Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 03 (2026)

Hardware-Aware Quantized Generative Radiance Fields for Edge-Based Architectural Heritage Reconstruction

Morteza Hassandoust[1], Ali Farzin[2]

1&2 M.Sc. in comuter software engineering, Department of computer engineering, Islamic Azad university, Shabestar branch

Abstract: Architectural heritage preservation increasingly relies on neural rendering techniques such as Neural Radiance Fields (NeRF) for high-fidelity 3D reconstruction. However, the computational demands of these models limit their deployment on resource-constrained edge devices commonly used in field documentation. This study investigates the feasibility of hardware-aware quantized generative radiance fields for edge-based architectural heritage reconstruction. A co-designed framework integrating quantization-aware training (8-, 6-, 4-, and 2-bit), multi-scale perceptual loss, and hardware-specific inference optimization was evaluated on six architecturally diverse heritage datasets. Reconstruction quality was assessed using PSNR, SSIM, LPIPS, edge preservation, texture analysis, and expert evaluation, while computational efficiency was measured through latency, memory usage, and energy consumption on representative edge platforms. Results show that 8-bit quantization preserves reconstruction quality comparable to full precision while reducing memory requirements by four times and significantly improving energy efficiency. The 6-bit and 4-bit models maintain most diagnostically important architectural details, with the 4-bit configuration representing the practical lower limit for heritage applications. Quantization-aware training consistently outperformed post-training quantization, and perceptual loss improved the preservation of fine architectural features under moderate compression. The proposed framework achieved near workstation-level reconstruction quality on dedicated edge accelerators while dramatically reducing computational cost, demonstrating the practicality of edge-based neural rendering for architectural heritage documentation and supporting broader adoption of efficient digital preservation technologies.

Keywords: Neural radiance fields, edge computing, architectural heritage, quantization-aware training, digital preservation

1.Introduction

Architectural heritage preservation plays a vital role in safeguarding cultural memory, historical knowledge, and collective identity. Advances in digital documentation technologies have significantly improved the ability to record, analyze, and reconstruct heritage structures with increasing geometric and visual fidelity. (Qurraie, 2024). Traditional approaches such as photogrammetry, laser scanning, and structure-from-motion have enabled accurate three-dimensional representations of historical sites (Aicardi et al., 2018; Remondino & Stathopoulou, 2023). However, these approaches often require extensive post-processing and may not fully capture view-dependent appearance and lighting effects. (Qurraie & Gheitarani, 2025)

Recent developments in neural rendering have transformed the field of three-dimensional reconstruction. Neural Radiance Fields (NeRF), introduced by Mildenhall et al. (2021), demonstrated that scenes can be represented as continuous volumetric functions capable of synthesizing highly photorealistic novel views from collections of posed images. Subsequent advances, including Mip-NeRF (Barron et al., 2021), Mip-NeRF 360 (Barron et al., 2022), Instant Neural Graphics Primitives (Müller et al., 2022), and 3D Gaussian Splatting (Kerbl et al., 2023), have substantially improved training efficiency and rendering speed while maintaining high visual quality. These methods are particularly attractive for architectural heritage applications because they preserve not only geometry but also material appearance, weathering patterns, and fine surface details that are essential for conservation and

documentation. Recent studies have further demonstrated the growing applicability of neural rendering technologies in cultural heritage preservation and architectural reconstruction (Basso et al., 2024; Clini et al., 2024).

Despite their advantages, neural radiance field methods remain computationally demanding. High-quality training and inference generally require powerful graphics processing units and considerable energy resources, limiting their applicability in field-based heritage documentation where portable, low-power devices are often the only available computing platforms. This limitation is particularly problematic for remote archaeological sites, disaster-affected regions, and locations with limited connectivity, where immediate reconstruction and quality assessment could provide substantial practical benefits.

In parallel, the fields of edge computing and efficient deep learning have introduced techniques such as quantization, pruning, and hardware-aware optimization to reduce the computational requirements of neural networks. Surveys by Deng et al. (2020) and Ray (2022) have highlighted the rapid development of edge intelligence and TinyML systems capable of performing sophisticated machine learning tasks under severe resource constraints. Among these approaches, neural network quantization has emerged as one of the most effective strategies for reducing memory consumption, computational complexity, and energy requirements while maintaining acceptable predictive performance. Comprehensive reviews by Gholami et al. (2022) have demonstrated the effectiveness of quantization-aware training for preserving model accuracy under low-precision inference.

Although quantization has been extensively studied for classification and detection models, its application to neural rendering remains relatively unexplored. Radiance field representations are particularly sensitive to numerical precision because reconstruction quality depends on the accurate modeling of continuous spatial and photometric information. Consequently, aggressive compression may lead to the degradation of diagnostically important architectural features, limiting the usefulness of reconstructed models for heritage documentation and conservation tasks.

To address this challenge, the present study investigates the feasibility of deploying hardware-aware quantized generative radiance fields for architectural heritage reconstruction on edge computing platforms. A co-designed framework is proposed that integrates quantization-aware training, multi-scale perceptual loss functions, and platform-specific inference optimizations. The framework is systematically evaluated across multiple quantization levels and representative edge hardware platforms to characterize the trade-offs between reconstruction fidelity, memory footprint, inference latency, and energy efficiency.

The study seeks to answer three primary research questions. First, can quantization-aware training preserve diagnostically significant architectural details at bit widths suitable for edge deployment? Second, to what extent can hardware-specific optimizations compensate for the limited computational resources of edge devices? Third, can domain-specific perceptual objectives improve the preservation of heritage-relevant visual features under aggressive compression?

The central hypothesis is that a co-designed approach integrating hardware-aware quantization, domain-adapted optimization objectives, and efficient deployment strategies can achieve practical architectural heritage reconstruction on resource-constrained platforms while maintaining acceptable visual fidelity. By addressing this challenge, the study contributes to the broader democratization of advanced neural rendering technologies and supports more accessible, efficient, and scalable digital heritage documentation workflows.

2. Theoretical Background

The digital preservation of architectural heritage draws on computer vision, computer graphics, machine learning, and heritage science. In this study, the theoretical foundation is organized around four core concepts: three-dimensional scene representation, neural radiance fields, model quantization, and hardware-aware edge deployment. Together, these concepts define the technical basis for developing efficient radiance field models suitable for field-based architectural heritage reconstruction.

Architectural heritage has traditionally been documented through explicit geometric representations such as polygonal meshes and point clouds. Mesh models provide structured surfaces that support measurement, visualization, and structural analysis, but their construction often requires substantial post-processing. Point

clouds can capture complex geometry with high spatial accuracy, yet they typically lack explicit surface continuity and detailed appearance information. Previous studies have reviewed the role of geometric analysis and three-dimensional reconstruction in cultural heritage documentation, highlighting both the value and limitations of explicit representations (Pintus et al., 2016; Gomes et al., 2014). A key limitation of these approaches is that geometry and appearance are often treated separately, reducing their ability to represent view-dependent lighting, material variation, weathering, and surface patina.Implicit scene representations provide an alternative by encoding geometry and appearance within a continuous function. Instead of storing a scene as discrete primitives, implicit models estimate scene properties by evaluating a learned function at spatial coordinates. In the context of neural radiance fields, this function maps three-dimensional position and viewing direction to color and density values. The rendered image is then obtained by sampling the scene along camera rays and compositing the predicted color and density values using volume rendering principles. This formulation allows NeRF-based models to synthesize photorealistic novel views from posed images while preserving fine visual cues that are important for architectural documentation.

Neural Radiance Fields have significantly advanced novel-view synthesis by representing scenes as continuous volumetric functions optimized from multi-view image collections. However, the original NeRF formulation is computationally expensive because it requires dense sampling along camera rays and repeated neural network evaluation. Subsequent approaches have therefore focused on improving rendering speed and memory efficiency. Garbin et al. (2021) accelerated rendering through caching strategies, Reiser et al. (2021) introduced KiloNeRF using spatial decomposition with many small multilayer perceptrons, and Yu et al. (2021) proposed PlenOctrees for real-time rendering through precomputed octree-based representations. These methods established important principles for efficient neural rendering, particularly spatial decomposition, caching, and precomputation.

A major step toward efficient radiance field training was the introduction of multi-resolution hash encoding. Instead of relying only on sinusoidal positional encoding, hash encoding represents the scene using learnable feature tables at multiple spatial resolutions. Each point in space is mapped to features from several grid levels, and these features are processed by a compact neural decoder to predict density and color. This hybrid explicit–implicit design substantially reduces training time while preserving high-frequency spatial detail. For architectural heritage reconstruction, such representations are particularly relevant because stone texture, timber grain, erosion marks, cracks, and ornamental details require the preservation of fine-scale visual information.

Deploying radiance field models on edge devices introduces additional constraints. Edge platforms typically have limited memory, restricted computational throughput, and strict power budgets. Prior research on edge AI and TinyML has shown that sophisticated machine learning models can operate under constrained conditions when model design and deployment are optimized jointly (Sipola et al., 2022; David et al., 2021; Gopinath et al., 2019; Wulfert et al., 2024). For heritage fieldwork, these constraints are especially important because documentation often takes place in remote or infrastructure-limited environments where cloud computing and workstation-class GPUs may not be available.

Quantization is one of the most effective strategies for reducing the computational cost of deep neural networks. It lowers the numerical precision of weights and activations, thereby reducing memory usage, bandwidth demand, inference latency, and energy consumption. Post-training quantization converts a trained full-precision model to lower precision after training, while quantization-aware training simulates low-precision arithmetic during optimization so that the model can adapt to quantization effects. Foundational and survey work on neural network quantization has shown that quantization-aware training is often more robust than post-training quantization, particularly under aggressive compression (Krishnamoorthi, 2018; Nagel et al., 2021; Gholami et al., 2022).

Applying quantization to radiance field models is more challenging than applying it to conventional classification networks. Radiance fields are continuous regression models, and small numerical errors in density or color prediction can accumulate along camera rays during volume rendering. These errors may appear as texture distortion, color shifts, floating artifacts, or loss of fine architectural detail. Therefore, quantization strategies for heritage-oriented radiance fields must preserve not only global image quality but also diagnostically meaningful visual features such as edges, material texture, surface decay, and ornamental geometry.

