IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 38, No. 6, December, 2021, pp. 1719-1726
Journal homepage: http://iieta.org/journals/ts

Three-Dimensional Image Reconstruction for Virtual Talent Training Scene

Tanbo Zhu[1*], Die Wang[2], Yuhua Li[1], Wenjie Dong[3]

[1] State Grid Shandong Electric Power Company, Jinan 250001, China
[2] State Grid Shandong Electric Power Company Electric Power Research Institute, Jinan 250002, China
[3] Shandong Luruan Digital Technology Co., Ltd., Jinan 250001, China

Corresponding Author Email: zhutanbo@sd.sgcc.com.cn

https://doi.org/10.18280/ts.380615

Received: 12 August 2021
Accepted: 10 November 2021

Keywords:
virtual training, three-dimensional (3D) image, image reconstruction

ABSTRACT

In real training, the training conditions are often undesirable, and the use of equipment is severely limited. These problems can be solved by virtual practical training, which breaks the limit of space, lowers the training cost, while ensuring the training quality. However, the existing methods work poorly in image reconstruction, because they fail to consider the fact that the environmental perception of actual scene is strongly regular by nature. Therefore, this paper investigates the three-dimensional (3D) image reconstruction for virtual talent training scene. Specifically, a fusion network model was deigned, and the deep-seated correlation between target detection and semantic segmentation was discussed for images shot in two-dimensional (2D) scenes, in order to enhance the extraction effect of image features. Next, the vertical and horizontal parallaxes of the scene were solved, and the depth-based virtual talent training scene was reconstructed three dimensionally, based on the continuity of scene depth. Finally, the proposed algorithm was proved effective through experiments.

1. INTRODUCTION

With the continuous development of the times, the computer simulation technique of virtual reality (VR) has been widely applied in military, medical, teaching, and many other fields [1-10]. In the field of education, the VR-based virtual talent training becomes an emerging form of training, and captures widespread attention from the education community, thanks to its diverse contents and flexible arrangements [11-15]. The traditional training model mainly imparts knowledge with the aid of slides and videos. By contrast, virtual talent training effectively improves the knowledge learning and skill training effects through interaction and experience perception [16-19]. Compared with the current real training, virtual practical training breaks the limit of space, lowers the training cost, while ensuring the training quality. It provides a good solution to the problems of real training, e.g., undesirable training conditions and limited use of equipment [20-23]. Some training programs are a bit risky, such as electricity training and vehicle driving training. In these programs, virtual training provides a stronger safety guarantee for the trainees than the traditional training model [24, 25]. Image reconstruction is very important to the construction of virtual talent training scene. Experts and scholars have paid much attention to improving the accuracy and completeness of the three-dimensional (3D) reconstruction of virtual talent training scene.

In the industrial sector, VR applications can be used to support training in highly risky or costly environments, which cannot be replicated in real life. Bellemans et al. [26] described a recent VR application built through the close cooperation between Royal Military Academy Sandhurst, the Belgian Navy, and the industrial community. The VR application allows future firefighters to be trained in a virtual replicated ship cabin. VR and augmented reality (AR) are very useful tools for developing new training tools, for they facilitate the creation and maintenance of multiple scenes and environments. AR/VR-based training can reduce the travel and living costs incurred when students are brought to the central training facility, and offer them an immersible training environment. Gluck et al. [27] attempted to integrate artificial intelligence (AI) into VR-based immersive combatant training environment, developed an AI-assisted VR system for training ground soldiers, which help soldiers walk in the environment without being detected. Chen et al. [28] designed an industrial robot training platform based on VR and mixed reality (MR). The platform solves multiple problems of industrial robot training: the high purchase cost of training equipment, the presence of hidden hazards, and the lack of teaching resources. Guptaa and Vargheseb [29] proposed a design and development framework for security training VR platform. As a design file, the framework conceptualizes the accident scene according to the recognized situation of the accident, and requires every trainee to analyze the simulation condition, identify the risks in each scene, and decides the right mitigation measures for the accident outcome. Khwanngern et al. [30] developed a VR application for simulating mandible surgery, which visualizes the operating room in a highly real VR environment. The user of the application can clamp, cut, drill, connect, and compare the 3D skull model, using a motion controller.

Concerning the existing studies on virtual scene reconstruction, there are some methods that utilize the principles of camera imaging and the basic theories on 3D image reconstruction. However, none of them considers the fact that the environmental perception of actual scene is

strongly regular by nature. As a result, two-dimensional (2D) target detection has never been combined with semantic segmentation to reduce the demand for data, lower the cost of data labeling, and improve the quality of the reconstructed image. Therefore, this paper investigates the 3D image reconstruction for virtual talent training scene. The main contents are as follows: Section 2 identifies images on virtual talent training scene, designs a fusion network model, and explores the deep-seated correlation between target detection and semantic segmentation for images taken in 2D scenes, aiming to enhance the extraction effect of image features. Section 3 solves the vertical and horizontal parallaxes of the scene, and reconstructs the 3D virtual talent training based on depth, using the continuity of scene depth. Finally, experiments were carried out to verify the effectiveness of the proposed algorithm [31, 32].

