ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Research on Immersive Digital Media Art Scene Generation and Interaction Design Driven by Neural Rendering Technology

Di Luo[1,*]

[1] College of Information Engineering, Chongqing Vocational and Technical University of Mechatronics, Chongqing 404100, Chongqing, China

SUMMARY: *With the extension of immersive digital media art to 3D real-time interactive space, scene generation requires higher precision reconstruction, low-latency rendering and behavior-driven feedback. Multi-source visual data, depth map, camera pose, and semantic annotation are used to complete unified coding. Neural radiance field is combined to achieve continuous 3D reconstruction. At the same time, the multi-modal fusion of head display pose, gaze, gesture and control command is introduced, and the interaction intention recognition and closed-loop update of scene state are realized through attention weight allocation, intention recognition and state scheduling. Experiments show that the average reconstruction error of four types of art scenes is 1.61 cm, the PSNR of digital exhibition hall reaches 32.4 dB, the SSIM is 0.936, the frame rate of 1080p high-complexity scenes is 59.7 FPS, and the overall recognition accuracy of interactive tasks reaches 94.5%. This study provides technical support for intelligent scene construction and real-time interaction design of immersive digital media art.*

KEYWORDS: *Neural rendering; Immersive digital media art; Scene generation; Interaction design*

1 Introduction

As digital media art gradually shifts from two-dimensional visual presentation to three-dimensional immersive space, the generation method, rendering efficiency and interactive experience of art scenes are being jointly influenced by computer graphics, deep learning and human-computer interaction technology [1]. Traditional digital scene production mostly relies on manual modeling, mapping and fixed illumination configuration. Although it can ensure a certain degree of visual controllability, it still has problems such as long production cycle, high resource consumption and insufficient scene reuse ability in complex space reconstruction, detailed texture restoration, dynamic perspective generation and real-time interactive feedback [2, 3]. Especially in virtual exhibitions, immersive imaging, digital art installations and interactive narrative Spaces, scenes need to change in real time according to the user's position, gaze, gesture or control command, and it is difficult to meet the design requirements of high realism and high responsiveness by simply relying on static models and preset animations [4].

Neural rendering technology provides a new technical path for digital media art scene generation. This technology learns the mapping relationship between images, camera poses, geometric structures, material textures and lighting relations through deep neural networks,

*13637929791@163.com
https://doi.org/10.65102/is2026976
Published: 30 April, 2026

and can reconstruct continuous 3D space based on limited view data, and generate visual results with real light and shadow and spatial consistency [5]. Neural radiance field can describe the density and color distribution of the scene through implicit expression, which is suitable for multi-view reconstruction of complex art space. 3D Gaussian representation improves real-time display efficiency through explicit point cloud and differentiable rendering mechanism, and provides support for rapid rendering and interactive update of immersive scenes [6, 7]. Introducing the above methods into digital media art design can reduce the pressure of manual modeling, improve the automation level of scene generation, and enhance the detail representation and visual coherence of virtual space.

However, the application of neural rendering for immersive digital media art scenes still faces several key problems [8]. Artistic scenes often contain irregular shapes, complex materials, local strong light and shadow, and stylized textures. Common 3D reconstruction methods are prone to boundary blurring, texture drift, and unstable viewpoint switching. In the real-time rendering phase, there is an obvious contradiction between high-resolution output and low-latency interaction, and the model parameter scale, video memory occupation and rendering frame rate will directly affect the system operation effect. At the interaction design level, how to establish the mapping relationship between user behavior data and neural rendering scene states, and how to ensure the scene update accuracy and visual stability after interactive commands are triggered also need further research [9, 10].

Based on this, this paper focuses on the technical link of "scene generation-real-time rendering-interactive response", and constructs an immersive digital media art scene generation and interaction design method driven by neural rendering. Starting from multi-source visual data acquisition and semantic coding of art scene, this paper uses neural radiance field to complete 3D space reconstruction, and fuses 3D Gaussian representation to optimize real-time rendering efficiency. On this basis, a user behavior perception mechanism is introduced to transform perspective movement, gesture input and interaction instructions into scene dynamic update parameters, forming an immersive interaction process with real-time feedback. The experiment verifies the scene reconstruction quality, visual generation effect, system response efficiency and dynamic interaction stability, which provides a feasible technical reference for intelligent scene construction and interaction design in digital media art creation.

2 Related work

Neural rendering has become an important research direction in the field of 3D visual computing and immersive scene generation. Liao et al. systematically reviewed the development of neural radiance fields, implicit representation, view synthesis and accelerated training methods, and pointed out that NeRF can achieve high-quality new view generation through continuous space modeling, which provides a basis for complex digital scene reconstruction [11]. Šlapak et al. further extended the neural radiance field to industrial and robotics scenarios, emphasizing its potential application in complex environment reproduction, spatial perception and virtual simulation [12]. Li et al. proposed the Magic NeRF Lens framework to enhance the user's ability to observe NeRF scenes through interactive focus and context exploration, so that neural rendering is no longer limited to static view generation, but gradually enters an immersive space that can be manipulated and explored [13]. Stacchio et al. discussed the credibility of neural rendering from the perspective of cultural heritage and creative industries, and pointed out that data sources, generation authenticity and artistic expression boundaries would affect the communication value of digital content [14].

Research on neural rendering for virtual reality and extended reality pays more attention to real-time performance and visual perception efficiency. Deng et al. proposed FoV-NeRF to reduce the computational load in VR scenes by preferentially rendering the gaze area, and improve the display efficiency while ensuring the quality of the central field of view [15]. Wang et al. constructed the VPRF method to introduce the visual perception mechanism into the radiation field generation process, so that the image generation results were more in line with the human eye attention distribution [16]. Shi et al. proposed a scene-aware gaze rendering method to further adjust rendering resource allocation by using scene structure differences and enhance stability under complex viewpoint switching [17]. Li et al. reviewed the radiation field applications in XR and pointed out that neural rendering has strong adaptability in head-mounted display, spatial reconstruction, remote collaboration and virtual interaction, but it is still limited by frame rate, delay and device computing power [18].

At the specific application level, Kleinbeck et al. used neural digital twin to reconstruct complex medical environment and used it for spatial planning in virtual reality, demonstrating that neural rendering can serve high-precision spatial reproduction tasks [19]. Fabra et al. used neural radiance field to represent 3D models in the industrial metaverse and verified its usability in virtual-real fusion scenarios [20]. Atik compares the performance of neural radiance field and 3D Gaussian splash in UAV image point cloud generation, and shows that 3D Gaussian representation has advantages in reconstruction efficiency and explicit expression, which can provide technical support for real-time artistic scene rendering [21].

In terms of immersive digital media art, Mills and Brown believe that virtual reality can change the way of expression in digital media creation, and creators can complete cross-modal conversion through spatialized media [22]. Paatela-Nieminen discussed the restructuring effect of fully immersive VR on real and imagined content in art education, reflecting the expansion value of immersive space for artistic narrative and perceptual experience [23]. Serna-Mendiburu and Guerra-Tamez pointed out that VR is promoting the transformation of art and design learning and helping to form a more participatory creation environment [24]. Coruh further focuses on the role of immersive digital creation environment in design education, emphasizing the important influence of interactive space, real-time feedback and collaboration of digital tools on creative generation [25]. In summary, the existing research has formed a good foundation in neural rendering, XR display, and immersive art creation, but there is still room for expansion of the research on unifying NeRF reconstruction, 3D Gaussian real-time rendering, and user behavior interaction into the digital media art scene generation process.

3 Neural rendering-driven immersive digital media art scene generation and interaction design method

3.1 Multi-source visual data acquisition and artistic scene semantic feature coding

The immersive digital media art scene has the characteristics of complex spatial structure, rich material changes, continuous switching of viewpoints and real-time change of interactive behaviors. It is difficult for a single image or static model to fully express its visual relationship. In order to ensure stable input for subsequent neural radiance field reconstruction and 3D Gaussian real-time rendering, we construct a scene dataset from five sources: multi-view RGB images, depth data, camera pose, artistic semantic annotation and user behavior data. Multi-view RGB images are used to preserve color, texture, stroke and device

boundary information. Depth map and point cloud data are used to supplement spatial distance and geometric structure. Camera intrinsic and extrinsic parameters are used to establish the mapping relationship between 2D image and 3D space. Artistic semantic annotation is used to distinguish the subject, background, light and shadow area and material type. The user behavior data is used for the dynamic state control of the subsequent interactive response phase.

