IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 28, 2020, accepted July 22, 2020, date of publication July 27, 2020, date of current version August 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3012207

Value of Virtual Reality Technology in Image Inspection and 3D Geometric Modeling

LONGYU LU, JINKAI MA, AND SHUYING QU
School of Civil Engineering, Yantai University, Yantai 264005, China

Corresponding author: Shuying Qu (qsytmx@ytu.edu.cn)

This work was supported by the Graduate Innovative Education Funds of Shandong Province under Grant SDYJG19067.

ABSTRACT Aiming at the poor expressive ability of image statistical information during the reconstruction process of traditional 3D image reconstruction method based on virtual reality technology, resulting in low accuracy of 3D image after reconstruction, a new image detection and 3D image reconstruction based on virtual reality technology are studied method. This paper first proposed a new two-level cascade convolutional neural network structure. The first level of the network predicts target positioning based on the image-level labels of the training image, generates a bounding box of the target in the original image, and generates a cropped image. The cropped image is input to the second-level network. The cropped image may contain areas where the target is stuck in the original image. Level 2 networks only use the adhesion area as training data. Secondly, the visualization software development platform and virtual reality 3D image processing software are selected as the platform for 3D image reconstruction. After the original image is imported into the computer through data input and file analysis steps, the original image is detected. The detected image is in the virtual in the real software, the bounding box method is first used to construct the three-dimensional data field of image reconstruction, and the three-dimensional direct volume of the image is drawn according to the three-dimensional data field of image reconstruction. Preferably, the three-dimensional image reconstruction output formula is obtained through the three-dimensional image direct volume to realize the three-dimensional image reconstruction based on the virtual reality technology. The simulation results show that the method proposed in this paper can effectively detect images. The average traversal coverage of 3D image reconstruction is up to 0.979, and the reconstruction accuracy is higher than 0.97.

INDEX TERMS Image classification, distributed network representation learning, deep learning, neighbor reconstruction.

I. INTRODUCTION

Virtual reality technology is a high-tech technology born with the development of science and technology. It refers to the formation of interactive three-dimensional dynamic simulation that simulates the real environment through the computer, enabling users to realize the immersive virtual reality effect through the computer [1]–[3]. The image has irregularities and is easily interfered by noise during the reconstruction process. Therefore, 3D image reconstruction is more difficult [4], [5]. In recent years, with the rapid development of technologies such as virtual reality and 3D printing, the application of 3D models has become more and more widespread [6]–[9]. Traditional three-dimensional scene modeling methods mainly include modeling techniques based on geometric modeling and modeling techniques that use instrumentation to obtain scene depth information [10]–[11]. The instrument-based modeling technology obtains the depth information of the scene by using laser, infrared and other ranging technologies to realize the three-dimensional modeling of the target object [12]. The modeling method based on a single image avoids complex image matching and constraints, data acquisition is more convenient and extensive, and has a wider application field, which is the current research hotspot [13]–[14].

Kansal and Mukherjee [15] proposed a method based on single view metrology. By using the geometric information in the image, such as the vanishing line of the reference surface

The associate editor coordinating the review of this manuscript and approving it for publication was Zhihan Lv.

and the vanishing point of the rays not parallel to the reference surface, the scene was inferred the affine structure can be modeled from a single image without the need to calibrate the camera. Baldos *et al.* [16] proposed a layered modeling method. This method divides the image into different layers according to depth, and constructs a three-dimensional scene corresponding to the image layer by layer, depending on the user to label the depth of each layer. In view of the common symmetry in buildings, He *et al.* [17] reconstructed a three-dimensional building model with texture information from a single image. Using the symmetry of the building, a camera calibration method based on the bevel model was proposed. Kaur and Gandhi [18] performed wavelet transformation on the input image and then used entropy to reduce the dimensionality of the feature. At the same time, the spider web graph was used to further reduce the input feature. Then they used a probabilistic neural network to detect the image. Song *et al.* [19] changed the classifier, used kernel support vector machine, and used particle swarm optimization algorithm to train KSVM. This document compares the proposed PSO-KSVM with the optimized classification method, and the classification performance of the improved classifier is obviously improved. In recent years, with the development of convolutional neural network (CNN) [20], the progress in the field of computer vision has been promoted. The CNN model is by far one of the most effective models in the field of visual recognition. CNNs have been successfully applied to image detection problems under strong supervised learning [21], [22]. CNN has a large number of adjustable parameters and requires a large number of pixel-level tags to improve the practicality and robustness of the algorithm. However, strongly supervised training data is only provided in a small number of public datasets. If the task annotation category is increased later, the image must be re-labeled. In addition, for some practical application areas such as medical imaging, data needs to be manually annotated, and certain professional knowledge is required. The annotation workload is heavy and subjective, and the accuracy and accuracy of the annotation cannot be guaranteed. Cai *et al.* [23] proposed a new algorithm for class activation mapping (CAM), which not only outputs predictions, but also visualizes the most interesting target areas, which significantly improves the ability to detect image targets. Currently, weakly supervised image detection tasks are mainly based on the CAM algorithm to obtain image positioning and classification information. The output form of the positioning is a heat map. Detection is to find visually consistent images in the training image. Usually the detection task may contain multiple targets in a single image, resulting in multi-target adhesion problems.

Although the CAM algorithm can improve the accuracy of target detection, the problem of multi-target adhesion is difficult to solve. In order to make better use of the characteristic information implied in the data, this paper proposes an image detection and three-dimensional image reconstruction method based on virtual reality technology. Specifically, the technical contributions of this article can be summarized as follows:

First: In this paper, four gradient images in the horizontal direction, vertical direction and two diagonal directions are used as edge information, and multi-directional gradient information is used as the input of CNN. The feature information of the 4 gradient directions is input to the convolutional neural network to make the extracted image features more effective, and a batch normalization algorithm is added before the classification layer, which further reduces the error rate of image recognition.

Second: Secondly, this paper proposes a new two-level cascade convolutional neural network structure. The first level of the network predicts target positioning based on the image-level labels of the training image, generates a bounding box of the target in the original image, and generates a cropped image. The cropped image is input to the second-level network. The cropped image may contain areas where the target is stuck in the original image. Level 2 networks only use the adhesion area as training data.

Third: Select the Visual C++ visualization software development platform and virtual reality 3D image processing software as the 3D image reconstruction implementation platform, and draw the image 3D direct volume based on the image reconstruction 3D data field.

The rest of this article is organized as follows. Section 2 analyzes related concepts. Section 3 constructs image detection based on virtual reality technology and image 3D geometric modeling. Section 4 carried out simulation experiments. Section 5 summarizes the full text.

II. RELATED CONCEPTS

A. VIRTUAL REALITY TECHNOLOGY

Virtual reality technology mainly refers to a virtual artificial media space constructed with a computer, and all the images presented in the space are virtual forms. But at the same time, it also gives people a certain sense of reality. It can be either a simulation of the real world or a conception of the virtual world. Therefore, it has been widely used in my country's construction, entertainment, military and medical areas [24].

Virtual reality technology mainly has three significant characteristics of authenticity, interactivity and virtual imagination [25], Virtual reality technology module as shown in Figure 1.

Authenticity: Through the combination of computer image technology and three-dimensional modeling technology, the images in the virtual space are presented in front of people's eyes in three-dimensional and three-dimensional forms.

Interactivity: Through the use of virtual interactive devices, users can interact with the virtual world and perform interactive operations.

Virtual imagination: The establishment of the virtual world is mainly based on people's requirements and imagination of the virtual world, and finally the concept is presented in

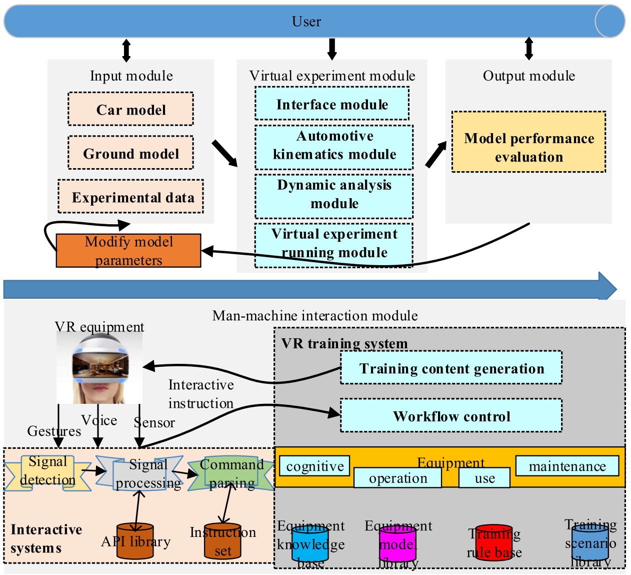

FIGURE 1. **Virtual reality technology module.**

a virtual way through computer technology. The car model uses virtual reality technology, which can be simulated using a display screen to make the user feel as if they were in a real driving scene.

B. IMAGE-BASED THREE-DIMENSIONAL MODELING

Image-based three-dimensional modeling technology is mainly image-based graphics algorithms, and algorithms in this area have achieved outstanding results in many fields [26], [27]. The scene modeling of virtual reality technology mainly uses the method based on the panorama, as shown in Figure 2.

