Int. J. Simul. Multidisci. Des. Optim. **16**, 14 (2025)
© Y. Qin et al., Published by EDP Sciences, 2025
https://doi.org/10.1051/smdo/2025015

IJSMDO

Available online at:
https://www.ijsmdo.org

Multi-modal Information Learning and Analytics on Cross-Media Data Integration

Research Article

Open Access

Optimization algorithm for 3D image visual communication based on digital image reconstruction

Yunchu Qin[1,*], Fugui Luo[2], Mingzhen Li[3], Peigang Guo[2], and Liangliang Li[2]

[1] School of Big Data and Computer, Hechi University, Yizhou, 546300 Guangxi, China
[2] School of Artificial Intelligence, Nanning Vocational and Technical University, Nanning, 530000 Guangxi, China
[3] College of Software, Henan University of Engineering, Zhengzhou, 450000 Henan, China

Received: 25 June 2025 / Accepted: 18 August 2025

Abstract. Existing three-dimensional (3D) reconstruction methods have deficiencies in image depth estimation accuracy, texture mapping continuity, and illumination consistency, which lead to obstacles such as geometric distortion and texture fracture in the visual communication of 3D images, affecting the efficiency of information acquisition and immersive experience. To solve these problems, this paper proposes a 3D image visual communication optimization algorithm that integrates neural implicit modeling and a multi-scale visual perception mechanism. By jointly encoding the image depth map and Red-Green-Blue (RGB) map into the Neural Radiance Fields(NeRF) voxel hashing network, the continuity of spatial structure expression and the integrity of texture restoration are improved. Structural similarity constraints and perceptual consistency loss functions are introduced to enhance visual stability and subjective quality under different viewing angles, and context completion and detail enhancement of edge texture missing areas are achieved through graph neural networks. User evaluation results show that this method can shorten the average recognition time by up to 33.1% in all target recognition tasks, improve the average subjective immersion score by up to 58.2%, and reduce the Root Mean Square Error (RMSE) of depth reconstruction in occluded areas to 0.164 meters. The Structural Similarity Index Measure(SSIM) of high-frequency texture areas reached 0.871, and the Learned Perceptual Image Patch Similarity(LPIPS) was stable at 0.162 under ±45° viewing angle offset, which effectively improved the image's structural restoration quality and functional visual communication performance, and provided algorithm support for high-precision virtual expression in multiple scenarios.

Keywords: Three-dimensional reconstruction / visual communication optimization / neural radiance fields / depth estimation network / texture enhancement

1 Introduction

Against the background of the rapid development of digital image processing and computer vision, the construction and expression of three-dimensional images have shown broad application prospects in many fields such as virtual reality, augmented reality, medical imaging, industrial inspection, and cultural heritage digitization. As a three-dimensional expression of visual information, three-dimensional images [1,2] carry information on geometric structures and also require highly realistic visual effects in multiple dimensions, such as color, lighting, and texture, to meet the comprehensive needs of realism, interactivity, and immersive experience. Most of the existing mainstream 3D reconstruction methods [3,4] rely on technical means such as multi-view geometric reasoning, structured light scanning, and voxel grid construction [5,6]. Although some achievements have been made in terms of geometric accuracy, they still face great challenges in terms of the quality and expressiveness of visual communication [7,8]. The most prominent difficulties in the current 3D reconstruction process are limited depth estimation accuracy, insufficient texture mapping continuity, and insufficient processing of illumination changes [9,10]. These problems can cause structural deformation, color breakage, and light-dark separation in the reconstruction results, which can increase the user's perceptual burden, hinder spatial understanding and target positioning, and reduce the consistency and immersion of the interactive experience. When faced with sparse image input or complex scenes, the traditional reconstruction method based on geometric modeling [11,12] can seriously reduce the accuracy of the depth map, resulting in distortion or loss

* e-mail: 04004@hcnu.edu.cn

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

of the reconstructed geometric structure. The texture mapping stage often lacks a context-aware mechanism, which leads to problems such as color misalignment, blurred edges, and texture breakage during the mapping process. In addition, in natural lighting environments, traditional methods [13,14] often cannot effectively model lighting changes, and the reconstructed 3D images visually show obvious color inconsistency and lighting discontinuity, which in turn weakens the authenticity and visual communication ability of the images. The combination of these problems reduces the expression efficiency of 3D images and limits their promotion and use in high-demand application scenarios [15,16]. Therefore, there is an urgent need for an optimization algorithm that can further improve the visual quality and communication capabilities while ensuring geometric accuracy, so as to achieve high-quality and high-fidelity reconstruction of 3D visual information [17,18].

To address the above difficulties and improve the expressiveness of three-dimensional images in visual communication, this paper designs and implements a three-dimensional image visual communication optimization algorithm based on digital image reconstruction, and proposes to integrate a multi-scale visual perception mechanism and image structure feature constraints under the neural implicit expression framework, to achieve higher-fidelity image reconstruction and visual expression optimization. In this method, the depth map corresponding to the source image is first extracted through a highly robust single-view depth estimation network, and the Mixed Frequency Data Sampling Regression Models (MiDaS) model is used as the basic architecture to fuse multi-level feature maps to achieve cross-scene depth regression. After the depth map is obtained, a joint representation of the RGB image and the depth map is constructed and input into the neural radiation field model based on the voxel hash grid structure. Multi-resolution position encoding is used for spatial mapping to build a complete three-dimensional implicit expression model. In the reconstruction process, structural similarity constraint loss and perceptual distance loss are introduced to optimize the perceptual consistency between geometry and texture, effectively suppressing texture distortion and color drift problems. The algorithm structure takes the fixation area as the core of visual priority, enhances the texture fidelity and boundary stability of the salient area, and makes the visual presentation of key parts clearer and more continuous during the interaction process. To solve the problem of blurred boundaries and missing local details of reconstructed textures, a texture enhancement module based on a graph neural network is further introduced, which uses the context transfer mechanism of the graph structure between nodes to refine and complete the edge area, and enhance the texture coherence and visual resolution. The method proposed in this paper constructs a technical closed loop of "modeling mechanism-communication optimization" in three aspects: enhancing the continuity of structure restoration through joint modeling of depth map and RGB map, optimizing the consistency and visual stability of texture expression through perceptual consistency loss, and completing the broken areas and edge details through graph structure context modeling to improve expression completeness and interaction fluency. From the user's perspective, the recognition clarity of the image structure and the coherence of the scene perception are improved, and the fluency of information acquisition and the comfort of visual carrying in interactive tasks are optimized. The proposed method provides a systematic solution to the problems of unclear expression and poor realism in the current three-dimensional image visual communication. At the same time, the ability to maintain structural stability and texture coherence in interactive dynamic scenes provides modeling support for the adaptability of visual communication under complex perspective changes and real-time rendering conditions, and has high practical value and research significance.

2 Related works

Three-dimensional reconstruction methods continue to advance the evolution of visual expression quality, and the core challenge focuses on the trade-off between geometric accuracy and structural fidelity. In the field of three-dimensional image reconstruction, researchers have proposed a variety of methods from different perspectives, trying to strike a balance between improving geometric accuracy, texture fidelity, and reconstruction efficiency. In response to the need to reconstruct small and medium-sized objects, Cui et al. [19] systematically sorted out the key technologies for high-precision three-dimensional reconstruction based on line structured light scanning along the technology development path, and focused on the application performance of such methods in actual scenarios, making up for the lack of in-depth discussion of technical principles in previous work. In addition, to meet the needs of rapid assessment of post-disaster scenes, Hong et al. [20] used a deep learning-based Cascade Cost Volume for High-Resolution Multi-View Stereo Network multi-view 3D reconstruction method to reconstruct building damage from drone images and verified its applicability in terms of multi-view consistency and computational efficiency. Although such methods have advantages in global consistency, they lack a structure completion mechanism, and the depth continuity of the boundary area decreases during the reconstruction process. This paper introduces a graph neural structure enhancement module to achieve contextual reasoning completion of local details and alleviate the depth ambiguity problem. In terms of systematic evaluation of methods, Phang et al. [21] conducted a comprehensive comparison of single-image-based deep learning 3D reconstruction methods from encoders, decoders, training mechanisms, to datasets and evaluation indicators, and pointed out the shortcomings of various methods in terms of expressiveness and generalization. In terms of model type comparison, Fu et al. [22] used statistical models, discriminative models, and generative models as objects, and conducted a systematic evaluation on public datasets from multiple dimensions such as input form, matching accuracy, precision, and recall, providing a reference for model selection and actual deployment. In addition, in order to fully cover the current mainstream directions, Lee et al. [23] sorted out the dense

reconstruction technology routes based on geometry, optics, and deep learning, and focused on its adaptability and technical bottlenecks in complex dynamic environments. In summary, the existing 3D image reconstruction methods still have significant deficiencies in multi-view consistency, texture expression integrity, and robustness in extreme scenes [24–26]. Existing methods show edge blurring and structural jumps in complex texture areas. The depth estimation stage does not consider the consistency constraints of multi-scale semantics, making it difficult to ensure the overall continuity and spatial consistency of geometric restoration.

The visual communication quality of three-dimensional images continues to develop under the guidance of multimodal feature modeling and subjective perception, and the research focus has shifted to the expression focus control and communication stability construction of image structure. In the study of 3D image visual communication and perceptual quality optimization, research work in multiple directions is dedicated to improving the subjective quality and experience comfort of 3D image content based on human perception characteristics. In response to the temporal instability problem that often occurs during the playback of 3D synthetic videos, Zhang et al. [27] proposed a convolutional neural network denoising method combined with a perceptual quality measurement mechanism to reduce flicker distortion and improve visual stability from the temporal and spatial dimensions. Considering the accuracy and spatial consistency of structural restoration in three-dimensional scenes, Xia et al. [28] constructed a spatial structure similarity index based on elastic energy modeling to measure spatial distortion in a way that is closer to the laws of perception. At the same time, dynamic indicators were introduced to incorporate visual discomfort into the evaluation system to improve the experience continuity in the virtual environment. In terms of stereoscopic perception, Coskun et al. [29] designed a 3D video experience quality evaluation mechanism that integrates spatial resolution and depth guidance features. By simulating the stereoscopic perception sensitivity of the human eye, it demonstrated strong evaluation stability under different real-time playback conditions. Shen et al. [30] promoted the development of reference-free quality assessment methods based on the study of the human eye's dual-channel visual fusion mechanism. They introduced natural scene statistical features, adaptive dimensionality reduction strategies, and binocular fusion competition mechanisms to achieve accurate judgment of stereo image quality without references. Some generative models do not perform structural weighting on semantically salient areas, resulting in a blurred center and weakened communication effect of focal areas during texture mapping. This paper introduces a structure-aware constraint mechanism to improve the texture fidelity and communication clarity of important areas. At the same time, considering the image modeling accuracy that three-dimensional visual communication relies on, Jiang et al. [31] sorted out the adaptation problems of spherical image-based three-dimensional reconstruction technology in feature matching and dense restoration. They emphasized the potential of this type of technology in expressing complex spatial information and pointed out the technical challenges it faces in maintaining visual consistency. Existing work generally still has insufficient research in visual comfort, multimodal perception feature modeling, and real-time communication consistency in dynamic environments [32–34]. Most current studies do not introduce active perception mechanisms to model multi-scale salient regions, and ignore the weighted regulation of the human eye's perception, focusing on structural expression, resulting in significant distortion of key semantic areas in visual communication, affecting the stability of user experience and information capture efficiency.

3 Methods

3.1 Depth estimation guided by visual communication

To obtain high-quality geometric structure information, the input image is first passed through the MiDaS depth estimation network for single-view depth prediction. The network adopts a hybrid scale transformation structure, integrates multi-scale feature representation, and has the ability to collaboratively model global structure and local details. MiDaS introduces the ResNeXt-101 backbone network in the encoder stage for semantic-aware feature extraction. The output features are passed through the Transformer module for global context modeling, which enhances the depth consistency across regions. Finally, a dense depth map is generated by the multi-resolution decoder. In order to avoid depth estimation offset caused by image texture interference, an edge-preserving depth regularization term is introduced to suppress the interference signal caused by strong gradients in the texture area, and the initial depth result is smoothed and reconstructed through image-guided filtering.

