Enhancing Cultural Heritage Digitalization through 3D Graphics Algorithm and Immersive Visual Communication Technology

Fang Yuan*

Guangxi Normal University for Nationalities, College of Art, Chongzuo, China

* *Corresponding author: Fang Yuan (yuanfang16316@163.com)*

Submitted: 17 Jun 2025 **Accepted:** 06 Oct 2025 **Published:** 16 Feb 2025

Abstract With the continuous advancement of digital technology, cultural and creative product design is shifting from static presentation to dynamic immersive experience. The research aims to address the challenges faced by traditional modeling methods in accurately restoring complex textures and cross platform visual communication. The neural radiation field algorithm was enhanced by introducing a multi-level cost volume fusion module and a Gaussian uniform mixture sampling strategy. Furthermore, a collaborative visual communication framework integrating augmented reality and virtual reality was constructed, achieving a transition from single image input to high-precision 3D reconstruction, and then to dynamic interaction. The experiment showed that the improved algorithm achieved peak signal-to-noise ratios of 30.63 and 30.15 on the UoM-Culture3D and Bootstrap 3D synthetic datasets, respectively, with structural similarity indices of 0.88 and 0.89, respectively. Field deployment tests have shown that integrating AR and VR technologies into visual communication strategies significantly improves spatial perception consistency, prolongs user engagement time, and enhances detail recognition accuracy. This research emphasizes the potential of combining deeply coupled 3D graphics algorithms with immersive technology, which can help improve the digital restoration accuracy and cultural dissemination efficiency of cultural and creative products, thereby supporting the modern inheritance of traditional culture.

Keywords: 3D graphics algorithm, visual communication technology, cultural and creative product design, NeRF, VR, AR.

1. Introduction

With the adoption of digital technology in the industry, cultural and creative product design is facing a transition from static output to dynamic immersive experience. The popularization of Virtual Reality (VR) hardware and the advancement of real-time graphics computing have made the digital revitalization of cultural heritage a new direction. Through technological means, it is possible to break through the physical limitations of physical exhibitions, allowing historical patterns and traditional techniques to gain cross temporal and spatial dissemination power. This cultural and creative product design has put forward new requirements and urgently needs to break through the limitations of traditional two-dimensional expression, establish a multidimensional design system that integrates high-precision modeling, dynamic narrative, and interactive experience [14]. Currently, the Neural Radiation Field (NeRF) technology in the field of 3D graphics algorithms combines ray tracing and deep learning to achieve high fidelity digital reconstruction of complex cultural carriers such as cultural relics patterns and historical scenes [21]. However, this technology relies on dense input of hundreds of images in a single scene and time-consuming training on a scene by scene basis, making it difficult to adapt to the fast iterative design process of cultural and creative products [12, 31]. The augmented reality (AR) and VR technologies in the field of visual communication can create a virtual real fusion experience environment. However, most of the existing schemes use a single mode, which has problems such as large spatial alignment error, homogenization of interaction forms, and shallow semantic analysis of cultural symbols [27, 32]. To this end, a multidimensional design method for cultural and creative products based on the Improved NeRF (INeRF) algorithm and the integration of AR and VR is proposed. By integrating multi-level cost structures and utilizing cross-scale feature fusion techniques, geometric reasoning capabilities are strengthened. Furthermore, the implementation of a Gaussian uniform mixture sampling strategy optimizes the efficiency of surface detail reconstruction. Consequently, a seamless interactive experience across AR and VR platforms is attained within the visual communication layer. The research aims to enhance the cultural connotation expression and user experience of cultural and creative products, and promote the development of the cultural and creative industry towards digitalization and multidimensionality. The innovation of the research lies in introducing a multi-level geometric feature fusion mechanism and a mixed sampling strategy into the NeRF framework. Meanwhile, through AR-VR collaborative interactive design, the organic unity of cultural symbols in spatial, temporal, and perceptual dimensions is achieved, providing practical and expressive methodological support for the digital innovation of cultural and creative products.

2. Related works

High-precision 3D reconstruction is the cornerstone of cultural heritage digitization. NeRF technology has garnered significant attention for its ability to fuse ray tracing with deep learning, enabling high-fidelity reconstruction of the complex textures and structures of cultural relics. However, classical NeRF and its variants generally suffer from significant limitations: their training process heavily relies on hundreds of dense multi-view images from a single scene and time-consuming scene-by-scene optimization, which severely restricts their applicability in cultural and creative product design workflows requiring rapid iteration. To address reconstruction challenges in specific domains, researchers have proposed various optimization schemes. To achieve texture synthesis optimization, Houdard et al. [9] proposed a general framework named GOTEX. By constraining the local feature statistical distribution and utilizing the optimal transport semi-dual formula to control the feature distribution, high-quality texture synthesis and restoration were achieved. To improve the accuracy and efficiency of 3D reconstruction of ancient buildings, Ge et al. [7] introduced depth supervision into the NeRF framework, combining a truncated signed distance function and an incremental training strategy, effectively enhancing the accuracy and efficiency of 3D reconstruction of ancient buildings. In the field of dynamic scene reconstruction, Qiu et al. [19] innovatively combined NeRF with signed distance fields to achieve realistic reconstruction of dynamic ship models, demonstrating its potential for dynamic modeling of specific objects. Mazzacca et al. [15] further validated the effectiveness of NeRF in reconstructing cultural heritage datasets, particularly in handling uniform textures or shiny surfaces, expanding the documentation pathways for digital heritage.

Visual communication technology serves as a bridge connecting digital reconstruction outcomes with user experience. AR and VR technologies enable the creation of immersive cultural experience environments that blend virtual and real elements. To enhance the visual communication effectiveness of digital animated advertisements, Fang et al. [5] proposed a multimodal visual communication system model based on multimodal video emotion analysis. This model dynamically adjusts digital animated advertisement content according to user emotions, enhancing the personalization and appeal of interactions, and demonstrating the potential of emotion-driven content adaptation. Liu et al. [13] conducted an in-depth analysis of visual communication strategies for cultural imagery in rural environments, emphasizing the importance of environmental perception in experiencing cultural spirit through the integration of art intervention institutions, and providing insights for cultural narratives in specific spaces. In terms of communication effectiveness evaluation, the video data analysis system by Yachnaya et al. [26] can identify and assess paralinguistic and non-verbal components in communication, providing tools for quantifying user experience. Yudhanto et al. [30] advocate a visual communication design philosophy grounded in culture and communication, emphasizing the importance of researching the target audience's values, norms, language, beliefs, and visual elements to enhance the cultural relevance and effectiveness of design.

As can be seen from the above, although three-dimensional graphics algorithms and visual communication technologies have made significant progress in their respective fields, there remains a lack of cross-platform, multi-modal integrated design methods for the digitization of cultural heritage. Existing solutions often struggle to balance high-precision texture restoration, real-time interactive performance, and visual consistency across multiple devices. This research gap leads to issues such as experience discontinuity and information loss in the actual dissemination of cultural and creative products. To address this, the study proposes a multi-dimensional design framework for cultural and creative products based on an improved INeRF and the deep integration of AR and VR, providing a solution that combines precision and expressiveness for the innovative transformation and dissemination of cultural heritage.