Multi-source data need to be aligned with timestamp, unified resolution, noise filtering and unified coordinate system after acquisition. Suppose that the ith view data collected at the same time is composed of image, depth, camera parameters, semantic label and interaction state, and its unified data expression is as follows:

$$\mathcal{D}_t^i = \left\{ I_t^i, D_t^i, K^i, R^i, t^i, S_t^i, B_t \right\} \tag{1}$$

where, I_t^i represents the RGB image collected at the ith view point at time t, D_t^i represents the corresponding depth map, K^i is the camera internal reference matrix, R^i and ti represent the rotation matrix and translation vector respectively, S_t^i is the artistic semantic label, B_t is the user behavior state. Through this structure, visual, spatial, and behavioral information can be put into the same data frame, providing a unified input for subsequent neural rendering modeling.

In the 3D spatial encoding stage, 2D pixels need to be converted to a unified world coordinate system. For pixel point (p,q) in the image, its depth value is $d_t^i(p, q)$, and the 3D world coordinates corresponding to this point can be expressed as follows:

$$x_w = (R^i)^{-1}\left(d_t^i(p, q)(K^i)^{-1}[p, q, 1]^T - t^i \right) \tag{2}$$

where, x_w represents the world coordinates of the pixel point after back-projection, and $[p, q, 1]^T$ represents the homogeneous pixel coordinates. This process associates the texture information in multi-view images with the 3D spatial position, which can reduce the view drift and spatial dislocation in subsequent scene reconstruction.

In the visual feature encoding stage, this paper uses convolutional network and vision Transformer to jointly extract local texture and global structure features, and synchronously embed deep features into a unified vector space:

$$f_v^i = \mathrm{LN}\left(\phi_{cnn}(I_t^i) \oplus \phi_{vit}(I_t^i) \oplus \psi(D_t^i) \right) \tag{3}$$

where, ϕ_{cnn} represents the convolutional feature extraction function, which is used to capture edges, textures, and local material variations. ϕ_{vit} represents the Vision Transformer feature extraction function for modeling long-range spatial relationships; Let ψ denote the depth encoding function; $\oplus$ represents feature concatenation; LN denotes the layer normalization operation. This encoding method can preserve the local brushwork details and the overall spatial layout in the art scene.

The main area, light and shadow areas, and device boundaries in art scenes have a great impact on the rendering quality, so it is necessary to introduce a semantic attention mechanism to enhance the expression of key areas. Let the KTH semantic prototype be s_k and the visual query vector be q_i, then the semantic attention weight is as follows:

$$\alpha_{i,k} = \frac{\exp\left(q_i^T s_k / \sqrt{d} \right)}{\sum_{k=1}^{K} \exp\left(q_i^T s_k / \sqrt{d} \right)} \tag{4}$$

where, $\alpha_{i,k}$ represents the attention degree of the ith view feature to the KTH semantic region,k represents the total number of semantic categories, and d represents the feature dimension. This weight can improve the participation intensity of subject boundaries, material changes, and lighting transition regions in subsequent rendering.

After integrating visual features, 3D coordinates, semantic weights and user behavior states, the final scene encoding vector is formed as follows:

$$h_i = \mathrm{MLP}\left([f_v^i, x_w, \sum_{k=1}^{K} \alpha_{i,k}\, s_k, B_t]\right) \tag{5}$$

where h_i represents the integrated feature vector input to the neural rendering network, MLP represents the multilayer perceptron, and $[\cdot]$ represents the vector cascade. The feature not only contains image texture and spatial geometry relationship, but also contains artistic semantics and interaction state, which can support subsequent scene reconstruction, real-time rendering and dynamic interaction design.

In order to facilitate the illustration of the role of different data types in the coding process, this paper arranges the configuration of multi-source visual data acquisition and semantic feature coding, as shown in Table 1.

Table 1: Configuration table of multi-source visual data acquisition and semantic feature encoding

Data Type	Collected Content	Preprocessing Method	Encoded Features	Technical Function
Multi-view RGB Images	1920×1080 art scene images and local texture images	Deblurring, color normalization, view filtering	Texture features, color features, edge features	Preserves visual details of digital media art scenes
Multi-view Video Frames	Continuous scene video sequences at 30 fps	Key-frame extraction, timestamp alignment, duplicate-frame removal	Temporal visual features, viewpoint variation features	Supports scene reconstruction under continuous viewpoint changes
Depth Maps and Point Clouds	Depth camera data, sparse point clouds, local spatial distance information	Outlier removal, voxel downsampling, coordinate unification	Geometric structure features, spatial position features	Improves spatial stability of 3D reconstruction
Camera Pose Data	Camera intrinsic parameters, extrinsic parameters, shooting trajectories	Pose optimization, scale calibration, coordinate registration	View direction features, spatial transformation features	Establishes the mapping relationship between 2D images and 3D space
Artistic Semantic Annotations	Subject regions, background regions, material types, light-shadow regions	Semantic segmentation, manual verification, label unification	Regional semantic features, style control features	Supports region-wise rendering and artistic style representation
User Behavior Data	Headset pose, gesture input, gaze focus, control commands	Kalman filtering, outlier removal, state encoding	Interaction state features, behavior triggering features	Drives dynamic scene feedback and interactive response

As can be seen from Table 1, the multi-source data is hierarchically encoded around the scene generation and interaction feedback requirements. The RGB images and video frames provide the visual foundation, the depth and pose data ensure the spatial consistency, the semantic annotation enhances the expression of the art area, and the user behavior data enables the generated scene to be dynamically updated according to the interaction state.

Through this encoding flow, immersive digital media art scenes can be transformed from static visual materials into computable, reconstructable, and interactive neural rendering inputs.

3.2 3D Reconstruction method of immersive artistic scene Based on Neural radiance Field

After the multi-source visual data and semantic feature encoding are completed, the neural radiance field is further used to construct the continuous 3D representation of the immersive art scene. Different from traditional mesh modeling that relies on explicit geometric facets, neural radiance field describes the density and color distribution of spatial points through implicit neural networks, which can recover art installations, light and shadow structures, material textures, and spatial hierarchical relationships under the constraints of multi-view images. For immersive digital media art scenes, translucent media, strongly reflective materials, irregular boundaries and local stylized textures are often included in the scene. Directly using sparse point clouds or ordinary 3D meshes is prone to holes, edge fracture and texture dislocation. Neural radiance field can learn view-dependent color changes in continuous coordinate space, so it is more suitable for expressing illumination transition and visual continuity in complex art Spaces.

In this paper, the integrated feature vector obtained in Section 3.1 is used as the conditional input, and the radiation field mapping relationship is established by combining the 3D space coordinates with the viewing direction. Let the spatial sampling point be x, the viewing direction be d, and the semantic encoding vector be hi, then the neural radiance field network can be expressed as follows:

$$F_{\Theta}(\gamma(x), \gamma(d), h_i) = (\sigma, c) \tag{6}$$

where, F_{Θ} represents the neural radiance field network composed of multi-layer perception mechanisms, Θ is the network parameter, $\gamma(\cdot)$ represents the position encoding function, σ is the spatial volume density, and c is the view-dependent color. The position encoding can map the low-dimensional space coordinates to the high-frequency feature space, so that the network can better learn the detail edges, texture strokes, and local light and shadow changes in the art scene.

In the 3D reconstruction process, each pixel corresponds to a spatial ray emitted from the center of the camera. Suppose the camera optical center is o, the ray direction is d, and the sampling distance is u. Then any sampling point on the ray can be expressed as follows:

$$r(u) = o + ud, \quad u \in [u_n, u_f] \tag{7}$$

where, r(u) represents the 3D sampling point on the ray, and u_n and u_f represent the distance between the near clipping surface and the far clipping surface respectively. By stratified sampling in this interval, the system can continuously model the foreground subject, background structure, and transparent artistic medium, avoiding missing spatial hierarchy caused by a single depth estimation.

The neural radiance field accumulates the color and density of multiple sampling points on a ray into pixel color by volume rendering. The predicted color of a pixel after discrete sampling is calculated as follows:

$$\hat{C}(r) = \sum_{j=1}^{N} T_j \left(1 - \exp(-\sigma_j \delta_j)\right) c_j \tag{8}$$

Among them:

$$T_j = \exp\left(-\sum_{m=1}^{j-1} \sigma_m \delta_m\right) \tag{9}$$

where, $\hat{C}(r)$ represents the predicted color of ray r, N is the number of ray sampling points, σ_j and c_j represent the volume density and color of the JTH sampling point, δ_j represents the distance between adjacent sampling points, and T_j represents the cumulative transmittance of the ray before it reaches the JTH sampling point. The calculation process can simulate the occlusion, transparency and color superposition effects of light passing through different spatial regions, so that the immersive art scene can maintain good visual consistency under multi-view switching.

