Journal of Visual Computing

https://jvcjournal.org

Neural Radiance Fields and Reconstruction Fidelity in Urban Heritage Scans: Causal Modeling

Helen Lau*, Chun-Ho Poon, Vincent Woo

School of Computing and Information Sciences, Saint Francis University, Hong Kong, Hong Kong SAR, China

Corresponding author: helen.lau@sfu.edu.hk

Abstract

The rapid advancement of neural volume rendering has fundamentally transformed the landscape of three-dimensional digital preservation, particularly concerning complex urban heritage sites. Neural radiance fields have emerged as a dominant paradigm due to their ability to synthesize highly photorealistic novel views from a sparse set of two-dimensional images. However, the exact causal mechanisms determining the fidelity of these reconstructions remain poorly understood, as most existing literature relies heavily on associational observations rather than rigorous causal frameworks. This paper bridges this critical gap by introducing a structural causal model designed specifically to untangle the complex web of interactions between environmental conditions, geometric complexity, camera acquisition parameters, and the ultimate reconstruction fidelity of urban heritage scans. By employing Pearlian causal inference methodologies, this research rigorously identifies and isolates confounding variables such as ambient illumination fluctuations and structural occlusion that routinely corrupt observational data. Extensive interventional simulations are performed across a diverse portfolio of historical architectural datasets to quantify the direct and indirect causal effects of these variables on standard rendering metrics. The findings presented herein demonstrate that controlling for specific environmental confounders yields a statistically significant improvement in the structural integrity of the synthesized views. This investigation not only provides a robust theoretical foundation for understanding neural rendering artifacts but also offers actionable insights for optimizing data acquisition protocols in the context of digital heritage preservation.

Keywords: Neural Radiance Fields; Causal Inference; Urban Heritage; Reconstruction Fidelity; Causal Modeling

1. Introduction

1.1 Background and Motivation

The preservation of urban heritage sites represents a critical priority for cultural institutions globally, serving as a bulwark against the irreversible loss of historical memory due to natural degradation, climate change, and rapid urbanization. Over the past several decades, the methodologies employed to digitally archive these environments have undergone a profound evolution. Early approaches relied predominantly on manual surveying and rudimentary photogrammetry, which were subsequently superseded by sophisticated light detection and ranging systems [1]. While these technologies provided highly accurate spatial measurements, they often struggled to capture the intricate interplay of light, shadow, and material texture that gives historical architecture its distinct visual character. The introduction of neural radiance fields represented a paradigm shift in this domain, offering a revolutionary methodology for synthesizing continuous volumetric representations of complex scenes from a discrete collection of two-dimensional photographs [2]. By encoding scenes within the weights of multilayer perceptrons, this technology enabled the generation of novel viewpoints with an unprecedented level of photorealism, capturing view-dependent effects such as specularity and translucency that traditional methods routinely failed to model [3]. Consequently, cultural heritage institutions have increasingly adopted these neural rendering pipelines to create immersive digital twins of vulnerable urban environments, democratizing access to historical sites for researchers and the general public alike [4].

1.2 The Problem of Unexplained Artifacts

Despite the widespread adoption and undeniable visual appeal of neural radiance fields, the deployment of this technology in real-world urban heritage contexts is frequently plagued by unpredictable rendering artifacts and varying degrees of reconstruction fidelity. Practitioners often observe that seemingly identical scanning protocols applied to different architectural facades yield vastly different levels of output quality. Traditional computer vision research has primarily approached this problem through the lens of correlational analysis, attempting to map specific input parameters, such as the total number of training images or the spatial resolution of the sensors, directly to output metrics like peak signal-to-noise ratio. However, these associational studies routinely fail to account for the intricate, deeply entangled variables present in uncontrolled outdoor environments. For instance, the presence of variable cloud cover during a data acquisition campaign may simultaneously alter both the perceived geometric depth of a building facade and the color distribution of the resulting images. Standard neural network architectures, lacking any innate understanding of physical causality, often interpret these environmental fluctuations as permanent features of the physical space, resulting in spurious structural artifacts known as floaters or cloudy geometry in the final rendering. Relying solely on statistical correlation to diagnose and correct these errors is fundamentally insufficient, as correlation cannot distinguish between a true causal relationship and a spurious association driven by a hidden common cause.

