ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume X-M-2-2025
30th CIPA Symposium “Heritage Conservation from Bits:
From Digital Documentation to Data-driven Heritage Conservation”, 25–29 August 2025, Seoul, Republic of Korea

3D Gaussian Splatting for Enhanced Documentation of Cultural Artifacts

Jungmin Lee[1], Sojeong Park[1], Jihyun Min [2], Jongwon Choi [1, *]

[1]Dept. of Advanced Imaging, GSAIM, Chung-Ang University, Seoul, Republic of Korea - (jngmnlee, dazzsoj, choijw)@cau.ac.kr
[2]Technology Research Institute for Culture & Heritage, Daejeon, Republic of Korea - jhmin@tric.or.kr

Keywords: 3D Gaussian Splatting, Novel View Synthesis, 3D Reconstruction, Digital Documentation.

Abstract

3D Gaussian splatting has shown promise for high-fidelity 3D modeling in cultural heritage d ocumentation. However, applying 3DGS to cultural artifacts faces challenges, including image alignment errors in dynamic capture environments and background interference from scene-level reconstruction. We propose an optimized 3DGS framework tailored for artifact modeling, introducing two key innovations: (1) COLMAP-free for robust image alignment under non-static conditions, and (2) implementing targeted isolation to separate artifacts from extraneous backgrounds. Our approach enhances reconstruction quality, preserving intricate geometric details and critical textures for cultural artifacts. Comprehensive evaluations against traditional photogrammetry and standard 3DGS demonstrate superior performance in accuracy and visual fidelity. Through enhanced 3D Gaussian splatting, the proposed method achieves precise 3D documentation of cultural artifacts, enabling diverse digital applications.

1. Introduction

3D modeling has emerged as an indispensable tool for preserving and utilizing cultural artifacts. This approach facilitates accurate documentation of the artifacts' current state and converts the artifacts into digital assets for various applications. Recent advances in computer graphics have highlighted novel view synthesis (NVS) as a transformative approach to 3D modeling. As shown in Figure 1, NVS generates photorealistic images from novel viewpoints using limited 2D image data. Methodologies such as neural radiance fields (NeRF) (Mildenhall et al., 2021) and 3D Gaussian splatting (3DGS) (Kerbl et al., 2023) have received substantial interest from the research community.

The rapid evolution of NVS has catalyzed the adoption of methods based on artificial intelligence for 3D modeling in cultural heritage (Murtiyoso and Grussenmeyer, 2023; Croce et al., 2023). Among these approaches, 3DGS stands out by enabling real-time rendering and high-fidelity visualization. 3DGS employs structure-from-motion (SfM) algorithms (Schonberger and Frahm, 2016) to align 2D images and leverages point clouds to represent 3D objects through Gaussian distributions. These distributions facilitate exceptional reproduction of intricate geometric details and surface textures, thereby overcoming fundamental limitations of NeRF and achieving superior performance in representing fine-grained details critical for cultural artifacts.

Originally developed for general scene representation, 3DGS presents significant challenges when applied to object-level targets, such as cultural artifacts. First, image alignment frequently fails in dynamic capture environments. 3DGS depends on SfM processing through COLMAP (Schönberger et al., 2016), assuming a static subject with a moving camera. However, inconsistent repositioning of artifacts or camera position changes during capture sessions cause significant image alignment errors that degrade reconstruction quality. Second, 3DGS inherently captures entire scenes rather than isolated objects. Consequently, backgrounds such as exhibition halls or conservation laboratories frequently appear in artifact captures, introducing extraneous visual elements that detract from the digital representation. Third, cultural artifacts often have regions with limited or no photographic coverage due to access constraints or varying levels of photographic expertise, resulting in incomplete reconstructions with missing data.

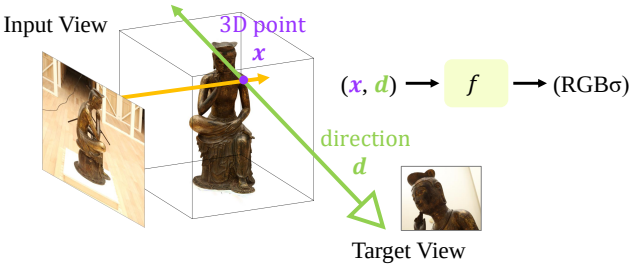


Figure 1. Overview of novel view synthesis.

To address these critical challenges, we propose a modified 3DGS approach specifically optimized for cultural artifact digitization. Our method introduces two key innovations: (1) replacing COLMAP with DUSt3R (Wang et al., 2024) to enhance image alignment accuracy for non-static capture scenarios substantially, (2) implementing targeted isolation techniques using SAM2 (Ravi et al., 2024) that effectively separate artifacts from their surroundings during the modeling process, and (3) developing a data completion strategy based on GaussianObject (Yang et al., 2024) to reconstruct missing regions where image data is insufficient. To evaluate the effectiveness of our method, we conduct comprehensive comparisons with traditional photogrammetry and standard 3DGS. These methodological advances aim to significantly enhance the digital preservation and utilization of cultural heritage assets, potentially establishing a novel paradigm for cultural heritage research that seamlessly integrates technical innovation with preservation.

