How to display products in the Metaverse? A scheme based on improved NeRF 3D reconstruction.

Yu Jincheng
Faculty of Science and Technology
University of Macau
https://orcid.org/0000-0002-3100-5838

***Abstract*—Metaverse and virtual store are the development direction of future e-commerce market. In order to solve the problem of high cost and low quality of 3D model production of products in the current virtual store, which leads to poor shopping experience for customers, this paper proposes a method of 3D reconstruction of products based on improved NeRF. Merchants only need to collect pictures of products from different angles to quickly complete high-quality 3D reconstruction of products, which provides technical support for the display of products in the future virtual store.**

Keywords—Metaverse, 3D reconstruction, E-commerce, NeRF, Computer vision

I. INTRODUCTION

A. Introduce of Virtual marketplace

With the development of Internet technology, people's shopping channels have changed from offline stores before the emergence of electronic markets to online shopping with the help of powerful electronic shopping platforms. Today's electronic shopping platform has been very perfect in terms of functions [9], which can basically meet the needs of online shopping. With the demand of consumers for consumption experience, the integration degree of e-commerce market is getting higher and higher, and the virtual shopping platform based on the metaverse has become the development trend of future e-commerce. Shen et al. [10] conducted a comprehensive study on virtual commerce from two aspects of application design and consumer behaviour research in the metaverse. On this basis, future research directions are proposed, including investigating the diversity of boundary factors and immersive technology, forming an organic behavioural design researcher circle, virtual consumption, and the evolution trend of virtual worlds. The study of Luna-Nevarez et al. [11] pointed out that immersion, enjoyment, trust and VR self-efficacy have potential indirect effects on consumers' purchase intention and company visit intention. The enhanced functionality and interactivity of the metaverse can solve the problem of lack of contact and face-to-face interaction with products in current e-commerce. Xi Et al. [13] pointed out that future studies should increase the interaction between shoppers and objects in the virtual reality shopping environment to create a more natural and real simulated shopping experience. Future researchers should consider creating more natural virtual shopping scenarios and allowing users to interact with products and environments as necessary, such as by designing interactive programs to rotate, move, touch, drop, pick and shake virtual goods, as well as using more advanced and versatile interactive devices. Therefore, the virtual store in the meta-universe can further expand the functions of the e-commerce market, enhance the interaction with users, enhance the user's immersion and thus affect the user's consumption intention. Virtual immersion technology based on computer vision is a key technology to support the future metaverse e-commerce market [12].

B. Introduce of 3D reconstruction

Three-dimensional reconstruction refers to the process of reconstructing and restoring the shape, position and texture of an object in 3D space by collecting multi-view images or other effective information of the object. The basic method is to use the 2D image information obtained from different perspectives to restore the 3D geometry, surface texture and other information of the object through technical means such as computer vision and computer graphics, so as to realize the 3D digital representation of the object. Traditional 3D reconstruction methods include: traditional 3D reconstruction based on depth image (DRGB) [1-3], and traditional Multi-View Stereo (MVS) [4-6]. Depth image-based methods use the depth images of the scene collected by a depth camera or a lidar to reconstruct the scene, where the RGBD camera generally has an RGB sensor and a depth sensor, so it can collect depth information and colour information at the same time. By combining the two kinds of information for reconstruction, the accuracy of registration can be improved, the reconstructed model has colour attributes, and the realism and reduction of reconstruction can be increased. The traditional multi-angle reconstruction method is pure RGB modelling of 3D reconstruction through RGB images from different angles, by collecting the scene image and calculating the camera pose, and then performing dense reconstruction (MVS) of the scene through the image and the corresponding pose information, and finally performing model UV mapping. With the continuous development of deep learning, its excellent performance, ability to deal with multi-tasks and elegant and efficient end-to-end pipeline mode make 3D reconstruction methods based on deep learning become a research hotspot. Neural Radiance Fields (NeRF) is a 3D reconstruction method based on deep learning [7], which is a deep learning 3D reconstruction method based on multi-angle images. NeRF is different from traditional 3D reconstruction methods, which represent the scene as a point cloud, grid or voxel display. It uses an implicit representation to map the scene to a five-dimensional radiance field neural network. The neural network model is trained from different views of the scene and used for scene rendering in the later stage. Due to the large time overhead of traditional nerf training, Müller et al. [8] improved the traditional NeRF, optimized the neural network training process of NeRF, and significantly improved the training speed of nerf under the premise of ensuring the quality.