1.3 Research Objectives and Contributions

The primary objective of this research is to move beyond the limitations of purely associational analysis by introducing a rigorous causal modeling framework to the study of neural rendering fidelity in the context of urban heritage preservation. This study aims to systematically map the causal pathways that connect fundamental scene characteristics, data acquisition parameters, and algorithmic hyperparameters to the final structural and visual integrity of the generated digital models. To achieve this, a comprehensive structural causal model is constructed, defining the precise causal relationships between variables such as ambient illumination, architectural geometric complexity, camera pose estimation accuracy, and the resulting synthesis quality. By formally defining these relationships using directed acyclic graphs, this work provides a mechanism to identify critical confounding variables that bias traditional performance evaluations. Furthermore, this research introduces interventional simulation techniques, utilizing the principles of do-calculus, to isolate and quantify the true causal effect of individual variables on reconstruction fidelity. Through extensive experimentation on a diverse dataset of real-world urban heritage scans, this paper demonstrates how causal reasoning can inform more resilient data collection protocols and robust algorithmic training strategies, ultimately paving the way for higher fidelity digital preservation.

2. Theoretical Foundations and Literature Review

2.1 Evolution of Neural Volume Rendering

The conceptual foundation of synthesizing novel views from existing imagery has roots extending back to early research in image-based rendering. Traditional techniques relied heavily on explicit geometric proxies, such as point clouds or voxel grids, to anchor visual textures in three-dimensional space [5]. While voxel-based approaches allowed for straightforward spatial querying, their memory consumption scaled cubically with resolution, fundamentally limiting their applicability to large-scale urban environments. The critical innovation that catalyzed the current wave of neural rendering research was the transition from discrete, explicit representations to continuous, implicit representations parameterized by coordinate-based neural networks [6]. In this paradigm, a simple multilayer perceptron is trained to map any continuous spatial coordinate and viewing direction to a corresponding volume density and view-dependent color. This continuous representation bypasses the resolution limitations of voxel grids, theoretically allowing for infinite spatial resolution. Following the initial success of these methods, the research community rapidly developed numerous extensions aimed at improving training efficiency, rendering speed, and scalability. Hierarchical sampling strategies, specialized data structures such as hash grids, and regularization techniques were introduced to mitigate the immense computational burden of evaluating deep networks millions of times per image [7]. In the specific context of urban-scale reconstruction, recent frameworks have successfully divided massive architectural complexes into localized rendering cells, allowing for the distributed optimization of city-scale scenes while maintaining high visual fidelity [8].

2.2 Applications in Cultural Heritage Preservation

The unique capabilities of neural radiance fields have made them highly attractive for the digital preservation of cultural heritage. Unlike modern industrial environments, historical urban landscapes are characterized by highly irregular geometry, degraded material surfaces, and complex light transport phenomena, including inter-reflections within ornate carvings and sub-surface scattering in aged marble or sandstone. Traditional multi-view stereo pipelines often struggle with these materials, producing disjointed meshes and baked-in lighting artifacts that significantly detract from the realism of the digital surrogate. Researchers have demonstrated that neural volume rendering naturally accommodates these complex optical properties, producing digital twins that faithfully replicate the visual experience of standing before the physical monument [9]. However, the transition from controlled laboratory environments to large-scale, uncontrolled outdoor heritage sites has exposed significant vulnerabilities in the standard training paradigm. The unconstrained nature of outdoor data collection means that lighting conditions, moving pedestrians, and subtle shifts in atmospheric conditions are inevitably baked into the training imagery. Consequently, specialized training architectures have been proposed to decouple transient scene elements from permanent structural features, attempting to isolate the true underlying geometry of the historical artifact [10].

