npj | heritage science

Article

https://doi.org/10.1038/s40494-026-02424-8

3DSynBrush a high quality 3D reconstruction framework for single Dunhuang murals

Check for updates

Xianlin Peng[1], Jingyu Wang[2], Qiyao Hu[2,3,4] ✉, Nuo Xu[5], Jun Wang[2], Shuyi Qu[2] & Jinye Peng[2,6]

The Dunhuang Mogao Grottoes are among the most significant historical and cultural sites in China, with the *Dunhuang Mural* serving as key representative artifacts. Due to natural erosion, human activities, and other factors, it is crucial to protect these frescoes authentically and accurately to ensure their preservation. To address this, we propose a technological framework, 3DSynBrush, for high-quality 3D reconstruction of Chinese painting data. First, single elements are extracted from the mural through the Perspective-Driven Synthesis (PDS) Module, and a sparse perspective is generated. The sparse view is then passed through the Neural Rendering Synthesizer to obtain a continuous view image, which ultimately meets the input requirements of the Light Field Fusion Meshing Module to achieve 3D reconstruction. 3DSynBrush aims to reconstruct a realistic model of the scene depicted using only a single primary view of the fresco. Moreover, we develope a high-quality Chinese Mural Elements dataset, termed CME. Compared to current 3D reconstruction algorithms, 3DSynBrush produces visually coherent results while using only 40% of the vertices and triangles, thereby minimizing computational resources.

The history of the Dunhuang cultural heritage dates back to 366 CE. It is a splendid crystallization accumulated over a millennium of Chinese history. It is widely regarded as one of the largest, oldest, most content rich, and best preserved repositories of cultural heritage in the world. As an unparalleled treasure within the Buddhist grotto cultural heritage of China, the Dunhuang Mogao Grottoes are renowned for their vast collection of murals and lifelike sculptures. These grottoes are not only halls of art but also faithful witnesses to history. They profoundly reflect ancient social life, religious evolution, and the integration and collision of diverse Eastern and Western cultures[1]. Thus, with their unique cultural foundation and outstanding artistic achievements, they have become an indispensable symbol of Chinese civilization. These murals enjoy a prestigious reputation in art history for exquisite techniques and grand themes. Particularly during the *Tang* Dynasty, their techniques and forms of expression reached a zenith. This makes them not only possess extremely high artistic and aesthetic value but also serve as precious historical documents containing rich information. Today, the *Dunhuang Murals* are not only a cultural heritage in urgent need of protection but also a valuable cultural tourism resource that inspires future generations and attracts the world.

The artistic achievements of the *Dunhuang murals* are embodied in all encompassing depictions. They reach a pinnacle of excellence in multiple domains. In architectural representation, the murals span over a thousand years. They present an extremely rich variety of stupa architectural forms[2]. According to the research of *Sicheng Liang*, the stupas in the *Dunhuang murals* can be subdivided into six types. These include single story and multi story wooden stupas, sloped roof brick and stone stupas, single story and multi story brick and stone stupas, and the brick and stone stupa known in Japan as a hall stupa[2]. Among these, the depiction of the *Hall-style Stupa* not only inherits the architectural features of early stupas but also further enriches stupa architecture forms. Specific grotto examples vividly confirm this point: the architectural complex in Mogao Grotto 257 embodies the unique style of integrating the stupa and temple in early Buddhist monasteries[3]. This design created a novel spatial experience. It fully demonstrates the ingenuity of ancient architects[4].

Furthermore, in the *Dunhuang Murals of the Tang Dynasty*, the decorative function of ornamentation developed significantly. It reached an extremely high level of artistry[5]. As shown in Figs. 1 and 2, the decorative bands on the niche walls and the aureole of the Buddha statue together symbolize divine radiance. The aureole of the Buddha is composed of a

[1]School of Art, Northwest University, Xi'an, China. [2] School of Electronic Information, Northwest University, Xi'an, China. [3]State-Province Joint Engineering and Research Center of Advanced Networking and Intelligent Information Services, Northwest University, Xi'an, China. [4]Shaanxi Silk Road Cultural Heritage Digital Protection and Inheritance Collaborative Innovation Center, Xi'an, China. [5]Xi'an Museum, Xi'an, China. [6]Shaanxi Key Laboratory of Higher Education Institution of Generative Artificial Intelligence and Mixed Reality, Xi'an, China. ✉e-mail: huqiyao@nwu.edu.cn

Fig. 1 | The nine-colored deer rescuing a drowning man. Northern Wei Dynasty, Mural from Cave 257 of the Mogao Caves. In this Jataka narrative, the Nine-Colored Deer, embodying beauty and compassion, selflessly rescues a man who has fallen into the river. Confronting the torrent and hidden whirlpools without hesitation, the deer leaps into the water and, guided by its pure and benevolent nature, carries the drowning man to safety.

Fig. 2 | Architectural elements in the Maitreya Sutra transformation. Tang Dynasty, Yulin Cave 25, North Wall. This scene illustrates Maitreya's descent and the three Dharma assemblies, emphasizing architectural details. The central assembly features a gem-canopied lectern supported by a two-tiered structure with a rotating pedestal. In the surrounding assemblies, incense altars and palatial buildings are framed by heavenly pavilions and divine treasures. The architecture reflects the idealized order of Maitreya's Pure Land.

single flame pattern. It has varying complexity in each layer, showcasing exquisite craftsmanship[6]. The murals also demonstrate superb skill in creating a sense of space and three dimensional 3D effect. The landscape painting on the west side of the south wall in Grotto 217 is an excellent example. It vividly depicts a scene of springtime travelers heading to a fortress in the Western Regions. It uses 3D elements like overlapping mountains and cliffs, waterfalls, and blooming peach and plum blossoms to create a profound sense of depth[7]. Similarly, the depiction of the Pure Land on the south wall of Grotto 220 presents rich spatial layers and a vivid 3D quality. It does this through portrayal of the Three Holy Ones of the West and their retinue, combined with grand scenes such as the jeweled pond and musical performances[8]. Finally, the world renowned music and dance scenes in the Dunhuang murals are also a concentrated manifestation of their artistic charm.

However, this magnificent palace of art and civilization, which carries a thousand years of history, is constantly subjected to the combined effects of natural weathering, erosion, and human activities. It faces increasingly severe risks of damage. The fading, flaking, and deterioration of the murals mean that valuable historical and cultural information is gradually being lost. Therefore, it is imperative to initiate a comprehensive and in depth digital preservation process for this precious heritage site. In recent years, with the rapid development of information technology, how to deeply integrate traditional cultural heritage with modern science and technology has become an important research area[9,10]. The cultural preservation of the Mogao Grottoes has become a research hotspot in this context[11].

Traditional preservation methods often prove inadequate for complex demands of restoring and displaying cultural artifacts. Image based 3D reconstruction technology provides a revolutionary solution. This technology utilizes neural network algorithms[12] and artificial intelligence powerful generative capabilities[13]. It accurately converts static two dimensional image information into interactive, high fidelity 3D models. Its core advantage is that the entire process requires no physical interaction with fragile artifacts[14]. Thus, it fundamentally avoids risk of secondary damage. This provides researchers with extremely realistic digital models. It offers

crucial technical support for in depth studies of the murals. In this way, many precious cultural artifacts at risk of destruction due to natural aging, environmental changes, or human caused damage[15] can achieve digital perpetual preservation.

With advancement of modern digital technologies, ancient Dunhuang architecture images can be accurately restored. This allows these precious historical artifacts to radiate new vitality in the modern era[16,17]. In terms of historical value, the *Dunhuang Mural* are not only an artistic heritage but also documentary materials of significant historical value[18]. As a hub on the ancient Silk Road, the murals of the Mogao Grottoes deeply integrate Central Plains culture with Western cultures. They vividly embody intersection and collision of diverse cultures.

