Automation in Construction 165 (2024) 105517

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

3D reconstruction of building structures incorporating neural radiation fields and geometric constraints

Depeng Cui [a,b,c], Weidong Wang [a,b,c], Wenbo Hu [d,e], Jun Peng [a,b,c], Yida Zhao [a,b,c], Yukun Zhang [a,b,c], Jin Wang [a,b,c,*]

[a] School of Civil Engineering, Central South University, Changsha 410075, China
[b] MOE Key Laboratory of Engineering Structures of Heavy-haul Railway, Central South University, Changsha 410075, China
[c] Center for Railway Infrastructure Smart Monitoring and Management, Central South University, Changsha 410075, China
[d] National Rail Transit Electrification and Automation Engineering Technology Research Center (Hong Kong Branch), Hong Kong 999077, China
[e] Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China

ARTICLE INFO

Keywords:
Building structure
NeRF
3D reconstruction
Geometric constraints
Manhattan world model

ABSTRACT

Neural Radiance Fields (NeRF) techniques demonstrate potential for reconstructing complex architectural scenes in three dimensions. However, applying NeRF poses challenges, such as prolonged training processes and limited detailed characterization. To address these issues, this paper introduces NeRFusion by integrating Mipmapped Neural Radiance Fields (Mip-NeRF) and Instant Neural Graphics Primitives (iNGP). NeRFusion enhances training efficiency, rendering quality, and reduces jaggedness through multisampling, weight reduction, and loss function optimization. Geometric constraints—normal consistency, plane fitting, and vertical/horizontal constraints—improve 3D reconstruction accuracy. Using the Manhattan World Model (MWM), NeRFusion accurately extracts primary building components and dimensions. Experimental results show NeRFusion achieved a Peak Signal-to-Noise Ratio (PSNR) of 36.23 dB and completed training in 50 min. Structural Similarity Index Measure (SSIM) improved by 13.41% over Mip-NeRF and 50.00% over iNGP, with PSNR increasing by 10.45% and 25.10%. NeRFusion significantly reduces training time by a factor of 15 compared to the original NeRF.

1. Introduction

The field of three-dimensional (3D) scene reconstruction has gained prominence due to advancements in computer vision and graphics technology [1,2]. This process involves extracting detailed 3D geometric information from images to enable the generation of images or videos from multiple viewpoints. Applications span diverse domains including autonomous driving, robotic navigation, augmented reality, building evaluation [3–5]. Current methods for image-based 3D scene reconstruction are categorized into two main types: geometry-based and deep learning-based. Geometry-based methods rely on principles of multi-view geometry to determine camera poses and scene point clouds through feature extraction and matching. Although effective, they necessitate complex processing workflows, substantial computational resources, and face limitations in capturing detailed scene intricacies. Conversely, recent advancements in deep learning approaches [6–8] employ deep neural networks to efficiently reconstruct 3D scenes from images, capitalizing on image features and geometric relationships. These models, divided into explicit and implicit methods based on their 3D scene geometry representation during the reconstruction process [9,10], demonstrate exceptional expressive capabilities.

3D reconstruction techniques, particularly explicit-based methods, have significantly advanced, supporting applications from structural disease detection to construction process simulation [11–13]. These methods utilize instruments like laser scanning [14–16], structured light [17,18], and Time of Flight (TOF) [19,20] to acquire depth information, essential for precise 3D modeling in areas such as engineering surveying and medical imaging. In addition, Chen et al. [21] employed aerial imagery from drones for bridge reconstructions, improving inspections through data quality assessments and damage detection, thereby underscoring the technique's utility in elongated features identification like railings. Moreover, explicit methods have proven versatile in structural monitoring and urban planning, as evidenced by Poku et al. [22] and Pan et al. [23], who used image processing and point cloud data for detailed 3D representations, significantly enhancing structural reconstruction accuracy and efficiency.