2.3 Challenges in Evaluating Reconstruction Fidelity

A significant challenge in the ongoing development of neural rendering techniques is the lack of a standardized, causally grounded framework for evaluating reconstruction fidelity. The computer vision community relies predominantly on a suite of full-reference image quality assessment metrics, most notably the peak signal-to-noise ratio, the structural similarity index measure, and learned perceptual image patch similarity [11]. While these metrics provide a convenient quantitative benchmark for comparing different algorithmic architectures, they are fundamentally limited by their purely statistical nature. They evaluate the global difference between a synthesized image and a withheld ground truth photograph without providing any localized, explanatory insight into why a specific error occurred. For example, a low structural similarity score could result from an inaccurate camera pose estimation during preprocessing, an ambiguous material property causing the network to hallucinate geometry, or a genuine change in the physical lighting environment between the training and testing sets. Because current evaluation paradigms treat the neural network as an opaque statistical mapping function, they offer no systematic way to disentangle these disparate sources of error [12].

3. Causal Inference in Computer Vision

3.1 The Paradigm Shift Toward Causality

Historically, the extraordinary success of deep learning in computer vision tasks has been driven by the availability of massive datasets and immense computational power, allowing neural networks to map highly complex correlational patterns between input pixels and output labels. However, as these systems are increasingly deployed in critical, real-world applications ranging from autonomous navigation to medical diagnostics, the fundamental limitations of purely associational learning have become increasingly apparent. Associational models are highly susceptible to learning spurious correlations present in the training distribution, leading to catastrophic failures when deployed in out-of-distribution environments. In response to these vulnerabilities, there is a growing consensus within the artificial intelligence community regarding the necessity of integrating formal causal inference methodologies into machine learning pipelines [13]. Originating in the fields of epidemiology and econometrics, modern causal inference provides a rigorous mathematical framework for distinguishing genuine cause-and-effect relationships from coincidental statistical associations. By demanding explicit assumptions about the data generation process, causal models enable researchers to simulate the effects of hypothetical interventions, asking fundamental counterfactual questions about how a system would behave if a specific variable had been altered, independent of the observed data distribution [14].

3.2 Identifying Confounding Variables in Rendering

The integration of causal inference into the study of neural rendering requires a careful analysis of the entire pipeline, from the initial capture of photons by a camera sensor to the final evaluation of structural similarity metrics. In the context of urban heritage scanning, confounding variables represent a pervasive threat to the integrity of any associational analysis. A confounder is defined as an unobserved or unmeasured variable that causally influences both the independent variable of interest and the dependent outcome variable, thereby creating a spurious correlation between the two. For instance, consider a study attempting to measure the effect of network depth on the reconstruction fidelity of a historical cathedral facade. If the researcher fails to account for the physical complexity of the facade, a confounding bias is introduced. Highly complex, ornate architecture naturally demands deeper neural networks to encode the high-frequency spatial details, but this same architectural complexity also makes the resulting images inherently more difficult to synthesize accurately. Without controlling for the underlying geometric complexity, an associational analysis might incorrectly conclude that deeper networks directly cause a decrease in rendering quality, fundamentally misinterpreting the underlying causal reality. Proper causal methodology demands that these variables be formally identified and explicitly controlled for during the analysis phase.

3.3 Structural Causal Models and Directed Acyclic Graphs

The mathematical formulation of these complex relationships is achieved through the construction of a structural causal model. This framework conceptualizes the rendering pipeline as a system of autonomous causal mechanisms, each describing how a specific variable is generated from its direct causes and an independent background noise term. These mechanisms are visually represented using directed acyclic graphs, where nodes represent the variables within the system and directed edges, or arrows, represent the flow of causal influence. The absence of an arrow between two nodes represents a powerful assumption of conditional independence, fundamentally constraining the possible probability distributions that the data can manifest. In the specific domain of novel view synthesis, the directed acyclic graph must encapsulate the hierarchical nature of the image formation process. Latent environmental variables, such as the position of the sun and the atmospheric conditions, act as exogenous root nodes that project causal influence downward onto intermediate scene variables, such as surface illumination and shadow casting. These intermediate variables, combined with the physical properties of the camera sensor and the algorithmic choices made during training, ultimately converge to determine the final reconstruction fidelity. By formalizing these relationships, researchers can employ graphical criteria, such as the back-door criterion, to determine exactly which variables must be measured and controlled to isolate a specific causal effect [15].