However, this magnificent palace of art and civilization, which carries a thousand years of history, is constantly subjected to the combined effects of natural weathering, erosion, and human activities. It faces increasingly severe risks of damage. The fading, flaking, and deterioration of the murals mean that valuable historical and cultural information is gradually being lost. Therefore, it is imperative to initiate a comprehensive and in depth digital preservation process for this precious heritage site. In recent years, with the rapid development of information technology, how to deeply integrate traditional cultural heritage with modern science and technology has become an important research area[9,10]. This integration benefits from cutting-edge advancements in the field of computational intelligence, including multiscale generative adversarial learning in complex system control[19], adaptive multichannel graph recurrent networks for spatio-temporal fusion[20], and robust federated learning of large language models in challenging environments[21]. These technologies collectively drive a trend across diverse domains, including the crucial area of cultural heritage preservation, toward utilizing complex algorithms to address intricate problems. The cultural preservation of the Mogao Grottoes has become a research hotspot in this context[11].

Recently, the emergence of Neural Radiance Fields NeRF, a neural network technique for synthesizing new perspectives of images, has opened up new possibilities for digital rebirth of murals. NeRF learns the 3D representation of a scene. It generates high quality 3D model from multiple 2D images by learning radiance fields of light in space.

Thanks to large 3D datasets such as Objaverse, Objaverse-xl, and OmniObject3D, recent datasets like ShapeNet and Geometric offer improved model size and quality. Many methods based on these datasets show strong potential in generation speed and quality. Dreamfusion uses Score Distillation Sampling SDS to control diffusion. It optimizes NeRF for 2D to 3D upscaling. This approach can still suffer from 3D consistency issues. To address this, improved novel view synthesis models exist. Examples include Zero1to3, Zero123plus, InstantMesh, and mvdream. These models enhance 3D consistency. They use strategies like improved viewpoint interpolation, depth guidance, and regularization. Yet, these methods often achieve high quality generation in only a small number of cases.

Traditional NeRF algorithms are renowned for high fidelity image synthesis. They typically require many images for training. However, their ability to generate continuous views is limited with sparse input views. To address this, researchers have proposed innovative methods. They combine these methods with generation of artistic images from sparse views. This broadens NeRF applications.

SparseNeRF[22] enhances performance with limited views. It optimizes sparse input strategies. It minimizes the need for numerous input images through feature extraction and view interpolation techniques. This enables generation of high quality views even with sparse artistic images. Depth guided robust fast point cloud fusion NeRF[23] incorporates depth information. It improves view synthesis accuracy using point cloud data. This method more effectively captures scene geometry under sparse view conditions. It results in more realistic and detailed synthesis of artistic images.

RegNeRF[24] enhances model stability with sparse data. It uses regularization techniques. It effectively models relationships between views. It generates smooth and continuous, art like images even with fewer inputs. By integrating these methods, NeRF not only improves performance under sparse view conditions. It also creates new perspective images with an artistic style. These advancements have significant implications for 3D reconstruction, virtual reality, and cultural heritage creation. They make NeRF technology more flexible and efficient for various innovative applications.

In the Mogao Caves at Dunhuang, notable examples include the Kabuki Bodhisattva in Cave 148, the Bouncing Pipa Dance in Cave 112, and the Karenga Kabuki in Cave 172, a variation of the Guanmiguanshou Sutra. NeRF[25] can now be used to restore these music and dance scenes from the *Dunhuang Mural*. It presents fantastical scenes in 3D. NeRF[26] allows us to view restored scenes from any 360 degree angle.

The *Dunhuang Mural* music and dance images[27] feature beautiful gestures. They create a fantastic and magnificent scene. Different emotional intentions are conveyed through various gestures. NeRF can achieve high precision detail restoration in 3D rendering. It accurately reproduces intricate details. Additionally, NeRF can render images from any 360 degree. It brings unreal scenes depicted in the murals to life.

In the *Dunhuang Mural*, use of color is crucial[28]. Pigments, made from minerals such as ochre and cinnabar, create a natural harmony. They present a unique Dunhuang palette that is pure yet not garish, muted yet not dull. Additionally, colors of the murals change with variations in light. NeRF allows us to examine impact of light on visual effects. It integrates different lighting models. It employs deep learning to learn color distribution and lighting conditions of a scene. By training the network, NeRF can generate high quality color details. It reproduces complex lighting effects.

While NeRF offers promising avenues for digital heritage, applying it to artistic images like Dunhuang murals still faces distinct difficulties. These problems limit the wider use of 3D reconstruction algorithms in cultural heritage. They need urgent research and exploration.

One major obstacle is the lack of 3D reconstruction datasets for artistic images. Existing datasets focus on real world objects. This makes it hard to reconstruct cultural heritage images. Heritage images have unique styles, complex compositions, and diverse representations. Specialized datasets are needed. They will better capture and understand rich visual features of artworks.

Second, cultural heritage images face the Janus problem in 3D reconstruction. This means images have multiple meanings. It leads to inconsistent or incorrect results. Abstract characteristics and diverse styles worsen this problem. This increases reconstruction difficulty.

Third, art image subjectivity challenges 3D reconstruction algorithms. Current research lacks targeted innovation. It needs to address unique characteristics of heritage images. This includes capturing meaningful color and morphological features. So, 3D reconstruction often fails. It cannot accurately reproduce color gradients or morphological details of original artworks.

Building on theoretical foundation and research progress, we selected a combination of diffusion and NeRF. This extracts 3D information of Chinese frescoes. We performed appropriate preprocessing on a complete fresco to meet input requirements of the network. Classical 3D representations are explicit, such as point clouds, voxels, and meshes. We extracted mesh representations to broaden application scenarios by rendering new perspectives with NeRF. This process completed 3D reconstruction of Chinese mural elements.

Our core contribution lies not in the invention of novel foundational algorithms, but in the innovative and effective integration and adaptation of existing advanced techniques to address the unique and challenging task of single-view three dimensional reconstruction of Dunhuang mural elements. This specific problem within cultural heritage digitization demands a delicate balance between computational efficiency, artistic fidelity, and the practical usability of the resulting models. Specifically, Dunhuang murals present distinct obstacles that traditional reconstruction methods struggle to overcome, including non-photorealistic perspectives, weak texture information caused by aging pigments and flat painting styles, and complex thin-wall structures found in elements like floating ribbons or halos.

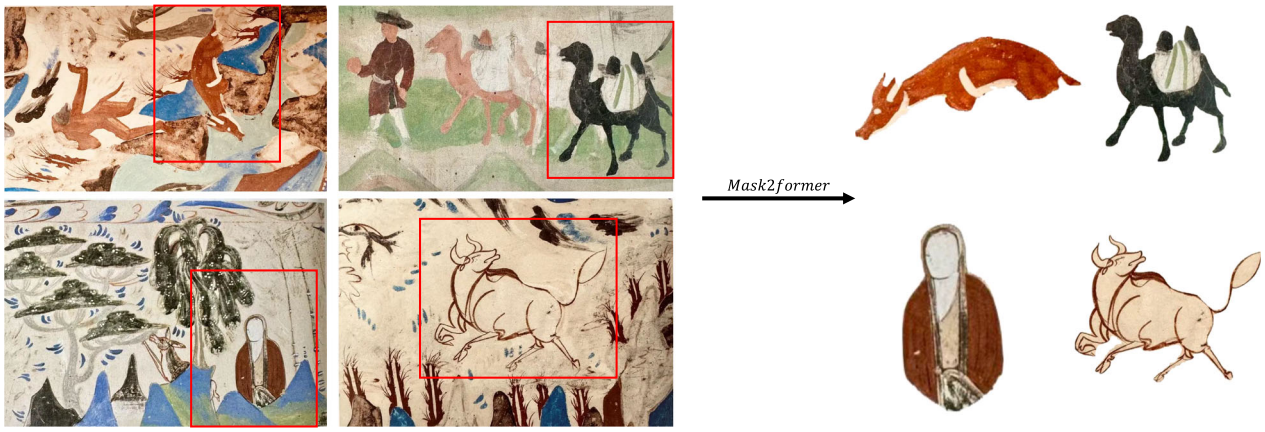

Fig. 3 | Data preprocessing and mural element separation. Example from the CME dataset, showing data preprocessing and using Mask2Former to crop and separate mural elements.