* Corresponding author at: School of Civil Engineering, Central South University, Changsha 410075, China.
E-mail address: jerryw@csu.edu.cn (J. Wang).

https://doi.org/10.1016/j.autcon.2024.105517
Received 28 December 2023; Received in revised form 23 May 2024; Accepted 31 May 2024
Available online 18 June 2024

Despite their extensive applications, explicit 3D reconstruction methods face challenges with complex geometries or dynamic lighting conditions, resulting in inaccuracies [24,25]. These issues are compounded by the high costs and technical demands of data acquisition equipment, such as laser scanners and TOF cameras, limiting their broad adoption. Hu et al. [26] addressed some of these limitations by integrating deep learning with 3D reconstruction for cable-stayed bridges, showcasing the potential of structure-aware learning methods. However, the feasibility of explicit methods is still limited by their significant computational resource requirements and the difficulty in managing occluded or intricate details. This highlights the continuous need for innovative solutions that optimize detail, accuracy, and efficiency in 3D structural modeling, guiding future research towards improving algorithmic approaches and data integration techniques.

In deep learning-based 3D reconstruction, Neural Radiance Fields (NeRF) signify a substantial advancement in capturing complex scene geometries with high fidelity [27]. Unlike traditional explicit methods limited by geometric and lighting complexities, NeRF employs a Multi-Layer Perceptron (MLP) for inferring a scene's continuous volumetric representation from multi-view images and camera poses, thus obviating the need for explicit 3D depth information. This methodology exceeds prior 3D neural representations and view synthesis techniques [28–31] in generating photorealistic images, while also simplifying the reconstruction process by eliminating the dependence on costly and technically demanding data acquisition equipment. Furthermore, ongoing research aimed at enhancing NeRF's foundational algorithms [32,33] has solidified its status as an efficient tool for 3D scene reconstruction. NeRF's advantage over other deep learning methods lies in its capacity to synthesize scenes with unprecedented detail and realism from a sparse set of images. This efficiency stems from its innovative continuous volumetric scene representation, which allows for the capture of intricate details that models based on discrete representations or simpler learning paradigms cannot achieve. Furthermore, NeRF's flexibility in adapting to diverse lighting conditions and perspectives distinguishes it from techniques less capable of handling such variations. By leveraging image-based data without heavy reliance on depth sensors, NeRF sets a new standard for realism and efficiency in accessible and versatile 3D reconstruction applications.

In summary, although explicit methods have their merits, they significantly struggle with the complex geometries and dynamic lighting conditions of architectural structures. These challenges are further compounded by the high costs associated with data acquisition equipment and the complexity of technical requirements, which typically lead to a lack of precision and efficiency. The NeRF method suggests a potential solution, yet its application in architectural reconstruction is still in its nascent stages, highlighting an urgent research necessity to leverage NeRF's potential to address these issues. Currently, comprehensive research on utilizing NeRF for 3D reconstruction of architectural structures remains markedly limited. This paper addresses the aforementioned issues by presenting NeRFusion, a rapid 3D reconstruction framework for architectural structures utilizing NeRF. NeRFusion aims to integrate two enhanced NeRF models, Mip-NeRF (Mipmapped Neural Radiance Fields) and iNGP (Instant Neural Graphics Primitives), to facilitate modeling of expansive building structures in large-scale scenes. The Mip-NeRF model enhances the representation of multiple scales by considering the color and density of conical cross-sections along rays, enabling high-fidelity reconstructions and notably improved antialiasing capabilities. The iNGP model employs a pyramid-structured grid for mapping coordinates to colors and densities, potentially expediting NeRF's training process but presenting challenges in jaggedness. NeRFusion integrates Mip-NeRF's multi-scale approach with iNGP's efficient mesh representation, incorporating a novel sampling strategy, a down-weighting mechanism, and optimized loss function to address jaggedness, thus enhancing rendering quality and training speed. To incorporate the fused model into 3D reconstruction of building structures, NeRFusion is enhanced with geometric constraints, aligning closely with the real building's geometric characteristics through normal consistency, plane fitting, and vertical/horizontal constraints. Utilizing the Manhattan World Model (MWM), the paper extracts the primary structure of buildings in 3D scenes to determine the main outline's dimensions, presenting a comprehensive approach to rapid architectural reconstruction with NeRF.

