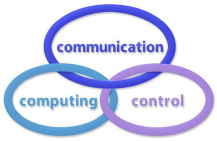
INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 19, Issue: 2, Month: April, Year: 2024
Article Number: 6234, https://doi.org/10.15837/ijccc.2024.2.6234



CCC Publications

Enhancing Automated Loading and Unloading of Ship Unloaders through Dynamic 3D Coordinate System with Deep Learning

L.F. Wang, Q. Li, W. Fu, F. Jiang, T. X. Song, G. B. Pi, S. J. Sun*

Lufeng Wang
Department of Artificial Intelligence and Big Data
Chongqing Industry Polytechnic College
Chongqing, China, 401120
wanglf@cqipc.edu.cn

Qu Li
Chongqing Tiancheng Digital
Technology Co., Ltd.
Chongqing, China, 401120
liq@tianchengdt.cn

Wei Fu
Department of Artificial Intelligence and Big Data
Chongqing Industry Polytechnic College
Chongqing, China, 401120
fuwei@cqipc.edu.cn

Fei Jiang
Chongqing Tiancheng Digital
Technology Co., Ltd.
Chongqing, China, 401120
ynthjf@163.com

Tianxing Song
Chongqing Tiancheng Digital
Technology Co., Ltd.
Chongqing, China, 401120
songtx0713@sina.com

Guangbo Pi
Chongqing Tiancheng Digital
Technology Co., Ltd.
Chongqing, China, 401120
pigb@tianchengdt.cn

Shijie Sun*
Chongqing Tiancheng Digital
Technology Co., Ltd.
Chongqing, China, 401120
*Corresponding author: monkey-114@163.com

Abstract

This paper proposes a deep learning approach for accurate pose estimation in ship unloaders, improving grasping accuracy by reconstructing 3D coordinates. A convolutional neural network optimizes depth map prediction from RGB images, further enhanced by a conditional generative adversarial network to refine quality. Evaluation of simulated ship unloading tasks showed over 90% grasping success rate, outperforming baseline methods. This research offers valuable insights into advanced visual perception and deep learning for next-generation automated cargo handling.

Keywords: 3D coordinates, Convolutional, GAN, Codec, Loading and unloading.

1 Introduction

A ship unloader is a specialized machinery that uses continuous conveying machinery to lift loose materials and transport them to the rack area [1]. Among them, the accuracy of robotic arm grasping

and placement has a significant impact on the equipment. Although this is only a very simple action in the human body, it is more complex in machine simulation, which can be roughly divided into target perception, pose estimation, and trajectory planning. Pose estimation is the key point in the entire process, which first requires perception through the camera. Traditional image technology has high limitations on the structural environment and complexity of actions, but deep learning algorithms have made breakthroughs in this field, enabling machines to have higher autonomous sensing capabilities. This is a technology that achieves new object recognition capabilities through the continuous learning of multi-layer network neurons. Its excellent fitting ability helps to complete more efficient work [2]. Visual perception optimization can simulate the human eye, reconstruct three-dimensional coordinates, and assist machines in locating target objects [3]. Attitude estimation mainly refers to algorithms to determine the orientation and position of an object in 3D space. In recent years, the development of deep learning techniques has greatly contributed to the advancement of this field. However, problems include reduced accuracy when dealing with complex backgrounds and occlusion situations, and high computational load when processing high-resolution images in real-time. The method of applying this technology to robotic arms is not yet mature, and the progress made is insufficient. Therefore, a deep learning-based dynamic construction method for 3D coordinate systems is proposed, and the feasibility of convolutional neural network (CNN) in estimating image depth information is analyzed. This basic analysis deepens the theoretical understanding of the application of deep learning in 3D visual reconstruction. To address the issue of low feature utilization in traditional encoder-decoder networks, an innovative multi-level integrated encoder-decoder network is proposed. This network improves depth estimation accuracy by efficiently fusing and utilizing features. The inclusion of a residual upper projection module enhances high-level semantic information decoding, resulting in improved depth prediction accuracy. Additionally, the application of conditional generative adversarial network (cGAN) to single image depth estimation introduces a fresh perspective on depth map generation. The overall process involves introducing the concept of depth camera, designing 3D coordinate reconstruction, and employing a multilayer integrated encoder-decoder network. By optimizing the generator and discriminator of the pix2pix network, incorporating jumping convolution down-sampling module and pyramid matching network, and enforcing conditional constraints, the model achieves enhanced performance. The study is divided into four parts. The first part studies the current situation of ship unloaders and 3D coordinate reconstruction. The second part designs a deep learning based dynamic construction method for 3D coordinate systems. The third part verifies the proposed method through experiments, and the fourth part summarizes the experimental results.

2 Related works

Currently, various types of machines are tending towards automation, and ship unloaders are no exception. Essentially, they are autonomous control of robotic arms. Li L et al. proposed a point cloud pose estimation method based on the dual hypothesis RPM-Net for grasping metal workpieces. This method utilized PCL to segment workpiece point clouds from the scene and used a trained network for the dual pseudo point cloud registration and pose estimation. In the experiment, a six-axis industrial robot and a binocular structured light sensor were used to establish an experimental platform, and the success rates of three real datasets were tested. The results showed that this method accurately estimated the pose of discrete and less textured metal workpieces. At the same time, this study also proposed a method for generating training datasets using the PCL visualization algorithm [4]. Kushwaha V et al. used a Pix 2 Pix cGAN to generate grasping poses/rectangles for intelligent robot grasping tasks. This network took object images as input and generated grab rectangles labeled with objects as output. By extracting the posture of the generated grasping rectangle as the centroid of the object and using a limited number of Cornell grasping datasets for training, the accuracy of grasping the rectangle was significantly improved. The experimental results showed that the generative model-based method had good potential in successfully grasping and processing invisible objects [5]. Shengqian LI et al. proposed a hand-eye calibration method based on ROS, which precisely controlled the rotation of the robot arm at different positions through the ROS system and used the robot head camera to capture images of fixed points. Then, a nonlinear equation system was established based

on the image and pose information of the ROS system, and the hand-eye relationship was optimized using the least squares method. The experimental results validated the effectiveness of this method [6]. Higo R et al. proposed a high-speed hand grip strategy on a two-dimensional plane, using non-contact fingers and a high-speed visual system. The re-clamping strategy included three stages: rotation, release, and capture, and utilized high-speed cameras for visual feedback. The experimental results showed that in 60 re-grasping experiments on three cubes at different initial grasping positions, the success rate reached 100%, and each re-grasping time was less than 0.2 seconds [7].

