

Reconstruction of Nepali Statues: A Comparative Evaluation of Neural Radiance Fields and 3D Gaussian Splatting

Aayushma Kafle

*Dept. of Computer Engineering,
Gandaki College of Engineering and Science,
Pokhara, Nepal
be2022ce197@gces.edu.np*

Achhyut Baral

*Dept. of Computer Engineering,
Gandaki College of Engineering and Science,
Pokhara, Nepal
be2022ce175@gces.edu.np*

Jasmine Timilsina

*Dept. of Computer Engineering,
Gandaki College of Engineering and Science,
Pokhara, Nepal
be2022ce187@gces.edu.np*

Rajendra Bahadur Thapa

*Dept. of Computer Engineering,
Gandaki College of Engineering and Science,
Pokhara, Nepal
rajendra@gces.edu.np*

Abstract—Digital preservation of cultural-heritage artifacts such as statues has traditionally depended on specialized 3D scanning hardware – Light Detection and Ranging (LiDAR) scanners, structured-light sensors, or depth cameras – that are costly and largely inaccessible to smaller institutions. This paper presented an accessible reconstruction pipeline for generating 3D models of statues from ordinary multi-view 2D photographs, combining Structure-from-Motion (SfM) camera-pose estimation via COLMAP with two neural rendering paradigms: Nerfacto, an accelerated Neural Radiance Field (NeRF) variant, and 3D Gaussian Splatting (3DGS). A controlled, statue-specific comparative evaluation of Nerfacto and 3DGS was conducted on two independently captured, self-collected statue datasets (146 images captured per statue, 145 retained after quality filtering), trained under a matched 15,000-iteration budget on a single Tesla T4 GPU. Across both statues, 3DGS achieved higher reconstruction fidelity than Nerfacto – average Peak Signal-to-Noise Ratio (PSNR) of 20.80 dB and 11.15 dB versus Nerfacto’s 10.91 dB and 9.64 dB, with corresponding gains in Structural Similarity Index Measure (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) – while training $5\times$ – $7\times$ faster (17–24 minutes versus 120 minutes per statue). Given the small sample, these results were treated as preliminary evidence favoring 3DGS under resource-constrained training budgets rather than a general claim. To the authors’ knowledge, this was among the first reported applications of this comparison to Nepali cultural statues. A COLMAP camera-model incompatibility between the distortion-aware OPENCV model and the distortion-free 3DGS rasterizer, and a non-uniform default train/test split identified for one dataset, were also documented and discussed as methodological limitations.

Keywords—Neural Radiance Fields, Nerfacto, 3D Gaussian Splatting, Structure-from-Motion, Cultural Heritage Digitization, Computer Vision, Novel View Synthesis

I. INTRODUCTION

The digitization of cultural artifacts has been recognized as a growing priority for museums, universities, and heritage-

conservation bodies, yet the dominant reconstruction technologies – Light Detection and Ranging (LiDAR) scanning, structured-light scanning, and depth-camera capture – have remained expensive, bulky, and dependent on specialized operator expertise [1]. This constraint has disproportionately affected institutions in resource-limited settings, where the cost of a professional 3D scanner could be prohibitive even when the cultural material being preserved was of significant historical value.

Recent advances in neural rendering have begun to close this accessibility gap. Neural Radiance Fields (NeRF) [2] were shown to enable photorealistic novel-view synthesis from ordinary posed photographs alone. Building on this, 3D Gaussian Splatting (3DGS) [3] replaced NeRF’s implicit, ray-marched representation with an explicit set of optimizable 3D Gaussian primitives rendered through differentiable rasterization, reported to achieve comparable visual quality at substantially lower training and rendering cost. Both techniques were found to depend on Structure-from-Motion (SfM) [4] for the camera-pose estimation that precedes reconstruction.