This paper delineates the primary contributions and advancements of this research from three distinct perspectives:

1) Introduction of the NeRFusion model, aimed at accelerating the training process and enhancing rendering fidelity.
2) Incorporation of geometric constraints to augment the NeRFusion model's capacity for reconstructing 3D building structures, allowing for the inclusion of actual building geometric features and resulting in improved accuracy and quality of the reconstructed building.
3) Utilization of the MWM to identify and extract the primary structure of the building, facilitating the calculation of the main building outline's dimensions and thereby providing more precise data for structural analysis.

The manuscript is structured as follows: Section 1 provides a comprehensive contextual framework. Section 2 reviews relevant 3D reconstruction technologies. Section 3 details the NeRFusion model algorithm. Section 4 demonstrates the application of the NeRFusion model in a real-world 3D modeling context. Section 5 conducts a comparative analysis of the NeRFusion model with Mip-NeRF and iNGP in cluster-building scenarios. Section 6 summarizes the experimental results and discusses the limitations of the NeRFusion model. Section 7 concludes the paper, summarizing the key findings and proposing directions for future research.

2. Implicit 3D modeling and NeRF basics

This section examines implicit-based 3D reconstruction techniques, emphasizing NeRF's role in advancing the modeling of complex geometries from sparse data. Implicit methods inherently provide a significant advancement over traditional explicit approaches by enabling the direct inference of volumetric details from 2D images. These methods facilitate the generation of detailed and accurate 3D models, thereby enhancing the fidelity and efficiency of 3D scene reconstruction. Initially, the foundational aspects of implicit reconstruction are discussed, followed by an exploration of NeRF's specific contributions to the field.

2.1. Implicit 3D reconstruction methods

Implicit 3D reconstruction methods extract geometric and appearance features of 3D scenes from 2D images without dedicated depth sensors. These approaches rely on computer vision and graphics principles, analyzing extensive image data. The primary challenge involves determining objects' extent and spatial arrangement from images of various perspectives. Implicit 3D reconstruction includes techniques such as stereo vision, structure from motion, multi-view stereo vision, and deep learning approaches. These methods extract depth and geometric information from images captured at various angles. Notably, innovations such as NeRF, within deep learning, enhance 3D reconstruction's efficiency and accuracy by leveraging neural networks to incorporate light and color data. This section traces NeRF's evolution from its inception to its latest advancements.

The rise of NeRF is attributed to rapid advancements in neural rendering, which applies neural networks to establish implicit surface representations. This section introduces NeRF, delineating its approach to representing implicit scenes via neural networks. NeRF diverges from other 3D image generation methods like voxel, mesh, and point cloud representations by utilizing continuous spatial functions to define implicit surfaces. This approach efficiently encodes 3D structures, using

continuous functions to implicitly portray objects, as illustrated in Fig. 1. Various studies, such as those on Occupancy Networks (OPNs) [34], Implicit Modeling Network (IM-NET) [35], and Deep Signed Distance Functions (DeepSDF) [36], have explored similar approaches. For instance, OPNs use feature vectors and spatial points to predict binary occupancy, employing deep neural network classifiers to represent 3D surfaces as continuous decision boundaries. This OPNs introduces a new paradigm in 3D reconstruction and object representation, eliminating the need for explicit surface representation and streamlining object modeling.

DeepSDF, in contrast to the OPNs principle, employs a method that implicitly represents continuous 3D surfaces through direct regression of the signed distance function. It demonstrates the capability to represent complex shapes without discretization errors and with lower memory requirements, thereby establishing a valuable benchmark for future research in implicit scene representation through neural networks. Notably, these approaches necessitate supervision through prior analysis of the 3D object's structure and exhibit limitations in model generalization and scene applicability. Consequently, researchers [37,38] have explored the development of implicit function-based representations for directly training networks using images as supervised data. Furthermore, studies [39,40] have introduced differentiable rendering, enabling explicit representation of the geometric properties of scene space. This approach eliminates the need to acquire depth and shape information, allowing for direct network training with 2D images and camera poses. Mildenhall et al. [27] introduced NeRF, which combines neural networks with MLPs to encode implicit surfaces. This method utilizes neural fields for parameterization and integrates positional encoding to capture intricate details. By leveraging a substantial volume of scene images as supervisory data, NeRF generates photorealistic images from 2D representations of diverse scenes. This study demonstrates the capacity to produce and synthesize photorealistic images from 2D representations of various scenes, marking a significant advancement in the field.