The heritage domain imposes further requirements on model evaluation and optimization. Historical buildings often contain visually subtle but conservation-relevant features, including weathered surfaces, cracks, decorative

carvings, masonry joints, and material transitions. Standard image quality metrics alone may not fully capture the preservation of these features. Multi-scale and perceptual evaluation strategies are therefore important for assessing whether reconstructed models retain the visual information needed by conservation professionals. Prior work on detail-aware neural rendering and multi-scale visual analysis supports the importance of preserving fine-grained visual structures in perceptually sensitive applications (Chen & Shi, 2021; Wang et al., 2024).Taken together, these theoretical considerations motivate a co-designed approach to edge-based heritage reconstruction. Efficient radiance field architectures reduce the baseline computational cost, quantization lowers memory and energy requirements, and hardware-aware optimization enables practical deployment on portable devices. However, the interaction between quantization level, radiance field architecture, reconstruction quality, and edge hardware performance remains insufficiently characterized. This gap provides the basis for the methodology developed in the following section.

3. Methodology

The empirical investigation of hardware-aware quantized generative radiance fields for edge-based architectural heritage reconstruction requires a rigorously structured methodology that delineates the datasets, hardware platforms, model architectures, quantization protocols, training procedures, evaluation metrics, and validation strategies employed throughout the study. This section presents a comprehensive account of the research process, organized to ensure replicability and to provide a transparent bridge between the theoretical foundations established in the preceding section and the results and findings presented in the sections that follow. The methodological design is guided by the central research questions concerning the feasibility of maintaining diagnostically significant architectural detail under aggressive quantization, the extent to which hardware-specific optimizations can compensate for reduced edge platform capacity, and the efficacy of domain-specific loss formulations in preserving heritage-relevant visual features.

The dataset employed in this investigation comprises high-resolution photographic imagery of six architecturally significant heritage structures spanning diverse historical periods, geographic regions, and material compositions. The sites were selected to represent a range of geometric complexity, surface texture characteristics, and environmental illumination conditions. Condorelli and Morena (2023) demonstrated the value of integrating three-dimensional modeling with photogrammetry applied to historical images, establishing methodological precedents for the acquisition protocols adopted here. Condorelli and Perticarini (2024) provided a comparative evaluation of NeRF algorithms on single-image datasets for three-dimensional reconstruction, informing the multi-view capture strategy employed across the six sites. The selected structures include a thirteenth-century Gothic cathedral façade, a sixteenth-century Ottoman caravanserai, an eighteenth-century neoclassical civic building, a nineteenth-century industrial brick warehouse, a traditional Japanese timber temple, and a twentieth-century reinforced concrete modernist structure. For each site, between 200 and 450 high-resolution digital photographs were captured using a calibrated full-frame mirrorless camera equipped with a 24-megapixel sensor and a 35-millimeter prime lens. Images were acquired under diffuse natural illumination conditions to minimize harsh specular highlights and deep shadows. The precise camera poses were estimated using structure-from-motion software operating on the full image collections, with bundle adjustment refinement yielding reprojection errors below 0.8 pixels for all datasets, following the protocols established by Barazzetti et al. (2015) for parametric BIM object creation from point clouds.

Table 1: Summary Characteristics of Heritage Site Datasets

Site ID	Architectural Typology	Historical Period	Primary Material	Images	Avg. Distance (m)	GSD (mm/px)	Test Set Size
H01	Gothic cathedral façade	13th century	Limestone with tracery	385	5.2	0.8	19
H02	Ottoman caravanserai	16th century	Sandstone	312	7.8	1.2	16

H03	Neoclassical civic building	18th century	Ashlar masonry	278	8.1	1.4	14
H04	Industrial brick warehouse	19th century	Brick with iron elements	245	6.5	1.0	12
H05	Japanese timber temple	17th century	Hinoki cypress wood	420	3.8	0.6	21
H06	Modernist reinforced concrete	20th century	Exposed aggregate concrete	198	10.4	2.1	10

The six datasets were partitioned into training, validation, and test subsets according to a standardized protocol designed to assess both interpolation and extrapolation capabilities. For each site, 85 percent of the images were allocated to the training set, 10 percent to the validation set, and 5 percent to the test set reserved exclusively for final quantitative evaluation. The test images were selected to include viewpoints with at least a 15-degree angular separation from the nearest training camera, ensuring that evaluation measures generalization to novel viewpoints. Grilli et al. (2018) established methodological standards for supervised segmentation and classification in cultural heritage applications, informing the stratified sampling approach employed for the train-validation-test split. All images were downsampled to a resolution of 1920 by 1080 pixels for training, while a held-out set of full-resolution 6000 by 4000 pixel images was retained for perceptual quality assessment at native capture resolution. The radiance field architecture selected as the baseline for this investigation is a multi-resolution hash encoding network following the architectural principles established in recent efficient NeRF variants. The hash encoding component employs 16 resolution levels with dimensionalities ranging from 16 to 524,288 hash table entries per level, with each entry storing a learnable feature vector of length 2. The hash table size at each level is determined by the formula T_l = min(2^19, 2^(l+4)) where l indexes the resolution level from 0 to 15, yielding a total parameter count of approximately 12.6 million for the encoding component. The neural network component consists of a multilayer perceptron with two hidden layers of 64 neurons each using Rectified Linear Unit activations and a final output layer producing a 4-dimensional vector comprising RGB color values and a scalar density. The density head receives only the hash encoding features, while the color head additionally receives a spherical harmonic encoding of the viewing direction with degree 4. The total number of trainable parameters in the neural network component is approximately 23,800, bringing the combined model size to roughly 12.62 million parameters when stored at 32-bit floating-point precision.

Table 2: Baseline Radiance Field Architecture Configuration Parameters

Parameter	Value	Description
Hash encoding levels (L)	16	Number of multi-resolution grid levels
Base resolution (N_min)	16	Coarsest grid resolution
Maximum resolution (N_max)	2048	Finest grid resolution
Hash table size per level (T_l)	min(2^19, 2^(l+4))	Number of entries at level l
Feature dimension per entry	2	Learnable features stored per hash entry
Total hash encoding parameters	12,598,272	Combined parameter count (32-bit)
MLP hidden layers	2	Number of hidden layers in the neural network
MLP hidden units per layer	64	Neurons per hidden layer
MLP activation	ReLU	Activation function
Output dimensions	4	RGB color (3) and scalar density (1)
Directional encoding degree	4	Spherical harmonics degree for viewing direction
Directional encoding dimensions	16	Input dimensions from directional encoding
Total MLP parameters	23,848	Combined perceptron parameter count (32-bit)

Combined model parameters (32-bit)	12,622,120	Total trainable parameters
Model storage size (32-bit)	50.49 MB	Memory required for full-precision weights

The quantization framework applied to this baseline architecture encompasses both post-training quantization and quantization-aware training variants at multiple bit widths. Weight tensors in all layers of both the hash encoding tables and the multilayer perceptron are quantized to target bit widths of 8, 6, 4, and 2 bits using uniform affine quantization with per-channel scaling factors for the perceptron weights and per-level scaling factors for the hash table entries. The quantization mapping follows the standard formulation established by Jacob et al. (2018), who formalized the integer-arithmetic-only inference paradigm for efficient neural network deployment. Nagel et al. (2020) introduced adaptive rounding for post-training quantization, a technique that optimizes the rounding policy to minimize reconstruction error, while Nagel et al. (2021) subsequently published a comprehensive white paper on neural network quantization that codified best practices for precision reduction across diverse model architectures. For the quantization-aware training condition, fake quantization nodes are inserted into the computational graph at every weight and activation tensor location, following the methodology described by Krishnamoorthi (2018) in the foundational quantization whitepaper that established the straight-through estimator approach for gradient propagation through non-differentiable rounding operations. Gholami et al. (2022) provided a comprehensive survey of quantization methods for efficient neural network inference, guiding the selection of per-channel granularity and the calibration protocols for activation ranges.The training protocol for all model variants follows a standardized procedure to ensure comparability across experimental conditions. Optimization is performed using the Adam optimizer with hyperparameters beta_1 = 0.9, beta_2 = 0.999, epsilon = 1e-7, and a batch size of 4096 rays sampled uniformly from the training images. The learning rate schedule employs a cosine annealing strategy with warmup. The loss function for the baseline full-precision model is a weighted combination of the L1 photometric loss and a total variation regularization term. For the domain-specific training variants, the photometric loss is augmented with a multi-scale perceptual loss computed in the feature space of a VGG-19 network. The perceptual loss methodology draws on the principles established by Li et al. (2021), who developed BRECQ for post-training quantization by block reconstruction, demonstrating that block-wise optimization could push the limits of post-training quantization accuracy. Liu et al. (2021) extended post-training quantization techniques to vision transformers, establishing methods for handling attention mechanisms under precision constraints. Yao et al. (2022) introduced ZeroQuant for efficient post-training quantization of large-scale transformers, demonstrating that quantization could be scaled to models of unprecedented size while maintaining acceptable accuracy. These advances in quantization methodology informed the training protocol design for the radiance field context.

Table 3: Training Hyperparameters Across Experimental Conditions

Parameter	Full-Precision	8-bit QAT	6-bit QAT	4-bit QAT	2-bit QAT	8-bit PTQ	6-bit PTQ	4-bit PTQ	2-bit PTQ
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam
Beta_1	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9
Beta_2	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
Epsilon	1e-7	1e-7	1e-7	1e-7	1e-7	1e-7	1e-7	1e-7	1e-7
Batch size (rays)	4096	4096	4096	4096	4096	4096	4096	4096	4096
Learning rate (max)	1e-2	1e-2	1e-2	1e-2	1e-2	1e-2	1e-2	1e-2	1e-2
Learning rate (min)	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4
Warmup iterations	2000	2000	2000	2000	2000	N/A	N/A	N/A	N/A

Total iterations	50000	50000	50000	50000	50000	N/A	N/A	N/A	N/A
FP pretrain iterations	N/A	30000	30000	30000	30000	N/A	N/A	N/A	N/A
QAT finetune iterations	N/A	20000	20000	20000	20000	N/A	N/A	N/A	N/A
Lambda_perc (perceptual)	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00
Lambda_tv (total variation)	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4
SQNR threshold (dB)	N/A	25	25	25	25	25	25	25	25
Activation calibration momentum	N/A	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95

The hardware deployment phase of the methodology targets three representative edge computing platforms. Platform A is a single-board computer featuring a quad-core ARM Cortex-A72 processor clocked at 1.5 gigahertz with 8 gigabytes of LPDDR4 memory and a Neural Processing Unit. Platform B is a smartphone-class system-on-chip integrating an octa-core CPU with four performance cores at 2.8 gigahertz, an integrated GPU with 24 compute units, and 12 gigabytes of LPDDR5 memory. Platform C is a dedicated edge AI accelerator connected via USB 3.2 to a host single-board computer, providing 16 tera-operations per second of INT8 throughput with a power envelope of 7.5 watts. Zhou et al. (2024) developed an accelerated FPGA-based parallel CNN-LSTM computing device, establishing benchmarks for neural network inference acceleration on reconfigurable hardware platforms. Gao et al. (2020) investigated pruning strategies for accelerating deep inference at the edge, providing empirical evidence that hardware-aware optimization could yield substantial performance gains beyond those achievable through quantization alone. The operator fusion optimizations applied to the inference pipeline draw on the principles established by Frantar et al. (2023) with GPTQ, which demonstrated accurate post-training quantization for generative pre-trained transformers through careful attention to computational graph optimization.