Despite this progress, the published literature on neural-rendering-based heritage digitization was found to remain concentrated on buildings and architectural sites; direct, controlled comparisons of NeRF-family methods against 3DGS specifically for statue-class objects were found to be sparse, and to the authors’ knowledge – based on the literature surveyed in Section II rather than a formal systematic search – no prior peer-reviewed study was identified that applied either technique to Nepali cultural statues. This paper presented an evaluation pipeline built to narrow this gap. The contributions of this work were: (i) an accessible, end-to-end statue reconstruction pipeline that combined COLMAP-based Structure-from-Motion camera-pose estimation and sparse point-cloud

generation with neural rendering, requiring only a consumer camera and a cloud GPU; (ii) a controlled comparative evaluation of Nerfacto (an accelerated NeRF variant) and 3DGS, conducted under a matched iteration budget on two independently captured statue datasets, reporting PSNR, SSIM, LPIPS, and training time, together with an explicit analysis of the evaluation split itself; and (iii) practical implementation findings – including a COLMAP camera-model compatibility issue – considered relevant to future accessible heritage-digitization pipelines in Nepal and similar resource-constrained settings. Given the small sample of two statues evaluated, these findings were presented as preliminary evidence rather than a general claim of superiority.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the pipeline architecture, reconstruction methodology, and implementation, including the evaluation protocol. Section IV presents the experimental results and discussion. Section V discusses limitations, and Section VI concludes with directions for future work.

II. RELATED WORK

A. Traditional and Geometric Reconstruction

Early 3D reconstruction relied on active sensing hardware such as LiDAR and structured-light scanners, which measured geometry directly but were found to be costly and impractical for large or delicate artifacts [1]. Photogrammetry was introduced as a comparatively affordable alternative, reconstructing geometry by triangulating shared features across overlapping photographs [5], and SfM automated this into a pipeline that jointly recovered camera poses and a sparse point cloud from unordered images [6], [7]. COLMAP [4] formalized SfM as an incremental bundle-adjustment process and has remained the standard pose-estimation front end for neural reconstruction pipelines, including the one used in this study. However, photogrammetry and SfM alone were found to yield only sparse, moderately dense geometry rather than the dense, photorealistic output required for immersive visualization [8], which motivated the shift toward neural rendering.

B. Neural Radiance Fields

Mildenhall et al. [2] introduced NeRF, which represented a scene implicitly as a continuous volumetric function encoded in a multilayer perceptron (MLP); novel views were synthesized by ray-marching through the volume and integrating predicted color and density. NeRF was found to produce high visual fidelity but was computationally expensive, often requiring hours of GPU training per scene. Subsequent work targeted this bottleneck: Instant-NGP [9] replaced NeRF’s dense positional encoding with a multi-resolution hash grid, reducing training time from hours to minutes, and mip-NeRF [10] addressed aliasing by reasoning over conical frustums. Nerfstudio’s Nerfacto model, used as the NeRF baseline in this paper, combined hash-grid encoding, a learned proposal-sampling network, and per-image appearance embeddings; it was therefore an accelerated NeRF variant rather than

the original formulation, and it is referred to as “Nerfacto” throughout the remainder of this paper to avoid conflating it with vanilla NeRF. Even with such optimizations, NeRF-family methods were found to remain structurally slower than rasterization-based alternatives.

C. 3D Gaussian Splatting

Kerbl et al. [3] introduced 3DGS, which represented a scene explicitly as millions of anisotropic 3D Gaussians initialized from a sparse SfM point cloud and optimized via gradient descent with adaptive densification. Rendering was performed through differentiable rasterization rather than per-ray sampling, which was reported to enable real-time frame rates while matching or exceeding NeRF-family visual quality on several benchmarks. Open-source re-implementations such as gsplat [11] were shown to further reduce training time and memory footprint, making 3DGS particularly suitable for low-resource pipelines such as the one used in this study.

D. Comparative Studies and Heritage Applications

Recent (2024–2026) comparative studies were found to consistently report that 3DGS achieved faster training and real-time rendering with visual quality comparable to, and sometimes exceeding, NeRF-family methods, while NeRF-family methods retained an edge in memory efficiency and geometric consistency under sparse views [12]. Applications to cultural heritage were found to have grown rapidly but remained concentrated on buildings and architectural sites [13]; work specifically targeting statues and small artifacts – the closest prior research to this study – had only begun to emerge [14], and recent studies noted that standard SfM-based pose estimation could fail on small, texture-poor, or reflective sculptural surfaces, which motivated learned dense-correspondence alternatives such as DUST3R [15]. Within Nepal, institutional preservation efforts (the Kathmandu Valley Preservation Trust, the Department of Archaeology, and the National Museum of Nepal) were found to rely on conventional photographic and archival documentation. Based on the literature surveyed here – not a formal systematic search across all venues, languages, and repositories – no study was identified that applied NeRF-family methods or 3DGS to Nepali cultural statues specifically; this contribution was therefore framed as an early application to the authors’ knowledge, rather than an assertion of priority.