2. Related Works

2.1 Photogrammetry

Photogrammetry has dominated the digital documentation of cultural heritage for decades through the ability to convert 2D

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-M-2-2025-215-2025 | © Author(s) 2025. CC BY 4.0 License.

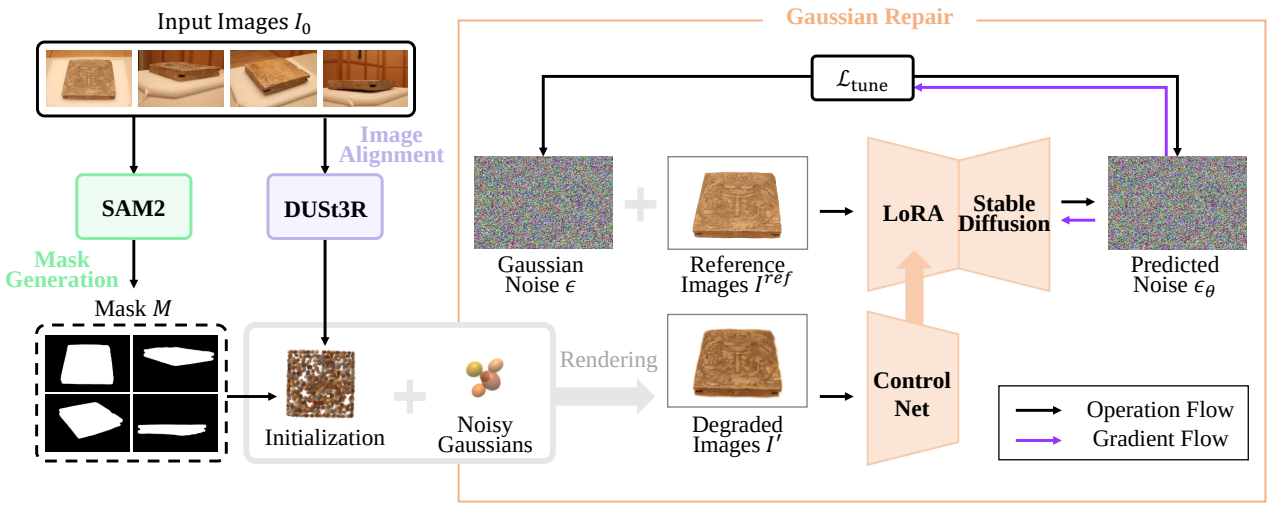


Figure 2. Framework of our proposed method.

images to 3D models (Mikhail et al., 2001). The SfM algorithms (Schonberger and Frahm, 2016) used in COLMAP (Schönberger et al., 2016), Metashape (Agisoft LLC, 2023), and Reality Capture (Capturing Reality, 2023) estimate camera positions, create point clouds, and transform point clouds into 3D models through meshing. Researchers have successfully applied photogrammetry to various cultural artifacts, including museum objects of various materials (Nicolae et al., 2014) and small-scale artifacts measuring 5-10 cm (Sapirstein, 2018).

Despite widespread adoption, photogrammetry encounters several significant limitations. Photogrammetry cannot reconstruct viewpoints absent from the original images and requires numerous images to achieve accurate 3D reconstructions. The quality of the resulting models depends on precise calculations of overlapping regions between captured images, which require considerable operator expertise. The requirement of specialized skills constitutes a substantial barrier to broader implementation in cultural heritage preservation initiatives.

2.2 Novel View Synthesis

NeRF (Mildenhall et al., 2021) appeared as an innovative technology for NVS, utilizing neural networks to render scenes from arbitrary viewpoints based on 2D images. Researchers have effectively applied NeRF to digitize artifacts using relatively few images, allowing preservation, virtual exhibitions, and archaeological analysis (Condorelli et al., 2021; Croce et al., 2023; Murtiyoso and Grussenmeyer, 2023). However, computational complexity causes excessively long training and rendering times, creating significant constraints for practical implementation in real-world applications.

3DGS (Kerbl et al., 2023) recently emerged as a methodology to address the computational bottlenecks of NeRF, allowing real-time rendering while maintaining visual quality. This approach significantly improves rendering efficiency and detail representation capabilities. However, applications of 3DGS in cultural heritage focus primarily on large-scale archaeological sites and architectural environments (Clini et al., 2024; Wang and Weijia, 2024), with limited case studies involving individual artifacts. Some researchers have attempted to extract an object from 3D scenes (Dahaghin et al., 2024), yet this method cannot capture base surfaces, limiting complete artifact documentation. Furthermore, existing studies fail to resolve image alignment errors when capturing artifacts with moving cameras and repositioned objects. This situation also creates problems where background elements surrounding the artifacts appear in the resulting models. These technical limitations hinder the effective utilization of 3DGS for the precise documentation of heritage artifacts, particularly in museums and research settings.