The depth camera is a part of the loading and unloading system that realizes pose detection and plays a crucial role in the loading and unloading accuracy of the ship unloader. Jin X et al. designed a prototype of a low-loss transplanting robot for hole seedlings based on machine vision and proposed combining edge recognition technology with the transplanting robot to achieve path planning. Through experiments, it was confirmed that compared with the conventional transplantation group, the machine vision group had an 11.11% reduction in injury rate and an increase in transplantation time of 0.029 seconds; The injury rate increased by 0.46% and the transplantation time decreased by 0.238 seconds [8]. He G et al. believed that with the increase in urban cable tunnel construction, unmanned aerial vehicle (UAV) patrol cable tunnels became the preferred choice. A UAV positioning and navigation method that integrated ultra-wideband positioning and depth cameras was used for comparative experiments with optical flow methods. The results indicated that the ultra-wideband positioning method had a smaller yaw angle, with a minimum of only 0.16°, providing an important reference for UAV patrols in cable tunnels [9]. He B et al. conducted a lightweight design on YOLOv3 for fruit recognition and localization in complex environments. By improving the T-Net algorithm, the position and diameter of mature tomatoes were obtained in the robot operating system. The experimental results showed that T-Net achieved an average accuracy of 99.2% and an F1 score of 98.9% at a detection rate of 104.2 FPS. Compared to YOLOv3, T-Net had better performance and a detection speed of 1.8 times faster. This model provided an effective method for real-time detection and localization of tomatoes [10]. Zhou C et al. proposed using a depth camera as a detection device to address the issue of human position detection in mechanical safety. Real-time human position detection was achieved by collecting depth camera image information for human presence detection and distance measurement. Using a detection method based on an Intel RealSense depth camera and MobileNet-SSD algorithm, the results showed that depth cameras replaced traditional devices, achieved accurate and reliable human position detection, and distinguished between human and mobile devices, improving security [11, 12].

It can be seen that depth cameras have a wide range of applications and are a research hotspot in the academic community. However, automatic ship unloader faces many challenges in the actual operating environment, such as different types and sizes of cargo, complex stacking methods, and changes in various environmental factors. Existing research may not have adequately addressed the impact of these complexities and uncertainties on attitude detection and grasp planning. Meanwhile, attitude detection usually relies on various sensors, such as cameras, lidar, and so on. These sensors may not provide sufficiently accurate or reliable data under certain conditions (such as bad weather, or low illumination), which can affect the accuracy of detection and planning. Automatic ship unloaders require fast and efficient processing of large amounts of data for attitude detection and grasp planning. Current technology may not meet the requirements of high efficiency and real-time performance, especially in complex scenes, with limited generalization ability. The system's adaptability and learning ability present significant challenges, and traditional research lacks sufficient autonomy in learning. Therefore, a 3D coordinate system dynamic construction technology based on deep learning is proposed in this paper. The mature CNN technology is used to estimate the three-dimensional depth map of the cargo. The traditional depth map is transformed through point cloud reconstruction and modules such as residual upper projection are introduced to enhance the recognition effect and efficiency under different conditions.

3 Research method

The key of an automated ship unloader is to detect and identify the target cargo, in order to achieve the conversion from the original image to the depth map, and finally transmit it to the controller to guide cargo loading and unloading. Therefore, a visual 3D coordinate reconstruction technology based on deep learning is proposed. Firstly, the working principles of the ship unloader and depth camera are introduced, the coordinate reconstruction process is designed, and then the encoding and decoding system is optimized through deep learning.

3.1 Visual 3D coordinate and point cloud reconstruction and optimization of encoding and decoding system construction

Port terminals usually use ship unloaders for the loading and unloading of large goods, which requires equipment with strong lifting capacity and small work limitations to ensure smooth loading and unloading of goods of different shapes and weights. The principle of mechanically grasping the target object is shown in Figure 1.

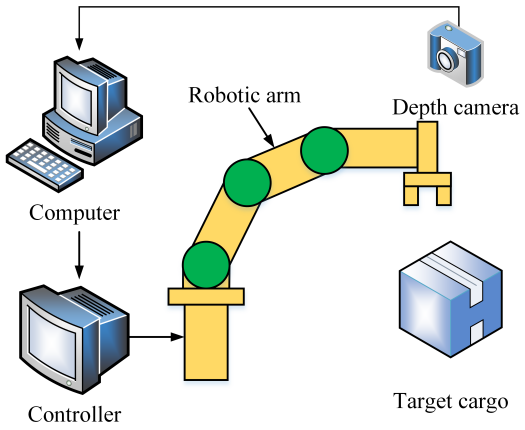
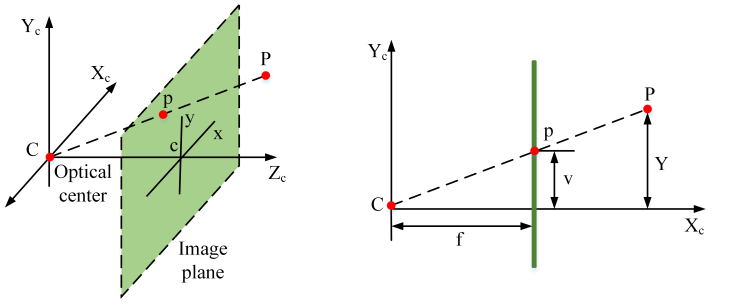


Figure 1: Schematic diagram of loading and unloading of ship unloader

The workflow of loading and unloading can be roughly divided into three steps. Firstly, a depth camera is used to capture the relevant depth data of the goods, and then the depth image is fed back to the computer for recognition. The information is further processed through pose detection technology, packaged, and transmitted to the controller. Finally, the controller sends commands to drive the equipment to achieve loading and unloading of the goods [12]. Among them, the key to affecting the operational accuracy of the ship unloader is the target detection step, including the construction of the three-dimensional coordinate system of the cargo by the depth camera, as well as the transformation of the pose coordinate system. The research is achieved through visual 3D reconstruction, and the principle of camera imaging is the small-hole imaging model. The target plane extends along the Z-axis to the front position point, as shown in Figure 2.