E. Positioning Relative to Prior Approaches

Table I summarizes how this work was positioned relative to the traditional and neural-rendering approaches discussed above. Active scanning was found to offer the most reliable geometric output but at prohibitive hardware cost, while photogrammetry and SfM alone remained inexpensive but geometry-limited. Prior neural-rendering applications to cultural heritage were found to concentrate on buildings and architectural sites rather than statues, and the single prior study identified that applied 3DGS to statues [14] integrated a web-based viewer but did not benchmark against any NeRF-family

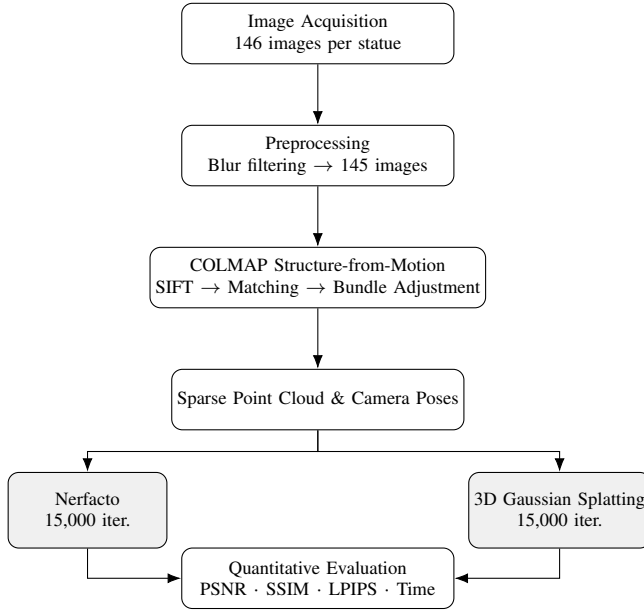


Fig. 1. Reconstruction and evaluation pipeline, from image acquisition through COLMAP-based SfM pose estimation and sparse point-cloud generation, to parallel Nerfacto/3DGS reconstruction, and quantitative evaluation.

method. The specific quantitative trade-off between Nerfacto and 3DGS for statue-class objects, and its application within a Nepali cultural-heritage context, was therefore identified as unaddressed prior to this work.

III. PIPELINE ARCHITECTURE AND METHODOLOGY

A. Overview

The reconstruction pipeline was implemented in Python and proceeded from image acquisition, through COLMAP-based camera-pose estimation and sparse point-cloud generation, to parallel Nerfacto and 3D Gaussian Splatting reconstruction, and finally to quantitative evaluation (Fig. 1).

B. Data Acquisition

All photographs used in this study were collected by the authors specifically for this work, rather than sourced from an existing public dataset. Each statue was captured as a set of overlapping photographs taken along a circular arc around the subject at multiple elevation rings, following capture guidelines drawn from the photogrammetry literature [5]: consistent diffuse lighting, an uncluttered composition, and tripod-stabilized capture to minimize motion blur. Images were captured with a handheld consumer-grade smartphone camera at a fixed 4:3 aspect ratio, chosen to reflect the low-cost, accessible capture scenario targeted by this study. Statue-1 and Statue-2 were each captured with 146 raw images; after automated blur/quality filtering, 145 images per statue were retained for reconstruction (preprocessing time ≈ 20 minutes per statue).

Unlike many controlled NeRF/3DGS benchmark captures, each statue was deliberately photographed against its natural, unmodified surrounding environment rather than a plain

or texture-less backdrop. This choice was informed by established guidance in the Structure-from-Motion literature: texture-poor or feature-sparse scenes were found to provide fewer distinctive keypoints for reliable feature matching and were reported to degrade downstream camera-pose estimation, with the recommendation that additional background objects be included when the primary subject alone lacked sufficient texture [16]. This finding was corroborated by a dedicated study that integrated background context directly into a COLMAP-based reconstruction pipeline for small objects, which found that retaining a textured background improved both the number of reconstructed points and the median reprojection error relative to reconstructing the same object without background, and further noted that camera poses were better estimated for objects with many surrounding features than for small, feature-poor objects captured in isolation [17]. Retaining each statue’s natural surroundings was therefore intended to supply additional trackable features to improve registration robustness, consistent with this evidence.

