Application of Structured Light 3D Reconstruction Technology in Industrial Automation Scenarios in the Context of Digital Transformation

Lihong Wang[1]*, *Jun* Cai[1], *Yanan* Zhang[1], *Dong* Chen, *Lv* Ge[1], *Yu* Zhang[1]

[1]ZheJiang Gongshang University Hangzhou College of Commerce, China

Abstract. In this work, a dynamic structured light 3D reconstruction method adapted to random phase shift step size is proposed for the problem of dynamic workpiece 3D reconstruction in industrial manufacturing scenes. The method firstly utilizes a structured light sinusoidal grating pattern to match the relative phase shift of the workpieces moving along the perpendicular direction of the stripes on the conveyor belt. Secondly, in order to solve the problem of non-uniform phase shift step length due to external vibration, electromagnetic radiation and other disturbing factors of the conveyor belt and other mechanical devices, the method proposes a dynamic structured light 3D reconstruction model based on RPSNet, which is based on the CycleGAN network model, and uses the AIR2U-net model proposed in this paper as the generator, and the multilayer convolutional neural network CNN as the discriminator to realize the grating map. discriminator to realize the conversion of raster map to depth map. Finally, for the problem that the network model needs a large amount of data for training and it is difficult to collect data in the actual industrial scene, this paper uses the Thing10k dataset, and the dataset made by Blender simulation software for model training. Finally, a higher quality 3D reconstruction of the workpiece is realized.

1 Introduction

Structured light as an optical non-contact 3D shape measurement technique has been widely used in intelligent manufacturing, reverse engineering, and heritage digitization. FPP[1] (Fringe Projection Profilometry) based structured light projection technology is one of the most popular optical 3D imaging technologies due to its simple hardware structure, flexible implementation, and high measurement accuracy. With the improved performance of imaging and projection equipment, it is possible to realize dynamic 3D shape measurement based on FPP structured light technology. To realize 3D measurement in dynamic scenes, it is usually necessary to reduce the number of images required for each reconstruction to improve the measurement efficiency. The traditional structured light measurement method theoretically requires at least three phase-shifted images to complete a reconstruction, and in the actual reconstruction process, to obtain a higher reconstruction quality, it often requires five or more phase-shifted images, and the quality of the 3D reconstruction and the number of phase-shifted images are to a certain extent directly proportional. The quality of 3D reconstruction and the number of phase-shifted images are to some extent proportional. However, this traditional method has good reconstruction effect in static scenes and non-real-time measurement scenes, but in dynamic measurement or real-time measurement scenes, this method cannot reach the expected effect, this is because the acquisition of multiple phase-shifted images takes time, which is unacceptable for applications requiring high real-time performance, and what is more important is that due to the motion of the object to be measured, the error between the multiple phase-shifted images is caused by the motion of the object, and this is not acceptable for the applications requiring high real-time performance. More importantly, since the object is in motion, there are errors between multiple phase-shifted images due to object motion, which leads to unsatisfactory results in the final 3D reconstruction.

Since the structured light three-dimensional measurement technology was proposed in the 1970s, it has been rapidly developed due to its features of high measurement accuracy, high real-time performance, and easy extraction of grating images. The principle of the structured light method is to firstly form a three-dimensional reconstruction system with the streak projector, image collector and the object to be measured according to a certain positional relationship; secondly, project the encoded structured light map on the object surface of the object to be measured; then use the visual sensor for image acquisition, so as to obtain the structured light image projection information on the surface of the object to be measured as well as on the object's reference plane; lastly, utilize the principle of triangulation, Finally, the acquired image data are

*Corresponding author: wlh@zjhzcc.edu.cn

processed using triangulation principle, image processing and other technologies to calculate the distance information of the object surface from the camera, thus realizing the conversion of two-dimensional images to three-dimensional images. Common structured light measurement techniques can be categorized into point structured light, line structured light, and surface structured light according to the different projection patterns, and surface structured light is chosen for 3D reconstruction in this paper because of its higher measurement accuracy.

Due to the excellent characteristics of structured light measurement technology, it has aroused the research interest of many researchers and achieved fruitful results. For example, Qian[2] et al proposed a method for unwrapping the phase of a single frame of structured light based on deep learning and geometrical constraints, which improves the robustness of the measurement and effectiveness. Spoorthi[3] et al proposed PhaseNet as well as PhaseNet2.0 network model, which uses DenseNet neural network to predict the number of phase unfolding steps for each pixel from the phase unfolding map, thus achieving efficient phase unfolding, and the method is also robust to noise. Yang[4] et al proposed a deep learning-based generalized phase error compensation method, which can effectively eliminate the influence of nonlinear effects in Phase Shifting Profilometry (PSP) on the phase error in 3D measurements. Zhang[5] et al solved the problem of the loss of phase information due to the large difference between light and dark by designing a specialized convolutional neural network, so that the loss of phase information can be realized by neural network using the phase expansion stages. and can realize 3D reconstruction in images with oversaturated or low brightness by neural network. In addition, several 3D structured light sensing cameras developed by the ORBBEC are widely used in face recognition, 3D measurement, map reconstruction and other fields.