2. IMAGE IDENTIFICATION

In a virtual talent training scene, the images shot in the scene (hereinafter referred to as scene images) are the most important information source for perceiving the virtual training environment. The images taken by cameras in the virtual training environment can be imported to the convolutional neural network (CNN). The network output will assist with the adjustment of the training strategy, providing an important guarantee to training quality. Normally, two independent CNNs are selected to detect the targets and segment the semantics of the training scene, respectively. However, the labeling of semantic segmentation data is too costly, given the limited number of training samples. It is difficult for the independently trained semantic segmentation model to achieve an ideal effect of image feature extraction. To overcome the difficulty, this paper designs a fusion network model, which improves the image feature extraction effect by mining the deep-seated correlation between target detection and semantic segmentation for images taken in 2D scenes.

In this paper, the deep residual network (DRN) is employed as the feature extraction module for scene images. Let $G(a)$ denote the residual. To prevent network degradation, the proposed deep neural network is transformed into a shallow neural network through the following identity mapping:

$$F(a)=G(a)+a \tag{1}$$

To reduce the difficulty for the neural network model to directly learn identity mapping, formula (1) is converted equivalently into:

$$a=F(a)-G(a) \tag{2}$$

Formula (2) shows that the identity mapping $F(a)=a$ can be constructed, as long as $G(a)=0$ holds. Table 1 shows the network structure of the feature extraction module for the scene images.

Table 1. Network structure of the feature extraction module for the scene images

Module number	0	1	
Layer structure	*Conv*(3, 32, [4, 4])	*MaxPooling*([4, 4])	[*Conv*(64, 32, [3, 3]) *Conv*(64, 32, [3, 3]) ×4 *Conv*(64, 32, [3, 3])]
Modul number	2	3	4
Layer structure	[*Conv*(128, 64, [4, 4]) *Conv*(64, 64, [3, 3])×2 *Conv*(64, 256, [4, 5])]	[*Conv*(256, 128, [2, 2]) *Conv*(128, 128, [5, 5]) ×2 *Conv*(128, 512, [2, 2])]	[*Conv*(512, 64, [2, 2]) *Conv*(256, 256, [2, 2]) ×2 *Conv*(256, 512, [2, 2])]

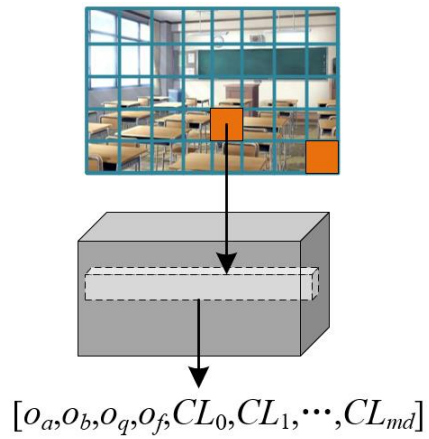


Figure 1. Target-grid mapping

After the feature mapping is completed by the feature extraction module, the center of the target in the scene image will fall within a grid in the feature map (Figure 1). The prediction of the target will be carried out based on grids. During the prediction, the corresponding grid will generate an m-anchor box that approximates the true bounding box for each prediction vector. Let (d_a, d_b) be the coordinates of upper left corner of the grid; (o_a, o_b) be the coordinates of the center of the true bounding box to be predicted; (o_q, o_f) be the size of the true bounding box; o^*_a, o^*_b, o^*_q and o^*_f be the abscissa, ordinate, width, and height of the predicted bounding box, respectively; t_q and t_f be the width, and height of the anchor box, respectively; $\varepsilon(.)$ be the sigmoid function that maps the input to the interval (0, 1). Then, the true bounding box can be predicted based on the information of the anchor box by:

$$\begin{cases} o_a^* = \varepsilon(p_a)+d_a \\ o_b^* = \varepsilon(p_b)+d_b \\ o_q^* = t_q s^{pq} \\ o_f^* = t_f s^{pf} \end{cases} \tag{3}$$

In fact, the neural network needs to predict $\varepsilon(p_a)$, $\varepsilon(p_b)$, s^{pq}, and s^{pf}. Let (p_a, p_b) and (p_q, p_f) be the true coordinates and true size of the predicted bounding box, respectively. After obtaining the values of $\varepsilon(p_a)$, $\varepsilon(p_b)$, s^{pq}, and s^{pf}, (p_a, p_b) and (p_q, p_f) can be restored through reverse deduction by formula (3).

Confidence CL_o is calculated by sigmoid function, and used to judge whether any target exists in the predicted bounding

box. If CL_o>0.5, the target exists in the box; if CL_o<0.5, the target does not exist in the box. The confidence of each type of targets in the scene image, denoted as CL_1~CL_{md}, can also be computed by sigmoid function. The predicted class is the class corresponding to the target with the highest confidence in the scene image:

$$TO(X,Y)=\frac{|X\cap Y|}{|X\cup Y|}=\frac{|X\cap Y|}{|X|+|Y|-|X\cap Y|} \tag{4}$$