The rationale behind our selected components and their combined workflow is designed to explicitly tackle these mural-specific challenges. First, to address the lack of three dimensional information and the non-photorealistic nature of the input, we employ the Perspective-Driven Synthesis PDS Module utilizing Zero123plus. Traditional photogrammetry fails with single views, and standard generative models often force artistic inputs into strictly physical geometries, losing their charm. PDS is chosen because its diffusion priors allow for the hallucination of geometrically consistent multi-view images that respect the stylized, non-rigid perspective of the murals. This module effectively resolves the data scarcity issue by synthesizing a robust sparse view set that maintains artistic logical consistency.

Second, to overcome the challenge of weak textures and prevent geometric collapse in flat-colored regions, we utilize the Neural Rendering Synthesizer NRS powered by ZeroRF. Murals often lack the high-frequency texture details required for traditional feature matching, leading to holes or noise in standard reconstruction. ZeroRF is selected for its deep image prior mechanism, which parameterizes feature grids via a neural network. This provides intrinsic regularization, allowing the model to reconstruct smooth and coherent geometry even in regions with weak texture or subtle inconsistencies in the synthetic views, thereby avoiding the floating artifacts common in sparse-view NeRF training.

Third, to capture intricate geometric details like thin-wall structures and ensure the model is usable, we integrate the Light Field Fusion Meshing LFFM Module based on NeRF2mesh. Implicit representations are difficult to edit or render in standard engines. NeRF2mesh is essential here not just for conversion, but for its adaptive refinement capability. By iteratively adjusting vertex density based on rendering errors, it can allocate more geometric resources to complex boundaries. This allows our framework to preserve delicate thin-wall structures that might otherwise be over-smoothed or lost during coarse mesh extraction, balancing detail preservation with mesh lightweighting.

In summary, 3DSynBrush represents a synergistic divide-and-conquer pipeline. PDS handles the artistic perspective synthesis. NRS solves the geometry estimation under weak texture conditions. LFFM ensures the preservation of fine structural details and model usability. This integrated strategy effectively addresses the specific constraints of mural digitization, offering a generalizable framework for transforming static cultural heritage into interactive digital assets.

Methods

Dataset

To facilitate the training and evaluation of our framework, we constructed a high-quality dataset, termed the **C**hinese **M**ural **E**lements (**CME**) dataset. The raw images in this dataset were primarily collected from the "Digital Dunhuang" Open Material Library and other publicly available digital mural archives. The selected elements were meticulously screened and categorized to ensure diversity in artistic style and representativeness of structural complexity. As shown in Fig. 3, the CME dataset comprises over 2000 high-resolution images of individual mural elements extracted from traditional Chinese wall paintings.

Given that current 3D reconstruction and view synthesis algorithms often impose constraints on input resolution (typically 512 × 512), we preprocessed the data to ensure compatibility. Specifically, we resized and cropped the mural elements into fixed-size frames while preserving their structural integrity. Furthermore, to eliminate the interference of complex mural backgrounds which are unsuitable for element-level analysis, we employed Mask2Former[29] to perform semantic segmentation. This process isolated each mural element, resulting in images with transparent backgrounds (RGBA format), which are directly compatible with downstream 3D reconstruction pipelines.

$$I_{out} = M(I_{in})$$
$$M \in \{Resize, Divide, Mask2Former\} \quad (1)$$

To ensure wide coverage, the CME dataset includes mural elements from different dynasties and regional schools. We categorize them into three primary classes: *Character*, *Animal*, and *Plant*. Each major category contains hundreds to thousands of elements, providing sufficient data for both recognition and reconstruction tasks.

The CME dataset aims to digitally preserve the artistic essence of Chinese mural elements. Each element carries cultural symbolism and reflects the aesthetic principles of ancient mural art. By systematically selecting, segmenting, and curating these elements, we intend to both preserve their authenticity and enable their application in modern digital environments.

The diversity of CME lies in its rich stylistic variations and its coverage of multiple dynasties, illustrating the artistic evolution of Chinese murals and emphasizing cultural continuity across history.

Multi-view image generation

Multi-view images provide rich spatial and depth cues, playing an indispensable role in 3D reconstruction[30]. In our technical framework, the Perspective-Driven Synthesis Module generates geometrically consistent multi-view projections through a viewpoint-conditioned diffusion process, thereby laying the foundation for high-fidelity 3D reconstruction. The overall pipeline is illustrated in Fig. 4.

(a) Perspective-Driven Synthesis (PDS) Module. This module employs a viewpoint-conditioned diffusion model[31] to synthesize six sparse views from a single input mural element. This generation occurs from standardized, predefined camera poses inherent to the model's training, rather than by dynamically computing unique camera poses for each input. These poses

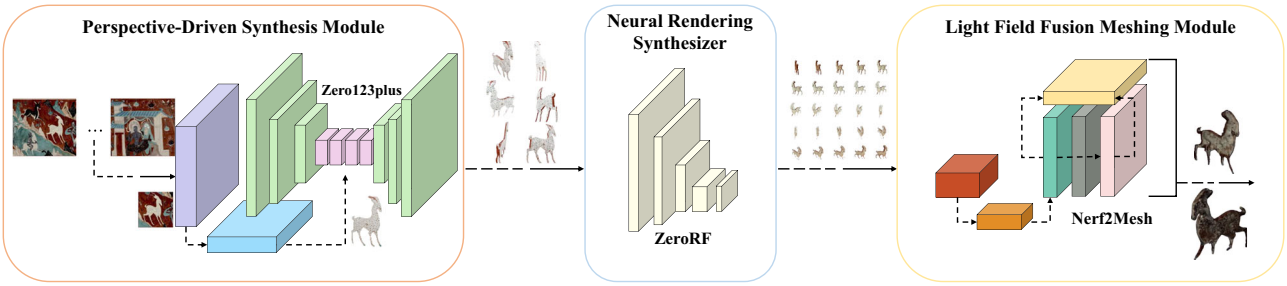

Fig. 4 | Overall framework structure. In the PDS model, we obtain six viewpoints of the image through a number of strategies. In the Neural Rendering Synthesizer, we use NeRF to predict continuous viewpoints. In the Light Field Fusion Meshing Module, we use the continuous viewpoints to generate a 3D mesh.

provide comprehensive spatial coverage around the object. The azimuth, elevation, and distance for these views are intrinsic parameters of the Zero123plus model design, engineered for robust 3D reconstruction. Our PDS module thus precisely leverages this inherent capability to generate these standardized views. By incorporating the joint probability distribution of different viewpoints during the diffusion process, Zero123plus simultaneously produces a mosaic of multi-angle projections. This design notably enhances 3D consistency over prior methods such as Zero1to3, a critical factor for accurate neural radiance field training and mesh generation, especially for mural artworks. While direct experimental data optimizing this specific view configuration for murals is not provided, the robust qualitative and quantitative results of our framework empirically validate its efficacy for this application.

(b) Neural Rendering Synthesizer. This component adopts the parameter-efficient paradigm of ZeroRF[25], constructing neural radiance fields via hash-encoded volumetric rendering. Its hybrid optimization strategy, compared with the original NeRF, significantly enhances generalization to non-photorealistic artistic scenes, thereby achieving more faithful scene representations.

(c) Light Field Fusion Meshing Module. At this stage, light field data in LLFF format[32] are incorporated as input for the reconstruction algorithm. The algorithm leverages NeRF2Mesh[33] to convert implicit neural fields into explicit mesh surfaces. After generating an initial coarse mesh via the Marching Cubes algorithm, the mesh is further refined using NeRF. Simultaneously, texture refinement under optimized lighting conditions enhances the realism of the reconstructed model, effectively aligning with the visual characteristics of mural elements in real-world contexts.