C. Camera Pose Estimation

Camera poses were recovered using COLMAP’s incremental SfM pipeline [4]. SIFT [18] keypoints were extracted from every image and matched across overlapping pairs; bundle adjustment then jointly optimized all camera poses $\{P_i\}$ and 3D point positions $\{X_j\}$ by minimizing the total reprojection error

$$E = \sum_{i,j} \rho(\|x_{ij} - \pi(P_i, X_j)\|^2) \quad (1)$$

where x_{ij} is the observed 2D keypoint of 3D point X_j in image i , $\pi(\cdot)$ is the camera projection function, and $\rho(\cdot)$ is a robust (Huber) loss [4]. For Statue-1, the OPENCV distortion-aware camera model was used, registering 140 of 145 images (96.6%) with a mean reprojection error of 1.03 px and a resulting sparse cloud of 17,663 points. For Statue-2, SfM was initially run with the OPENCV model but was re-run using the distortion-free PINHOLE model after the 3DGS rasterizer rejected the OPENCV distortion parameters (Section III-G); the PINHOLE run registered 126 of 145 images (86.9%) and produced 27,815 sparse points. COLMAP was executed headlessly (QT_QPA_PLATFORM=offscreen) to operate within a GUI-less cloud-notebook environment.

D. Nerfacto Reconstruction

The NeRF-family baseline was trained using Nerfstudio’s Nerfacto model – an accelerated variant combining hash-grid encoding, proposal-network sampling, and per-image appearance embeddings, distinct from the original NeRF formulation [2] (Section II-B). Rays were cast from each camera through image pixels; each ray was sampled at N points via the learned proposal network, and an MLP predicted a color c_i and density σ_i at each sample i . As in the original NeRF formulation [2], samples were composited into a final pixel color by discretized volume rendering,

$$\hat{C}(r) = \sum_{i=1}^N T_i (1 - \exp(-\sigma_i \delta_i)) c_i \quad (2)$$

TABLE I
QUALITATIVE POSITIONING RELATIVE TO REPRESENTATIVE PRIOR APPROACHES

Approach	Hardware Cost	Target Domain	NeRF-vs-3DGS Comparison	Nepal Context
Active scanning (LiDAR / structured light) [1]	High	General heritage objects	N/A	Not applied
Photogrammetry / SfM only [5]–[7]	Low	General scenes	N/A	Not applied
Neural rendering for architecture [13]	Low–medium	Buildings, monuments	Limited / qualitative	Not applied
3DGS for statues, web-integrated [14]	Low	Damaged statues	Not compared to NeRF	Not applied
This work	Low (camera + cloud GPU)	Statues	Quantitative: PSNR, SSIM, LPIPS, time	First application, to our knowledge

with accumulated transmittance

$$T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right) \quad (3)$$

where δ_i was the distance between adjacent samples [2]. The network was optimized against ground-truth pixels $C_{gt}(r)$ over the set of training rays $\mathcal{R}$ by minimizing the photometric mean-squared error

$$L = \sum_{r \in \mathcal{R}} \left\| \hat{C}(r) - C_{gt}(r) \right\|_2^2 \quad (4)$$

via the Adam optimizer. Both statues were trained for 15,000 iterations.

E. 3D Gaussian Splatting Reconstruction

The 3DGS pipeline followed Kerbl et al. [3]. The sparse COLMAP point cloud was used to initialize a set of anisotropic 3D Gaussians, each defined in world space as

$$G(x) = \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right), \quad \Sigma = R S S^T R^T \quad (5)$$

where μ was the Gaussian’s mean position and the covariance Σ was factorized into a rotation R and scale S to keep it positive semi-definite during optimization [3]. Each Gaussian additionally carried an opacity and view-dependent color via spherical-harmonic coefficients. Adaptive densification spawned additional Gaussians in under-reconstructed regions while all parameters were refined via gradient descent against the same photometric loss form as (4). Rendering was performed through tile-based differentiable rasterization: Gaussians overlapping a given pixel were depth-sorted and alpha-composited, front to back,

$$C = \sum_{i=1}^N c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \quad (6)$$

which mirrored the discrete Nerfacto rendering equation (2) structurally, but replaced per-ray point sampling with per-pixel Gaussian compositing, avoiding computation in empty space and enabling substantially faster training and real-time

rendering [3]. Both statues were trained for 15,000 iterations, initialized from 17,663 (Statue-1) and 27,815 (Statue-2) COLMAP points respectively.