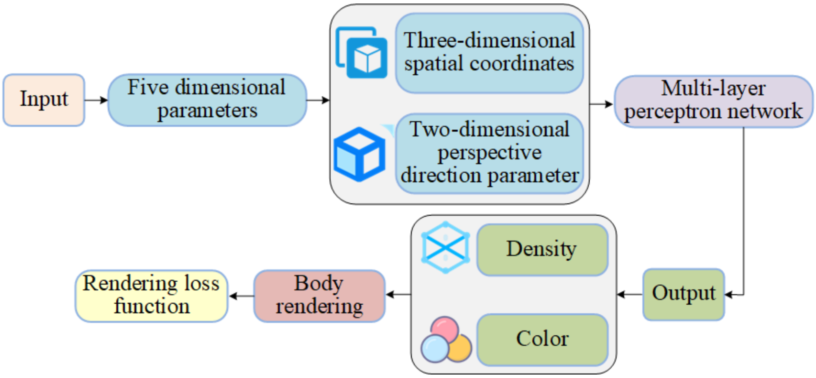
Input
Five dimensional parameters
Three-dimensional spatial coordinates
Two-dimensional perspective direction parameter
Multi-layer perceptron network
Output
Density
Color
Body rendering
Rendering loss function

Fig. 1. Schematic diagram of NeRF algorithm (icons designed by Freepik [6]).

3. Methods and materials

3.1. Design of 3D graphics algorithm based on INeRF

Three-dimensional graphics algorithms are driving the transformation of cultural and creative products toward multi-dimensional design. NeRF technology combines deep learning and ray tracing to achieve high-fidelity three-dimensional reconstruction of cultural relics and historical scenes, effectively restoring complex textures and material effects, and solving the challenges of traditional modeling in reproducing complex materials and intricate patterns [11, 16, 28]. The basic structure of NeRF technology is shown in Fig. 1. This technology first receives a five-dimensional input parameter consisting of spatial position coordinates and the angle of light incidence. This parameter is then mapped by a multi-layer perceptron network into RGB color values and density parameters. Subsequently, the system emits rays from the viewpoint, continuously sampling points along the path, and uses a volume rendering formula to calculate the transmittance and color contribution of each point, thereby synthesizing a realistic lighting effect. Finally, the model is optimized using a pixel-level rendering loss function to approximate the optical properties of the real-world scene. Among them, the NeRF mapping function [10] is

$$F(x, y, z, \theta, \phi) \to (R, G, B, \sigma) \,, \tag{1}$$

where x, y, and z represent three-dimensional spatial coordinates, θ and ϕ represent the angle parameters of the incident direction of light rays, R, G, and B represent the RGB color values of the sampling points, and σ represents the medium density of the sampling point. The function predicts the optical properties of each sampling point based on the light and scene geometry characteristics, thereby providing basic data for

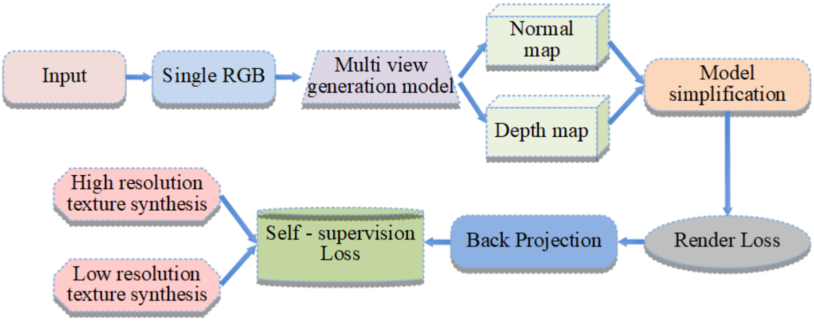

Fig. 2. The basic framework of INeRF.

volume rendering. The rendering expression is

$$C(r) = \int_{t_n}^{t_f} T(t) \cdot \sigma(r(t)) \cdot c(r(t), d)\, \mathrm{d}t\,, \tag{2}$$

where $C(r)$ represents the cumulative color of light, $T(t)$ indicates the transmittance of light from the starting point to the current point, $\sigma(r(t))$ represents the density of path point $r(t)$, $c(r(t), d)$ indicates the color of the path point $r(t)$ in the direction d, and t_n and t_f represent the starting and ending points of the light. This formula achieves optically realistic image synthesis by accumulating the color and transparency of each sampling point along the light path. The expression of the rendering loss function is

$$\mathcal{L}_{\text{render}} = \sum_{p} \left\| \hat{C}(p) - C_{\text{gt}}(p) \right\|^2 , \tag{3}$$

where $\mathcal{L}_{\text{render}}$ represents pixel-level rendering loss, $\hat{C}(p)$ represents the color of pixels in the generated image, and $C_{\text{gt}}(p)$ represents the color of pixel p in real multi-view images. Although NeRF technology can achieve high-precision 3D reconstruction, it relies on a large number of input images from a single scene and time-consuming scene by scene optimization training, which makes it difficult to meet the design requirements for rapid iteration of cultural and creative products. Therefore, the study proposed the INeRF algorithm, whose basic framework is shown in Figure 2.

The INeRF algorithm starts with a single RGB input and extends the model to multi view data through multi view generation. It combines camera parameters to drive the 3D reconstruction module to generate normal maps and depth maps. During the process, supervised and soft supervised loss optimization is used to optimize depth and RGB prediction, and geometric consistency is ensured through backprojection. The rendering loss function further optimizes the lighting and material performance of the

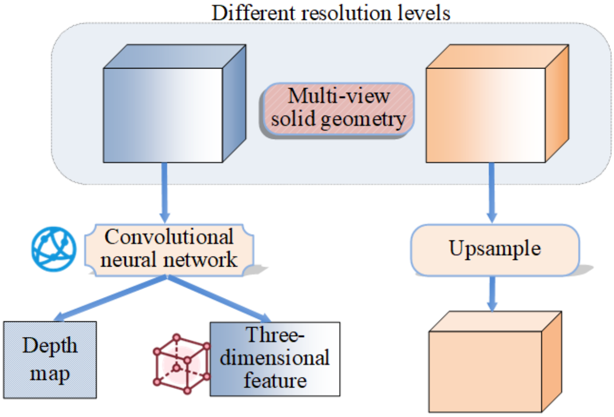
Different resolution levels
Multi-view solid geometry
Convolutional neural network
Upsample
Depth map
Three-dimensional feature

Fig. 3. The basic structure of the cost body (icons designed by Freepik [6]).

model, followed by high-resolution texture synthesis to enhance details, and ultimately balances accuracy and efficiency through model simplification techniques to output high-quality 3D models.