In the application of synthesizing 3D perspectives with diffusion models, the challenge lies in the lack of 3D consistency in the generated images. This means that the new perspectives generated by the diffusion model for the same object differ in detail. There are two approaches to solving this problem. The first is to use large-scale 3D data for training, which requires extremely complex networks and large-scale computational resources. The other approach, adopted by Zero123++[31], involves modifying the structure of the diffusion model to generate new perspectives in a one-to-many manner, significantly reducing the demand for training resources. This method can significantly improve the 3D consistency of the results, although it relatively reduces the resolution of each output. In the context of 3D reconstruction of Chinese paintings, stronger 3D consistency is crucial for both the effectiveness and the success rate of the reconstruction.

Based on these findings, we use the Zero123++ pre-trained model[31] to synthesize new perspectives from segmented images of Chinese painting elements. Previous work has highlighted the challenges in achieving 3D consistency in diffuse perspectives. Zero-1-to-3[30] cannot directly generate 360° views suitable for 3D reconstruction. This significantly improves the 3D consistency of the generated results, thereby increasing the reconstruction success rate. Luckily, Zero123++ fine-tunes the diffusion model to generate six multi-views by considering the joint probability density. This makes it possible to generate consistent multi-view images as input for NeRF training.

This strong 3D consistency stems from the model training on large-scale 3D datasets, enabling it to learn and predict geometric relationships between different viewpoints. Although the model strives to generate consistent views, structural distortions or semantic errors can still arise when processing non-photorealistic artistic images such as Dunhuang murals. The model might interpret two-dimensional paintings as possessing 3D depth, leading to discrepancies with the original artistic style. Furthermore, the limited output resolution can result in the loss of fine details within the murals. While pursuing geometric consistency, the model may regularize the artistic expressions of the murals to conform more closely to real-world geometric rules, which could lead to a distortion of certain artistic intentions. Our objective is to obtain a geometrically plausible interpretation as the basis for 3D reconstruction, rather than perfectly preserving every stylistic nuance of the 2D artwork.

In the next step, the generated sparse views are fed into Neural Rendering Synthesizer is designed to use NeRF to generate continuous views between sparse views, resulting in 360° video output. Recently, the emergence of Neural Radiance Fields, a neural network technique for synthesizing new perspectives of images, has opened up new possibilities for the digital rebirth of murals. NeRF learns the 3D representation of a scene and generates high-quality 3D model from multiple 2D images by learning the radiance fields of light in space.

The core idea of NeRF is to represent a continuous 3D scene by means of a multilayer perceptron (MLP). NeRF aims at predicting the colour and density of each point in the space. Specifically, NeRF takes as input the 3D coordinates (x, y, z) and the viewing direction (θ, ϕ) and outputs the colour (r, g, b) and density σ. The input to the neural network is a five-dimensional vector, and the outputs are colour and density: $F_\Theta(x, y, z, \theta, \phi) \rightarrow (r, g, b, \sigma)$ where θ denotes the parameters of the network. The predicted colours and densities are incorporated into the final image using a volume rendering technique. The formula is:

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

where $T(t)$ is the cumulative transmittance from the starting point of the ray to t, $\sigma(t)$ is the density, and $c(t)$ is the colour, NeRF is able to generate high-quality 3D model with complex light and shadow variations, suitable for 3D reconstruction of a wide range of scenes and ideal for digital preservation of cultural relics. When subtle geometric contradictions exist among the input generated views, NeRF tends to learn a compromised 3D model rather than amplifying these contradictions. This compromise may manifest as geometric blurring, floating artifacts, or an averaging of color and density fields. Our choice of ZeroRF is critical due to its enhanced capabilities in handling such challenges. ZeroRF employs hash-encoded volumetric rendering and a hybrid optimization strategy, significantly improving its generalization ability to non-photorealistic artistic scenes. Its predefined depth prior is particularly instrumental in mitigating the impact of missing data caused by geometric contradictions between generated views. This prior guides NeRF's learning process, enabling it to produce a more stable and reasonable compromised 3D model when faced with inconsistent

regions, rather than exacerbating the inconsistencies. ZeroRF's robustness allows it to converge to a plausible 3D representation even with slight inconsistencies, though severe or persistent inter-view contradictions can still degrade the overall reconstruction quality.

Based on this, ZeroRF[25] uses predefined Gaussian noise as input to the deep generative network. It outputs planar and vector features in a Tensor-VM, describing the properties of the 3D scene. For state-of-the-art images, the predefined depth prior effectively mitigates the impact of missing state-of-the-art image data on the reconstruction results. The generated results also reduce the occurrence of multi-faceted artifacts. The decomposed Tensor-VM improves computational efficiency while maintaining high-quality reconstruction. Six multi-view images are processed for NeRF representation. The generation from sparse perspective to continuous perspective is achieved. The new view images are uniformly rendered over 360 degrees, generating 100 frames of video. This video is used as input for mesh representation in subsequent steps.

There are fundamental differences between ZeroRF and traditional NeRF training based on real photos. The core idea of ZeroRF is to solve the challenge of 3D reconstruction in sparse views by integrating customized depth image priors. ZeroRF parameterizes the feature grid through a neural network generator that takes frozen standard Gaussian noise samples as input. This design utilizes the high impedance of neural networks to noise and artifacts, enabling them to better generalize under sparse supervision. The standard volume rendering process and simple mean square error loss drive training, but the key lies in the parameterization of the feature grid, which enables NeRF to learn stable and coherent 3D scene representations when facing input images generated by PDS modules that may contain minor inconsistencies.

Unlike methods that rely on explicit regularization or external pre training modules to address sparsity or noise, ZeroRF achieves this without any pre training or additional regularization. The optimization process of NeRF relies on the smoothing ability of its multi-layer perceptron and the modeling of three-dimensional spatial continuity to find a single, continuous three-dimensional geometry and radiation field that best explains all views. In this case, the architecture design of ZeroRF provides an implicit regularization of input noise, tending to produce smoother and more consistent representations. The PDS module itself significantly improves the 3D consistency of generating multi view images and reduces input noise levels from the source by utilizing Zero123plus compared to earlier models. For artistic murals, the concept of geometric "reality" usually corresponds to artistic intent rather than strict physical accuracy. Therefore, NeRF aims to learn a three-dimensional representation that faithfully reflects the artistic and geometric features conveyed by the views generated by PDS modules, rather than enforcing a purely physical world geometry. The qualitative and quantitative results presented in our study provide empirical validation for the effectiveness and artistic rationality of this method in reconstructing mural elements.

Through the collaboration of these modules, we successfully transform mural images lacking prior data into high-quality continuous-view images that conform to realistic logic.

High-quality 3D reconstruction

Since NeRF is not commonly used for displaying 3D representations, we use NeRF2mesh[33] to synthesize explicit mesh representations based on multi-view information. The VT-LLFF[32] module converts the multiview output into a dataset format suitable for LLFF. LLFF[32] stores point clouds, poses, sparse models, and other data in a simple file format that can be easily recognized by the computer. By inputting the LLFF file, we ensure that Python understands our input. Colmap[34] is used to obtain the camera pose and sparse point cloud, formatting the data as an LLFF dataset. In the previous step, 100 multi-view images with consistent 3D appearance were generated for 3D reconstruction. We choose the NeRF2mesh[33] algorithm for mesh reconstruction.

NeRF2mesh[33] optimizes the appearance and texture of geometric models to accurately recover the colors of Chinese paintings. The NeRF2mesh structure involves a two-stage mesh refinement process, commencing with geometry learning through a density field.