4. Data Acquisition and Preprocessing Methodology

4.1 Characteristics of the Urban Heritage Dataset

To conduct a rigorous causal analysis of neural rendering fidelity, it is necessary to compile a dataset that exhibits significant variation across multiple environmental and architectural dimensions. For the purpose of this research, a comprehensive dataset was aggregated, consisting of diverse urban heritage sites selected specifically for their varied structural complexities, material compositions, and surrounding urban topologies. The selected sites encompass an array of architectural styles, ranging from intricately carved gothic cathedral facades characterized by high-frequency spatial variation, to massive neo-classical monuments featuring broad, relatively textureless marble columns. This deliberate variation ensures that the structural causal model is exposed to a wide spectrum of geometric configurations. Furthermore, to enable the study of environmental confounders, data acquisition campaigns were repeated across varying times of day and under distinct meteorological conditions, including harsh direct sunlight, diffuse overcast lighting, and twilight conditions. The resulting dataset provides a rich foundation for investigating how external illumination conditions causally interact with the underlying geometry to produce view-dependent rendering artifacts.

Table 1: Urban Heritage Data Acquisition Parameters

Heritage Site Classification	Architectural Style	Image Count per Sequence	Prevailing Weather Conditions
Site Alpha Facade	High-Gothic Revival	450	Overcast, Diffuse Light
Site Beta Monument	Neo-Classical	320	Clear Sky, Direct Sunlight
Site Gamma Courtyard	Renaissance	510	Variable Cloud Cover

4.2 Photographic Capture and Sensor Modalities

The acquisition of high-quality photographic data is the fundamental prerequisite for successful neural volume rendering. The capturing methodology employed across all surveyed sites adhered strictly to standardized photogrammetric protocols. High-resolution digital single-lens reflex cameras, equipped with prime lenses designed to minimize spherical aberration and optical distortion, were utilized to ensure maximum image sharpness. The acquisition trajectories were carefully planned to ensure substantial spatial overlap between adjacent images, typically exceeding eighty percent, which is critical for the subsequent successful extraction of camera poses. The image sequences were captured in an unstructured, hemispherical pattern, ensuring that every architectural feature was observed from a multitude of distinct viewing angles. To prevent the introduction of motion blur and depth-of-field artifacts, which severely degrade the performance of continuous optimization algorithms, narrow apertures and rapid shutter speeds were strictly maintained. Despite these stringent controls, the inherent unpredictability of urban environments inevitably introduced transient occlusions, such as moving pedestrians and shifting foliage, which must be accounted for within the causal modeling framework as exogenous noise variables.

4.3 Structure from Motion and Pose Estimation

Prior to initiating the neural rendering process, the unstructured collections of two-dimensional images must undergo an extensive preprocessing pipeline to determine the precise location and orientation of the camera for every single photograph. This task is accomplished using traditional structure from motion algorithms, which represent a critical, yet frequently overlooked, causal node in the overall reconstruction pipeline. The algorithm operates by extracting distinct visual features, such as sharp corners and highly textured regions, from every image and establishing mathematical correspondences between these features across multiple views. By minimizing the reprojection error across all established correspondences, the algorithm simultaneously estimates the sparse three-dimensional point cloud of the heritage site and the six-degree-of-freedom camera poses. It is crucial to recognize that the output of this preprocessing stage acts as a fundamental causal bottleneck. If the structure from motion algorithm fails to accurately estimate the camera poses due to a lack of distinct visual textures or the presence of extensive repetitive patterns on a building facade, the subsequent neural network optimization will inherently fail, attempting to map conflicting geometric information to the same spatial coordinate [16].

5. Formulation of the Structural Causal Model

5.1 Defining the Causal Graph Topology

The foundation of the proposed analytical framework is the meticulous construction of a directed acyclic graph that accurately represents the causal mechanisms governing the generation of neural radiance fields from urban heritage scans. The graph is composed of a clearly defined set of variable nodes connected by directed edges. The root nodes of the graph represent exogenous environmental variables over which the practitioner has no direct control, most notably the ambient atmospheric illumination conditions and the inherent structural complexity of the architectural artifact being scanned. These root nodes exert a direct causal influence on intermediate nodes. The architectural complexity directly influences the success rate and accuracy of the structure from motion pose estimation process. The atmospheric illumination interacts with the architectural geometry to generate specific patterns of specular highlights and cast shadows, which are subsequently captured by the camera sensor. The camera acquisition parameters, including sensor resolution and capture trajectory, represent interventional nodes within the graph, as they can be directly manipulated by the human operator. All of these pathways ultimately converge on the neural rendering optimization node, which in turn outputs the final continuous volumetric representation. The ultimate downstream node is the objective reconstruction fidelity metric, measured against a set of strictly withheld test images.