The output depth map of this process adopts an anti-normalization remapping strategy to remap the depth value interval regressed by the network to a distance space consistent with the actual scene structure to ensure the physical consistency of the depth field. The depth map and the original RGB image are spliced according to the channel dimension to generate the input tensor, which constitutes the joint expression basis for the subsequent modeling stage. In order to enhance the boundary saliency of geometric details, a gradient enhancement operator is applied to the depth map to explicitly highlight the structural contour area and improve the geometric response ability of the reconstructed model to the edge position. The three-dimensional spatial coordinates of each pixel in the image are analytically calculated through the projection function. Assuming the camera intrinsic parameter matrix is K, the image coordinates are (u, v), and the depth value is $d(u, v)$, the corresponding three-dimensional coordinates (X, Y, Z) are determined by the following formula:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = d(u,v)\cdot K^{-1}\cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}. \tag{1}$$

This projection model completes the mapping of two-dimensional pixels to three-dimensional space, providing an accurate spatial positioning basis for subsequent NeRF

voxel space encoding. In order to further improve the continuity of the depth map in spatial structure modeling, the context consistency loss function is introduced in the depth prediction stage. By constructing the similarity term of the depth gradient distribution in the local window, the depth jumps in different regions are suppressed. The loss function is defined as follows:

$$\mathcal{L}_{ctx} = \sum_{i \in \Omega} \sum_{j \in \mathcal{N}(i)} w_{ij} \cdot \|D_i - D_j\|^2 \tag{2}$$

where D_i represents the depth value of pixel i, $N(i)$ is its neighborhood, and w_{ij} is a weight function based on RGB space distance and gradient direction, which is used to control the degree of influence between different regions.

The image depth map generation stage provides accurate input for 3D geometric modeling and directly determines the benchmark framework for subsequent texture mapping and lighting modeling. In visual communication tasks, the accuracy of depth information directly affects the quality of NeRF voxel interpolation. Therefore, in the generation stage, it is necessary to ensure the semantic alignment of the depth map and the image content. The multi-channel feature input, composed of the image depth map and the RGB map, can further extract spatial structure information and color texture information in the subsequent multi-scale feature fusion module, building a solid data foundation for high-quality three-dimensional visual expression.

In order to enhance the perception ability of the depth estimation model in the structural boundary area, this method introduces an edge enhancement processing mechanism based on the output of the MiDaS network to strengthen the edge gradient response and improve the model's structural recognition accuracy for boundary mutations. By introducing the context consistency loss function in the depth regression stage, the depth jump problem in high-frequency boundaries and occluded edge areas is effectively alleviated, allowing the model to show strong structural adaptability in dealing with different types of boundary conditions. In areas where structural contours and texture variations coexist, the feature expression obtained through global modeling of the Transformer module improves semantic continuity and enhances the stability of depth regression in complex boundary areas, providing a more reliable geometric prior foundation for subsequent continuous modeling of three-dimensional structures.

3.2 Multi-channel image feature fusion

In order to achieve the coordinated encoding of image color information and spatial structure information, the RGB image and its corresponding depth map are spliced in the channel dimension to construct a fusion tensor. The original image size is set to $\mathrm{I} \in \mathbb{R}^{H \times W \times 3}$, and the dense depth map $D \in \mathbb{R}^{H \times W \times 1}$ is obtained by the MiDaS model, which is then spliced to form a four-channel input $F_0 = \mathrm{concat}(I, D) \in \mathbb{R}^{H \times W \times 4}$. In order to obtain local structures and global perception features across scales, three groups of convolution operations with different receptive fields are applied to F_0 in turn. The convolution kernel sizes are 3×3, 5×5, and 7×7, respectively. The boundary size is kept unchanged, and the fused feature tensor $F_m \in \mathbb{R}^{H \times W \times C}$ is obtained by processing with a unified number of channels. The weight of the convolution module is denoted as θ_c, and the overall mapping process can be expressed as:

$$F_m = \phi(F_0; \theta_c) \tag{3}$$

where ϕ represents the composite mapping function of multi-scale convolution and feature compression. In order to enhance the response strength of the structural area and the selectivity of the semantic layer features, the channel attention and spatial attention modules are further introduced to model the cross-channel feature importance and spatial position response weight respectively. In the channel attention path, global average pooling and maximum pooling are first performed on F_m to generate two one-dimensional channel description vectors, which are transformed and merged through a shared fully connected network to obtain a channel weight map $M_c \in \mathbb{R}^{1 \times 1 \times C}$, and a channel-by-channel weighted operation is performed to obtain the output feature $F_c = F_m \otimes M_c$. In the spatial attention path, averaging and maximum pooling are first performed along the channel dimension, and then concatenated and convolved to generate a spatial weight map $M_s \in \mathbb{R}^{H \times W \times 1}$, which is then multiplied element-wise with F_c to obtain the fused output:

$$F_{cs} = (F_m \otimes M_c) \otimes M_s. \tag{4}$$

In order to improve the ability of fusion features to maintain the edge of the structure, the gradient consistency loss constraint is introduced to make the RGB channel and the depth channel maintain direction consistency in the local change trend. The gradient consistency loss function is defined as:

$$\mathcal{L}_{\mathrm{grad}} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} (|\nabla_x I_{i,j} - \nabla_x D_{i,j}| + |\nabla_y I_{i,j} - \nabla_y D_{i,j}|) \tag{5}$$

∇_x and ∇_y are the gradient operators in the horizontal and vertical directions, respectively, and $I_{i,j}$ and $D_{i,j}$ are the grayscale values of the RGB image and the depth map at pixel (i, j), respectively.

Figure 1 shows the overall process of multi-channel image feature fusion, including the complete path from concatenating the input image and the depth map, extracting structural semantic features through multi-scale convolution, and then weighted fusion through the attention module. The dimensional changes of each intermediate tensor are marked, which helps to clearly understand the structural position and role of this module in the entire algorithm process.

To regulate the contribution of features at different receptive fields, a learnable weight allocation mechanism is embedded into the multi-scale convolution module. Each of the three convolution branches with kernel sizes of 3×3, 5×5, and 7×7 is assigned an independent trainable weight

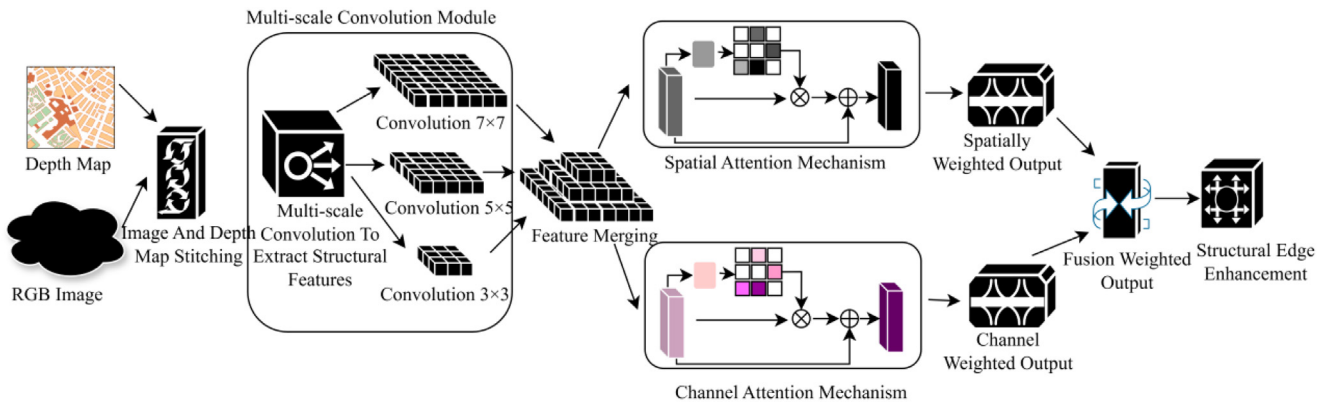

Fig. 1. Multi-channel perception-driven image feature fusion structure.

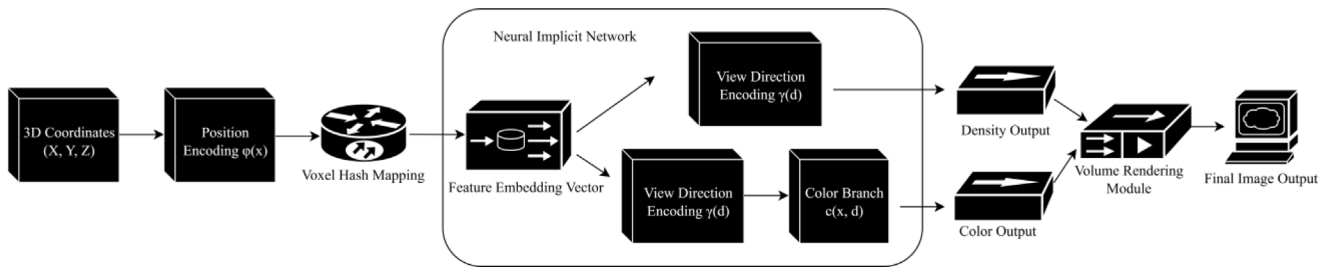

Fig. 2. Schematic diagram of the neural implicit modeling structure based on voxel hashing.

scalar, which adjusts the relative response strength of local texture, mid-range semantic structure, and global contextual information. These weights are jointly optimized with the network parameters through backpropagation, initialized with uniform values and updated dynamically during training. The fusion process applies weighted summation across the feature maps before entering the attention modules, allowing the network to adaptively prioritize salient features from each scale level depending on the structural complexity of the input region. This design aims to enhance the granularity of local detail modeling while maintaining holistic perceptual consistency across scales.

3.3 Perception-driven neural implicit modeling

In order to improve the geometric continuity and texture expression accuracy of three-dimensional images, an improved NeRF voxel hashing network is constructed as the core structure of neural implicit modeling based on the joint features of depth and image. The voxel hashing mechanism is used to discretize the three-dimensional space, and the scene space is divided into sparse voxel grids to minimize memory redundancy and computational redundancy, thereby improving the efficiency of high-resolution modeling. Each voxel position is input through three-dimensional coordinates, and after hash mapping, it is connected to the multi-layer perceptron network to express the RGB color field and density field in the form of implicit functions. The position embedding of spatial points adopts the position encoding strategy, and the high-frequency details are enhanced by mapping the input coordinates $\mathbf{x}\in\mathbb{R}^3$ into the frequency domain. The formula is:

$$\gamma(\mathbf{x}) = [\sin(2^0\pi\mathbf{x}), \cos(2^0\pi\mathbf{x}), \ldots, \sin(2^{L-1}\pi\mathbf{x}), \cos(2^{L-1}\pi\mathbf{x})] \tag{6}$$

where L represents the number of frequency levels. For hash coding, a multi-resolution voxel grid is adopted, and a spatial hashing strategy based on Morton coding is used to map the high-dimensional space to a low-dimensional hash table, which greatly reduces memory consumption while retaining positional distinctiveness. During network training, the distribution of active voxels is dynamically adjusted through the voxel hierarchical update mechanism to improve the modeling resolution of texture edges and depth boundary areas.

Figure 2 shows the voxel hashing-based spatial coding structure and its coupling with the NeRF implicit network, which characterizes the mapping path between sparse coding and full-space modeling, and effectively supports the efficient expression of complex scene information during the reconstruction process.

In order to achieve coupled reconstruction of texture and geometry, a dual-branch network structure is designed, one network is responsible for predicting density $\sigma(x)$, and the other network predicts color $c(x,d)$, where d is the viewing direction. The color prediction network integrates perspective encoding to improve texture coherence under perspective dependence. After introducing the

direction-aware feature $\gamma(d)$ in color estimation, the overall color estimation process of the network can be described as:

$$\mathbf{c} = \mathcal{F}_c\Big(\gamma(\mathbf{x}), \gamma(\mathbf{d}), \mathbf{f}_x\Big) \tag{7}$$

where F_c represents the color prediction subnetwork, and f_x is the spatial feature encoding. The density function is output by another network F_σ, whose input includes the embedded feature $\gamma(x)$ of the spatial point and the spatial feature encoding f_x, reflecting the spatial occupancy probability of the voxel, which is independent of the observation direction. Its formula is:

$$\sigma = \mathcal{F}_\sigma\Big(\gamma(\mathbf{x}), \mathbf{f}_x\Big). \tag{8}$$

The density function $\sigma(x)$ and the color function $c(x,d)$ are used together for the volume rendering function:

$$\hat{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma\Big(\mathbf{r}(t)\Big)\mathbf{c}\Big(\mathbf{r}(t), \mathbf{d}\Big)dt \tag{9}$$

where $r(t)$ represents the light from the viewpoint, and $T(t)$ represents the transmittance function, which reflects the probability that the light from the camera to the depth t is not blocked.