To address the issue of insufficient geometric information in single-view input, a multi-level cost volume fusion module based on convolutional attention was designed, as shown in Figure 3. Its the schematic diagram is based on multi-view solid geometry. Firstly, feature maps are extracted from input images of different resolution levels, and the three-dimensional geometric information of the scene is captured by constructing a multi-scale cost volume. In the feature fusion stage, low-resolution cost bodies encode global semantics, while high-resolution cost bodies retain details. Cross-layer feature interaction is achieved through convolutional attention, and channel and spatial attention are used to optimize the coordination of local and global information. Ultimately, a geometric neural field with both spatial accuracy and semantic integrity is formed, providing multi-level feature support for rendering. The formula for multi-level cost volume fusion is

$$F_{\text{fused}} = \sum_{l=1}^{L} w_l \cdot F_l^{\text{up}} + F_{\text{res}} , \tag{4}$$

where F_{fused} represents the fused multi-level features, F_l^{up} represents the features after upsampling at layer l, w_l represents the feature weight calculated through attention mechanisms, F_{res} represents the residual connection feature, and L indicates the total number of feature levels. In the feature decoding and rendering optimization stage, IN-eRF achieves efficient and accurate volume rendering by improving the sampling strategy and loss function design. In response to the problem of insufficient density in traditional uniform sampling, a Gaussian uniform mixture sampling strategy is proposed. Based on the depth prior information inferred from multi-view solid geometry, Gaussian distribution dense sampling is used in the surface area of the object, while maintaining uniform sampling density in non critical areas. The expression for Gaussian uniform mixture sampling distribution is [18]

$$P(s) = \lambda \cdot \mathcal{N}(s \mid \mu_d, \sigma_d) + (1 - \lambda) \cdot \mathcal{U}(s \mid s_{\min}, s_{\max}) \,, \tag{5}$$

where $P(s)$ represents the probability density function of the sampling point s, $\mathcal{N}$ is the Gaussian distribution, $\mathcal{U}$ represents the uniform distribution, λ represents mixed weight coefficients, μ_d represents the depth mean, and σ_d represents variance.

Meanwhile, a deep self-supervised loss function was designed to generate pseudo depth maps using multi-view consistency constraints. The pixel information of the source view was distorted to the target perspective through differentiable reprojection, and a self-supervised signal without the need for real depth annotation was constructed. Moreover, during the feature decoding stage, the algorithm spatially aligns the three-dimensional local features generated by the geometric neural field with the two-dimensional global features. It then incorporates the encoded information of light ray directions, dynamically decoding the color and density values for each sampling point via a multi-layer perceptron. Finally, it synthesizes the pixel color and depth information of the target viewpoint using a differentiable rendering equation, thereby establishing an end-to-end trainable framework. Through this framework, designers can quickly convert historical images, physical photos, or 2D drawings into interactive 3D models, greatly improving the responsiveness and flexibility of the creative production process.

The pseudocode of the INeRF algorithm is presented in the Appendix A.

3.2. Design of cultural and creative products based on visual communication technology

After completing high-precision digital reconstruction based on 3D graphics algorithms, visual communication technology has become the core supporting means in multi-dimensional expression of cultural and creative products. To achieve deep dissemination and innovative expression of cultural values, a deep integration strategy based on AR and VR has been studied and designed. The overall framework is shown in Figure 4. In the data generation layer, the system relies on the INeRF algorithm to construct a high-precision 3D model from a single image, obtaining multidimensional data including geometry, normal maps, and depth maps, laying the foundation for subsequent visual presentation. The visual expression layer focuses on the graphic rendering and semantic visualization processing of 3D models, mapping digital models into recognizable and culturally significant visual content through lighting simulation, material mapping, and color coding, and adapting to AR and VR platforms for dynamic presentation [29].

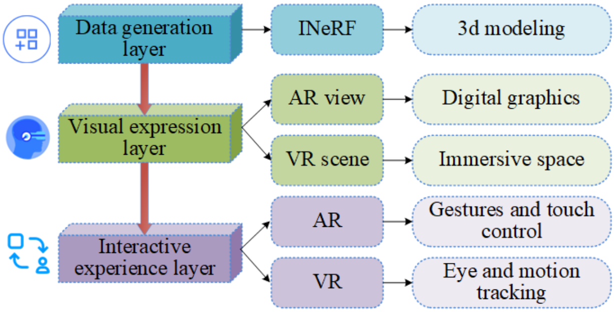

Fig. 4. The overall framework of visual communication strategy of cultural and creative products (icons designed by Freepik [6]).

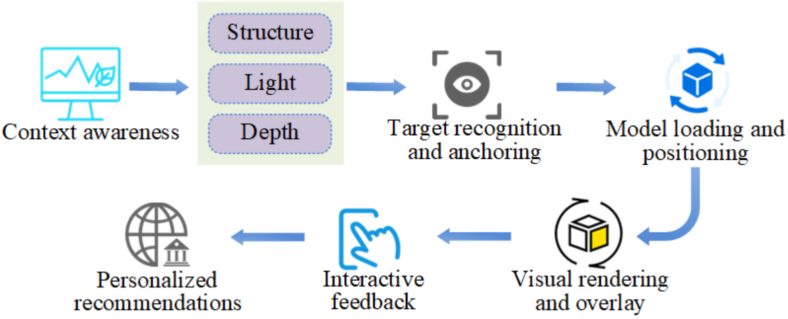
Context awareness
Structure
Light
Depth
Target recognition and anchoring
Model loading and positioning
Visual rendering and overlay
Interactive feedback
Personalized recommendations

Fig. 5. Visual communication flow chart of AR-based cultural and creative products (icons designed by Freepik [6]).

The interactive experience layer revolves around user perception, combining the real-time positioning and virtual real overlay capabilities of AR, as well as the immersive spatial construction characteristics of VR, to achieve dynamic calling and multi-modal interaction design of cultural and creative graphic content.

The basic process of visual communication for AR-based cultural and creative products is shown in Figure 5. The system uses the RGB-D sensor built into the AR device to collect data on the geometric structure, depth distribution, and lighting conditions of the user's surroundings. It then uses feature point matching algorithms to identify and anchor targets, accurately locating physical objects such as display cases, cultural and creative packaging, and interior walls, and setting attachment points for virtual elements [22]. During the graphic deployment phase, the 3D models generated by INeRF are compressed and optimized for lightweight performance, then loaded into the augmented reality platform. The system automatically adjusts the orientation based on the on-site

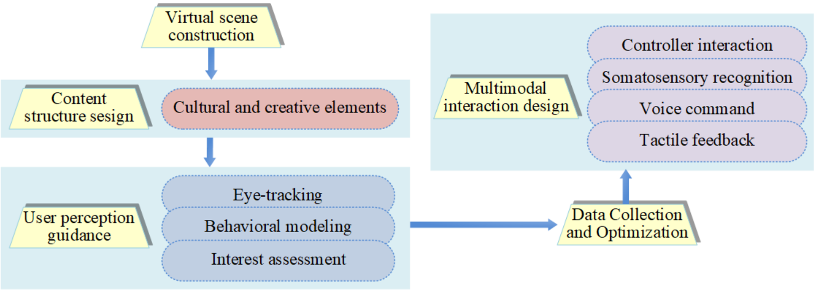

Fig. 6. Framework diagram of cultural and creative space construction based on VR (icons designed by Freepik [6]).