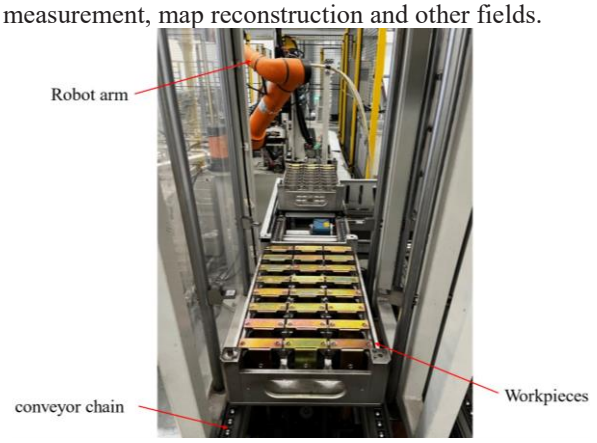

Fig. 1. Application Scenario

In recent years, with the improvement of neural network algorithms and the arithmetic power of computer hardware, deep learning presents a powerful fitting ability. Many studies have proved that deep learning is superior to traditional algorithms in terms of speed and robustness. On the research field of structured light, deep learning is also widely used in the direction of stripe denoising, stripe analysis, phase unfolding and so on.

The application scenario diagram of this paper is shown in Fig. 1. The method proposed in this paper will replace manual labor to yard the scattered workpieces in the yarding trays, which will be transported by the conveyor chain to the designated location and subsequently processed. For this application scenario, this paper proposes a dynamic structured light 3D reconstruction method adapted to random phase shift, which makes the structured light 3D reconstruction in dynamic scenarios achieve better results. Specifically, the traditional static structured light measurement uses the phase shift technique to realize the 3D measurement, generally the measured object is kept stationary and the grating pattern is displaced by N steps (N>3), the essence of which is to use the relative displacement between the grating pattern and the measured object to make the pixels with different brightness in the grating pattern to be projected onto the object, and the grating pattern is modulated by the height of the object after the grating pattern is projected onto the object, and then the corresponding pixels with different brightness can be found using the height of the object is modulated by the height of the raster pattern. Currently, the phase of the corresponding pixels is derived from the pixels of different brightness, and then the height of the object is derived by matching the pixels of the same phase in the left and right cameras. Since the phase shift technique requires relative displacement between the object to be measured and the grating pattern, the grating pattern remains unchanged during the measurement process, and the directional displacement of the object to be measured also meets the requirements of the phase shift, and the movement of the object in the scene happens to be one-dimensional, i.e., the object is only moving along the conveyor belt in the direction of the movement, which provides the conditions for the realization of the above relative phase shift. The specific scene schematic is shown in Fig. 2.

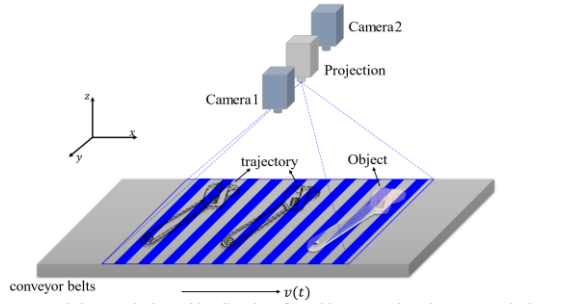

Fig. 2. Scene Schematic

Although the above method transforms the original motion error due to object motion into one of the elements of relative phase shift, it also introduces a new error because in a complex factory environment, external vibration, electromagnetic interference to the motor, and other factors will affect the uniform motion of the object to be measured, making it difficult to ensure that the relative phase-shift step length is uniform. The traditional three-step phase shift, twelve-step phase shift and other methods need to be used in the case of ensuring that the phase shift step is uniform, which requires a method that can adapt to the random phase

shift step. As a result, this paper proposes a network model based on RPSNet to accommodate the measurement of random phase-shift step length. The model uses CycleGAN[6] as the network framework, in which the generator uses the AIR2U-net network model, the AIR2U-net model is based on the U-net[7] network model, and the IRR module and Attention Mechanism[13] proposed in this paper are used in order to highlight the key features and inhibit the non-key features; while the selection of the discriminator needs to be compatible with the discriminator needs to be adapted to the generator, and after experimental comparison, this paper chooses to use multilayer convolutional neural network to realize the discrimination of true and false images. Finally, this method is applied to the 3D reconstruction scene of the actual factory workpiece and achieves high quality reconstruction effect.

2 Proposed approach

The dynamic structured light measurement model of RPSNet proposed in this chapter is shown in Fig. 3, which is based on CycleGAN network model. CycleGAN is an unsupervised machine learning algorithm that can be used for image generation and transformation of unpaired pictures. The generator of the RPSNet model uses AIR2U-net network model proposed in this paper, which uses U-net as the AIR2U-net network model can output the depth image of the object to be measured based on multiple phase-shifted images as input. The discriminator is implemented using a multi-layer convolutional neural network, which is trained to determine whether the input image is the depth map output by the generator, or the real depth map based on a single input image. The biggest advantage of the RPSNet model is that it can reconstruct the object in 3D in a dynamic scene and adapt to the random phase-shift step size.