To improve the subjective visual quality and perceptual stability of 3D images in the multi-view reconstruction process, this section introduces a perceptual consistency optimization mechanism, which realizes the perceptual guidance of depth reconstruction and texture mapping by constructing a joint loss function, and optimizes the performance of the results in detail preservation, structure alignment, and texture consistency. The perceptual consistency constraint is based on the sensitivity of the human visual system to structural and texture changes during image reconstruction. It constrains the network output from two dimensions: structure preservation and perceptual distance, to avoid the problem that the traditional L2 loss cannot capture high-dimensional perceptual errors.

During the training process, a structural similarity loss is introduced between each pair of reconstructed images and the corresponding target image, and the brightness, contrast, and structural similarity between images are measured based on local statistics. This part uses SSIM as the optimization term, and its mathematical definition is as follows:

$$\mathcal{L}_{SSIM} = 1 - \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{10}$$

where μ_x and μ_y represent the local means of the reconstructed image and the target image, μ_x^2 and μ_y^2 represent the local variance, σ_{xy} is the covariance, and the constants c_1 and c_2 are used to avoid the denominator being zero.

On this basis, LPIPS perceptual loss is further introduced to quantify the similarity of images in the deep feature space. This loss uses a pre-trained convolutional network to extract high-level semantic features and reflects the perceptual difference through the L2 distance between features. It is defined as follows:

$$\mathcal{L}_{LPIPS} = \sum_l w_l \|\phi_l(I_{pred}) - \phi_l(I_{gt})\|_2^2 \tag{11}$$

where ϕ_l represents the feature map extracted at the lth layer, w_l is the weight coefficient of the layer, I_{pred} and I_{gt} are the predicted image and the target image, respectively.

The image quality is limited by the illumination changes and occlusion interference between different perspectives. In order to further improve the stability of the reconstructed image under multi-view conditions, a multi-view consistency loss term is designed to impose consistency constraints through the projection relationship between corresponding pixels under different perspectives. Suppose there are multiple view images $\{I_i\}_{i=1}^N$, and the corresponding reconstructed depth map is $\{D_i\}_{i=1}^N$. By knowing the camera external parameter and the internal parameter matrix K, the pixel (u, v) under the i-th view can be mapped to the projection point (u', v') of the j-th view, and the reprojection error can be calculated:

$$\mathcal{L}_{cons} = \sum_{i \neq j} \|I_i(u,v) - \hat{I}_j(u',v')\|_1 \tag{12}$$

where $\hat{I}_j$ is the image sampling value after the pixel of the i-th perspective is reprojected to the j-th perspective according to the depth information. The reprojection operation uses bilinear interpolation for sub-pixel level mapping.

To ensure the convergence of the overall optimization and the synergy of perceptual constraints, a joint loss function is constructed as the training objective:

$$\mathcal{L}_{total} = \lambda_{ssim}\mathcal{L}_{SSIM} + \lambda_{lpips}\mathcal{L}_{LPIPS} + \lambda_{cons}\mathcal{L}_{cons} \tag{13}$$

where λ_{ssim}, λ_{lpips}, and λ_{cons} are the weight coefficients of each loss item, which are adjusted according to the degree of perception dominance.

3.4 Semantic-guided texture enhancement

In order to solve the problem of local texture breakage and edge blur in 3D reconstructed images, this paper introduces a texture detail enhancement module based on a graph neural network. This module is based on the fused image feature tensor and constructs a graph structure to model the relationship between local and global pixels in the texture space. The specific composition method adopts the pixel block division strategy to divide each image area into a set of equally spaced graph nodes. The connection between nodes is determined based on the Euclidean distance in the pixel space and the color similarity, and the adjacency matrix $A \in \mathbb{R}^{N \times N}$ is constructed, where N is the number of graph nodes. The adjacency matrix is calculated by the Gaussian kernel function, and the formula is:

$$A_{ij} = \exp\left(-\frac{\|x_i - x_j\|_2^2}{\sigma_x^2} - \frac{\|f_i - f_j\|_2^2}{\sigma_f^2}\right) \tag{14}$$

where x_i and x_j are the spatial positions of nodes i and j in the image, f_i and f_j are their corresponding color features, and σ_x and σ_f are the normalization factors of the position and color space.

In the graph neural network modeling stage, this paper uses a multi-layer graph convolutional network to perform feature propagation and context fusion on the constructed graph structure. The basic graph convolution operations used are as follows:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right) \tag{15}$$

where $H^{(l)}$ represents the node feature representation of the lth layer, $\tilde{A} = A + I$ is the adjacency matrix after adding self-loops, $\tilde{D}$ is the corresponding degree matrix, $W^{(l)}$ is the learnable weight parameter, and σ is the nonlinear activation function.

In order to improve the structural recovery ability of texture detail areas, the graph neural network module introduces multi-structure mapping methods and different types of graph convolution operators when constructing it to establish a more sophisticated context modeling mechanism. The graph structure is constructed based on the composite relationship between image pixel features and spatial positions. Three connection strategies are adopted: fixed neighborhood pixel map, regional aggregation map, and adaptive edge weight map based on similarity learning, which respectively reflect the isometric, semantic block-level, and perception-driven texture dependency modeling characteristics. The topological structure of the graph is constructed by calculating the color difference and spatial coordinate distance between nodes to form a weighted adjacency matrix. The weight function adjusts the similarity distribution in the high-dimensional embedding space to form a differential representation of the texture relationship.

In the construction of the graph convolution module, the design considers the balance between propagation depth and expression ability, and controls the coverage and detail retention of the feature receptive field by setting different numbers of convolution layers. The shallow structure emphasizes local continuity modeling, and the deep structure enhances cross-region dependency capture. In the overly deep structure, feature smoothing is prone to occur, which affects the clarity of edge response. The graph convolution kernel function is designed using a variety of heterogeneous schemes. By comparing and analyzing standard graph convolution, neighborhood aggregation graph convolution, and weighted attention mechanism, the response characteristics of various mechanisms in boundary recognition and context coupling are explored. Among them, the mechanism based on inter-node attention learning enhances the weight distinction ability of significant edge areas in the graph and improves the consistency of texture transition in structural mutation areas. In the overall module design, a two-layer graph convolution structure with adaptive edge weight expression and learnable attention weight is finally adopted to meet the needs of detail recovery and structure-guided expression in texture-damaged areas. This structure has strong edge response ability and context integration efficiency, and has good training convergence stability and adaptation performance, which meets the consistency requirements of texture enhancement in multiple types of scenes.

Fig. 3. Texture enhancement and structural refinement results produced by the graph neural network module.

In order to further enhance the edge structure clarity of the reconstructed image, a gradient-guided attention mechanism is added to modulate the edge sensitivity of the graph convolution output. The specific method is to calculate the image grayscale gradient map G through the Sobel operator, introduce the weight modulation function $M = \tanh(\alpha G)$, and apply it to the graph convolution output feature map H to form the edge enhancement representation $H' = M \odot H$. The enhanced texture features output by the graph neural module can be fused with the initial NeRF output results to construct the final texture image with complete details.

Figure 3 illustrates the enhancement process of local structural regions using the proposed graph neural network. The upper part shows reconstructed textures before and after enhancement in edge and occluded areas, while the lower part presents the corresponding mesh topology with visible improvements in triangle density and continuity near geometric discontinuities. The fusion of graph-based context modeling and perceptual modulation effectively restores missing textures and refines boundary structures.

Each pixel of the final output texture map is determined by the fusion of three-way information, namely the weighted synthesis of the NeRF network output value, the graph neural network enhancement result, and the depth-guided weight map. The fusion method adopts a dynamic weighted average. The weight vector $\omega = [\omega_1, \omega_2, \omega_3]$ is generated by the local texture gradient and pixel confidence estimation module to meet the

normalization constraint $\omega_1 + \omega_2 + \omega_3 = 1$. The optimization goal is to minimize the perceptual difference between the enhanced texture and the real texture. The loss function is as follows:

$$\mathcal{L}_{\text{fuse}} = \lambda_1 \cdot \text{SSIM}(I_{\text{fuse}}, I_{\text{gt}}) + \lambda_2 \cdot \text{LPIPS}(I_{\text{fuse}}, I_{\text{gt}}) \quad (16)$$

where I_{fuse} is the fused image, I_{gt} is the reference image, λ_1 and λ_2 are loss weight coefficients, which control the optimization weights of structural similarity and perceptual consistency.

4 Experiment

4.1 Experimental environment and platform configuration

To ensure that the implementation process of the 3D image visual communication optimization algorithm has good operating efficiency and repeatability, the experiment was completed on a high-performance computing platform, combining modern deep learning frameworks and graphics processing units to support the model training and reasoning process. In terms of hardware environment, the experimental platform is equipped with an NVIDIA RTX 4090 graphics card with a memory capacity of 24GB, which supports highly parallel tensor computing tasks and is suitable for the training of large-scale voxel networks and graph neural structures. The Central Processing Unit (CPU) uses an Intel Core i9-13900K processor with a main frequency of 3.0GHz and multi-threaded computing capabilities, providing stable computing power support for data preprocessing and feature extraction. The memory configuration is 128GB to ensure data reading and writing efficiency during multitasking.

In terms of software environment configuration, the operating system uses Ubuntu 22.04, which has stable driver support and deep learning dependent environment compatibility. The main programming language is Python 3.10, the deep learning framework is PyTorch 2.1, and Compute Unified Device Architecture 12.1 (CUDA 12.1) and CUDA Deep Neural Network library 8.9 (cuDNN 8.9) are used to accelerate neural network calculations. OpenCV 4.7 and Albumentations libraries are used for data processing and image enhancement to complete image standardization and cropping operations. During the construction of the voxel hash coding module, the Tiny CUDA Neural Networks (tiny-cuda-nn) and Instant Neural Graphics Primitives (instant-ngp) toolsets are used to achieve high-speed voxel query and implicit function training. The graph neural network part relies on the PyTorch Geometric framework, version 2.4, and cooperates with Deep Graph Library (DGL) for graph structure modeling and node relationship construction.

In order to achieve multi-module joint training, the experimental process adopts a modular training strategy, integrating the MiDaS depth estimation network, voxel hash implicit modeling network, and graph neural texture enhancement network into a unified training framework. Each module is initialized with the same random seed to ensure the reproducibility of the training process. In order to optimize the memory usage and improve the training efficiency, mixed precision training is enabled during the training process, and the Automatic Mixed Precision (AMP) automatic mixed precision strategy is used for floating point control of forward and backward propagation. The model parameters are saved using the PyTorch Checkpoint mechanism, and the intermediate model is saved regularly according to the training rounds and verification performance, which is convenient for subsequent comparison and performance review.

4.2 Dataset and preprocessing process

This paper uses a public 3D visual reconstruction dataset as the data basis for training and evaluation, covering indoor and outdoor scene images taken from multiple angles. Each scene in the dataset provides high-resolution RGB images, depth maps, and camera pose information. The resolution of the RGB images used is unified to 768×512 pixels, and the number of viewing angles is distributed between 6 and 20 frames per scene. The depth map is given in meters, and its correspondence has been spatially aligned using calibrated camera parameters. The original image data has problems such as inconsistent color style, obvious differences in brightness distribution, blurred edges, and uneven occlusion areas. In order to ensure that the network training converges stably and the output results have a uniform scale, the input image and its corresponding depth map are systematically preprocessed.

The original RGB image is uniformly processed by global color mapping through a linear stretching operation to enhance the overall contrast and correct the hue shift. The image is then smoothed by edge preservation using bilateral filtering to suppress noise interference while retaining structural information. In the depth map processing process, invalid values and hole areas are masked and filled with inverse distance weighted interpolation to avoid structural distortion caused by sparse areas. The RGB image and the depth map are spatially remapped and aligned through the external parameter matrix to ensure the one-to-one correspondence between pixels and uniformly construct the channel input tensor.

After the completion of the primary image preprocessing steps, two critical strategies were applied to optimize model performance: color mapping and geometric enhancement. The color mapping process involved a global linear stretching operation to adjust the image hue and contrast, addressing lighting inconsistencies and improving texture alignment. This step ensured uniformity across input images, which directly influenced the model's ability to preserve texture integrity in different scenes.

Geometric enhancement aimed to improve the accuracy of the depth maps by mitigating structural distortions. An edge-preserving depth regularization method was applied to smooth the depth map, particularly in regions with sharp gradients, thereby improving the model's capacity to capture fine geometric details. In addition, inverse distance-weighted interpolation was employed to address gaps in the depth map and ensure precise spatial alignment between the RGB images and corresponding depth data. These preprocessing strategies contributed to the overall

Table 1. Statistics of the multi-scene dataset.