3. Method

We propose a modified 3DGS approach optimized for cultural artifact documentation that addresses three key challenges: inaccurate image alignment in non-stationary acquisition environments, unwanted background elements in artifact reconstructions, and incomplete data with insufficient photographic coverage. Our framework is detailed in Figure 2.

3.1 Preliminary

3DGS represents scenes using a collection of 3D Gaussian primitives (Kerbl et al., 2023). Each Gaussian is defined by a center position $\boldsymbol{\mu} \in R^3$, a covariance matrix $\boldsymbol{\Sigma} \in R^{3\times3}$, an opacity $\alpha \in [0, 1]$, and spherical harmonics coefficients for view-dependent appearance. This anisotropic representation enables precise modeling of complex surfaces, making 3DGS highly suitable for reconstructing cultural heritage artifacts with intricate details.

The 3DGS pipeline begins with SfM (Schonberger and Frahm, 2016) processing via COLMAP (Schönberger et al., 2016) to estimate camera parameters and generate an initial point cloud. These points initialize the 3D Gaussians, which are optimized to minimize the discrepancy between rendered and ground-truth images. During rendering, each 3D Gaussian is projected onto the 2D image plane, with the Gaussian contribution determined by a projected mean $\boldsymbol{\mu}_i' \in R^2$ and covariance $\boldsymbol{\Sigma}_i' \in R^{2\times2}$, computed as $\boldsymbol{\Sigma}_i' = \boldsymbol{J}_i\boldsymbol{\Sigma}_i\boldsymbol{J}_i^T$, where $\boldsymbol{J}_i \in R^{2\times3}$ is the Jacobian of the projection function at $\boldsymbol{\mu}_i$. The rendered image at a pixel $\boldsymbol{x} \in R^2$ is obtained by alpha-blending the contributions of depth-sorted Gaussians:

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-M-2-2025-215-2025 | © Author(s) 2025. CC BY 4.0 License.

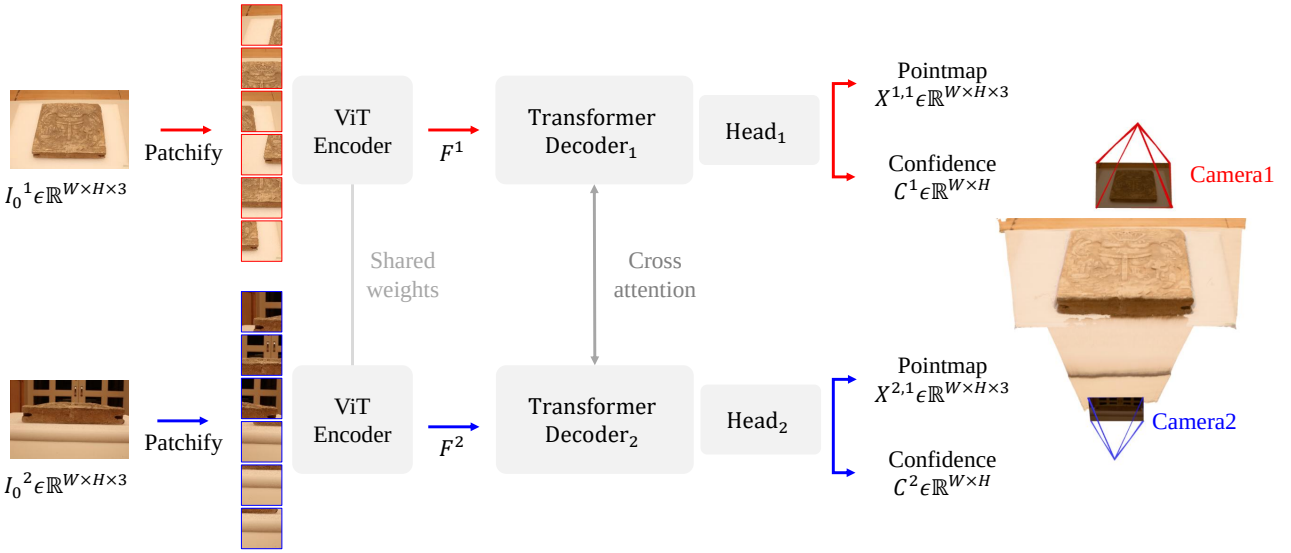


Figure 3. Architecture of DUSt3R.

$$I(\boldsymbol{x}) = \sum_{i=1}^{N} \boldsymbol{c}_i \alpha_i \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i', \boldsymbol{\Sigma}_i') \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{1}$$

where $\boldsymbol{c}_i \in R^3$ is the view-dependent color derived from spherical harmonics.