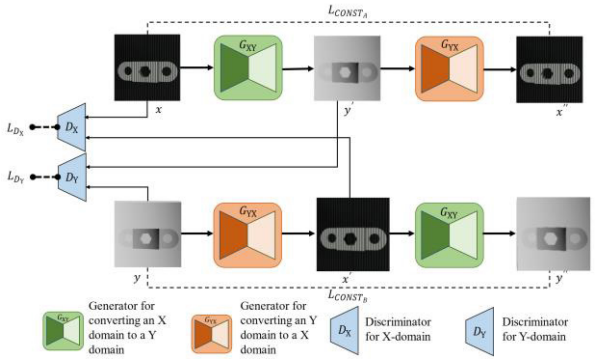

Fig. 3. RPSNet model architecture

As can be seen from Fig. 3, the whole RPSNet model can be divided into two parts, the first part is that the raster map x of the measured object is converted into the depth map y' of the measured object by the generatorG_{XY}, and then the raster map x is generated by the generator G_{YX}, the results of the two conversions of the process, x' and y', will be discriminated as true or false using the discriminators D_X and D_Y respectively, and the generator calculates the confrontation loss based on the generator calculates the adversarial loss based on the output of the discriminators and thus adjusts the output. In addition, to prevent the generator's output from being too aggressive and losing the features of the original input image, a consistency L_{CONST_A} is added in the process of generating the image. The second part is like the first part, which converts the depth map y to raster map x' and then to depth map y''. This process also requires constraints using discriminators and consistency loss functions. By alternating the training of the above process, eventually the whole network model can learn the key features and output high-quality depth images.

2.1 Generator network model

The generator network in the RPSNet model uses the AIR2U-net network model. The architecture of the AIR2U-net network model is shown in Fig. 3. This network model is based on the U-net model as the network architecture and incorporates the attention mechanism and IRR module to improve the model performance. The whole network obtains the feature map through the feature extraction and downsampling operation of the IRR module in the encoder, then through the feature extraction and upsampling process of the IRR module in the decoder, and through the hopping connection, the feature map of the encoding process is added to the decoding process through the attention mechanism to get the final output result. The addition of the attention mechanism enables the whole network to focus on the target features and suppress irrelevant features, which greatly enhances the performance of the network. The proposed IRR module enables the whole network to feel the multi-scale information and enhance the effective features cyclically, so that the whole model has a strong capability of recognizing object features.

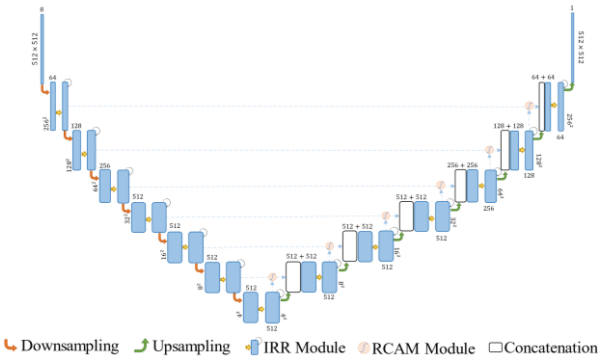

Fig. 3. AIR2U-net model architecture

2.1.1 IRR module

In traditional U-net networks, both the encoder part and the decoder part use simple convolution or inverse convolution to extract features. Due to the diversity and complexity of image applications, simple convolution can no longer meet the requirements of various applications for accuracy, speed, and other performance indicators. In this paper, for the application of dynamic structured light in factory scenarios, the IRR module is designed to improve the feature extraction operation in

U-net, which ensures that the model can enhance the target features and inhibit irrelevant features while minimizing the parameters of the network to achieve a higher detection speed.

The structure of the IRR module is shown in Fig. 4. The IRR module uses the Inception-ResNet[8] module as the base network framework and adds the Recurrent Convolution Block (RCB) to enhance the features. In the design of the IRR module, the operation of the recurrent convolution gives the model the ability to recognize the object features, which improves the detection accuracy of the model. As can be seen from the figure below, the input of the IRR module is divided into four ways to be sent to the residual connection, 1×1 RCB, 3×3 RCB, and 5×5 RCB, and in order to reduce the computation brought by the model parameters, the outputs of the three RCB modules are all input into the 1×1 bottleneck layer, and the output of the bottleneck layer will be input into the residual connection module along with the inputs of the IRR module, and the final output of the whole IRR The final output is the result of the whole IRR processing.

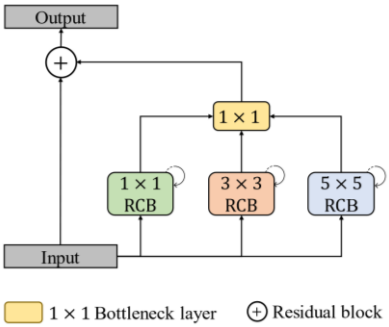

Fig. 4. IRR model architecture

The residual mechanism[9] in the IRR module can be the whole module to avoid the problem that as the number of layers of the network increases, the data may be lost and distorted in the process of transmission of information, resulting in a decrease in the accuracy of the network. Specifically, the residual mechanism, by introducing cross-layer connections, allows the data to be passed directly to the subsequent layers across several layers, thus solving the problem of gradient disappearance and gradient explosion in deep neural networks, in addition, the residual module can also accelerate the training of the network. Because the training of the network is usually realized by back propagation algorithm, and the cross-layer connection of the residual module can make the back propagation algorithm pass the gradient information more easily, thus accelerating the training process of the model.

A multi-scale RCB module is also introduced in the IRR module, which is a mechanism that can utilize RCB modules with different convolutional kernel sizes to extract features of different sizes, and then fuse these features to improve the sensory field and feature expression of the network. Meanwhile, the use of 1×1 bottleneck layer for channel dimension reduction and dimension upgrading can effectively reduce the number of parameters in the network to avoid overfitting sending and reduce the computational burden.

The structure diagram of the RCB module mentioned above is shown in Fig. 5, the input of this module will be used as the input of each convolutional layer, and the output of the convolutional layer of the previous layer will also be input to the convolutional layer of the next layer if the number of loops, t>0, where t denotes the number of loops of convolution, which is generally set to 2.