Scene ID	Number of images (frames)	Average number of views	Depth map completeness (%)	Image resolution (pixels)
S001	128	10	92.4	768×512
S002	96	8	89.7	768×512
S003	144	12	93.5	768×512
S004	72	6	85.2	768×512
S005	120	10	91.1	768×512

performance by enhancing both visual consistency and geometric accuracy, enabling the model to generate high-quality 3D reconstructions.

In order to enhance the generalization ability of the model, an image enhancement strategy is introduced in the training stage. Before the image is input, brightness perturbation, color temperature transformation, and Gaussian blur processing can be applied probabilistically to simulate the input changes under different lighting and imaging conditions. At the same time, geometric enhancement methods such as random cropping and horizontal flipping are used to expand the scene perspective and increase the spatial diversity of training samples. All image enhancement processes are performed online during the data loading phase to ensure that the input image samples of each epoch are different. The image tensor is normalized before entering the neural network to make the RGB value distribution uniform to zero mean unit variance, improving the network's robustness to color changes.

To more clearly show the basic composition, image size, and multi-view coverage of the dataset, Table 1 gives the statistical information, such as the number of images, average number of views, and depth map completeness of various scenes in the training and testing stages.

After image preprocessing, all samples are randomly divided into training set and test set according to the scene division method, with a ratio of 8:2. The training set data is used for joint optimization learning of each module of the neural network, and the test set is used as the basis for evaluating the generalization performance of the model, running through multiple evaluation links such as geometric structure accuracy, texture restoration ability and multi-view consistency, to ensure that the entire method is reproducible and engineering deployable under a unified data processing process.

4.3 Parameter setting and training strategy

To ensure that the 3D image visual communication optimization algorithm converges stably during the training process and achieves the expected reconstruction quality, the parameter setting and training strategy must be strictly designed in a unified manner based on the functional characteristics and convergence behavior of each module. In the neural implicit modeling part, the NeRF voxel hashing network uses the Adam optimizer for weight update, and the initial learning rate is set to 0.001, combined with a learning rate decay strategy. When the number of training rounds reaches half, the learning rate is linearly reduced to 0.0001 to enhance the fineness of the model's late convergence. The batch size is set to 16 to balance the memory usage and training stability. The number of training rounds is fixed to 1200 rounds to ensure that the model fully learns the geometric and texture distribution relationship in the scene in the multi-scale feature space. The loss function is designed in the form of a multi-weighted combination, and the overall loss is composed of deep reconstruction error, structural similarity loss, and perceptual consistency loss. Among them, the depth reconstruction error uses the L1 norm to constrain the pixel-by-pixel difference between the predicted depth map and the input MiDaS depth map, and the weight is set to 0.5. The structural similarity loss uses the SSIM indicator to guide the local reconstruction quality in the image structure space, with a weight of 0.3. The perceptual consistency part introduces the LPIPS loss to measure the perceptual difference between the reconstructed image and the original image from the feature space, with a weight of 0.2. The loss weight coefficient remains unchanged throughout the training process to ensure that various error indicators guide the training process in a balanced manner.

The graph neural network structure used in the texture detail enhancement stage adopts two layers of graph convolutional neural units, with 64 and 128 output channels in each layer. The node feature dimension is initialized to 256, and the ReLU activation function is used. It is updated synchronously with the main network during training. The graph structure is constructed based on the pixel adjacency relationship, and the edge weight is jointly determined by the color difference and spatial position. The connectivity of low-contrast boundaries is enhanced through edge normalization operations. In order to adapt to the computational overhead of graph convolution for large-scale image processing, the texture enhancement images were uniformly cropped into 256 × 256 local areas for training.

During the training process, all images were uniformly standardized before input, with the mean and standard deviation set to 0.5 and 0.5, respectively, and normalized to the [–1,1] interval to improve the stability of gradient propagation. Data loading is done in a multi-threaded parallel mode. The training process is performed on the NVIDIA RTX 4090 Graphics Processing Unit (GPU) platform. PyTorch 2.1 is used to build the entire neural network structure. Mixed precision calculation is used during training to improve efficiency and reduce video memory pressure.

4.4 User behavior experiment

In order to systematically evaluate the actual performance of the proposed 3D image visual communication optimization algorithm at the user perception level, a user behavior experiment combining subjective scoring and task execution was constructed. The experimental goal is to establish a corresponding mechanism between perception dimensions and interactive performance, to quantify the impact of different 3D reconstruction results on visual experience and operational efficiency, and then verify the effectiveness and applicability of the model at the human factor adaptation level.

The experiment selected subjects with a basic understanding of images to ensure that the scoring judgment and operational behavior have controllable consistency. The experimental scene covers multiple typical 3D reconstruction results. The images are all from the same shooting conditions and scene space, and are uniformly processed by scale normalization, lighting calibration, and color mapping to ensure the alignment of different reconstructed images in objective content and eliminate interference from non-reconstructed factors.

To examine the user's visual adaptation ability under dynamic changing conditions, this experiment introduces a dynamic scene perturbation mechanism in the occluded target positioning task and the multi-view consistency judgment task, and performs slight scale transformation, occlusion simulation, and gradual perspective shift operations on the target image, thereby constructing observation conditions that approximate dynamic interaction. In the specific design, three levels of perturbation are set by applying rotation, scaling, or partial occlusion that changes continuously between frames to the target area to simulate the visual instability caused by viewpoint changes or object occlusion in real scenes.

In the behavioral operation stage, an image target recognition task is set. After the system prompts the target area, visual positioning is performed, and the time required to complete the positioning task and the pixel-level error offset value are collected. The task interface records click events and response times in real time. All operation instructions are trained and familiarized before the formal experiment to avoid data bias caused by differences in operating experience. All image tasks are carried out under the same content density and target size conditions to ensure the structural consistency and comparability of the behavioral tasks.

The experimental data is automatically recorded by the system to generate the original scoring matrix and behavioral response log, and the sampling frequency is controlled at the millisecond level to ensure that the behavioral time data has precision support. The subjective perception indicators and behavioral task data do not interfere with each other in structure. The corresponding images and users are identified by numbers to achieve accurate matching and analysis based on multi-dimensional experimental data. The experimental data is standardized and sorted for subsequent result analysis, which further reveals the user adaptation efficiency and perceptual performance advantages of the reconstruction method in visual communication.

5 Result analysis

5.1 3D reconstruction geometric accuracy

To verify the depth restoration capabilities of various 3D image reconstruction algorithms in different geometric structure areas, this experiment compares the performance of five representative depth modeling methods in four typical spatial regions. Method 1: COLMAP is based on sparse point cloud matching, and its accuracy is limited by feature density and viewing angle baseline; Method 2: Dense Prediction Transformer (MiDaS-DPT) relies on a large-scale monocular deep pre-training model, has good global consistency but lacks local geometric accuracy; Method 3: NeRF-Baseline uses a volume rendering method of ray rendering, which has continuity in structural recovery but still lacks texture and boundary areas; Method 4: Instant-NGP uses multi-resolution hash coding to accelerate the reconstruction process, but the high-frequency geometric details are slightly weakened. Method 5: Graph Neural Network-NeRF (GNN-NeRF) is a method proposed in this paper, which combines voxel hashing implicit modeling with the contextual structure enhancement mechanism of graph neural networks, and demonstrates stronger robustness in reconstructing complex boundaries and occluded areas. The region division covers typical spatial structures such as planes, edges, occlusions, and long distances, and can reflect the differences in the model in dimensions such as structural uniformity, depth continuity, and occlusion processing. In terms of evaluation indicators, Root Mean Square Error (RMSE) is used to measure the overall magnitude of geometric deviation and is more sensitive to extreme errors. Absolute Relative Error (Abs Rel) reflects the normalized ratio between the prediction error and the true value, reflecting the relative accuracy. Mean Absolute Error (MAE) measures the average level of the overall prediction deviation to avoid being dominated by extreme values. The three constitute a complementary error evaluation system. The three bar graphs in Figure 4 present the performance comparison of the three indicators in each area, showing the error distribution and stability of each method.

The GNN-NeRF method has the lowest RMSE value in all four types of areas. The RMSE in the occluded area is 0.164 meters, which is 0.043 meters lower than that of NeRF-Baseline, showing better geometric reconstruction ability in areas with missing information. This advantage can be attributed to the role of its graph structure perception mechanism in complementing contextual relationships. In the Abs Rel indicator, the error of this method in the edge area is 0.103, which is significantly lower than MiDaS-DPT's 0.151, indicating that this method is more accurate in depth ratio recovery and avoids the problem of over-smoothing of monocular depth models in texture change areas. In terms of MAE, GNN-NeRF maintains 0.119 meters in the long-distance area, which is lower than Instant-NGP's 0.150 meters, verifying its modeling effect on far-field depth continuity. Overall, GNN-NeRF's modeling capabilities in multi-scale structures and high-frequency details effectively suppress reconstruction errors, providing a more stable geometric basis for subsequent 3D visual expression.

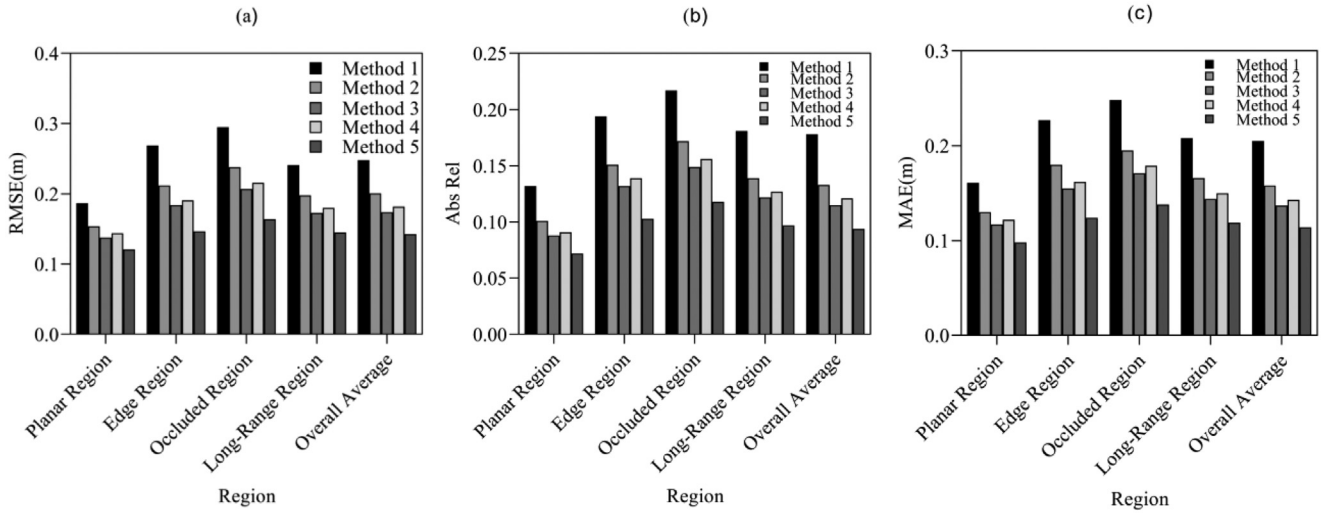

Fig. 4. Error comparison analysis of five depth reconstruction methods in typical structural areas. (a) RMSE error comparison in different areas. (b) Abs Rel error comparison in different areas. (c) MAE error comparison in different areas.

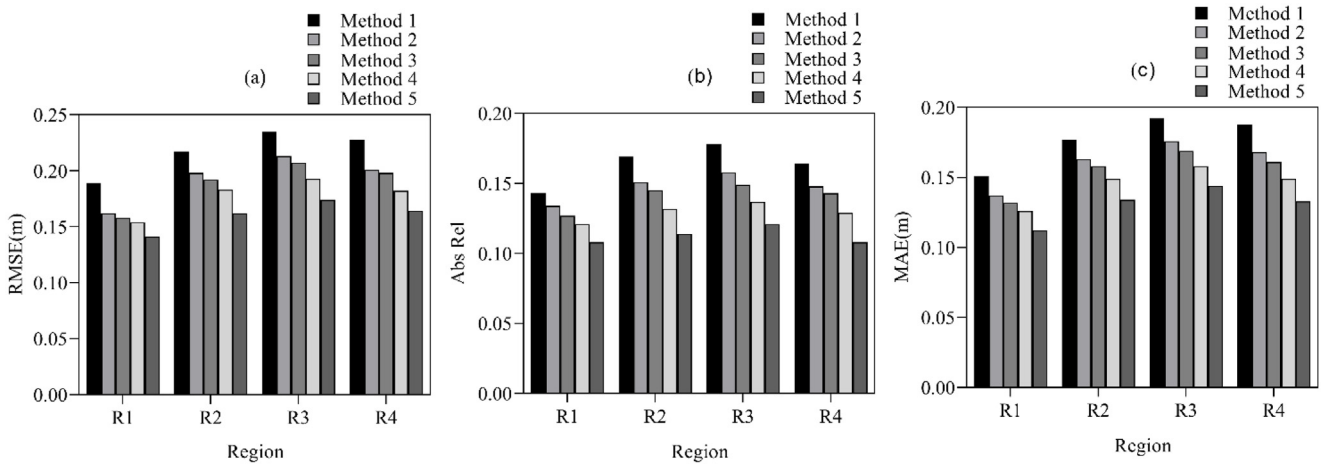

Fig. 5. Depth estimation error comparison of different 3D reconstruction methods in complex boundary areas (a) RMSE error comparison. (b) Abs Rel error comparison. (c) MAE error comparison.

