10.48047/jocaaa.2026.35.01.78

# Hallucinating History: Multi-Modal Reconstruction of Lost Cultural Heritage via Generative Adversarial Networks

**Sai Teja Battula**
Independent Researcher, USA

## Abstract

Cultural heritage sites that have been destroyed through armed conflict, natural disasters, and long-term neglect represent permanent losses to historical knowledge. Traditional digital preservation requires comprehensive pre-destruction documentation either through laser scanning or through systematic photography. Such documentation exists for relatively few heritage monuments globally. Fragmented archival materials, including tourist photographs, historical postcards, amateur sketches, and travel literature, frequently survive site destruction. Multi-modal generative frameworks enable threedimensional reconstruction from heterogeneous archival sources. Text-to-image generators synthesise visual content for undocumented structural regions. Neural Radiance Fields integrate authentic and generated imagery into continuous volumetric representations. Textual descriptions extracted from historical accounts guide the interpretation of ambiguous photographic regions. Visual evidence grounds abstract linguistic references to specific spatial configurations. Architectural historians' degree reconstruction outputs against standards that encompass structural plausibility, stylistic coherence, proportional accuracy, and decorative authenticity. Expert assessment confirms substantial qualitative improvements in structural fidelity compared to photograph-only baselines. The developed framework operates with minimal specialised documentation requirements. Public domain photography and historical publications constitute sufficient input materials. Transparent communication regarding the generative nature of outputs maintains scholarly integrity. Digital repatriation becomes possible for communities separated from physical monuments through destruction or displacement.

**Keywords:** Cultural Heritage Preservation, Generative Adversarial Networks, Neural Radiance Fields, Multi-Modal Learning, Digital Archaeology, Architectural Reconstruction

## I. Introduction

It represents the irreparable destruction of cultural heritage sites and is a great loss to historical scholarship and collective human memory. Architectural monuments have been systematically targeted during armed conflicts. Natural disasters have erased structures that survived for centuries. Neglect and urban development have demolished buildings of significant cultural value. Each loss eliminates tangible connections to past civilisations and their artistic achievements.

Digital preservation has emerged as a crucial response to the vulnerability of heritage. Three-dimensional modelling technologies enable the creation of permanent digital records for endangered or destroyed sites. Image-based approaches typically apply photogrammetric principles for which overlapping photographs are used to extract geometric information. On the other hand, scanner-based methods utilise laser or structured light systems that capture surface measurements with high accuracy. Both approaches have matured significantly over recent decades, benefiting from advances in sensor technology and computational processing [1]. However, these methodologies share a fundamental limitation. Accurate reconstruction requires comprehensive documentation captured before destruction.

Photogrammetric reconstruction requires some photographs taken from different perspectives with enough overlap. Laser scanning requires physical access to the structure and ordered survey protocols. High-

10.48047/jocaaa.2026.35.01.78

resolution documentation exists for only a small fraction of global heritage sites. Major monuments in economically developed regions have received extensive digital attention. Countless structures in conflict zones and developing nations lack any systematic record. The absence of pre-destruction documentation renders conventional three-dimensional reconstruction impossible for these sites.

Alternative archival sources offer partial evidence for lost structures. Tourist photographs capture informal views from popular vantage points. Historical postcards preserve commercial imagery from earlier periods. Amateur sketches document architectural features through artistic interpretation. Travel literature contains descriptive passages detailing structural characteristics and spatial relationships. Personal correspondence and diary entries record subjective impressions of visited monuments. Scholarly publications include measured drawings and analytical descriptions. Collectively, these heterogeneous materials preserve fragmentary information regarding destroyed architecture.

The field of digital heritage has evolved considerably in addressing such reconstruction challenges. Research communities spanning computer science, archaeology, architecture, and museum studies have contributed diverse methodological perspectives. Interdisciplinary collaboration has produced innovative approaches for heritage documentation and visualisation. Survey analysis reveals substantial growth in digital heritage scholarship across multiple decades [2]. The geographical distribution of this research activity shows concentration in institutions in Europe and North America. Thematic foci of the papers have been identified to include 3D reconstructions, virtual museums, and augmented reality applications.