(a) A simplified pinhole imaging model (b) Side view of keyhole imaging along the X-axis

Figure 2: Schematic diagram of pinhole imaging

Small hole imaging includes a camera coordinate system, world coordinate system, and image coordinate system, respectively represented by $X_cY_cZ_c/X_wY_wZ_w/xyz$, and assuming that the origin of the coordinate system is consistent with the center of the image. $P = (x_w, y_w, z_w)^T \in R^3$ represents any point in a three-dimensional coordinate system, corresponding to a $p = (u, v)^T \in R^2$ point in a two-dimensional coordinate system. Figure 2 (b) shows the view along the camera X_c axis, and the matrix form of the geometric relationship is shown in equation (1) [13].

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \frac{1}{Z} \begin{pmatrix} f & 0 & c_x & 0 \\ 0 & f & c_y' & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{1}$$

In equation (1), $\left(c_x', c_y'\right)$ is the offset of the image point, as the assumption that the image coordinate origin coincides with the image center is difficult to achieve in reality. At the same time, the unit of the image coordinate system is pixels, while the rest uses unit length. k_x/k_y is introduced to represent its unit length pixels, as shown in equation (2).

$$\begin{cases} f_x = k_x \cdot f & f_y = k_y \cdot f \\ c_x = k_x \cdot c_x^{'} & c_y = k_y \cdot c_y^{'} \end{cases} \tag{2}$$

By converting it to the camera coordinate system, the camera's internal parameter matrix can be obtained as shown in equation (3).

$$C = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \tag{3}$$

To further simplify the model, assuming that the world coordinate system and camera coordinate system have the same origin and do not rotate, the translation matrix and rotation matrix of the system can be obtained as shown in equation (4) [14].

$$\begin{cases} S = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T \\ R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{cases} \tag{4}$$

In equation (4), $T = \begin{bmatrix} R & S \end{bmatrix}$ represents the external parameter matrix of the camera. By substituting equation (1), the final world coordinate system can be obtained as shown in equation (5).

$$\begin{cases} x_w = Z\left(\mathcal{A} - c_x\right)/f_x \\ y_w = Z\left(\mathcal{A} - c_y\right)/f_y \\ z_w = Z \end{cases} \tag{5}$$

Equation (5) is the foundation for realizing the transformation from a depth map to a point cloud. This process requires first converting the two-dimensional map into a depth map through a deep learning network, then obtaining the corresponding coordinates of the world coordinate system based on the point depth d, and finally completing the three-dimensional point cloud mapping. The axial error of this point is shown in equation (6).

$$\begin{cases} \Delta x = x - x^* \\ \Delta y = y - y^* \\ \Delta z = z - z^* \end{cases} \tag{6}$$

In equation (6), (x^*, y^*, z^*) represents the three-dimensional point cloud coordinates. The input and output dimensions of loading and unloading goods are too large, requiring the network to have better structure and hyperparameters, and the task is essentially a transformation in the image domain.

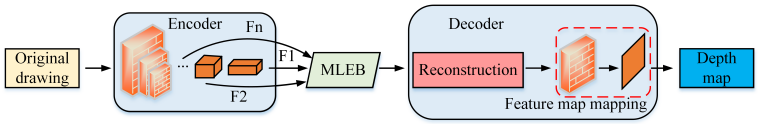


Figure 3: Schematic diagram of MLEED-Net

Depth map prediction belongs to dense prediction, where encoders decoders are widely used, while deep CNNs excel in dense estimation optimization [15, 16]. Therefore, the study utilizes CNN to optimize the encoder and decoder and proposes an end-to-end high-level integrated encoder-decoder network (MLEED-Net), as shown in Figure 3.

In Figure 3, F represents output numbers at different levels. The encoder extracts features through convolution and converts the data from low-level to high-level semantics. The improvement in feature abstraction also strengthens the requirements for computing networks, and deep CNN can meet this requirement. ResNet- 50 is chosen as its backbone network. The Multi Layer Ensembled Block (MLEB) takes the encoder output as input information and achieves feature integration through spatial aggregation. The reconstruction module in the decoder introduces a Residual Up Projection Block (RUPB) for advanced semantic decoding. The ResNet-50 network will increase the output features and increase the computational burden, so it is directly mapped to the depth map through the feature map. The prediction of depth maps by CNN is essentially a regression task, with consistent input and output sizes. Therefore, by directly inputting the original RGB image, the depth map can be output. Deep prediction also requires a loss function to optimize network training. Common loss functions include L_1 loss function, L_2 loss function, and inverse loss function [17, 18]. Among them, the loss function has the highest fitness with the MLEED-Net network, so this function is selected as the loss function, as shown in equation (7).

$$L_1 = \frac{1}{N}\sum_{i=1}^{N}|d_i - d_i^*| \tag{7}$$

In equation (7), d_i represents the predicted depth map, d_i^* represents the actual depth map, and N represents the number of features.

3.2 Depth map quality optimization design based on conditional generative adversarial network

To further optimize the quality of depth maps, an improved model based on the cGAN is proposed. This network is adept at processing high-resolution image tasks, and its structure is mainly divided into two parts: a generator and a discriminator. The former realizes the transformation from random noise to image-generated samples, while the latter identifies and identifies the data to distinguish between actual data and predicted data. By continuously enhancing the performance of both through training, the model converges when the discriminator is unable to distinguish between predicted and actual values. However, random noise input can increase the bias of predicted data and cannot be controlled. Therefore, the cGAN is introduced for further optimization, aiming to limit the generation of generated data by constraining condition y [19]. Compared to traditional network optimization, this model only uses color images as input values for the generator, and estimates and outputs depth maps through depth information. The entire process does not include post-processing steps, thus greatly optimizing network efficiency, as shown in Figure 4.