To deeply characterize the geometric modeling performance of different algorithms in complex boundary areas, a detailed evaluation set of four types of boundaries is further constructed based on the above methods 1 to 5, covering four types of scenes with typical local geometric challenges: smooth contours (R1), high-frequency texture breaks (R2), depth mutation areas (R3), and occluded boundaries (R4). The method system represents five mainstream modeling mechanisms: sparse point cloud matching, monocular depth pre-training, volume rendering baseline, multi-resolution hash coding, and graph structure context enhancement, with different boundary processing capabilities and modeling expression strategies. Each boundary type presents spatial characteristics such as strong texture continuity, obvious structural mutation, and severe occlusion faults. The test data extracts representative boundary fragments from the same scene and conducts comparative analysis of local depth errors to evaluate the response accuracy and stability of the algorithm under non-uniform structural conditions. The performance differences of each method in terms of RMSE, Abs Rel, and MAE are shown in Figure 5.

In the RMSE error dimension, GNN-NeRF always maintains the lowest error value in the four types of regions. The RMSE in the occluded boundary area is 0.164 meters, which is significantly lower than 0.198 meters of NeRF-Baseline and 0.201 meters of MiDaS-DPT. The error compression is mainly attributed to its context graph structure to complete the missing area and the boundary continuity modeling advantage. In terms of Abs Rel, this method reaches 0.114 in the high-frequency edge area, which is better than 0.132 of Instant-NGP and 0.169 of COLMAP, showing better relative depth prediction stability, reflecting its encoding mechanism with stronger robustness to gradient fluctuations caused by texture fracture. The MAE indicator further verifies this trend. In the depth mutation area, the error of this method is 0.144 meters, which is significantly lower than 0.192 meters of

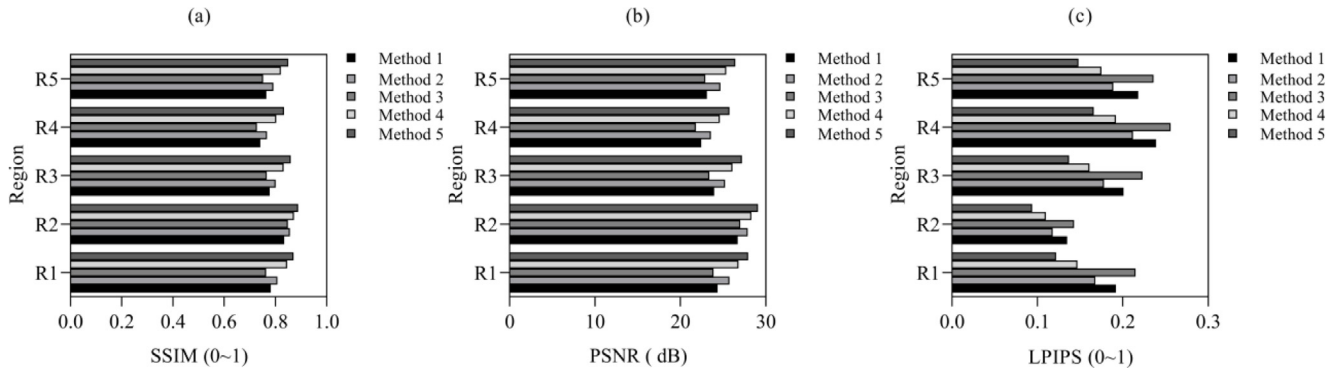

Fig. 6. Comparison of subjective and objective indicators of texture restoration performance of multiple methods in typical areas. (a) Structural similarity evaluation results of each method in typical areas. (b) Peak signal-to-noise ratio evaluation results of each method in typical areas. (c) Perceptual distance evaluation results of each method in typical areas.

COLMAP, reflecting its smoother modeling response to geometric jump structures in high gradient areas. The overall results show that the joint modeling method with enhanced graph structure has higher geometric expression fidelity and structural error suppression capabilities when facing discontinuous boundary areas.

5.2 Texture clarity and detail restoration effect

To comprehensively evaluate the performance of different 3D reconstruction methods in texture restoration, this experiment introduced three subjective and objective indicators: SSIM, Peak Signal-to-Noise Ratio (PSNR), and LPIPS, and quantitatively evaluated them from the perspectives of structural fidelity, image clarity, and perceptual consistency. The experiment selected five typical regions, including high-frequency texture (R1), low-frequency smoothness (R2), edge transition (R3), occlusion fracture (R4) and illumination change (R5), which respectively represent visual difficulties such as texture complexity, surface uniformity, geometric transition, depth continuity and radiation robustness, and conducted a fine-grained comparison of the regional adaptability of the reconstruction results. Method 1: NeRF is based on volume rendering, emphasizing implicit expression but lacking local texture detail modeling; Method 2: Multi-View Stereo Network (MVSNet) adopts a multi-view stereo strategy, relying on geometric consistency but suffering from information compression loss in texture preservation; Method 3: COLMAP combines sparse reconstruction with explicit texture mapping, with a clear overall structure but facing the problem of incomplete occlusion; Method 4: TensoRF introduces tensor decomposition to achieve a balance between lightweight structure and expression efficiency; Method 5 is the method in this paper, which integrates depth perception and graph structure detail recovery mechanism based on neural implicit modeling. Figure 6 shows the performance of each method under three indicators for five types of regions.

In Figure 6, in the high-frequency texture area, the SSIM of this method is 0.871, significantly higher than NeRF's 0.782 and MVSNet's 0.808. The PSNR also reaches 27.94dB, far exceeding other methods, indicating that this model has a stronger ability to preserve local features when processing areas with complex details, thanks to its multi-scale convolution fusion and structure-aware optimization mechanism. In occluded and broken areas, the performance of traditional methods is generally weak, with most SSIM values below 0.77, while the proposed method remains at 0.834 and the PSNR reaches 25.74dB, indicating that its graph neural network's ability to complete context in discontinuous areas improves the overall structural continuity. In terms of the perceptual index LPIPS, the method in this paper has the lowest score in all areas, with a score of 0.148 in the illumination change area, which is lower than 0.175 of TensoRF and 0.218 of NeRF, indicating that its texture restoration under the guidance of perception is more consistent with human visual consistency. Combined with the performance of the above indicators, the method in this paper shows strong texture expression ability and structural restoration robustness in different areas, effectively alleviating the trade-off problem between detail integrity and subjective quality in existing methods.

Figure 7 shows the reconstruction results under four viewpoints, where the upper row displays the original RGB images and the lower row presents the corresponding reconstructed outputs generated by the proposed algorithm. The restored geometry maintains structural consistency with the input while preserving high-frequency details such as edge contours and surface texture. The mesh density and texture alignment in occluded and curved regions remain visually stable, confirming the effectiveness of the joint depth and texture modeling approach.

In order to quantify the influence of graph neural network structure on texture refinement quality, an internal comparative experiment was constructed under controlled graph convolution depth and topology variations. The evaluation results are summarized in Table 2. When adopting a fixed-radius pixel adjacency graph, the SSIM metric increased from 0.812 to 0.843 as the number of graph convolution layers rose from one to two, accompanied by a reduction in LPIPS from 0.188 to 0.176, indicating enhanced structural alignment and lower perceptual deviation. Extending the depth to four layers

Fig. 7. Multi-view comparison of original input images and reconstructed outputs produced by the proposed method. (a) Input RGB images from different viewpoints. (b) Reconstructed outputs by GNN-NeRF method.

Table 2. Effect of graph structure and depth on texture restoration metrics.

Graph Conv Layers	Graph Type	SSIM (HF Region)	LPIPS (Occlusion Region)
1	Fixed-radius pixel graph	0.812	0.188
2	Fixed-radius pixel graph	0.843	0.176
4	Fixed-radius pixel graph	0.838	0.181
2	Superpixel-based graph	0.854	0.169
2	Adaptive affinity graph	0.871	0.162

did not yield further improvements, with SSIM marginally decreasing to 0.838 and LPIPS increasing to 0.181, reflecting a degradation attributed to feature oversmoothing effects. Under the condition of fixed two-layer graph convolution, adjustments to the graph topology led to further differentiation in performance. When replacing the fixed-radius graph with a superpixel-based graph, SSIM improved to 0.854 and LPIPS declined to 0.169. The adaptive affinity graph structure achieved the highest SSIM at 0.871 and the lowest LPIPS at 0.162, indicating that structural modeling guided by feature-driven affinity contributes to higher fidelity in edge continuity and context preservation. These comparative results demonstrate that both the depth of graph convolution and the design of inter-node connectivity play a critical role in the accuracy of texture structure recovery. The observed performance variation under different configurations confirms the necessity of structural optimization in the design of graph-based texture enhancement modules.

5.3 Multi-view consistency evaluation

To comprehensively evaluate the visual consistency performance of the proposed method under multi-view conditions, the experiment compared and analyzed the fluctuations of the structural similarity (SSIM) and perceptual distance (LPIPS) indicators of five typical 3D image reconstruction algorithms under different viewing directions from the frontal view and multiple angles on both sides. Among the compared methods, method A and method B are based on traditional multi-view geometry and rough depth estimation based on image block matching, respectively. Methods C and D use neural implicit network modeling. The former does not perform texture optimization, and the latter introduces a local texture restoration strategy. Method E in this paper introduces a depth perception feature fusion mechanism in the structural modeling stage, and combines structural similarity with perceptual loss for collaborative optimization. The original intention of the design is to improve the coherence and detail stability of the reconstructed image under different viewing angles. In Figure 8, the horizontal axis is the viewing angle, and the vertical axis represents the image structure similarity and perceptual distance, respectively. A double Y-axis line chart is used to distinguish the nature of the indicators. The left axis represents SSIM, and the right axis corresponds to LPIPS.

From the overall trend, all methods perform best in the central perspective, but have varying degrees of performance degradation under offset angles. Among them, method E always maintains a high SSIM value and a low LPIPS value in the range of $-45°$ to $+45°$, showing its reconstruction stability under multiple perspectives. The structural similarity of method E at a viewing angle of 0° reaches 0.927, which is slightly lower than 0.942 of method D, but it remains above 0.821 at angles of $\pm 30°$ and $\pm 45°$.

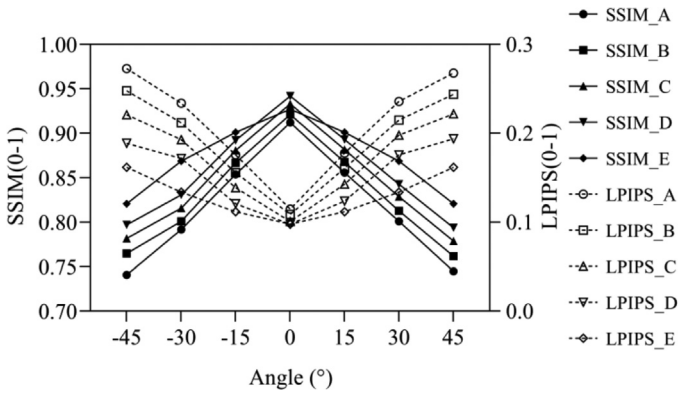

Fig. 8. Comparison of structural similarity and perceptual distance index trends of various methods under multiple perspectives.