Our feature extraction model fuses target detection and semantic segmentation. In the target detection module, the loss function defines four prediction errors: center offset, size, confidence, and class confidence:

$$\begin{aligned} Loss_O &= \sum_{i=1}^{q\times f}\sum_{j=1}^{m}\varphi_{ij}^{O}\left[\left(o_{ai}-o_{ai}^{*}\right)^2+\left(o_{bi}-o_{bi}^{*}\right)^2\right] \\ &+\sum_{i=1}^{q\times f}\sum_{j=1}^{m}\varphi_{ij}^{O}\left[\left(o_{qi}-o_{qi}^{*}\right)^2+\left(o_{fi}-o_{fi}^{*}\right)^2\right] \\ &+\sum_{i=1}^{q\times f}\sum_{j=1}^{m}CL_{0ij}log\left(CL_{0ij}^{*}\right) \\ &+\sum_{i=1}^{q\times f}\sum_{j=1}^{m}\sum_{d=1}^{m_d}CL_{dij}log\left(CL_{dij}^{*}\right) \end{aligned} \tag{5}$$

where, q and f are the width and height of feature map, respectively; ϕ_{ij}^{O} is a binary function (any value greater than 0.5 is set to 1, and any value smaller than 0.5 is set to 0); o_a, o_b, o_q, and o_f are the abscissa, ordinate, width, and height of the true bounding box, respectively; CL_{dij} and CL_{dij}^{*} are the true and predicted class confidences, respectively; CL_{0ij} and CL_{0ij}^{*} indicate whether the true and predicted targets are confident, respectively.

The semantic segmentation module consists of a feature extraction module of the scene images, and a spatial pyramid pooling module containing dilation convolution layers and pooling layers. The flow of the semantic segmentation module is explained in Figure 2. The network structure of spatial pyramid pooling module is given in Table 2.

Table 2. Network structure of spatial pyramid pooling module

Serial number	1	2	3
Structure	*Conv*(1024, 128, [2, 2], *dilate*=2)	*Conv*(1024, 128, [5, 5], *dilate*=7)	*Conv*(1024, 128, [5, 5], *dilate*=14)
Serial number	4	5	
Structure	*Conv*(1024, 128, [5, 5], *dilate*=19) average pooling	*Conv*(1024, 128, [2, 2], *dilate*=2)	

In the spatial pyramid pooling module, the spliced output of each network layer is up-sampled through bilinear interpolation. Let $g(W_{11})$, $g(W_{12})$, $g(W_{21})$, and $g(W_{22})$ be the values of function g(.) at points $W_{11}(a_1, b_1)$, $W_{12}(a_1, b_2)$, $W_{21}(a_2, b_1)$, and $W_{22}(a_2, b_2)$, respectively. To predict the value of g(.) at interpolation point T(a, b), the first step is to perform linear interpolation along the a-axis:

$$\begin{cases} g(V_1)=\dfrac{a_2-a}{a_2-a_1}g(W_{11})+\dfrac{a-a_1}{a_2-a_1}g(W_{21}) \\ g(V_2)=\dfrac{a_2-a}{a_2-a_1}g(W_{12})+\dfrac{a-a_1}{a_2-a_1}g(W_{22}) \end{cases} \tag{6}$$

Then, another linear interpolation should be implemented on points V_1 and V_2 along the b-axis:

$$g(T)=\frac{b_2-b}{b_2-b_1}g(V_1)+\frac{b-b_1}{b_2-b_1}g(V_2) \tag{7}$$

The final result of semantic segmentation can be obtained by splicing the output of the feature extraction module with the output of up-sampling, and performing bilinear interpolation again on the spliced result. Let B be the true label of semantic segmentation data; B* be the predicted semantic segmentation results. In the fusion feature extraction model, the loss function of the semantic segmentation module can be realized by the cross-entropy function below:

$$Loss_{SEM}=\sum_{i=1}^{q}\sum_{j=1}^{f}B_{ij}log\left(B_{ij}^{*}\right) \tag{8}$$

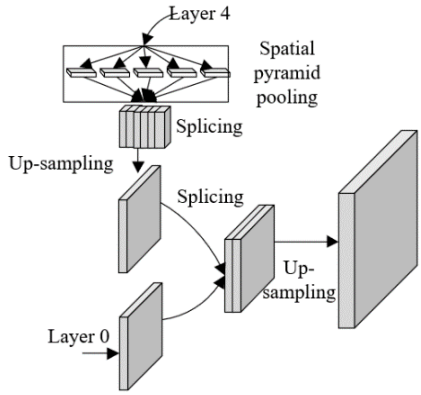


Figure 2. Flow of semantic segmentation module

3. 3D RECONSTRUCTION

The corresponding points on two 2D planar scene images can be constrained by polar lines. However, these points on the 3D reconstructed scene images for virtual training cannot be solved under the constraint of polar lines. To better interact with the scene, and realize fast, accurate, and dense correspondence, this paper solves the vertical and horizontal parallaxes of the scene, and reconstructs the 3D virtual training scene based on depth, using the continuity of the depth of the scene.

3.1 Solving vertical and horizontal parallaxes

Let (0, 0, 0), and (0, 0, -v_1) be the coordinates of U and U_1, respectively; (e_0, r_0) be the coordinates of T_0 in image SA_0. Then, the corresponding coordinates in the global coordinate system can be expressed as T_0(-g sin(e_0), -(F/2-r_0), -g cos(e_0)), where g and F are the focal length and height of the panoramic

image shot in scene SA_0, respectively. The polar plane passing through T_0UU_1 can be described by:

$$\begin{vmatrix} a & b & c \\ -g\ sin(u_0) & -(F/2-r_0) & -g\ cos(e_0) \\ 0 & 0 & -v \end{vmatrix} = 0 \tag{9}$$

The cylindrical surface with U_1 as the center can be expressed as:

$$(e_0, r_0)a^2 + (c+v)^2 = g^2 \tag{10}$$

Formula (10) can be converted into a parametric equation:

$$\begin{cases} a = -gsin(\omega) \\ c = -(g(\omega)+v) \end{cases} (0 \le \omega \le 2\pi) \tag{11}$$

Point T_1 must exist on the quadratic curve, where the cylindrical surface with U_1 as the center intersects the polar plane. Combining formulas (9) and (11), the intersecting line can be expressed as:

$$b = -\frac{(F/2-r_0)sin(\omega)}{sin(e_0)} (0 \le \omega \le 2\pi) \tag{12}$$

The same vertical parallax can be obtained by restoring the depth of any point on SA_0, based on image SA_2. Figure 3 presents the principle of calculating the vertical parallax.