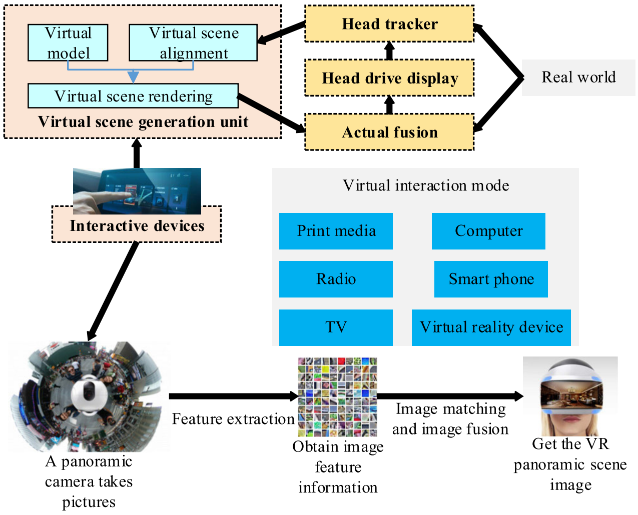

FIGURE 2. **VR interaction of panorama.**

The goal of image-based 3D reconstruction is to obtain 3D geometric models from sequence images. The process of 3D reconstruction is divided into the following steps:

1) Raw data acquisition, usually using a handheld camera or video camera to take a series of images around a static scene.

2) Pre-processing of image sequence (including image selection, noise removal, lighting processing, etc.).

3) Obtain a three-dimensional model of the scene from the image sequence.

4) Finally, draw a 3D model or export a 3D model in a special format.

In general, 3D reconstruction emphasizes the work of step 3, the four main parts of 3D reconstruction, and their processes, as shown in Figure 3.

Unlike traditional images, VR panoramic images are not characterized by the same data format when displayed, transmitted, and stored. Some projection transformation is required when viewing VR panoramic images. During encoding transmission, the VR panoramic image may be converted to other projection domains, for example, cubic projection [28]. In different projection domains, the sampling frequency and spatial structure of VR panoramic images will change. The sampling frequency and spatial structure of the image directly affect human visual attention to the image. Therefore, projection transformation also affects the saliency detection performance of VR panoramic images.

Another significant difference between VR panoramic images and traditional images is the user's viewing behavior. When viewing a VR panoramic image, there is an interaction process between the user and the image. In the process of viewing VR panoramic images, in addition to eye movements, users also have head movements. The user is in the center of the field of view, and can freely move his head to select the scene he wants to see. For the user, the angle of view of the person is limited. During head-up, only the area near the equator of the spherical surface can be seen, and the polar regions of the spherical surface cannot be seen. However, it is uncomfortable to look down or look up for a long time. This results in content near the equator being more easily followed by users.

III. IMAGE DETECTION AND 3D GEOMETRIC MODELING BASED ON VIRTUAL REALITY TECHNOLOGY

A. PROBLEM DESCRIPTION

Given an image set pair $T_l = \{t_{l1}, t_{l2}, \ldots, t_{lN}\}$and an image set pair$g_i = \{x, y, w, h\} \in G = \{g_1, g_2, \ldots, g_N\}$, the image $t_{li}$and the corresponding target $t_{ri}$are labeled as, and N is the number of samples. The symbol x, the symbol y, the symbol w, and the symbol h represent the center coordinates, width and height of the target frame in the image, respectively, and G represents all the marked sample information. Model training includes two processes: candidate region feature extraction and classification prediction, namely:

$$C = \Pr(class_i|object) \tag{1}$$

Among them, C represents the target prediction type. The matching of two convolutional features of the same layer is regarded as the process of optimally solving the correlation

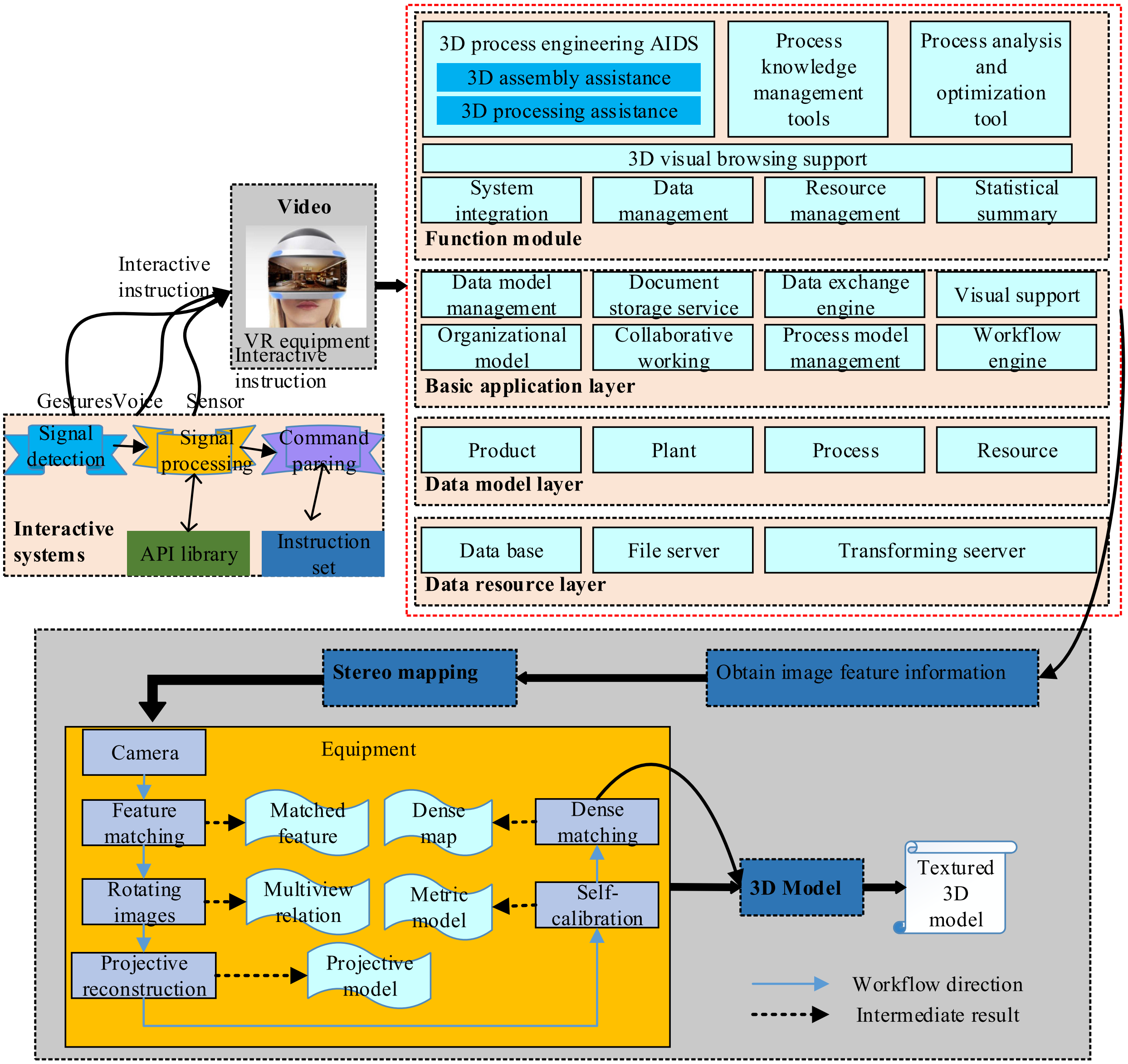

FIGURE 3. The framework of 3D reconstruction technology.

filter w, namely

$$w^* = \arg\min \sum ||w \cdot v_{m,n} - y_{m,n}||^2 + \lambda||w||_2^2 \tag{2}$$

Among them, each sample $v_{m,n}$of the feature vector v corresponds to a Gaussian label$y_{m,n}$, the variable λ is a regular term coefficient, and λ ≥0. The Gaussian label is obtained by formula (3):

$$y_{m,n} = e^{(-\frac{(m-M/2)^2+(n-N)^2}{2\sigma^2})} \tag{3}$$

Among them, the variable σ represents the width of the Gaussian convolution kernel, and variable M and variable N represent the width and height of the feature layer, respectively.

B. MULTI-LAYER FEATURE FUSION SSD STRUCTURE

Convolution is a commonly used algorithm in image recognition, which means that each pixel in the output image is obtained by weighting the pixels of a small area at the corresponding position of the input image. This small area is called the local receptive field, and the weight of the area is called convolution [29]. After the convolution operation is performed on the input image, the offset term is added, and its feature map is obtained through the activation function. The form of the convolution layer:

$$X_j^l = f(\sum_{i \in M_j} X_i^{l-1} \times K_{ij}^l + b_j^l) \tag{4}$$

Among them, the variable 1 is the number of layers. The variable $X_j^l$is the j-th feature map in the 1 layer of the convolution layer. The variableM_j is the receptive field of the input layer. The variable K is the convolution kernel. The variable b is the offset. The variable f is the activation function of the neuron. The activation function selected in this paper is a modified linear unit (ReLU) function. The ReLU activation function can prevent the gradient from disappearing, alleviate the phenomenon of overfitting, and achieve better results.