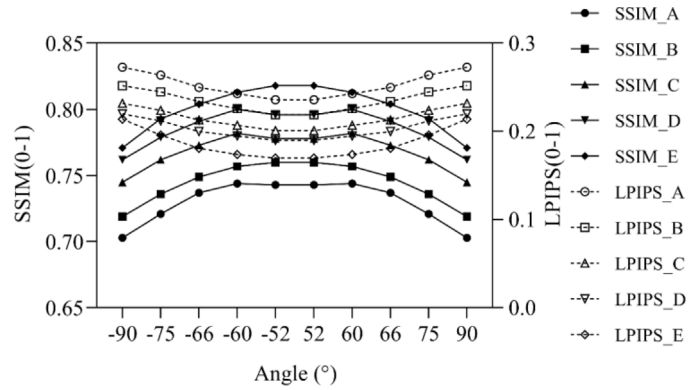

Fig. 9. Comparison of consistency indicators of multiple methods under the expansion of viewing angle offset.

In contrast, the highest value of method A at the same angle is only 0.801, indicating that method E can still maintain a high structural consistency under large viewing angle offsets. In terms of perceptual distance, method E is 0.098 at 0° and only 0.162 at ±45°, while method A is as high as 0.273, indicating that the optimization of texture modeling by method E effectively suppresses local perceptual distortion. This consistency advantage across angles is attributed to the joint expression of depth information and color features, as well as the contextual completion mechanism of the graph neural network for edge texture break areas. It can be seen that the proposed method exhibits higher visual communication stability and texture coherence under complex perspective changes.

To further verify the structural consistency and perceptual stability of the proposed method under larger viewing angle offsets, the experiment expanded the multi-view consistency evaluation range from the original ±45° to ±90°. A denser sampling point was constructed in the angle range of ±52° to ±90°, and the structural similarity and perceptual distance indicators of each method under these viewing angles were calculated, respectively, aiming to systematically examine the performance trend and stability differences under large viewing angles. The results are shown in Figure 9.

In terms of structural similarity, the proposed method still maintains a high level of 0.818 at ±52°, slightly lower than ±45° but maintaining a smooth trend, and then drops to 0.771 at ±90°, but is generally better than other methods, and the decline is controlled within 0.050. The fluctuations of the comparison methods in the same viewing angle range are more drastic. Method A is only 0.743 at ±52° and drops to 0.703 at ±90°, and the structural expression is significantly weakened. Although Methods B to D maintain a certain level, the slope of the decline curve is generally greater than that of the proposed method, indicating that their structural consistency is unstable under larger viewing angle offsets. In terms of perceptual distance, the LPIPS of the proposed method rises gently from 0.170 at ±52° to 0.214 at ±90°, maintaining the lowest value while maintaining a stable trend, indicating that it has stronger texture retention ability and local continuity adaptability at the perceptual level. The other methods all showed a significant increase in LPIPS at ±75° and ±90° viewing angles, while Method A reached 0.273 at ±90°, with significant subjective quality degradation, indicating that it is difficult to maintain texture consistency under large angle offsets. This shows that the proposed method still has good structural expression ability and perceptual stability under a wide range of viewing angle changes, is more adaptable, and has limited error diffusion.

5.4 Comparative analysis of perceptual quality indicators

In order to comprehensively evaluate the performance of different 3D reconstruction methods in terms of subjective visual perception quality, a perceptual consistency comparison experiment was designed. Based on the three indicators of SSIM, LPIPS, and Natural Image Quality Evaluator (NIQE), the ability differences of each method in the dimensions of structure restoration, texture realism, and visual naturalness were analyzed. The first method is NeRF, which uses ray sampling and voxel density field modeling for implicit rendering. It shows basic modeling capabilities in free-view synthesis, but has great limitations in geometric structure continuity and texture detail restoration. The second method is Multum in Parvo-Neural Radiance Fields (Mip-NeRF), which introduces conical rays and Mip-level sampling strategies based on NeRF to reduce aliasing artifacts in reconstruction, making scale changes and edge blurring smoother. The third method is VolRecon, which constructs a dense voxel grid and uses a volume rendering fusion strategy for geometric reconstruction, which improves local consistency and spatial restoration capabilities, but the texture expression is still limited by the modeling method. The fourth method is TensoRF, which constructs a low-rank structured tensor field through tensor factor decomposition to achieve efficient compression and fast rendering, and achieves good results in texture clarity. Method 5 is a 3D reconstruction algorithm that integrates depth-guided neural implicit modeling and graph neural network texture enhancement, jointly encodes the RGB and depth information of the image, and constructs a NeRF voxel network with structure-aware constraints. At the same time, the graph structure context is introduced to complete

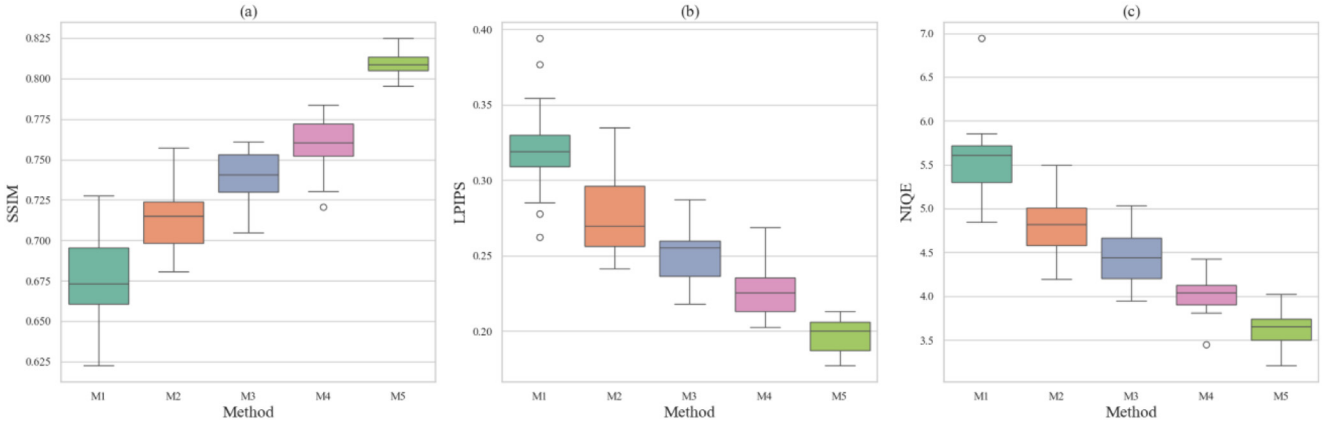

Fig. 10. Comparison of perceptual quality distribution of different 3D reconstruction methods under SSIM, LPIPS, and NIQE indicators. (a) Comparison of structural similarity indicator distribution of each method. (b) Comparison of the perceptual distance indicator distribution of each method. (c) Comparison of the no-reference image quality indicator distribution of each method.

the completion of texture breaks and edge areas, enhancing the perceptual consistency in visual communication. Figure 10 shows the box-line statistical distribution of the three indicators under the five methods. The horizontal axis is the comparison method, and the vertical axis corresponds to the SSIM, LPIPS, and NIQE indicators, respectively. The box in the figure represents the main distribution range. The median line reflects the central trend of the indicators of each method, and the box whisker range depicts its extreme value fluctuations, which helps to reveal the stability and performance differences of the algorithm in the subjective quality perception dimension.

The median of method 5 in the SSIM dimension is about 0.8089, which is higher than 0.7604 of method 4 and 0.7406 of method 3, and the distribution box range is the smallest, indicating that it has higher stability in structural consistency. In terms of LPIPS, the median value of method 5 is about 0.1999, which is much lower than 0.3191 of method 1 and 0.2697 of method 2, and the upper and lower quartile distances are shorter, suggesting that it has better consistency in the multi-view perception space. This performance improvement is mainly attributed to the introduction of perceptual consistency constraints and deep joint encoding in the model during the training phase, which makes the texture expression more spatially coherent. In terms of NIQE indicators, the median of method 5 is 3.6588, which is significantly better than 4.0362 of method 4 and 4.4435 of method 3. This shows that its texture restoration is more visually natural, with fewer artifacts in the reconstructed image and more coordinated details. The box-line distribution of the three indicators shows that method 5 achieves better texture integrity and perceptual quality while ensuring structural fidelity, which is closely related to its introduction of depth-guided encoding and graph neural texture completion mechanism.

5.5 Method stability and generalization ability test

In order to evaluate the performance of the three-dimensional image visual optimization algorithm proposed in this paper in terms of structural restoration and perceptual communication, a multi-method comparison experiment with SSIM and LPIPS as evaluation indicators was constructed. The compared algorithms cover the current mainstream neural implicit modeling and multi-view reconstruction methods, which are as follows: Method 1 is NeRF, Method 2 is the enhanced version of Neural Radiance Field (NeRF in the Wild, NeRF-W), Method 3 is MVSNet based on multi-view volume reconstruction, Method 4 is COLMAP+Texture Mapping (COLMAP +TM) which combines multi-view matching with traditional texture fusion, and Method 5 proposed in this study is an improved algorithm that combines voxel hashing neural modeling with perceptual consistency optimization mechanism. The five public datasets used have typical scene differences. Among them, ScanNet is for indoor scenes, Tanks & Temples, and Blended Multi-View Stereo (BlendedMVS) represent complex outdoor and synthetic environments, Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) focuses on urban traffic perspectives, and Danish Technical University (DTU) provides densely sampled standard model scenes. In order to fully reflect the robustness and visual fidelity of different methods in different scenarios, Tables 3 and 4 show the SSIM and LPIPS index scores of each method on the above five datasets.

In terms of structural restoration performance, the improved method that integrates voxel hash coding and a multi-scale-aware optimization mechanism achieved the highest SSIM score on all datasets. It performed particularly well on the DTU dataset, reaching 0.934, significantly higher than NeRF's 0.901 and COLMAP+TM's 0.884, indicating that this method has stronger structural consistency modeling capabilities in geometrically dense areas. NeRF-W achieved similar results in ScanNet and Tanks&Temples, 0.892 and 0.887, respectively, indicating that the image feature regularization introduced by it helps improve the robustness in complex texture environments. Although MVSNet maintained an upper-middle level in KITTI and BlendedMVS, it performed slightly worse on the boundaries of structural details due to the lack of a texture continuity modeling mechanism. In the perceptual

Table 3. Structural similarity index of each reconstruction method on five typical 3D datasets.

Dataset	Method 1	Method 2	Method 3	Method 4	Method 5
ScanNet	0.871	0.892	0.878	0.845	0.915
Tanks&Temples	0.862	0.887	0.869	0.841	0.902
KITTI 2015	0.845	0.867	0.852	0.831	0.887
DTU	0.901	0.915	0.896	0.884	0.934
BlendedMVS	0.868	0.889	0.873	0.842	0.908

Table 4. Perceptual image distance indicators of each reconstruction method on five typical 3D datasets.

Dataset	Method 1	Method 2	Method 3	Method 4	Method 5
ScanNet	0.194	0.162	0.175	0.208	0.137
Tanks&Temples	0.188	0.158	0.169	0.204	0.142
KITTI 2015	0.204	0.177	0.189	0.223	0.158
DTU	0.169	0.144	0.158	0.184	0.121
BlendedMVS	0.192	0.159	0.172	0.211	0.139

consistency evaluation, the improved method outperforms other methods in LPIPS on all datasets. It obtains 1 − LPIPS = 0.879 on DTU, which is significantly better than NeRF-W's 0.856. This shows that the dual perceptual constraint strategy combining SSIM and LPIPS can effectively reduce the perceptual deviation of reconstructed images under texture breaks and illumination changes. COLMAP+TM performed the weakest among all indicators, indicating that the traditional texture mapping scheme suffers from a large loss of perceptual consistency when faced with large viewing angle deviations or incomplete depth inputs. Comprehensive analysis shows that the method that integrates deep structure priors, perceptual guided training, and graph neural texture completion mechanisms exhibits high structural accuracy and perceptual stability in multiple scenarios, and has strong versatility and generalization capabilities.