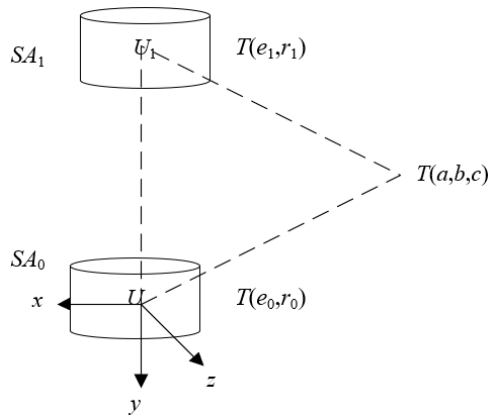


Figure 3. Calculation principle of vertical parallax

Suppose $T_0(x_2, r_i)$ is the right adjacent pixel of $T_0(x_1, r_i)$ in SA_0, $T_1(\alpha_1, r_{i1})$ be the corresponding point of $T_0(\alpha_1, r_i)$ in SA_1, and $T'_1(\alpha_1, r'_{i1})$ be the corresponding point of $T'_0(\alpha_1, r'_{i1})$ in SA_1. Let δ_1 and δ_2 be the depth and height of T_0, respectively. Provided that $l=\delta_2/\delta_1$, we have:

$$\delta_1 = e_1 sin(\alpha_1)/sin(\alpha_1 - \beta_1) \tag{13}$$

$$\delta_2 = e_1 sin(\alpha_2)/sin(\alpha_2 - \beta_2) \tag{14}$$

Given $\delta_2=l\delta_1$, formulas (13) and (14) can be combined into:

$$\frac{le_1 sin(\alpha_1)}{sin(\alpha_1 - \beta_1)} = \frac{e_1 sin(\alpha_2)}{sin(\alpha_2)cos(\beta_2) - cos(\alpha_2)sin(\beta_2)} \tag{15}$$

Let Q be the width of SA_0. Substituting $\beta_1=\beta_2-2\pi/Q$ into formula (15):

$$\alpha_2 = arctg\left(\begin{matrix} l\ sin(\alpha_1)sin(\beta_2) \\ /\left(\begin{matrix} l\ cos(\beta_2)sin(\alpha_1) \\ -sin(\alpha_1 - \beta_2 + 2\pi/Q) \end{matrix} \right) \end{matrix} \right) \tag{16}$$

It can be seen that the range of T'_1 is related to β, α_1, and l. Formula (16) can be simplified as:

$$\alpha_2(l) = arctg\left(\begin{matrix} l\ sin(\alpha_1)sin(\beta_2) \\ /\left(\begin{matrix} l\ cos(\beta_2)sin(\alpha_1) \\ -sin(\alpha_1 - \beta_2 + 2\pi/Q) \end{matrix} \right) \end{matrix} \right) \tag{17}$$

If $0<\beta<\pi$, then $\alpha(l)\le\alpha_2\le\alpha(1/l)$. Otherwise, if $\alpha(l)<0$ or $\alpha(1/l)<0$, the angle should be adjusted by $\alpha(l)+\pi$ or $\alpha(1/l)+\pi$. When $\pi<\beta<2\pi$, if $\alpha(l)>0$ and $\alpha(1/l)>0$, then $\alpha(1/l)+\pi\le\alpha_2\le\alpha(l)+\pi$. Otherwise, the angle should be adjusted accordingly. Similarly, we have:

$$\alpha'_2 = arctg\left(\begin{matrix} l\ sin(\alpha'_1)sin(\beta_2) \\ /\left(\begin{matrix} l\ cos(\beta_2)sin(\alpha'_1) \\ +sin(-\alpha'_1 + \beta_2 - 2\pi/Q) \end{matrix} \right) \end{matrix} \right) \tag{18}$$

$$\alpha'_2(l) = arctg\left(\begin{matrix} l\ sin(\alpha'_1)sin(\beta_2) \\ /\left(\begin{matrix} l\ cos(\beta_2)sin(\alpha'_1) \\ -sin\left(\begin{matrix} -\alpha'_1 - \beta_2 \\ -2\pi/Q \end{matrix} \right) \end{matrix} \right) \end{matrix} \right) \tag{19}$$

Figure 4 presents the top view of horizontal parallax calculation.