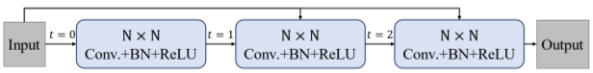

Fig. 5. RCB model architecture

The RCB module utilizes the idea of circular convolution to accumulate the features extracted from the network efficiently, which greatly improves the model's ability to learn the features can be expressed by Equation 1:

$$O_{ijk}^{s}(t) = \left(W_k^f\right)^T \times x_s^{f(i,j)}(t) + \left(W_k^r\right)^T \times x_s^{r(i,j)} + b_k \quad (1)$$

where i, j are used to locate the pixels of the input feature map, k denotes the kth channel of the feature map, s denotes the size of the convolutional kernel, t denotes the number of loops, f denotes the feature map for the input of each standard convolutional layer, and r denotes the feature map for the input of the whole RCB module. $O_{ijk}^{s}(t)$ denotes the output of the standard convolutional layer in the RCB module, $x_s^{f(i,j)}(t)$ and $x_s^{r(i,j)}$ are the inputs of the standard convolutional layer and the inputs of the RCB module, respectively, W_k^f and W_k^r are the weights of the standard convolutional layer with respect to the two inputs, and b_k is the corresponding bias of that standard convolutional layer.

This output will then be fed into the standard ReLU activation function, formulated as shown in Equation 2 below:

$$F(x_s, w_s) = f\left(O_{ijk}^{s}(t)\right) = \max\left(0, O_{ijk}^{s}(t)\right) \quad (2)$$

where $F(x_s, w_s)$ represents the output of the RCB module with a convolution kernel size of s. After multiple scales of convolution kernels respectively perform the same operations described above on the inputs, the outputs are fed into a 1×1 convolution operation, and finally, the input x and the output of this convolution operation are residually spliced, and the result of the splicing is used as the output of the entire IRR module. The formulaic expression is shown in Equation 3:

$$X_{l+1} = B\left(F\left(x_{s_1}, w_{s_1}\right)\right) \circ B\left(F\left(x_{s_3}, w_{s_3}\right)\right) \circ B\left(F\left(x_{s_5}, w_{s_5}\right)\right) + X_l \quad (3)$$

where X_{l+1} and X_l are the inputs of the X_{l+1} and lth layers, respectively, X_(l+1) is also the output of the lth, $F\left(x_{s_1}, w_{s_1}\right)$、$F\left(x_{s_3}, w_{s_3}\right)$、$F\left(x_{s_5}, w_{s_5}\right)$ and are the outputs of the RCB module with convolution kernel sizes of 1 × 1, 3 × 3, and 5 × 5, respectively. outputs, $B(\cdot)$is the Bottleneck Layer function, and ◦ denotes the matrix splicing along the depth direction (Filter Concatenation).

2.1.2 RCA module

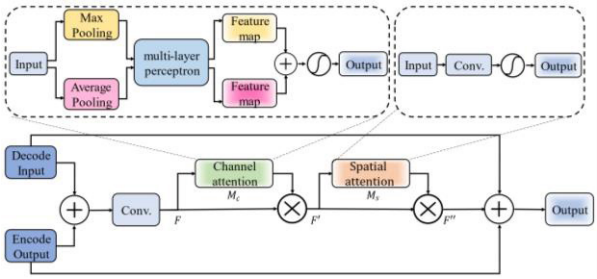

Fig. 6. RCAM model architecture

The structure diagram of the RCAM (Residual Convolutional Attention Module) module network proposed in this paper is shown in Fig. 6, The model takes the outputs of all levels of encoding and decoding modules as inputs, which are spliced and fed into the convolutional layer, and then processed using the channel attention mechanism and spatial attention mechanism, and finally the outputs are spliced with the encoded outputs to obtain the outputs of the entire RCAM module. The channel attention mechanism firstly does global maximum pooling[11] and average pooling[12] for each input channel, and then uses a fully connected layer with shared weights to compute the weights of different channels for the pooled feature maps, obtaining the weights of the two feature maps in different channels, summing up the weights, and obtaining the final weights through the activation function. The channel attention mechanism mainly focuses on the importance of image features in the channel dimension, and dynamically adjusts the channel weights to enhance the feature representation of important channels and suppress the feature representation of unimportant channels. And the channel attention mechanism can effectively reduce the number of parameters of the network and improve the computational efficiency of the model while enhancing the expressive ability of the network. The spatial attention mechanism, on the other hand, needs to compute the weight values under different spatial coordinates in the 2D feature map after dimensionality reduction of the channels of the feature map. Specifically, the spatial attention mechanism first uses maximum pooling and average pooling to pool the channels, the two feature maps obtained are spliced and then the weights are calculated using a convolutional network, and finally the obtained weights are output through the activation function. The spatial attention module is mainly used to improve the network's attention to features at different spatial locations, through a convolution operation and a global average pooling to learn the importance weight of each spatial location to enhance the representation of important spatial locations. RCAM combines the advantages of the two attention mechanisms to effectively improve the model's learning ability. The whole process of the RCAM module can be expressed using the equation 4:

$$\begin{cases} F = Conv(f_d + f_e) \\ F' = M_c(F) \otimes F \\ F'' = M_s(F') \otimes F' \\ O = F'' + f_d + f_e \end{cases} \tag{4}$$

where f_d, f_e denote the feature maps of the decoder and encoder outputs, respectively, Conv(·) denotes the convolution operation on the inputs, $M_c(\cdot)$, $M_s(\cdot)$ are the channel-attention and spatial-attention operations on the inputs, respectively, $\otimes$ denotes the element-by-element multiplication of the left and right inputs, and O denotes the final output of the module.