The predicted depth map output by the generator is represented as $\times$ (fake), with the purpose of "deceiving" the discriminator. The actual depth map is represented as $\times$ (real), and is compared with the predicted depth map in the discriminator. Among them, the generator learns the mapping function based on the observed sample x and noise z, as shown in equation (8).

$$G : \{x, z\} \to y \tag{8}$$

The demand for model training for actual depth maps is inevitable, as it serves as supervisory information, which is the constraint conditions introduced in the model, to achieve the relevant limitations

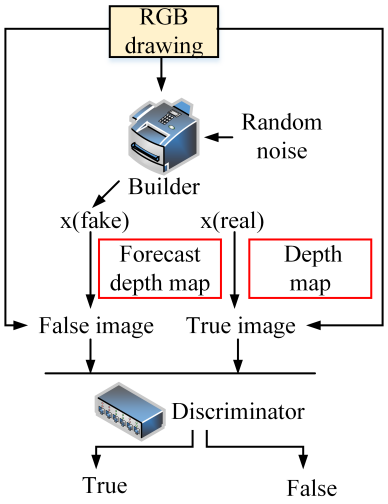


Figure 4: Basic architecture of cGAN

of generator output [20]. It can be seen that the generator network is essentially an end-to-end dense prediction task. The study introduces a pix2pix network that can achieve image style conversion. This is a network structure improved by U-Net technology with superior image processing performance, as shown in Figure 5.

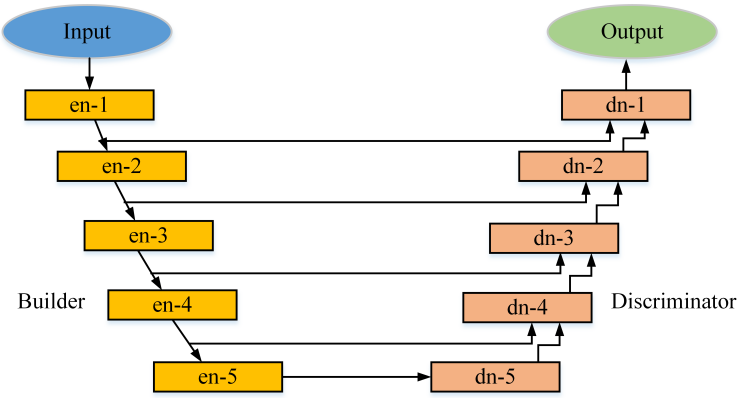


Figure 5: pix2pix network infrastructure

In Figure 5, pix2pix-G represents the generator in the model. The difference between this network and the initial U-Net model lies in the changes in the encoding and decoding structure. The MLEED-Net model designed earlier was selected and designed according to a consistent structure, only changing the number of convolutional filters. Moreover, the connection methods of encoding and decoding between the two models are also different. The former connects through an uncut feature map and only uses a layer of convolution with a size of $4 * 4$ in the encoder. However, this feature extraction performance is poor. Therefore, the Skip-Convolution Down-Sampling Module (SCDSM) is introduced in the study, which is compared with the bottleneck module structure in ResNet, as shown in Figure 6.

In Figure 6, conv-2 and k3s1 represent a convolution size of 3×3 with a step size of 1 . It can be seen that the SCDSM module has a relatively similar structure to the bottleneck module, while the former mostly introduces convolutional jump connections and mean pooling. When extracting features from the data, the encoder utilizes hierarchy to reduce image resolution while increasing the number of channels. It performs advanced semantic encoding to match conv- 3 with skip-connected channels and introduces conv- 3 to increase the number of channels. Mean pooling is also aimed at reducing the spatial resolution of the graph, with a window size of $2 * 2$ and a step size of 2. This module is used to replace the encoder in the MLEED-Net model, but without changing the decoder

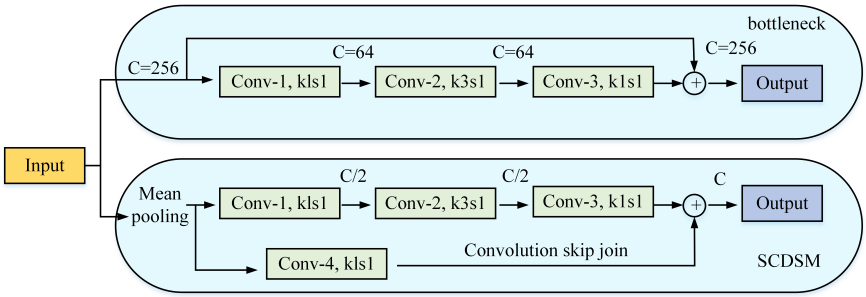


Figure 6: Comparison structure of the SCDSM module and bottleneck module

structure. The discriminator generally uses CNN structure for feature extraction and uses a sigmoid activation function to output similarity, represented by pix2pix-D, including five convolutional layers. After the convolutional structure, it is necessary to use standardized output at different levels and use multi-scale feature data to unify the spatial resolution of each part. Finally, the average response is used as the fusion data, as shown in equation (9) [21].

$$F_m = \frac{1}{n}\sum_{i=1}^{n} S\left(f_i\right) \tag{9}$$

In equation (9), F_m is the multi-feature fusion information and $S(\mathcal{K}$ is the downsampling function. The downsampling adopts the nearest neighbor difference method. The input for training the cGAN is random noise and image, and its objective function is shown in equation (10).

$$L_{cGAN}(G,D) = E_{x,y}[\log D(x,y)] + E_{x,z}[\log(1 - D(x, G(x,z)))] \tag{10}$$

The generator needs to minimize L_{cGAN}, as shown in equation (11).

$$\min_{G} = C_g \cdot \frac{1}{M}\sum\nolimits_{i=1}^{M}\left(\log(1 - D(I_r, I_p))\right) + C_c \cdot L_1 \tag{11}$$

In equation (11), $C_g = 1$ represents the weight coefficient of the objective function, $C_c = 100$ represents the weight coefficient of the loss function, I_r represents the color image, I_p represents the generator output, and M represents the number of sample pairs. Among them, the sample pairs are shown in equation (12) [22].

$$(I_r, I_g) = \left(I_r^{(1)}, I_g^{(1)}\right), \ldots, \left(I_r^{(M)}, I_g^{(M)}\right) \tag{12}$$

The discriminator needs to minimize L_{cGAN}, as shown in equation (13) [23].

$$\max_{D} = \frac{1}{M}\sum_{i=1}^{M}\left(\log\left(1 - D\left(I_r, I_p\right)\right) + \log D\left(I_r, I_g\right)\right) \tag{13}$$

In equation (13), I_g is the actual depth map. To optimize the quality of production images, a L_1 loss function is introduced, and the optimized objective function is shown in equation (14).

$$G^* = \arg\min_{G}\max_{D} C_g L_{cGAN} + C_c L_1 \tag{14}$$

In equation (14), G^n represents the objective function of the cGAN optimized by the loss function.