The down-sampling layer is also called the pooling layer, which divides the image into small and small areas, calculates a value for each area, and then arranges the calculated values in sequence to output a new image. This process is equivalent to fuzzy filtering and can increase the robustness of image feature extraction. The pooling method used in this paper is average pooling.

In this paper, first, the input image is passed through the Sobel operator to obtain four gradient images in the horizontal direction, vertical direction, and two diagonal directions. Then, input 4 multi-layer convolutional neural networks to learn the characteristics of 4 gradient images in different directions. Then, the features in 4 different directions are subjected to randomized feature fusion, and the information after feature fusion has better robustness. In this paper, after randomized feature fusion, batch input standardization is applied to the data input to the classifier, so that the data input to the classifier can be dispersed from centralized to achieve better recognition results. To a certain extent, overfitting can be prevented.

For a 32 × 32 image, the gradient information in four directions, all of which are 32 × 32, can be obtained. For each channel, the image gradient first passes through 8 5 × 5 convolution kernels to obtain 8 28 × 28 feature graphs. Then, sampling was carried out under the pooled size of 2 × 2 to obtain eight 14 × 14 characteristic graphs. After 16 5 × 5 convolution kernels, 16 10 × 10 characteristic graphs were obtained. Finally, 16 5 × 5 feature graphs were obtained by sampling under the pooling size of 2 × 2. Each network can obtain 16 5 × 5 feature graphs. Then, the feature graphs of four channels are randomly fused. The process of stochastic feature fusion is shown in Figure 4.

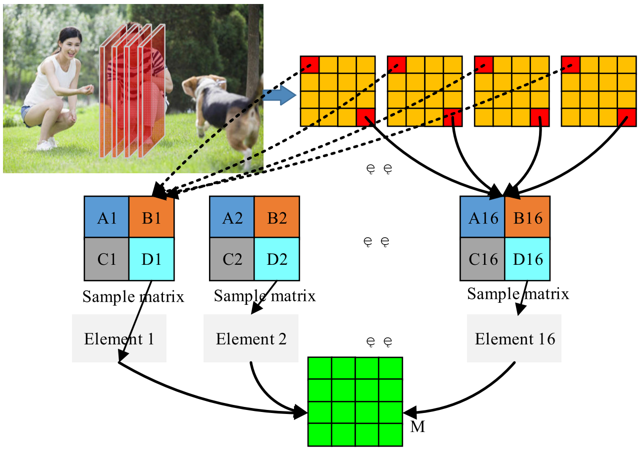

FIGURE 4. Stochastic feature fusion process.

A, B, C, and D respectively represent feature maps of different channels in Figure 4, and the superscript j represents the j-th feature map of each channel. The symbols a, b, c, and d represent the elements from the corresponding feature maps A, B, C, and D, respectively, and the subscript numbers represent the positions of the corresponding feature map elements. The symbol M represents the feature map after the randomized feature fusion. Take 4 feature maps at the same position in 4 networks at a time, then take 4 elements at the same position in 4 feature maps, and then use the sampling matrix to select one element. The sampling matrix is randomly selected according to its probability value, that is, the probability of being selected with a large element value is large.

In order to enhance the semantic information of the low-level feature layer, a feature pyramid model is constructed on the feature layer of the SSD. The convolutional features of the same level are first fused and enhanced, and then top-down up sampling is used to achieve the fusion of the high-level features and the low-level features. First, in the network feedforward convolution operation, multiple convolutional layers in the same level are used to enhance the feature map. Assume that any hidden layer in the convolutional feature layer is represented as ConvN, and ConvN-1 is the hidden level feature output layer. After convolution and pooling, ConvN-2 is obtained, and then ConvN-1 and ConvN-2 are convolved. The fusion yields ConvN-3. Finally, the ConvN-x (x=1, 2, 3) is fused, and the stacking effect of multiple fusions is eliminated through a 1 × 1 dimensionality reduction convolution kernel to achieve information enhancement within the feature level.

In this paper, the same layer horizontal connection and the difference layer residual processing are added between the fourth to seventh layer feature layers of the convolutional neural network to build a feature pyramid. Then superimpose the sixth layer feature layer to obtain a new sixth layer feature, and finally add an additional 1 × 1 convolution process to reduce the feature dimension. By analogy, the fourth layer of new features is obtained, and finally the 3 × 3 convolution operation is used to eliminate multiple feature aliasing effects, as shown in Figure 5. An anchor mechanism

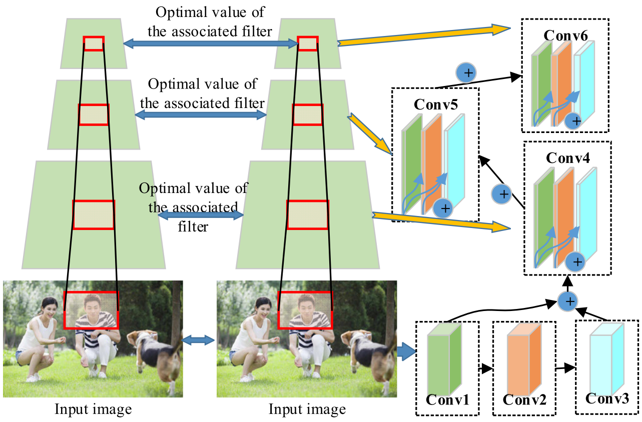
Optimal value of the associated filter
Optimal value of the associated filter
Optimal value of the associated filter
Conv6
Conv5
Conv4
+
+
+
+
+
+
Input image
Input image
Conv1
Conv2
Conv3

FIGURE 5. Convolutional layer fusion in feature hierarchy based on SSD.

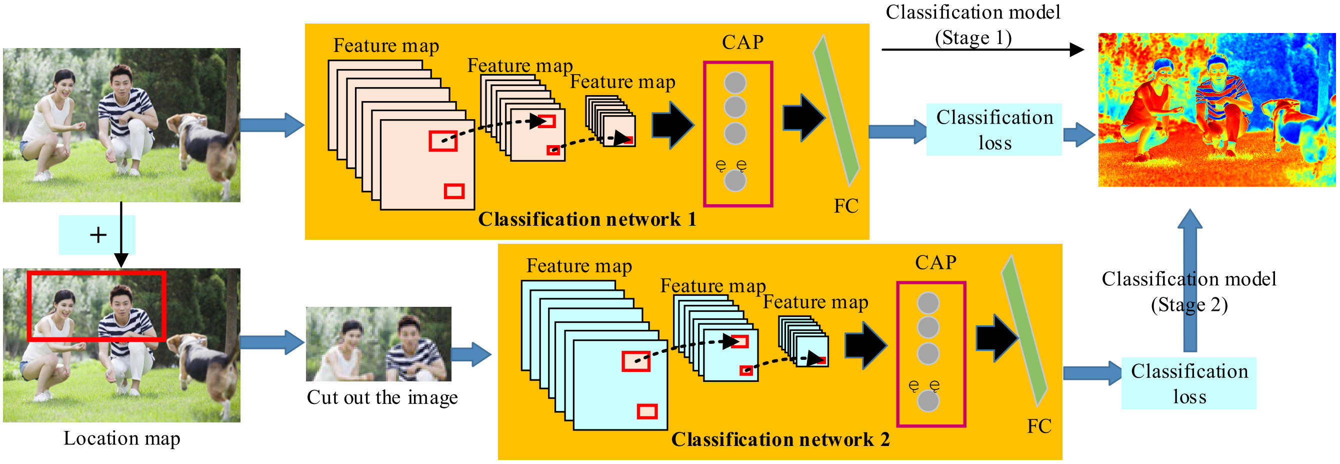

FIGURE 6. Cascade deep convolutional neural network model.

is used on the new feature layer to extract the image target candidate frame and output it to the fully connected layer. According to the predicted image target classification information, the non-maximum suppression is used to remove the overlapping image targets.

This paper first calculates the sixth layer of feature matching results to find the region with the highest feature correlation, that is, the optimal solution of the correlation filter w is obtained by Fourier transform:

$$f_L = \Psi^{-1}(\sum_{d=1}^{D} w^d \odot \bar{v}^d) \tag{5}$$

Among them, the variable $f_L$is the L-level binocular image feature correlation matrix. The variable D is the number of L-layer characteristic channels. The variable Ψ^{-1}represents the inverse Fourier transform. The variable $\bar{v}^d$represents complex conjugate. The symbol $\odot$ indicates the Hadamard product. At this time, when the maximum value of f_L is obtained, the corresponding point position$v_{m,n}$ is the matching point position of the feature layer. According to the receptive field characteristics of the network, the position of the points $v_{m,n}$corresponding to the low-level feature points$(\hat{m}, \hat{n})$ must meet the following conditions:

$$|m - \hat{m}| + |n - \hat{n}| \leq r \tag{6}$$

Among them, (m, n) is the feature point coordinates of the current layer, and r is the receptive field coverage, which can be obtained according to the up-sampling scaling ratio.

Traverse the points in the fifth-level search area to obtain low-level matching features. By repeating this search process, the fourth layer of matching features can be quickly obtained. Finally, use the receptive field to map the fourth layer of matching features to the original image to obtain the matching information of the pixels, and use the arithmetic mean to obtain the fused pixel values of the same-named point pairs in the image to achieve image fusion.