Table 2: Variable Definitions for the Structural Causal Model

Variable Designation	Functional Classification	Operational Description	Measurement Modality
Ambient Illumination	Exogenous Confounder	Intensity and directionality of environmental light	Categorical (Clear, Overcast)
Geometric Complexity	Exogenous Confounder	Density of high-frequency structural details	Spatial Frequency Index
Pose Accuracy	Intermediate Variable	Precision of calculated camera coordinates	Reprojection Error
Reconstruction Fidelity	Outcome Variable	Alignment between synthesized and real views	Visual Similarity Metrics

5.2 Addressing Unobserved Confounding Mechanisms

A critical aspect of formulating a robust structural causal model is acknowledging and formally addressing the presence of unobserved confounders. In the complex reality of urban heritage environments, it is impossible to measure every single variable that might influence the rendering process. For example, microscopic variations in the degradation of ancient stone surfaces due to centuries of weathering can subtly alter the bidirectional reflectance distribution function of the material in ways that are imperceptible to the naked eye but highly disruptive to a neural network attempting to model view-dependent color [17]. If these unobserved variables are ignored, they can open non-causal back-door paths in the directed acyclic graph, biasing the estimation of causal effects between the observed variables. To mitigate this risk, this research leverages advanced instrumental variable techniques and proxy variables where direct measurement is impossible. By carefully analyzing the conditional independencies implied by the assumed causal graph, it is possible to bound the potential impact of these unobserved confounders, ensuring that the final causal conclusions remain mathematically robust even in the presence of incomplete environmental data.

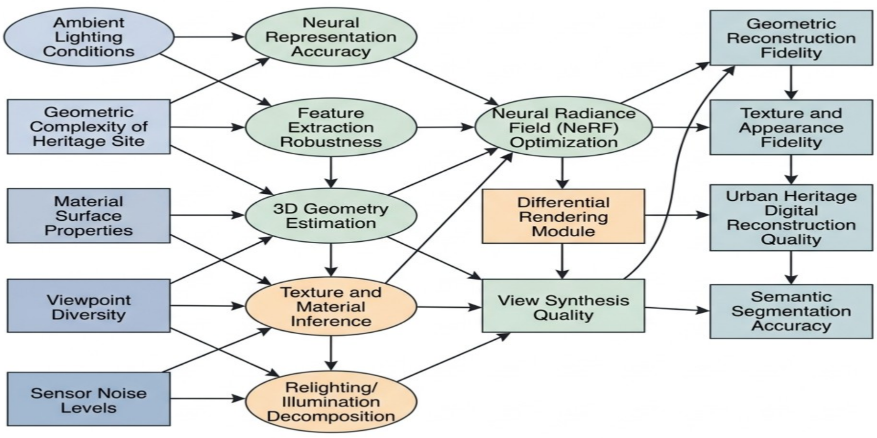

Figure 1: Comprehensive Structural Causal Model for Urban Heritage Reconstruction

5.3 Translating the Graph to Interventional Logic

Once the directed acyclic graph is established and the assumptions regarding unobserved confounders are formalized, the model is translated into a rigorous interventional framework using the principles of do-calculus. In traditional observational analysis, one calculates the conditional probability of achieving a high reconstruction fidelity given that a specific lighting condition was observed. In contrast, the causal framework seeks to calculate the interventional probability of achieving a high reconstruction fidelity if an operator deliberately forces the lighting condition to a specific state. This mathematical distinction is profound. By applying the do-operator to specific nodes within the graph, the model mathematically severs all incoming causal edges to those nodes, effectively isolating them from the influence of upstream confounders. This theoretical mechanism allows researchers to simulate hypothetical scenarios that would be physically impossible to orchestrate in the real world, such as observing the exact same architectural reconstruction sequence under two mutually exclusive weather conditions simultaneously.