2.2. NeRF principles

The NeRF model represents the 3D scene via a neural network approximation of the radiance field. This field characterizes the color and volume density at every point in the scene and from every viewing direction. The stationary scene is represented as a continuous five-dimensional vector function, as delineated below:

$$F_\theta(x,d)\rightarrow(c(r,g,b),\sigma) \tag{1}$$

Where $x=(x,y,z)$ denotes the spatial 3D coordinates within the scene; $d=(\theta,\phi)$ represents the viewing orientation of the input image, and $c(r,g,b)$ represents the color emitted from that location in the direction of d. σ represents the volume density, which can be likened to a differentiable opacity and signifies the total amount of radiation emitted by the light as it traverses (x,y,z). F_θ represents the implicit expression of this five-dimensional function using one or more MLP neural networks. By taking as input a sequence of images captured from different viewpoints, this function outputs the color c and volume density σ corresponding to the 3D spatial location.

NeRF employs the technique of body painting to acquire the color information of camera rays $r(t)=o+td$ based on their volume density and color. Subsequently, it generates a novel view by tracking the color $C(r)$ of camera rays associated with each pixel in the image. This process is facilitated through an integral representation:

$$C(r)=\int_{t_0}^{t_n} T(t)\cdot\sigma(r(t))\cdot c(r(t),d)dt \tag{2}$$

Where o represents the location of the optical center of the camera. The function $T(t)$ is used to describe the cumulative transmittance of light from point t_0 to point t:

$$T(t)=e^{-\int_{t_0}^{t}\sigma(o+ud)du} \tag{3}$$

The original NeRF implementation and most of the later methods used a hierarchical sampling approach to compute the value of the integral $C(r)$. This approach is used to compute the value of the integral $C(r)$ by dividing the light into N equal regions. By dividing the light into N equal regions, sampling uniformly from each region, and subsequently obtaining the value of the integral based on the samples that have been sampled using body plotting, Eq. (2) is expressed as:

$$C(r)=\sum_{i=1}^{N}(1-e^{\sigma_i\delta_i})\cdot c_i\cdot e^{-\sum_{j=1}^{i=1}\sigma_i\delta_j} \tag{4}$$

Where δ_i represents the distance between sample points i and $i+1$, which are selected from a specific region. Meanwhile, σ_i and c_i represents the predicted density and color, respectively, of sample point i along a particular ray. NeRF optimizes the parameters of the MLP for every pixel by employing photometric consistency constraints based on squared error. Additionally, it enhances the scene details through the utilization of positional coding. For a comprehensive understanding of the implementation and process, please refer to the original research study.

NeRF is trained through the integration of the aforementioned representations, utilizing multi-view images with known poses. This process adeptly translates visual data from these images into precise 3D object representations, resulting in high-quality 3D image generation. The introduction of the NeRF algorithm has spurred widespread adoption of the implicit modeling, garnering increasing attention for its ability to tackle complex rendering challenges. Although the initial NeRF model excels in producing detailed, high-quality images, its applicability is constrained in specific scenes or applications due to its focus on static environments. Consequently, targeted enhancements and optimizations are necessary to enhance NeRF's rendering effectiveness and adaptability across various scenes. This study conducts a comprehensive analysis of various NeRF-based improvement algorithms, emphasizing their directions and experimental outcomes. Results are summarized in Table 1, with "–" indicating missing data as reported in the literature. This table presents average PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index Measure) values for multiple test scenes. It is crucial to consider this table as a reference due to variations in GPU capabilities across different experiments reported in