Despite these methodological advances, significant gaps persist between the available archival evidence and the requirements of computational reconstruction. Fragmented photographic collections lack the systematic coverage necessary for conventional photogrammetry. Textual descriptions provide rich qualitative information but resist direct geometric translation. Sketches and drawings introduce interpretive variations from original structures. No single archival source type contains sufficient information for comprehensive reconstruction.

Multi-modal generative frameworks offer potential solutions to these evidentiary limitations. Integration of visual and textual information enables mutual constraint satisfaction across modalities. Machine learning architectures can identify complementary patterns within heterogeneous datasets. Generative models synthesise coherent outputs from incomplete inputs. This research develops such a framework specifically tailored to cultural heritage reconstruction applications. The approach addresses the fundamental gap between fragmentary archival survival and the data requirements of established reconstruction methodologies.

## II. Related Work and Methodological Framework

Digital heritage preservation has progressed substantially through advances in three-dimensional modelling technologies. Image-based photogrammetry and laser scanning constitute established techniques for heritage documentation. However, these conventional methods demand comprehensive pre-destruction records unavailable for most destroyed monuments. Vision-language models have demonstrated remarkable capabilities for joint representation learning across modalities. Pre-trained architectures establish aligned feature spaces enabling cross-modal reasoning between visual and textual content. Knowledge-enhanced visual models exhibit improved generalisation through external information integration.

The proposed framework synthesises advances from multiple domains into a coherent reconstruction pipeline. The architecture employs a two-stage generative approach that distinguishes between twodimensional image completion and three-dimensional volumetric reconstruction. Generative Adversarial Networks and diffusion models operate in the first stage to repair damaged photographs and synthesise missing viewpoints from textual conditioning. These models produce two-dimensional imagery

10.48047/jocaaa.2026.35.01.78

that fills evidential gaps within archival collections. Neural Radiance Fields operate in the subsequent stage to integrate both authentic archival photographs and generated synthetic imagery into unified threedimensional representations. This separation of responsibilities enables each computational component to address its optimal task domain.

Archival material collection encompasses photographs, postcards, sketches, and textual descriptions from historical publications. Natural language processing extracts architecturally relevant passages through named entity recognition and spatial relationship identification. Text-to-image generators produce synthetic visual content for undocumented structural regions conditioned upon extracted descriptions. Neural Radiance Fields integrate authentic and generated imagery into continuous volumetric representations through differentiable rendering optimisation. Expert evaluation by architectural historians validates reconstruction quality against criteria including structural plausibility, stylistic coherence, and proportional accuracy. The framework transforms fragmentary archival survival into coherent three-dimensional visualisations. Minimal specialised documentation requirements enable applications to heritage sites globally. Transparent designation of outputs as interpretive visualisations maintains scholarly integrity throughout preservation efforts.

## III. Theoretical Framework for Multi-Modal Heritage Reconstruction

### A. Limitations of Unimodal Approaches

Single-modality reconstruction systems face inherent constraints when applied to heritage documentation. Image-only approaches depend entirely on visual patterns extracted from photographic evidence. These systems perform adequately when processing high-quality imagery with comprehensive coverage. Performance degrades substantially when encountering damaged, faded, or incomplete photographs. Historical heritage imagery rarely meets modern computational standards. Amateur photographs contain motion blur and exposure inconsistencies. Archival materials suffer from chemical degradation and physical damage.

Generative models trained exclusively on visual data exhibit characteristic failure modes. Ambiguous regions within degraded photographs trigger hallucination artefacts. The models synthesise plausible visual content that lacks historical accuracy. Architectural elements may violate period-appropriate construction techniques. Decorative features might reflect anachronistic stylistic influences. Regional building traditions become conflated with geographically distant conventions. Such errors undermine the scholarly utility of generated reconstructions.

Knowledge integration offers pathways to address these limitations. External knowledge sources provide contextual information absent from visual data alone. Textual descriptions, encyclopaedic entries, and scholarly analyses constitute rich semantic content. Recent research demonstrates that augmenting visual models with external knowledge improves recognition accuracy. The integration of linguistic knowledge enables better generalisation across visual categories [3]. Knowledge-enhanced models exhibit improved performance on novel visual concepts. This principle extends directly to heritage reconstruction applications.