In the first stage, a coarse mesh is created by extracting an initial mesh from NeRF's density field using Marching Cubes. An empirically determined fixed isosurface density threshold is employed for this coarse mesh extraction. For subjects like Dunhuang mural elements, which often possess soft, non-rigid boundaries, a fixed threshold inherently involves a trade-off between preserving fine details and preventing spurious holes or artifacts. However, the primary objective of this initial stage is to establish a topologically accurate foundational geometry.

The more critical refinement is subsequently carried out in the second stage, where the mesh is refined to enhance appearance and detail. This stage leverages an iterative surface refinement algorithm that adaptively adjusts the vertex positions and face density of the mesh based on reprojected 2D rendering errors. Regions exhibiting larger rendering errors are subdivided to capture more detail, while those with smaller errors are simplified, or decimated, to maintain conciseness. This adaptive mechanism enables our method to effectively handle the diverse stylistic variations of mural elements, striking a balance between detail preservation and geometric integrity, without relying on a single fixed threshold across the dataset.

Notably, the mesh generated by our method utilizes approximately 40% of the vertices compared to traditional methods, achieving significant lightweighting. This lightweighting is an inherent characteristic of the iterative refinement process of the NeRF2Mesh algorithm, rather than a simple post-processing mesh simplification algorithm. By dynamically allocating geometric resources and concentrating limited computational and mesh representation capacity on regions critical for visual quality and geometric accuracy, this effectively reduces the overall number of vertices and faces while preserving crucial details, thereby achieving an excellent balance between efficiency and detail capture.

The texture representation is learned and decomposed into diffuse and specular components. To clarify the origin and refinement mechanism of the final mesh texture, we state that it is derived directly from the color field learned by NeRF2mesh, rather than through direct projection texture mapping from original input images or generated multi-view images. The NeRF2mesh framework decomposes the color field into a viewpoint-independent diffuse color and a viewpoint-dependent specular component, a process that commences in the model's first stage.

The texture refinement discussed in the framework is an iterative optimization process conducted jointly with geometric refinement. The model achieves texture refinement by optimizing parameters of its three dimensional color field and minimizing photometric error between rendered images and training views. This embeds the optimization of texture color within NeRF's intrinsic color prediction mechanism, thereby naturally addressing potential texture color inconsistencies in multi-view inputs. The strength of NeRF lies in its ability to learn a consistent three dimensional radiance field from multi-view inputs.

Regarding the optimization of illumination and color, NeRF2mesh handles this by decomposing appearance into diffuse and specular terms. The diffuse color can be directly baked into standard RGB image textures. The specular reflection serves as an intermediate feature, generating viewpoint-dependent specular color through a small multilayer perceptron combined with viewpoint direction. This decomposition allows the model to implicitly optimize lighting effects by learning surface responses under different lighting conditions, without explicitly estimating ambient illumination.

Finally, upon model convergence, the fine mesh is exported with expanded UV coordinates. NeRF2mesh then bakes these decomposed textures, specifically the diffuse color and specular features, into two dimensional images via UV unwrapping for subsequent real-time rendering and editing.

Currently, our framework does not explicitly incorporate prior artistic knowledge specific to Dunhuang murals or Chinese painting. This means the model primarily relies on its general 3D understanding to reconstruct the 3D structure of the murals.

However, this does not mean the model will simply guess or unintentionally correct deliberately non-physical perspective effects in the murals. We employ the following mechanisms to ensure the model faithfully reproduces the artistic language of the murals:

First, the PDS module is crucial. We choose the Zero123plus model, whose training data and generation mechanism allow it to maintain the style and relative geometric relationships of the input images when generating multiple perspectives from a single image. For artworks like murals, which inherently contain stylized perspective or non-physical geometry, Zero123plus tends to generate multi-view images that maintain consistency in these stylized features. It does not attempt to force these artistic features into strict physical perspective.

Second, ZeroRF and its underlying NeRF learn radiation fields. During training, NeRF strives to find a 3D representation that fits all input 2D views to the greatest extent possible. If the views generated by the PDS module inherently possess the non-physical perspective effects unique to murals, then NeRF aims to learn a 3D field capable of faithfully rendering these views. In other words, NeRF learns the 3D "artistic logic" embodied in the input views, rather than a set of strict physical geometric laws. This allows the model to capture and reproduce the unique geometric and spatial expressions of the murals, rather than forcibly "correcting" them.

Therefore, our model aims to faithfully reproduce the artistic 3D features embodied in the input views, rather than achieving strict geometric accuracy in the physical world. It constructs a 3D representation that is logically coherent by learning the artistic patterns presented in the input data. Nevertheless, we acknowledge that this is an important direction for future research: exploring how to explicitly integrate the unique artistic a priori knowledge of Dunhuang murals extracted from art studies into the 3D reconstruction framework to achieve a deeper understanding and reconstruction of artistic language.

Through these steps, we have achieved high-quality 3D reconstruction using cross-domain information fusion, successfully representing the unique artistic style of Chinese paintings in digital form.

This work takes a single mural element as input and employs a four-stage pipeline to transform it from 2D to 3D: The first stage uses Mask2Former in conjunction with our constructed CME dataset for semantic segmentation and thin structure enhancement to obtain clear element masks and boundary priors; the second stage generates sparse multi-view images through perspective-driven synthetic PDS to supplement the viewpoint geometric cues; the third stage uses Zero123++ to expand the sparse viewpoints into a consistent multi-view image sequence and densifies it into a continuous neural radiation field by ZeroRF; the fourth stage uses mesh generation based on light field fusion and iterative surface refinement to convert the continuous representation into an explicit mesh and perform texture reprojection to preserve the original tones.

Digital Restoration and Virtual Reality

Combining 3DSynBrush with traditional Chinese paintings creates a unique and profound cultural heritage experience. Through 3D reconstruction, viewers can appreciate the details and layers of Chinese paintings from any 360°, experiencing the deep cultural heritage within the works. This technology allows the composition, color, and mood of Chinese paintings to be intuitively understood from various perspectives.

VR technology offers audiences an immersive experience. 3DSynBrush provides high-quality 3D model for VR, helping to construct the unreal scenes depicted in the *Dunhuang Mural*. This allows people to vividly recreate historical scenes and interact with their cultural heritage.

In digital heritage restoration, 3DSynBrush aids in restoring damaged artifacts. Its ability to accurately predict 3D reconstruction of invisible perspectives through diffusion modeling allows experts to synthesize information about these hidden areas. This data provides a scientific basis for restoration work, enabling it to be conducted according to established benchmarks.

Results

Visual analysis

The generated six views and multiple views exhibit strong 3D consistency, effectively training NeRF to synthesize new views. The results are largely consistent with the original input images. The model adheres to real-world a priori rules, demonstrating a high degree of reasonableness. The generated texture images are finely detailed with smooth color transitions, closely matching the original images. The model is visually balanced, with well-proportioned details, and the gloss and reflection of materials are handled appropriately.

NeRF training

The output of ZeroRF[25] is shown in Fig. 5 for the input of six views. These figures depict the rendering results of ZeroRF, categorized into NeRF effects for multi-view images of quadrupeds and NeRF generation for multi-view images of human figures. As observed, the results demonstrate excellent visual quality and robustness, with minimal color variations and almost no walls or floating objects. The basic reconstruction of the 3D images is both realistic and logical.

Mesh generation

The NeRF2mesh[33] method excels in complex topological scenarios, producing high-quality surface meshes that render texture images more compact and intuitive than those generated by other methods. The mesh visualization is shown in Figs. 6 and 7.

We conduct both qualitative and quantitative comparisons against various state-of-the-art (SOTA) single-view mesh generation methods. These comparisons further demonstrate that our framework is more adept at reconstructing non-photorealistic artistic images, while simultaneously avoiding common pitfalls encountered during generalization.