4 Result and discussion

To verify the effectiveness of the proposed deep learning-based 3D coordinate dynamic construction method for loading and unloading machinery, simulation analysis experiments were conducted on it. Firstly, performance analysis experiments were conducted on the design algorithm itself to ensure the reliability of its model. Then, it was applied to practical scenarios of loading and unloading goods to ensure its feasibility in practical applications.

4.1 Codec CNN network and depth map recognition cGAN error and accuracy performance analysis

The study first conducted experimental validation on CNN and cGAN models, and the experimental environment and related parameter settings are shown in Table 1.

Table 1: Experimental environment and related parameter Settings

Name		Configure/Selective value
Processor		E5-2620 v4
Graphics card		NVIDIA Titan XP
Internal memory		32GB
Software environment		lensorflow
Network training optimizer		Adam
Initial learning rate		10-3
Batch size		16
Convolution kernel		3*3
Pooled window		2*2
Convolution/Pooling step		2
Convolutional filter number		64/128/256/512/1
Training rounds		2000
Loss function		L_1
Data set	Indoor scene	NYUD-v2
	Outdoor scene	KITTI

The NYUD-v2 dataset contained original color maps and corresponding depth maps, with a total of 414 scenes and image resolutions of 640 * 480. However, some depth maps had some missing phenomena in points and noise, so the bilateral filtering method was used to correct them [24]. This dataset was indoor scene data, according to the official classification of the training and testing scenes were 249 and 215 respectively. In addition, there was a difference in the viewing angle between the original depth image and the color image, so it was also necessary to correct the field of view of the depth map, and the size of each image was uniformly 226*302. The KITTI dataset data was used as an outdoor scene, and the distance of the depth map reached a maximum of 100m, which was 10 times that of the NYUD-v2 dataset, unlike the former that was captured with sensors and cameras, the KITTI dataset was captured by LIDAR, and its depth map was more sparse. Unlike the former, which utilized sensors and cameras, the KITTI dataset was captured by LIDAR, and its depth map was more sparse, so on top of the bilateral filtering process, it also needed to be cropped, and the dimensions of each image were uniformly 226*454. The ratio of the training set to the test set images for the two datasets was 8:2. The selection of hyperparameters was first verified and the experimental results for batch size are shown in Figure. 7.

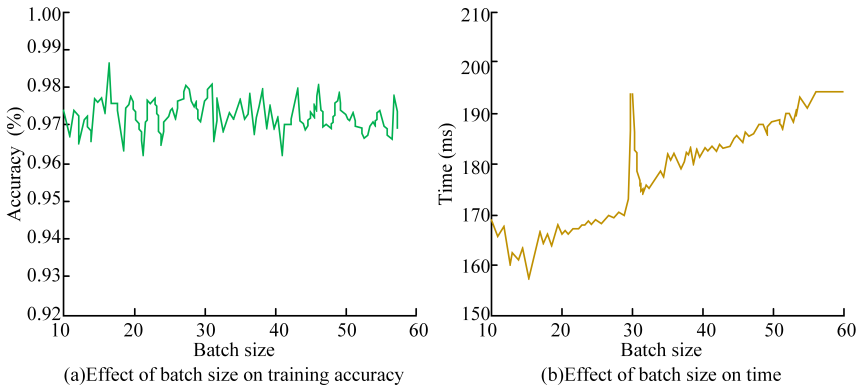


Figure 7: Effect of parameter settings on modeling

Figure 7 shows the training accuracy and running time of the model under different batch sizes, respectively, whose accuracy fluctuated in the interval of [0.96,0.99], but reached the maximum accu-

racy of 98.98% when the batch size was equal to 16. The change curve of running time with batch size showed a steadily increasing trend in general, but there was a more obvious decreasing trend at the beginning, i.e., when the batch size was equal to 16, the running time of the model was the shortest, which was 155.46ms. The study first compared the CNN-based MLEED-Net model with the VGG16 (Visual Geometry Group 16) network and Google Net network and used Mean Relative Error (MRE), Logarithmic mean error (LMAE), Root mean square error (RMAE), and accuracy as performance evaluation criteria. The experimental results are shown in Figure 8.