In order to improve the representation ability of the reconstruction results for artistic semantic regions, this paper introduces color reconstruction error, depth constraint error and semantic boundary constraint error in the training stage. The combined loss function is defined as follows:

$$\mathcal{L}_{nerf} = \mathcal{L}_{rgb} + \lambda_d \mathcal{L}_{depth} + \lambda_s \mathcal{L}_{sem} \tag{10}$$

where, $\mathcal{L}_{rgb}$ represents the error between the predicted color and the real image color, $\mathcal{L}_{depth}$ is used to constrain the consistency between the predicted spatial depth and the acquired depth, and $\mathcal{L}_{sem}$ is used to strengthen the reconstruction effect of the subject boundary, material partition, and light and shadow transition region, λ_d and λ_s is the loss weight. Through the joint optimization of multiple losses, the model can not only restore the overall structure of the scene, but also enhance the expression accuracy of the edges, texture distribution and spatial illumination of the art installation.

To illustrate the 3D reconstruction process of this paper, the 3D reconstruction process of immersive art scene based on neural radiance field is shown in Figure 1.

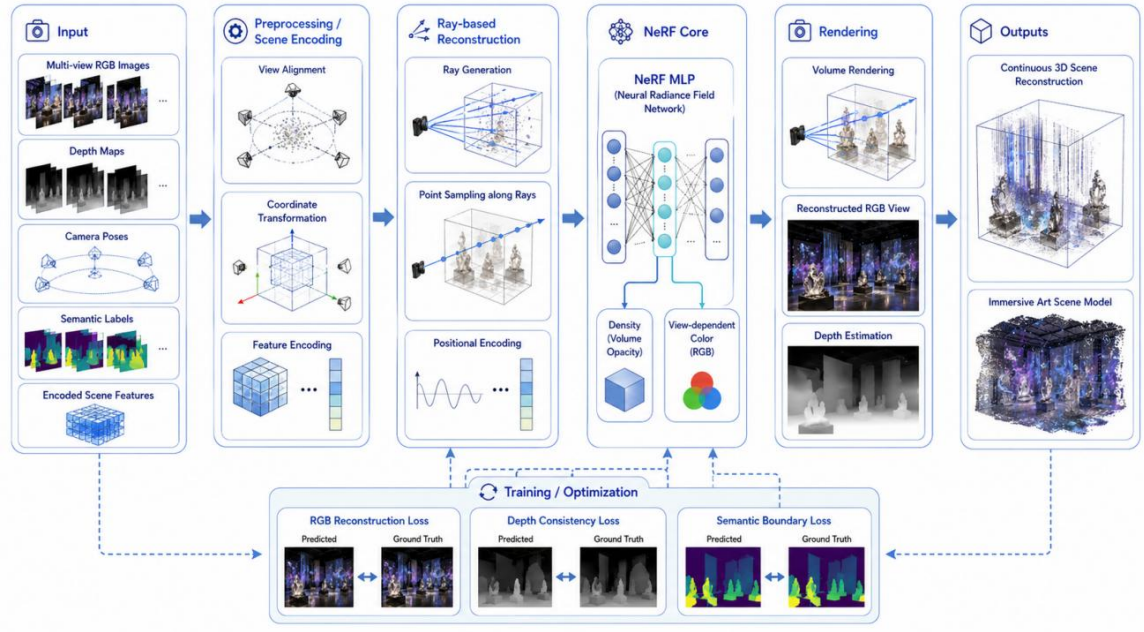

Figure 1: Flowchart of 3D reconstruction of immersive artistic scene based on neural radiance field

As can be seen from Figure 1, the proposed 3D reconstruction method forms a continuous spatial expression through camera pose, ray sampling, semantic feature fusion, and volume rendering optimization. This process can provide stable initial geometric structure and view-consistent texture information for subsequent 3D Gaussian real-time rendering, and also lay the foundation for dynamic scene update in interactive state.

3.3 Real-time Neural Rendering Optimization with 3D Gaussian Representation

After the 3D reconstruction of the neural radiation field is completed in Section 3.2, the scene has the ability to express continuous space. However, NeRF requires a large number of network queries in the process of ray-by ray sampling and volume rendering integration, which is easy to cause frame rate degradation and response delay increase when directly used in immersive digital media art scenes. In order to improve real-time rendering performance, we introduce 3D Gaussian representation based on NeRF reconstruction results, transform the continuous implicit radiance field into an explicitly optimized set of Gaussian primitives, and achieve fast image synthesis through differentiable rasterization. The proposed method can reduce the computational cost in multi-view interactive display while preserving the lighting, material and spatial hierarchy of the art scene.

The 3D Gaussian representation decomposes the scene into a number of Gaussian primitives with spatial position, scale, rotation, transparency, and color attributes. Let the NTH Gaussian primitive be:

$$G_n = \{\mu_n, \Sigma_n, \alpha_n, c_n\} \tag{11}$$

where μ_n represents the 3D center position of the NTH Gaussian primitive, Σ_n represents the covariance matrix, α_n represents the transparency, and c_n represents the color feature. Through this explicit primitive representation, the system can convert the sculpture contour, projected texture, translucent device and spatial light spot in the art scene into data units that can be rendered in parallel, thus improving the computational efficiency of the GPU.

In order to ensure that the Gaussian primitive can express the spatial deformation in different directions, the covariance matrix is decomposed into a combination of a scale matrix and a rotation matrix:

$$\Sigma_n = R_n S_n S_n^T R_n^T \tag{12}$$

where, R_n denotes the rotation matrix of the Gaussian primitive and S_n denotes the scale matrix. This decomposition method can make the Gaussian primitive form an ellipsoidal distribution in three-dimensional space, and adapt to different forms of spatial objects in the art scene. For example, the elongated light band can be represented by Gaussian primitives with large stretching scale, the local point reflection can be represented by small scale primitives with high transparency, and the complex material surface can be superimposed by multiple Gaussian primitives to form a continuous visual effect.

In the real-time rendering phase, 3D Gaussian primitives need to be projected into 2D screen space. Let the camera transformation matrix be W and the local Jacobian matrix of the projection function be J. Then the two-dimensional screen space covariance can be expressed as follows:

$$\Sigma_n' = JW\Sigma_n W^T J^T \tag{13}$$

where, Σ_n'represents the 2D Gaussian distribution range after projection. This process maps 3D ellipsoidal primitives to elliptical blobs on the screen, which facilitates batch rendering by the GPU through the rasterization pipeline. Compared with point-by-point ray sampling, 2D projection and rasterization are more suitable for real-time graphics rendering process, and can significantly reduce the computation time of a single frame in immersive interaction.

The screen pixel color is obtained by transparency blending of multiple Gaussian primitives in depth order. Suppose that the pixel position is p, and there are M Gaussian primitives involved in the rendering of the pixel, then the color composition process is as follows:

$$\hat{C}(p) = \sum_{m=1}^{M} \tau_m(p)\alpha_m(p)c_m \tag{14}$$

Among them:

$$\tau_m(p) = \prod_{r=1}^{m-1}(1 - \alpha_r(p)) \tag{15}$$

where, $\hat{C}(p)$ represents the predicted color of pixel p, $\alpha_m(p)$ represents the transparency contribution of the MTH Gaussian primitive to the pixel, and $\tau_m(p)$ represents the cumulative transmission influence of the preceding primitive on the current primitive. Through transparency blending, the system can represent the occlusion relationship, light and shadow superposition and translucent media effects in art scenes, so that the real-time rendering results have a strong sense of spatial hierarchy.

In order to improve the rendering quality and operation efficiency, this paper jointly optimizes the Gaussian primitive parameters. The optimization objective simultaneously considers color reconstruction error, structural similarity error and primitive sparsity constraint:

$$\mathcal{L}_{gs} = \|\hat{\mathbf{C}} - \mathbf{C}^*\|_1 + \lambda_q(1 - \mathrm{SSIM}(\hat{\mathbf{C}}, \mathbf{C}^*)) + \lambda_r \sum_{n=1}^{N} \|\boldsymbol{\Sigma}_n\|_F \tag{16}$$

where, C^* represents the real image color, SSIM is used to measure the structural consistency between the generated image and the real image, $\|\boldsymbol{\Sigma}_n\|_F$ represents the norm constraint of the covariance matrix, and λ_q and λ_r are the weight coefficients. The loss can suppress the excessive diffusion of Gaussian primitivity, reduce boundary blur and texture drift, and make digital media art scenes maintain good visual quality at high frame rate rendering.

To illustrate the real-time neural rendering optimization process of this paper, the real-time neural rendering optimization framework fusing 3D Gaussian representation is shown in Figure 2.