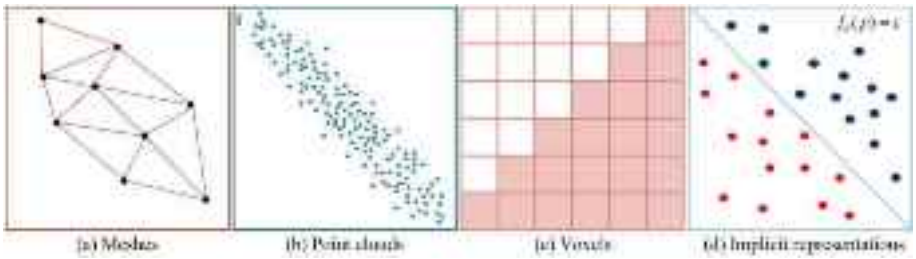


Fig. 1. Different 3D spatial expressions.

Table 1
Comparison of experimental results of NeRF and important improved methods.

Algorithms	Year of publication	Directions for improvement	Location coding method	Data set	Training time	PSNR	SSIM
NeRF [27]	2020	/	Fourier coding	NeRF dataset Synthetic	>12 h	31.01	0.94
Ref-NeRF [41]	2021	Image quality	Integrated position coding and integrated directional coding	NeRF dataset Synthetic	–	33.98	0.96
FastNeRF [42]	2021	Inference speed	Fourier coding	NeRF dataset Synthetic	>12 h	29.97	0.94
PixelNeRF [43]	2021	Sparse view	Fourier coding	ShapeNet	–	24.69	0.92
DietNeRF [44]	2021	Sparse view	Fourier coding	NeRF dataset Synthetic	–	23.59	0.86
MVSNeRF [45]	2021	Sparse view	Fourier coding	NeRF dataset Synthetic	~15 min	27.07	0.93
PlenOctrees [46]	2021	Reasoning speed and real-time rendering	Fourier coding	NeRF dataset Synthetic	>12 h	31.71	0.95
SNeRG [47]	2021	Reasoning speed and real-time rendering	Fourier coding	NeRF dataset Synthetic	>12 h	30.38	0.95
NerfingMVS [48]	2021	Add dense depth map guidelines guide the NeRF rendering process	Fourier coding	ScanNet	–	31.55	0.94
DS-NeRF [49]	2021	Sparse view	Fourier coding	ScanNet	~5 h	23.4	0.95
Mip-NeRF [32]	2021	Image quality and inference speed	Integrated Position Coding	NeRF dataset Synthetic	~3 h	33.09	0.96
RapNeRF [50]	2022	Image quality	Fourier coding	NeRF dataset Synthetic	>6 h	27.63	0.92
Point-NeRF [51]	2022	Add dense point cloud Enhancement	Frequency position coding	Synthetic	~40 min	33.31	0.97
iNGP [33]	2022	Inference speed	Hash encoding	NeRF dataset Synthetic	~5 min	30.05	–

the literature.

This paper provides a concise overview of recent advancements in NeRF, highlighting critical improvements in image quality, inference speed, spatial representation, and lighting modeling. Through a detailed examination of a diverse set of NeRF models, including iNGP, Point-NeRF, Ref-NeRF, Mip-NeRF, FastNeRF, PixelNeRF, DietNeRF, MVSNeRF, PlenOctrees, SNeRG, NerfingMVS, DS-NeRF, and RapNeRF, this study elucidates the unique contributions of each model to enhancing the efficiency and precision of 3D reconstructions, as delineated in Table 1. Models such as Mip-NeRF and Ref-NeRF are recognized for their advanced positional encoding methods that improve image fidelity, whereas iNGP's hash encoding technique significantly minimizes inference times. Furthermore, FastNeRF and PlenOctrees have played a crucial role in increasing rendering speeds, reflecting the diverse strategies employed to achieve real-time rendering. The adoption of sparse view synthesis by PixelNeRF and DietNeRF addresses the constraints of limited data, illustrating ongoing efforts to tackle the intricate challenges inherent in modeling large-scale indoor and outdoor environments. Collectively, these developments signify progress towards achieving more accurate, efficient, and comprehensive 3D architectural models.