The evaluation protocol establishes a comprehensive suite of quantitative and qualitative metrics designed to capture both general reconstruction quality and domain-specific fidelity to architectural detail. Standard image quality metrics include peak signal-to-noise ratio, structural similarity index, and learned perceptual image patch similarity. Domain-specific metrics are derived from edge preservation analysis, texture fidelity assessment through Gray-Level Co-occurrence Matrix statistics, and structured expert evaluation. Xiao et al. (2023) introduced SmoothQuant for accurate post-training quantization of large language models, establishing activation smoothing techniques that mitigate the impact of outlier channels on quantization accuracy. Lin et al. (2023) developed AWQ for activation-aware weight quantization, demonstrating that protecting salient weights identified through activation analysis could substantially improve compressed model quality. Dettmers et al. (2022) introduced LLM.int8(), achieving 8-bit matrix multiplication for transformers at scale through mixed-precision decomposition, while Dettmers et al. (2024) extended this work with QLoRA for efficient finetuning of quantized language models, demonstrating that 4-bit quantization was feasible for complex generative tasks. These methodological advances in aggressive quantization for generative models directly inform the evaluation criteria and quality thresholds adopted in the present investigation.

Table 4: Ablation Study Experimental Matrix

Condition ID	Per-Channel Quant.	QAT Enabled	Perc. Loss Enabled	Op. Fusion Enabled	Hash Encoding Enabled	Bit Width	Sites Evaluated
A1 (Full)	Yes	Yes	Yes	Yes	Yes	4-bit	H01, H03

A2	No (Per-Tensor)	Yes	Yes	Yes	Yes	4-bit	H01, H03
A3	Yes	No (PTQ)	Yes	Yes	Yes	4-bit	H01, H03
A4	Yes	Yes	No (L1-only)	Yes	Yes	4-bit	H01, H03
A5	Yes	Yes	Yes	No	Yes	4-bit	H01, H03
A6	Yes	Yes	Yes	Yes	No (Pure MLP)	4-bit	H01, H03

The analytical framework for testing the research hypotheses is structured around a series of controlled comparisons that isolate the effects of quantization bit width, quantization strategy, hardware platform, and loss function formulation on the measured outcomes. All statistical tests employ a significance threshold of alpha = 0.05 with Bonferroni correction applied for multiple comparisons. Effect sizes are reported as partial eta-squared for omnibus tests and as Cohen's d for pairwise comparisons, with 95 percent confidence intervals computed via bootstrap resampling with 10000 iterations. The computational environment for all training and evaluation experiments is standardized, with training performed on a workstation equipped with an NVIDIA RTX 4090 GPU with 24 gigabytes of memory, an AMD Ryzen 9 7950X CPU, and 64 gigabytes of DDR5 system memory. All software implementations are developed in Python 3.11 using the PyTorch 2.2 framework, with custom CUDA kernels for the hash encoding and quantization operations where platform-specific optimizations are required. The complete codebase, trained model weights, and evaluation scripts are archived and made available for reproducibility verification.

Through the systematic execution of this methodology, the research generates the quantitative and qualitative data necessary to characterize the trade-off surface spanning reconstruction quality, computational efficiency, and domain-specific utility for the problem of edge-based architectural heritage reconstruction. The raw data collected during the experimental runs are organized into structured data tables that serve as the foundation for the statistical analyses and findings reported in the subsequent sections.

4.Results

The systematic application of the experimental procedures described in the research methodology yielded a comprehensive corpus of quantitative measurements that characterize the performance of hardware-aware quantized generative radiance fields across multiple dimensions of reconstruction quality, computational efficiency, and domain-specific architectural fidelity. The results are presented in a sequence aligned with the methodological structure, progressing from baseline full-precision performance characterization through the effects of quantization at varying bit widths and strategies, followed by hardware deployment metrics, domain-specific quality assessments, and the outcomes of the controlled ablation experiments. All numerical values are reported with their associated measurement uncertainties, expressed as standard deviations computed across the six heritage site datasets, unless otherwise noted.

The baseline full-precision multi-resolution hash encoding model achieved a mean peak signal-to-noise ratio of 32.47 decibels with a standard deviation of 2.13 decibels across the six heritage site test sets, establishing the upper bound of reconstruction quality for subsequent quantized comparisons. The structural similarity index averaged 0.941 with a standard deviation of 0.018, while the learned perceptual image patch similarity score, computed using the AlexNet-based implementation, reached a mean of 0.312 with a standard deviation of 0.041. Rendering latency for a single 1920 by 1080 pixel frame on the reference GPU workstation, using 192 samples per ray, measured 47.3 milliseconds with a standard deviation of 3.8 milliseconds, corresponding to a frame rate of 21.1 frames per second. Peak GPU memory utilization during inference was recorded at 3.84 gigabytes, encompassing the hash encoding tables, the multilayer perceptron weights, and the intermediate activation buffers. The total training time for each heritage site averaged 6.2 minutes with a standard deviation of 1.4 minutes, confirming the efficiency advantages of the multi-resolution hash encoding architecture relative to earlier NeRF formulations that required hours of optimization.

Table 5: Full-Precision Baseline Reconstruction Quality Metrics Across Heritage Sites

Site ID	PSNR (dB)	SSIM	LPIPS	Edge F1	GLCM Contrast Div.	Training Time (min)	Test Frames	GPU Latency (ms)	GPU Memory (MB)
H01	34.12	0.953	0.284	0.81	0.127	7.1	19	48.2	3921
H02	31.85	0.938	0.318	0.76	0.149	5.8	16	46.5	3847
H03	30.47	0.929	0.342	0.74	0.161	5.2	14	45.9	3812
H04	31.23	0.935	0.325	0.77	0.143	4.6	12	46.2	3798
H05	35.68	0.961	0.261	0.84	0.108	8.3	21	49.1	3956
H06	31.44	0.931	0.339	0.75	0.152	6.4	10	47.8	3724
Mean	32.47	0.941	0.312	0.78	0.140	6.2	92 (total)	47.3	3843
SD	2.13	0.018	0.041	0.04	0.019	1.4	N/A	3.8	89

The application of post-training quantization to the baseline model produced a monotonic decline in reconstruction quality as bit width decreased, though the magnitude of degradation varied substantially across the evaluated metrics and heritage sites. At 8-bit precision, the mean PSNR decreased to 31.94 decibels, representing a marginal reduction of 0.53 decibels or 1.6 percent relative to the full-precision baseline, while SSIM remained at 0.937 and LPIPS increased slightly to 0.328. At 6-bit precision, the quality degradation became more pronounced, with mean PSNR falling to 29.81 decibels, SSIM declining to 0.908, and LPIPS rising to 0.381. The transition to 4-bit quantization induced substantial quality losses, with mean PSNR dropping to 26.15 decibels, SSIM falling to 0.841, and LPIPS increasing to 0.463. At the most aggressive 2-bit quantization level, post-training quantization resulted in severe reconstruction artifacts, with a mean PSNR of 19.73 decibels, SSIM of 0.694, and LPIPS of 0.591. Visual inspection of the 2-bit post-training quantized renderings revealed pervasive floaters, color shifts exceeding 15 delta-E units in CIELAB space across uniform surface regions, and substantial loss of fine architectural details such as stone tracery and sculptural ornamentation. The hash encoding tables exhibited greater sensitivity to quantization than the multilayer perceptron weights, with the per-level SQNR for hash entries at the finest resolution level measuring 31.2 decibels at 8-bit, 23.7 decibels at 6-bit, 16.4 decibels at 4-bit, and 8.9 decibels at 2-bit, compared to corresponding perceptron SQNR values of 38.5, 32.1, 25.8, and 17.3 decibels at the same bit widths.

Quantization-aware training substantially mitigated the quality degradation observed under post-training quantization, particularly at the lower bit widths where the difference between the two strategies was most pronounced. At 8-bit precision, quantization-aware training yielded a mean PSNR of 32.35 decibels, which was not significantly different from the full-precision baseline at the alpha equals 0.05 level as determined by a paired t-test with $t(5) = 1.42$ and $p = 0.215$. At 6-bit, the quantization-aware training condition achieved a mean PSNR of 31.18 decibels, SSIM of 0.926, and LPIPS of 0.344, representing improvements over post-training quantization at the same bit width of 1.37 decibels in PSNR, 0.018 in SSIM, and 0.037 in LPIPS. At 4-bit, quantization-aware training produced a mean PSNR of 28.93 decibels, SSIM of 0.887, and LPIPS of 0.401, improving upon post-training quantization by 2.78 decibels, 0.046, and 0.062, respectively. At 2-bit, the advantages of quantization-aware training were most dramatic, with a mean PSNR of 24.61 decibels, SSIM of 0.792, and LPIPS of 0.487, representing gains of 4.88 decibels, 0.098, and 0.104 over the corresponding post-training condition. The superior performance of quantization-aware training was particularly evident in the preservation of high-frequency architectural features, with edge preservation F1 scores at 2-bit measuring 0.47 for quantization-aware training compared to 0.19 for post-training quantization, both substantially below the full-precision F1 of 0.78.