In the qualitative analysis, LN3Diff[35] successfully reconstructs a plausible 3D structure but fails to preserve the overall shape fidelity. Methods such as DreamGaussian[36], TripoSR[37], and Magic123[38] focus exclusively on the frontal view during training, resulting in generated meshes that lack reasonable 3D consistency. Furthermore, due to the scarcity of mural content in public datasets, their reconstructed meshes often resemble other similar objects rather than being generated based on real-world priors. In contrast, our method achieves a comprehensively superior visual effect in terms of plausible 3D structure, overall shape, and texture appearance.

Regarding quantitative metrics, as shown in Table 1, for object (a), our method utilizes only 718 faces, a significant reduction of 98.56% compared to Magic123 50,000, which has the lowest face count among the other methods. Our vertex count is 352, representing a substantial 98.59% decrease from Magic123, the lowest among the competitors. Our method achieves the highest SSIM score of 0.957, an improvement of 1.06% over the runner-up, LN3Diff. The PSNR value reaches 22.229, which is 6.96% higher than the second-best DreamGaussian.

As shown in Table 1, for object (b), our method similarly outperforms across all metrics, achieving a dual optimization of model complexity and generation quality. It uses 21,875 faces, a 12.29% reduction compared to DreamGaussian, the most efficient alternative. The vertex count is 10,928, which is 24.93% lower than that of DreamGaussian 14,558. The SSIM of our method is 8.29% higher than the next best, TripoSR. Moreover, its PSNR of 19.413 marks a significant 14.83% improvement over the second-highest, TripoSR 16.906.

The reduction in model complexity is quantified by comparing our method face and vertex counts against the most efficient competing methods for each object. For faces, our method uses approximately 1.44 percent of Magic123 count for object a and about 87.71 percent of DreamGaussian count for object b. Averaging these usage percentages for faces yields approximately 44.58 percent. Similarly for vertices, our method uses approximately 1.41 percent of Magic123 count for object a and about 75.07 percent of DreamGaussian count for object b. Averaging these usage percentages for vertices yields approximately 38.24 percent.

Fig. 5 | Outputs of the PDS module and neural rendering synthesizer. Sparse view images are first generated based on the segmented extracted single-element images. Subsequently, these sparse view images are used to predict continuous view images between the sparse viewpoints.

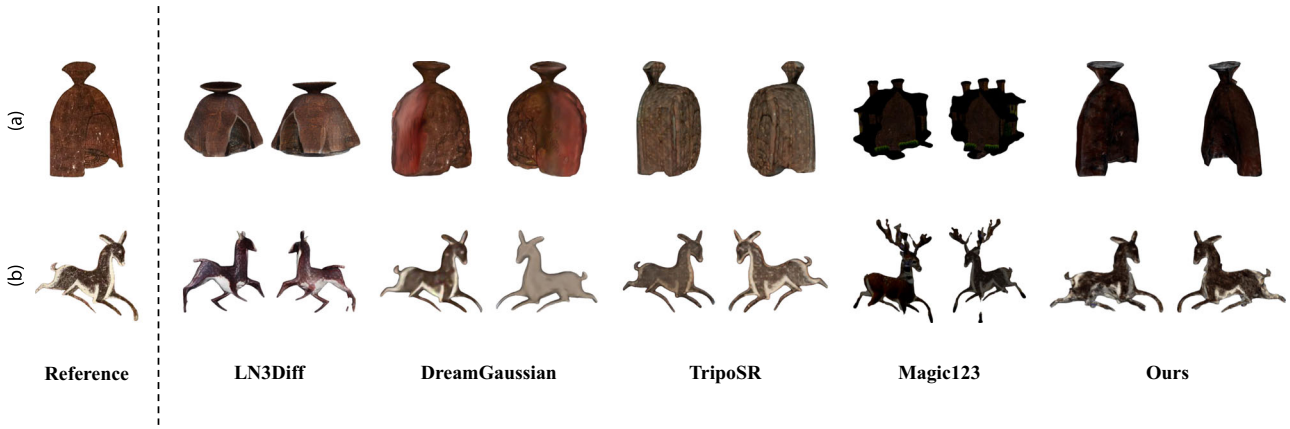

Fig. 6 | Qualitative comparison with state-of-the-art (SOTA) approaches. The leftmost column displays the input single-view image. The results from our method (Ours), alongside those from LN3Diff, DreamGaussian, TripoSR, and Magic123, are shown on the right. Our method not only generates 3D shapes with plausible geometry but also recovers the high-fidelity textured appearance of the model under realistic lighting conditions.

Therefore, on average, our method requires only about 40 percent of the vertices and about 45 percent of the triangles compared to the most efficient baselines.

The experimental results confirm the superior suitability of our method for reconstructing artistic works, as exemplified by murals. For both objects (a) and (b), our method utilizes substantially lower face and vertex counts than all competing methods, demonstrating that we achieve comparable visual quality with significantly fewer resources. Concurrently, it generates 3D models with higher SSIM and PSNR scores with respect to the input view. This indicates that our method possesses significant advantages in both efficiency and performance.

Robustness analysis

To comprehensively evaluate the reliability of the 3DSynBrush framework in practical applications, its robustness was tested under various adverse conditions. These conditions included Gaussian noise, variations in illumination intensity, and image occlusion. The impact of these adverse conditions on mesh shape was assessed by calculating the Chamfer Distance, hereafter referred to as CD, and the effect on texture quality was measured using the Structural Similarity Index, hereafter referred to as SSIM.

The robustness of 3DSynBrush was first evaluated under varying levels of Gaussian noise. Figure 8 illustrates the progressive introduction of

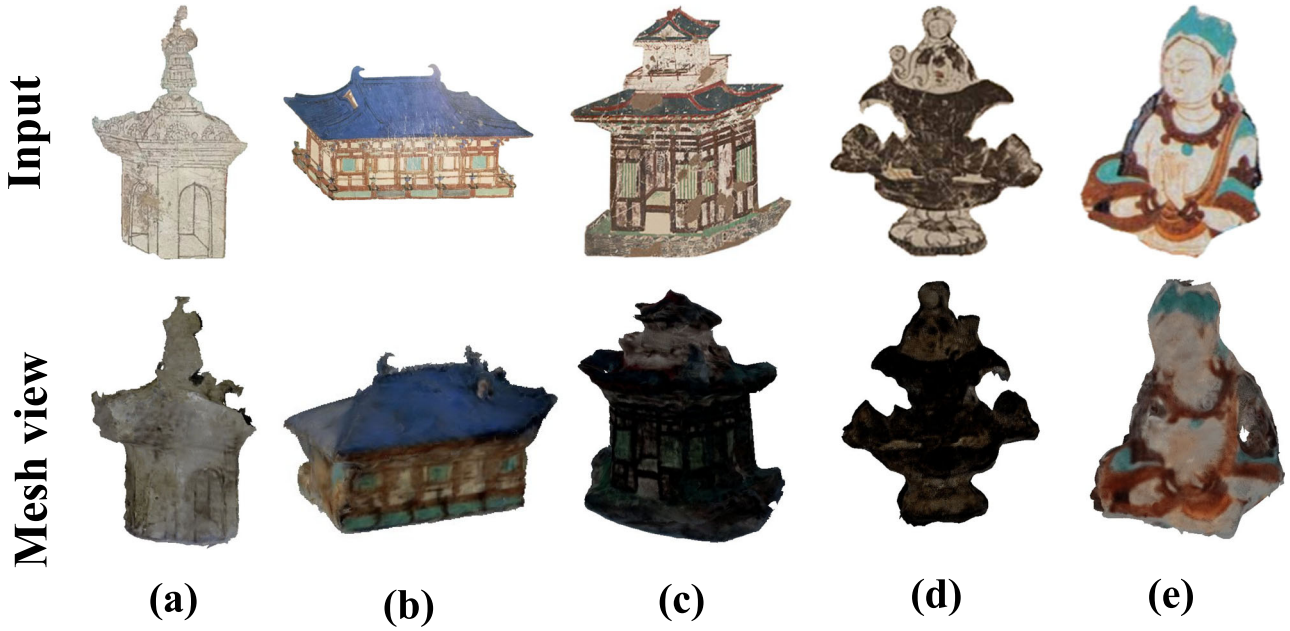

Fig. 7 | Supplementary experimental results. This figure demonstrates additional qualitative evaluations validating the robustness and visual fidelity of our proposed framework.