C. Our work introduction

There are many computer vision research directions applied to the virtual e-commerce market in the metaverse, such as the construction of virtual scenes, the 3D reconstruction of goods, and the interaction between users and commodity models. This paper focuses on the construction of 3D models of commodities. The interactive 3D model of commodity is the core of the virtual mall, and the user shopping in the virtual mall must interact with the commodity. The quality of the reconstructed 3D model of commodity directly affects the user's shopping experience, and is also the basic measure of the level of immersion in the virtual mall. Virtual shopping is divided into four main levels, Input device, Output device, Representation, Interaction[14]. This paper focuses on the representation level, innovatively applies the improved nerf method to the product modelling of the virtual store in the metaverse, and proposes the process framework for the display of goods in the virtual store in the metaverse.

In the existing virtual market, the modelling of virtual goods is relatively rough [14-16]. In the traditional virtual shop display scheme, the goods are manually modelled and tiled, and the high-precision model generally has a large file and a high threshold for modelling. Therefore, some virtual shops choose to use simple low-precision models and low-resolution maps considering cost and performance, which makes the virtual goods look very rough and greatly affects the user experience. The starting point of this paper is to reduce the cost and threshold of the virtual commodity model, and use the improved NeRF to model the commodity, so that sellers only need to upload the product image to complete the 3D model of the product and put it on the metaverse virtual store.

II. METHOD

A. Overview

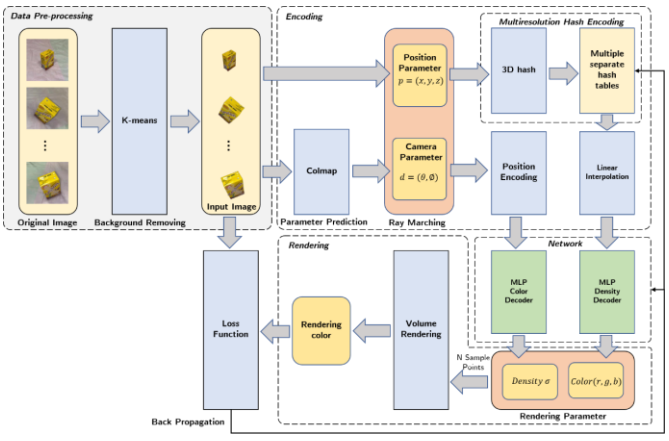

Fig. 1. The figure of the improved neural radiance fields method

As shown in Fig. 1, our method contains four main modules: Data pre-processing module, Encoding module, Network module and Rendering module. Among them, the Data pre-processing module is responsible for the background removal processing of multi-angle commodity pictures and the removal of pictures with high similarity. The Encoding module is responsible for encoding the camera pose and the position of the spatial sampling point corresponding to the image. The Rendering module performs volume rendering according to the density and colour mapped by the camera pose and spatial sampling points, and generates the picture of the corresponding perspective. Finally, the loss value between the rendered image and the real image is calculated and backpropagated to update the network and encoder parameters.

B. Data pre-processing

The data pre-processing in this paper uses the k-means algorithm twice to process the multi-angle product pictures. The first clustering divides the images into n classes using the intersection ratio between each image as the distance.

$$Distant = IoU$$

The images in each category are the group with the highest similarity, and only one image of each category is encoded in the subsequent encoding. The second clustering operation uses the colour and position of each pixel in each image as a distance to classify the pixels in the image into two classes, one is the product and the other is the background.

$$Distant = \alpha Pos + \beta RGB, \alpha \text{ and } \beta \text{ are weights}$$

A mask is generated based on the clustering results, and only commodity pixels are processed in the subsequent operations.