The optimization of Gaussian parameters $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$, $\boldsymbol{c}_i$, α_i is performed through gradient descent to minimize a photometric loss function:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_1 + \lambda \mathcal{L}_{\text{D-SSIM}}, \tag{2}$$

where $\mathcal{L}_1$ represents the L1 loss between the rendered image I and the ground truth image, $\mathcal{L}_{\text{D-SSIM}}$ computes the structural similarity loss, and λ is a weighting factor. Then, an adaptive density control adds Gaussians in regions with high reconstruction error and removes redundant Gaussians, optimizing computational efficiency for complex artifact geometries.

3.2 Image Alignment

The initialization phase employs DUSt3R (Wang et al., 2024) to generate point cloud representations from input images, addressing critical limitations of traditional 3DGS workflows. DUSt3R eliminates the dependency on prior camera calibration and viewpoint pose estimation while providing robust image alignment in dynamic capture scenarios. Standard 3DGS implementations rely on COLMAP for Structure-from-Motion processing and assume static scenes. However, cultural artifact documentation often involves varying camera positions and object repositioning during capture sessions, causing COLMAP-based alignment to fail. DUSt3R directly addresses these challenges by performing robust correspondence estimation and depth prediction without requiring static scene assumptions. The method integrates DUSt3R as a direct replacement for COLMAP in the 3DGS pipeline, with the overall architecture illustrated in Figure 3.

DUSt3R analyzes pairs of RGB images to regress corresponding 3D pointmaps. Figure 3 shows that the network consists of two identical branches, each containing a vision transformer (ViT) encoder, a transformer decoder, and a regression head. For input images $I_0^1, I_0^2 \in R^{W \times H \times 3}$, a shared ViT encoder analyzes the images in a Siamese configuration and generates token representations $F^1, F^2 \in R^{N \times D}$, where N represents the number of tokens and D denotes the feature dimension. ViT-Base decoders, equipped with cross-attention mechanisms, process these representations to enable continuous information exchange between views. The decoders generate aligned 3D pointmaps $X^{1,1}, X^{2,1} \in R^{W \times H \times 3}$, defined in the coordinate frame of I_0^1, along with corresponding confidence maps $C^{1,1}, C^{2,1} \in R^{W \times H \times 1}$.

Building on pairwise predictions, DUSt3R applies global alignment in multi-view scenarios to create a cohesive 3D scene representation. For a set of input images $\{I_0^1, I_0^2, \ldots, I_0^N\}$, DUSt3R constructs a connectivity graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where vertices $\mathcal{V}$ represent images and edges $\mathcal{E}$ indicate image pairs with sufficient visual overlap. DUSt3R determines overlap using off-the-shelf image retrieval or confidence-based filtering through the network, as detailed in Figure 3. For each edge $e = (n, m) \in \mathcal{E}$, DUSt3R computes pairwise pointmaps $X^{n,n}, X^{m,n} \in R^{W \times H \times 3}$ and confidence maps $C^{n,n}, C^{m,n} \in R^{W \times H \times 1}$. Global alignment optimizes a set of world-coordinate pointmaps $\{\chi^n \mid n = 1, \ldots, N\}$ by minimizing 3D projection errors. This approach defines the per-pixel projection error for each edge e and vertex v as:

$$\mathcal{L}_i^{v,e} = C_i^{v,e} \left\| \chi_i^v - \sigma_e P_e X_i^{v,e} \right\|_2^2, \tag{3}$$

and formulates the global optimization as:

$$\chi^* = \arg \min_{\{\chi^n\}, \{P_e\}, \{\sigma_e\}} \sum_{e \in \mathcal{E}} \sum_{v \in \{n,m\}} \sum_{i=1}^{W \cdot H} \mathcal{L}_i^{v,e}, \tag{4}$$

where $P_e \in R^{3 \times 4}$ and $\sigma_e > 0$ represent the relative pose and scale for edge e, respectively, with the constraint $\prod_{e \in \mathcal{E}} \sigma_e = 1$ to ensure scale consistency. This optimization facilitates robust 3D reconstruction of cultural artifacts using a pinhole camera model.

3.3 Cultural Artifact Modeling

3.3.1 Mask Guided Optimization We incorporate mask-guided optimization to separate artifacts from surroundings using SAM2 (Ravi et al., 2024). This approach reduces noise in the 3D reconstruction pipeline by providing clean inputs, focusing solely on cultural artifacts of interest. The implementation follows a process where SAM2's transformer-based attention mechanisms identify artifact boundaries and automatically generate masks that isolate target objects from non-relevant elements. For an input RGB image $I_0^i \in R^{W\times H\times 3}$, SAM2 generates a segmented output image $I_i \in R^{W\times H\times 3}$ through the following process:

$$I_i = \text{SAM2}(I_0^i) = I_0^i \odot M, \tag{5}$$

where $M \in \{0,1\}^{W\times H}$ is the binary mask produced by SAM2, identifying the foreground artifact, and $\odot$ denotes element-wise multiplication with M broadcast across the color channels of I_0^i.