C. CASCADED DEEP CONVOLUTIONAL NEURAL NETWORK

After obtaining the fused feature information, it needs to be sent to the classification neural network for detection and analysis. Therefore, this paper proposes a new two-level cascade DCNNs structure, the only supervision signal given in training is image-level labels, that is, weak supervision information. The proposed method is divided into two stages, as shown in Figure 6. Stage 1 training classification network, the training data is the original image and its image-level labels, predict the target category and visual positioning map. Generate cropped images as input for stage 2. The cropped image generated in stage 1 has obvious target adhesion phenomenon, and stage 2 trains the classification network of the same structure to learn the information of the adhesion area, and solves the problem that multiple categories of targets in a single image cannot be clearly distinguished.

1) DATA PREPROCESSING

Since the training image contains multiple types of targets, image-level label selection is defined as a multi-category label format, that is, the category labels of the training data are represented by 0 and 1, "1" indicates that the target exists in the image, and "0" indicates that there is no such target in the image. In this paper, the image-level label of the training image is converted into a two-dimensional vector format. The expression of the two-dimensional label vector is:

$$y = \{y_1, y_2, \ldots, y_c\} \in \{0, 1\} \tag{7}$$

Due to the limitation of label type and format, this paper chooses sigmoid cross entropy as the loss function. The expression is:

$$loss = -\frac{1}{N}\sum_{n=1}^{N}[P_n \log p_n + (1 - P_n)\log(1 - p_n)] \tag{8}$$

The training images and corresponding multi-category labels are randomly perturbed by the same rules to ensure the effectiveness of gradient descent and learning during training.

2) CAM CALCULATION

In this paper, CAM is used to highlight the local interest region of the target to obtain the required positioning information. As shown in Figure 6, the structure of the CAM algorithm during the training phase is to add a global average pooling layer (GAP) after the last convolutional layer, followed by a fully connected layer for classification. The average value of each unit feature map of the last convolution layer and all parameters of the fully connected layer are output, and the heat map is obtained by weighting the feature map obtained by the GAP operation. At this time, the output size of the heat map and the size of the feature map are the same. The heat map of the final output in Figure 6 is a superposition of the heat map and the original image. Using M_c to define the CAM of category c, each element of the space is:

$$M_c(x, y) = W_c^T f(x, y) \tag{9}$$

In the formula, the variable W_c^T is the classification weight associated with class c. The variable f(x, y) represents the feature value at (x, y) on the feature map of the last convolutional layer.

D. APPLYING VIRTUAL REALITY TECHNOLOGY TO 3D IMAGE RECONSTRUCTION

We use of virtual reality technology in the computer to achieve three-dimensional image reconstruction. First import the original image into the computer, use image preprocessing technology to perform image filtering, image segmentation and image interpolation on the imported original image, and use the above neural network model to reconstruct the detected image in the VTK software for 3D image reconstruction to achieve three-dimensional visualization of images.

In this paper, the bounding box method is selected as a three-dimensional image reconstruction algorithm based on virtual reality technology. Choose the bounding box method to draw the image in 3D in VTK software. The process of 3D image rendering is the process of 3D image reconstruction.

Using the local image gradient information to obtain the three-dimensional coordinates of the image data volume, set the original image coordinates to (x, y), the contour length formula of the three-dimensional image reconstruction is as follows:

$$E = \gamma E^{LBF} + (1-\gamma)E^{LGF} + vL(\phi) + \mu P(\phi) \tag{10}$$

In the formula, the variables γ and $L(\phi)$ respectively represent the gray weight coefficient and the edge contour length constraint of each pixel neighborhood in the local image. The variable $P(\phi)$ represents the sparse regular term. Variable E^{LBF} and variableE^{LGF} represent local gray information and local gradient energy terms, respectively. Both the variable v and the variable μ are constants greater than 0, which represent the weight coefficients of the scan constraint vectors in each local image space.

By implementing the smoothness evolution game on the image target area and the background area, the box model is obtained as:

$$L(\phi) = \int \delta(\phi)|\nabla\phi|dx \tag{11}$$

In the formula, the variable $\delta(\phi)$ and the variable $\nabla\varphi$ represent the pixel sparsity regularization terms of the image target area and the background area, respectively. The variable dx represents the horizontal coordinate position of the original image target point.

The local Gaussian probability distribution of the pixels on the edge of the image is obtained by the bounding box method, the formula is:

$$P(\phi) = \int \frac{1}{2}(|\nabla\phi| - 1)^2 dx \tag{12}$$

The formula for obtaining the three-dimensional coordinates of the image data volume by formula (12) is as follows:

$$\begin{aligned} &E^{LBF}(\phi, f_1, f_2) \\ &\quad = \lambda_1 \int [B_\sigma(x-y)|I - f_1(x)|^2 H(\phi)dy]dx \\ &\quad\quad + \lambda_2 \int [B_\sigma(x-y)|I - f_2(x)|^2(1 - H(\phi))dy]dx \end{aligned} \tag{13}$$

$$\begin{aligned} &E^{LGF}(\phi, f_1^G, f_2^G) \\ &\quad = \lambda_1 \int [B_\sigma(x-y)|I^G - f_1^G(x)|^2 H(\phi)dy]dx \\ &\quad\quad + \lambda_2 \int [B_\sigma(x-y)|I^G - f_2^G(x)|^2(1 - H(\phi))dy]dx \end{aligned} \tag{14}$$

In the formula, variable I^G, variable f_1^G, and variable f_2^G respectively represent the gradient mode of the reconstructed surface, the gradient mode of the curved part of the 3D image reconstruction grid surface and the gradient mode of the linear part of the 3D image reconstruction grid surface. The variable λ1 and the variable λ2 are both constants greater than 0, indicating the weight coefficients of the scan constraint vectors in each partial image space. The variable B_σ and the variable $H(\phi)$ represent the standard deviation of the Heaviside function and the three-dimensional nonlinear space mapping, respectively. Variable $f_1(x)$ and variable $f_2(x)$ reconstruct the gray value of the three-dimensional image. The variable dy represents the ordinate position of the original image target point.

Using the above formula to construct the three-dimensional data field formula for image reconstruction using Euler Lagrange equation is as follows:

$$\begin{aligned} \frac{\partial\phi}{\partial t} &= -\delta(\phi)[\theta(\lambda_1 e_1^{LBF} - \lambda_2 e_2^{LBF}) \\ &\quad + (1-\theta)(\lambda_1 e_1^{LBF} - \lambda_2 e_2^{LBF})] \\ &\quad + v\delta(\phi)div(\frac{\nabla\phi}{|\nabla\phi|}) + \mu(\nabla^2\phi - div(\frac{\nabla\phi}{|\nabla\phi|})) \end{aligned} \tag{15}$$

In formula, the calculation formulas of variable e_1^{LBF}, variable e_2^{LBF}, variable e_1^{LGF}, and variable e_2^{LGF} are as follows:

$$\begin{aligned} e_1^{LBF} &= \int B_\sigma(y-x)|I-f_1(x)|^2dy \\ e_2^{LBF} &= \int B_\sigma(y-x)|I-f_2(x)|^2dy \\ e_1^{LGF} &= \int B_\sigma(y-x)|I^G-f_1^G(x)|^2dy \\ e_2^{LGF} &= \int B_\sigma(y-x)|I^G-f_2^G(x)|^2dy \end{aligned} \tag{16}$$

In the above formula, the Heaviside function $H(\phi)$ and Dirac function $\delta(\phi)$ represent the dynamic points of three-dimensional nonlinear space mapping and image mapping, respectively.

The 3D direct volume of the image is drawn according to the 3D data field of the image reconstruction. The 3D texture map and the 3D array coordinate parameters are reconstructed from the image in the 3D data field. The pixels in the image distribution are regarded as templates and matched. The matching formula is as follows:

$$H_\varepsilon(z) = \frac{1}{2}[1+\frac{2}{\pi}\arctan(\frac{z}{\varepsilon}) \tag{17}$$

$$\delta_\varepsilon(z) = \frac{1}{\pi}\cdot\frac{\varepsilon}{\varepsilon^2+z^2}) \tag{18}$$

In the formula, variable $H_\varepsilon(z)$ and variable $\delta_\varepsilon(z)$ are the reconstructed 3D texture map and 3D array coordinate parameters, respectively. To fix the ray direction ε in the image template, by minimizing the variablef_1, variable f_2, variable f_1^G, and variable f_2^G respectively, the 3D image reconstruction output formula using the 3D direct volume of the image is:

$$\begin{aligned} f_1(x) &= \frac{B_\sigma(x)\times H_\varepsilon(\phi(x))I(x)}{B_\sigma(x)\times H_\varepsilon(\phi(x))} \\ f_2(x) &= \frac{B_\sigma(x)\times(1-H_\varepsilon(\phi(x)))I(x)}{B_\sigma(x)\times(1-H_\varepsilon(\phi(x)))} \\ f_1^G(x) &= \frac{B_\sigma(x)\times H_\varepsilon(\phi(x))I^G(x)}{B_\sigma(x)\times H_\varepsilon(\phi(x))} \\ f_2^G(x) &= \frac{B_\sigma(x)\times(1-H_\varepsilon(\phi(x)))I^G(x)}{B_\sigma(x)\times(1-H_\varepsilon(\phi(x)))} \end{aligned} \tag{19}$$

Using this method to perform three-dimensional reconstruction of images in VTK software can effectively improve the accuracy of three-dimensional image reconstruction and the ability to express statistical information.