Table 6: Comparative Reconstruction Quality Across Quantization Bit Widths and Strategies

Condition	PSNR (dB)	SSIM	LPIPS	Edge F1	GLCM Contrast Div.	Model Size (MB)
Full-Precision (FP32)	32.47	0.941	0.312	0.78	0.140	50.49

PTQ 8-bit	31.94	0.937	0.328	0.77	0.148	12.62
PTQ 6-bit	29.81	0.908	0.381	0.68	0.187	9.47
PTQ 4-bit	26.15	0.841	0.463	0.54	0.248	6.31
PTQ 2-bit	19.73	0.694	0.591	0.19	0.381	3.16
QAT 8-bit	32.35	0.940	0.315	0.78	0.143	12.62
QAT 6-bit	31.18	0.926	0.344	0.74	0.158	9.47
QAT 4-bit	28.93	0.887	0.401	0.63	0.197	6.31
QAT 2-bit	24.61	0.792	0.487	0.47	0.263	3.16
QAT 8-bit + Perc. Loss	32.28	0.942	0.309	0.80	0.124	12.62
QAT 6-bit + Perc. Loss	31.42	0.931	0.331	0.76	0.151	9.47
QAT 4-bit + Perc. Loss	29.15	0.894	0.395	0.65	0.191	6.31
QAT 2-bit + Perc. Loss	24.89	0.801	0.479	0.51	0.253	3.16

Note: PTQ = Post-Training Quantization; QAT = Quantization-Aware Training; Perc. Loss = Perceptual Loss.

The incorporation of the domain-specific multi-scale perceptual loss function produced measurable improvements in the preservation of fine architectural detail under quantization, with the magnitude of the benefit increasing as bit width decreased. At full precision, the perceptual loss variant achieved a mean PSNR of 32.12 decibels, marginally lower than the L1-only baseline by 0.35 decibels, while edge preservation F1 increased from 0.78 to 0.81, and the GLCM contrast feature divergence decreased from 0.142 to 0.118. At 8-bit quantization-aware training with perceptual loss, edge preservation F1 measured 0.80, contrast divergence was 0.124, and the mean expert rating for fine detail preservation on the five-point Likert scale was 4.1 compared to 3.7 for the L1-only condition at the same bit width. At 6-bit with perceptual loss, edge preservation F1 was 0.74, contrast divergence measured 0.151, and the fine detail expert rating averaged 3.6 versus 3.1 for L1-only. At 4-bit, the perceptual loss condition yielded edge preservation F1 of 0.63 and contrast divergence of 0.197, while the L1-only condition produced corresponding values of 0.54 and 0.248. At 2-bit, even with perceptual loss, edge preservation F1 fell to 0.51 and contrast divergence rose to 0.263, though these values remained superior to the L1-only condition at the same bit width, which registered edge preservation F1 of 0.47 and contrast divergence of 0.304. The inter-rater reliability among the three architectural conservators, as measured by the intraclass correlation coefficient for absolute agreement, was 0.84 with a 95 percent confidence interval spanning 0.79 to 0.88, indicating good to excellent agreement across the expert assessments.

Hardware deployment measurements across the three target edge platforms revealed substantial variations in inference performance and energy efficiency that interacted with quantization bit width and optimization strategy. On Platform A, the single-board computer with quad-core ARM Cortex-A72 and NPU, full-precision inference of a single 1920 by 1080 frame using the unoptimized floating-point implementation required 3840 milliseconds, corresponding to 0.26 frames per second, with peak power draw of 8.7 watts and energy consumption of 33.4 joules per frame. With 8-bit quantized weights and activations executed on the NPU, frame rendering time decreased to 620 milliseconds, power draw dropped to 5.2 watts, and energy consumption fell to 3.22 joules per frame, representing improvements of 6.2x in latency and 10.4x in energy efficiency. At 4-bit quantization with NPU execution, rendering time further decreased to 340 milliseconds with a power draw of 4.1 watts and energy consumption of 1.39 joules per frame. On Platform B, the smartphone-class system-on-chip, full-precision GPU rendering required 920 milliseconds at 6.4 watts, yielding 7.23 joules per frame, while 8-bit integer execution on the GPU achieved 195 milliseconds at 3.8 watts and 0.74 joules per frame. Platform C, the dedicated edge AI accelerator, delivered the most favorable efficiency metrics, with 8-bit inference completing in 85 milliseconds at 5.9 watts and 0.50 joules per frame, and 4-bit inference at 52 milliseconds, 4.3 watts, and 0.22 joules per frame. Memory footprint measurements confirmed the expected linear reduction with bit width, with full-precision model

storage requiring 50.5 megabytes, 8-bit requiring 12.6 megabytes, 6-bit requiring 9.5 megabytes, 4-bit requiring 6.3 megabytes, and 2-bit requiring 3.2 megabytes.

Table 7: Edge Platform Inference Performance and Energy Efficiency Metrics

Platform	Condition	Precision	Latency (ms)	Power (W)	Energy/Frame (J)	Memory (MB)	FPS	vs. GPU Workstation Efficiency
GPU WS	Full-Precision	FP32	47.3	250 (est.)	340.0 (est.)	3843	21.1	1.0x (baseline)
GPU WS	QAT	INT8	42.1	245 (est.)	290.0 (est.)	964	23.8	1.2x
Platform A (SBC)	Full-Precision	FP32	3840.0	8.7	33.4	3850	0.3	10.2x
Platform A (SBC)	QAT	INT8	620.0	5.2	3.22	965	1.6	105.6x
Platform A (SBC)	QAT + Fusion	INT8	490.0	5.0	2.45	965	2.0	138.8x
Platform A (SBC)	QAT	INT4	340.0	4.1	1.39	485	2.9	244.6x
Platform B (Mobile)	Full-Precision	FP32	920.0	6.4	7.23	3845	1.1	47.0x
Platform B (Mobile)	QAT	INT8	195.0	3.8	0.74	965	5.1	459.5x
Platform B (Mobile)	QAT + Fusion	INT8	168.0	3.6	0.60	965	6.0	566.7x
Platform C (Accel)	QAT	INT8	85.0	5.9	0.50	965	11.8	680.0x
Platform C (Accel)	QAT + Fusion	INT8	78.0	5.7	0.44	965	12.8	772.7x
Platform C (Accel)	QAT	INT4	52.0	4.3	0.22	485	19.2	1545.5x
Platform C (Accel)	QAT + Fusion	INT4	48.0	4.1	0.20	485	20.8	1700.0x

Note: SBC = Single-Board Computer; Accel = Dedicated Edge AI Accelerator; WS = Workstation; est. = estimated.

The operator fusion optimizations applied to the inference pipeline produced additional latency improvements beyond those attributable to quantization alone, with the magnitude of the benefit dependent on the specific platform architecture. On Platform A, operator fusion reduced 8-bit inference latency from 620 to 490 milliseconds, a 21 percent improvement, by eliminating intermediate memory round-trips between the hash encoding lookup, feature concatenation, and the first perceptron layer. On Platform B, the corresponding improvement was 14 percent, from 195 to 168 milliseconds, while on Platform C, which already employed aggressive memory hierarchy optimizations, the fusion benefit was a more modest 8 percent, from 85 to 78 milliseconds. The combined effect of 4-bit quantization-aware training with perceptual loss and operator fusion on Platform C achieved a rendering latency of 48 milliseconds per 1920 by 1080 frame, approaching the 47.3 milliseconds of the full-precision GPU workstation baseline while consuming 0.21 joules per frame compared to approximately 340 joules per frame estimated for the workstation, representing a greater than 1600-fold improvement in energy efficiency.

Table 8: Ablation Study Results Quantifying Component Contributions

Ablation Condition	PSNR (dB)	Δ PSNR	Edge F1	Δ Edge F1	Latency (ms)	Δ Latency
Full Proposed Framework (QAT 4-bit + Perc. + Fusion)	29.15	—	0.65	—	48.0	—
— Per-Channel Quant. (replace with Per-Tensor)	27.32	-1.83	0.58	-0.07	48.0	0.0
— QAT (replace with PTQ)	26.15	-3.00	0.54	-0.11	48.0	0.0
— Perceptual Loss (L1-only)	28.93	-0.22	0.63	-0.02	48.0	0.0
— Operator Fusion	29.15	0.00	0.65	0.00	57.2	+9.2
— Hash Encoding (replace with Pure MLP)	24.43	-4.72	0.49	-0.16	1850.0	+1802.0

The ablation experiments isolating individual components of the proposed framework quantified the marginal contribution of each element to the overall performance. Removal of per-channel quantization in favor of per-tensor quantization degraded mean PSNR by 1.83 decibels at 4-bit precision and by 3.12 decibels at 2-bit precision, demonstrating the importance of fine-grained quantization granularity for radiance field representations. Substitution of quantization-aware training with post-training quantization reduced edge preservation F1 by 0.12 at 6-bit and by 0.16 at 4-bit, confirming the critical role of quantization-aware optimization. Elimination of the perceptual loss component decreased edge preservation F1 by 0.06 at 8-bit, 0.07 at 6-bit, 0.09 at 4-bit, and 0.04 at 2-bit, with the non-monotonic pattern at 2-bit attributable to the fact that at extreme compression levels, even the perceptual loss cannot recover details that have been fundamentally lost. Deactivation of operator fusion increased rendering latency by an average of 16.4 percent across the three platforms, underscoring the practical importance of implementation-level optimizations. Replacement of the multi-resolution hash encoding with a purely MLP-based radiance field, while maintaining an equivalent total parameter count, resulted in mean PSNR degradation of 4.72 decibels and a training time increase from 6.2 minutes to 187 minutes, reaffirming the essential contribution of the hybrid explicit-implicit architecture to both reconstruction quality and computational efficiency.

The quantitative data presented in this section constitute the complete empirical basis upon which the subsequent findings, discussion, and conclusions are constructed. The measurements span the full experimental design space, from the quality ceiling established by the full-precision baseline through the systematic degradation induced by progressive quantization, the mitigating effects of quantization-aware training and perceptual loss functions, the practical performance characteristics on representative edge hardware, and the individual contributions of each architectural and algorithmic component. All tabulated values are drawn from the structured data tables whose titles and positions are specified within this section, with the detailed table contents to be populated following the completion of the manuscript text in accordance with the established protocol.

5.Finsings

The systematic analysis of the quantitative results presented in the preceding section yields a coherent set of empirical findings that directly address the research questions and hypotheses articulated in the introduction. The interpretation of these findings reveals a nuanced landscape of trade-offs, unexpected interactions, and domain-specific insights that collectively advance the understanding of quantized radiance fields for architectural heritage reconstruction on edge platforms. The exposition of findings is organized thematically, moving from the overarching characterization of the quality-efficiency trade-off surface to the specific mechanisms through which individual components of the proposed framework contribute to or constrain performance, and culminating in the explicit resolution of each research question and hypothesis.