coordinate system. Subsequently, the system performs real-time graphic rendering and visual overlay, utilizing dynamic lighting estimation and reflection maps to ensure high consistency between virtual images and the real-world environment. Users can interact with virtual graphics through gesture recognition, voice input, or touch operations to obtain multi-dimensional feedback. The system finally combines user behavior trajectories and preference patterns to achieve personalized push notifications for cultural and creative content, further enhancing the targeting and engagement of visual communication [4]. To achieve seamless integration between virtual and real environments, the study developed an AR-VR hybrid interaction framework. When users transition from an AR scene to a VR scene, the system retains their operational state and interaction history through a spatial state caching mechanism, enabling state restoration and content continuity within the virtual space. First, the system uses the RGB-D sensor and IMU data from AR devices for real-time environmental mapping and user localization. Second, a virtual scene mapping model is established on the VR platform to ensure that the scene geometry aligns with the real-world spatial coordinates [3]. Finally, a state caching and synchronization mechanism is designed to save user interaction operations and object states, enabling seamless cross-device switching. In terms of on-site deployment, the system considers lighting matching, dynamic occlusion handling, and device load optimization to ensure stable operation in exhibition or cultural and creative experience spaces.

The AR-VR hybrid interaction framework is shown in Figure 6, which is the framework for constructing cultural and creative spaces based on VR. It systematically outlines the methodological path of VR technology in multidimensional cultural and creative design. Firstly, the designer relies on a 3D model database and INeRF generated results to construct a virtual environment that covers historical block restoration, cultural festival scenes, and immersive exhibition spaces for cultural relics, forming a virtual field with cultural depth. At the level of content structure, cultural and creative elements are orderly embedded into spatial nodes, forming multiple types of information units, including decorative shapes, interactive objects, semantic labels, and dynamic animations, thus establishing a rich cultural narrative space. The system integrates gaze tracking and behavior modeling modules to dynamically adjust the visual hierarchy and dynamic parameters of virtual content based on users' attention paths and interest preferences, guiding users to naturally integrate into the narrative process. In terms of interaction, the platform integrates controller control, speech recognition, motion capture, and tactile feedback technology to provide users with multi-channel immersive interaction methods, enhancing the degree of freedom and realism of the experience. Meanwhile, the system continuously collects user behavior data in the virtual space in the background, including field of view movement, dwell time, and interaction frequency, providing data support and model basis for subsequent scene structure adjustment and visual information optimization, thereby achieving iterative updates and precise push of the design system.

4. Results

4.1. Performance verification of 3D graphics algorithm based on INeRF

To verify the effectiveness of multi-dimensional design of cultural and creative products based on 3D graphics algorithms and visual communication technology, a 3D reconstruction and visual communication system for cultural and creative products based on INeRF algorithm and AR/VR fusion was constructed in an experimental environment with GPU acceleration capability.

The image datasets used in the experiment included the UoM-Culture3D dataset [25] and the Bootstrap3D synthetic dataset [23, 24]. The UoM-Culture3D dataset contains multi-perspective images of historical artifacts and cultural scenes, with a resolution of 1920 × 1080, suitable for high-quality 3D reconstruction. The Bootstrap3D synthetic dataset contains millions of multi-view images covering creative objects such as fictional creatures and cultural symbols.

The specific experimental environment and parameter configuration are shown in Table 1. Based on this experimental environment, the study compared the introduction of raw NeRF [16,28], NeRF based on multi-resolution texture pyramid (Mip-based, Mip-NeRF) [2], Instant Neural Graphics Primitives with a multi-resolution hash encoding (Instant-NGP) [17], and INeRF model proposed in this paper.

Firstly, using Peak Signal to Noise Ratio (PSNR) as a comparison metric, tests were conducted on different datasets, and the results are shown in Figure 7, where the PSNR comparison performance of four 3D reconstruction models on two datasets are displayed. In Figure 7a, on the UoMCult3D dataset, NeRF had the weakest performance with a PSNR of 25.82 at the 500th iteration. Mip-NeRF and Instant-NGP reached 28.19 and

Tab. 1. Experimental environment and parameter configuration.

Type	Name	Version
Hardware equipment	CPU	Intel Xeon Gold 6248R, 3.0 GHz, 24C
	GPU	NVIDIA RTX 3090, 24 GB RAM
	RAM	128 GB DDR4
	Memory device	2 TB NVMe SSD
Software equipment	Operating system	Ubuntu 20.04 LTS
	DL framework	PyTorch 1.13
	Graphics rendering	Unity 2022.3 (HDRP line pipe)
	AR develop	ARCore 1.35, ARKit 5.0
	VR develop	SteamVR 2.0, OpenXR 1.0
Parameter name	Learning rate	0.001
	Batch size	1024
	Render resolution	800 × 800 pixels
	Real-time render target frame rate	$\geq$ 30 FPS

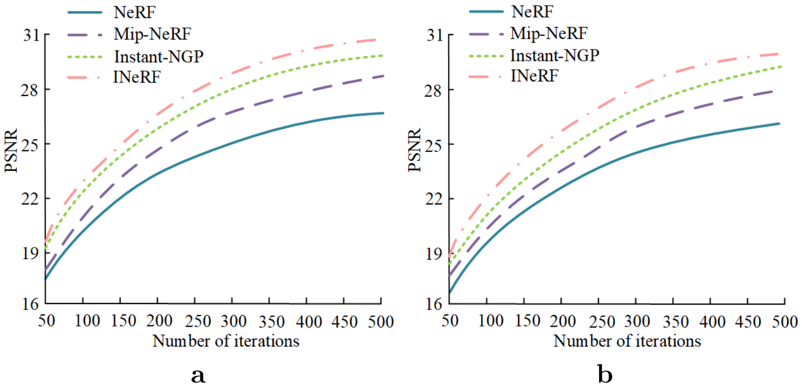

Fig. 7. PSNR comparison of four models with different data sets: (**a**) UoM-Culture3D, (**b**) Bootstrap3D.