IV. SIMULATION EXPERIMENT

A. EXPERIMENTAL ENVIRONMENT

The VR equipment used in the experiment is HTC Vive, which includes a headset, a handle on each side, and two lighthouse base stations. Among them, the effective resolution of the single-head headset is 1200 × 1080, the combined resolution of the two eyes is 2160 × 1200, the screen refresh rate is 90Hz, and the delay is 22ms. The left and right hands hold a handle, and the two lighthouse base stations are placed diagonally in the room to ensure that there is no obstruction. The camera uses a Kinect V2 color camera with a resolution of 1920 × 1080 and a frame rate of 30fps. It is placed on the horizontal ground at the edge of the room to ensure that most areas of the room can be seen.

In this paper, images from four different datasets are used for the experiment: UCAS-AOD, RSOD, DOTA and self-built datasets, which are respectively recorded as datasets A-D. They are all public data sets. The images of the UCAS-AOD and RSOD datasets are both small-scale airport area slices. The UCAS-AOD dataset has a larger image repetition rate, a larger aircraft target size, and less interference from background and other artificial objects. The RSOD data set images have a certain inclination, the imaging quality is poor, and there are a large number of dense small targets. Both DOTA and self-built datasets select large-scale remote sensing images that can cover the entire airport. Using 70% of the image training in dataset A, a total of 30 positive samples was selected, some 30 positive samples, and 90 negative samples. The remaining images are used as test data to verify the generalization and practicability of the model for images with different resolutions, sizes, and quality.

The experimental platform is HP workstation, based on Ubuntu operating system, the system model is OptiPlex 990, the central processing unit is Intel Core i7-2600, equipped with 4G memory, programming software is Python2.7, using Caffe framework, dependent Open-CV library and PIL. The initial learning rate of the DCNNs network is 0.01, the batch size is 64.

B. VR DEVICE SENSOR OPTIMIZATION

Due to the noise in the sensor measurement process, the jitter is generated during the movement of the model, and the position deviation between the two frames is unstable. The human eye is usually sensitive to low speeds, and is particularly sensitive to high-speed lag, so we use a cut-off frequency adaptive low-pass filter. By estimating the speed of the signal, the cut-off frequency of the low-pass filter is adjusted for each new sample. Although noise signals are usually sampled at a fixed frequency, filtering does not always follow the same speed. Consider the actual time interval between samples, according to the following formula:

$$\alpha = (1+\frac{\tau}{T_e})^{-1} \tag{20}$$

Formula (20) calculates α based on the sampling period T_e and the time constant τ (in seconds).

$$\tau = \frac{1}{2\pi f_c} \tag{21}$$

Obtain the cut-off frequency fc (Hertz) according to equation (21).

$$\hat{X}_i = (X_i + \frac{\tau}{T_e}\hat{X}_{i-1})1 + \frac{\tau}{T_e} \tag{22}$$

$$f_c = f_{c_{\min}} + \beta|\hat{X}_i| \tag{23}$$

Then according to formula (22) and formula (23), the adaptive cutoff frequency fc can be calculated. Use low fc at low signal speeds, and to reduce hysteresis, fc increases with speed. Use the sampling rate to calculate the speed from the original signal value, and then use the selected cutoff frequency for low-pass filtering. We tested this time filtering with other filtering methods in experiments, and the effect comparison is shown in Figure 7.

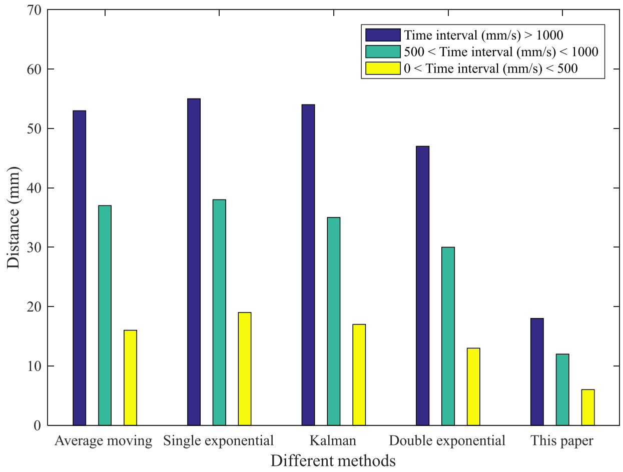

FIGURE 7. **The average distance between the filtered position and the actual cursor position in each time interval.**

C. PERFORMANCE OF IMAGE DETECTION ALGORITHM

1) TEST RESULTS

In single-label image classification tasks, top-1 accuracy, or top-5 accuracy is generally used to measure the accuracy of classification. The evaluation method of multi-label image classification tasks generally adopts a method similar to that in information retrieval. The average precision (AP) is calculated as follows:

$$AP = \int_0^1 PRdr \tag{24}$$

$$P = \frac{TP}{TP + FP} \tag{25}$$

$$P = \frac{TP}{TP + FN} \tag{26}$$

where, TP represents positive class predicted as positive class, FN represents negative class predicted as negative class, FP represents negative class predicted as positive class, and TN represents negative class predicted as negative class.

In the formula, the variable P represents the precision rate. The variable R represents the recall rate. The variable P is a function that takes R as a parameter. The experiment in this paper first uses the trained model to output the category confidence of all test images, and then sorts the category confidence. In this order, as the number of selected images N gradually increases, P and R under each N are calculated. The category confidence greater than 0.5 in this process is considered to be the positive prediction. Finally, output the largest P under any R to get a certain type of AP. The simulation experiment results are shown in Figure 8.

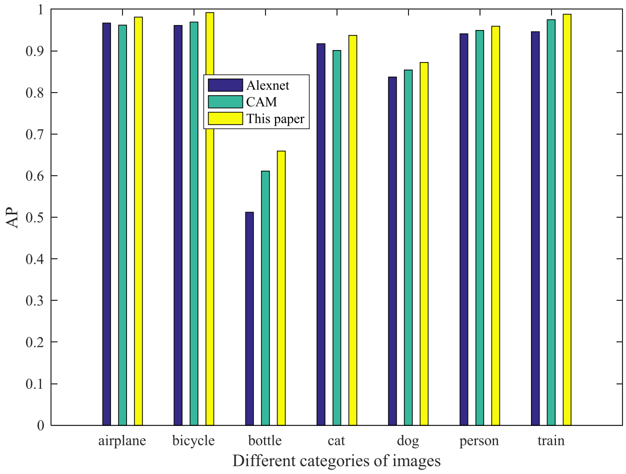

FIGURE 8. **AP results for different image types.**

Literature [30] pointed out that deleting the additional layer has a great impact on the classification performance of AlexNet. The experimental results in Figure 8 show that in the case of using weakly supervised labels, although the method in this paper cuts out two fully connected layers relative to AlexNet, the classification accuracy differs by only 0.1%, and the classification capabilities of the two are comparable, while the relative for CAM, the classification accuracy of this method is improved by 0.028.

In order to verify the effectiveness of the cascaded convolutional neural network model and training method designed in this paper, using the same training and test data, five target detection methods based on convolutional neural networks were compared: SSD300 [31], YoLoV2 [32], FRCNN [33], RetinaNet [34] and MTCNN algorithm [35]. SSD300 uses VGG16 [36] as the backbone network, and YoLov2, FRCNN and RetinaNet use ResNet50 as the backbone network. Both MTCNN and the method in this paper can detect images of any size, and algorithms such as FRCNN cannot directly detect large images. Therefore, when using algorithms such as FRCNN to detect datasets C and D, two types of 512 × 512 pixels and 1024 × 1024 pixels are used split way.

The five algorithms are used to detect the images in the four public datasets, and the PR curve is drawn according to the accuracy and recall in the detection results, as shown in Figure 9. Based on the detection results of the four datasets, the performance of the SSD300 algorithm is the worst. The main reasons are:

1) The training data comes from dataset A. The image size of dataset A is 1280 × 659 ∼ 1372 × 940 pixels. When training SSD300 algorithm, the image scaling is too drastic, which seriously affects the model training and detection effect.

2) In this paper, part of the images in dataset A is used as the training set to test the remaining three datasets. The images of different datasets have large differences in size and resolution, resulting in a decline in the detection results.