2.2 Discriminator network model

In the RPSNet model, the discriminator uses a simple multilayer convolutional network structure, and its network architecture diagram can be shown in Fig. 7, The input of this discriminator is the output of the generator G(x) or the label y of the dataset, and after a few convolution operations, a feature map with 512 channels is obtained, which is then fed into a fully connected neural network, which outputs a feature map with 1 channel, and finally a scalar is obtained after an average pooling operation. The discriminator can determine whether the input data is real or not after learning. The output of the generator will calculate the loss function with the label and update the parameters of the generator after backpropagation, and the updated generator again outputs the generated image to the discriminator for true or false judgment, and so on. This adversarial learning process can help the generator to continuously learn the feature distribution of the real data to generate more realistic data.

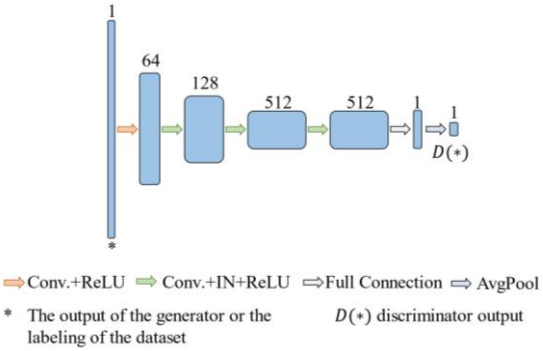

Fig. 7. Discriminator Network Architecture

The RPSNet proposed in this paper is a kind of CycleGAN, and the discriminator of CycleGAN is different from the discriminator of ordinary GAN[14] in that it needs to judge the relationship between two sets of data at the same time. Specifically, the discriminator of RPSNet has two discriminators, one for discriminating the real image and the generated image, and the other for discriminating the transformation result of the real style image and the generated image. The structure of these two discriminators is the same, both are deep neural networks composed of multiple convolutional and fully connected layers, only that both act on different image domains respectively. It should be noted that in CycleGAN, the discriminator is not used to directly optimize the generator's loss function but acts as an adversarial target for the generator.

2.3 Loss function

RPSNet is trained according to the loss function, and since RPSNet is divided into two parts, the generator and the discriminator, the loss function of this model also consists of two parts. Where the loss function of the generator contains three parts which are Adversarial loss, Cycle consistency loss and Identity loss. The formulaic expression is shown in Equation 5. Two generators are used in RPSNet to perform the conversion between the original domain and the target domain, so the three loss functions mentioned above have two components, while the total loss function has six components. L_{GAN_X} and L_{GAN_Y} denote the adversarial generator functions for the X and Y domains, respectively. L_{cycX} and L_{cycY} denote the adversarial generator functions for the X and Y domains, respectively. L_{cyc_X} and L_{cyc_Y} denote the cyclic consistency loss functions in the X and Y domains, respectively, and λ_X and λ_Y denote the weights of both, respectively. L_{idt_Y} and L_{idt_Y} denote the constant loss functions in the X and Y domains, respectively, and μ_X and μ_Y denote the weights of both, respectively.

$$L_G = L_{GAN_X} + L_{GAN_Y} + \lambda_X L_{cyc_X} + \lambda_Y L_{cyc_Y} + \mu_X L_{idt_Y} + \mu_Y L_{idt_Y} \quad (5)$$

The purpose of the adversarial loss function is to make the generator produce samples that are like the real samples so that the discriminator cannot distinguish between the generated samples and the real samples. The purpose of cyclic consistency loss is to ensure that the image generated through the generator can be reduced to the image in the original domain. Whereas, the purpose of the constant loss is to maintain the overall characteristics of the original input image, such as features like hue, lightness, and darkness. The loss function of the discriminator is used to determine the authenticity of the output image of the generator and is an important part of the loss function in the whole RPSNet model.

3 Experiments and results

3.1 Datasets

To build effective deep learning models, in addition to excellent computational modules, a large amount of training data is also essential, but for many deep learning methods based on FPP methods, obtaining enough training data is a tricky problem. Fortunately, in recent years, computer graphics has made great progress, and its simulation of some real scenarios has reached a point where people can hardly distinguish them. As a result, this graphical technique is beginning to be used in the fields of data enhancement and data generation, and many experiments have shown that the models trained in this way can also be applied to real-world scenarios. This provides inspiration for this paper to solve the problem that the RPSNet model requires many data sets for training, and the collection of data sets becomes efficient and convenient by building a virtual FPP system using computer graphics-based simulation and modeling software. Specifically, this section will explain in detail how to use the 3D scene creation software Blender to build the virtual FPP system and show some data samples made based on this virtual system.