30.11, respectively, while INeRF performed the best, stabilizing at 30.63, with an average improvement of 9.24% compared to the other three models. In Figure 7b, INeRF still had a significant advantage on the Bootstrap3D dataset, with a PSNR of 30.15 at the 500th iteration, an average increase of 8.17% compared to other models. This indicated that INeRF had good universality and reconstruction stability in stylized data and cultural images. On this basis, the graphic loading speed and Root Mean Square Error (RMSE) of four models on the AR platform were tested, and the results are shown in Figure 8. According to Figure 8a, as the number of experiments increased, the loading speed of

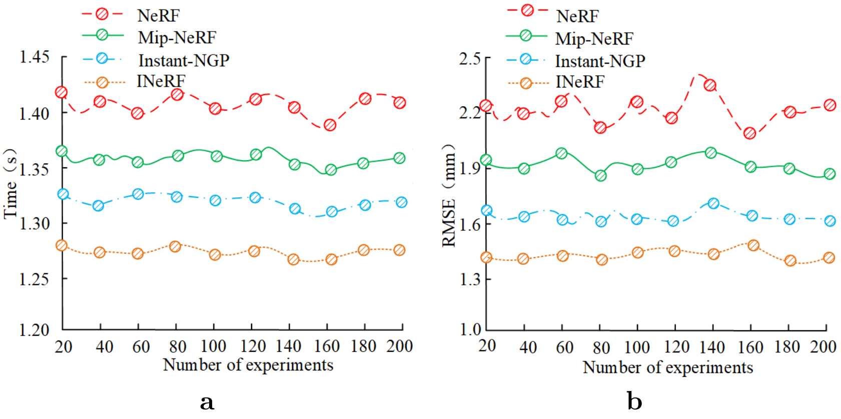

Fig. 8. Comparison of parameters of four models: (**a**) graphic loading time, (**b**) RMSE.

the INeRF model remained at a relatively low level of about 1.26 s, demonstrating high stability and efficiency. In contrast, the NeRF model had the longest loading time, close to 1.41 seconds, and it fluctuated greatly. This might have been due to its reliance on a large number of input images and scene-by-scene optimization training, which resulted in high computational complexity and a slow speed during the loading process. Based on Figure 8b, INeRF had the lowest RMSE value among 200 experiments, stabilizing at around 1.42 mm, with an average reduction of 25.28% compared to other models. Overall, the balance between speed and accuracy of INeRF validated the effectiveness of its improved architecture, providing a reliable technical path for high-fidelity digitization of cultural heritage.

Meanwhile, the Structural Similarity Index Measure (SSIM) of four models on different datasets were compared, and the results are shown in Figure 9. Figure 9a shows the SSIM comparison of four models on the UoM-Culture3D dataset. As the number of iterations increased, the SSIM value of INeRF gradually rose and tended to stabilize. When the number of iterations reached 500, the SSIM value of INeRF remained stable at around 0.88, significantly better than the other three models. Figure 9b presents the SSIM comparison of four models for the Bootstrap3D dataset. NeRF performed better than other models on the Bootstrap3D dataset. When the number of iterations reached 500, the SSIM value of INeRF reached 0.89. This indicates that INeRF can effectively integrate geometric features of different scales, enhancing the model's perception and reconstruction ability of complex image structures.

To directly validate the accuracy of the INeRF algorithm in 3D structure reconstruction, a quantitative evaluation based on point cloud comparison was conducted on the

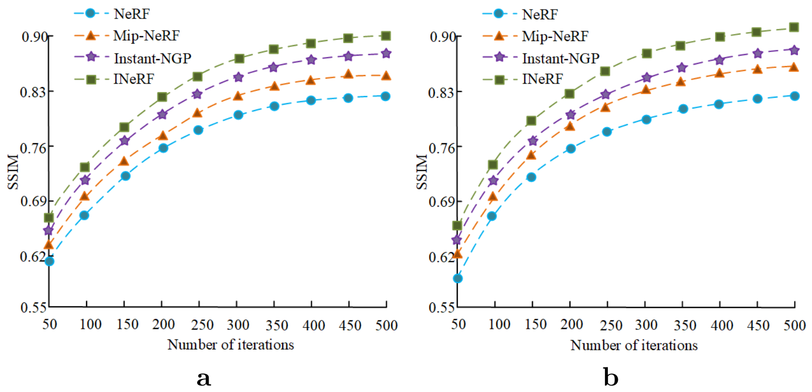

Fig. 9. Comparison of SSIM of four models for different data sets: (**a**) UoM-Culture3D, (**b**) Bootstrap3D.

Tab. 2. Comparison of 3D geometric reconstruction effects of different 3D reconstruction techniques. Asterisks '*' and '**' indicate statistically significant differences compared to INeRF at $p < 0.05$ and $p < 0.01$, respectively.

Model	NeRF	Mip-NeRF	Instant-NGP	INeRF	3DGS	DiffRF
Chamfer dist. [mm]	2.12*	1.88*	1.55*	1.42	1.60	1.50
Hausdorff dist. [mm]	6.48**	5.92**	5.11*	4.78	5.05	4.92
F1-score @0.05	0.42**	0.51**	0.61*	0.65	0.62	0.63
F1-score @0.1	0.59**	0.65**	0.74*	0.78	0.75	0.76
F1-score @0.2	0.71**	0.76**	0.83*	0.86	0.84	0.85
Normal Consistency	0.78**	0.81**	0.85*	0.88	0.86	0.87
Training time [h]	12.42	9.71	4.15	5.31	6.27	6.82
Peak vRAM [GB]	18.60	16.24	9.83	11.41	12.58	13.16

UoM-Culture3D dataset. Two emerging 3D reconstruction techniques were also introduced for comparison: the 3D Gaussian Splatting (3DGS) model and the Rendering-Guided 3D Radiance Field Diffusion Model (DiffRF). Marching Cubes algorithm was used to extract meshes from the density fields predicted by each model, and 50 000 vertices were uniformly sampled to generate point clouds for evaluation. The results are shown in Table 2. It can be seen that the NeRF model performs the worst in various indicators, reflecting its insufficient ability to reconstruct complex textures and details in sparse views, as well as high resource requirements. Mip NeRF improved feature expression through multi-resolution texture pyramids, reducing Chamfer Distance to

1.88 mm and Hausdorff Distance to 5.92 mm. However, there were still significant differences ($p < 0.05$) between the improvements and INeRF. Instant NGP further optimized the point cloud distribution under dense feature encoding, with a Chamfer Distance of 1.55 mm and a normal consistency of 0.85. Although the overall accuracy is close, the difference with INeRF is still significant ($p < 0.05$). In contrast, INeRF achieved the best performance on all indicators, with the lowest Chamfer Distance being 1.42 mm, the Hausdorff Distance dropping to 4.78 mm, F1-scores reaching 0.78 and 0.86 at the 0.1 and 0.2 thresholds, respectively, and a normal consistency of 0.88. The INeRF maintains high accuracy while controlling the training time to 5.31 h, with a video memory usage of only 11.41 GB. Although slightly higher than Instant NGP, it still demonstrates good deployability in resource constrained environments, reflecting the balance advantage between accuracy and efficiency. The difference from most methods is significant or highly significant, thanks to the collaborative optimization of multi-level cost volume fusion and Gaussian uniform mixture sampling strategy in details and global structure. For emerging technologies, the 3D Gaussian jet model and rendering guided radiation field diffusion model approach Instant NGP on Chamfer Distance and F1-score, with no significant difference compared to INeRF, but slightly lower in performance, indicating that there are still subtle geometric errors in sparse input and complex texture scenes.

Based on various indicators and statistical analysis, INeRF exhibits excellent performance in point cloud accuracy, surface normal consistency, and F1-score at different scales. It also shows strong advantages in computational resource utilization, verifying its robustness and reliability in high-precision 3D reconstruction. At the same time, it demonstrates strong adaptability to complex textures and geometric structures in the process of cultural heritage digitization.