SAM2 uses a hierarchical vision transformer to extract feature embeddings from each frame and references memories across multiple images to maintain consistent artifact segmentation. The model's multi-scale processing handles images captured from different viewpoints, which 3D modeling requires. SAM2 segments artifacts across camera angles spanning 360° horizontally and ±45° vertically while maintaining consistent boundary detection. The hierarchical feature extraction recognizes artifact characteristics regardless of perspective changes, enabling viewpoint-invariant segmentation. This segmentation keeps only the artifact while eliminating background clutter.

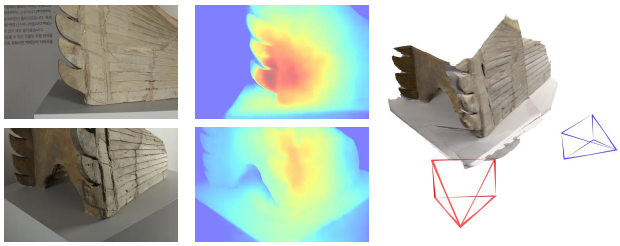
3.3.2 Gaussian Repair We adopt the Gaussian repair model from the GaussianObject framework (Yang et al., 2024) to address regions with limited or inconsistent observations in the reconstruction of cultural artifacts. This approach produces high-fidelity 3D models by correcting rendering artifacts such as missing textures, distorted surfaces, and inaccurate geometric details caused by incomplete photographic coverage. The Gaussian repair model employs a 2D diffusion-based architecture (Rombach et al., 2021) to restore intricate details across multi-view scenarios with varying camera angles and lighting conditions.

The training process applies a *leave-one-out* strategy to create paired data for multi-view inputs. For a reference image set $I^{\text{ref}} = \{I_i\}_{i=1}^N$ (where $N > 4$), we train a preliminary 3DGS using all images except I_i. We generate a corrupted rendering I_i' for the corresponding view, reflecting missing information. To simulate artifacts common in artifact reconstruction, we inject 3D noise into Gaussian attributes sampled from a Gaussian distribution $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The repair model adjusts the noise parameters to mimic surface irregularities and texture loss, prevalent in complex or reflective artifact geometries. This process produces a dataset pairing corrupted renderings I_i' with reference images I_i, capturing multi-view reconstruction errors.

The Gaussian repair model utilizes a ControlNet architecture (Zhang et al., 2023) to condition a pre-trained diffusion model on corrupted renderings I_i'. The repair model leverages paired data with I_i' to train artifact detail reconstruction. During training, the repair model minimizes a denoising loss:

(a) Front point cloud fragment (b) Back point cloud fragment

Figure 4. Failed alignment of Chimi Using RealityCapture.



(a) RGB image (b) Depth map (c) 3D reconstruction

Figure 5. Images alignment of Chimi using DUSt3R.

$$\mathcal{L}_{\text{tune}} = E_{I_i, I_i', t, \epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\alpha_t} I_i + \sqrt{1-\alpha_t}\epsilon, I_i', t\right)\right\|_2^2\right], \tag{6}$$

where ϵ_θ denotes the noise predictor, t represents the diffusion timestep, α_t signifies the scheduler coefficient and I_i' acts as the conditioning image. This loss enables mapping low-quality renderings to high-fidelity images, preserving features like engravings and weathered surfaces. The training process fine-tunes ControlNet weights, potentially using LoRA (Hu et al., 2022) for efficiency, while keeping the diffusion backbone frozen to optimize artifact-specific details and maintain generalization.

After training, we freeze the Gaussian repair model and apply the model to correct problematic rendered views. Distance-aware sampling selects views with high deviation from reference images, often indicating sparse observations or complex geometry. We process corrupted renderings to generate repaired images, recovering missing details and correcting distortions. We optimize 3D Gaussians using repaired images and SAM2-segmented reference images. This iterative refinement improves the reconstruction quality in underrepresented regions, ensuring that the final 3D model accurately represents the geometry and appearance of the artifact.

4. Experiments

4.1 Implementation Detail

Our framework is built on GaussianObject (Yang et al., 2024). Instead of the GaussianObject's visual hull initialization strategy, we utilize DUSt3R (Wang et al., 2024) to handle a large number of images. To process masks, the framework incorporates SAM2 (Ravi et al., 2024). The entire process was tested on two A6000 GPUs.

4.2 Dataset

To evaluate the effectiveness of the proposed method on cultural artifacts, we employ six cultural artifacts from two museums.