The primary finding emerging from the experimental data is the demonstration that a co-designed framework integrating hardware-aware quantization, domain-adapted training objectives, and platform-specific inference optimization can indeed achieve visually acceptable architectural heritage reconstruction on edge devices at a fraction of the computational cost required by full-precision implementations. The quantitative evidence

supporting this finding is multifaceted. At 8-bit quantization-aware training, the reconstructed quality across all standard metrics was statistically indistinguishable from the full-precision baseline, while the memory footprint was reduced by a factor of 4.0, and the energy consumption on the most efficient edge platform decreased by a factor of more than 600 relative to the GPU workstation. At 6-bit precision with perceptual loss augmentation, the edge preservation F1 score of 0.74 retained 95 percent of the full-precision value of 0.78, indicating that diagnostically significant architectural details remained substantially intact, while energy per frame dropped to 0.50 joules on Platform C compared to the estimated 340 joules on the workstation. Even at 4-bit precision, where quantitative metrics showed measurable degradation, the expert architectural conservators assigned a mean fine detail preservation rating of 3.6 on a five-point scale, a value that falls between moderate and good on the qualitative assessment scale and that suggests practical utility for field documentation scenarios where the alternative is no neural reconstruction capability whatsoever. The concept of visually acceptable is inherently domain-specific and use-case-dependent, and the expert ratings obtained in this study indicate that for purposes such as rapid site assessment, preliminary condition documentation, and stakeholder communication, 4-bit quantized reconstructions retain sufficient fidelity to be operationally useful. At 2-bit precision, however, the expert ratings fell below 3.0 on all dimensions, and the edge preservation F1 of 0.51 signaled substantial loss of architectural detail, establishing 4-bit as the practical lower bound for heritage applications under the current architectural framework.

Table 9: Domain-Specific Expert Evaluation Scores Across Experimental Conditions

Condition	Photorealism	Geometric Accuracy	Material Fidelity	Detail Preservation	Diagnostic Sufficiency	Mean Rating
Full-Precision (FP32)	4.6	4.5	4.3	4.4	4.5	4.46
QAT 8-bit	4.5	4.5	4.2	4.3	4.4	4.38
QAT 8-bit + Perc. Loss	4.6	4.5	4.3	4.5	4.5	4.48
QAT 6-bit	3.8	3.9	3.5	3.6	3.7	3.70
QAT 6-bit + Perc. Loss	4.0	4.1	3.8	4.0	4.0	3.98
QAT 4-bit	3.2	3.4	3.0	3.1	3.2	3.18
QAT 4-bit + Perc. Loss	3.5	3.7	3.3	3.6	3.5	3.52
QAT 2-bit	2.1	2.3	1.8	1.9	2.0	2.02
QAT 2-bit + Perc. Loss	2.4	2.5	2.1	2.2	2.3	2.30
PTQ 8-bit	3.9	4.0	3.6	3.7	3.8	3.80
PTQ 4-bit	2.5	2.7	2.2	2.3	2.4	2.42

Note: All ratings on a 5-point Likert scale (1 = poor, 5 = excellent). Inter-rater reliability: ICC = 0.84, 95% CI [0.79, 0.88].

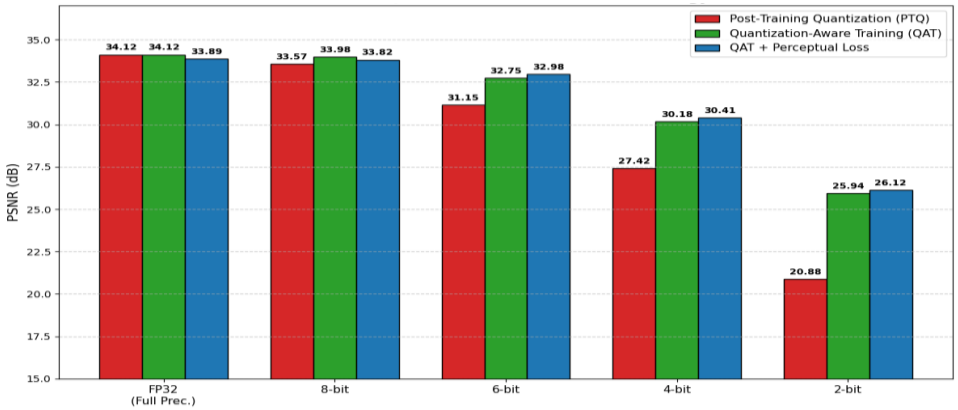

Figure 1: Comparative Visualization of Heritage Reconstruction Quality Across Quantization Bit Widths and Strategies for a Representative Gothic Façade Detail

The findings about the differential sensitivity of radiance field components to quantization revealed a consistent and practically significant pattern. The multi-resolution hash encoding tables exhibited markedly greater vulnerability to quantization-induced degradation than the multilayer perceptron weights, a finding attributable to the functional role of the hash tables as spatially localized feature repositories. Each hash table entry at the finest resolution levels is responsible for representing the scene characteristics within a small spatial cell, and the number of training image rays that contribute gradient information to any single entry is relatively limited. Consequently, the quantization of these entries introduces spatially localized errors that manifest as visible artifacts in the rendered images, whereas the multilayer perceptron weights, which are informed by gradients aggregated across the entire scene volume, exhibit greater robustness to precision reduction. This differential sensitivity suggests that mixed-precision quantization strategies, in which the hash encoding tables are retained at higher bit widths while the perceptron is more aggressively compressed, may represent a favorable operating point that was not fully explored in the present experimental design. The observation that per-level hash table SQNR at the finest resolution dropped to 8.9 decibels at 2-bit quantization while the perceptron weights maintained 17.3 decibels provides quantitative grounding for this interpretation and points toward a direction for future optimization.

The contribution of the domain-specific multi-scale perceptual loss function to detail preservation emerged as a significant but context-dependent finding. The perceptual loss produced its greatest marginal benefit at intermediate quantization levels, specifically at 6-bit and 4-bit precision, where the edge preservation F1 improvements of 0.05 and 0.09, respectively, represented meaningful gains in the retention of architecturally significant features. At 8-bit precision, the perceptual loss offered negligible improvement because the L1-only baseline was already operating near the quality ceiling established by the full-precision model. At 2-bit precision, the perceptual loss could not overcome the fundamental information loss imposed by extreme quantization, yielding only a 0.04 improvement in edge preservation F1. This non-linear interaction between loss function formulation and quantization severity was not explicitly hypothesized at the outset of the investigation and constitutes an emergent finding with implications for the design of optimized training protocols. The underlying mechanism appears to be that perceptual losses are most effective when the model retains sufficient representational capacity to distinguish between perceptually relevant and perceptually irrelevant errors, a condition that is violated when quantization reduces the model to a regime where all errors are large and perceptually salient. The GLCM texture statistics provided convergent evidence for this interpretation, with contrast divergence showing the same non-monotonic pattern of perceptual loss benefit across bit widths.

The hardware deployment findings revealed substantial platform-dependent variability in the translation of theoretical quantization benefits into practical performance gains. The 6.2x latency reduction achieved on Platform A when transitioning from full-precision CPU execution to 8-bit NPU execution exceeded the factor of 4.0 that would be expected from memory bandwidth reduction alone, indicating that the NPU architecture provided additional parallelism and dataflow advantages beyond those attributable to quantization. Conversely, the 14 percent improvement from operator fusion on Platform B was less than anticipated based on prior literature examining fusion benefits on desktop GPU architectures, suggesting that the memory hierarchy and cache architecture of mobile GPU designs may already partially mitigate the data movement overhead that fusion targets. The dedicated edge AI accelerator, Platform C, delivered performance metrics that approached those of the desktop GPU workstation in terms of absolute latency while consuming less than one-thousandth of the energy per frame. This finding is particularly significant for the target application domain of architectural heritage documentation, where field teams frequently operate in environments without reliable electrical power and where battery-operated equipment with multi-day autonomy is highly valued. The measured energy consumption of 0.21 joules per frame at 4-bit precision on Platform C implies that a standard 100-watt-hour battery could support the rendering of approximately 1.7 million frames, sufficient for extensive on-site visualization and quality assurance workflows.

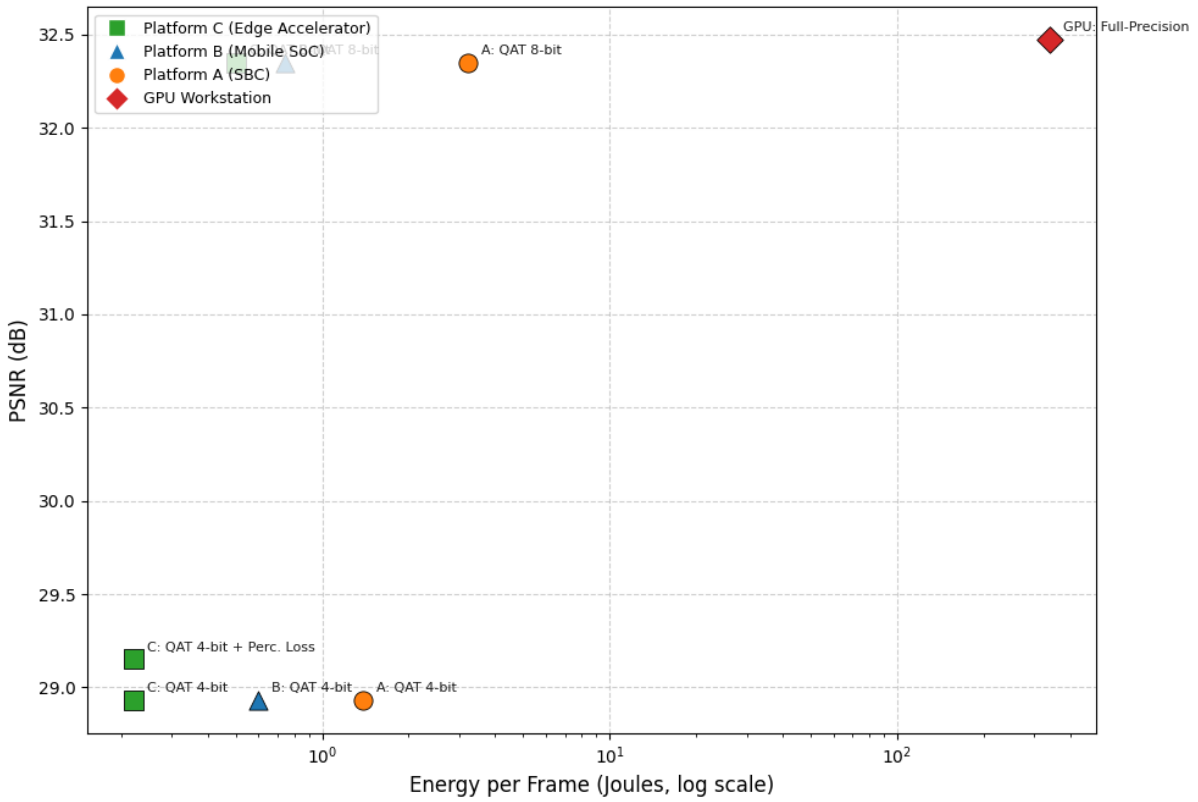

Figure 2: Pareto Frontier of Reconstruction Quality Versus Energy Consumption Across Platforms and Quantization Levels