4.2. Visual communication effect verification

To validate the effectiveness of the proposed visual communication strategy integrating AR and VR in actual deployment, the study conducted on-site deployment tests in museum exhibition spaces. The deployment included: AR end: Using ARCore/ARKit devices to scan the exhibition area, accurately anchor the location of exhibits, and overlay virtual information. VR end: Using SteamVR devices to construct virtual exhibitions of historical scenes, allowing users to freely interact in the virtual space. The actual measurement data covers indicators such as spatial perception consistency, interaction fluidity, immersion, and cultural understanding perception (out of 10 points) for 30 test subjects. The study compared the traditional 2D display, single VR, and single AR technologies with the proposed fusion strategy, and tested the spatial perception consistency and average dwell time of the four technologies in four cultural and creative scenes: porcelain, murals, ancient architecture, and bronze ware. The results are shown in Figure 10.

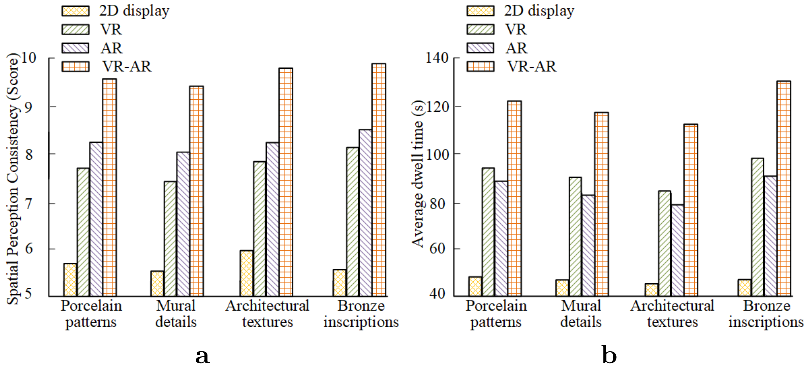

Fig. 10. Comparison of visual communication effects in different cultural and creative scenes: (**a**) spatial consistency ratings; (**b**) average dwell time.

According to Fig. 10a, the scores for integrating VR–AR technology in the four cultural and creative scenes of porcelain, mural, ancient architecture, and bronze ware were 9.50 points, 9.40 points, 9.60 points, and 9.83 points, respectively. When compared with the other three single visual communication technologies, the average scores had increased by 32.72%, 38.35%, 32.27%, and 34.25%, respectively. This indicated that the integration of VR–AR technology achieved better real-world mapping and spatial positioning of three-dimensional structures under the fusion of virtual and real environments. Meanwhile, based on Fig. 10b, the average dwell time of the fusion strategy in the four cultural and creative scenes of porcelain, mural, ancient architecture, and bronze ware was 121.24 s, 118.16 s, 115.67 s, and 130.13 s, respectively. This was an average increase of 59.67%, 58.32%, 57.19%, and 62.25% compared to the other three technologies. By constructing an immersive virtual space and implementing a personalized interactive content push mechanism, users were able to form a deeper sense of participation and cultural context immersion during the experience, which in turn extended their stay time.

Finally, the interactive experience and cultural perception effects of visual communication technology integrating AR and VR were studied, and the results are shown in Table 3. Visual communication technology that integrates VR and AR significantly outperforms single 2D display, VR, or AR solutions in terms of scene detail recognition, interaction fluidity, visual immersion, cultural compatibility, and memory retention. Most of these differences are highly significant ($p < 0.01$), confirming its advantages. Specifically, the scene detail recognition accuracy of the proposed fusion VR and AR visual communication technology reached 92.36%, an average improvement of 23.37% compared to the other three methods, indicating that it had higher accuracy in visual clarity and spatial recognition. In terms of interaction fluency and visual immersion, Tab. 3. Comparison of interactive experience and cultural perception effect of different visual communication technologies. Asterisks '*' and '**' indicate statistically significant differences compared to INeRF at $p < 0.05$ and $p < 0.01$, respectively.

Index	Scene detail recognition accuracy [%]	Interaction fluency [points]	Visual immersion [part]	Cultural fit [%]	Memory retention [%]
2D display	64.37**	5.18**	4.92**	61.25**	58.63**
VR	78.45**	7.86*	7.32**	76.12**	71.40**
AR	81.78*	7.12**	8.47*	80.56*	74.93*
VR–AR	92.36	9.14	9.68	90.42	86.71

the fusion strategy achieved scores of 9.14 and 9.68, respectively, with an average improvement of 36.07% and 39.94% compared to the other three methods. This indicated that it had advantages in operational response and system feedback, while also providing a more immersive cultural experience. In addition, the cultural fit and memory retention of fusion technology were 90.42% and 86.71%, respectively, with an average improvement of 24.47% and 26.92%, indicating that it was more accurate in conveying cultural connotations and symbol fit, and had a stronger effect on retaining cultural information. Overall, the integration of VR and AR technology had significant advantages in enhancing user immersion, improving cultural understanding and memory retention, which validated the scientific and practical nature of the visual communication strategies proposed in the study.

5. Discussion

The research is dedicated to addressing the challenge of synergistically optimizing high-precision reconstruction and cross-platform immersive communication in the digitization of cultural heritage. While 3D reconstruction technology has made progress in multiple fields, it still faces limitations in scene adaptability: a new real-time 3D reconstruction framework significantly enhances maritime situational awareness by integrating temporal 2D video data. Its optimized dynamic reconstruction pipeline enables real-time computation on GPU-accelerated embedded devices. However, it lacks the ability to predict the pose of semi-static objects, making it difficult to capture the geometric continuity of cultural relics under micro-movement conditions [20]. Visual tracking technology based on real-time localization and mapping serves as the core support for augmented reality localization. While it can real-time obtain user pose information, it faces inherent limitations in static scenes due to global localization drift and translation dependency, leading to insufficient spatial anchoring stability in cultural heritage sites [1]. In the field of medical imaging, three-dimensional reconstruction methods for brain tumors based on magnetic resonance imaging demonstrate efficient and precise visualization capabilities. However, when faced with the multi-layered composite texture structure of cultural relics, their topological adaptability remains weak [8]. The aforementioned technologies are either constrained by the integrity of dynamic modeling, limited by the robustness of static localization, or lack the generalization capability for heterogeneous structures, and thus fail to bridge the dual demands of millimeter-level precision reconstruction and multi-modal immersive narrative in cultural heritage digitization.

Therefore, this study aims to establish an integrated system that combines high-precision digital reconstruction with immersive cultural communication, proposing the INeRF algorithm and a multi-dimensional design method that integrates AR and VR technologies. By introducing a multi-level cost-volume fusion module, it achieves collaborative optimization of geometric features across scales, and adopts a Gaussian-uniform hybrid sampling strategy to enhance computational efficiency. Additionally, it combines AR and VR technologies to construct a three-tier communication system encompassing data generation, visual expression, and interactive experience. At the technical implementation level, the system uses multi-sensor fusion to achieve real-time positioning and environmental perception. It also uses dynamic lighting matching, object posture adjustment, and content stream optimization to ensure the accurate presentation of virtual objects on different platforms and in different exhibition environments. Finally, the study validated the feasibility of the integrated AR–VR strategy through field deployment. Field tests demonstrated that the system could achieve stable virtual overlay and multimodal interaction in real exhibition spaces, and user feedback showed significant improvements in cultural information understanding and immersive experiences.