F. Evaluation Protocol and Train/Test Split

Both pipelines were trained on the same COLMAP-derived poses and image set for each statue, under a matched 15,000-iteration budget on identical hardware (single Tesla T4, 16 GB). Reconstruction quality was measured with three standard metrics on a per-image basis, averaged across a held-out test split not used for training. Peak Signal-to-Noise Ratio was computed from the mean-squared pixel error as

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right) \quad (7)$$

where MSE was the mean-squared pixel error between the rendered and ground-truth images and MAX_I was the maximum possible pixel value. Structural Similarity [19] additionally compared local luminance, contrast, and structure between the rendered image x and ground truth y ,

$$\text{SSIM}(x, y) = \frac{(2\mu_x \mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (8)$$

and Learned Perceptual Image Patch Similarity [20] measured perceptual distance in a pretrained deep feature space,

$$\text{LPIPS}(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{y}_{hw}^l - \hat{y}_{0,hw}^l)\|_2^2 \quad (9)$$

Test views were not selected via an explicit random split designed for this study; rather, each toolchain applied its own default index-based holdout convention over the registered, capture-ordered image sequence (Nerfstudio for Nerfacto; the reference 3DGS codebase’s convention of holding out a fixed subset of indices for 3DGS), a common default in the NeRF/3DGS literature. For Statue-2, this produced a near-uniform test set of 7 images spaced approximately every 7th captured frame (image indices 3920, 3927, 3934, 3941, 3948, 3965, 3972), identical for both methods, with the remaining 138 images used for training – a reasonably well-distributed split across the capture arc. For Statue-1, by contrast, inspection of the 14 held-out indices used for 3DGS evaluation (6, 7,

8, 22, 23, 26, 28, 35, 55, 57, 59, 62, 70, 108) showed that they were not uniformly spaced: three consecutive frames (6, 7, 8) and other closely spaced pairs (22–23, 57–59–62) appeared together in the test set, alongside larger gaps elsewhere (e.g., 70→108). Because consecutive captured frames were designed to maintain substantial visual overlap (Section III-B), adjacent test frames were likely to share close viewpoints with images in the training set, which could inflate reported reconstruction quality relative to a split with guaranteed angular separation between train and test views. This was reported explicitly as a methodological limitation (Section V) rather than adjusting the already-computed metrics, since re-evaluation under a controlled, angularly-separated split was left for future work.

G. Implementation

Development followed an iterative, sprint-based workflow, with work divided into one-to-two-week cycles that covered data preprocessing, pose estimation, and reconstruction modules. The pipeline was implemented in Python 3.9+ with PyTorch and CUDA for model training, COLMAP for pose estimation, and OpenCV/Open3D for image and point-cloud processing. Training and evaluation were carried out on Google Colab Pro with a single Tesla T4 GPU (16 GB), with model checkpoints saved to Google Drive to guard against session disconnections. Version control was maintained through Git/GitHub, and Draw.io was used for system diagrams.

Image preprocessing was implemented with OpenCV, where each raw photograph was scored for sharpness using the variance of its Laplacian; frames falling below an empirically chosen threshold were discarded as blurred prior to Structure-from-Motion, yielding the 145 retained images per statue reported in Section III-B. Nerfacto reconstructions were trained and rendered through Nerfstudio’s command-line utilities, while 3D Gaussian Splatting reconstructions used the official reference implementation released alongside Kerbl et al. [3], both installed within a shared Conda environment to keep dependency versions consistent across runs. Quantitative evaluation was implemented as a separate Python script that loaded each method’s rendered test-view images alongside the corresponding ground-truth photographs, computed PSNR and SSIM directly from pixel statistics and LPIPS [20] using a pretrained deep-feature backbone, and wrote per-image and averaged results to JSON and CSV logs for both statues; these logs formed the basis of Tables II and III.