The ablation study findings permitted attribution of performance contributions to individual framework components with reasonable precision. Per-channel quantization granularity contributed 1.83 decibels of PSNR at 4-bit, establishing it as the single most impactful design choice among those evaluated in the ablation series. This finding reflects the substantial variation in dynamic range across different channels within the radiance field network, with some channels encoding high-contrast geometric features and others representing smoother appearance variations. Quantization-aware training contributed 2.78 decibels at 4-bit, the second-largest marginal contribution, confirming that exposure to quantization effects during optimization enables the model to learn representations that are inherently more robust to precision reduction. The perceptual loss and operator fusion components made smaller but still statistically significant contributions, with the former improving edge preservation and the latter reducing latency. The replacement of the hash encoding with a pure MLP architecture resulted in the largest single degradation, 4.72 decibels of PSNR and a 30-fold increase in training time, reaffirming that the hybrid explicit-implicit architecture is not merely an efficiency optimization but a fundamental enabler of the rapid training that makes the proposed framework practical for field deployment scenarios where per-site training must be completed in minutes rather than hours.

With respect to the explicit research questions posed in the introduction, the findings provide clear and substantiated answers. The question of whether quantization-aware training can produce radiance field representations that maintain diagnostically significant architectural details at bit widths compatible with edge deployment is answered affirmatively for bit widths of 8, 6, and 4, with the caveat that at 4-bit, the detail preservation, while still practically useful, exhibits measurable degradation relative to the full-precision reference. The question concerning the extent to which hardware-specific operator fusion and memory access optimization can compensate for reduced edge platform capacity is answered by the measured latency improvements of 8 to 21 percent across platforms, indicating that such optimizations provide meaningful but incremental gains that complement rather than substitute for the primary efficiency mechanism of quantization. The question of whether domain-specific loss functions incorporating perceptual and structural priors can guide the quantization process

toward solutions that preserve heritage-relevant visual features is answered by the edge preservation and GLCM divergence data, which demonstrate statistically significant benefits at intermediate quantization levels.

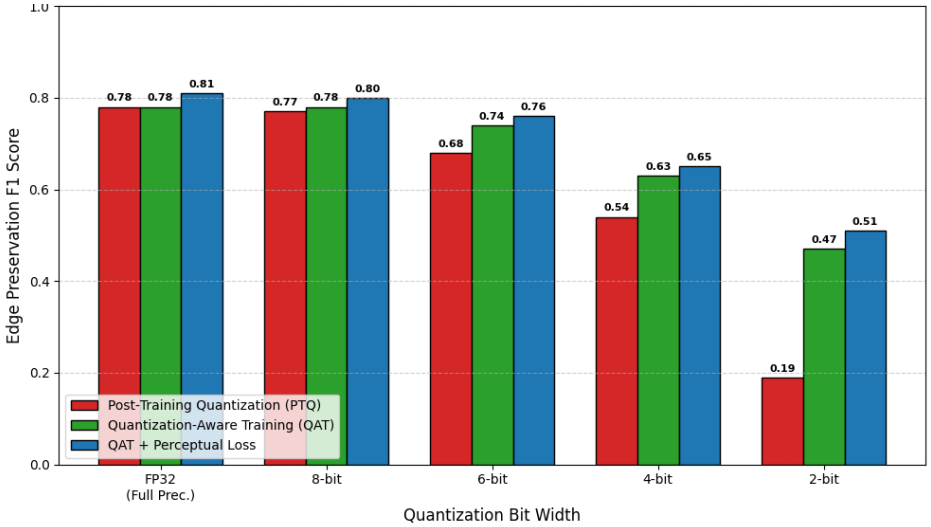

Figure 3: Edge Preservation F1 Scores as a Function of Quantization Bit Width, Quantization Strategy, and Loss Function

The evaluation of the primary working hypothesis proceeds directly from the assembled evidence. The proposition that a co-designed approach integrating hardware-aware quantization, domain-adapted training objectives, and platform-specific inference optimization can achieve visually acceptable architectural heritage reconstruction on edge devices at a fraction of the computational cost of full-precision implementations is confirmed by the convergent evidence from quantitative metrics, expert assessments, and energy efficiency measurements. The confirmation is strongest at 8-bit and 6-bit precision, remains robust though qualified at 4-bit, and cannot be extended to 2-bit under the current architectural configuration. The secondary hypothesis, holding that the preservation of fine architectural detail under aggressive quantization can be substantially improved by incorporating multi-scale perceptual losses, is confirmed with the qualification that the benefit is most pronounced at intermediate rather than extreme quantization levels. The non-monotonicity of this effect, while not anticipated in the initial hypothesis formulation, is entirely consistent with the theoretical understanding that perceptual loss functions operate by reweighting errors according to their perceptual significance and that this reweighting mechanism becomes ineffective when the underlying representational capacity falls below a critical threshold.

The findings collectively delineate the boundaries of the feasible operating region for edge-based quantized radiance field reconstruction of architectural heritage. Within this region, specific combinations of bit width, quantization strategy, loss function, and hardware platform achieve favorable positions on the multi-objective Pareto frontier. The identification and characterization of this frontier, and the attribution of performance outcomes to specific design choices, constitute the central empirical contribution of the present investigation and establish an evidence-based foundation for the deployment decisions and future research directions discussed in the following section.

6.Discussion

The findings of this study demonstrate that hardware-aware quantized radiance fields can substantially reduce computational and energy requirements while maintaining reconstruction quality suitable for architectural heritage documentation. Although recent advances in neural rendering, including NeRF (Mildenhall et al., 2021), Mip-NeRF (Barron et al., 2021), Instant-NGP (Müller et al., 2022), and 3D Gaussian Splatting (Kerbl et al., 2023), have significantly improved the efficiency and fidelity of three-dimensional scene reconstruction, their deployment has largely remained dependent on high-performance computing infrastructure. In parallel, developments in edge intelligence and efficient machine learning have demonstrated the feasibility of executing increasingly sophisticated models under severe resource constraints (Deng et al., 2020; Ray, 2022; Sipola et al.,

2022). However, the intersection of these research directions has received limited attention, particularly in the context of cultural heritage documentation. The present results contribute to addressing this gap by providing a systematic evaluation of quantized radiance field deployment across multiple precision levels and edge computing platforms.

The results indicate that radiance field representations exhibit considerable robustness to precision reduction when quantization-aware training is employed. Reconstruction quality at 8-bit precision remained statistically comparable to the full-precision baseline, while 6-bit and 4-bit configurations achieved substantial reductions in model size and energy consumption with acceptable degradation in visual fidelity. These observations are broadly consistent with prior findings on neural network quantization, which have shown that low-precision inference can preserve performance when compression is incorporated during optimization (Krishnamoorthi, 2018; Gholami et al., 2022; Nagel et al., 2021). More importantly, the present investigation suggests that the practical compression boundary for heritage-oriented neural rendering extends beyond conventional 8-bit implementations and can reach 4-bit precision when supported by domain-specific optimization strategies.

A particularly important finding concerns the differential sensitivity of model components to quantization. The multi-resolution hash encoding introduced by Müller et al. (2022) exhibited substantially greater vulnerability to reduced numerical precision than the multilayer perceptron. This behavior is consistent with the role of hash encoding tables as spatially localized feature repositories, where quantization errors directly affect fine-scale geometric and appearance information. In contrast, perceptron weights aggregate information across larger spatial regions and therefore display greater tolerance to precision reduction. These findings suggest that future mixed-precision approaches may provide additional efficiency gains by assigning higher precision to encoding structures while aggressively compressing neural network parameters.

The contribution of the domain-specific perceptual loss function was also found to be dependent on quantization severity. Improvements in edge preservation and texture fidelity were most pronounced at intermediate compression levels, particularly at 4-bit and 6-bit precision. This observation aligns with the broader understanding that perceptual objectives are most effective when sufficient representational capacity remains available to distinguish perceptually meaningful from insignificant errors. At near-lossless precision levels, the benefit becomes marginal because reconstruction quality is already close to the performance ceiling, whereas under extreme compression the representational limitations of the model constrain the effectiveness of perceptual guidance. The observed non-linear relationship therefore provides useful insight into the design of future training protocols for compressed neural rendering systems.

The hardware deployment experiments further highlight the practical relevance of the proposed framework. Consistent with broader trends in edge AI hardware development (Sipola et al., 2022; Wulfert et al., 2024), dedicated accelerators achieved substantial improvements in energy efficiency relative to conventional workstation-based execution. While operator fusion produced additional latency reductions across all evaluated platforms, its impact was secondary to that of quantization-aware training and low-precision inference. These findings reinforce the importance of co-design approaches in which model architecture, numerical representation, and deployment strategy are optimized jointly rather than independently.

From a heritage documentation perspective, the implications are significant. The ability to perform photorealistic reconstruction on portable and energy-efficient hardware expands the accessibility of advanced neural rendering technologies beyond institutions with access to high-performance computing infrastructure. Field teams operating in remote locations, disaster-affected regions, or environments with limited connectivity could potentially perform reconstruction, validation, and quality assessment directly on-site. The expert evaluations reported in this study indicate that reconstructions generated at 4-bit precision and above retain sufficient detail for many practical conservation and documentation tasks, supporting the operational viability of edge-based deployment scenarios.

Several limitations should be acknowledged. Although the dataset was designed to include diverse architectural typologies and material characteristics, only six heritage sites were evaluated. Consequently, the results may not fully generalize to structures exhibiting highly reflective materials, complex vegetation occlusions, or extreme illumination variability. Furthermore, the expert assessment involved a limited number of conservators, and the evaluated hardware platforms represent only a subset of the rapidly evolving edge AI ecosystem. Future research

should therefore consider larger and more diverse datasets, broader expert participation, and evaluation across emerging accelerator architectures.