Name of Artifact	Institution	# Images	Resolution
Wooden Carving of Human Face	NMK	257	8688×5792 pixel
Celadon Incense Burner with Lion-shaped Lid	NMK	275	8688×5792 pixel
Gilt-bronze Pensive Maitreya Bodhisattva	NMK	782	8688×5792 pixel
Patterned Tile from Oe-ri	BMK	185	8688×5792 pixel
Bronze Miniature Pagoda	BMK	461	8688×5792 pixel
Chimi (Ridge-end Roof Tile)	BMK	167	8688×5792 pixel

Table 1. Details of cultural artifacts used in our experiments.

Name of Artifact	# Images	RealityCapture	Colmap	DUSt3R
Wooden Carving of Human Face	2			✓
	10			✓
	100			✓
	257			
Celadon Incense Burner with Lion-shaped Lid	2			✓
	10			✓
	100			✓
	275	✓	✓	
Gilt-bronze Pensive Maitreya Bodhisattva	2			✓
	10			✓
	100			✓
	782	✓	✓	
Patterned Tile from Oe-ri	2			✓
	10			✓
	100			✓
	185			
Bronze Miniature Pagoda	2			✓
	10			✓
	100			✓
	461			
Chimi (Ridge-end Roof Tile)	2			✓
	10			✓
	100			✓
	167			

Table 2. Alignment performance of COLMAP, RealityCapture, and DUSt3R across image counts.

The National Museum of Korea (NMK) provided three artifacts, and the Buyeo Museum of Korea (BMK) contributed the remaining three artifacts. These artifacts represent various materials and geometric complexities, providing comprehensive evaluation cases. Table 1 presents detailed information about each artifact.

4.3 Image Alignment

To evaluate the robustness of different image alignment methods, we compared the performance of COLMAP (Schönberger et al., 2016), RealityCapture (Capturing Reality, 2023), and DUSt3R across varying numbers of input images. We conducted experiments with 2, 10, 100 images per artifact, and all captured images. As shown in Table 2, both RealityCapture and COLMAP failed to properly align images in all experiments except for static camera environments (Celadon Incense Burner with Lion-shaped Lid and Gilt-bronze pensive Maitreya Bodhisattva). Although there was insufficient overlap between images, COLMAP and RealityCapture exhibited poor alignment capability even in static environments.

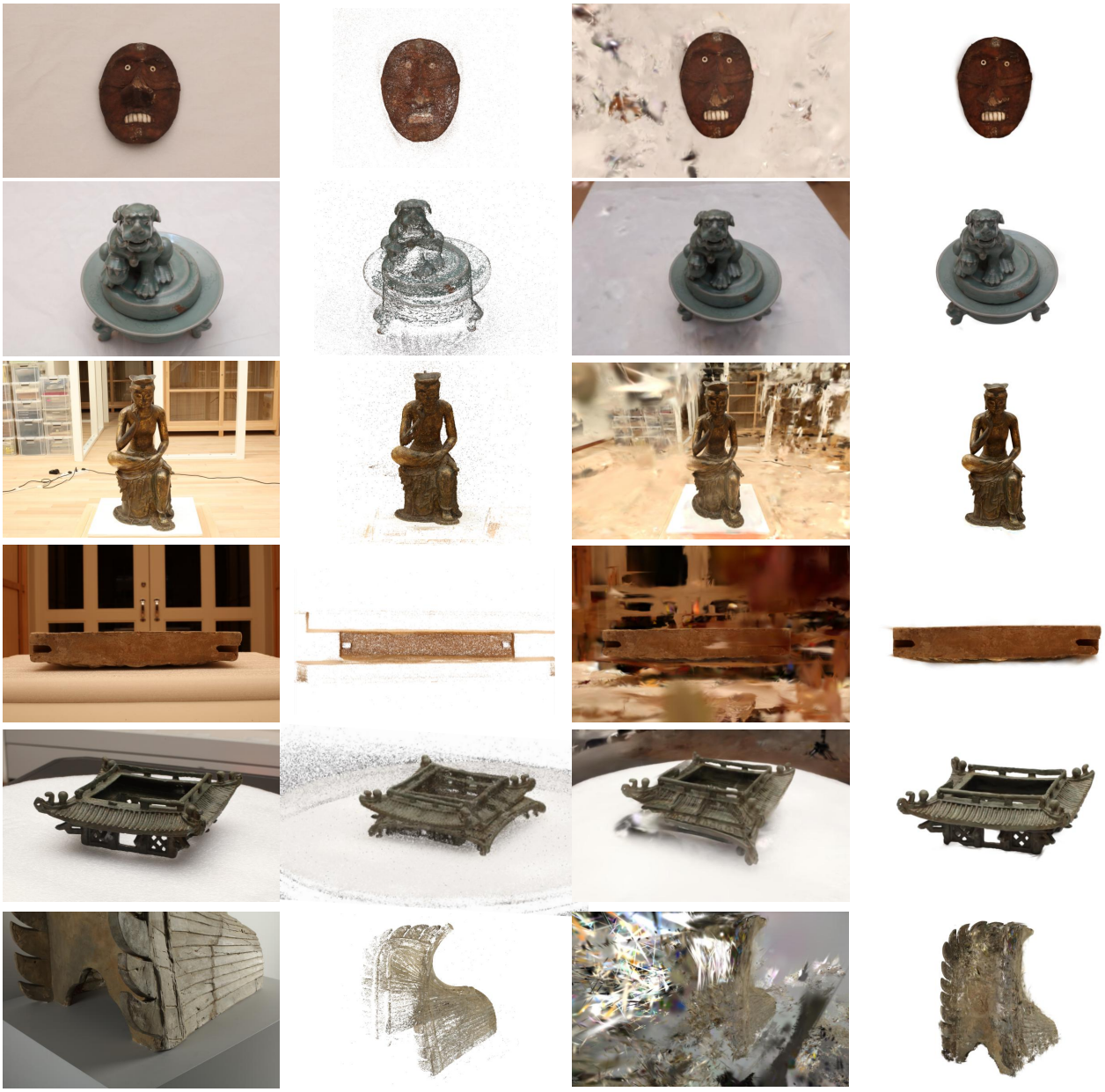
Figure 4 demonstrates this limitation clearly, where the Chimi appeared as two separate fragments because the algorithms could not establish correspondences between the front and back regions due to a lack of overlapping visual information. This highlights the fundamental limitations of conventional SfM methods in cultural heritage documentation. These limitations typically necessitate time-consuming manual alignment procedures or extensive post-processing to remove noise artifacts, significantly increasing the workload for professionals.