One notable implementation finding concerned camera-model compatibility: COLMAP’s OPENCV camera model, which estimated radial and tangential lens-distortion coefficients, produced sparse reconstructions that were incompatible with the reference 3DGS rasterizer, which assumed distortion-free (PINHOLE) camera intrinsics. For Statue-1 this was not an issue; for Statue-2, SfM had to be re-run with the PINHOLE model specifically to enable Gaussian Splatting training, at the cost of a lower image-registration rate (126/145 vs. the 140/145 achieved with OPENCV). This was identified as a practical constraint worth anticipating in

TABLE II
IMAGE CAPTURE AND STRUCTURE-FROM-MOTION STATISTICS FOR THE TWO RECONSTRUCTED NEPALI STATUES

Statue	Registered	Camera	3D Points
Statue-1	140/145 (96.6%)	OPENCV	17,663
Statue-2	126/145 (86.9%)	PINHOLE	27,815

TABLE III
COMPARATIVE RECONSTRUCTION QUALITY AND TRAINING TIME BETWEEN NERFACTO AND 3D GAUSSIAN SPLATTING FOR THE TWO RECONSTRUCTED NEPALI STATUES (15,000 ITERATIONS, SINGLE TESLA T4 GPU)

Statue	Method	PSNR	SSIM	LPIPS	Time
Statue-1	Nerfacto	10.91	0.163	0.819	120 min
Statue-1	3DGS	20.80	0.578	0.511	17 min
Statue-2	Nerfacto	9.64	0.123	0.857	120 min
Statue-2	3DGS	11.15	0.170	0.743	24 min

future accessible reconstruction pipelines that pair COLMAP with 3DGS.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Dataset and Structure-from-Motion Statistics

Table II summarizes the image-capture and COLMAP reconstruction statistics obtained for both statues. Both datasets began from an identical capture size (146 images, 145 retained after preprocessing), but differed in registration rate and sparse point density, a difference driven primarily by the camera-model change described in Section III-G.

B. Comparative Reconstruction Quality and Training Time

Table III reports the quantitative comparison obtained between Nerfacto and 3D Gaussian Splatting, both trained for 15,000 iterations on identical hardware. Across both statues, 3DGS was found to outperform Nerfacto on every quality metric while training substantially faster. For Statue-1, 3DGS achieved a PSNR of 20.80 dB versus Nerfacto’s 10.91 dB – an improvement of just under 10 dB – alongside higher SSIM (0.578 vs. 0.163) and lower (better) LPIPS (0.511 vs. 0.819), while training was completed in 17 minutes versus Nerfacto’s 120 minutes, a $7.1\times$ speedup. For Statue-2, the gap narrowed but 3DGS still led on every metric (PSNR 11.15 dB vs. 9.64 dB; SSIM 0.170 vs. 0.123; LPIPS 0.743 vs. 0.857), with training completed in 24 minutes versus 120 minutes, a $5\times$ speedup. As discussed in Section III-F, the Statue-1 test split was found to include several closely spaced (potentially correlated) frames, so the magnitude of Statue-1’s PSNR gap in particular was interpreted with more caution than that of Statue-2, whose test split was closer to uniform. Fig. 2 visualizes the PSNR, SSIM, and LPIPS comparison across both statues and methods alongside the tabulated values in Table III.

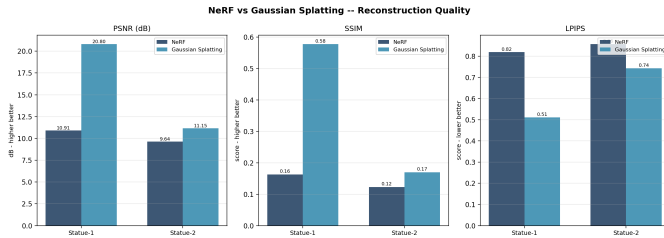


Fig. 2. Bar chart comparing PSNR, SSIM, and LPIPS for Nerfacto and 3D Gaussian Splatting across Statue-1 and Statue-2.

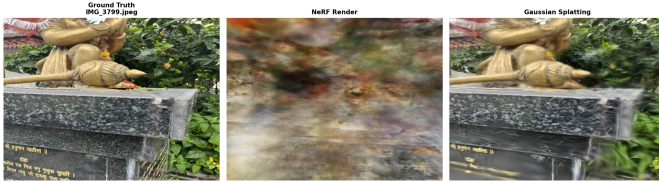


Fig. 3. Statue-1 qualitative comparison: ground-truth photograph (left), Nerfacto render (center), and 3D Gaussian Splatting render (right) for a held-out test viewpoint.