Text-only generative approaches present complementary weaknesses. Natural language descriptions capture qualitative architectural characteristics effectively. Spatial relationships, proportional systems, and decorative programmes receive detailed textual treatment. However, textual descriptions resist precise geometric interpretation. The statement that a column appears "tall and slender" conveys an aesthetic impression without dimensional specificity. Volumetric reconstruction requires exact spatial coordinates.

10.48047/jocaaa.2026.35.01.78

Text-only systems generate semantically coherent but geometrically inconsistent outputs. Architectural elements may occupy impossible spatial configurations despite accurate individual descriptions.

## B. Multi-Modal Integration Principles

The combination of visual and textual modalities enables mutual constraint satisfaction. Each modality compensates for limitations inherent in the other. Textual descriptions guide the interpretation of ambiguous photographic regions. Visual evidence grounds abstract linguistic references to specific spatial configurations. This bidirectional information flow reduces reconstruction errors substantially.

Vision-and-language models have demonstrated remarkable capability for joint representation learning. Pre-training approaches establish aligned feature spaces for visual and textual content. The ViLBERT architecture introduced co-attentional transformer layers for processing image-text pairs [4]. Separate processing streams handle visual and linguistic inputs independently. Co-attention mechanisms enable selective information exchange between modalities. Visual features attend to relevant textual tokens. Textual representations incorporate salient visual regions. This architectural design preserves modalityspecific characteristics whilst enabling cross-modal reasoning.

Task-agnostic pre-training produces representations applicable across diverse downstream applications.

Transfer learning enables adaptation to heritage reconstruction without massive domain-specific datasets. Pre-trained models capture general principles of visual-linguistic correspondence. Fine-tuning aligns these representations with architectural terminology and spatial concepts. The resulting systems interpret heritage documentation through integrated multimodal understanding. Ambiguous photographic content receives disambiguation through textual context. Vague textual descriptions acquire geometric precision through visual grounding. Multi-modal integration thus establishes the theoretical foundation for heritage reconstruction from fragmentary archival evidence.

| Aspect | Image-Only Approach | Text-Only Approach | Multi-Modal Approach |
|---|---|---|---|
| Input Requirements | High-quality photographs with comprehensive coverage | Descriptive textual documentation | Combined visual and textual archival materials |
| Spatial Precision | High geometric accuracy when adequate imagery exists | Limited dimensional specificity | Balanced geometric and semantic accuracy |
| Ambiguity Handling | Produces hallucination artefacts in degraded regions | Generates geometrically inconsistent outputs | Mutual constraint satisfaction reduces errors |
| Stylistic Coherence | Dependent on visual pattern recognition alone | Captures qualitative characteristics effectively | Bidirectional information flow ensures consistency |
| Historical Accuracy | May violate periodappropriate construction principles | Semantically appropriate content generation | Textual conditioning guides historically plausible outputs |

10.48047/jocaaa.2026.35.01.78

| | | | |
|---|---|---|---|
| Failure Modes | Anachronistic element introduction in ambiguous areas | Impossible spatial configurations despite accurate descriptions | Reduced reconstruction errors through crossmodal reasoning |

Table 1. Comparative Analysis of Unimodal and Multi-Modal Reconstruction Approaches [3, 4].

# IV. Computational Pipeline Architecture

## A. Data Curation and Preprocessing

The reconstruction pipeline commences with the systematic collection of available archival materials. Target structures require comprehensive evidence gathering across multiple source categories. Photographic archives yield historical imagery from institutional and private collections. Digital repositories provide scanned postcards and tourist photographs. Library holdings contain illustrated travel accounts and scholarly publications. Each source type contributes distinct evidential value to the reconstruction process. Visual materials undergo rigorous quality assessment upon collection. Image resolution determines utility for geometric extraction. Degradation patterns indicate chemical stability and preservation requirements. Temporal classification establishes chronological relationships between captured views. Earlier photographs may depict structural configurations altered by subsequent modifications. Quality metrics inform relative reliability weightings for downstream processing.