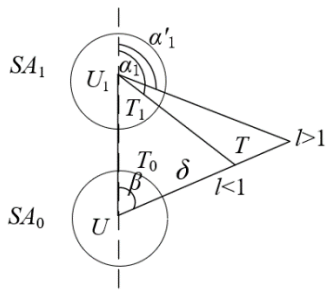


Figure 4. Top view of horizontal parallax calculation

3.2 Depth-based 3D reconstruction

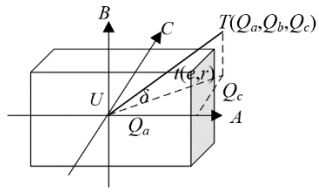


Figure 5. Calculation of the 3D coordinates of the scene image

Suppose $SA_{N\times M}$ is the generated scene image of virtual talent training, whose resolution is $N\times M$. The depth image of $SA_{N\times M}$ is denoted as $\delta_{N\times M}$ (Figure 5). Taking the center of the panoramic image shot in the scene as the origin, a regular coordinate system *U-ABC* is constructed. In addition, the camera coordinate system of *SA* is established as *S-ERQ*. It is assumed that the origins of the two coordinate systems coincide. Let $VP(VP_a, VP_b, VP_c)$ be the position of the view point. For any pixel $t(e, r)$ in the panoramic image, its coordinates in *U-ABC* can be recorded as $T(Q_a, Q_b, Q_c)$. Let T' be the projection of T on AUC plane; δ be the depth of point T solved by quadratic polar curve. Then, we have:

$$\begin{aligned} Q_a &= \delta cos\left(\frac{2\pi e}{Q}\right) \\ Q_c &= \delta sin\left(\frac{2\pi e}{Q}\right) \\ Q_b &= \frac{\delta}{g}\left(\frac{F}{2}-r\right) \end{aligned} \tag{20}$$

where, g can be calculated through calibration. Any four adjacent pixels $T(i,j)$, $T(i+1,j)$, $T(i+1,j+1)$ and $T(i,j+1)$ of the panoramic image, plus the corresponding view points $T(i, j)$, $T(i+1, j)$, $T(i+1, j+1)$, and $T(i, j+1)$, form a space quadrangle, i.e., the reconstructed 3D scene.

On the panoramic scene image, if there exists an adjacent pixel on the same plane with pixel $T(e, r)$, then the depth difference between $T(e, r)$ and that pixel should be constant. In the real training scene space, the depth δ mutates only on the edge of a plane. Thus, this paper adopts a second-order differential operator to process the image:

$$\begin{aligned} &\Delta^2\delta(i,j) = 4\delta(i,j)-\delta(i,j-1) \\ &-\delta(i,j+1)-\delta(i-1,j)-\delta(i+1,j) \end{aligned} \tag{21}$$

The resulting second-order differential image contains many areas with the gradient of zero. For the special pixels in the noisy depth image, this paper sets a relatively small threshold, which is always greater than the second-order difference of the pixels in large planar areas of the panoramic scene image. In this way, the similarly judgment can be completed on the special pixels.

Firstly, it is necessary to establish the covariance matrix of all points r_i in the $l*l$ neighborhood of any special point T' on the panoramic scene image:

$$DE = \sum_{i=1}^{M}\left((r_i-C)^T\cdot(r_i-C)\right) \tag{22}$$

where, C is the centroid of the adjacent point set. The offset of point T' from the fitted plane is characterized by the minimum eigenvalue of the covariance matrix. If the offset is smaller than the preset threshold, then point T' and its adjacent points both belong to the fitted plane, and the normal vector of that plane is the corresponding eigenvector. This operation can effectively eliminate the poorly fitted points, and obtain the normal vector of the fitted pixels. Let L be the number of randomly selected points; S_j be the normal vector of the fitted plane at the j-th random point. Then, the normal vector NV of the i-th initial fitted plane can be calculated by:

$$NV_i = \frac{1}{L}\sum_{j=1}^{L} S_j \tag{23}$$

Let NV_1 and NV_2 be the normal vectors of two adjacent fitted planes in the reconstructed 3D scene, respectively; THR be the preset threshold. For the two planes to merge into one plane, the following condition must be satisfied:

$$\left|NV_1\times NV_2\right| < THR \tag{24}$$

The above method can satisfactorily segment the panoramic scene image based on planar features.

The surfaces in the training scene are reconstructed in the following manner. Firstly, the panoramic scene image is segmented based on planar features. After that, the grids of the scene are reconstructed through triangular expanding. Let T be the point in the 3D space corresponding to the center pixel of any region; NV' be the normal vector of the fitted plane for the region. Then, the four spatial triangles adjacent to T are tested. Let NV_i' be the normal vector of the i-th spatial triangle. Then, the seed triangle can be expressed as:

$$value = \arg\min_{0\le i\le 3}\left|NV'\times NV_i'\right| \tag{25}$$

Let $T_1(a_1, b_1, c_1)$ and $T_2(a_2, b_2, c_2)$ be the 3D coordinates of points T_1 and T_2, respectively; NV'_{SC} be the normal vector of the fitted plane for the semi-circular search area; $T_i(a_i, b_i, c_i)$ be the 3D coordinates of any unexpanded pixel in that region. If there is a boundary point on the fitted plane, any boundary point will be taken as the new vertex of the plane; otherwise, the point corresponding to the minimum of the following formula will be taken as the new vertex of the plane:

$$\begin{aligned} &NV_{SC}^{'} = \{a_1-a_i, b_1-b_i, c_1-c_i\} \\ &\vec{l} = \{a_2-a_i, b_2-b_i, c_2-c_i\}, \\ &V_{NEW} = min\left(\left|NV_{SC}^{'}\times\left(NV_{SC}^{'}\times\vec{l}\right)\right|\right) \end{aligned} \tag{26}$$

The above analysis shows that the resolution of the triangular grid model is directly influenced by the length of the extension lines of T_1 and T_2. The longer these lines, the larger the semi-circular search area of the new vertex, and the better the resolution of the generated triangular grid model.