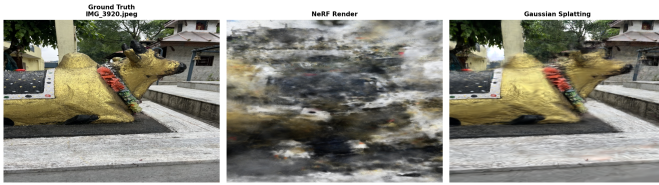


Fig. 4. Statue-2 qualitative comparison: ground-truth photograph (left), Nerfacto render (center), and 3D Gaussian Splatting render (right) for a held-out test viewpoint.

C. Qualitative Comparison

Figs. 3 and 4 present side-by-side renders obtained from held-out test viewpoints for Statue-1 and Statue-2 respectively, comparing ground-truth photographs against Nerfacto and 3DGS reconstructions at 15,000 iterations.

D. Discussion

Three observations followed from these results. First, 3DGS’s advantage in both quality and speed over Nerfacto was directionally consistent with trends reported in the broader comparative literature [12], and, for Statue-1, was larger than typically reported – plausibly because 3DGS’s explicit, rasterization-based representation converged faster than an implicit, ray-marched representation under a constrained, matched iteration budget. Given that only two statues were evaluated, and that the Statue-1 split included closely spaced test frames (Section III-F), this was presented as preliminary, consistent-direction evidence for 3DGS’s practicality under resource-constrained training rather than a generalizable claim across statue types, materials, or capture conditions.

Second, the absolute PSNR values obtained here (9.6–10.9 dB for Nerfacto; 11.2–20.8 dB for 3DGS) were lower than commonly reported in published NeRF/3DGS benchmarks (often 25–35 dB), most plausibly because both

methods were trained for only 15,000 iterations – a fraction of the 100,000–200,000-iteration budgets typical in the literature – due to the time and compute constraints of a single Colab Pro GPU. This constraint was found to affect Nerfacto considerably more than 3DGS, since implicit MLP-based representations converged more slowly per iteration; the comparison should therefore be read as a relative comparison under matched, resource-constrained conditions, not as an attempt to reproduce state-of-the-art absolute quality.

Third, the substantially lower quality obtained on Statue-2 relative to Statue-1 across both methods (e.g., 3DGS PSNR of 11.15 dB vs. 20.80 dB) was correlated with Statue-2’s lower image-registration rate (86.9% vs. 96.6%) and the camera-model change discussed in Section III-G, even though Statue-2’s test split was itself more uniform than Statue-1’s. This suggested that registration completeness, rather than split quality alone, was likely the dominant factor behind Statue-2’s lower scores, and reinforced observations in the literature that COLMAP-based pose estimation was a common failure point for accessible reconstruction pipelines [15], which motivated the DUST3R-based fallback discussed as future work in Section VI.

V. LIMITATIONS

- This evaluation covered only two statues; the consistent direction of the Nerfacto/3DGS gap across both should be read as preliminary evidence, not a generalizable claim about statue-class reconstruction. A general claim that 3DGS was “the” practical method for cultural-heritage statues was deliberately avoided.
- The held-out test split for each statue followed each toolchain’s own default index-based holdout convention rather than a split designed and validated for this study. For Statue-1 in particular, several held-out frames were consecutive or closely spaced in the capture sequence, meaning nearby, highly correlated training frames likely remained in the training set; this could inflate reported quality relative to an angularly separated split, and was a methodological limitation of the reported numbers, most relevant to Statue-1.
- Both methods were trained for only 15,000 iterations due to single-GPU time constraints; absolute reconstruction quality, particularly for Nerfacto, was below what longer training budgets typically achieved in the literature.
- Structure-from-Motion registration was incomplete for both statues (96.6% and 86.9% of captured images), and Statue-2 required a camera-model change (OPENCV → PINHOLE) to remain compatible with the 3DGS rasterizer, at the cost of registering fewer images.
- The pipeline targeted static, non-moving subjects captured under controlled conditions, and was evaluated only on held-out views from the same capture session; cross-session or cross-lighting generalization was not tested.
- Both datasets were captured with a natural (non-plain) background to aid SfM feature matching (Section III-B); this choice was not itself ablated against

a plain-background capture of the same statues, so its quantitative contribution to the reported registration rates could not be isolated in this study.