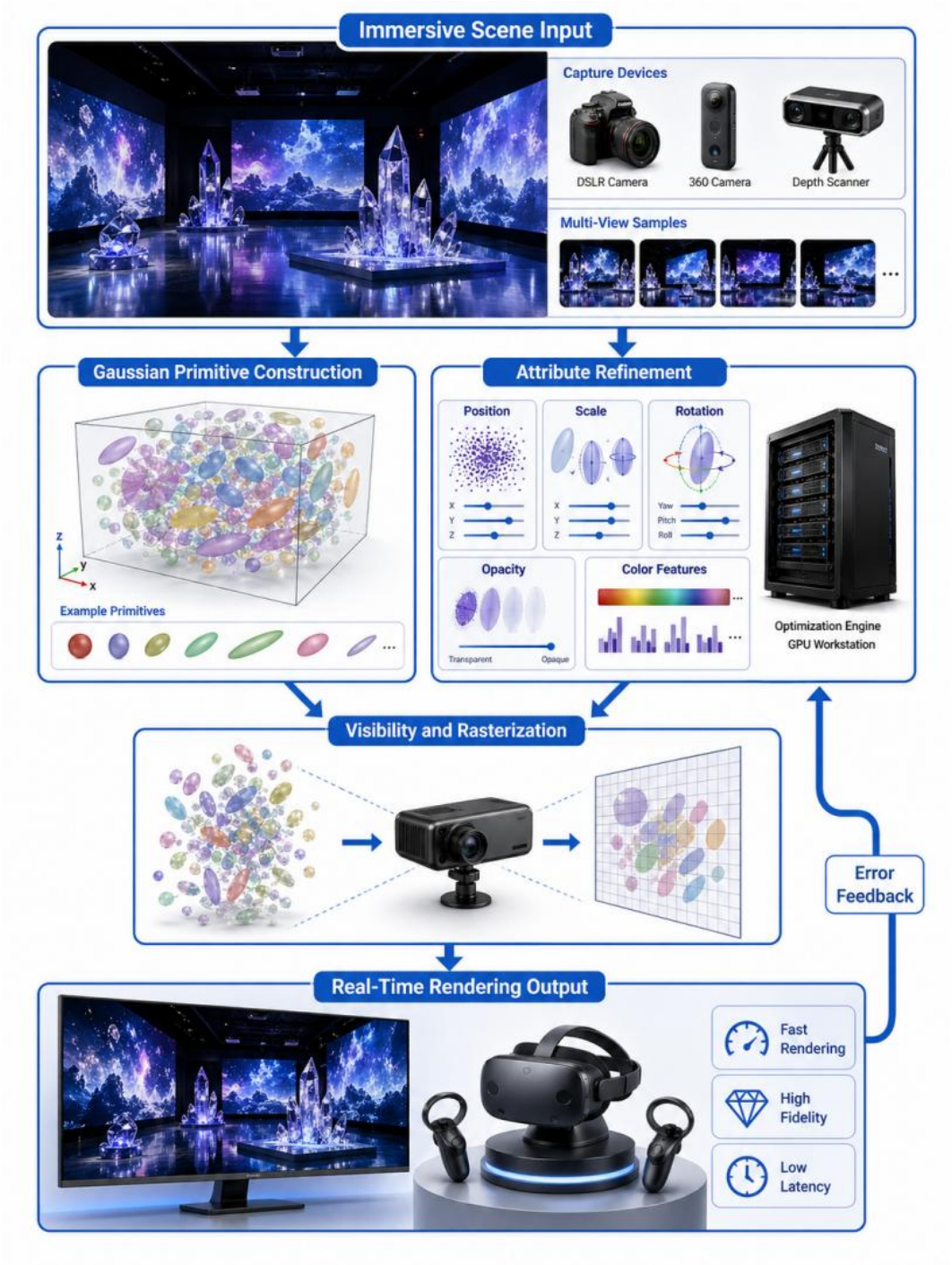

Figure 2: Framework diagram of real-time neural rendering optimization fusing 3D Gaussian representations

As can be seen from Figure 2, starting from the NeRF reconstruction results, the system initializes the scene structure to a set of 3D Gaussian primitives, and jointly optimizes the position, scale, rotation, transparency, and color features. Then, the three primitiles are projected into screen space by differentiable rasterization to generate real-time rendering images, and the rendering errors are fed back to the optimization module. The proposed flow combines implicit reconstruction accuracy with explicit rendering efficiency, enabling immersive artistic scenes to remain continuously displayed during head-mounted viewpoint movement, scene scaling, and interactive control. For digital media art creation, 3D Gaussian

representation not only improves the real-time rendering frame rate, but also enables scene light and shadow, material textures and spatial devices to participate in interactive updates in a more flexible way, which provides an efficient graphics computing basis for subsequent dynamic interactive response mechanisms.

3.4 Immersive Interaction mapping and response generation method for user behavior perception

The interactive effect of immersive digital media art scene depends not only on the rendering quality, but also on whether the system can accurately perceive the user's behavior and complete the scene feedback under low latency conditions. Aiming at the characteristics of the change of view Angle, gesture operation, eye focus and control command at the same time, this paper constructs an interaction mapping and response generation method for user behavior perception, which converts multimodal behavior input into control signals that can drive scene update, and improves interaction stability through state prediction and feedback scheduling mechanism. The system collects head pose, hand joint points, gaze points, controller states and voice trigger information in the input layer, completes behavior fusion, intention recognition and action mapping in the computing layer, and updates view parameters, object states, light and shadow effects and art content playback logic in the rendering layer, thus forming a closed-loop interaction process of "perception, understanding, response and feedback".

Let the user behavior state vector at time t be as follows:

$$b_t = [p_t, q_t, g_t, h_t, u_t] \tag{17}$$

where p_t represents the head spatial position, q_t represents the head pose quaternion, g_t represents the gaze feature, h_t represents the gesture feature, and u_t represents the controller or voice command encoding. This vector unifies spatial actions and interaction instructions into the same feature space, which provides the input basis for subsequent intention recognition.

Due to the jitter, occlusion and sampling error in the original behavior signal, this paper uses a temporal smoothing strategy to update the user state:

$$\tilde{b}_t = \beta \tilde{b}_{t-1} + (1-\beta) b_t \tag{18}$$

where, $\tilde{b}_t$ represents the behavior state after smoothing and β is the smoothing coefficient. This processing can reduce the interference of head micro-jitter and gesture noise on interaction judgment, and make scene switching more stable. The smoothed behavior sequence is sent to the multi-modal fusion module. This paper uses the attention mechanism to dynamically assign the importance of different modalities, and its fusion is expressed as follows:

$$z_t = \sum_{m=1}^{M} \omega_t^m \phi_m(\tilde{b}_t^m), \quad \omega_t^m = \frac{\exp(e_t^m)}{\sum_{m=1}^{M} \exp(e_t^m)} \tag{19}$$

where, M represents the number of modalities, ϕ_m represents the feature mapping function of the MTH modality, ω_t^m represents the corresponding weight, and z_t is the fused interactive semantic vector. This mechanism can automatically increase the participation degree of gaze, gesture or voice information according to the current task, thus enhancing the context adaptability of interaction judgment.

In the intention recognition stage, the system generates interaction action categories based

on the fused features and maps them into the scene control set:

$$a_t = \arg\max\ (\text{Softmax}(W_a z_t + b_a)) \tag{20}$$

where, a_t represents the interactive action category at time t, W_a and b_a are the classification layer parameters. The recognition results can be used for view navigation, object selection, content triggering, parameter adjustment and scene switching. After the action is generated, the system updates the rendering control parameters based on the current scene state st:

$$\mathbf{s}_{t+1} = f(\mathbf{s}_t, a_t, \mathbf{z}_t), \mathcal{R}_t = \lambda_1 A_t + \lambda_2 C_t - \lambda_3 D_t 3 \tag{21}$$

where, $f(\cdot)$ represents the scene state update function, s_{t+1} is the updated scene state, Rt is the interactive response quality score, A_t represents the accuracy of action recognition, Ct represents the scene feedback consistency, D_t represents the response delay, λ_1, λ_2, λ_3

Is the weight coefficient. The objective can simultaneously constrain the interaction accuracy, visual continuity and response speed, so that the system can maintain a natural interactive experience in a complex art space.

To illustrate the interactive response link in this paper, the immersive interactive response mechanism oriented to user behavior perception is shown in Figure 3.