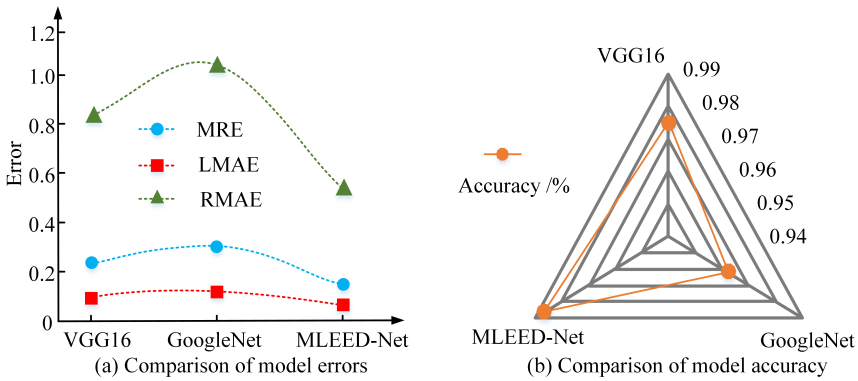


Figure 8: Performance comparison of codec models

From Figure 8(a), it can be seen that the RMAE error values of each model were relatively high, with values of 0.824, 1.04, and 0.539, respectively. The MLEED-Net model proposed in the study was 34.59% and 48.17% lower than the VGG16 network and Google Net network in this indicator. The MRE values of each model were 0.23, 0.305, and 0.145, respectively, which reduced the difference compared to RMAE. However, the error values of the MLEED-Net model were still lower than those of the VGG16 network and Google Net network, with 8.5% and 52.46%, respectively. The LMAE value of the MLEED-Net model was 0.063, which was on average 38.715% lower than the other two models. Therefore, the error of the CNN-based MLEED-Net model proposed in the study was always the lowest. From Figure 8 (b), it can be seen that the accuracy of each model reached over 90%, which was 97.5%, 96.3%, and 98.8% in order, respectively [25]. Therefore, the accuracy of the MLEED-Net model was on average 1.9% higher than that of the other models. Further experimental verification was conducted on the cGAN for depth map recognition, and pyramid matching network (PMN) and radial basis function network (RBF) were introduced for comparative analysis. The experimental results are shown in Figure 9.

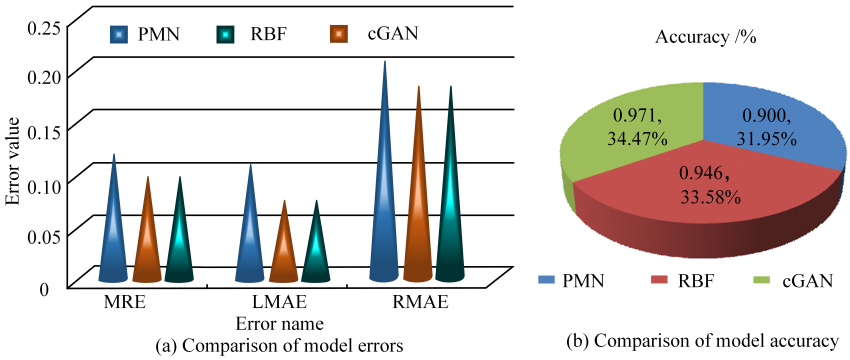


Figure 9: cGAN model performance control experiment

Figure 9(a) shows the values of each model in different errors. Among the MRE error values, the values of each model were 0.115, 0.102, and 0.097, respectively. The cGAN proposed in the study was 15.65% and 4.90% lower than the PMN and RBF models, respectively. The LMAE value of the cGAN was 0.044, which was 60.36% and 42.1% higher than the other two algorithms, respectively. The RMAE value of the cGAN was 0.153, and the difference between the RBF model and the model

was not significant, only 0.0034. However, the RMAE value of the PMN model reached 0.212, which was 27.83% higher than that of the cGAN. Figure 9(b) shows the accuracy of each model, all at 90% or above, while the cGAN proposed in the study reached 95%, which was 7.8% higher than the PMN model and 2.65% higher than the RBF model. The above data indicated that the CNN network optimized as an encoder and decoder, as well as the depth map recognition network cGAN, achieved excellent performance, proving the effectiveness of this optimization algorithm for depth map recognition based on 3D coordinate reconstruction.

4.2 CNN-cGAN model overall performance analysis and practical application in simulated loading and unloading systems

The study further applied the depth map recognition network to the simulation of loading and unloading machines, where the error generated by model training varied with iteration, and the accuracy varied with iteration, as shown in Figure 10.

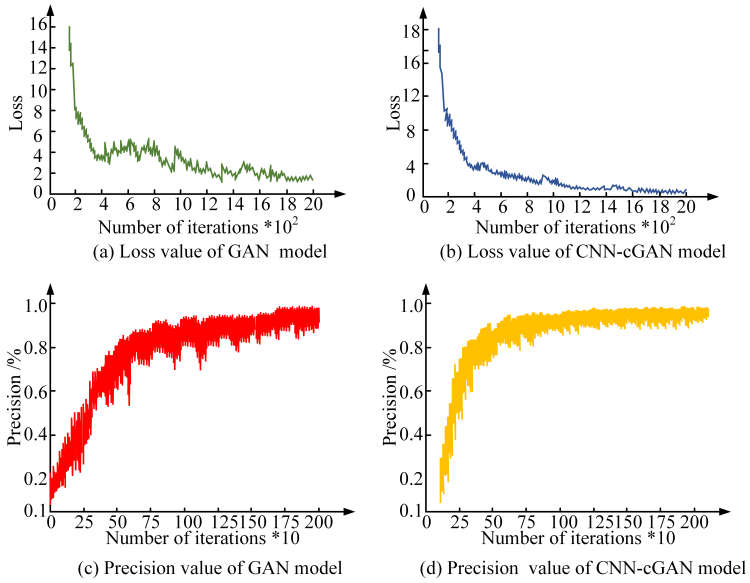
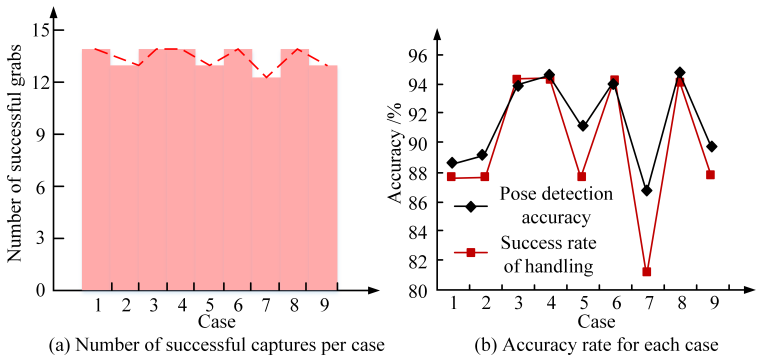