5.6 User-side communication performance evaluation

In order to deeply evaluate the response efficiency of different 3D reconstruction models in visual communication tasks, this paper designed a user experiment including five typical recognition tasks. These tasks include occluded object positioning (T1), heterogeneous texture recognition (T2), spatial structure judgment (T3), multi-view consistency judgment (T4), and edge detail discrimination (T5). This covers the common information communication obstacles in actual virtual visual scenes, especially the recognition difficulties caused by geometric continuity interruptions and texture breaks. Five representative models were compared in the experiment: Model 1 is the traditional Structure from Motion- Multi-View Stereo (SfM-MVS) system COLMAP, which has stable geometric reconstruction capabilities but lacks texture continuous expression; Model 2 is the original NeRF framework, which is modeled based on neural radiation fields and has strong perspective consistency but insufficient geometric details;

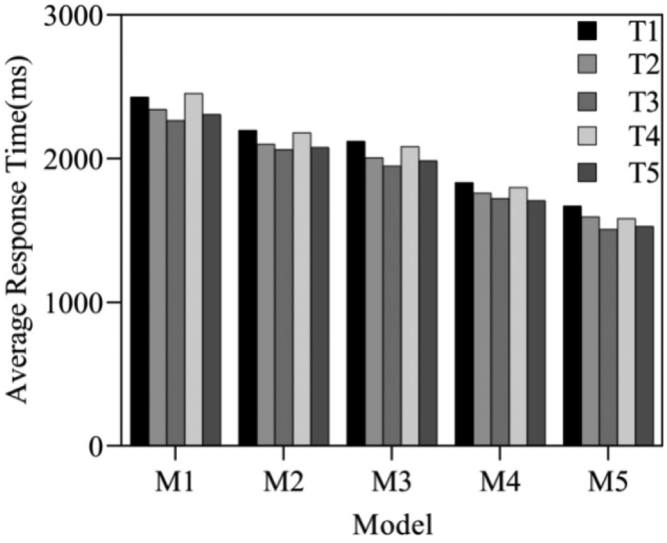

Fig. 11. Average user reaction time of different 3D reconstruction models in five types of visual communication tasks.

Model 3 introduces NeRF+DepthBaseline with depth map prior, integrating certain structural guidance but not optimizing texture and perceptual characteristics; Model 4 is a variant of the proposed method that removes the graph neural network module, aiming to verify the impact of the context completion mechanism on edge and texture expression; Model 5 is the complete optimization algorithm of this paper, which combines MiDaS depth estimation, perceptual consistency loss and graph neural enhancement to achieve multi-level optimization from geometric structure to subjective perception. Figure 11 shows the average reaction time of each model under different tasks.

From the experimental results, Model 5 showed the lowest average reaction time in all tasks. In the edge detail judgment task, the response speed reached 1530 milliseconds, which was significantly lower than COLMAP's

Table 5. Comparison of user performance of T1 and T4 tasks under dynamic perturbation conditions.

Task type	Disturbance level	Recognition accuracy (%)	Task completion rate (%)	Error variation (%)
T1	No Disturbance	87.3	93.2	5.1
T1	Mild Disturbance	83.2	89.1	6.3
T1	Moderate Disturbance	78.4	84.5	8.2
T1	High Disturbance	72.1	77.8	10.7
T4	No Disturbance	85.5	91.8	6.4
T4	Mild Disturbance	81.4	87.9	7.5
T4	Moderate Disturbance	76.9	83.4	9
T4	High Disturbance	70.6	76.1	11.3

2310 milliseconds. This difference is mainly due to the structure completion and edge refinement of local texture fracture areas by the graph neural network, which enables users to quickly identify targets in structurally blurred areas. Although Model Three introduces depth priors, due to the lack of a targeted texture enhancement mechanism, the response time in heterogeneous texture recognition tasks is still 2008 milliseconds, which is significantly slower than the 1598 milliseconds of model five. Model 1 and Model 2 performed relatively poorly in the multi-view consistency task, with average response times exceeding 2100 milliseconds, reflecting that neither the traditional method nor the original NeRF effectively solved the visual instability problem caused by geometric discontinuity. The overall results show that in complex visual communication scenes, the joint optimization of structure completion and perception guidance significantly improves the efficiency of user information extraction.

In the experiment of dynamic tasks, the goal is to evaluate the adaptability of dynamic perturbations to users in task execution, especially in the two tasks of occluded target positioning (T1) and multi-view consistency judgment (T4). The experiment introduces different levels of perturbations to simulate the dynamic changes in real scenes and examine the visual stability and task execution efficiency of users under different perturbation intensities. In the task design, mild, moderate, and high perturbations gradually increase the complexity of the task through changes in view angle offset and target occlusion, and then observe their impact on user performance. The results are shown in Table 5.

It can be seen from the experimental data that as the perturbation level increases, the user's performance in the T1 and T4 tasks decreases significantly. In the T1 task, mild perturbation reduces the recognition accuracy from 87.3% to 83.2% and the task completion rate from 93.2% to 89.1%. Moderate perturbation further reduces the recognition accuracy to 78.4% and the task completion rate to 84.5%. High perturbation causes the accuracy to drop sharply to 72.1%, and the task completion rate to 77.8%. The error change increases to 10.7% under high perturbation conditions, reflecting the significant interference of occlusion on task execution. Similar trends are also verified in the T4 task, where mild perturbation reduces the recognition accuracy from 85.5% to 81.4% and the task completion rate from 91.8% to 87.9%. Under moderate perturbation conditions, the recognition accuracy is 76.9% and the task completion rate is 83.4%, while under high perturbation conditions, the recognition accuracy drops to 70.6% and the task completion rate is 76.1%. The error variation also increased to 11.3%, showing the instability of task execution under high disturbance conditions. The experimental results show that as the disturbance intensity increases, the user's performance in task completion ability decreases significantly, especially in tasks involving visual occlusion and perspective changes. The data reveals the interference of visual instability on cognitive processing and decision-making processes. As the disturbance intensity increases, the user's error rate and response delay in task execution increase significantly. Under highly disturbed conditions, it is difficult for users to maintain a stable flow of visual information, resulting in a decrease in target recognition accuracy and affected consistency in task execution. These results show that the intensity of dynamic disturbance is negatively correlated with the user's ability to adapt to complex visual environments, and environments with strong disturbances significantly weaken the accuracy and efficiency of task execution.

In order to further verify the performance advantage of the optimization algorithm at the user perception level, a multi-dimensional subjective immersion scoring experiment was set up on the basis of completing the response efficiency evaluation. This experiment continues the five model settings in Figure 7 and comprehensively describes the user's immersive experience in the process of watching the reconstructed image through five perception dimensions: Scene realism (D1) measures the overall realism of the image, spatial consistency (D2) reflects the stability of multi-view geometric expression, and texture naturalness (D3) evaluates whether the texture generation has continuity and realism. Edge clarity (D4) focuses on the visual sharpness of the object boundary, and interactive integration (D5) reflects the user's state of involvement in the viewing process. These dimensions correspond to the key communication barriers in image generation and have a direct impact on the improvement of perceived quality. The experiment presents the scores of each model in five dimensions in the form of a grouped bar chart, and superimposes the average score line for overall trend judgment. Different colors are used to distinguish the scoring dimensions in Figure 12. The X-axis represents the

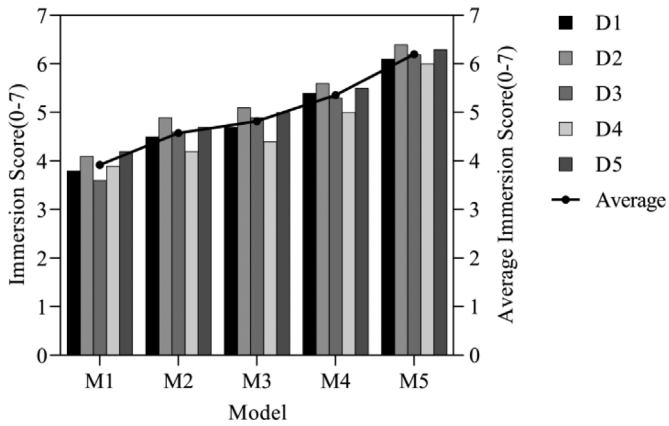

Fig. 12. Comparison of subjective evaluations of different 3D reconstruction models in the five dimensions of immersion.

names of the five models, the left side of the Y-axis corresponds to the dimension score value, and the right side corresponds to the average immersion score, which is used to assist in observing the correlation between multi-dimensional performance and overall evaluation.

Model 5 received the highest evaluation in all dimensions, with an average score of 6.20 points. It scored 6.1 points and 6.3 points in the dimensions of scene realism and interactive integration, respectively, showing the significant effectiveness of the graph neural completion mechanism and perceptual consistency optimization in enhancing the subjective immersive experience. Model 4 scored 5.3 in terms of texture naturalness, which is significantly lower than Model 5's 6.2. This shows that after removing GNN, the expression of the texture area still has local structural defects, which affect the user's overall immersion judgment. Although Model 3 introduced a depth prior and scored 5.1 in terms of spatial consistency, which is an improvement over NeRF, it did not reach 5 points in terms of edge clarity and scene realism, and still did not effectively alleviate the problems of geometric discontinuity and texture drift. The traditional MVS method (model 1) is in the lowest range in the five dimension scores, with an average score of less than 4 points. Its poor texture restoration ability directly weakens the user's perception of the real scene. The results show that structural perception optimization and contextual texture completion have a key impact on the subjective immersive experience. The improvement of the optimization model in the logical consistency of image construction and the ability to express details jointly support the significant enhancement of immersion.

5.7 Influence of multi-scale fusion weights

To investigate the impact of multi-scale feature fusion weight configuration on 3D image structural perception and visual consistency, this paper conducted experiments perturbing the weight parameters of the three convolution kernel branches within a trained multi-scale convolutional module by ±10%. The three sets of convolution kernels correspond to 3×3 (w_1), 5×5 (w_2), and 7×7 (w_3) scales, focusing on local image details, regional semantics, and global context, respectively. The perturbation experiments reduced or increased the weight of one of the parameters while keeping the other parameters unchanged to analyze the differences in the role of features at different scales in the overall modeling. Evaluation metrics included structural similarity (SSIM), perceptual distance (LPIPS), and peak signal-to-noise ratio (PSNR). The various perturbation schemes were compared using a unified dataset and training configuration. The experimental results are shown in Table 6.

Table 6. The impact of multi-scale fusion weight perturbation on high-frequency regional structure and perception indicators.

Perturbation condition	SSIM ↑	LPIPS ↓	PSNR (dB) ↑
Baseline (no change)	0.871	0.162	27.94
w_1 −10%	0.85	0.176	26.43
w_1 +10%	0.872	0.161	28.01
w_2 −10%	0.862	0.168	27.62
w_2 +10%	0.868	0.165	27.89
w_3 −10%	0.869	0.163	27.9
w_3 +10%	0.873	0.16	28.03

Among the various perturbations, decreasing the w_1 parameter resulted in a 0.021 decrease in SSIM, a 0.014 increase in LPIPS, and a 1.51 dB decrease in PSNR, demonstrating a significant impact on structural continuity and texture consistency. The 3×3 convolutional kernel primarily targets local edge regions of the image, and its weight changes directly impact the model's detail recovery capabilities. When its contribution is weakened, structural representation and perceptual alignment performance in high-frequency regions significantly degrade. In contrast, perturbing w_2 results in smaller fluctuations in these metrics, indicating that the mid-scale semantic structure modeled by the 5×5 convolutional kernel plays a moderate role in maintaining local stability. Perturbing w_3 results in the smallest fluctuations in structural and perceptual metrics, indicating that the global perceptual features modeled by the 7×7 convolutional kernel are less sensitive in high-frequency regions. Experimental results demonstrate that in multi-scale feature fusion, low-scale channels dominate the modeling of image details and structural boundaries. Properly configuring weights is key to improving image clarity and perceptual consistency.