Table 1 | Comparison results of 3DSynBrush and other methods on Number of Vertices and Faces

Method	Object	Number of faces ↓	Number of vertices ↓	SSIM ↑	PSNR ↑
LN3Diff	(a)	522350	261379	0.947	17.918
	(b)	100310	50176	0.851	15.303
DreamGaussian	(a)	71536	43180	0.876	20.782
	(b)	24942	14558	0.839	14.834
TripoSR	(a)	106492	53324	0.862	15.764
	(b)	42056	21055	0.868	16.906
Magic123	(a)	50000	25001	0.782	10.547
	(b)	50000	24900	0.784	10.398
Ours	(a)	**718**	**352**	**0.957**	**22.229**
	(b)	**21875**	**10928**	**0.940**	**19.413**

*Optimal results are displayed in **bold**. Red and Blue are used to distinguish (a) from (b).

Gaussian noise into the LLFF image dataset. The noise was characterized by a mean of 0 and variances of 500, 1000, 2000, and 3000, which progressively increased the noise intensity. Quantitative results are presented in Table 2, with Fig. 8a providing a visual representation. As Gaussian noise increased, the image surface texture gradually blurred. Specifically, as noise variance increased from 0 to 3000, the CD metric rose from 0 to 0.013, indicating a slight decrease in model precision. SSIM values remained robust, measuring 0.814 under no-noise conditions and 0.809 under maximum noise conditions, suggesting a minimal impact on texture quality. Statistically, the 95% confidence interval for CD ranged from 0.002 to 0.014, and for SSIM it ranged from 0.799 to 0.813. These findings underscore the significant robustness of the 3DSynBrush framework to noisy input conditions.

Illumination effects were subsequently investigated to assess system generation performance under different lighting intensities. As depicted in Fig. 8(b) and quantitatively summarized in Table 2, illumination effects were progressively introduced into the LLFF image dataset. An illumination intensity of $E = 1$ represents normal lighting conditions. Under low illumination conditions where $E < 1$, specifically $E = 0.5$ and $E = 0.75$, the CD values were 0.035 and 0.020, respectively, while SSIM values were 0.805 and 0.795, respectively. This indicates that the method maintained high fidelity under low illumination, producing results closely comparable to those without illumination effects. Under high illumination conditions where $E > 1$, at $E = 1.25$ and $E = 1.5$, the CD values were 0.021 and 0.022, respectively, and SSIM values were 0.848 and 0.842, respectively. These results suggest that while the framework can recover textures associated with enhanced lighting, it may not fully restore the appearance observed under normal illumination. The 95% confidence interval for CD ranged from 0.004 to 0.035, and for SSIM it ranged from 0.792 to 0.850. This analysis confirms that the model robustly performs despite variations in illumination intensity.

(a)Effects of Gaussian noise

$\sigma = 0$ $\sigma = 500$ $\sigma = 1000$ $\sigma = 2000$ $\sigma = 3000$

σ represents the Gaussian noise intensity.

(b)The effect of light

$E = 1$ $E = 0.5$ $E = 0.75$ $E = 1.25$ $E = 1.5$

E represents the light intensity.

(c)The impact of obstructions

$d = 0$ $d = 10$ $d = 20$ $d = 30$ $d = 40$

d represents the side length of the obstruction.

Fig. 8 | Robustness testing of 3DSynBrush under various adverse conditions. a Noise intensity modulated by variance *σ*, where increased variance corresponds to heightened intensity; **b** illumination intensity regulated by E, with E=1 denoting an absence of illumination; and **c** occlusions created using black squares of side length d.

Finally, the impact of image occlusion was evaluated by simulating square black patches within the LLFF image dataset. As shown in Fig. 8(c) and detailed in Table 2, the side length d of the black patch, expressed in pixels, served as a measure of occlusion size. Performance was evaluated for d values of 10, 20, 30, and 40. For small occlusions, specifically when d = 10, CD was 0.033 and SSIM was 0.805, demonstrating that the system maintained high generation quality. As occlusion size increased, CD values remained notably stable at 0.034, 0.035, and 0.036 for d = 20, 30, and 40, respectively, indicating a very limited impact on the reconstructed mesh shape. However, SSIM values decreased to 0.775, 0.750, and 0.725 for d = 20, 30, and 40, respectively, reflecting a more significant challenge in capturing texture details due to increasing occlusion. Even under extreme occlusion, with d = 40, the framework still yielded a competitive SSIM level, significantly exceeding results achievable through random guessing. Statistically, the 95% confidence interval for CD ranged from 0.008 to 0.047, and for SSIM it ranged from 0.728 to 0.820. While occlusion reduced texture accuracy, the 3DSynBrush framework demonstrated strong robustness in maintaining overall shape coherence, effectively mitigating the challenges posed by the loss of localized information.

Discussion

This paper introduces an innovative framework for the three-dimensional reconstruction of individual elements within ancient Chinese artworks, such as the Dunhuang murals. The framework systematically addresses the key challenge of generating high-quality 3D models from a single, non-photorealistic image. Our pipeline begins with the CME dataset, a specialized collection we built using meticulous segmentation with Mask2Former. It then leverages Zero123++ to generate sparse, 3D-consistent multi-view images, which are densified into a 360-degree video by ZeroRF. Finally, NeRF2mesh converts the resulting neural radiance field into an explicit 3D mesh, transforming static mural elements into interactive, spatially coherent digital assets.

A key advantage of our method is its divide-and-conquer strategy. By first ensuring multi-view consistency, it significantly improves the success rate and quality of the subsequent NeRF training and mesh generation. The use of a fixed depth prior is particularly suitable for non-photorealistic art, where data scarcity makes learning accurate priors difficult. Furthermore, the framework's output of a standard mesh model offers significant application potential, from creating virtual museums to providing morphological references for heritage restoration.

Table 2 | The quantitative test results under adverse conditions

Conditions	Gaussian noise				
σ	0	500	1000	2000	3000
CD	0	0.007	0.009	0.011	0.013
SSIM	0.814	0.805	0.799	0.803	0.809
95% CI	CD:	(0.002, 0.014)	SSIM:	(0.799, 0.813)	
Conditions	Illumination				
E	1	0.5	0.75	1.25	1.5
CD	0	0.035	0.020	0.021	0.022
SSIM	0.814	0.805	0.795	0.848	0.842
95% CI	CD:	(0.004, 0.035)	SSIM:	(0.792, 0.850)	
Conditions	Occlusion				
d	0	10	20	30	40
CD	0	0.033	0.034	0.035	0.036
SSIM	0.814	0.805	0.775	0.750	0.725
95% CI	CD:	(0.008, 0.047)	SSIM:	(0.728, 0.820)	

We use CD to represent the impact of adverse conditions on the shape of the Mesh. We use SSIM to measure the impact on texture. The confidence interval of the metrics is measured to represent the statistical range of the metrics.

At the same time, this study also recognizes that there are some directions for the current model to be optimized:

First, the pipeline's reliance on the initial Mask2Former segmentation means that any inaccuracies, such as blurred edges or missed details, are propagated and amplified in the final model, causing morphological distortions. Future work could explore end-to-end joint optimization of the segmentation and generation stages or introduce an interactive correction step to ensure source data fidelity.

Second, the framework is currently limited by the 512×512 resolution of the upstream multi-view generation model (Zero123++). This cap leads to a loss of intricate details common in Dunhuang murals, diminishing the artistic fidelity of the final reconstruction. Adopting higher-resolution generation models is a critical next step.