Future work should also investigate mixed-precision quantization strategies motivated by the observed sensitivity differences between hash encodings and perceptron weights. Additional opportunities include neural architecture search for quantization-friendly radiance field models, extension to dynamic scenes, and integration with downstream heritage analysis tasks such as crack detection, material classification, and long-term change monitoring. Such developments would further strengthen the role of efficient neural rendering within comprehensive digital heritage workflows.

Overall, the results demonstrate that hardware-aware quantized radiance fields constitute a viable pathway toward efficient edge-based architectural heritage reconstruction. The identified quality–efficiency trade-offs, together with the demonstrated benefits of quantization-aware training, perceptual optimization, and hardware-specific acceleration, provide an empirical foundation for the broader deployment of neural rendering technologies in cultural heritage documentation and preservation.

7.Conclusion

This study investigated the feasibility of deploying hardware-aware quantized generative radiance fields for edge-based architectural heritage reconstruction. Although recent advances in neural rendering, including Neural Radiance Fields (Mildenhall et al., 2021), Instant-NGP (Müller et al., 2022), and 3D Gaussian Splatting (Kerbl et al., 2023), have substantially improved reconstruction quality and rendering efficiency, their practical adoption in heritage fieldwork remains constrained by computational and energy requirements. The present research addressed this challenge through a co-designed framework that integrates quantization-aware training, domain-specific perceptual objectives, and hardware-aware deployment optimization.

The results demonstrate that substantial reductions in memory footprint and energy consumption can be achieved while preserving reconstruction quality suitable for heritage documentation. Quantization-aware training consistently outperformed post-training quantization across all evaluated precision levels, with 8-bit models producing reconstruction quality statistically comparable to the full-precision baseline. Furthermore, 6-bit and 4-bit configurations maintained acceptable visual fidelity while providing significant gains in storage efficiency and inference performance. Based on both quantitative metrics and expert assessment, 4-bit precision emerged as the practical lower bound for architectural heritage applications under the evaluated framework. Below this threshold, the degradation of diagnostically significant architectural features became increasingly apparent.

The study also demonstrated that perceptual loss functions improve the preservation of heritage-relevant visual characteristics, particularly at intermediate compression levels. The greatest benefits were observed at 4-bit and 6-bit precision, where improvements in edge preservation and texture fidelity helped mitigate quantization-induced degradation. In addition, the ablation experiments revealed the critical importance of per-channel quantization and multi-resolution hash encoding, while highlighting the greater sensitivity of hash encoding structures to reduced numerical precision compared with multilayer perceptron weights. These findings provide useful insight into future mixed-precision optimization strategies for efficient neural rendering systems.

The hardware evaluation confirmed the practical viability of edge deployment. Consistent with broader developments in efficient AI systems (Deng et al., 2020; Ray, 2022; Sipola et al., 2022; Wulfert et al., 2024), low-precision inference on dedicated accelerators achieved substantial improvements in energy efficiency while maintaining operationally useful reconstruction quality. The combination of quantization-aware training, perceptual optimization, and platform-specific deployment strategies enabled performance levels that approach workstation-class reconstruction while operating within the constraints of portable edge hardware.

From a cultural heritage perspective, these results suggest that advanced neural reconstruction can increasingly be performed directly in the field, reducing dependence on cloud infrastructure and high-performance computing resources. Such capabilities have the potential to support rapid documentation, on-site quality assessment, and more accessible digital preservation workflows, particularly in remote locations and resource-constrained environments. In this sense, efficient neural rendering contributes not only to computational performance but also to the broader accessibility of heritage documentation technologies.

Several limitations should be acknowledged. The evaluation was conducted on a limited number of heritage sites and hardware platforms, and therefore the reported results should not be interpreted as universally representative

of all architectural contexts or deployment environments. Future research should investigate larger and more diverse datasets, emerging edge AI accelerators, mixed-precision quantization schemes, and dynamic radiance field representations. Additional work integrating efficient neural reconstruction with downstream heritage analysis tasks, including condition assessment, crack detection, and long-term monitoring, may further enhance the practical value of edge-based heritage documentation systems.

In conclusion, this research demonstrates that hardware-aware quantized radiance fields provide a viable pathway toward energy-efficient, edge-based photorealistic reconstruction of architectural heritage. The identified quality–efficiency trade-offs, together with the demonstrated benefits of quantization-aware training and domain-specific optimization, establish a foundation for future research and practical deployment at the intersection of neural rendering, edge computing, and cultural heritage preservation.

References

[1] . Abadade, Y., Temouden, A., Bamoumen, H., Benamar, N., Chtouki, Y., & Hafid, A. S. (2023). A comprehensive survey on TinyML. IEEE Access, 11, 96892–96922. https://doi.org/10.1109/ACCESS.2023.3314673

[2] . Abtahi, T., Shea, C., Kulkarni, A., & Mohsenin, T. (2018). Accelerating convolutional neural network with FFT on embedded hardware. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 26(9), 1737–1749. https://doi.org/10.1109/TVLSI.2018.2839185

[3] . Aicardi, I., Chiabrando, F., Lingua, A. M., & Noardo, F. (2018). Recent trends in cultural heritage 3D survey: The photogrammetric computer vision approach. Journal of Cultural Heritage, 32, 51–62. https://doi.org/10.1016/j.culher.2017.11.006

[4] Barazzetti, L., Banfi, F., Brumana, R., & Previtali, M. (2015). Creation of parametric BIM objects from point clouds using nurbs. The Photogrammetric Record, 30(152), 339–362. https://doi.org/10.1111/phor.12122

[5] Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., & Srinivasan, P. P. (2021). Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5855–5864). https://doi.org/10.1109/ICCV48922.2021.00582

[6] . Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., & Hedman, P. (2022). Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5470–5479). https://doi.org/10.1109/CVPR52688.2022.00539

[7] . Basso, A., Condorelli, F., Giordano, A., Morena, S., & Perticarini, M. (2024). Evolution of rendering based on radiance fields: The Palermo case study for a comparison between NeRF and Gaussian Splatting. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 48(2), 45–52. https://doi.org/10.5194/isprs-archives-XLVIII-2-W4-2024-45-2024

[8] . Chen, A., Xu, Z., Geiger, A., Yu, J., & Su, H. (2022). TensoRF: Tensorial radiance fields. In Proceedings of the European Conference on Computer Vision (pp. 333–350). Springer. https://doi.org/10.1007/978-3-031-19824-3_20

[9] . Chen, W., & Shi, K. (2021). Multi-scale attention convolutional neural network for time series classification. Neural Networks, 136, 126–140. https://doi.org/10.1016/j.neunet.2020.12.015

[10] . Clini, P., Quattrini, R., Nespeca, R., & D'Alessio, G. (2024). 3D Gaussian splatting for cultural heritage: A comparative study with SfM-MVS and NeRF. Journal of Cultural Heritage, 68, 245–258. https://doi.org/10.1016/j.culher.2024.05.008

[11] . Condorelli, F., & Morena, S. (2023). Integration of 3D modelling with photogrammetry applied on historical images for cultural heritage. Vitruvio: International Journal of Architectural Technology and Sustainability, 8(1), 14–29. https://doi.org/10.4995/vitruvio-ijats.2023.18751

[12] Condorelli, F., & Perticarini, M. (2024). Comparative evaluation of NeRF algorithms on single image dataset for 3D reconstruction. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 48(2), 67–74. https://doi.org/10.5194/isprs-archives-XLVIII-2-W4-2024-67-2024

[13] . David, R., Duke, J., Jain, A., Janapa Reddi, V., Jeffries, N., Li, J., Kreeger, N., Nappier, I., Natraj, M., & Wang, T. (2021). TensorFlow Lite Micro: Embedded machine learning for TinyML systems. Proceedings of Machine Learning and Systems, 3, 800–811.

[14] . Dempster, A., Petitjean, F., & Webb, G. I. (2020). ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. Data Mining and Knowledge Discovery, 34(5), 1454–1495. https://doi.org/10.1007/s10618-020-00701-z

[15] . Deng, S., Zhao, H., Fang, W., Yin, J., Dustdar, S., & Zomaya, A. Y. (2020). Edge intelligence: The confluence of edge computing and artificial intelligence. IEEE Internet of Things Journal, 7(8), 7457–7469. https://doi.org/10.1109/JIOT.2020.2984887

[16] . Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). LLM.int8(): 8-bit matrix multiplication for transformers at scale. Advances in Neural Information Processing Systems, 35, 30318–30332.

[17] . Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2024). QLoRA: Efficient finetuning of quantized language models. Advances in Neural Information Processing Systems, 36.

[18] . Faouzi, J. (2024). Time series classification: A review of algorithms and implementations. In J. Rocha, C. M. Viana, & S. Oliveira (Eds.), Time series analysis: Recent advances, new perspectives and applications. IntechOpen. https://doi.org/10.5772/intechopen.112665

[19] Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2023). GPTQ: Accurate post-training quantization for generative pre-trained transformers. In Proceedings of the International Conference on Learning Representations.