It should be noted that there are still certain limitations in the experimental and validation of the research. Firstly, the test object mainly focuses on the 3D reconstruction of static scenes. However, with the continuous expansion of digital demand for cultural heritage, dynamic cultural heritage such as dance, ceremony, and performance have gradually become research hotspots. For scenes with temporal variability, relying solely on static modeling cannot fully capture their temporal features and dynamic details. Secondly, there are certain limitations to the user research conducted. The current experiment only involves 30 participants, with a relatively limited sample size and a relatively small group composition, which may affect the universality of the research conclusions to some extent and not fully reflect the real experiences of users with different backgrounds.

Future research will further expand the applicability of the INeRF framework in dynamic modeling, such as by introducing temporal consistency constraints and combining optical flow or skeleton driven motion modeling methods to achieve high fidelity reconstruction and presentation of dynamic cultural heritage. At the same time, it is necessary to expand the sample size in user research, increase the dual participation of experts in cultural heritage protection and ordinary visitors, in order to obtain a more comprehensive evaluation. With further validation of the system in multi-user collaboration and dynamic exhibition scenarios, its universality and sustainability in digital protection and cross platform dissemination of cultural heritage are expected to be greatly improved.

6. Conclusion

Compared with existing methods, the INeRF based method improves reconstruction accuracy by 9%, reduces RMSE to 1.42 mm, and enhances visual immersion by nearly 40%. AR–VR integration significantly enhances cultural detail recognition and user engagement. Although research still has limitations in terms of static scene adaptability and small user sample size, future work will explore lightweight network architectures and broader user testing to achieve more universal applications and higher dynamic scene adaptability.

Funding

This paper was supported by the 2024 Guangxi University Research Foundation Capacity Enhancement Project for Middle-aged and Young Teachers: *Research on the Collection, Organization and Digital Inheritance of Endangered Materials on the Feitao Art of Maonan Ethnic Group in Guangxi* (No. 2024KY0765).

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability

The data supporting the findings of this study are referenced in the literature.

References

[1] L. Baker, J. Ventura, T. Langlotz, S. Gul, S. Mills, et al. Localization and tracking of stationary users for augmented reality. *The Visual Computer* 40(1):227–244, 2024. doi:10.1007/s00371-023-02777-2.

[2] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, et al. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5835–5844, 2021. doi:10.1109/ICCV48922.2021.00580.

[3] J. Bast. Managing the image. The visual communication strategy of European right-wing populist politicians on Instagram. *Journal of Political Marketing* 23(1):1–25, 2024. doi:10.1080/15377857.2021.1892901.

[4] J.-J. Cao, S.-M. Fang, and H. Contreras. Multimodal fusion visual communication method based on genetic algorithm. *Journal of Network Intelligence* 10(2):1071–1083, 2025. `https://bit.kuas.edu.tw/~jni/2025/vol10/s2/34.JNI-S-2024-05-019.pdf`.

[5] J. Fang and X. Gong. Application of visual communication in digital animation advertising design using convolutional neural networks and big data. *Peerj Computer Science* 9:e1383, 2023. doi:10.7717/peerj-cs.1383.

[6] FREEPIK. Find icons that go together. Fast. `https://www.freepik.com/icons`.

[7] Y. Ge, B. Guo, P. Zha, S. Jiang, Z. Jiang, et al. 3D reconstruction of ancient buildings using UAV images and neural radiation field with depth supervision. *Remote Sensing* 16(3):473, 2024. doi:10.3390/rs16030473.

[8] M. A. Guerroudji, K. Amara, M. Lichouri, N. Zenati, and M. Masmoudi. A 3D visualization-based augmented reality application for brain tumor segmentation. *Computer Animation and Virtual Worlds* 35(1):e2223, JAN 2024. doi:10.1002/cav.2223.

[9] A. Houdard, A. Leclaire, N. Papadakis, and J. Rabin. A generative model for texture synthesis based on optimal transport between feature distributions. *Journal of Mathematical Imaging and Vision* 65(1):4–28, 2023. doi:10.1007/s10851-022-01108-9.

[10] Z. Jia, B. Wang, and C. Chen. Drone-nerf: Efficient nerf based 3D scene reconstruction for large-scale drone survey. *Image and Vision Computing* 143:104920, 2024. doi:10.1016/j.imavis.2024.104920.

[11] X. Liao, X. Wei, M. Zhou, and S. Kwong. Full-reference image quality assessment: Addressing content misalignment issue by comparing order statistics of deep features. *IEEE Transactions on Broadcasting* 70(1):305–315, 2023. doi:10.1109/TBC.2023.3294835.

[12] J. Lin, G. Sharma, and T. N. Pappas. Toward universal texture synthesis by combining texton broadcasting with noise injection in StyleGAN-2. *e-Prime – Advances in Electrical Engineering, Electronics and Energy* 3:100092, 2023. doi:10.1016/j.prime.2022.100092.

[13] F. Liu, B. Lin, and K. Meng. Design and realization of rural environment art construction of cultural image and visual communication. *International Journal of Environmental Research and Public Health* 20(5):4001, 2023. doi:10.3390/ijerph20054001.

[14] W. Liu, Y. Zang, Z. Xiong, X. Bian, C. Wen, et al. 3D building model generation from MLS point cloud and 3D mesh using multi-source data fusion. *International Journal of Applied Earth Observation and Geoinformation* 116:103171, 2023. doi:10.1016/j.jag.2022.103171.

[15] G. Mazzacca, A. Karami, S. Rigon, E. Farella, P. Trybala, et al. Nerf for heritage 3D reconstruction. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 48(M-2-2023):1051–1058, 2023. doi:10.5194/isprs-archives-XLVIII-M-2-2023-1051-2023.

[16] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, et al. NeRF: representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* 65(1):99–106, 2021. doi:10.1145/3503250.

[17] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics* 41(4):102, 2022. doi:10.1145/3528223.3530127.

[18] M. Pepe, V. S. Alfio, and D. Costantino. Assessment of 3D model for photogrammetric purposes using AI tools based on NeRF algorithm. *Heritage* 6(8):5719–5731, 2023. doi:10.3390/heritage6080301.

[19] S. Qiu, S. Wang, X. Chen, F. Qian, and Y. Xiao. Ship shape reconstruction for three-dimensional situational awareness of smart ships based on neural radiation field. *Engineering Applications of Artificial Intelligence* 136:108858, 2024. doi:10.1016/j.engappai.2024.108858.

[20] F. Sattler, B. Carrillo-Perez, S. Barnes, K. Stebner, M. Stephan, et al. Embedded 3D reconstruction of dynamic objects in real time for maritime situational awareness pictures. *The Visual Computer* 40(2):571–584, 2024. doi:10.1007/s00371-023-02802-4.