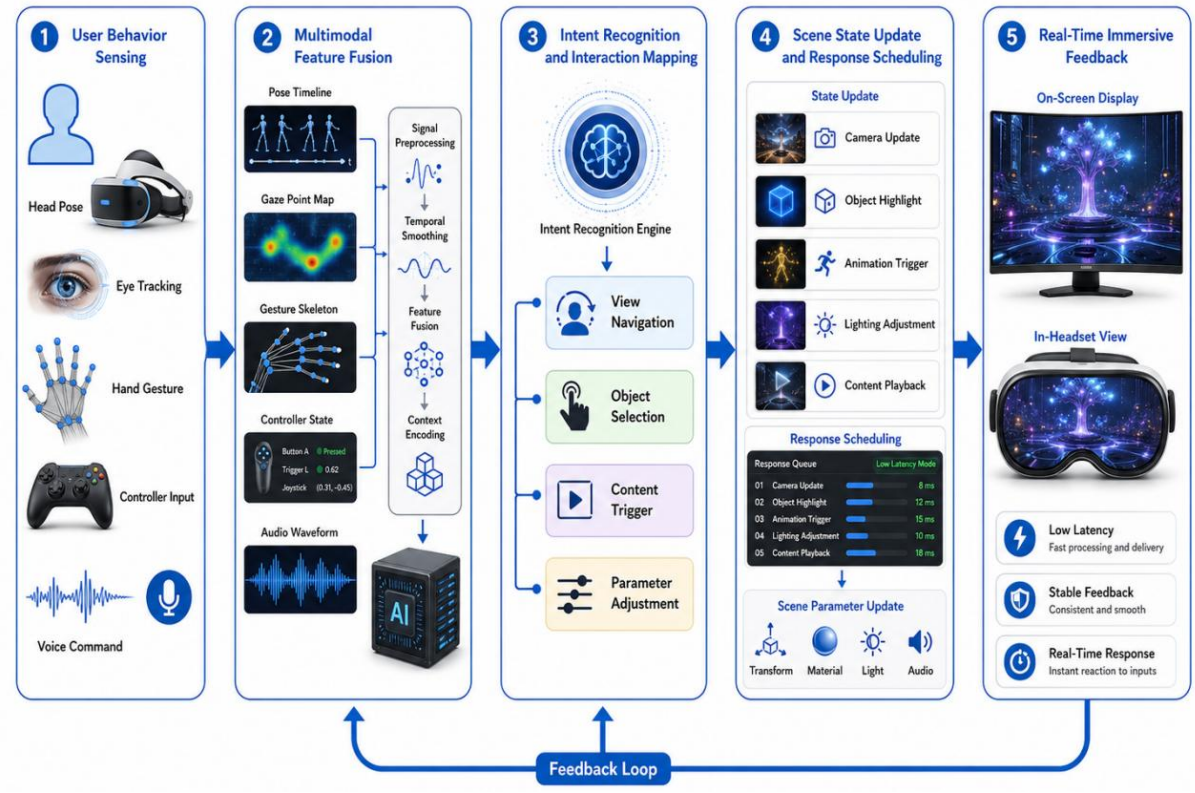

Figure 3: Diagram of immersive interactive response mechanism for user behavior perception

4 Experimental Verification

4.1 Experimental environment configuration and evaluation index setting

In order to verify the effectiveness of the proposed method in immersive digital media art

scene generation and interaction design, an experimental platform is built, which integrates multi-view data processing, neural radiance field reconstruction, 3D Gaussian real-time rendering and interactive response testing. The hardware environment uses Intel Core i9-13900K processor, 64 GB memory and NVIDIA RTX 4090 GPU to ensure that neural network training, differentiable rendering calculation and high-resolution scene display can be completed on a unified platform. The software environment is configured as Ubuntu 22.04, Python 3.10, PyTorch 2.1, CUDA 12.1, and Unity 2022.3, where PyTorch is used for neural rendering model training, CUDA is used for GPU parallel acceleration. Unity is used for immersive interactive scene deployment and real-time display testing.

The experimental data consists of four types of samples: projection art space, digital exhibition hall, installation art space and complex light and shadow scene. A total of about 6200 multi-view RGB images are collected, and the depth map, camera pose, semantic region label and user interaction log are synchronously recorded. All images are uniformly adjusted to 1920×1080 resolution. In the training stage, random cropping, exposure perturbation, view perturbation and color normalization are used to improve the adaptability of the model to complex illumination, partial occlusion and multi-view changes. In the interactive test part, the head pose, eye focus, gesture input and controller commands are collected to evaluate the dynamic feedback effect of the scene driven by user behavior.

The evaluation indicators revolve around visual generation quality, real-time rendering efficiency, computing resource occupation and interactive responsiveness. The visual quality is measured by PSNR, SSIM and LPIPS, which reflect image clarity, structural consistency and perceptual difference respectively. Real-time performance is evaluated by FPS and single frame rendering delay, focusing on whether the system meets the requirements of continuous frame rate and low delay for immersive display. The computational load is recorded by GPU memory and neural rendering pipeline stage time. Interaction effects were analyzed using response accuracy, scene update error, and stability score. Through the above configuration, the experiment can comprehensively verify the performance of the method from three levels of reconstruction accuracy, rendering speed and interactive stability.

4.2 Art scene reconstruction quality and visual generation effect verification

In order to verify the reconstruction adaptability of neural rendering method for different types of digital media art scenes, this paper selects projection art space, digital exhibition hall, installation art space and complex light and shadow scenes for testing, and analyzes the reconstruction effect from three aspects: spatial error, texture restoration and illumination consistency. Figure 4 shows the spatial distribution of neural reconstruction errors under different art scene types.

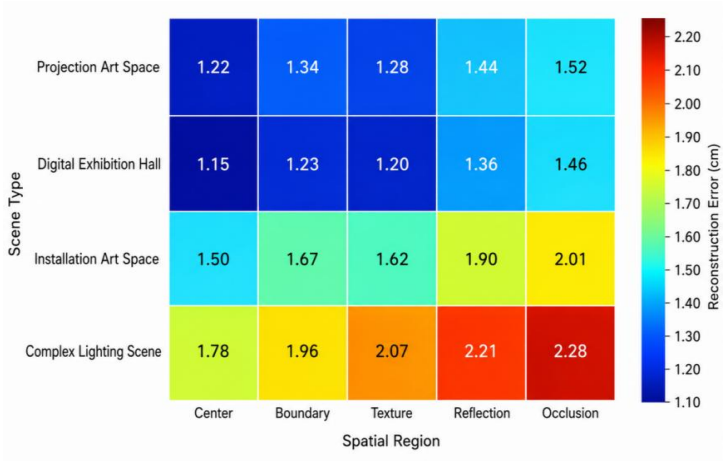

Figure 4: Heat maps of spatial distribution of neural reconstruction errors for different art scene types

Figure 4 shows that the reconstruction error distribution of the proposed method is relatively concentrated in the digital exhibition hall and the projection art space, and the average spatial error is 1.28 cm and 1.36 cm respectively, indicating that the regular display structure and continuous projection texture can be well modeled by the neural radiance field. Due to the existence of many irregular surfaces and translucent materials in the installation art space, the average error rises to 1.74 cm. The complex light and shadow scene is affected by strong reflection, partial occlusion and dark part noise, and the average error is 2.06 cm. The high error area is mainly concentrated in the edge projection and mirror reflection position, and the main structure still maintains a relatively complete spatial continuity. Overall, the average reconstruction error of the four types of scenes is controlled at 1.61 cm, indicating that the method has a good ability to reproduce immersive art space.

After the spatial reconstruction analysis is completed, the generation effect is further verified from the correlation between the objective quality index and the subjective visual evaluation. The correlation of subjective and objective indicators of art scene generation quality is shown in Figure 5.

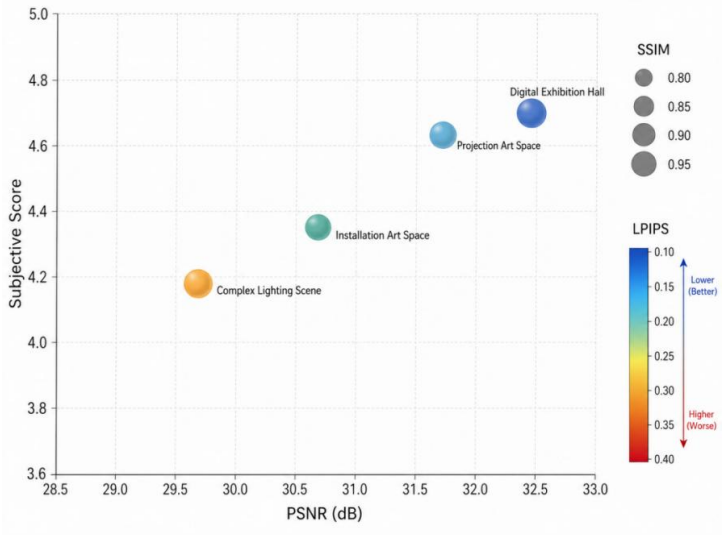

Figure 5: Bubble plot of correlation between subjective and objective indicators of artistic scene generation quality

As can be seen from Figure 5, PSNR and SSIM are positively correlated with subjective scores as a whole, while LPIPS is negatively correlated with subjective scores, indicating that the structural consistency and perceptual difference of generated images can better reflect users' judgment on the visual quality of artistic scenes. The PSNR of the projection art space reaches 31.8 dB, the SSIM is 0.924, and the subjective score is 4.62. The PSNR of the digital exhibition hall is 32.4 dB, the SSIM reaches 0.936, and the subjective score is the highest, 4.71. The LPIPS of installation art space and complex light and shadow scenes are 0.132 and 0.158, respectively, which are slightly higher than those of the first two categories of scenes, but still remain in a low range. The results show that the proposed method achieves a good balance between visual clarity, structure preservation and artistic perception quality.