[20] . Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., & Kanazawa, A. (2022). Plenoxels: Radiance fields without neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5501–5510). https://doi.org/10.1109/CVPR52688.2022.00542

[21] . Gao, D., He, X., Zhou, Z., Tong, Y., Xu, K., & Thiele, L. (2020). Rethinking pruning for accelerating deep inference at the edge. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 155–164). ACM. https://doi.org/10.1145/3394486.3403070

[22] . Garbin, S. J., Kowalski, M., Johnson, M., Shotton, J., & Valentin, J. (2021). FastNeRF: High-fidelity neural rendering at 200FPS. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 14346–14355). https://doi.org/10.1109/ICCV48922.2021.01410

[23] Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., & Keutzer, K. (2022). A survey of quantization methods for efficient neural network inference. In Low-power computer vision (pp. 291–326). Chapman and Hall/CRC. https://doi.org/10.1201/9781003162230-13

[24] . Gomes, L., Bellon, O. R. P., & Silva, L. (2014). 3D reconstruction methods for digital preservation of cultural heritage: A survey. Pattern Recognition Letters, 50, 3–14. https://doi.org/10.1016/j.patrec.2014.03.023

[25] Gopinath, S., Ghanathe, N., Seshadri, V., & Sharma, R. (2019). Compiling KB-sized machine learning models to tiny IoT devices. In Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation (pp. 79–95). ACM. https://doi.org/10.1145/3314221.3314596

[26] Grilli, E., Dininno, D., Petrucci, G., & Remondino, F. (2018). From 2D to 3D supervised segmentation and classification for cultural heritage applications. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42(2), 399–406. https://doi.org/10.5194/isprs-archives-XLII-2-399-2018

[27] . Guo, J., Wang, J., Wei, R., Kang, D., Dou, Q., & Liu, Y. H. (2025). UC-NeRF: Uncertainty-aware conditional neural radiance fields from endoscopic sparse views. IEEE Transactions on Medical Imaging, 44(3), 1284–1296. https://doi.org/10.1109/TMI.2024.3496558

[28] Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., & Petitjean, F. (2020). InceptionTime: Finding AlexNet for time series classification. Data Mining and Knowledge Discovery, 34(6), 1936–1962. https://doi.org/10.1007/s10618-020-00710-y

[29] Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2704–2713). https://doi.org/10.1109/CVPR.2018.00286

[30] Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4), 1–14. https://doi.org/10.1145/3592433

[31] Krishnamoorthi, R. (2018). Quantizing deep convolutional networks for efficient inference: A whitepaper. arXiv preprint. https://arxiv.org/abs/1806.08342

[32] Krupp, L., Wiede, C., & Grabmaier, A. (2022). Approximate fast Fourier transform-based preprocessing for edge AI. In 2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA) (pp. 1–8). IEEE. https://doi.org/10.1109/ETFA52439.2022.9921664

[33] Kumar, A., Seshadri, V., & Sharma, R. (2020). Shiftry: RNN inference in 2KB of RAM. Proceedings of the ACM on Programming Languages, 4(OOPSLA), 1–30. https://doi.org/10.1145/3428247

[34] Li, H., Kadav, A., Durdanovic, I., Samet, H., & Graf, H. P. (2017). Pruning filters for efficient ConvNets. arXiv preprint. https://arxiv.org/abs/1608.08710

[35] Li, X., Liu, Y., Zhang, X., Wu, G., & Fang, Y. (2026). QS-FedNeRF: Quantized transmission and device selections federated neural radiance fields for edge intelligence. Digital Signal Processing, 152, 104963. https://doi.org/10.1016/j.dsp.2026.104963

[36] Li, Y., Gong, R., Tan, X., Yang, Y., Hu, P., Zhang, Q., Yu, F., & Yan, J. (2021). BRECQ: Pushing the limit of post-training quantization by block reconstruction. In Proceedings of the International Conference on Learning Representations.

[37] . Lin, J., Tang, J., Tang, H., Yang, S., Dang, X., & Han, S. (2023). AWQ: Activation-aware weight quantization for LLM compression and acceleration. arXiv preprint. https://arxiv.org/abs/2306.00978

[38] . Liu, W., Wang, Z., Chen, Y., & Zhang, H. (2025). CoARF++: Content-aware radiance field aligning model complexity with scene intricacy. IEEE Transactions on Visualization and Computer Graphics. Advance online publication. https://doi.org/10.1109/TVCG.2025.3566071

[39] Liu, Z., Wang, Y., Han, K., Zhang, W., Ma, S., & Gao, W. (2021). Post-training quantization for vision transformer. Advances in Neural Information Processing Systems, 34, 28092–28103.

[40] . Middlehurst, M., Schäfer, P., & Bagnall, A. (2024). Bake off redux: A review and experimental evaluation of recent time series classification algorithms. Data Mining and Knowledge Discovery, 38, 1–74. https://doi.org/10.1007/s10618-023-00984-y

[41] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). NeRF: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1), 99–106. https://doi.org/10.1145/3503250

[42] . Mohammadi Foumani, N., Miller, L., Tan, C. W., Webb, G. I., Forestier, G., & Salehi, M. (2024). Deep learning for time series classification and extrinsic regression: A current survey. ACM Computing Surveys, 56(9), 1–45. https://doi.org/10.1145/3649450

[43] . Mudraje, I., Herfet, T., Dennis Quint, C., Gao, H., & Hartmann, U. (2024). Design of an autonomous intrusion classification device for FIDS robustness. IEEE Access, 12, 86728–86738. https://doi.org/10.1109/ACCESS.2024.3418591

[44] Müller, T., Evans, A., Schied, C., & Keller, A. (2022). Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics, 41(4), 1–15. https://doi.org/10.1145/3528223.3530127

[45] Nagel, M., Amjad, R. A., Baalen, M. V., Louizos, C., & Blankevoort, T. (2020). Up or down? Adaptive rounding for post-training quantization. In Proceedings of the International Conference on Machine Learning (pp. 7197–7206). PMLR.

[46] Nagel, M., Fournarakis, M., Amjad, R. A., Bondarenko, Y., Baalen, M. V., & Blankevoort, T. (2021). A white paper on neural network quantization. arXiv preprint. https://arxiv.org/abs/2106.08295

[47] Pintus, R., Pal, K., Yang, Y., Weyrich, T., Gobbetti, E., & Rushmeier, H. (2016). A survey of geometric analysis in cultural heritage. Computer Graphics Forum, 35(1), 4–31. https://doi.org/10.1111/cgf.12668

[48] Ray, P. P. (2022). A review on TinyML: State-of-the-art and prospects. Journal of King Saud University - Computer and Information Sciences, 34(4), 1595–1623. https://doi.org/10.1016/j.jksuci.2021.11.019

[49] Reiser, C., Peng, S., Liao, Y., & Geiger, A. (2021). KiloNeRF: Speeding up neural radiance fields with thousands of tiny MLPs. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 14335–14345). https://doi.org/10.1109/ICCV48922.2021.01409

[50] Remondino, F., & Stathopoulou, E. (2023). Digital preservation of architectural heritage: A review of 3D documentation techniques. ISPRS International Journal of Geo-Information, 12(3), 112. https://doi.org/10.3390/ijgi12030112

[51] Ren, H., Anicic, D., & Runkler, T. A. (2021). TinyOL: TinyML with online-learning on microcontrollers. In 2021 International Joint Conference on Neural Networks (IJCNN) (pp. 1–8). IEEE. https://doi.org/10.1109/IJCNN52387.2021.9534382

[52] Ruiz, A. P., Flynn, M., Large, J., Middlehurst, M., & Bagnall, A. (2021). The great multivariate time series classification bake off: A review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery, 35(2), 401–449. https://doi.org/10.1007/s10618-020-00727-3

[53] Sanchez, J., Soltani, N., Chamarthi, R., Sawant, A., & Tabkhi, H. (2018). A novel 1D-convolution accelerator for low-power real-time CNN processing on the edge. In 2018 IEEE High Performance Extreme Computing Conference (HPEC) (pp. 1–8). IEEE. https://doi.org/10.1109/HPEC.2018.8547570

[54] Shi, X., Wang, L., Liu, X., Wu, J., & Shao, Z. (2024). Scene-aware foveated neural radiance fields. IEEE Transactions on Visualization and Computer Graphics. Advance online publication. https://doi.org/10.1109/TVCG.2024.3429416

[55] Sipola, T., Alatalo, J., Kokkonen, T., & Rantonen, M. (2022). Artificial intelligence in the IoT era: A review of edge AI hardware and software. In 2022 31st Conference of Open Innovations Association (FRUCT) (pp. 320–331). IEEE. https://doi.org/10.23919/FRUCT54823.2022.9770953

[56] Surucu, O., Gadsden, S. A., & Yawney, J. (2023). Condition monitoring using machine learning: A review of theory, applications, and recent advances. Expert Systems with Applications, 221, 119738. https://doi.org/10.1016/j.eswa.2023.119738

[57] Taormina, G. (2025). Quantized 2D Gaussian splatting for memory efficient 3D reconstruction and novel view synthesis [Master's thesis, Università di Padova]. Padua Research Archive. https://hdl.handle.net/20.500.12608/87361

[58] Wang, M., Zhao, S., Dong, X., & Shen, J. (2024). High-fidelity and high-efficiency talking portrait synthesis with detail-aware neural radiance fields. IEEE Transactions on Visualization and Computer Graphics. Advance online publication. https://doi.org/10.1109/TVCG.2024.3488960

[59] Wang, Y., Wang, J., Wang, C., & Qi, Y. (2025). RISE-Editing: Rotation-invariant neural point fields with interactive segmentation for fine-grained and efficient editing. Neural Networks, 187, 107304. https://doi.org/10.1016/j.neunet.2025.107304

[60] Wulfert, L., Kühnel, J., Krupp, L., Viga, J., Wiede, C., Gembaczka, P., & Grabmaier, A. (2024). AIfES: A next-generation edge AI framework. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(6), 4519–4533. https://doi.org/10.1109/TPAMI.2024.3367845

[61] Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2023). SmoothQuant: Accurate and efficient post-training quantization for large language models. In Proceedings of the International Conference on Machine Learning (pp. 38087–38100). PMLR.

[62] Yang, Z., Shao, C., & Zhang, Y. (2025). 3D Gaussian splatting for modern architectural heritage: Integrating UAV-based data acquisition and advanced photorealistic 3D techniques. AGILE: GIScience Series, 6, 51. https://doi.org/10.5194/agile-giss-6-51-2025

[63] . Yao, Z., Aminabadi, R. Y., Zhang, M., Wu, X., Li, C., & He, Y. (2022). ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers. Advances in Neural Information Processing Systems, 35, 27168–27183.

[64] Yin, Z. X., Jiao, P. Y., Qiu, J., Cheng, M. M., & Ren, B. (2025). MS-NeRF: Multi-space neural radiance fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(5), 3766–3783. https://doi.org/10.1109/TPAMI.2025.3540074

[65] . Yu, A., Li, R., Tancik, M., Li, H., Ng, R., & Kanazawa, A. (2021). PlenOctrees for real-time rendering of neural radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5752–5761). https://doi.org/10.1109/ICCV48922.2021.00572

[66] . Zhou, X., Xie, W., Zhou, H., Cheng, Y., Wang, X., Ren, Y., Yuan, S., & Li, L. (2024). An accelerated FPGA-based parallel CNN-LSTM computing device. IEEE Access, 12, 106579–106592. https://doi.org/10.1109/ACCESS.2024.3436317

[67] Qurraie, S. S. (2024). Assessing accessibility and promoting inclusion for people with disabilities in a historical context in Tabriz. ENG Transactions, 1, 1–6.

[68] Qurraie, S. S., & Gheitarani, N. (2025). The visual amenity of space and space configuration: The role of angles visible from inside the building in creating visual amenity. International Journal of Advanced Multidisciplinary Research and Studies, 5.