Geometric calibration addresses optical distortions present in historical photography. Lens characteristics varied substantially across camera technologies. Perspective distortions require correction before photogrammetric analysis. Camera position estimation enables spatial registration of multiple views. Calibrated imagery supports accurate three-dimensional inference.

Textual sources undergo parallel preprocessing for architectural information extraction. Natural language processing techniques identify relevant descriptive passages. Named entity recognition locates references to specific architectural elements. Spatial relationship extraction captures relative positioning statements. Semantic representation frameworks enable structured encoding of architectural knowledge. Research demonstrates effective translation from survey data to semantic models for heritage applications [5].

Recurring architectural elements receive standardised classification. Decorative vocabularies establish consistent terminology across textual sources. The resulting structured representations support computational reasoning about architectural configurations.

## B. Generative Synthesis Mechanism

Archival collections exhibit inevitable gaps in coverage. Certain structural elevations may lack photographic documentation entirely. Interior spaces often receive less attention than exterior facades. Decorative details escape capture in distant overview photographs. Generative synthesis addresses these evidential lacunae through conditional image production.

The framework employs a sequential two-stage architecture separating two-dimensional image synthesis from three-dimensional volumetric reconstruction. Text-to-image generators operating through Generative Adversarial Networks or diffusion models constitute the first stage. These generators produce synthetic photographs depicting undocumented structural regions based upon textual conditioning derived from archival descriptions. The generators restore degraded portions of authentic photographs through inpainting operations guided by surrounding visual context and textual specifications. Stylistic features derived from authenticated period imagery constrain aesthetic consistency across generated outputs. Architectural vocabulary specifications prevent anachronistic element introduction during the synthesis process.

Neural Radiance Fields constitute the second stage and require geometrically consistent input imagery for successful volumetric reconstruction. The fundamental challenge arises from the requirement that all input

10.48047/jocaaa.2026.35.01.78

photographs depict the identical three-dimensional structure from varying viewpoints. Independently generated synthetic images may exhibit mutual geometric inconsistencies that propagate errors into volumetric representations. The framework addresses this three-dimensional consistency problem through iterative refinement protocols.

Initial coarse geometry estimation proceeds from authenticated archival photographs alone. This preliminary reconstruction establishes spatial constraints for subsequent synthetic image generation. Textto-image generators receive geometric conditioning derived from the coarse model alongside textual architectural descriptions. Generated images must satisfy both semantic requirements from textual sources and geometric requirements from preliminary reconstruction. Consistency verification compares generated viewpoints against predicted appearances from the evolving Neural Radiance Field. Inconsistent generations undergo rejection and regeneration with adjusted conditioning parameters. This iterative loop progressively refines both synthetic imagery and volumetric representation until convergence criteria achieve satisfaction.

The Neural Radiance Field formulation enables remarkable view synthesis capabilities. Input photographs establish training signals for network optimisation. The network learns implicit scene geometry through differentiable rendering. Volume rendering techniques integrate density and colour along camera rays. Rendered images receive comparison against input photographs during training. Gradient descent optimisation refines network parameters iteratively [6].

## C. Mathematical Formulation of Loss Functions

The training objective encompasses multiple loss components addressing distinct reconstruction requirements. The total loss function combines photometric reconstruction accuracy with semantic consistency constraints derived from textual conditioning:

$$Ltotal = Lphotometric + \lambda semanticLsemantic + \lambda consistencyLconsistency$$

The photometric loss measures pixel-wise correspondence between rendered and observed imagery. For each camera ray $r$ passing through the scene volume, the Neural Radiance Field predicts colour values through volumetric integration. The photometric loss aggregates squared differences across all rays within the training batch:

$$Lphotometric = \sum r \in R\hat{C}(r - C(r)22$$

where $C(r)$ denotes the ground truth colour observed along ray $r$ from authentic archival photographs, $\hat{C}(r)$ represents the colour predicted through volume rendering, and $R$ constitutes the set of sampled rays within each training iteration.