4.3 Real-time rendering performance and system response efficiency analysis

Real-time rendering performance directly affects the interactive continuity of immersive digital media art scenes. In this paper, the system frame rate is tested by setting low complexity, medium complexity and high complexity scenes under four resolutions of 720p, 1080p, 2K and 4K. The low complexity scene includes a small amount of art installations and basic projection textures. The medium complexity scene adds multi-layer light and shadow, translucent materials, and dynamic particles. The real-time rendering frame rate variation under different resolutions and scene complexity is shown in Figure 6.

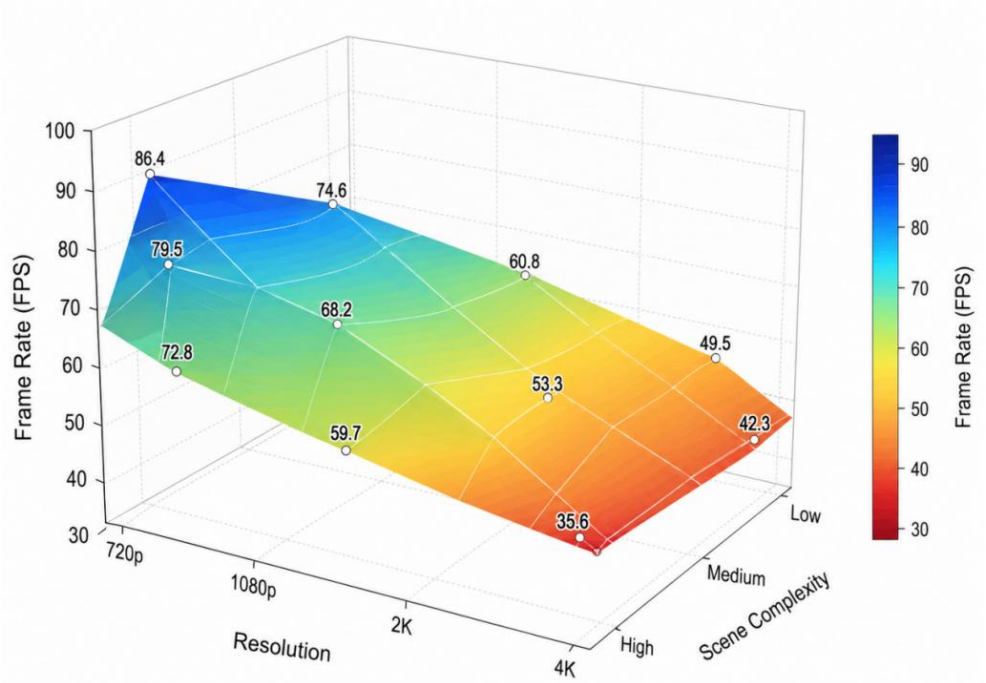

Figure 6: Real-time rendering of frame rate 3D surface plots at different resolutions and scene complexity

It can be seen from Figure 6 that with the simultaneous increase of resolution and scene complexity, the frame rate of the system shows a stepwise decline trend, but the overall system still maintains a good real-time rendering ability. Under the condition of 720p, the frame rate of low-complexity scene reaches 86.4FPS, and the frame rate of high-complexity scene still maintains 72.8FPS. Under the condition of 1080p, the frame rates of the three types of scenes are 74.6FPS, 68.2FPS and 59.7FPS, respectively, which can meet the basic

continuity requirements of immersive display. When the resolution is increased to 4K, the frame rate of the high-complexity scene is reduced to 35.6 FPS, which is mainly affected by the number of Gaussian primitives, transparency blending and light and shadow projection calculation. On the whole, the proposed method can balance the visual quality and interactive real-time performance in the range of 1080p to 2K, which is suitable for the dynamic display requirements of digital galleries and immersive art Spaces.

To further analyze the reasons for the frame rate variation, we calculate the time consumption of the main computational stages in the neural rendering pipeline, including data loading, feature encoding, Gaussian attribute optimization, differentiable raster, interactive state synchronization, and screen output. Figure 7 shows the time consumption of each stage of the neural rendering pipeline.

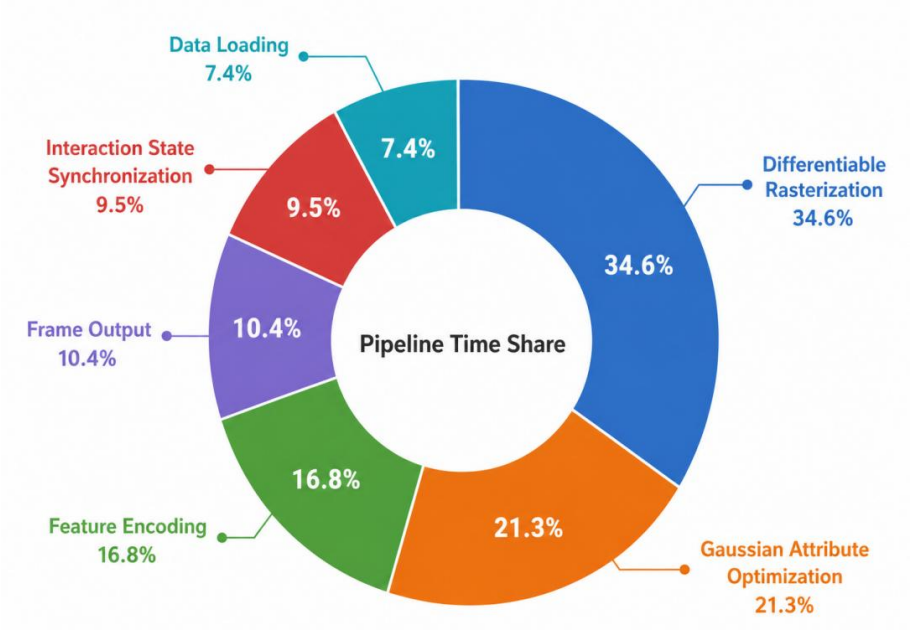

Figure 7: Circular diagram of the proportion of computation time consumed at each stage of the neural rendering pipeline

Figure 7 shows that differentiable rasterization accounts for the highest proportion of the overall time consumption, which is 34.6%, indicating that 3D Gaussian primitive projection and transparency blending are the key links affecting the efficiency of real-time rendering. Gaussian attribute optimization accounted for 21.3%, mainly from the parallel update of position, scale, rotation and color features. Feature encoding accounted for 16.8%, interactive state synchronization accounted for 9.5%, data loading and screen output accounted for 7.4% and 10.4%, respectively. In the whole test, the average single frame response delay of the system is 18.6 ms, and the GPU memory occupation is stable within 8.7 GB. The results show that 3D Gaussian representation can reduce the computational pressure caused by traditional NeRF ray-by-ray sampling, and improve the system response efficiency by GPU parallel rasterization.

4.4 Verification of response accuracy and scene update stability under dynamic interaction tasks

In order to verify the recognition accuracy of the proposed mutual response mechanism in dynamic tasks, five typical interaction tasks are set up, including view navigation, object selection, content triggering, parameter adjustment and scene switching. During the test, the

system synchronously collected the head display pose, eye focus, gesture trajectory, controller input and voice command, and completed the interaction intention recognition through the multimodal fusion module. The results of multimodal interaction task identification are shown in Figure 8.