5.8 Real-time performance and dynamic scene evaluation

In order to evaluate the runtime stability and visual structure continuity of the algorithm in dynamic scenes, a frame-level time series dataset was constructed to jointly analyze the inference delay and cross-frame image consistency. The experiment was measured under 60 frames of continuous dynamic input, recording the inference time of each frame and the structural similarity

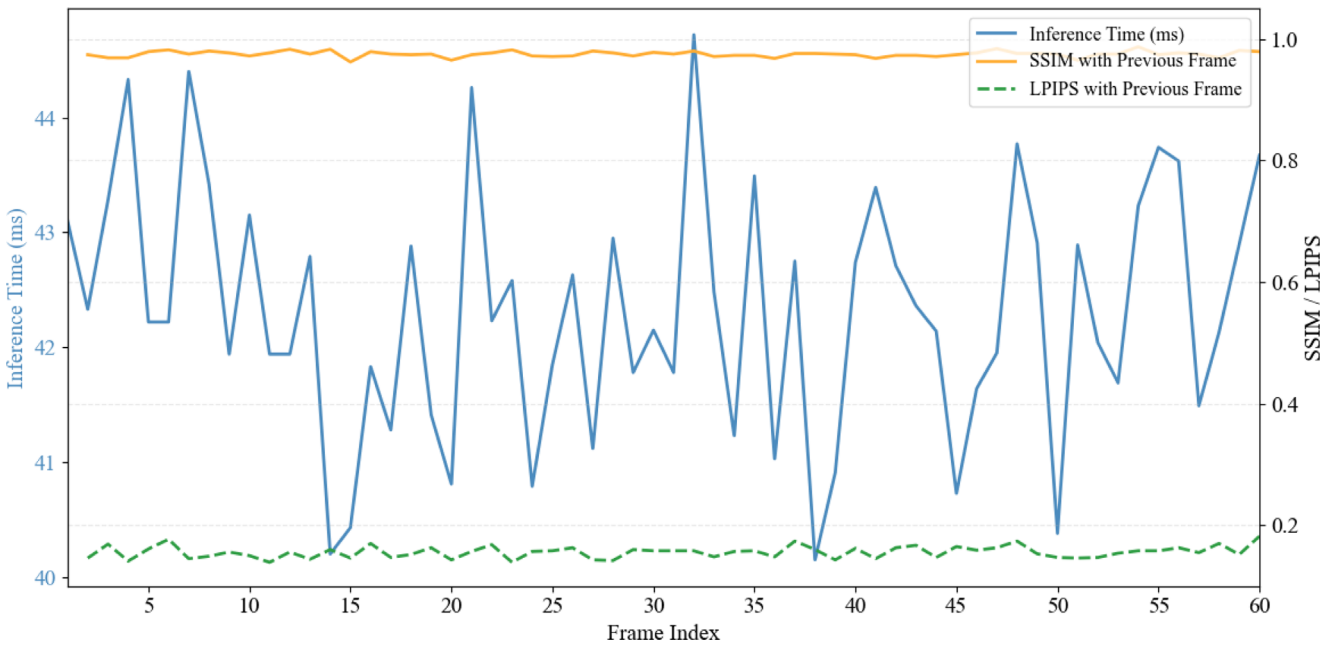

Fig. 13. Joint trends of the algorithm's frame-level inference latency, structural consistency, and perceptual stability in dynamic sequences.

and perceptual distance between adjacent frames, and simultaneously drawing a three-dimensional joint curve to show the temporal change trend of the system response and image consistency. Figure 13 shows the change trajectory of the three indicators of frame-level inference time, structural similarity (SSIM), and perceptual distance (LPIPS) in the dynamic frame sequence of the algorithm as the frame advances.

The inference delay curve fluctuates slightly between 42 milliseconds and 44 milliseconds, with an average of 42.7 milliseconds, reflecting that the method maintains stable frame processing capabilities in dynamic tasks. The SSIM curve remains between 0.972 and 0.980 for most of the time, with a small fluctuation range, and only a slight drop occurs at a few frame points, indicating that the reconstructed image performs well in terms of inter-frame structural continuity. The LPIPS index varies in the range of 0.145 to 0.165 as a whole, with a maximum value not exceeding 0.182, indicating that the consistency between images in the perceptual dimension has not been significantly disturbed. The stable trend of SSIM and LPIPS is closely related to the synchronous stability of the inference time, indicating that the algorithm effectively suppresses the structural and semantic drift caused by disturbances while maintaining high operating efficiency. The coupling performance between the three indicators supports the method's ability to be real-time and consistent across frames in dynamic scenes.

6 Conclusions

This paper proposes a 3D image visual communication optimization algorithm that integrates neural implicit modeling and multi-scale visual perception. By jointly encoding RGB images and MiDaS depth maps into the voxel hashing NeRF framework and introducing structural similarity constraints and perceptual consistency loss functions, the geometric accuracy and texture authenticity of 3D reconstruction are significantly improved. Experiments have verified that the RMSE of depth reconstruction in occluded areas of the algorithm is reduced to 0.164 meters, which is 0.043 meters lower than the traditional NeRF method; the SSIM in high-frequency texture areas reaches 0.871; in multi-view consistency evaluation, the LPIPS under a $\pm 45^\circ$ viewing angle offset is stable at 0.162, with the maximum reduction of 40.7% compared with the comparison method. These data show that the joint expression of depth information and color features and the context completion mechanism of graph neural networks for texture fracture areas effectively enhance the restoration ability of complex structures and the stability of cross-view perception. Current methods are limited to static scene reconstruction, lack adaptability to dynamic objects and extreme lighting, and consume high computing resources. In the future, it is necessary to explore lightweight real-time modeling architecture, combine physical lighting models to optimize the expression of material reflection characteristics, and expand to non-rigid scene reconstruction to support the construction of highly interactive virtual environments.

Funding

This work was supported by the Guangxi Nature Science Foundation Project 2021GXNSFBA220023, Guangxi College Youth Basic Ability Improvement Project 2021KY0620 and 2021KY1019, 2021 Hechi University High-level Talents Research Start-up Project 2021GCC013.

Conflicts of interest

The authors have nothing to disclose.

Data availability statement

This article has no associated data generated and/or analyzed.

Author contribution statement

Conceptualization, Y.Q., F.L., M.L., P.G. and L.L.; Methodology, Y.Q.; Software, X.X.; Validation, L.L.; Formal Analysis, M. L.; Investigation, F.L; Resources, P.G.; Writing – Original Draft Preparation, Y.Q., F.L., M.L., P.G. and L.L.; Visualization, Y.Q.; Supervision, P.G.

References

1. P. Maken, A. Gupta, 2D-to-3D: a review for computational 3D image reconstruction from X-ray images, Arch. Comput. Methods Eng. **30**, 85–114 (2023)
2. G. Yao, Z. Wang, G. Wei, F. Zhu, Q. Fu, Q. Yu, M. Wei, Multi-view three-dimensional reconstruction based on feature enhancement and weight optimization network, ISPRS Int. J. Geo-Inform. **14**, 43–58 (2025)
3. R. Jia, X. Chen, J. Cui, Z. Hu, MVS-T: a coarse-to-fine multi-view stereo network with transformer for low-resolution images 3D reconstruction, Sensors **22**, 7659–7673 (2022)
4. L.I. Xiangyu, Z.H.A.N.G. Xueqin, ORBTSDF-SCNet: an online 3D reconstruction method for dynamic scene, J. East China Univ. Sci. Technol. **49**, 284–294 (2023)
5. F. Luo, Y. Zhu, Y. Fu, H. Zhou, Z. Chen, C. Xiao, Sparse rgb-d images create a real thing: a flexible voxel based 3d reconstruction pipeline for single object, Visual Inform. **7**, 66–76 (2023)
6. Z. Li, M. Oskarsson, A. Heyden, Detailed 3D human body reconstruction from multi-view images combining voxel super-resolution and learned implicit representation, Appl. Intell. **52**, 6739–6759 (2022)
7. Y. Feng, R. Wu, X. Liu, L. Chen, Three-dimensional reconstruction based on multiple views of structured light projectors and point cloud registration noise removal for fusion, Sensors **23**, 8675–8692 (2023)
8. H. Luo, C. Pape, E. Reithmeier, Scale-aware multi-view reconstruction using an active triple-camera system, Sensors **20**, 6726–6749 (2020)
9. M. Chen, Z. Duan, Z. Lan, S. Yi, Scene reconstruction algorithm for unstructured weak-texture regions based on stereo vision, Appl. Sci. **13**, 6407–6432 (2023)
10. M. Oliveira, G.H. Lim, T. Madeira, P. Dias, V. Santos, Robust texture mapping using rgb-d cameras, Sensors **21**, 3248–3270 (2021)
11. T.J. Mu, H.X. Chen, J.X. Cai, N. Guo, Neural 3D reconstruction from sparse views using geometric priors, Comput. Visual Media **9**, 687–697 (2023)
12. Q. Zhou, J. Zuo, W. Kang, M. Ren, High-precision 3D reconstruction in complex scenes via implicit surface reconstruction enhanced by multi-sensor data fusion, Sensors **25**, 2820–2841 (2025)
13. N. Luo, L. Huang, Q. Wang, G. Liu, An improved algorithm robust to illumination variations for reconstructing point cloud models from images, Remote Sens. **13**, 567–588 (2021)
14. T. Madeira, M. Oliveira, P. Dias, Neural colour correction for indoor 3D reconstruction using RGB-D data, Sensors (Basel, Switzerland) **24**, 4141–4152 (2024)
15. G. Nayak Seetanadi, K.E. Årzen, M. Maggio, Control-based event-driven bandwidth allocation scheme for video-surveillance systems, Cyber-Phys. Syst. **8**, 111–137 (2021)
16. B. Yogi, S. Roy, A.K. Khan et al., IELTSoC: enhanced image encryption using combined logistic and Tinkerbell maps with second order cellular automata for internet of things, Discov. Internet Things **5**, 78 (2025)
17. H. Jianfeng et al., Enhanced YOLOv11 for image-based anomaly detection in freight train gate chains, Int. J. Intell. Inform. Technolog. **21**, 1–18 (2025)
18. P. Velagapalli, N. Parveen, Concurrent attentional reconstruction network for 3D point cloud reconstruction from single image, Appl. Soft Comput. **172**, 112821 (2025)
19. B. Cui, W. Tao, H. Zhao, High-precision 3D reconstruction for small-to-medium-sized objects utilizing line-structured light scanning: a review, Remote Sens. **13**, 4457–4490 (2021)
20. Z. Hong, Y. Yang, J. Liu, S. Jiang, H. Pan, R. Zhou, C. Zhong, Enhancing 3D reconstruction model by deep learning and its application in building damage assessment after earthquake, Appl. Sci. **12**, 9790–9804 (2022)
21. J.T.S. Phang, K.H. Lim, R.C.W. Chiong, A review of three dimensional reconstruction techniques, Multimedia Tools Appl. **80**, 17879–17891 (2021)
22. K. Fu, J. Peng, Q. He, H. Zhang, Single image 3D object reconstruction based on deep learning: a review, Multimedia Tools Appl. **80**, 463–498 (2021)
23. Y. Lee, Three-dimensional dense reconstruction: a review of algorithms and datasets, Sensors **24**, 5861–5882 (2024)
24. X. Zhang, Z. Zheng, D. Gao, B. Zhang, Y. Yang, T.S. Chua, Multi-view consistent generative adversarial networks for compositional 3d-aware image synthesis, Int. J. Computer Vision **131**, 2219–2242 (2023)
25. L. Zhou, Z. Zhang, H. Jiang, H. Sun, H. Bao, G. Zhang, DP-MVS: detail preserving multi-view surface reconstruction of large-scale scenes, Remote Sens. **13**, 4569–4589 (2021)
26. V. Leroy, J.S. Franco, E. Boyer, Volume sweeping: learning photoconsistency for multi-view shape reconstruction, Int. J. Comput. Vision **129**, 284–299 (2021)
27. H. Zhang, Y. Zhang, L. Zhu, W. Lin, Deep learning-based perceptual video quality enhancement for 3D synthesized view, IEEE Trans. Circ. Syst. Video Technol. **32**, 5080–5094 (2022)
28. Z. Xia, Q. Han, Y. Zhang, Y. Zhang, F. Hu, Objective quantification of dynamic spatial distortions for enhanced realism in virtual environments, Virtual Real. **29**, 1–10 (2025)
29. S. Coskun, G. Nur Yilmaz, F. Battisti, M. Alhussein, S. Islam, Measuring 3D video quality of experience (QoE) using a hybrid metric based on spatial resolution and depth cues, J. Imag. **9**, 281–309 (2023)
30. L. Shen, Y. Yao, X. Geng, R. Fang, D. Wu, A novel no-reference quality assessment metric for stereoscopic images with consideration of comprehensive 3D quality information, Sensors **23**, 6230–6254 (2023)

31. S. Jiang, K. You, Y. Li, D. Weng, W. Chen, 3D reconstruction of spherical images: a review of techniques, applications, and prospects, Geo-spatial Inform. Sci. **27**, 1959–1988 (2024)
32. W. Zhou, L. Yuan, T. Mu, Multi3D: 3D-aware multimodal image synthesis, Comput. Visual Media **10**, 1205–1217 (2024)
33. Q. Jia, L. Chang, B. Qiang, S. Zhang, W. Xie, X. Yang, M. Yang, Real-time 3D reconstruction method based on monocular vision, Sensors **21**, 5909–5926 (2021)
34. F.E. Fadzli, A.W. Ismail, S. Abd Karim, M.N.A. Nor'a Ishigaki, M.Y.F. Aladin, Real-time 3D reconstruction method for holographic telepresence, Appl. Sci. **12**, 4009–4024 (2022)

Cite this article as: Yunchu Qin, Fugui Luo, Mingzhen Li, Peigang Guo, Liangliang Li, Optimization algorithm for 3D image visual communication based on digital image reconstruction, Int. J. Simul. Multidisci. Des. Optim. **16**, 14 (2025), https://doi.org/10.1051/smdo/2025015