Figure 10: Changes of model training error and accuracy rate

Figure 10 is a visual image of model training. From Figures 10 (a) and 10 (b), it can be seen that the convergence speed of the model using the GAN network alone was significantly slower. After the error decreased rapidly with iteration, there were still significant fluctuations. This indicated that the model was not stable during the training process, and the number of iterations required for the curve area to be flat was higher. This indicated that the fitting speed of the model was not ideal. After 2000 iterations, The loss value of the GAN network fluctuated in a small range of around 1.41. The CNN-cGAN model proposed in the study had a faster convergence speed, a more stable training process, and a smoother region. Therefore, the model had a more stable performance and faster fitting speed. After 2000 iterations, the loss value of the CNN-cGAN model remained stable at around 0.78. Figures 10 (c) and 10 (d) show the accuracy variation curves of the two models. There were also differences in stability, fitting speed, and other aspects between the two models. The CNN-cGAN model performed better, and the final recognition accuracy reached 99.2%, an improvement of 4.3% compared to the previous one. There were failure cases in GAN network training, where the detection position differed significantly from the actual coordinates, resulting in loading and unloading failure. In successful cases, although some detection boxes were correct, there were still some gaps between their rotation angles and the actual optimal coordinates, ultimately leading to loading and unloading failure. Nine objects were selected for the study, including wooden blocks, water bottles, screwdrivers, boxes, drill bits, hexagonal nuts, gears, tape, and weights. They were each grabbed 15 times, and the experimental results are shown in Figure 11.

Figure 11 (a) shows the number of successful grabbing attempts in different cases. It can be seen



(a) Number of successful captures per case (b) Accuracy rate for each case

Figure 11: Experimental results captured by different cases

that except for gears, the number of successful grabbing attempts was 12, while the other objects were all 13 or 14, with an average success rate of 91.89%. This indicated that machines basically achieved grasping in actual scenarios, and some reasons for grasping failures may be related to the placement position and angle of the target object. In Figure 11 (b), the accuracy of object pose recognition and the success rate of grasping changed. The accuracy rate of pose recognition was basically above 90%, with an average pose detection accuracy of 93.1%. However, the success rate of loading and unloading was lower than the accuracy rate of pose detection. This is because although some detection boxes were correct, the optimal grasping point was not located, resulting in grasping failure. Although there were some cases of grasping failure in the experiment, the overall performance was relatively excellent, and the experiments mostly used smaller objects, which undoubtedly increased the difficulty of machine operation. In practical applications, larger objects were targeted, and their success rates were also significantly improved. To understand the performance differences between the proposed method and other methods, the Perceptual Index (PI), a high-resolution quality evaluation index Ma for depth maps, and a natural image evaluation index NIQE were further selected for evaluation. The methods proposed by Wang M et al. and Jiang C et al. were introduced, and the experimental results are shown in Table 2.

Table 2: Comparison between research methods and other methods

Method	PI	Ma	NIQE	Accuracy(%)
Wang M[19]	10.013	2.288	15.655	89.93
Jiang C[20]	8.788	1.518	11.865	90.14
Ours	7.295	1.392	10.246	91.89

Among them, depth map resolution played an important role in the quality of 3D reconstruction. High resolution corresponded to a more specific level of spatial detail, making features such as contours, edges, and surface texture of objects clearer, which was very important for the recognition and accurate grasping of complex objects. The PI metric was also an index for evaluating the clarity of depth maps. The NIQE metric, on the other hand, contained a comprehensive evaluation of several aspects of the peak signal-to-noise ratio of the image as well as the image quality index. From Table 2, the proposed method had the best performance in all aspects, with its PI index lower than the other two models by 27.14% and 16.99%, respectively. Its Ma index was 39.16% and 12.6% lower than the other two models, respectively. Its NIQE indicators were 34.55% and 13.65% lower than the other two models, respectively. The success rate of grasping was 2.18% and 1.94% higher than the other two models, respectively. In summary, it can be seen that the CNN-cGAN proposed in the study can achieve better depth map quality and a higher success rate in crawling.

5 Conclusion

To improve the performance of automated ship unloaders, a deep learning-based 3D coordinate system dynamic construction technology is proposed for optimization. A deep CNN network is used

to build a codec MLEED-Net model, which is embedded in the cGAN to achieve the final depth map prediction. To verify its performance, simulation experiments are conducted on CNN and cGANs. The experimental results showed that the RMAE/LMAE/MRE values of the MLEED-Net model were on average 36.5% lower than those of the VGG16 network and Google Net network. The MRE values of the cGAN were 0.097, LMAE values were 0.044, and RMAE values were 0.153, with an accuracy of 95%. Subsequently, the complete model was trained and applied in practice. The experimental results showed that after 2000 iterations, the loss value of the GAN network was around 1.41, while the CNN-cGAN remained stable at around 0.78, and its recognition accuracy reached 99.2%, an improvement of 4.3% compared to the previous one. The study conducted simulation experiments on grasping 9 objects such as wood, and the results showed that the average pose detection accuracy was 93.1%. In the 15 grasping attempts of each object, the number of successful attempts reached approximately 13, with an average success rate of 91.89%. In comparison with the methods proposed by Wang M et al. and Jiang C et al., the PI index of the proposed method was 27.14% and 16.99% lower than the other two models, respectively, and the capture success rate was 2.18% and 1.94% higher than the other two models, respectively. Therefore, it has better depth map quality and grasping ability. However, the study only considers the loading and unloading of individual objects, and subsequent research can further analyze the situation of object stacking.

Funding

(1)General Program of Natural Science Foundation of Chongqing: Research on the safety of automatic loading and unloading of lifting ship unloader through dynamic construction of three-dimensional coordinate system.(CSTB2023NSCQ-MSX0981)

(2)Science and Technology Research Program of Chongqing Education Commission: Research on key Technologies of Predictive Maintenance of Industrial Bobots based on operation Big data.(KJZD-M202203201)

Author contributions

The authors contributed equally to this work.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Milana, G.; Banisoleiman, K.; González, A. (2021). An investigation into the moving load problem for the lifting boom of a ship unloader, *Engineering Structures*, 234(1), 111899-111919, 2021.

[2] Zan, J. (2022). Research on robot path perception and optimization technology based on whale optimization algorithm, *Journal of Computational and Cognitive Engineering*, 1(4), 201-208, 2022.