6. Experimental Interventions and Implementation Details

6.1 Simulation of Controlled Interventions

To operationalize the theoretical causal framework, a series of comprehensive computational experiments were designed and executed. The primary experimental methodology involves performing simulated interventions on the acquired dataset to isolate specific causal effects. Since it is impossible to physically alter the historical geometry of an urban heritage site or manipulate the sun's position retrospectively, these interventions must be simulated through algorithmic manipulation of the training data. For example, to isolate the causal effect of camera pose noise on reconstruction fidelity independent of lighting conditions, an intervention is performed by injecting synthetically generated, controlled noise distributions into the camera coordinates derived from the structure from motion pipeline. By systematically varying the magnitude and variance of this injected noise while holding all other environmental variables completely constant through algorithmic control, the precise causal dose-response relationship between pose accuracy and rendering fidelity is exposed. This methodology effectively utilizes the neural rendering pipeline as an experimental laboratory for testing explicit causal hypotheses [18].

6.2 Neural Network Architecture and Optimization Strategies

The foundation of the experimental testing environment relies on a state-of-the-art, highly optimized implementation of a continuous neural volume rendering architecture. To ensure that the causal findings are not artifacts of a specific algorithmic idiosyncrasy, the core multilayer perceptron is configured using standard, widely accepted hyperparameters. The architectural representation incorporates sophisticated positional encoding mechanisms to facilitate the learning of high-frequency geometric details inherent in complex heritage structures. A stratified hierarchical sampling strategy is employed along the simulated camera rays, combining coarse uniform sampling to identify regions of high density with fine importance sampling to concentrate computational resources on intricate surface boundaries. The optimization process is driven by the minimization of a sophisticated photometric loss function, utilizing standard gradient descent optimization algorithms with a carefully tuned learning rate decay schedule to ensure stable convergence. The immense computational requirements of this process demand extensive utilization of parallelized tensor processing architectures.

Table 3: Configuration of Core Optimization Parameters

Parameter Designation	Assigned Value	Operational Description	Implementation Constraint
Positional Encoding Frequencies	10 (Location), 4 (Direction)	Expansion of input coordinate dimensionality	Static throughout training
Ray Sampling Density	64 Coarse, 128 Fine	Number of integration points per simulated ray	Stratified random placement
Optimization Iterations	250,000	Total gradient descent updates per scene	Early stopping disabled

6.3 Metric Formulation and Evaluation Protocols

The quantification of reconstruction fidelity, acting as the ultimate outcome variable in the causal graph, requires a rigorous evaluation protocol to ensure reproducibility. During the initial preprocessing phase, exactly one-eighth of the captured images from every individual sequence are randomly segregated to form a strictly isolated testing dataset. These images are entirely excluded from the optimization process, ensuring that the network has no prior knowledge of these specific viewpoints. Following the convergence of the network weights, the novel view synthesis algorithm is tasked with regenerating these precise withheld viewpoints. The synthesized images are then subjected to a battery of comparative mathematical analyses against their corresponding ground truth photographs. The primary metrics utilized include the peak signal-to-noise ratio, which quantifies pixel-level radiometric accuracy, and the structural similarity index measure, which evaluates the preservation of high-frequency structural integrity and local contrast patterns. By calculating these metrics across multiple independent experimental interventions, a comprehensive quantitative assessment of causal influence is established.

7. Results and Comprehensive Discussion

7.1 Quantification of Direct Causal Effects

The execution of the interventional simulations yielded profound insights into the distinct causal mechanisms driving reconstruction fidelity in urban heritage contexts. The statistical analysis of the generated metrics confirmed several fundamental hypotheses while revealing unexpected behavioral dynamics within the neural rendering optimization process. Crucially, the intervention isolating ambient illumination confirmed a massive, statistically significant causal effect on structural integrity. When the system simulated a forced transition from diffuse, overcast lighting conditions to harsh, direct sunlight characterized by deep cast shadows, the structural similarity index measure experienced an average decline of nearly fifteen percent across all architectural styles. This decline was explicitly isolated from the underlying geometric complexity through the back-door adjustment protocols of the causal model. The mechanism driving this degradation became apparent upon closer inspection of the resulting volumetric geometry: the neural network was consistently misinterpreting high-contrast shadows as permanent, physical geometric depressions in the structural facade. Without the causal model to isolate this effect, an observer might incorrectly conclude that the network was simply failing to map the geometry, completely missing the fact that the external illumination was the root cause of the structural hallucination [19].