4. EXPERIMENTS AND RESULTS ANALYSIS

Figure 6 shows the variation of the IoU of the scene image set with the growing number of iterations. The proposed model, which fuses target detection with semantic segmentation, went through four rounds of training and four rounds of testing. As shown in Figure 6, the proposed model achieved a better effect of semantic segmentation, when it was trained by the auxiliary data source, i.e., the target detection data.

Table 1 shows how the target confidence varies of different training scenes after the addition of semantic segmentation data. The experimental results show that, when the target had sufficient instances in the training set of scene images, the introduction of semantic segmentation could provide pixel-level labels for the target. Then, the trained model had a relatively high prediction confidence for such a target. By contrast, when the target had insufficient instances, the model would have a relatively low prediction confidence.

Table 2 shows the semantic segmentation test results of our model. The model performance was evaluated by seven metrics, including maximum F1-score, mAP, Precision, Recall, FPR, and FNR. Models 1-4 are respectively the proposed model, weak supervised semantic segmentation model, region-based semantic segmentation model, and fully convolutional network (FCN)-based semantic segmentation model. Two types of test sets were experimented, namely, outdoor training scene, and indoor training scene. The results show that our model outperformed the other models in maximum F1-score and Recall, and achieved the lowest FNR.

Both indoor and outdoor training scenes were tested. Three modes were designed for each scene: no dynamic target, a single dynamic target, and multiple dynamic targets. Tables 3 and 4 present the 3D reconstruction results of indoor and outdoor training scenes, respectively. It can be seen that the indoor training scene was reconstructed better than the outdoor training scene. The scenes with a single dynamic target were reconstructed better than those with multiple dynamic targets. The results confirm that our model can accurately identify 3D targets.

Figure 7 compares the mean back projection error of our method with that of traditional incremental reconstruction. The image adoption rate of our method reached 77.6%, which is 23.1% higher than that of incremental reconstruction. The mean back projection error of our method was around 0.5 pixel, which is 0.1 pixel smaller than that of the other method. The comparison further verifies the effectiveness of our reconstruction method.

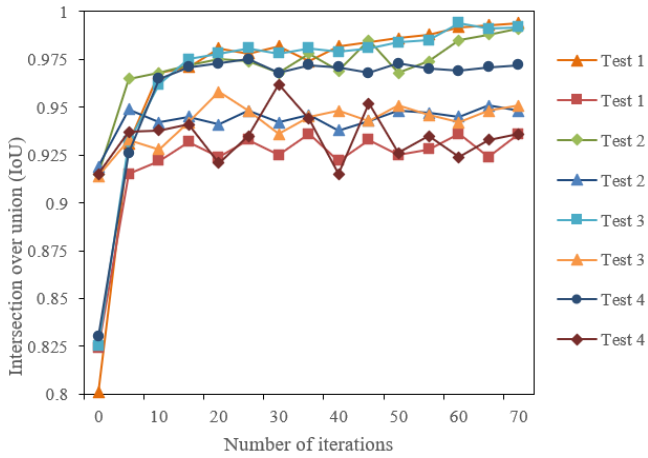


Figure 6. IoU curve of the scene image set

Table 1. Target confidence variation in different training scenes

Target class	Equipment	People	Desks and chairs
Confidence 1	0.9253	0.8549	0.712
Confidence 2	0.9855	0.9316	0.6827
Target class	Blackboard	Digital screen	Others
Confidence 1	0.8255	0.8746	0.7418
Confidence 2	0.7848	0.7318	0.6685

Table 2. Test results of semantic segmentation

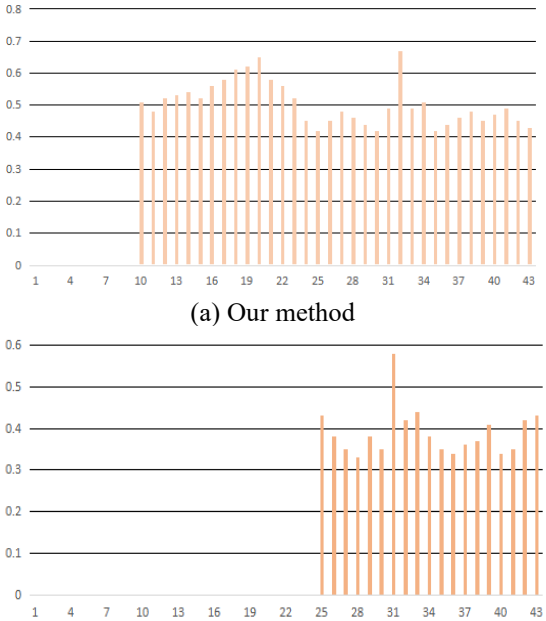
Image type	Outdoor				Indoor			
Model	1	2	3	4	1	2	3	4
Maximum F1-score	93.25%	92.35%	91.75%	90.75%	96.75%	92.18%	94.26%	91.37%
Mean average precision (mAP)	87.18%	84.27%	83.28%	85.74%	88.44%	89.48%	91.45%	92.37%
Precision	85.17%	91.22%	88.52%	92.38%	92.68%	94.27%	95.38%	93.27%
Recall	98.24%	95.48%	93.28%	96.15%	95.37%	99.22%	92.35%	88.29%
False positive rate (FPR)	6.14%	4.11%	5.36%	4.25%	8.42%	6.85%	5.39%	6.24%
False negative rate (FNR)	3.82%	6.75%	6.59%	8.48%	1.78%	8.48%	6.92%	11.48%