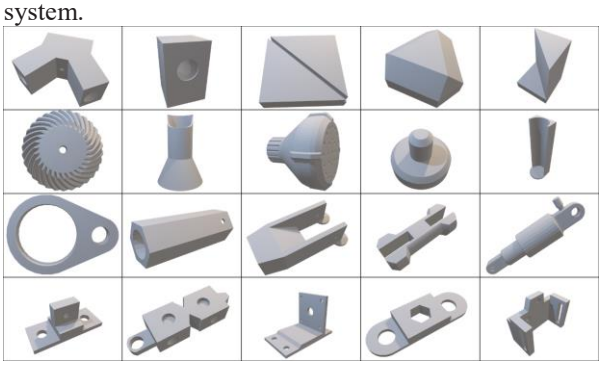

Fig. 8. Partial model of the Thing10k 3D model dataset

Currently, the commonly used 3D model datasets include ModelNet[15], ShapeNet[16], ABC[17], Thingi10K[18], and so on. When selecting the dataset, this paper mainly considers two points, firstly, the effective working distance of the FPP system under visible light, which is generally 1~2 meters, so the volume of the 3D model should not be too large; secondly, the application scenario of the model in this paper is in the industrial production process, and the selected model should be similar to the common workpieces in the industrial scenario. Based on the above two points, this paper chooses the Thingi10K dataset as the 3D model dataset used in this paper, which contains a variety of 3D models of many common objects such as workpieces, sculptures, vases, and so on, some of which are shown in Fig. 8, The diversity and scale of these models help to generate large and diverse data samples to train models with more generality and generalization capabilities.

The effect of virtual FFP system construction is shown in Fig. 9. Blender is an open-source 3D scene creation software, and it can batch process images through Python scripts. Using Blender simulation software, real-world scenes can be simulated in a virtual environment. In this virtual environment, by placing two virtual cameras and a virtual projector, and setting the projector to project sinusoidal stripes projected onto the object, the deformed stripes after the object height modulation are captured by the left and right cameras, and the raster maps captured by the virtual cameras can reach the degree of fakeness, and the real virtual system can be realized for the simulation of the real FFP system.

Using the virtual FFP system, the dataset required for model training can be generated. In the virtual FPP system, the projector is first needed to project the raster pattern, which can be achieved by setting the shader node. The shader node is a module in Blender that is used to color and render the model, which allows for different rendering effects. As shown in Fig. 10, a node named "Image Texture" needs to be set to select the source of the image to be projected by the projector. This node needs to select a picture with sinusoidal stripes as the input of the projected picture. In this way, the projection of raster pattern can be realized in the virtual FPP system, and the corresponding edge image and depth image can be generated.

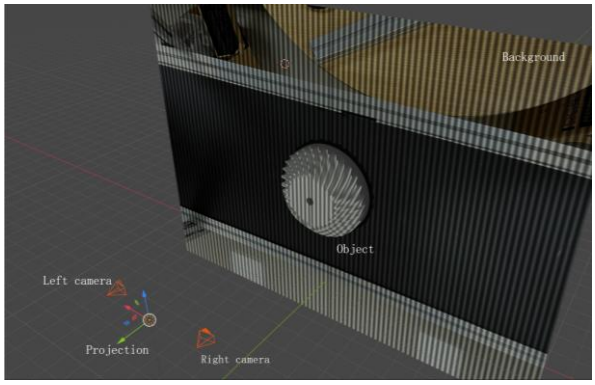

Fig. 9. Virtual FFP Scenario

The effect of virtual FFP system construction is shown in Fig. 9. Blender is an open-source 3D scene creation software, and it can batch process images through Python scripts. Using Blender simulation software, real-world scenes can be simulated in a virtual environment. In this virtual environment, by placing two virtual cameras and a virtual projector, and setting the projector to project sinusoidal stripes projected onto the object, the deformed stripes after the object height modulation are captured by the left and right cameras, and the raster maps captured by the virtual cameras can reach the degree of fakeness, and the real virtual system can be realized for the simulation of the real FPP system.

Using the virtual FFP system, the dataset required for model training can be generated. In the virtual FFP system, first a projector is needed to project the raster pattern, which can be achieved by setting up a shader node. The shader node is a module in Blender that is used to color and render the model, which allows for different rendering effects. As shown in Fig. 10, a node named "Image Texture" needs to be set to select the source of the image to be projected by the projector. This node needs to select a picture with sinusoidal stripes as the input of the projected picture. In this way, the projection of raster pattern can be realized in the virtual FPP system, and the corresponding edge image and depth image can be generated.

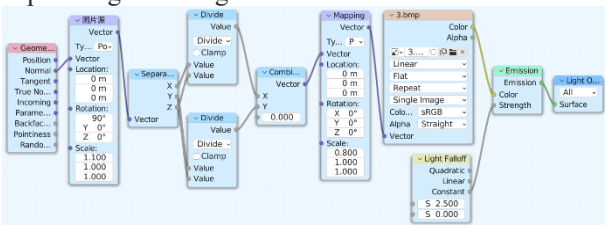

Fig. 10. Blender Shader setup

Subsequently, the raster map of the image captured by the camera is rendered. Specifically, the "image" or "depth" attributes in the "rendering layer" of the compositing node are output to the compositing node after the specification node to render a raster image and a depth image, respectively. depth image. The compositing node setup is shown in Fig. 10.