C. Position Encoding

Deep networks are biased toward learning lower frequency functions. Therefore mapping the inputs to a higher dimensional space using high-frequency functions before passing them to the network enables better fitting of data that contains high-frequency variation. The input to NeRF is a 5-dimensional vector. The three dimensions (x,y,z) represent the location of the sampling point, and (θ,ϕ) represent the observation direction at that point. So directly using a five-dimensional vector input to the network will not work well. NeRF does this by positionally encoding the initial five-dimensional vector:

$$\gamma(p) = (\sin(2^0\pi p), \cos(2^0\pi p), \ldots, \sin(2^{L-1}\pi p), \cos(2^{L-1}\pi p))$$

The five-dimensional vector (x,y,z,θ,ϕ) is encoded into a 2L vector.

D. Multiresolution hash encoding

We take the same space and represent it with different sizes of grids. Then we set the size T of the hash table to a fixed value. In this way, we obtain M hash tables of size T, which are combined to form multiple separate hash tables. When the grid is less than 64, there is always no collision, and the grid above 64 is ignored. The final density is obtained by mixing the density of each resolution grid with different

weights, which are optimized by backpropagation after finding the loss.

E. Rendering

Given the position (x,y,z) and the viewing direction (θ,ϕ) in the scene, the neural network will output a (c,σ), which represents the color c of the spontaneous light point in that direction and the voxel density σ of the point, and then render it using the following rendering equation:

$$C(r) = \int_{t_n}^{t_f} T(t)\sigma\big(r(t)\big)c(r(t), d)\, dt$$

$$T(t) = \exp\left(-\int_{t_n}^{t} \sigma\big(r(s)\big)ds\right)$$

The function T(t) denotes the accumulated transmittance along the ray from tn to t that is, the probability that the ray travels from tn to t without hitting any other particle. Rendering a view from our continuous neural radiance field requires estimating this integral C(r) for a camera ray traced through each pixel of the desired virtual camera.

After getting the picture under the perspective of the input camera pose by rendering, the model can be optimized by losing the result and the ground truth. The loss function is as follows:

$$Loss = \sum_{r\in R} \left\| \hat{C}(r) - C(r) \right\|_2^2$$

F. Metaverse virtual store demonstrate 3D product process

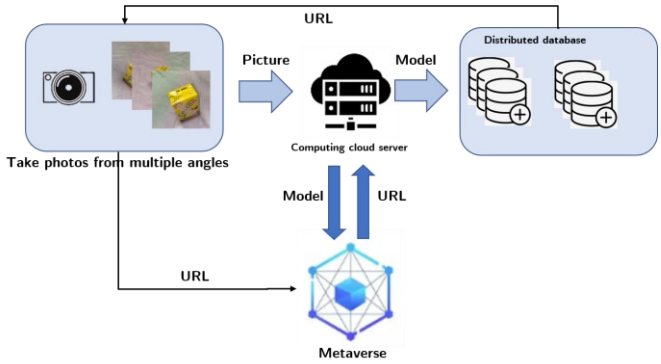

Fig. 2. System framework

Firstly, multi-angle photos or videos of products are collected through mobile terminal devices, such as mobile phones. If the video is collected, the multi-angle photos of the commodity are extracted from the video at intervals through the local client. The multi-angle photos of the commodity are uploaded to the cloud computing power server for 3D reconstruction, and the reconstructed model is stored in the distributed database, and the URL of the model is returned. At the same time, the corresponding interface should be developed in the metaverse. When the user enters the URL of the product model, the 3D model can be retrieved and displayed in the virtual store.

III. Experiments and Analysis

A. Dataset

The experiment in this paper uses the data of two products, one is Vita lemon tea and the other is Wusu beer, which contain 37 and 18 pictures taken from different angles respectively. The resolution of each image is 1280*720.