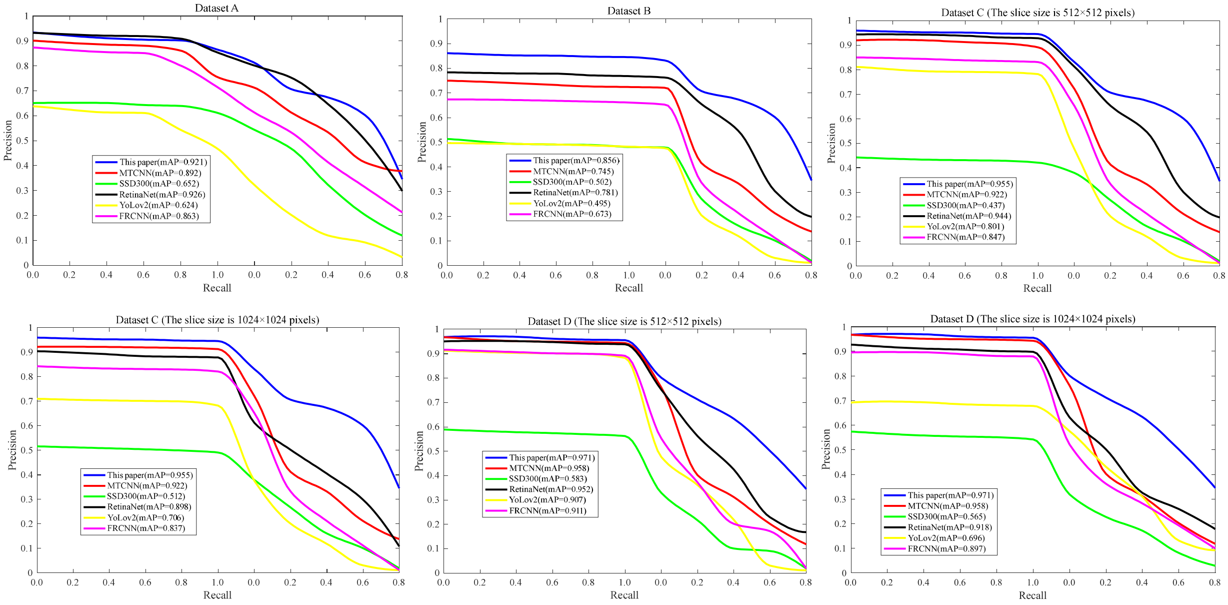

FIGURE 9. PR curves for four datasets with different methods.

3) The SSD300 algorithm uses the VGG16 model as the backbone network, and the VGG16 model itself has poor generalization. YoLoV2 input image size is 448 × 448 pixels, it is difficult to have a greater accuracy improvement for dataset A and dataset B with lower resolution. For the images in dataset C and dataset D, when divided into 512×512 pixels, the detection accuracy of more than 80% can be achieved; when divided into 1024 × 1024 pixels, the detection accuracy drops sharply. The detection accuracy of SSD300 and YOLov2 is weaker than that of FRCNN and RetinaNet. This is determined by its model design. The end-to-end detection method improves the detection speed, but sacrifices accuracy, especially weakens the detection ability of small targets. Under different datasets, FRCNN and RetinaNet algorithms can achieve relatively good detection results, and RetinaNet is superior to FRCNN. The higher accuracy of FRCNN and RetinaNet is mainly due to the idea of adopting regional suggestions, and the size of the training image is 1200 × 600 pixels. When the resolution of the detected image and the training image are close, a good detection result can be obtained. When the resolution of the detected image is improved, the accuracy can be further improved. The original cascaded network method can achieve higher accuracy than FRCNN in all four datasets without any improvement, indicating that the cascaded convolutional neural network benefits from being able to perform multi-scale traversal and search on the entire image, with natural precision advantages. The method in this paper further improves the accuracy of the original cascade network, especially improves the detection accuracy of the poor-quality images and dense small targets in dataset B. Except in Dataset A, the accuracy is slightly weaker than RetinaNet, and it is significantly better than RetinaNet in the remaining datasets.

2) POSITIONING RESULTS

Since the goal of this paper is to solve the problem of multi-target adhesion in a single image, only using CAM cannot mark the overall position of the object for certain classes. The CAM algorithm limits the output of the positioning results to the most interesting area. This paper further verifies the effectiveness of the algorithm through the evaluation method combining precision rate and recall rate. The location results of the recall rate and accuracy rate in the verification set are shown in Table 1.

TABLE 1. Comparison of positioning performance of validation sets evaluated using R and P.

Class	CAM		Stage 11		This paper	
	R	P	R	P	R	P
Airplane	0.361	0.644	0.414	0.561	0.459	0.678
Bird	0.135	0.308	0.285	0.289	0.276	0.337
Bicycle	0.216	0.353	0.235	0.639	0.265	0.642
Car	0.108	0.285	0.152	0.317	0.159	0.291
Cat	0.409	0.476	0.641	0.474	0.652	0.475
Dog	0.268	0.476	0.341	0.462	0.355	0.273
Sleep	0.054	0.325	0.109	0.419	0.167	0.492
Sofa	0.059	0.174	0.041	0.155	0.113	0.195

In Table 1, stage 11 indicates that the model is not retrained in stage 2, and the weights of stage 1 are still used for prediction. Compared with CAM, the accuracy of stage 11 and the method in this paper is only reduced by 0.02 and 0.014. The recall rate of the method in this paper is improved by about 0.041 relative to stage 11 and increased by 0.085 relative to the CAM algorithm. Experimental results show that the proposed method can obtain better target detection performance, the problem of multi-target adhesion is solved,

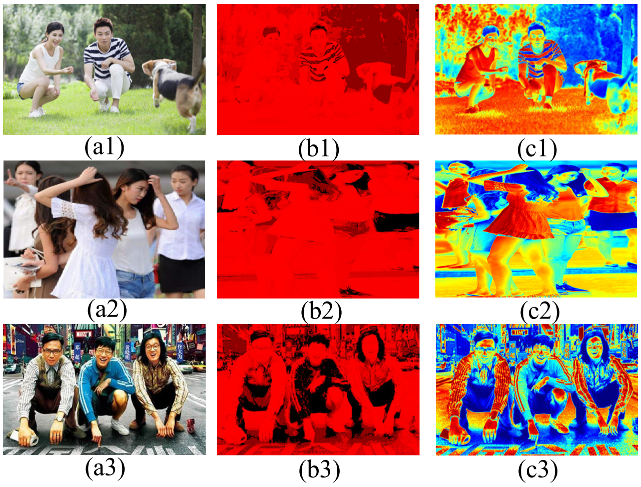

FIGURE 10. Visualization map.

and the accuracy is high. In order to more intuitively observe the performance of the algorithm in this paper, the generated visual positioning is shown in Figure 10.

The CAM positioning map in Figure 10 overlaps with the original image. The red highlighted part is the positioning area after visualization. Figure 10(a) can clearly see that there is a multi-target adhesion problem, and the positioning areas of two similar targets are stuck together. The problem of multi-target adhesion in Figure 10(b) begins to improve, and the similar targets of a certain type or a certain graph have been separated to a certain extent. The location diagram in Figure 10(c) shows that the multi-target adhesion problem has been solved. It shows that the method in this paper has obvious advantages for dealing with multi-objective adhesion problems. The algorithm in this paper can effectively solve the problem of multi-target adhesion. Since the CAM algorithm locates the area of most interest in a certain category, the positioning area may not include all parts of the target for certain categories, and the recall rate is improved by 0.09. As shown in Figure 10, for the "human" category, the higher the positioning accuracy, and the more concentrated the positioning area is on the person's head.

Figure 11 shows the recognition error rate of dataset A and dataset B under different algorithms. It can be seen from Figure 11 that, on the two databases, the multi-information input model proposed in this paper reduces the recognition error rate by 0.098 and 0.013 respectively compared to the single-information input model. CNN is superior to other deep learning algorithms in image recognition. This is mainly because CNN can effectively extract the features of images through local receptive fields and weight sharing, and has certain translation invariance. It can be seen from the curves of single information input and multiple information input in Figure 11 that it is better to input the multi-directional gradient information of the picture into the network and do randomized feature fusion than directly input the original picture information. This is mainly because the original data is directly input into the convolutional neural network, although it can also get a lower recognition error rate, but under the condition that the model structure is simple, it cannot effectively learn the hidden feature information in the dataset. The effectiveness directly determines the classification performance. The feature information of the four gradient directions can make more comprehensive use of the hidden feature information in the data, and the recognition error rate obtained on the two databases is lower, which also shows that the basic expression of multi-directional gradient information as edge information is effective. This paper compares the multi-input information model with or without batch normalization algorithm. After using batch normalization, the recognition error rate of dataset A after 40 iterations is reduced by 0.028. Due to the low recognition error rate of dataset B, the magnitude of the decrease is not very obvious, indicating that the concentrated data becomes more scattered, which is beneficial to prevent overfitting to a certain extent and obtain better results.

D. ANALYSIS OF 3D IMAGE MODELING RESULTS

In order to test the effectiveness of the three-dimensional image reconstruction method based on virtual reality technology in this paper, a simulation experiment was carried out

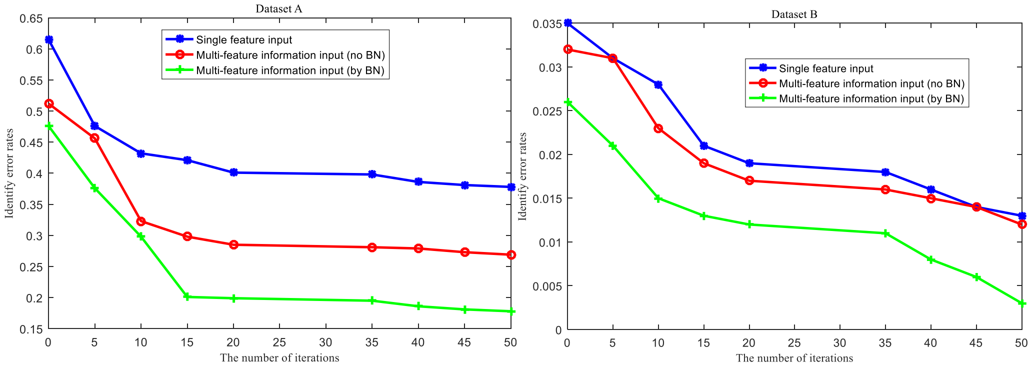

FIGURE 11. Recognition error rate of different datasets under different input information.