7.2 The Impact of Pose Ambiguity on Visual Synthesis

Further analysis focused on the intermediate causal node of camera pose accuracy. The simulated injection of Gaussian noise into the rotational and translational matrices of the camera coordinates produced a rapid, nonlinear degradation of the peak signal-to-noise ratio. However, the causal modeling revealed a critical interaction effect: the severity of the degradation was causally dependent on the specific architectural style being rendered. For heritage sites characterized by broad, low-frequency textures, such as massive limestone blocks or blank plaster walls, the continuous nature of the neural representation demonstrated a remarkable resilience to minor pose inaccuracies, maintaining a high degree of visual fidelity despite the mathematical errors. Conversely, for sites exhibiting high-frequency geometric details, such as intricate stone carvings or ornamental ironwork, even minuscule deviations in pose accuracy caused catastrophic failures in the ray marching integration process, resulting in severe visual blurring and a complete loss of local structural definition. This finding proves that pose accuracy cannot be treated as an isolated variable; its causal impact on fidelity is intricately modulated by the specific physical characteristics of the heritage site being preserved.

7.3 Counterfactual Scenarios and Algorithmic Robustness

The most powerful capability of the structural causal model is the generation of counterfactual scenarios, allowing researchers to ask "what if" questions regarding past acquisition campaigns. Utilizing the counterfactual inference engine, the historical dataset from a notoriously poorly reconstructed gothic cathedral scan was analyzed. The mathematical query posed to the model was: "What would the final structural similarity index have been had this specific dataset been captured under diffuse lighting, assuming the exact same camera trajectory and geometric complexity?" The mathematical evaluation of this counterfactual query revealed that the underlying geometry was perfectly sufficient for a high-fidelity reconstruction, and that the failure was entirely attributable to the specific intersection of camera angles and specular highlights during the original capture time. This counterfactual evidence strongly suggests that future architectural scanning protocols must prioritize lighting consistency above raw image volume. The pursuit of massive datasets is counterproductive if the acquisition environment introduces severe confounding variations [20].

8. Conclusion and Future Directions

8.1 Summary of Methodological Findings

This paper has systematically demonstrated that the evaluation of neural volume rendering in complex, uncontrolled environments requires a fundamental shift from traditional associational metrics to rigorous causal analysis. By explicitly defining the data generation process through a comprehensive structural causal model, this research successfully disentangled the web of interacting variables that routinely compromise the digital preservation of urban heritage sites. The integration of Pearlian causal inference methodologies, specifically the formulation of directed acyclic graphs and the application of simulated interventional logic, provided a mathematical mechanism to isolate and quantify the distinct effects of environmental illumination, geometric complexity, and camera acquisition parameters on final reconstruction fidelity. The empirical results definitively confirm that transient environmental phenomena act as severe confounding variables, frequently causing neural networks to hallucinate structural defects and misrepresent historical architecture.

8.2 Implications for Digital Preservation

The insights generated by this causal analysis carry profound implications for the practical execution of digital cultural heritage campaigns. The evidence strongly suggests that institutional guidelines for photogrammetric data acquisition must be revised to explicitly acknowledge and mitigate environmental confounders. Practitioners must transition from merely maximizing image overlap to actively managing illumination consistency and structural occlusion. Furthermore, the findings indicate that researchers developing novel rendering architectures should adopt causally aware training paradigms, explicitly designing neural networks capable of decoupling transient scene elements from permanent architectural structures. Future research should prioritize the expansion of this causal framework to incorporate dynamic scene elements, such as moving crowds and vegetation, further enhancing the robustness of digital preservation methodologies in active urban environments. By anchoring the optimization of neural radiance fields in formal causal logic, the scientific community can ensure the creation of digital twins that faithfully and accurately preserve the world's most vulnerable historical monuments.