The semantic consistency loss enforces alignment between generated visual content and textual architectural descriptions. Cross-modal embedding spaces enable comparison between visual and linguistic representations. The loss function minimises distance between rendered image embeddings and corresponding textual description embeddings:

$$Lsemantic = 1 - \| Evisual(\hat{I}) \| \| Etextual(T) \| Evisual(\hat{I}) \cdot Etextual(T)$$

where $E_visual$ denotes the visual encoder mapping rendered images $\hat{I}$ to embedding space, $E_textual$ represents the textual encoder mapping architectural descriptions $T$ to the aligned embedding space, and the cosine similarity formulation ensures scale-invariant comparison.

The geometric consistency loss addresses three-dimensional coherence across synthetically generated viewpoints. Rendered depth maps from adjacent viewpoints must exhibit mutual consistency when projected into common coordinate frames:

$$Lconsistency = \sum i, j \in V \| Di(p) - \pi i \leftarrow j(Dj(p)) \| 1$$

10.48047/jocaaa.2026.35.01.78

where $D_i$ and $D_j$ represent depth maps rendered from viewpoints $i$ and $j$ respectively, $\pi$ denotes the geometric projection operator transforming depth values between viewpoints, and $V$ constitutes the set of viewpoint pairs subject to consistency enforcement.

The weighting parameters $\lambda_semantic$ and $\lambda_consistency$ balance contribution magnitudes across loss components. Empirical calibration establishes optimal values based upon reconstruction quality metrics. Higher semantic weights strengthen adherence to textual specifications whilst potentially reducing geometric precision. Higher consistency weights enforce stricter three-dimensional coherence whilst potentially limiting stylistic flexibility. The framework employs adaptive weighting schedules that adjust parameters throughout training progression.

Trained networks support geometric extraction through density field analysis. High-density regions indicate solid architectural surfaces. Density gradients reveal surface orientations and edge locations. Mesh extraction algorithms convert implicit representations to explicit geometry. The resulting threedimensional models capture structural configurations from fragmentary photographic evidence. Multimodal conditioning ensures historical plausibility throughout the reconstruction process.

| Pipeline Stage | Input Materials | Processing Technique | Output Format |
|---|---|---|---|
| Data Collection | Photographic archives, digital repositories, library holdings | Systematic evidence gathering across source categories | Consolidated archival corpus |
| Visual Assessment | Historical photographs and postcards | Quality evaluation, degradation analysis, and temporal classification | Reliability-weighted image collection |
| Geometric Calibration | Raw archival imagery | Lens distortion correction, perspective adjustment, and camera position estimation | Spatially registered visual dataset |
| Textual Processing | Travel literature, scholarly publications, correspondence | Named entity recognition, spatial relationship extraction | Structured semantic representations |
| Gap Synthesis | Textual descriptions, authenticated period imagery | Text-to-image generation with stylistic conditioning | Synthetic visual content for undocumented regions |
| Volumetric Integration | Authentic and generated imagery | Neural Radiance Field optimisation through differentiable rendering | Continuous threedimensional representation |
| Geometric Extraction | Trained Neural Radiance Field | Density field analysis, mesh extraction algorithms | Explicit threedimensional architectural model |

Table 2. Computational Pipeline Components and Processing Functions [5, 6].

10.48047/jocaaa.2026.35.01.78

# V. Expert Evaluation Methodology

## A. Assessment Criteria

Computational reconstruction outputs require validation by domain experts. Algorithmic accuracy alone cannot guarantee historical authenticity. Architectural historians possess specialised knowledge essential for quality assessment. Their knowledge encompasses construction strategies, stylistic evolution, and local construction traditions. Formal evaluation protocols structure expert engagement with generated reconstructions.

Assessment criteria address multiple dimensions of reconstruction quality. Structural plausibility examines whether depicted configurations could physically exist. Load-bearing elements must follow rational structural logic. Column spacing must permit realistic spanning solutions. Wall thicknesses must accommodate known construction materials. Violations of structural principles indicate reconstruction failures regardless of visual appearance.