VI. CONCLUSION AND FUTURE WORK

This paper presented an accessible reconstruction and evaluation pipeline for generating 3D models of statues from ordinary multi-view photographs using COLMAP-based Structure-from-Motion together with Nerfacto and 3D Gaussian Splatting. Under a matched 15,000-iteration budget on a single Tesla T4 GPU, 3DGS was found to outperform Nerfacto on both statue datasets in PSNR, SSIM, and LPIPS while training $5\times-7\times$ faster. Given the sample of two statues and the evaluation-split caveats discussed in Sections III-F and V, this was treated as preliminary but consistent evidence favoring 3DGS under resource-constrained, accessible heritage-digitization settings, rather than a general claim. A practical camera-model incompatibility between COLMAP's distortion-aware OPENCV model and the 3DGS rasterizer was also documented, along with an evaluation-split irregularity for one dataset that was reported transparently as a limitation rather than a settled result. To the authors' knowledge, based on the literature surveyed in this paper, this was among the first reported comparisons of NeRF-family methods and 3DGS applied to Nepali cultural statues.

Future work includes: re-evaluating both methods under an angularly controlled, held-out test split designed specifically for this study, to remove the train/test correlation risk identified in Section III-F; extending the evaluation to a larger and more diverse set of statues to support broader claims; extending training to larger iteration budgets to narrow the gap with published benchmark quality; benchmarking against a public dataset such as Tanks and Temples [21] to support reproducibility and comparison with future work; ablating the effect of background texture on registration quality by capturing a matched plain-background dataset; integrating a DUST3R-based fallback [15] for scenes where COLMAP registration is incomplete; and accelerating Nerfacto-style training further for a fairer iteration-matched comparison.

REFERENCES

- [1] B. Curless and M. Levoy, "A Volumetric Method for Building Complex Models from Range Images," in *Proc. SIGGRAPH '96*, pp. 303–312, 1996.
- [2] B. Mildenhall et al., "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," in *Proc. European Conf. Computer Vision (ECCV)*, 2020.
- [3] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *ACM Trans. Graphics*, vol. 42, no. 4, Art. 139, 2023.
- [4] J. L. Schönberger and J.-M. Frahm, "Structure-from-Motion Revisited," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [5] D. C. Brown, "Close-Range Camera Calibration," *Photogrammetric Engineering*, vol. 37, no. 8, pp. 855–866, 1971.
- [6] S. Ullman, "The Interpretation of Structure from Motion," *Proc. Royal Society of London B*, vol. 203, no. 1153, pp. 405–426, 1979.
- [7] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo Tourism: Exploring Photo Collections in 3D," *ACM Trans. Graphics (SIGGRAPH)*, vol. 25, no. 3, pp. 835–846, 2006.
- [8] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2003.
- [9] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding," *ACM Trans. Graphics*, vol. 41, no. 4, Art. 102, 2022.
- [10] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields," in *Proc. IEEE/CVF ICCV*, pp. 5835–5844, 2021.
- [11] V. Ye, R. Li, J. Kerr, M. Turkulainen, B. Yi, Z. Pan, O. Seiskari, et al., "gsplat: An Open-Source Library for Gaussian Splatting," *arXiv:2409.06765*, 2024.
- [12] M. E. Atik, "Comparative Assessment of Neural Radiance Fields and 3D Gaussian Splatting for Point Cloud Generation from UAV Imagery," *Sensors*, vol. 25, no. 10, Art. 2995, 2025.
- [13] V. Croce, D. Billi, G. Caroti, A. Piemonte, L. De Luca, and P. Véron, "Comparative Assessment of Neural Radiance Fields and Photogrammetry in Digital Heritage: Impact of Varying Image Conditions on 3D Reconstruction," *Remote Sensing*, vol. 16, no. 2, Art. 301, 2024.
- [14] M. Khedr, M. Metawie, A. S. Eldin, et al., "3D Visualization of Damaged Statues Using Gaussian Splatting and Web Interface Integration," *npj Heritage Science*, vol. 13, Art. 587, 2025.
- [15] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "DUST3R: Geometric 3D Vision Made Easy," in *Proc. IEEE/CVF CVPR*, 2024.
- [16] J. Rowell, "Semantic Validation in Structure from Motion," *arXiv:2304.02420*, 2023.
- [17] P. Hruby and T. Pajdla, "Reconstructing Small 3D Objects in front of a Textured Background," *arXiv:2105.11352*, 2021.
- [18] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Key-points," *Int. J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [20] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [21] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction," *ACM Trans. Graphics*, vol. 36, no. 4, Art. 78, 2017.