Third, reconstructing stylized artistic representations poses a significant challenge. Dunhuang murals often employ non-realistic perspectives that current generative models, trained on real-world geometry, may attempt to correct. This can result in a geometrically plausible model that loses the artistic essence of the original work. A more profound challenge is enabling the model to understand and reproduce this unique artistic language rather than merely performing geometric reconstruction.

In our future work, we will continue to conduct more in-depth research on the optimization of multi-view synthesis. In summary, the single-view 3D reconstruction framework for mural art elements proposed in this study has achieved significant results in generating high-quality 3D models from a single non-photorealistic image, providing a new method for the field.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to copyright restrictions held by the Dunhuang Academy but are available from the corresponding author on reasonable request.

Code availability

The code used in this study is available from the corresponding author upon reasonable request.

Received: 24 September 2025; Accepted: 2 March 2026;
Published online: 13 March 2026

References

1. Lee, S. S. Repository of ingenuity: Cave 61 and artistic appropriation in tenth-century Dunhuang. *Art. Bull.* **94**, 199–225 (2012).
2. Sicheng, L. Ancient Chinese architecture as seen in the Dunhuang murals. *Cult. Relics.* **5**, 7 (1951). (in Chinese)
3. Shengliang, Z. & Zhaofu, Q. The flying apsaras: The central pillar front niche ceiling of cave 257 in theMogao Caves, Northern Wei. *Dunhuang Res.* **2018**, 2 (2018). (in Chinese)
4. Ru, S. The image of towers in the Dunhuang murals. *Dunhuang Stud.* **1996**, 1–16 (1996).
5. Zhang, H. Aesthetic contemplation of the Tang Dynasty Dunhuang frescoes elements on contemporary costume design. In: Cross-Cultural Design. User Experience of Products, Services, and Intelligent Environments (ed Rau, P. L. P.) 431–440 (Lecture Notes in Computer Science, Vol. 12192, Springer, Cham, 2020).
6. Youhui, G. Early patterns and decorations in the Dunhuang Mogao caves. *Journal of Dunhuang Studies* 101–107+125–126 (1980).
7. Shengliang, Z. et al. *Complete Works of the Dunhuang Caves: Landscape Paintings* (Commercial Press, Hong Kong, 2002).
8. Yagi, H. & Mei, L. About the Western Pure Land illustration on the south wall in the Mogao Cave 220 at Dunhuang. *Dunhuang Res.* **2012**, 9–15 (2012).
9. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10674–10685 (2021).
10. Saharia, C. et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Pro. 36th International Conference on Neural Information Processing Systems, NIPS '22* (Curran Associates Inc., Red Hook, NY, USA, 2022).
11. Zhu, Y.-L., Wang, X.-Y., Wan, T.-R. & Yang, Y.-Q. The analysis and creation of Mogao Caves' three-dimensional model. In *E-Learning and Games*, 191–198 (Springer International Publishing, Cham, 2017).
12. Wang, H., Du, X., Li, J., Yeh, R. A. & Shakhnarovich, G. Score Jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12619–12629 (2023).
13. Tang, J. et al. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 22762–22772 (2023).
14. Zhaowen, Q. & Tianwen, Z. Key techniques on cultural relic 3d reconstruction. *J. Electron.* **36**, 2423–2427 (2008).
15. Cheng, L., Gao, X. & Tian, X. The application of digital technology in mural preservation: A case study of Dunhuang animation. Art Style Art Cult. *Int. Mag.* **14**, 94–111 (2024).
16. Han, P.-H. et al. A compelling virtual tour of the Dunhuang cave with an immersive head-mounted display. *IEEE Comput. Graph. Appl.* **40**, 40–54 (2020).
17. Stone, R. & Ojika, T. Virtual heritage: What next? *IEEE Multimed.* **7**, 73–74 (2000).
18. Liu, B., He, F., Du, S., Zhang, K. & Wang, J. Dunhuang murals contour generation network based on convolution and self-attention fusion. *Appl. Intell.* **53**, 22073–22085 (2023).
19. Chen, B., Xu, H., Yin, Z., Zhou, C. & Yang, H. Cross-scale generative adversarial learning networks for intelligent hierarchical control of proton exchange membrane fuel cells systems. *Int. Commun. Heat. Mass Transf.* **169**, 109878 (2025).
20. Wang, H., Qiu, X., Xiong, Y. & Tan, X. Autogrn: An adaptive multi-channel graph recurrent joint optimization network with copula-based dependency modeling for spatio-temporal fusion in electrical power systems. *Inf. Fusion* **117**, 102836 (2025).
21. Wang, H. et al. Rofed-llm: Robust federated learning for large language models in adversarial wireless environments. *IEEE Trans. Netw. Sci. Eng.* **13**, 1084–1096 (2026).

22. Wang, G., Chen, Z., Loy, C. C. & Liu, Z. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9031–9042 (2023).
23. Guo, S. et al. Depth-guided robust point cloud fusion nerf for sparse input views. *IEEE Trans. Circuits Syst. Video Technol.* **34**, 8093–8106 (2024).
24. Niemeyer, M. et al. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5470–5480 (2022).
25. Shi, R., Wei, X., Wang, C. & Su, H. Zerorf: Fast sparse view 360° reconstruction with zero pretraining. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21114–21124 (2024).
26. Mildenhall, B. et al. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **65**, 99–106 (2021).
27. Zong, X., Li, Z. & Zhang, Q. A study on the stage image of "rebound lute behind the back" in Dunhuang, China. *Int. J. Adv. Cult. Technol.* **12**, 16–29 (2024).
28. Dazheng, Z. *Dunhuang Murals and the Colors of Chinese Painting* (People's Fine Arts Publishing House, Beijing, 2000).
29. Cheng, B., Misra, I., Schwing, A. G., Kirillov, A. & Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1280–1289 (2022).
30. Liu, R. et al. Zero-1-to-3: Zero-shot one image to 3d object. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 9264–9275 (2023).
31. Shi, R. et al. Zero123++: a Single Image to Consistent Multi-view Diffusion Base Model. *arXiv e-prints* arXiv:2310.15110 (2023).
32. Mildenhall, B. et al. Local light field fusion. *ACM Trans. Graph. (TOG)* **38**, 1–14 (2019).
33. Tang, J. et al. Delicate textured mesh recovery from nerf via adaptive surface refinement. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 17693–17703 (2023).
34. Schönberger, J. L. & Frahm, J.-M. Structure-from-motion revisited. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 4104–4113 (2016).
35. Lan, Y. et al. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *Computer Vision – ECCV 2024*, 112-130 (Springer Nature Switzerland, Cham, 2025).
36. Tang, J., Ren, J., Zhou, H., Liu, Z. & Zeng, G. Dreamgaussian: Generative Gaussian splatting for efficient 3d content creation. In Proc. International Conference on Learning Representations (ICLR, 2024).
37. Tochilkin, D. et al. TripoSR: Fast 3D object reconstruction from a single image. *arXiv e-prints* arXiv:2403.02151 (2024).
38. Qian, G. et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *International Conference on Representation Learning*, vol. 2024, 48142–48159 (2024).

Acknowledgements

This research was supported by the Yulin Science and Technology Plan Project (No. 2025-CXY-033, No. 2025-CXY-034), National Natural Science Foundation of China (No. 62471390, No. 62406247, No. 62306237), Key Project of Scientific Research Plan of Shaanxi Provincial Department of Education (No. 24JS052), Key Laboratory of Archaeological Exploration and Cultural Heritage Conservation Technology (Northwestern Polytechnical University, No. 2024KFT03).

Author contributions

Q.H., N.X., and S.Q. were responsible for conceptualization. Q.H. also developed the methodology. X.P. handled preparation and resources. J.W. wrote the original draft and performed the validation. Q.H. and J.W. were responsible for writing-review and editing. J.W. also provided supervision. J.P. was responsible for project administration.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Qiyao Hu.