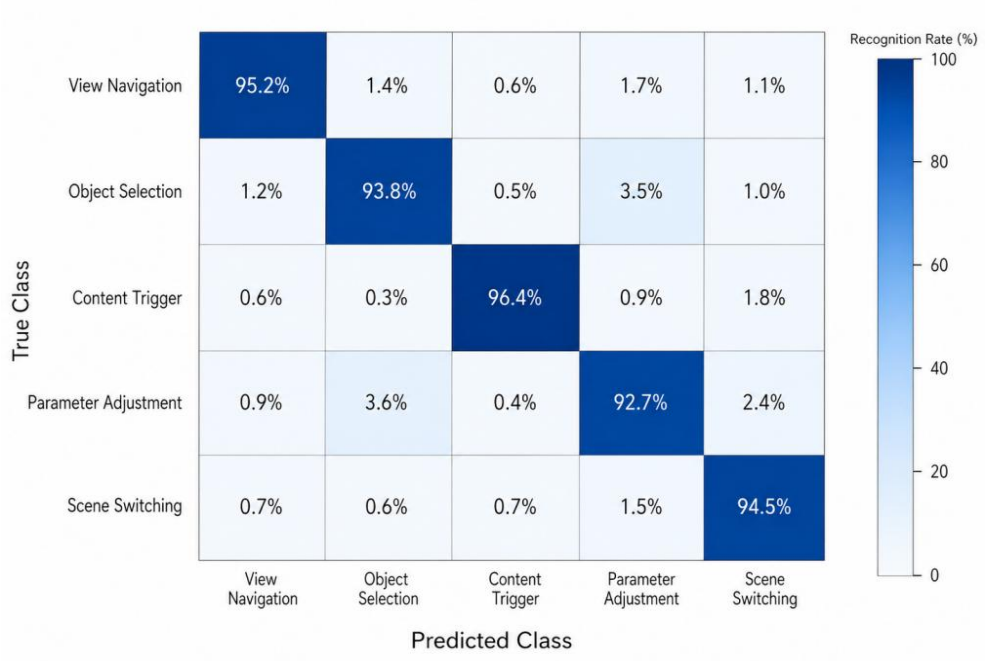

Figure 8: Confusion matrix diagram of multimodal interaction task recognition results

It can be seen from Figure 8 that the main diagonal values of the five types of interaction tasks are overall higher, indicating that the system can distinguish different user behavior intentions more accurately. Among them, the recognition accuracy of content-triggered tasks is the highest, reaching 96.4%. The main reason is that such tasks are usually triggered by controller keys or voice commands, and the input feature boundaries are relatively clear. The accuracy of view navigation task is 95.2%, and the accuracy of object selection task is 93.8%, which indicates that the fusion between head pose, eye focus and gesture positioning can effectively improve the ability of spatial interaction judgment. The accuracy of parameter adjustment task is relatively low, 92.7%, and the misjudgment is mainly concentrated between object selection and parameter adjustment. The reason is that the two types of tasks both include gesture stopping, dragging and fine-tuning actions, and there is a certain overlap in local behavior characteristics. The overall recognition accuracy reaches 94.5%, indicating that the proposed method can provide a stable basis for behavior understanding for real-time interactive control in immersive art scenes.

After the interactive task recognition is completed, the system also needs to update the scene state according to the user instructions, including camera perspective, object position, material parameters, light intensity, and content playback progress. To further examine the stability during continuous interaction, this paper records the scene state update error and drift trajectory under 50 rounds of continuous interaction. The scene state update drift during continuous interaction is shown in Figure 9.

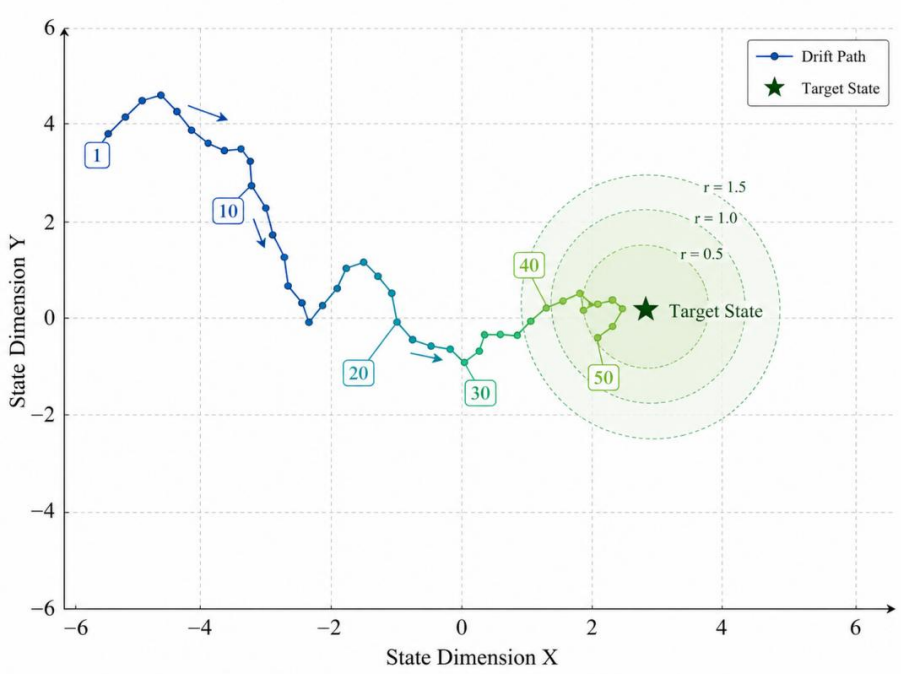

Figure 9: Scene state update drift trajectory diagram during continuous interaction

As can be seen from Figure 9, in the early stage of continuous interaction, the scene update drift is relatively obvious, and the average update error of the first 10 rounds is 0.46, which mainly comes from user gesture jitter, gaze jump and rendering state switching delay. With the continuous correction of the closed-loop feedback mechanism, the drift trajectory gradually converged to the target state, the average update error decreased to 0.24 after the 30th round, and further decreased to 0.18 after the 50th round, and the stability score increased from 88.6% to 96.1%. In the scene of high-frequency viewpoint switching and continuous object selection, there is no obvious screen tear or state jump, which indicates that the interactive feedback module can effectively suppress the cumulative error. Combining the results in Figs. 8 and 9, it can be seen that the proposed method can not only achieve high interactive intention recognition accuracy, but also maintain the smoothness and stability of scene update in continuous dynamic operation.

5 Discussion

This paper focuses on the application of neural rendering technology in immersive digital media art scene generation and interaction design. The method connects multi-source visual data encoding, neural radiance field reconstruction, 3D Gaussian real-time rendering and user behavior perception interaction as a complete technical link. The experimental results show that the proposed method can simultaneously take into account reconstruction quality, visual generation effect and interactive response efficiency in complex art scenes. The heat map of reconstruction error under different scene types shows that the average error of digital exhibition hall and projection art space is controlled at 1.28 cm and 1.36 cm, respectively, which indicates that regular space structure, continuous projection texture and stable illumination conditions are conducive to neural radiance field learning spatial density and color distribution. The errors in installation art space and complex light and shadow scenes increase, especially in the reflection area and occlusion area, the errors reach 2.21 cm and 2.28 cm, reflecting that translucent materials, specular reflection and local dark part noise still

cause interference to neural reconstruction.

From the perspective of visual generation quality, the digital gallery obtains a PSNR of 32.4 dB, a SSIM of 0.936, and a subjective evaluation of 4.71. The PSNR of the projected art space also reaches 31.8 dB, which indicates that the proposed method can better maintain the structural boundaries, color levels and spatial continuity in the art scene. The LPIPS of complex light and shadow scenes increases to 0.158, and the subjective score decreases to 4.18, indicating that the local perception difference of generated images is still obvious when strong light spots, dark background and multi-layer reflections appear at the same time. The results suggest that stronger illumination decomposition, reflection modeling, and dynamic exposure constraints can be introduced in the training phase to improve the adaptability of the model to non-uniform light environments.

In terms of real-time performance, 3D Gaussian representation effectively alleviates the computational pressure caused by traditional NeRF ray-by-ray sampling. At 1080p resolution, the frame rate of low, medium and high complexity scenes reaches 74.6FPS, 68.2FPS and 59.7FPS, respectively, which can meet the real-time display requirements of most immersive art displays. In 4K high-complexity scenes, the frame rate drops to 35.6 FPS, indicating that the high-resolution output is still limited by the number of Gaussian primitivity, transparency blending, and rasterization overhead. The ring graph results show that the time consumption of differentiable rasterization accounts for 34.6%, and Gaussian attribute optimization accounts for 21.3%, which constitute the computing bottleneck of the system. Therefore, cone cropping, importance sampling, hierarchical Gaussian compression, and asynchronous rendering scheduling can be further used in the future to reduce the proportion of invalid primitives participating in the computation.