Table 3. Reconstruction results of outdoor training scene

	Time consumption				Evaluation metrics		
	Scene depth calculation	3D coordinate calculation	Image segmentation	Triangulation	Maximum depth	Mean depth	Maximum error
No dynamic target	0.326	33.284	18.249	263.458	3.29585	0.12517	0.0748516
Single dynamic target	0.395	33.265	20.448	243.585	2.36258	0.152475	0.0518465
Multiple dynamic targets	0.362	33.451	21.367	258.162	1.02575	0.184633	0.0144756

Table 4. Reconstruction results of indoor training scene

	Time consumption				Evaluation metrics		
	Scene depth calculation	3D coordinate calculation	Image segmentation	Triangulation	Maximum depth	Mean depth	Maximum error
No dynamic target	0.362	32.158	16.285	162.37	1.25814	0.152485	0.045125
Single dynamic target	0.369	32.485	18.296	162.74	1.62835	0.132854	0.0484257
Multiple dynamic targets	0.355	31.4564	17.214	161.12	1.5474	0.142451	0.0387456



(a) Our method

(b) Incremental reconstruction

Figure 7. Mean back projection errors of different reconstruction methods

5. CONCLUSIONS

In this paper, 3D image reconstruction is studied in the context of virtual talent training scene. To improve image feature extraction, a fusion network model was designed to mine the deep-seated correlation between target detection and semantic segmentation for 2D scene images. On this basis, the vertical and horizontal parallaxes of the scene were solved, and the depth-based virtual talent training scene was reconstructed three dimensionally, based on the continuity of scene depth. Drawing on experimental results, the authors plotted the variation curve of the IoU of scene image set with the growing number of iterations, presented the target confidence change of different training scene images, and obtained the semantic segmentation test results. The relevant results confirm that our fusion model achieved better maximum F1-score and Recall than the other models, and realized the lowest FNR among all contrastive models. Finally, the reconstruction results of indoor and outdoor training scenes were collected, and the mean back projection errors of different reconstruction methods were summarized, which further demonstrate the effectiveness of our reconstruction method.

ACKNOWLEDGEMENTS

Supported by the Science and Technology Project of State Grid Shandong Electric Power Company: Research on Human Resource Intelligent Decision Analysis and Early Warning Technology Application Based on “New Heights of Talent Development” (Grant No.: 5206002000UR).

REFERENCES

[1] Lele, A. (2013). Virtual reality and its military utility. Journal of Ambient Intelligence and Humanized Computing, 4(1): 17-26. https://doi.org/10.1007/s12652-011-0052-4

[2] Marsili, M. (2021). Epidermal systems and virtual reality: Emerging disruptive technology for military applications. In Key Engineering Materials, 893: 93-101. https://doi.org/10.4028/www.scientific.net/KEM.893.93

[3] Gawlik-Kobylińska, M., Maciejewski, P., Lebiedź, J., Wysokińska-Senkus, A. (2020). Factors affecting the effectiveness of military training in virtual reality environment. In Proceedings of the 2020 9th International Conference on Educational and Information Technology, pp. 144-148. https://doi.org/10.1145/3383923.3383950

[4] Kot, T., Novák, P. (2018). Application of virtual reality in teleoperation of the military mobile robotic system TAROS. International Journal of Advanced Robotic Systems, 15(1): 1729881417751545. https://doi.org/10.1177/1729881417751545

[5] Bhagat, K.K., Liou, W.K., Chang, C.Y. (2016). A cost-effective interactive 3D virtual reality system applied to military live firing training. Virtual Reality, 20(2): 127-140. https://doi.org/10.1007/s10055-016-0284-x

[6] Georgieva-Tsaneva, G., Serbezova, I. (2020). Virtual Reality and Serious Games Using in Distance Learning in Medicine in Bulgaria. International Journal of Emerging Technologies in Learning (iJET), 15(19): 223-230.

[7] Sabalic, M., Schoener, J.D. (2017). Virtual reality-based technologies in dental medicine: knowledge, attitudes and practice among students and practitioners. Technology, Knowledge and Learning, 22(2): 199-207. https://doi.org/10.1007/s10758-017-9305-4

[8] Krpic, A., Savanovic, A., Cikajlo, I. (2014). Impact of virtual-reality feedback on human balance training when using a haptic support surface in rehabilitation medicine/Vpliv navidezne resnicnosti kot povratne informacije na vadbo ravnotezja cloveka ob uporabi hapticnih tal v rehabilitacijski medicini. Elektrotehniski Vestnik, 81(1/2): 15.

[9] Scerbo, M.W. (2004). Medical virtual reality simulation: Enhancing safety through practicing medicine without patients. Biomedical Instrumentation & Technology, 38(3): 225-228. https://doi.org/10.2345/0899-8205(2004)38[225:MVRSES]2.0.CO;2

[10] Law, L. (2002). Medicine: The new frontier for virtual reality. Advanced Imaging, 17(6): 36-37.

[11] Kamińska, D., Zwoliński, G., Wiak, S., Petkovska, L., Cvetkovski, G., Barba, P.D., Anbarjafari, G. (2021). Virtual reality-based training: Case study in mechatronics. Technology, Knowledge and Learning, 26(4): 1043-1059. https://doi.org/10.1007/s10758-020-09469-z

[12] Yin, J., Ren, H., Zhou, Y. (2021). The whole ship simulation training platform based on virtual reality. IEEE Open Journal of Intelligent Transportation Systems, 2: 207-215. https://doi.org/10.1109/OJITS.2021.3098932

[13] McIntosh, J. (2019). Virtual reality training immers students in welding skills. Welding Journal, 98(8): 44-47.