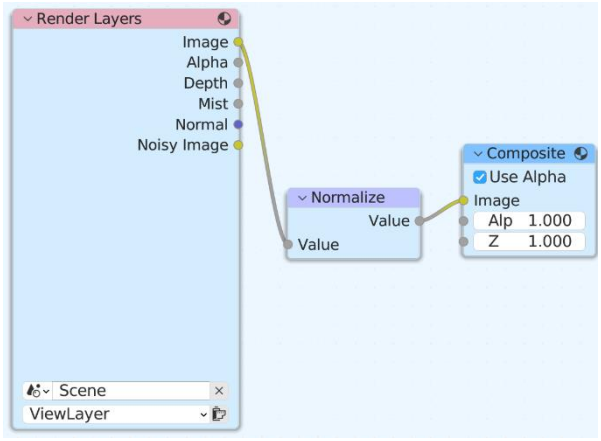

Fig. 11. Blender compositor setup

To make the virtual FFP system closer to the real factory environment, this paper adopts several methods to simulate the real environment. For example, to further enrich the dataset, this paper rotates the models in the 3D model dataset several times along all directions, which also simulates the messy placement of real workpieces. To enhance the realism of the virtual FFP system, a background plate is added to the scene, and photos of the real factory assembly line are rendered onto the background plate as a texture. The dataset obtained by some of the simulation means mentioned above can be maximally close to the dataset collected from the real environment, and the network model will be able to learn the input features better and avoid overfitting during training.

With the above-mentioned settings related to building an FFP virtual system in Blender, it is possible to capture many simulation datasets. However, in practice the above graphical setup is extremely cumbersome and inefficient. For example, to train the model to learn a grating pattern with random phase shift steps, it is necessary to randomly displace the object under test in a single direction, which is undoubtedly a huge amount of work if manually adjusted in the above graphical interface. Blender provides Python scripts to build the entire simulation system, which is extremely convenient for the user. Therefore, the above FFP virtual system is built using Python scripts, and some of the data sets generated are shown in Fig. 12. To train and evaluate the model correctly and efficiently, the dataset is divided into training dataset and testing dataset according to the ratio of 3:1 in this paper.

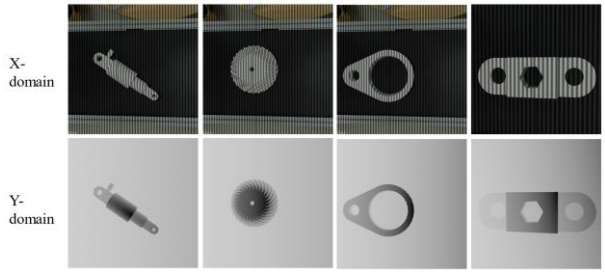

Fig. 12. Partial result of the datasets generated.

3.2 Comparison experiment

The experiments in this section aim to verify the effectiveness of the RPSNet model proposed in this paper. For this purpose, the method of Gray code combined with three-step phase shift, the method of multi-frequency outlier combined with twelve-step phase shift, and the method proposed in this paper are selected for comparison experiments. Since there is no universally applicable method with high reconstruction quality for the measurement of dynamic objects, the reconstruction algorithm of multifrequency aberration combined with twelve-step phase-shifting is chosen as Ground Truth, which can better inhibit the errors introduced by nonlinearities in projection, reflections of objects, etc., and thus ensures the reconstruction of 3D objects with high quality. 3D objects.

The specific reconstruction algorithms used for each of the schemes in the comparison experiments and the relevant settings for the corresponding experiments are described below:

• 3Step&Gray: This scheme uses a three-step phase-shift computation to put the wrapped phase and projects an additional grey code pattern for phase labelling, which is used for the resolution of the wrapped phase. In this scheme, the number of grey code patterns is set to 4, i.e., up to 16 cycles in the field of view are supported, which is sufficient for an image of 512-pixel size. The specific projection pattern is encoded according to the grey code to minimize the effect of the object indicating reflection on the grating pattern. In practical applications, considering the reflected light on the surface of the object under test and the uneven light in the environment, it is also necessary to project two additional patterns of all-black and all-white, which are used to normalize the luminance, so that it is easy to judge the brightness of pixel points under different light environments. In summary, the program needs to project a total of 3+4+2=9 patterns.

• 12Step&MultiFreq: In this scheme, the twelve-step phase shift is first used for the calculation of the wrapping phase, and the multi-frequency outlier method is used for phase unwrapping. In order to ensure that the total period obtained by multi-frequency outlier can cover the whole field of view, this scheme uses the algorithm of three-frequency outlier, which are 25 pixels, 27 pixels, and 29 pixels as a period of the sinusoidal pattern for the unwrapping calculation, and a total of 12×3=36 raster patterns need to be projected due to the twelve-step phase-shift used for the calculation of the wrapping phase.

• RPSNet: this scheme uses the dynamic structured light measurement model RPSNet adapted to the random phase shift step size proposed in this paper, which directly converts the input grating pattern into a depth image without going through the process of calculating the wrapped phase as well as unwrapping the phase.

• GT: In order to be able to highlight the reconstruction effect of individual programs, this comparison experiment adopts the reconstruction algorithm of twelve-step phase shift with multi-frequency outlier, and measures the object under the condition that the object is kept static, and the relevant experimental setups of this reconstruction algorithm and the 12Step & MultiFreq reconstruction program are kept in line with each other.