In contrast, DUSt3R successfully aligned artifact images in most experimental settings, demonstrating superior robustness in handling challenging capture conditions. Figure 5 illustrates how DUSt3R accurately recognized depth and correctly aligned even pairs of images with minimal overlap, a scenario where traditional methods typically fail. However, we observed that DUSt3R’s high GPU memory requirements became a limiting factor when processing larger image sets. Specifically, when processing more than 100 images, DUSt3R could not complete the entire dataset analysis due to memory constraints. This limitation indicates that while DUSt3R provides significant advantages for artifact image alignment, large-scale projects must address practical computational resource considerations.

4.4 3D Modeling

4.4.1 Qualitative Results Figure 6 compares the visual results of 3D reconstruction quality. RealityCapture generates a point cloud representation that captures only color and shape information. This method struggles with extraneous elements because RealityCapture incorporates unnecessary background and poorly aligned sections into the final shape. For example, the Patterned Tile from Oe-ri, captured by flipping the artifact, causes RealityCapture to render the floor surface both below and above the artifact, leading to visual confusion.

The standard 3DGS results demonstrate the capability to render color and shape in real-time as lighting conditions change within a viewer. Despite this advancement, 3DGS exhibits limitations similar to RealityCapture. Specifically, the Bronze Miniature Pagoda, captured by flipping the artifact, undergoes image alignment based on the floor surface, resulting in 3DGS rendering the



(a) RGB image (b) RealityCapture (c) 3DGS (d) Ours

Figure 6. Qualitative comparison of reconstruction methods.

	Carving		Incense Burner		Bodhisattva		Patterned Tile		Miniature Pagoda		Chimi	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
RealityCapture	27.34	0.31	25.62	0.26	27.80	0.27	24.89	0.25	23.52	0.26	22.17	0.24
3DGS	28.37	0.40	27.29	0.41	29.55	0.42	28.48	0.38	28.50	0.37	27.61	0.35
Ours	29.36	0.85	30.47	0.91	32.58	0.93	31.59	0.90	28.76	0.87	27.39	0.78

Table 3. Quantitative comparison of reconstruction methods.

pagoda upside down.

In contrast, the proposed method accurately renders essential information for artifact documentation. The method dynamically renders color and shape under varying lighting conditions, eliminates irrelevant background elements, corrects misaligned sections, and focuses solely on artifact geometry and appearance. The reconstructed artifacts maintain structural integrity and appear as unified objects without extraneous elements.

4.4.2 Quantitative Results Table 3 presents numerical evaluation results using PSNR and SSIM metrics(Wang et al., 2004). The proposed method consistently outperforms both RealityCapture and standard 3DGS across all artifacts. PSNR values demonstrate substantial improvements, with the proposed approach achieving scores ranging from 27.39 to 32.58 dB compared to RealityCapture’s 22.17 to 27.80 dB and standard 3DGS’s 27.29 to 29.55 dB. The bodhisattva artifact shows the most significant improvement, with the proposed method achieving 32.58 dB PSNR compared to 27.80 dB for RealityCapture and 29.55 dB

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-M-2-2025-215-2025 | © Author(s) 2025. CC BY 4.0 License.

for 3DGS.

