Information Sciences*, *522*, 241-258. https://doi.org/10.1016/j.ins.2020.02.067

[22] Yan, J., Wang, H., Yan, M. et al. (2019). IoU-adaptive deformable R-CNN: Make full use of IoU for multi-class object detection in remote sensing imagery. *Remote Sensing*, *11*(3), 286. https://doi.org/10.3390/rs11030286

Contact information:

Qianying ZOU, PhD, Professor
(Corresponding author)
Geely University of China,
No. 123, Section 2, Chengjian Avenue, East New District,
Chengdu, Sichuan Province
E-mail: zqy_bb@163.com

Fengyu LIU, Teaching Assistant
Geely University of China,
No. 123, Section 2, Chengjian Avenue, East New District,
Chengdu, Sichuan Province
E-mail: i@liufengyu.cn