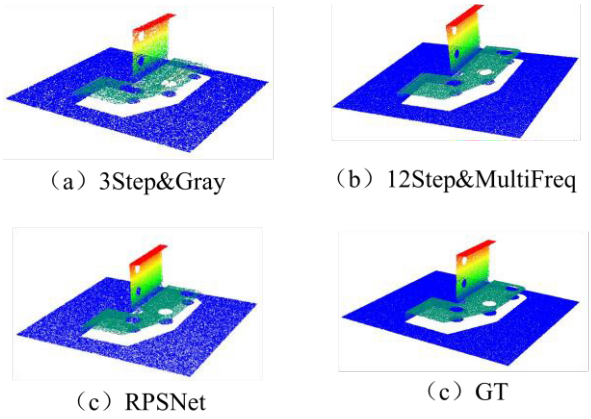

Fig. 13. Comparison of experimental result

Fig. 13 shows the results of this comparison experiment. From Fig. 13 (a), the 3Step&Gray structured light reconstruction scheme has uneven planes in the point cloud and the whole point cloud is sparse due to the existence of motion errors when facing the measurement of objects in a dynamic scene, and the reconstruction accuracy of the whole point cloud is also lower. This is since the three-step phase shift algorithm is unable to handle errors introduced by the nonlinearity of the projector, reflections on the surface of the object, and so on. Fig. 13 (b) shows the reconstruction results of the 12Step&MultiFreq scheme, although this scheme can have a better suppression effect on the environment as well as the equipment errors, but due to the large number of phase-shifting steps, it leads to a larger motion error. The RPSNet model proposed in this paper can better overcome the shortcomings of the above two reconstruction schemes, and its final reconstructed point cloud map is shown in Fig. 13 (c). It can be seen in the figure that although the point cloud data reconstructed by this scheme is not as dense as the twelve-step phase shift, in contrast, the reconstructed point cloud data of this scheme can provide accurate reconstruction of the basic features of the workpiece.

4 Conclusion

To realize 3D reconstruction of workpieces in motion, this paper proposes to utilize the idea of relative phase shift to convert the object motion factor, which originally introduces measurement error, into a necessary condition for relative phase shift. However, the resulting non-uniform phase shift will lead to the failure of the traditional solution phase method, so this paper proposes the RPSNet dynamic structured light measurement model adapted to random phase shift. The model is based on CycleGAN, using the AIR2U-net model proposed in this paper as the generator and the multilayer convolutional neural network model as the discriminator. The AIR2U-net model can generate depth maps based on the input raster image, and the depth maps generated by the generator and the real depth maps

will be inputted into the discriminator, which will learn the true and false depth maps to improve the discriminative ability, and the output of the discriminator is also used as the generator. The output of the discriminator is also used as the training input of the generator, and the two are trained against each other until the generator can output the real depth map with the fake one. For the model to obtain sufficient learning features, the network model often needs many data sets for training, and the collection of data sets in industrial scenarios is a major difficulty. To address this problem, this paper uses the simulation software Blender to produce the dataset and completes the dataset acquisition by building a virtual FPP system in this software and selecting the workpiece dataset from the Thing10k dataset as the object under test. In the acquisition process, we also set up a photo of the factory environment as the background of the workpiece and set up a random moving distance of the workpiece to simulate the data acquisition process in the real environment as much as possible. The above acquisition process can be realized by using the python script interface provided by Blender, and this way of acquiring data sets can ensure the validity of the data sets under the premise of improving the acquisition efficiency. Finally, the method proposed in this paper is compared with the three-step phase-shift with gray code method and the twelve-step phase-shift with multi-frequency outlier method, and the experimental results show that the dynamic structured light measurement method proposed in this paper can realize high-quality reconstruction of workpieces in dynamic scenes.

References

1. Gorthi, S. S., & Rastogi, P. (2010). Fringe projection techniques: whither we are? Optics and lasers in engineering, **48**(2)
2. Qian J, Feng S, Tao T, et al. Deep-learning-enabled geometric constraints and phase unwrapping for single-shot absolute 3D shape measurement[J]. Apl Photonics, **5**: 046105, (2020).
3. Spoorthi G E, Gorthi R K S S, Gorthi S. PhaseNet 2.0: Phase unwrapping of noisy data based on deep learning approach[J]. IEEE T Image Process, **29**: 4862-4872, (2020).
4. Yang H, Carlone L. A polynomial-time solution for robust registration with extreme outlier rates[J]. Robotics: Science and Systems, (2019).
5. Zhang L, Chen Q, Zuo C, et al. High-speed high dynamic range 3D shape measurement based on deep learning[J]. Opt Laser Eng, 134: 106245, (2020).
6. Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE ICCV (pp. 2223-2232) (2017).
7. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. MICCAI 2015: 18th International Conference, Proceedings, Part III **18** pp. 234-241, **(**2015).
8. Szegedy, Christian, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI conference on artificial intelligence, vol. **31**, no. 1. (2017).
9. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. CVPR, pp. 770-778. (2016).
10. Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. AISTATS, pp. 315-323. (2011).
11. Lin, Min, Qiang Chen, and Shuicheng Yan. Network in network. arXiv preprint arXiv:1312.4400 (2013).
12. Rota Bulo, Samuel, Gerhard Neuhold, and Peter Kontschieder. "Loss max-pooling for semantic image segmentation." CVPR, pp. 2126-2135. (2017).
13. Vaswani, Ashish, et al. Attention is all you need. Advances in neural information processing systems **30** (2017).
14. Goodfellow, Ian, et al. Generative adversarial nets. Adv. neural inf. process. syst **27** (2014).
15. Wu, Zhirong, et al. 3d shapenets: A deep representation for volumetric shapes. CVPR. (2015).
16. Wu, Zhirong, et al. 3d shapenets: A deep representation for volumetric shapes. CVPR. (2015).
17. Koch, Sebastian, et al. Abc: A big cad model dataset for geometric deep learning. CVPR. 2019.
18. Zhou, Qingnan, and Alec Jacobson. Thingi10k: A dataset of 10,000 3d-printing models. arXiv preprint arXiv:1605.04797 (2016).