Stylistic period coherence ensures temporal consistency across architectural elements. Decorative vocabularies evolved distinctly across historical periods. Mixing elements from different centuries produces anachronistic compositions. Expert evaluators identify stylistic inconsistencies invisible to computational systems. Regional variations within periods receive similar scrutiny. Building traditions differed substantially across geographic regions.

Proportional accuracy addresses dimensional relationships between elements. Classical architecture employed sophisticated proportional systems. Column heights are related mathematically to base diameters. Entablature dimensions followed codified ratios. Medieval and later traditions developed alternative proportional conventions. Expert assessment verifies adherence to period-appropriate dimensional relationships.

Three-dimensional models serve cultural heritage applications when meeting rigorous quality standards. Digital representations enable new forms of scholarly engagement with historical structures. Measurement capabilities support architectural analysis previously requiring physical access. Visualisation functions communicate spatial qualities to diverse audiences [7]. However, model utility depends entirely upon reconstruction accuracy. Expert validation thus constitutes an essential pipeline component.

## B. Qualitative Findings

Comparative evaluation protocols structure expert assessment activities. Evaluators receive paired presentations of reconstruction alternatives. Image-only baseline reconstructions demonstrate unimodal system capabilities. Multi-modal outputs incorporate textual conditioning throughout generation. Side-byside comparison reveals differential performance across reconstruction regions.

Expert assessment confirmed substantial quality improvements from textual conditioning. Regions with sparse photographic coverage benefited most significantly. Textual descriptions provided architectural guidance, absent from limited visual evidence. Period-appropriate elements appeared where image-only systems produced generic forms. Decorative details matched textual specifications rather than defaulting to common patterns.

Degraded photographic evidence similarly favoured multi-modal processing. Ambiguous visual regions received interpretation guided by the textual context. Image-only systems hallucinated plausible but historically incorrect features. Multi-modal systems constrained generation through documentary evidence. The resulting reconstructions demonstrated superior historical plausibility.

Digital cultural heritage research continues to advance preservation capabilities through technological innovation. Emerging technologies enable increasingly sophisticated documentation and reconstruction approaches. Multi-modal methods represent a significant progression beyond earlier unimodal techniques

10.48047/jocaaa.2026.35.01.78

[8]. Expert evaluation confirms practical utility alongside technical advancement. Architectural historians endorsed multi-modal reconstructions as valuable scholarly resources.

Ambiguous structural configurations presented particular evaluation interest. Certain architectural junctions permit multiple geometric interpretations from limited photographic angles. Image-only systems resolved ambiguity through statistical likelihood. Common configurations received preference regardless of specific historical evidence. Multi-modal systems incorporated textual descriptions of unusual features. Expert evaluators confirmed historically accurate resolution of ambiguous configurations through documentary integration. The evaluation findings validate multi-modal approaches for heritage reconstruction applications.

| Evaluation Criterion | Assessment Focus | Validation Indicators | Expert Domain Knowledge Required |
|---|---|---|---|
| Structural Plausibility | Physical feasibility of depicted configurations | Rational load-bearing logic, realistic spanning solutions, and appropriate wall thicknesses | Construction techniques, material properties |
| Stylistic Period Coherence | Temporal consistency across architectural elements | Absence of anachronistic compositions, period-appropriate decorative vocabulary | Stylistic evolution, historical periodisation |
| Proportional Accuracy | Dimensional relationships between elements | Adherence to codified ratios, period-appropriate proportional systems | Classical orders, regional proportional conventions |
| Decorative Element Authenticity | Historical accuracy of ornamental features | Matching textual specifications, regional tradition consistency | Decorative vocabularies, ornamental programmes |
| Regional Variation Recognition | Geographic consistency of building traditions | Absence of conflated distant conventions | Local construction practices, vernacular traditions |
| Ambiguity Resolution Quality | Handling of uncertain structural configurations | Historically accurate interpretation of junctions and transitions | Architectural junction typologies |

Table 3. Expert Evaluation Criteria and Assessment Dimensions [7, 8].