[3] Alam, M. S.; Kwon, K. C.; Kim, N. (2021). Implementation of a Character Recognition System Based on Finger-Joint Tracking Using a Depth Camera, *IEEE Transactions on Human-Machine Systems*, 51(3), 229-231, 2021.

[4] Li, L.; Chen, X.; Zhang, T. (2022). Pose estimation of metal workpieces based on RPM-Net for robot grasping from point cloud, *Industrial Robot*, 49(6), 1178-1189, 2022.

[5] Gao, H. N.; Shen, D. H.; Yu, L. ; Zhang, W. C. (2020). Identification of Cutting Chatter through Deep Learning and Classification, *International Journal of Simulation Modelling*, 19(4), 667-677.

[6] Kushwaha, V.; Shukla, P.; Nandi, G. C. (2023). Generating quality grasp rectangle using Pix2Pix GAN for intelligent robot grasping, *Machine Vision and Applications*, 34(1), 09821-09836, 2023.

[7] Shengqian, L.; Xiaofan, Z. (2022). Research on Hand-eye Calibration Technology of Visual Service Robot Grasping Based on ROS, *Chinese Journal of Instrument: English*, 9(001), 23-30, 2022.

[8] Higo, R.; Senoo, T.; Ishikawa, M. (2020). Dynamic In-Hand Regrasping Using a High-Speed Robot Hand and High-Speed Vision, *IFAC-PapersOnLine*, 53(2), 9796-9801, 2020.

[9] Jin, X.; Tang, L.; Li, R.; Zhao, B.; Ji, J.; Ma, Y. (2022). Edge recognition and reduced transplantation loss of leafy vegetable seedlings with Intel RealSense D415 depth camera, *Computers and Electronics in Agriculture*, 198(1), 107030-107044, 2022.

[10] Štampar, M.; Fertalj, K. (2022) Applied Machine Learning in Recognition of DGA Domain Names. *Computer Science and Information Systems*, 19(1), 205-227, 2022.

[11] He, G.; Huang, X.; Xu, Y.; Qin, J.; Sun, K. (2023). Cable tunnel unmanned aerial vehicle location and navigation method based on ultra wideband and depth camera fusion, *IET Electrical Systems in Transportation*, 13(1), 12068-12069, 2023.

[12] He, B.; Qian, S.; Niu, Y. (2023). Visual recognition and location algorithm based on optimized YOLOv3 detector and RGB depth camera, *The Visual Computer*, 240(1), 1-17, 2023.

[13] Zhou, C.; Ren, D.; Zhang, X.; Yu, C.; Ju, L. (2022). Human Position Detection Based on Depth Camera Image Information in Mechanical Safety, *Advances in Mathematical Physics*, 000(1), 9170642-9170652, 2022.

[14] Zou, L.; Zhuang, K. J.; Zhou, A.; Hu, J. (2023). Bayesian optimization and channel-fusion-based convolutional autoencoder network for fault diagnosis of rotating machinery, *Engineering Structures*, 280(1), 115708-115720, 2023.

[15] Cajas, Y. R. A.; Guisado, Y. Z.; Vergaray, A. D. (2023). Identify faults in road structure zones with deep learning. *Journal of System and Management Sciences*, 13(1), 63-84, 2023.

[16] Li, C.; Dong, G.; Wang, R. (2022). A three-dimensional reconstruction algorithm of nonwoven fabric based on an anthill model, *Textile Research Journal*, 92(11), 1876-1890, 2022.

[17] Ji, X.; Zhao, Q.; Cheng, J.; Ma, C. (2021). Exploiting spatio-temporal representation for 3D human action recognition from depth map sequences, *Knowledge-Based Systems*, 227(1), 107040-107040, 2021.

[18] Ni, J.; Zhou, Z.; Zhao, Y.; Han, Z.; Zhao, L. (2023). Tomato leaf disease recognition based on improved convolutional neural network with attention mechanism, *Plant Pathology*, 72(7), 1335-1344, 2023.

[19] Cheng, Y.; Wan, Y.; Sima, Y.; Zhang, Y.; Hu, S. ; Wu, S. (2022). Text Detection of Transformer Based on Deep Learning Algorithm. *Tehnički vjesnik*, 29(3), 861-866, 2022.

[20] Sun, X.; Zhou, K.; Shi, S.; Song, K.; Chen, X. (2022). A new cyclical generative adversarial network based data augmentation method for multiaxial fatigue life prediction, *International Journal of Fatigue*, 162(1), 106996-107013, 2022.

[21] Saeed, A.; Hayat, M. F.; Habib, T.; Ghaffar, D. A.; Qureshi, M. A. (2022). A novel multi-speakers Urdu singing voices synthesizer using Wasserstein Generative Adversarial Network, *Speech Communication*, 137(1), 103-113, 2022.

[22] Zhang, H.; Chen, X.; Lu, S.; Yao, L.; Chen, X. (2023). A contrastive learning-based attention generative adversarial network for defect detection in color-patterned fabric, *Coloration Technology*, 139(3), 248-264, 2023.

[23] Wang, M.; Luo, J.; Zheng, L.; Yuan, J.; Walter, U. (2020). Generate optimal grasping trajectories to the end-effector using an improved genetic algorithm, *Advances in Space Research*, 66(7), 1803-1817, 2020.

[24] Hong, L. C.; Tee, C.; Goh, M. K. O. (2022). Activities of daily living recognition using deep learning approaches. *Journal of Logistics, Informatics and Service Science*, 9(4), 129-148, 2022.

[25] Jiang, C.; Wang, D.; Zhao, B.; Liao, Z.; Gu, G. (2021). Modeling and inverse design of bio-inspired multi-segment pneu-net soft manipulators for 3D trajectory motion, *Applied Physics Reviews*, 8(4), 054468-054468, 2021.

Cite this paper as:
Wang, L.F.; Li, Q.; Fu, W.; Jiang, F.; Song, T.X.; Pi, G.B.; Sun, S.J. (2024). Enhancing Automated Loading and Unloading of Ship Unloaders through Dynamic 3D Coordinate System with Deep Learning, *International Journal of Computers Communications & Control*, 19(2), 6234, 2024.
https://doi.org/10.15837/ijccc.2024.2.6234