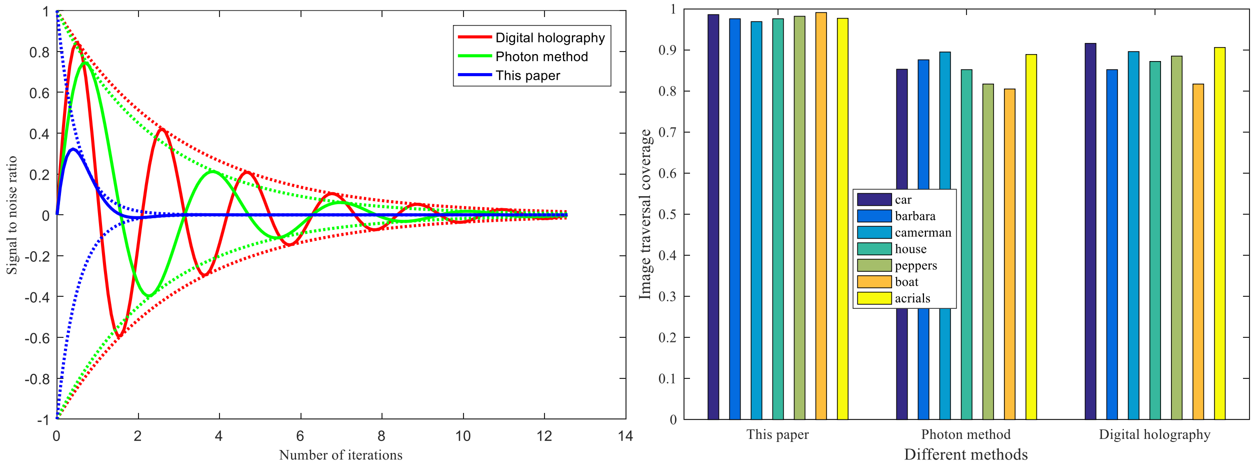

FIGURE 12. Comparison of the signal-to-noise ratio and traversal coverage of the output image of the three methods.

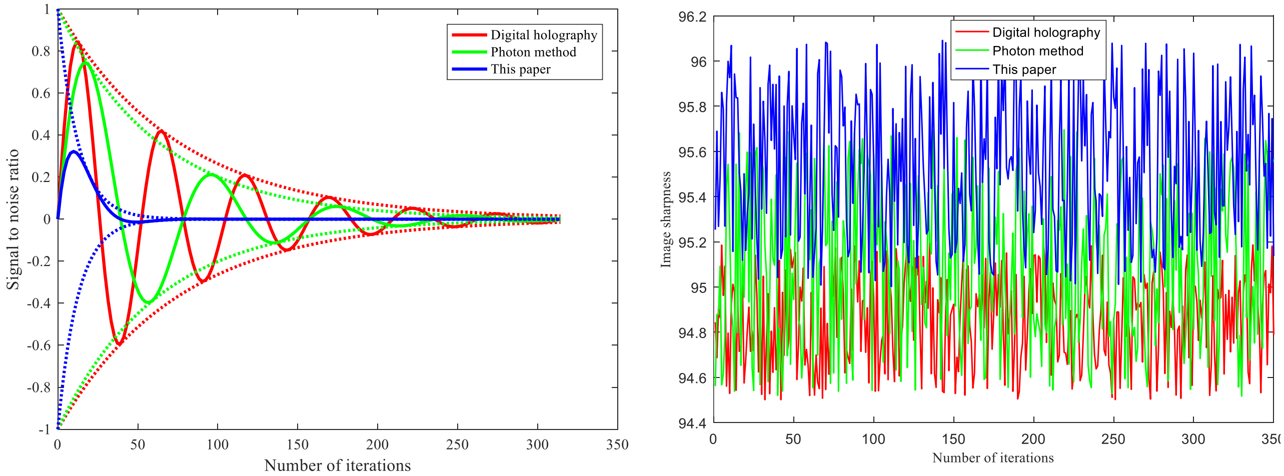

FIGURE 13. Comparison of precision and clarity of reconstructed images by three methods.

on the Matlab simulation platform. The experiment selected 10 representative images for three-dimensional reconstruction, and combined the method of this paper with the photon method [37] and digital holographic method [38] for comparison.

Three images are reconstructed three-dimensionally by three methods, and the signal-to-noise ratio results of the output three-dimensional images after the three methods are shown in Figure 12(a). The signal-to-noise ratio of the output image can directly reflect the size of the noise contained in the image. The greater the signal-to-noise ratio of the output image, the smaller the noise contained in the image and the higher the image quality; otherwise the opposite. It can be seen from the experimental results in Figure 12(a) that the three-dimensional reconstruction of 10 images using this method results in the highest signal-to-noise ratio of the output three-dimensional images, indicating that the image quality of the three-dimensional image reconstruction by this method is significantly higher than two other methods. The traversal coverage of the output image after three-dimensional reconstruction of 10 images by three methods, the results are shown in Figure 12(b). From the experimental results in Figure 12(b), it can be seen that the three-dimensional image reconstruction of the 10 images by the method of this paper, the output image traversal coverage area is more comprehensive, and the average traversal coverage of the reconstructed 10 three-dimensional images is as high as 0.979. The photon method and the digital holography method are used to reconstruct 10 images, and the average traversal coverage of the 10 images after reconstruction is only 0.839 and 0.873. The traversal coverage of the reconstructed image using this method is significantly higher than that of the other two methods, indicating that this method has higher reconstruction performance.

The three-dimensional image reconstruction method not only needs to have high reconstruction efficiency, reconstruction error rate and image clarity is important manifestations of the reconstruction performance. The reconstruction error rate and image clarity of the three-dimensional reconstructed image of this method with the photon method and digital holography method are compared. The comparison results are shown in Figure 13.

From the experimental comparison results in Figure 13, it can be seen that the reconstruction error rate and image clarity of this method are significantly better than the photon method and digital holography method for 10 images. The reconstruction error rate of this method is finally lower than 0.1, and the output image the clarity has been around 0.955, which further verifies the effectiveness of this method.

V. CONCLUSION

In the reconstruction process, traditional 3D image reconstruction method based on virtual reality leads to a new research method based on virtual reality technology for image detection and 3D image reconstruction after 3d image reconstruction with low precision. In this paper, a general weakly supervised learning multi-target detection algorithm based on two-level cascading deep convolutional neural network is proposed. All steps are learned end-to-end through deep neural networks, avoiding complex clustering or optimized initialization processes. Secondly, visual software development platform and virtual reality 3D image processing software are selected as the 3d image reconstruction implementation platform, and the 3d direct volume of the image is drawn according to the 3d data field of image reconstruction. Preferably, the output formula of 3D image reconstruction is obtained through the direct volume of 3D image to realize 3D image reconstruction based on virtual reality technology. In this paper, 3D image reconstruction based on virtual reality technology is realized, but at the cost of time efficiency. The next step of this paper is to improve the time efficiency while improving the accuracy of 3D image reconstruction.

REFERENCES

[1] C. Donghui, L. Guanfa, Z. Wensheng, L. Qiyuan, B. Shuping, and L. Xiaokang, "Virtual reality technology applied in digitalization of cultural heritage," *Cluster Comput.*, vol. 22, no. 4, pp. 1–12, Jul. 2019.

[2] Y. Zhang and X.-L. Ma, "Research on image digital watermarking optimization algorithm under virtual reality technology," *Discrete Continuous Dyn. Syst.*, vol. 12, nos. 4–5, pp. 1427–1440, Aug. 2019.

[3] W. Wei, B. Zhou, D. Połap, and M. Woźniak, "A regional adaptive variational PDE model for computed tomography image reconstruction," *Pattern Recognit.*, vol. 92, pp. 64–81, Aug. 2019.

[4] G. Roberts, N. Holmes, N. Alexander, E. Boto, J. Leggett, R. M. Hill, V. Shah, M. Rea, R. Vaughan, E. A. Maguire, K. Kessler, S. Beebe, M. Fromhold, G. R. Barnes, R. Bowtell, and M. J. Brookes, "Towards OPM-MEG in a virtual reality environment," *NeuroImage*, vol. 199, pp. 408–417, Oct. 2019.

[5] M. A. Martens, A. Antley, D. Freeman, M. Slater, P. J. Harrison, and E. M. Tunbridge, "It feels real: Physiological responses to a stressful virtual reality environment and its impact on working memory," *J. Psychopharmacol.*, vol. 33, no. 10, pp. 1264–1273, Jul. 2019.

[6] J. Torner, S. Skouras, J. L. Molinuevo, J. D. Gispert, and F. Alpiste, "Multipurpose virtual reality environment for biomedical and health applications," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 8, pp. 1511–1520, Aug. 2019.