[14] Abidi, M.H., Al-Ahmari, A., Ahmad, A., Ameen, W., Alkhalefah, H. (2019). Assessment of virtual reality-based manufacturing assembly training system. The International Journal of Advanced Manufacturing Technology, 105(9): 3743-3759.

https://doi.org/10.1007/s00170-019-03801-3
[15] Dodoo, E.R., Hill, B., Garcia, A., Kohl, A., MacAllister, A., Schlueter, J., Winer, E. (2018). Evaluating commodity hardware and software for virtual reality assembly training. Electronic Imaging, 2018(3): 468-1. https://doi.org/10.2352/ISSN.2470-1173.2018.03.ERVR-468
[16] Pereira, R.E., Gheisari, M., Esmaeili, B. (2018). Using panoramic augmented reality to develop a virtual safety training environment. In Construction Research Congress 2018, 29-39.
[17] Cecil, J., Gupta, A., Pirela-Cruz, M., Ramanathan, P. (2018). A network-based virtual reality simulation training approach for orthopedic surgery. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14(3): 1-21. https://doi.org/10.1145/3232678
[18] Choi, J.Y., Lee, J.H., Kim, Y.S., Kim, S. (2015). Virtual-reality-based operation training system for steel making process. Journal of Institute of Control, Robotics and Systems, 21(8): 709-712. https://doi.org/10.5302/J.ICROS.2015.15.0081
[19] Mikami, D., Takahashi, K., Saijo, N., Isogawa, M., Kimura, T., Kimata, H. (2018). Virtual reality-based sports training system and its application to baseball. NTT Technical Review, 16(3).
[20] Jiang, M., Zhou, G., Zhang, Q. (2018). Fire-fighting training system based on virtual reality. In IOP Conference Series: Earth and Environmental Science, 170(4): 042113. https://doi.org/10.1088/1755-1315/170/4/042113
[21] Maidenbaum, S., Amedi, A. (2015). Blind in a virtual world: Mobility-training virtual reality games for users who are blind. In 2015 IEEE Virtual Reality (VR), pp. 341-342. https://doi.org/10.1109/VR.2015.7223435
[22] Intraraprasit, M., Sunhem, W., Jinjakam, C. (2018). Interaction behavior of older adults with immersive virtual reality application for cognitive training. In 2018 3rd International Conference on Computer and Communication Systems (ICCCS), pp. 506-510. https://doi.org/10.1109/CCOMS.2018.8463223
[23] Ruthenbeck, G.S., Reynolds, K.J. (2015). Virtual reality for medical training: The state-of-the-art. Journal of Simulation, 9(1): 16-26. https://doi.org/10.1057/jos.2014.14
[24] Chao, C., Chalouhi, G.E., Bouhanna, P., Ville, Y., Dommergues, M. (2015). Randomized clinical trial of virtual reality simulation training for transvaginal gynecologic ultrasound skills. Journal of Ultrasound in Medicine, 34(9): 1663-1667. https://doi.org/10.7863/ultra.15.14.09063
[25] Grabowski, A., Jankowski, J. (2015). Virtual reality-based pilot training for underground coal miners. Safety Science, 72: 310-314. https://doi.org/10.1016/j.ssci.2014.09.017
[26] Bellemans, M., Lamrnens, D., De Sloover, J., De Vleeschauwer, T., Schoofs, E., Jordens, W., Van Steenhuyse, B., Mangelschots, J., Selleri, S., Hamesse, C., Freville, T., Haeltermani, R. (2020). Training Firefighters in Virtual Reality. 2020 International Conference on 3D Immersion, IC3D 2020 - Proceedings, December 15, 2020.
[27] Gluck, A., Chen, J., Paul, R. (2020). Artificial intelligence assisted virtual reality warfighter training system. In 2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), pp. 386-389. https://doi.org/10.1109/AIVR50618.2020.00080
[28] Chen, Z., Cao, Z., Ma, P., Xu, L. (2020). Industrial robot training platform based on virtual reality and mixed reality technology. In International Conference on Man-Machine-Environment System Engineering, pp. 891-898. https://doi.org/10.1007/978-981-15-6978-4_102
[29] Guptaa, A., Vargheseb, K. (2020). Scenario-based construction safety training platform using virtual reality. In ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction, 37: 892-899.
[30] Khwanngern, K., Tiangtae, N., Natwichai, J., Kattiyanet, A., Kaveeta, V., Sitthikham, S., Kammabut, K. (2019). Jaw surgery simulation in virtual reality for medical training. In International Conference on Network-Based Information Systems, pp. 475-483. https://doi.org/10.1007/978-3-030-29029-0_45
[31] Bhange, D., Dethe, C. (2020). Performance optimization of LS/LMMSE using swarm intelligence in 3D MIMO-OFDM systems. Traitement du Signal, 37(1): 107-112. https://doi.org/10.18280/ts.370114
[32] Özbay, E., Çınar, A. (2019). A comparative study of object classification methods using 3D Zernike moment on 3D point clouds. Traitement du Signal, 36(6): 549-555. https://doi.org/10.18280/ts.360610