[21] S. Shen, S. Xing, X. Sang, B. Yan, and Y. Chen. Virtual stereo content rendering technology review for light-field display. *Displays* 76:102320, 2023. doi:10.1016/j.displa.2022.102320.

[22] X. Shi and R. Villegas. AI technology in the virtual reality environment of graphic design of dynamic art visual communication frame. *Journal of Computational Methods in Sciences and Engineering* 25(3):2603–2616, 2025. doi:10.1177/14727978251321333.

[23] Z. Sun. BS-Objaverse. Hugging Face. `https://huggingface.co/datasets/Zery/BS-Objaverse/`.

[24] Z. Sun, T. Wu, P. Zhang, Y. Zang, X. Dong, et al. Bootstrap3D: Improving multi-view diffusion model with synthetic data. arXiv, arXiv:2406.00093v2, 2024. doi:10.48550/arXiv.2406.00093.

[25] Xinyi_Zheng. CULTURE3D: Cultural Landmarks and Terrain Dataset for 3D Applications. GitHub. `https://github.com/X-Intelligence-Labs/CULTURE3D`.

[26] V. O. Yachnaya, V. R. Lutsiv, and R. O. Malashin. Modern automatic recognition technologies for visual communication tools. *Computer Optics* 47(2):287–305, 2023. doi:10.18287/2412-6179-CO-1154.

[27] C. Yan, B. Gong, Y. Wei, and Y. Gao. Deep multi-view enhancement hashing for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43(4):1445–1451, 2020. doi:10.1109/TPAMI.2020.2975798.

[28] J.-W. Yang, J.-M. Sun, Y.-L. Yang, J. Yang, Y. Shan, et al. DMiT: Deformable Mipmapped Triplane representation for dynamic scenes. In: *Computer Vision – ECCV 2024*, pp. 436–453. Springer Nature Switzerland, Cham, 2025. doi:10.1007/978-3-031-73001-6_25.

[29] J. You and X. Lu. Visual communication design based on machine vision and digital media communication technology. *KSII Transactions on Internet & Information Systems* 19(6):1888–1907, 2025. doi:10.3837/tiis.2025.06.007.

[30] S. H. Yudhanto, F. Risdianto, and A. T. Artanto. Cultural and communication approaches in the design of visual communication design works. *Journal of Linguistics, Culture and Communication* 1(1):79–90, 2023. doi:10.61320/jolcc.v1i1.79-90.

[31] Z. Zhang, L. Li, G. Cong, H. Yin, Y. Gao, et al. From speaker to dubber: movie dubbing with prosody and duration consistency learning. In: *Proceedings of the 32nd ACM international conference on multimedia*, pp. 7523–7532, 2024. doi:10.1145/3664647.3680777.

[32] M. Zhao. Application of image reconstruction algorithm combining FCN and Pix2Pix in visual communication design. *Journal of Computational Methods in Sciences and Engineering* 25(4):3137–3151, 2025. doi:10.1177/14727978251319398.

A. Appendix

Pseudocode of INeRF algorithm

```
INeRF algorithm
Input:
  $I_{\text{src}}$
  $K$: Camera intrinsic matrix
Output:
  $M$: High-precision 3D mesh model (geometry + texture)
// Step 1: Multi-view generation (replaces multi-image input)
$I_{\text{views}} \leftarrow \text{MultiViewGenerator}(I_{\text{src}})$ //Generate $N$ virtual views $\{I_1, I_2, ..., I_N\}$
$\Theta_{\text{cam}} \leftarrow \text{EstimateCameraPoses}(I_{\text{views}}, K)$ //Estimate virtual view poses
//Step 2: Geometric reasoning (multi-level cost volume fusion)
$F_{\text{multi}} = [\,]$ //Initialize multi-scale feature list
for each $I_i$ in $I_{\text{views}}$
  for each scale $s$ in $[1, 2, 4]$//Multi-resolution feature extraction
    $F_s \leftarrow \text{CNN\_Encoder}(I_i, \text{scale} = s)$ //Extract features at scale $s$
    $F_{\text{multi}}[s] \leftarrow F_s$
  end for
  $C_i \leftarrow \text{BuildCostVolume}(F_{\text{multi}}, \Theta_{\text{cam}}[i])$ //Construct cost volume for view $i$
end for
$F_{\text{fused}} \leftarrow \text{MultiLevelFusion}(C_{\text{all}})$ //Fuse cost volumes (Eq. (4), Fig. 3)
//Step 3: Neural radiance field modeling
for each pixel $p$ in target view:
  ray $r \leftarrow \text{GenerateRay}(p, \Theta_{\text{cam\_target}})$
  //Gaussian-uniform hybrid sampling (Eq. (5))
  samples $\leftarrow \text{GaussUniformHybridSampling}(r, \text{depth\_prior=DepthMap}(F_{\text{fused}}),$
    $\mu = \text{depth\_mean}, \sigma = 0.2, \alpha = 0.7)$ //$\alpha$: Gaussian sampling weight
  $\sigma, c \leftarrow [\,]$ //Store density and color
  for each sample point $x$ in samples
    $\text{feat}_{3d} \leftarrow \text{Query3DFeature}(x, F_{\text{fused}})$ //Query 3D local feature
    $\text{feat}_{\text{dir}} \leftarrow \text{Encode}(\text{view\_dir})$ //View direction encoding
    $(\sigma_x, c_x) \leftarrow \text{MLP}_{\sigma c}(\text{feat}_{3d}, \text{feat}_{\text{dir}})$ //Predict density and color
    $\sigma.\text{append}(\sigma_x)$; $c.\text{append}(c_x)$
  end for
  //Volume rendering (Eq. (2))
  $\hat{C}_p \leftarrow \text{VolumeRendering}(\sigma, c, \text{samples})$
  $\hat{D}_p \leftarrow \text{DepthMapRendering}(\sigma, \text{samples})$ //Predict depth map
end for
//Step 4: Self-supervised optimization
$L_{\text{rgb}} \leftarrow \text{MSE}(\hat{C}, I_{\text{gt}})$ //RGB rendering loss (Eq. (3))
$L_{\text{depth}} \leftarrow \text{DepthConsistencyLoss}(\hat{D}, \text{FusedDepth})$ //Depth self-supervised loss
$L_{\text{total}} \leftarrow \lambda_1 L_{\text{rgb}} + \lambda_2 L_{\text{depth}}$ //$\lambda_1 = 1.0$, $\lambda_2 = 0.5$ (tunable)
Update $\text{MLP}_{\sigma c}$ via $\nabla L_{\text{total}}$ //Backpropagation update
//Step 5: High-res texture generation & model simplification
$M_{\text{highres}} \leftarrow \text{TextureSynthesis}(F_{\text{fused}}, \text{MLP}_{\sigma c})$ //Generate textured dense mesh
$M \leftarrow \text{MeshSimplification}(M_{\text{highres}}, \text{target\_faces=50k})$ //Simplify model
return $M$
```