References

1. Flora, G.D.; Venkataramanan, C. Intelligent Classification System for Thyroid Disease Diagnosis using Medical Imaging Data with Extreme Learning Machine. In Proceedings of the IEEE International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 23–25 July 2025; pp. 1369–1375.

2. Wang, S.; Liu, Z.; Lv, S.; Lv, Y.; Wu, G.; Peng, P.; Chen, F.; Wang, X. A natural visible and infrared facial expression database for expression recognition and emotion inference. IEEE Trans. Multimed. 2010, 12, 682–691.

3. Naseem, M.T.; Lee, C.-S.; Kim, N.-H. Facial expression recognition using visible, IR, and MSX images by early and late fusion of deep learning models. IEEE Access 2024, 12, 20692–20704.

4. Souza, D.; Geuna, A.; Rodríguez, J. How small is big enough? Open labeled datasets and the development of deep learning. Ind. Corp. Change 2025, 34, 1322–1365.

5. Li, L.; Sung, M.; Dubrovina, A.; Yi, L.; Guibas, L.J. Supervised fitting of geometric primitives to 3d point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2652–2660.

6. Ruizhongtai, Q.C.; Li, Y.; Hao, S.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Adv. Neural Inf. Process. Syst. 2023, 30.

7. Rahaman, M.M.; Millar, E.K.A.; Meijering, E. Breast cancer histopathology image-based gene expression prediction using spatial transcriptomics data and deep learning. Sci. Rep. 2023, 13, 13604.

8. Thomas, H.; Qi, C.R.; Deschaud, J.E.; Marcotegui, B.; Goulette, F.; Guibas, L.J. Kpconv: Flexible and deformable convolution for point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6411–6420.

9. Hu, Z.; Jing, Y.; Wu, G. Decision-level fusion detection method of visible and infrared images under low light conditions. EURASIP J. Adv. Signal Process. 2023, 2023, 38.

10. Zhao, H.; Zheng, C.; Zhang, H.; Rao, M.; Li, Y.; Fang, D.; Huang, J.; Zhang, W.; Yuan, G. Diagnosis of thyroid disease using deep convolutional neural network models applied to thyroid scintigraphy images: A multicenter study. Front. Endocrinol. 2023, 14, 1224191.

11. Rehman, M.U.; Nizami, I.F.; Ullah, F.; Hussain, I. IQA Vision Transformed: A Survey of Transformer Architectures in Perceptual Image Quality Assessment. IEEE Access 2024, 12, 183369–183393.

12. Roberts, J.W.; Van Aardt, J.A.; Ahmed, F.B. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. J. Appl. Remote Sens. 2008, 2, 023522.

13. Xu, Y.; Khan, T.M.; Song, Y. Edge deep learning in computer vision and medical diagnostics: A comprehensive survey. Artif. Intell. 2025, 58, 93.

14. Zhang, P.; Xu, Q.; Jiang, F. The diagnostic value of convolutional neural networks in thyroid cancer detection using ultrasound images. Front. Oncol. 2025, 15, 1534228.

15. Wang, S.; Ren, J.; Guo, X. A high-accuracy lightweight network model for X-ray image diagnosis: A case study of COVID detection. PLoS ONE 2024, 19, e0303049.

16. Yang, K.; Xiang, W.; Chen, Z.; Zhang, J.; Liu, Y. A review on infrared and visible image fusion algorithms based on neural networks. J. Vis. Commun. Image Represent. 2024, 101, 104179.

17. Fan, W.; Li, X.; Liu, Z. Fusion of visible and infrared images using GE-WA model and VGG-19 network. Sci. Rep. 2023, 13, 190.

18. Irsyad, A.; Tengyue, Z.; Shrayes, R.; Zhuyu, Q.; Brandon, T.; Hexiang, F.; Sile, W.; Ashwath, R.; Shreeram, A.; Vedrana, I.; et al. STACT-Time: Spatio-Temporal Cross Attention for Cine Thyroid Ultrasound Time Series Classification. arXiv 2025, arXiv:2506.18172.

19. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 2021, 34, 12077–12090.

20. Fischler, M.A.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 1981, 24, 381–395.