B. experimental result

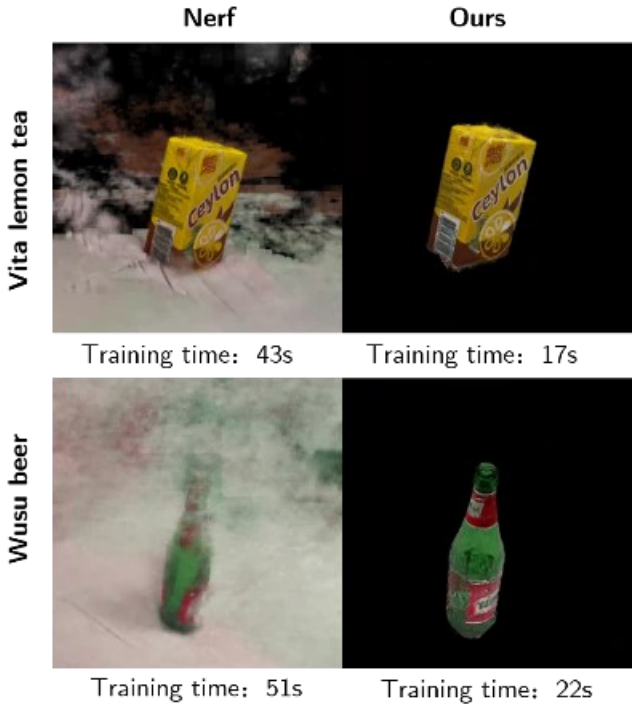

Fig. 3. With background removal module

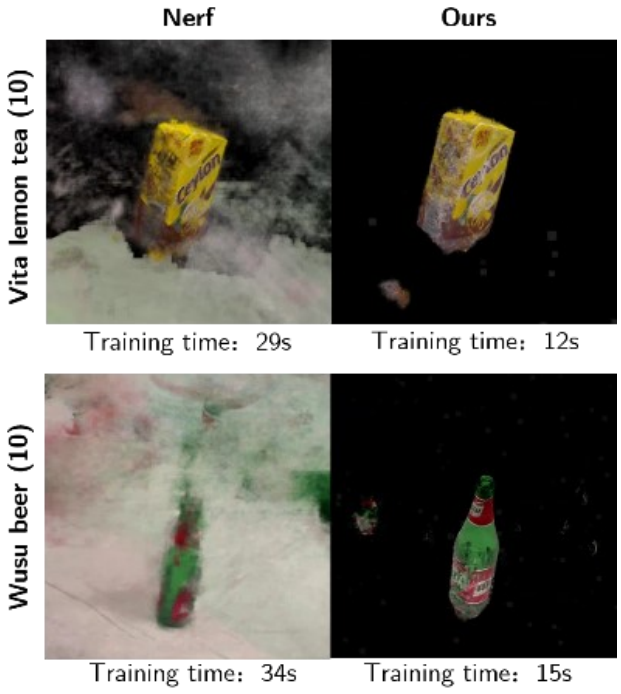

Fig. 4. With background removal module and similar images removal module

Fig. 3 shows a schematic comparison of traditional NeRF and our improved NeRF with only background removal module. It can be found that the reconstruction effect of our method in the main part of the commodity is improved compared with the traditional NeRF, because the traditional NeRF is disturbed by the background in the process of model optimization, and the main body fitting is affected, so there will be a large number of cloud-like points in the scene. Our method solves this problem well. At the same time, we also have advantages over traditional methods in model convergence time, because the pixels of the background part are not involved in the training optimization, so a lot of resources are saved.

Fig. 4 shows the comparison of traditional NeRF method and our improved NeRF with both background removal module and similar image removal module. We can see that both methods reduce the quality of the model compared to the previous results due to fewer images being used for training. The noise of the traditional method is further aggravated, and the fog noise caused by the poor fitting effect of our method also appears. However, from the perspective of model convergence time and training cost, there is a certain progress compared with the previous experiment.