Interactive experiments show that the multimodal behavior perception mechanism can accurately identify user intentions, and the overall recognition accuracy of five types of tasks reaches 94.5%. The accuracy of content triggering was the highest, 96.4%. The accuracy of parameter adjustment is relatively low (92.7%), and the misjudgment mainly occurs between object selection and parameter adjustment, which indicates that gesture stopping, dragging and fine-tuning actions have similarities in local features. The continuous interactive drift trajectory further shows that the closed-loop feedback can gradually suppress the state offset, the average update error is reduced from 0.46 in the first 10 rounds to 0.18 in the 50th round, and the stability score is increased from 88.6% to 96.1%. On the whole, the proposed method forms a good synergistic relationship among artistic scene generation, real-time rendering and dynamic interaction, but there is still room for further optimization in highly reflective scenes, 4K high-complexity rendering and fine-grained gesture discrimination.

6 Conclusion

Focusing on immersive digital media art scene generation and interaction design, this paper proposes a neural rendering method that integrates neural radiance field, 3D Gaussian representation and user behavior perception. Method We use multi-source RGB images, depth data, camera poses, semantic tags, and interaction logs as input to establish a unified representation of visual texture, spatial geometry, and behavior state. In the reconstruction stage, NeRF is used to complete the continuous space modeling. In the rendering stage, 3D Gaussian primitives and differentiable rasterization are used to improve the real-time display efficiency. The experimental results show that the average error between the digital exhibition hall and the projection art space is 1.28 cm and 1.36 cm, respectively, and the subjective score of the digital exhibition hall reaches 4.71. In 1080p high-complexity scenes, the average response time of a single frame is 18.6 ms, and the average performance is 59.7 FPS. In the

interactive verification, the accuracy of content triggering is 96.4%, the update error is reduced to 0.18 after 50 rounds of interaction, and the stability score is improved to 96.1%. The results show that the proposed method can balance visual quality, rendering efficiency and dynamic interaction stability.

Author's Profile

Di Luo was born in Meishan , Sichuan, China, in 1985.She is a Lecturer in Chongqing Vocational and Technical University of Mechatronics, She received the bachelor's degree from Sichuan Fine Arts Institute, her master's degree from Sichuan Fine Arts Institute. Her research interest include Digital Media Technology, Digital Media Art. E-mail: 13637929791@163.com

References

[1] Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R, Ng R. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis[J]. Communications of the ACM, 2022, 65(1): 99-106. DOI: 10.1145/3503250.

[2] Tewari A, Fried O, Thies J, Sitzmann V, Lombardi S, Sunkavalli K, et al. Advances in Neural Rendering[J]. Computer Graphics Forum, 2022, 41(2): 703-735. DOI: 10.1111/cgf.14507.

[3] Park K, Sinha U, Barron J T, Bouaziz S, Goldman D B, Seitz S M, Martin-Brualla R. HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields[J]. ACM Transactions on Graphics, 2021, 40(6): Article 238, 1-12. DOI: 10.1145/3478513.3480487.

[4] Müller T, Evans A, Schied C, Keller A. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding[J]. ACM Transactions on Graphics, 2022, 41(4): Article 102, 1-15. DOI: 10.1145/3528223.3530127.

[5] Reiser C, Szeliski R, Verbin D, Srinivasan P P, Mildenhall B, Geiger A, Barron J T, Hedman P. MERF: Memory-Efficient Radiance Fields for Real-Time View Synthesis in Unbounded Scenes[J]. ACM Transactions on Graphics, 2023, 42(4): Article 89, 1-12. DOI: 10.1145/3592426.

[6] Kerbl B, Kopanas G, Leimkühler T, Drettakis G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering[J]. ACM Transactions on Graphics, 2023, 42(4): Article 139, 1-14. DOI: 10.1145/3592433.

[7] Moenne-Loccoz N, et al. 3D Gaussian Ray Tracing: Fast Tracing of Particle Scenes[J]. ACM Transactions on Graphics, 2024, 43(6): Article 232, 1-19. DOI: 10.1145/3687934.

[8] Wu T, et al. Recent Advances in 3D Gaussian Splatting[J]. Computational Visual Media, 2024, 10(4): 613-642. DOI: 10.1007/s41095-024-0436-y.

[9] Luo J, Huang T, Wang W, Feng W. A Review of Recent Advances in 3D Gaussian Splatting for Optimization and Reconstruction[J]. Image and Vision Computing, 2024,

151: 105304. DOI: 10.1016/j.imavis.2024.105304.

[10] Fei B, Xu J, Zhang R, Zhou Q, Yang W, He Y. 3D Gaussian Splatting as a New Era: A Survey[J]. IEEE Transactions on Visualization and Computer Graphics, 2025, 31(8): 4429-4449. DOI: 10.1109/TVCG.2024.3397828.

[11] Liao Y, Di Y, Zhou H, Zhu K, Lu M, Duan Q, Liu J. A Survey on Neural Radiance Fields[J]. ACM Computing Surveys, 2025, 58(2): Article 41, 1-33. DOI: 10.1145/3758085.

[12] Šlapak E, Pardo E, Dopiriak M, Maksymyuk T, Gazda J. Neural Radiance Fields in the Industrial and Robotics Domain: Applications, Research Opportunities and Use Cases[J]. Robotics and Computer-Integrated Manufacturing, 2024, 90: 102810. DOI: 10.1016/j.rcim.2024.102810.

[13] Li K, Schmidt S, Rolff T, Bacher R, Leemans W, Steinicke F. Magic NeRF Lens: A Framework for Interactive Focus+Context Exploration of Neural Radiance Fields[J]. Frontiers in Virtual Reality, 2024, 5: 1377245. DOI: 10.3389/frvir.2024.1377245.

[14] Stacchio L, et al. An Ethical Framework for Trustworthy Neural Rendering in Cultural Heritage and Creative Industries[J]. Frontiers in Computer Science, 2024, 6: 1459807. DOI: 10.3389/fcomp.2024.1459807.

[15] Deng N, He Z, Ye J, Duinkharjav B, Chakravarthula P, Yang X, Sun Q. FoV-NeRF: Foveated Neural Radiance Fields for Virtual Reality[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(11): 3854-3864. DOI: 10.1109/TVCG.2022.3203102.

[16] Wang Z, Wu J, Fan R, Ke W, Wang L. VPRF: Visual Perceptual Radiance Fields for Foveated Image Synthesis[J]. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(11): 7183-7192. DOI: 10.1109/TVCG.2024.3456184.

[17] Shi X, Wang L, Liu X, Wu J, Shao Z. Scene-Aware Foveated Neural Radiance Fields[J]. IEEE Transactions on Visualization and Computer Graphics, 2025, 31(9): 5039-5054. DOI: 10.1109/TVCG.2024.3429416.

[18] Li K, Masuda T, Schmidt S, Mori S. Radiance Fields in XR: A Survey on How Radiance Fields Are Envisioned and Addressed for XR Research[J]. IEEE Transactions on Visualization and Computer Graphics, 2025, 31(11): 9709-9719. DOI: 10.1109/TVCG.2025.3616794.

[19] Kleinbeck C, Zhang H, Killeen B D, Roth D, Unberath M. Neural Digital Twins: Reconstructing Complex Medical Environments for Spatial Planning in Virtual Reality[J]. International Journal of Computer Assisted Radiology and Surgery, 2024, 19(7): 1301-1312. DOI: 10.1007/s11548-024-03143-w.

[20] Fabra L, Solanes J E, Muñoz A, Martí-Testón A, Alabau A, Gracia L. Application of Neural Radiance Fields for 3D Model Representation in the Industrial Metaverse[J]. Applied Sciences, 2024, 14(5): 1825. DOI: 10.3390/app14051825.

[21] Atik M E. Comparative Assessment of Neural Radiance Fields and 3D Gaussian

Splatting for Point Cloud Generation from UAV Imagery[J]. Sensors, 2025, 25(10): 2995. DOI: 10.3390/s25102995.

[22] Mills K A, Brown A. Immersive Virtual Reality for Digital Media Making: Transmediation Is Key[J]. Learning, Media and Technology, 2022, 47(2): 179-200. DOI: 10.1080/17439884.2021.1952428.

[23] Paatela-Nieminen M. Remixing Real and Imaginary in Art Education with Fully Immersive Virtual Reality[J]. International Journal of Education Through Art, 2021, 17(3): 415-431. DOI: 10.1386/eta_00077_1.

[24] Serna-Mendiburu G M, Guerra-Tamez C R. Shaping the Future of Creative Education: The Transformative Power of VR in Art and Design Learning[J]. Frontiers in Education, 2024, 9: 1388483. DOI: 10.3389/feduc.2024.1388483.

[25] Çoruh L. Immersive Digital Creation Environments in Design Education[J]. International Journal of Technology and Design Education, 2025, 35: 1841-1870. DOI: 10.1007/s10798-025-09961-6.