## VI. Implications for Digital Heritage Practice

### A. Scalability Considerations

The developed framework presents significant advantages for widespread heritage preservation efforts. Traditional reconstruction methodologies impose substantial documentation requirements. Laser scanning demands expensive equipment and trained operators. Systematic photogrammetric surveys require planned documentation campaigns. Such prerequisites limit application to well-resourced heritage institutions. The majority of endangered heritage sites lack access to professional documentation services. Multi-modal reconstruction operates with fundamentally different input requirements. Public domain photography provides accessible visual evidence. Historical publications contain descriptive textual content. Personal

10.48047/jocaaa.2026.35.01.78

travel accounts offer experiential documentation. Amateur sketches contribute interpretive visual records. These materials exist abundantly for heritage sites worldwide. Digital archives increasingly provide remote access to historical collections. The framework thus enables reconstruction attempts for previously inaccessible sites.

Visualisation techniques play critical roles in heritage interpretation and analysis. Three-dimensional representations communicate spatial relationships effectively. Complex archaeological data benefits from volumetric rendering approaches. Research demonstrates that advanced visualisation methods enhance interpretive capabilities significantly [9]. Interactive exploration supports hypothesis testing and analytical workflows. The framework produces outputs compatible with established visualisation platforms. Integration with existing heritage documentation systems proceeds without fundamental infrastructure changes.

Geographic scalability extends preservation capabilities globally. Heritage sites in conflict zones often lack systematic documentation. Post-destruction evidence gathering becomes the only viable approach. Distributed archival materials may reside across multiple international repositories. The framework accommodates geographically dispersed source collections. Collaborative reconstruction efforts can proceed across institutional boundaries.

## B. Ethical Dimensions

Generative reconstruction raises significant ethical considerations for heritage practice. Computational outputs differ fundamentally from documentary records. Photographs capture historical reality through optical processes. Generated content synthesises plausible interpretations from fragmentary evidence. This distinction requires transparent communication to all audiences.

Reconstructed models must receive a clear designation as interpretive visualisations. Authoritative representation claims would misrepresent the generative process. Probabilistic uncertainty pervades computationally synthesised content. Certain architectural features may reflect statistical likelihood rather than historical actuality. Viewers require an understanding of these epistemic limitations.

Documentation standards must accompany distributed reconstructions. Provenance information should identify the source materials utilised. Processing parameters merit preservation for reproducibility purposes. Confidence indicators should highlight regions of greater uncertainty. Alternative reconstructions might deserve parallel presentation. Such transparency prevents misattribution of certainty to inherently uncertain outputs.

Virtual reality technologies demonstrate effective approaches for heritage accessibility. Immersive experiences engage audiences with cultural content meaningfully. Research confirms that virtual environments improve heritage learning outcomes [10]. Accessibility considerations extend benefits to audiences unable to visit physical sites. Disabled visitors gain access previously denied by architectural barriers. Geographically distant communities reconnect with ancestral heritage through digital means. The framework thus enables digital repatriation for displaced communities. Destroyed monuments become accessible through reconstructed visualisations. Diaspora populations maintain cultural connections despite physical separation. Educational applications introduce heritage content to new generations. Ethical implementation ensures that reconstructive capabilities serve genuine preservation objectives. Transparent practices maintain scholarly integrity while substantially expanding access to heritage.

10.48047/jocaaa.2026.35.01.78

| Consideration Category | Traditional Methods | Multi-Modal Framework | Implementation Requirements |
|---|---|---|---|
| Documentation Requirements | Laser scanning equipment, trained operators, planned surveys | Public domain photography, historical publications, travel accounts | Digital archive access, natural language processing |
| Geographic Accessibility | Limited to wellresourced heritage institutions | Applicable to heritage sites globally, including conflict zones | Distributed archival material gathering |
| Infrastructure Dependencies | Expensive equipment, professional documentation services | Standard computational resources, existing visualisation platforms | Compatible with established heritage documentation systems |
| Output Designation | Documentary records of historical reality | Interpretive visualisations require transparent communication | Provenance documentation, confidence indicators |
| Audience Communication | Authoritative representation claims are appropriate | A clear distinction from documentary photographs is essential | Processing parameter preservation, uncertainty highlighting |
| Community Benefit | Physical site access required for engagement | Digital repatriation for displaced and diaspora populations | Educational application development, accessibility considerations |

Table 4. Scalability Factors and Ethical Considerations for Heritage Practice [9, 10].