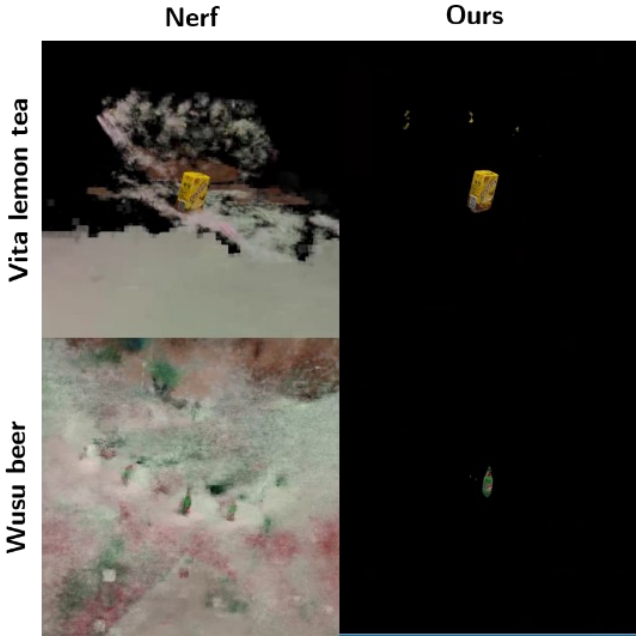

Fig. 5. Scene rendering situation diagram

Fig.5 shows the rendering of the whole scene, and it can be seen that the traditional method without the background removal module is very seriously disturbed by the background, and the commodity subject model fitting is seriously affected. Our method solves this problem well by adding an unsupervised background removal module. And different from introducing other supervised segmentation methods, we use clustering to plug and play, do not need to increase the training cost, and at the same time, the improvement effect is significant.

IV. Conclusion

This paper analyzes the problems of commodity display in the current virtual market and proposes a method based on improved NeRF. By introducing two clustering modules, this method can not only solve the problem of poor fitting of the commodity subject caused by background interference in the training process of NeRF, but also save training cost and time by simplifying the training set of images. At the same time, the generated 3D model of the commodity maintains excellent quality. For future research, we makes the following suggestions:

(1) Research on the interaction between commodity model and user. When the user touches the product, the product should also have a corresponding feedback to the user.
(2) Further improve the speed of 3D model reconstruction.
(3) Further optimize the space occupancy of high-quality 3D models. The algorithm to compress the model should be studied.
(4) Further improve the accuracy of model mapping and research on the reconstruction of commodity textures.
(5) Further improve the accuracy of the reconstructed model.
(6) The law of copyright protection for 3D virtual goods should be studied.
(7) Research on the security and tamper-proof of product 3D model.

References

[1] Newcombe, Richard A., et al. "Kinectfusion: Real-time dense surface mapping and tracking." *2011 10th IEEE international symposium on mixed and augmented reality*. Ieee, 2011.

[2] Dai, Angela, et al. "Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration." *ACM Transactions on Graphics (ToG)* 36.4 (2017): 1.

[3] Zollhöfer, Michael, et al. "State of the art on 3D reconstruction with RGB-D cameras." *Computer graphics forum*. Vol. 37. No. 2. 2018.

[4] Furukawa, Yasutaka, and Carlos Hernández. "Multi-view stereo: A tutorial." *Foundations and Trends® in Computer Graphics and Vision* 9.1-2 (2015): 1-148.

[5] Knapitsch, Arno, et al. "Tanks and temples: Benchmarking large-scale scene reconstruction." *ACM Transactions on Graphics (ToG)* 36.4 (2017): 1-13.

[6] Schonberger, Johannes L., and Jan-Michael Frahm. "Structure-from-motion revisited." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[7] Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *Communications of the ACM* 65.1 (2021): 99-106.

[8] Müller, Thomas, et al. "Instant neural graphics primitives with a multiresolution hash encoding." *ACM Transactions on Graphics (ToG)* 41.4 (2022): 1-15.

[9] An, Ran, and Jing Zhi Guo. "An empirical research on E-marketplace basic functions." *Applied Mechanics and Materials* 548 (2014): 1510-1523.

[10] Shen, Bingqing, et al. "How to promote user purchase in metaverse? A systematic literature review on consumer behavior research and virtual commerce application design." *Applied Sciences* 11.23 (2021): 11087.

[11] Luna-Nevarez, Cuauhtemoc, and Enda McGovern. "The rise of the virtual reality (VR) marketplace: exploring the antecedents and consequences of consumer attitudes toward V-commerce." *Journal of Internet Commerce* 20.2 (2021): 167-194.

[12] Nalbant, Kemal Gökhan, and Şevval UYANIK. "Computer vision in the metaverse." *Journal of Metaverse* 1.1 (2021): 9-12.

[13] Xi, Nannan, and Juho Hamari. "Shopping in virtual reality: A literature review and future agenda." *Journal of Business Research* 134 (2021): 37-58.

[14] Speicher, Marco. "Shopping in virtual reality." *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2018.

[15] Nasser, Nada, et al. "Social interaction in virtual shopping." *2021 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2021.

[16] Schnack, Alexander, Yinshu Zhao, and Nilufar Baghaei. "Introducing Shopper Avatars in a Virtual Reality Store." *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2023.