SSIM scores further validate the superior performance of the proposed approach. The method achieves SSIM values between 0.78 and 0.93, higher than RealityCapture, which ranges from 0.24 to 0.31, and standard 3DGS, which ranges from 0.35 to 0.42. The incense burner and bodhisattva artifacts demonstrate robust performance with SSIM scores of 0.91 and 0.93 respectively. These quantitative results confirm that the enhanced framework not only improves visual quality as observed in qualitative comparisons but also delivers measurable improvements in reconstruction fidelity and structural similarity to ground truth images.

5. Discussion

The proposed method addresses significant limitations in cultural artifact documentation but has certain constraints. Diffusion models may distort original artifact forms during reconstruction by generating areas that differ from the original. This risk increases with insufficient images. The method limits diffusion intervention to levels comparable to traditional photogrammetry's *hole-filling* stage. Securing adequate overlapping images remains critical for accurate documentation since preserving original forms is the primary objective, and diffusion models serve as supplementary tools.

Future research should improve diffusion model accuracy and fidelity to original forms. Retraining models with cultural heritage datasets could improve results in extreme scenarios with minimal image coverage. Training with artifact images would enable models to avoid arbitrary reconstruction in areas with insufficient images, instead generating details consistent with cultural heritage characteristics. Advanced training may reduce inaccurate detail generation in reconstructed regions and strengthen reconstruction reliability for cultural heritage applications. Enhanced accuracy would expand the method's applicable range to challenging documentation cases while maintaining fidelity to original artifact forms.

6. Conclusion

We present an enhanced 3DGS for cultural artifact documentation. The approach addresses two key limitations of existing methods: image alignment errors in non-stationary acquisition environments and unwanted inclusion of background elements. We introduce an integrated solution incorporating enhanced image alignment, artifact isolation through mask-guided optimization, and an artifact repair model.

Experiments utilizing six cultural artifacts from museums demonstrate superior performance. The method achieves excellent detail representation and accurate artifact isolation even with complex artifacts. The approach maintains structural integrity while eliminating extraneous elements that traditional methods struggle to handle.

We make a substantial contribution to the field of digital cultural heritage preservation and application. Future research will focus on providing a user-friendly workflow system accessible to non-experts. The research aims to address computational efficiency challenges, extend applicability to a broader range of cultural heritage types, and enhance the accuracy of diffusion models.

Acknowledgements

This research was partly supported by Culture Technology R&D Program through the Korea Creative Content Agency grant funded by Ministry of Culture, Sports and Tourism in 2023 (Project Name: Development of Intelligent Heritage Platform for Leading of Standardization on Digital Cultural Heritage, Project Number: RS-2023-00219579) and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)].

References

Agisoft LLC, 2023. Agisoft metashape. Computer software.

Capturing Reality, 2023. Realitycapture. Computer software.

Clini, P., Nespeca, R., Angeloni, R., Coppetta, L., 2024. 3D representation of Architectural Heritage: a comparative analysis of NeRF, Gaussian Splatting, and SfM-MVS reconstructions using low-cost sensors. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 93–99.

Condorelli, F., Rinaudo, F., Salvadore, F., Tagliaventi, S., 2021. A comparison between 3D reconstruction using nerf neural networks and mvs algorithms on cultural heritage images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 565–570.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. et al., 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2), 3.

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 139–1.

Mikhail, E. M., Bethel, J. S., McGlone, J. C., 2001. *Introduction to modern photogrammetry*. John Wiley & Sons.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), 99–106.

Murtiyoso, A., Grussenmeyer, P., 2023. Initial assessment on the use of state-of-the-art NeRF neural network 3d reconstruction for heritage documentation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 1113–1118.

Nicolae, C., Nocerino, E., Menna, F., Remondino, F., 2014. Photogrammetry applied to problematic artefacts. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40, 451–456.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L. et al., 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2021. High-resolution image synthesis with latent diffusion models.

Sapirstein, P., 2018. A high-precision photogrammetric recording system for small artifacts. *Journal of Cultural Heritage*, 31, 33–45.

Schonberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Schönberger, J. L., Zheng, E., Pollefeys, M., Frahm, J.-M., 2016. Pixelwise view selection for unstructured multi-view stereo. *European Conference on Computer Vision (ECCV)*.

Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J., 2024. Dust3r: Geometric 3d vision made easy. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20697–20709.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), 600–612.

Yang, C., Li, S., Fang, J., Liang, R., Xie, L., Zhang, X., Shen, W., Tian, Q., 2024. Gaussianobject: High-quality 3d object reconstruction from four views with gaussian splatting. *arXiv preprint arXiv:2402.10259*.

Zhang, L., Rao, A., Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models. *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.