## Conclusion

Destroyed cultural heritage monuments can now receive meaningful digital reconstruction through multimodal generative frameworks. The integration of visual and textual archival materials addresses fundamental gaps between fragmentary evidence and conventional photogrammetric requirements. Single-modality systems exhibit characteristic failure modes when processing degraded historical documentation. Image-only generators produce architectural hallucinations violating period-appropriate construction principles. Text-only systems generate semantically coherent yet geometrically inconsistent outputs. Multi-modal processing enables mutual constraint satisfaction across modalities. Textual descriptions disambiguate uncertain photographic regions. Visual evidence provides geometric grounding for abstract linguistic references. The two-stage architecture separates two-dimensional image synthesis through Generative Adversarial Networks or diffusion models from three-dimensional volumetric reconstruction through Neural Radiance Fields. Consistency enforcement mechanisms ensure that generated synthetic imagery maintains geometric coherence required for successful volumetric integration. Neural Radiance Fields synthesise continuous volumetric representations from combined authentic and generated imagery. Architectural historians confirm the structural plausibility and stylistic authenticity of multi-modal reconstruction outputs. The framework requires only publicly accessible archival materials for meaningful reconstruction attempts. Tourist photographs, historical publications, and personal travel accounts provide sufficient evidence. Heritage sites lacking systematic pre-destruction surveys become candidates for digital

10.48047/jocaaa.2026.35.01.78

preservation efforts. Ethical implementation demands transparent designation of outputs as interpretive visualisations rather than authoritative representations. Displaced communities gain reconnection with ancestral monuments through digital repatriation. Educational applications introduce heritage content to geographically distant audiences. Future development should address temporal reconstruction across historical periods and community-sourced evidence integration.

## References

[1] Styliani Verykokou and Charalabos Ioannidis, "An Overview on Image-Based and Scanner-Based 3D Modeling Technologies," MDPI, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/2/596

[2] S. Münste, "A survey on topics, researchers and cultures in the field of digital heritage," ISPRS Annals of Photogrammetry, 2017. [Online]. Available: https://isprs-annals.copernicus.org/articles/IV-2-W2/157/2017/isprs-annals-IV-2-W2-157-2017.pdf

[3] Sheng Shen et al., "K-LITE: Learning Transferable Visual Models with External Knowledge," 36th Conference on Neural Information Processing Systems, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/63fef0802863f47775c3563e18cbba17-PaperConference.pdf

[4] Jiasen Lu et al., "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Visionand-Language Tasks," 33rd Conference on Neural Information Processing Systems, 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/c74d97b01eae257e44aa9d5bade97bafPaper.pdf

[5] Valeria Croce et al., "From survey to semantic representation for cultural heritage: the 3D modelling of recurring architectural elements," ACTA IMEKO, 2021. [Online]. Available: https://arpi.unipi.it/bitstream/11568/1098977/1/From_survey_to_semantic_representation_for_Cultura.pdf

[6] Ben Mildenhall et al., "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," Communications of the ACM, 2022. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3503250

[7] PAOLO CIGNONI and ROBERTO SCOPIGNO, "Sampled 3D Models for CH Applications: A Viable and Enabling New Medium or Just a Technological Exercise?," ACM Journal on Computing and Cultural Heritage, 2008. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/1367080.1367082 [8] Cristina Portalés et al., "Digital Cultural Heritage," MDPI, 2018. [Online]. Available: https://www.mdpi.com/2414-4088/2/3/58

[9] Alexander Bornik and Wolfgang Neubauer, "3D Visualization Techniques for Analysis and Archaeological Interpretation of GPR Data," MDPI, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/7/1709

[10] ELMEDIN SELMANOVIĆ et al., "Improving Accessibility to Intangible Cultural Heritage Preservation Using Virtual Reality," ACM Journal on Computing and Cultural Heritage, 2020. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3377143