[7] L. Zhang, P. Shen, X. Peng, G. Zhu, J. Song, W. Wei, and H. Song, "Simultaneous enhancement and noise reduction of a single low-light image," *IET Image Process.*, vol. 10, no. 11, pp. 840–847, Nov. 2016.

[8] Q. Ke, J. Zhang, W. Wei, D. Połap, M. Woźniak, L. Kośmider, and R. Damaševičius, "A neuro-heuristic approach for recognition of lung diseases from X-ray images," *Expert Syst. Appl.*, vol. 126, pp. 218–232, Jul. 2019.

[9] B. Zhou, X. Duan, D. Ye, W. Wei, M. Woźniak, D. Połap, and R. Damaševičius, "Multi-level features extraction for discontinuous target tracking in remote sensing image monitoring," *Sensors*, vol. 19, no. 22, p. 4855, Nov. 2019.

[10] Y. Zhang, G. Fei, and G. Yang, "3D viewpoint estimation based on aesthetics," *IEEE Access*, vol. 8, pp. 108602–108621, Jun. 2020.

[11] İ. Bozcan and S. Kalkan, "COSMO: Contextualized scene modeling with Boltzmann machines," *Robot. Auto. Syst.*, vol. 113, pp. 132–148, Mar. 2019.

[12] C. Zou, R. Guo, Z. Li, and D. Hoiem, "Complete 3D scene parsing from an RGBD image," *Int. J. Comput. Vis.*, vol. 127, no. 2, pp. 143–162, Nov. 2019.

[13] X. Chen, F. He, and H. Yu, "A matting method based on full feature coverage," *Multimedia Tools Appl.*, vol. 78, no. 9, pp. 11173–11201, Sep. 2019.

[14] W. Li, W. Meng, Z. Tan, and Y. Xiang, "Design of multi-view based email classification for IoT systems via semi-supervised learning," *J. Netw. Comput. Appl.*, vol. 128, pp. 56–63, Feb. 2019.

[15] S. Kansal and S. Mukherjee, "Automatic single-view monocular camera calibration-based object manipulation using novel dexterous multi-fingered delta robot," *Neural Comput. Appl.*, vol. 31, no. 7, pp. 2661–2678, Oct. 2019.

[16] U. L. C. Baldos, F. G. Viens, T. W. Hertel, and K. O. Fuglie, "R&D spending, knowledge capital, and agricultural productivity growth: A Bayesian approach," *Amer. J. Agricult. Econ.*, vol. 101, no. 1, pp. 291–310, Jul. 2019.

[17] H. He, G. Li, Z. Ye, A. Mao, C. Xian, and Y. Nie, "Data-driven 3D human head reconstruction," *Comput. Graph.*, vol. 80, pp. 85–96, May 2019.

[18] T. Kaur and T. K. Gandhi, "Deep convolutional neural networks with transfer learning for automated brain image classification," *Mach. Vis. Appl.*, vol. 31, no. 3, pp. 1–16, Mar. 2020.

[19] G. Song, Z. Huang, Y. Zhao, X. Zhao, Y. Liu, M. Bao, J. Han, and P. Li, "A noninvasive system for the automatic detection of gliomas based on hybrid features and PSO-KSVM," *IEEE Access*, vol. 7, pp. 13842–13855, Jan. 2019.

[20] I. Cong, S. Choi, and M. D. Lukin, "Quantum convolutional neural networks," *Nature Phys.*, vol. 15, no. 12, pp. 1273–1278, Aug. 2019.

[21] J. Hanson, K. Paliwal, T. Litfin, Y. Yang, and Y. Zhou, "Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks," *Bioinformatics*, vol. 35, no. 14, pp. 2403–2410, Jul. 2019.

[22] Z. Zhang, B. Li, W. Zhang, R. Lu, S. Wada, and Y. Zhang, "Real-time penetration state monitoring using convolutional neural network for laser welding of tailor rolled blanks," *J. Manuf. Syst.*, vol. 54, pp. 348–360, Jan. 2020.

[23] J. Cai, F. Xing, A. Batra, F. Liu, G. A. Walter, K. Vandenborne, and L. Yang, "Texture analysis for muscular dystrophy classification in MRI with improved class activation mapping," *Pattern Recognit.*, vol. 86, pp. 368–375, Feb. 2019.

[24] D. Kim and Y. J. Ko, "The impact of virtual reality (VR) technology on sport spectators' flow experience and satisfaction," *Comput. Hum. Behav.*, vol. 93, pp. 346–356, Apr. 2019.

[25] C.-W. Shen, J.-T. Ho, P. T. M. Ly, and T.-C. Kuo, "Behavioural intentions of using virtual reality in learning: Perspectives of acceptance of information technology and learning style," *Virtual Reality*, vol. 23, no. 3, pp. 313–324, Sep. 2019.

[26] C. Li, K. Wang, and N. Xu, "A survey for the applications of content-based microscopic image analysis in microorganism classification domains," *Artif. Intell. Rev.*, vol. 51, no. 4, pp. 577–646, Apr. 2019.

[27] M. Chen, S. Lu, and Q. Liu, "Uniform regularity for a Keller–Segel–Navier–Stokes system," *Appl. Math. Lett.*, vol. 107, Sep. 2020, Art. no. 106476.

[28] S. Natarajan, A. Jain, R. Krishnan, A. Rogye, and S. Sivaprasad, "Diagnostic accuracy of community-based diabetic retinopathy screening with an offline artificial intelligence system on a smartphone," *JAMA Ophthalmol.*, vol. 137, no. 10, pp. 1182–1188, Aug. 2019.

[29] D.-X. Zhou, "Universality of deep convolutional neural networks," *Appl. Comput. Harmon. Anal.*, vol. 48, no. 2, pp. 787–794, Mar. 2020.

[30] S. Lu, Z. Lu, and Y.-D. Zhang, "Pathological brain detection based on AlexNet and transfer learning," *J. Comput. Sci.*, vol. 30, pp. 41–47, Jan. 2019.

[31] S. Woo, S. Hwang, H.-D. Jang, and I. S. Kweon, "Gated bidirectional feature pyramid network for accurate one-shot detection," *Mach. Vis. Appl.*, vol. 30, no. 4, pp. 543–555, Mar. 2019.

[32] Q. Xu, R. Lin, H. Yue, H. Huang, Y. Yang, and Z. Yao, "Research on small target detection in driving scenarios based on improved Yolo network," *IEEE Access*, vol. 8, pp. 27574–27583, Jan. 2020.

[33] M. Peebles, S. H. Lim, M. Duke, and B. McGuinness, "Investigation of optimal network architecture for asparagus spear detection in robotic harvesting," *IFAC-PapersOnLine*, vol. 52, no. 30, pp. 283–287, 2019.

[34] N. E. Freeman, J. P. Templeton, W. E. Orr, L. Lu, R. W. Williams, and E. E. Geisert, "Genetic networks in the mouse retina: Growth Associated Protein 43 and Phosphatase Tensin Homolog network," *Mol. Vis.*, vol. 17, no. 153, pp. 1355–1372, May 2011.

[35] Y. Pratama, M. Istoningtyas, and E. Rasywir, "Pengujian algoritma MTCNN (multi-task cascaded convolutional neural network) untuk sistem pengenalan wajah," *J. Media Inf. Budidarma*, vol. 3, no. 3, pp. 240–247, Jul. 2019.

[36] Z. Liu, J. Wu, L. Fu, Y. Majeed, Y. Feng, R. Li, and Y. Cui, "Improved kiwifruit detection using pre-trained VGG16 with RGB and NIR information fusion," *IEEE Access*, vol. 8, no. 1, pp. 2327–2336, Jan. 2020.

[37] J.-M. Wang, H.-H. Fang, and X.-X. Xu, "Two-photon Jaynes–Cummings model interacting with the squeezed vacuum state solved by dressed-state method," *Optik*, vol. 169, pp. 180–189, May 2018.

[38] M. Paturzo, V. Pagliarulo, V. Bianco, P. Memmolo, L. Miccio, F. Merola, and P. Ferraro, "Digital holography, a metrological tool for quantitative analysis: Trends and future applications," *Opt. Lasers Eng.*, vol. 104, pp. 32–47, May 2018.

LONGYU LU was born in Shandong, China, in 1998. He received the bachelor's degree from Yantai University, in 2019, where he is currently pursuing the master's degree in civil engineering. His research interests include construction equipment and the seismic analysis of soil.

JINKAI MA was born in Shandong, China, in 1994. He received the bachelor's degree from the Haidu College, Qingdao Agricultural University, in 2019. He is currently pursuing the master's degree in architecture and civil engineering with Yantai University. His research interest includes structural seismic research.

SHUYING QU was born in Shandong, China, in 1963. She received the master's degree in engineering from the Dalian University of Technology, in 1989. In 2000, she went to Korean Mokpo National University as a Senior Visiting Scholar. In 2002, she was promoted to a Professor of Yantai University. She currently works with Yantai University. She has completed two projects supported by the National Natural Science Foundation of China and three projects supported by the Natural Science Foundation of Shandong Province. She has published more than 50 academic articles in core journals of China and more than 30 articles can be indexed by the world four indexes for references such as the SCI and EI. Her research interests include higher education management and teaching equipment.