Despite the proficiency of the models listed in Table 1 in terms of training speed and reconstruction quality, their application to large-scale indoor and outdoor architectural environments has unveiled certain limitations. These limitations arise from:

1) The detailed architectural elements challenge the current mesh representations, underscoring the need for advanced modeling techniques to capture intricate details accurately.
2) The diverse surface textures and material properties in architectural scenes surpass the current lighting representation methods, calling for more sophisticated lighting and texture modeling techniques.
3) The complex occlusion relationships in architectural environments pose significant challenges to the existing distance function representations, necessitating improvements in spatial modeling.
4) The presence of both indoor and outdoor scenes requires robust multi-scale modeling techniques to address the wide range of spatial scales.

In response to these challenges, this paper presents an integrated approach that synergizes the strengths of Mip-NeRF and iNGP. This approach aims to enhance both the efficiency of training and the accuracy of detail representation. The examination of Table 1 reveals:

1) Mip-NeRF and iNGP offer complementary techniques—conical cross-sections for anti-aliasing and grid-based multi-scale representation, respectively—that together enhance model performance.
2) Mip-NeRF's high-quality output, albeit with slower training times, complements iNGP's efficient but slightly jagged results. Their integration promises a balance of quality and efficiency.
3) The integration of Mip-NeRF's sampling framework with iNGP's grid representation, through tailored sampling strategies and loss functions, ensures compatibility and performance enhancement.
4) Both models are united by the goal of achieving superior rendering quality, making their integration a strategic and coherent choice.

This paper offers a concise yet comprehensive overview of the state-of-the-art in NeRF research, with a particular focus on Mip-NeRF and

iNGP as highlighted in Table 1. Mip-NeRF introduces a novel approach to address the persistent challenge of jagged artifacts in NeRF models. By segmenting rays into tapered cross-sectional regions and approximating these using Gaussian distributions, Mip-NeRF facilitates artifact-free multi-scale modeling. This method employs a two-stage sampling strategy that not only enhances training efficiency but also ensures output consistency across various resolutions.

In contrast, iNGP accelerates rendering speeds by leveraging a multi-resolution hash coding technique. This technique efficiently encodes scene details as indexes in a hash table, significantly reducing training time while improving visual quality, thereby achieving real-time neurographic rendering performance. Notably, iNGP's optimization and hash coding approach enable a remarkable rendering rate of 60 frames per second, facilitating the incorporation of various ray tracing techniques and visual effects such as antialiasing, motion blur, and depth of field.

Integrating Mip-NeRF with iNGP addresses NeRF research limitations, especially in complex architecture. Supported by Table 1, this approach combines Mip-NeRF's precise modeling and iNGP's rapid rendering, advancing 3D modeling.

3. Proposed 3D reconstruction model

This paper introduces the NeRFusion model, a novel approach that amalgamates the strengths of Mip-NeRF and iNGP, incorporating geometric constraints to enhance learning of the combined geometric features of buildings. The MWM [54] is used to extract the primary structure of buildings and define their dimensions in a 3D scene, as depicted in Fig. 2.

NeRFusion, an innovative technique, mitigates aliasing artifacts and enhances training efficiency through a grid-based strategy. It integrates the overarching framework of Mip-NeRF with iNGP's characterization method, harnessing the grid hierarchy of iNGP to generate feature vectors. Multi-sampling and downweighting methods within the Mip-NeRF framework compute pre-filtered iNGP features, allowing the grid features of iNGP to capture scale information for anti-aliasing, in addition to optimizing the loss function to reduce aliasing through pre-filtering of the NeRF histograms at the inline distillation stage, significantly improve the training and rendering processes of NeRF. To further enhance 3D building reconstruction, geometric constraints [52,53], including normal consistency, planar fitting, and vertical/horizontal constraints, are incorporated. Their integration ensures the model closely aligns with the geometric characteristics of real-world structures,

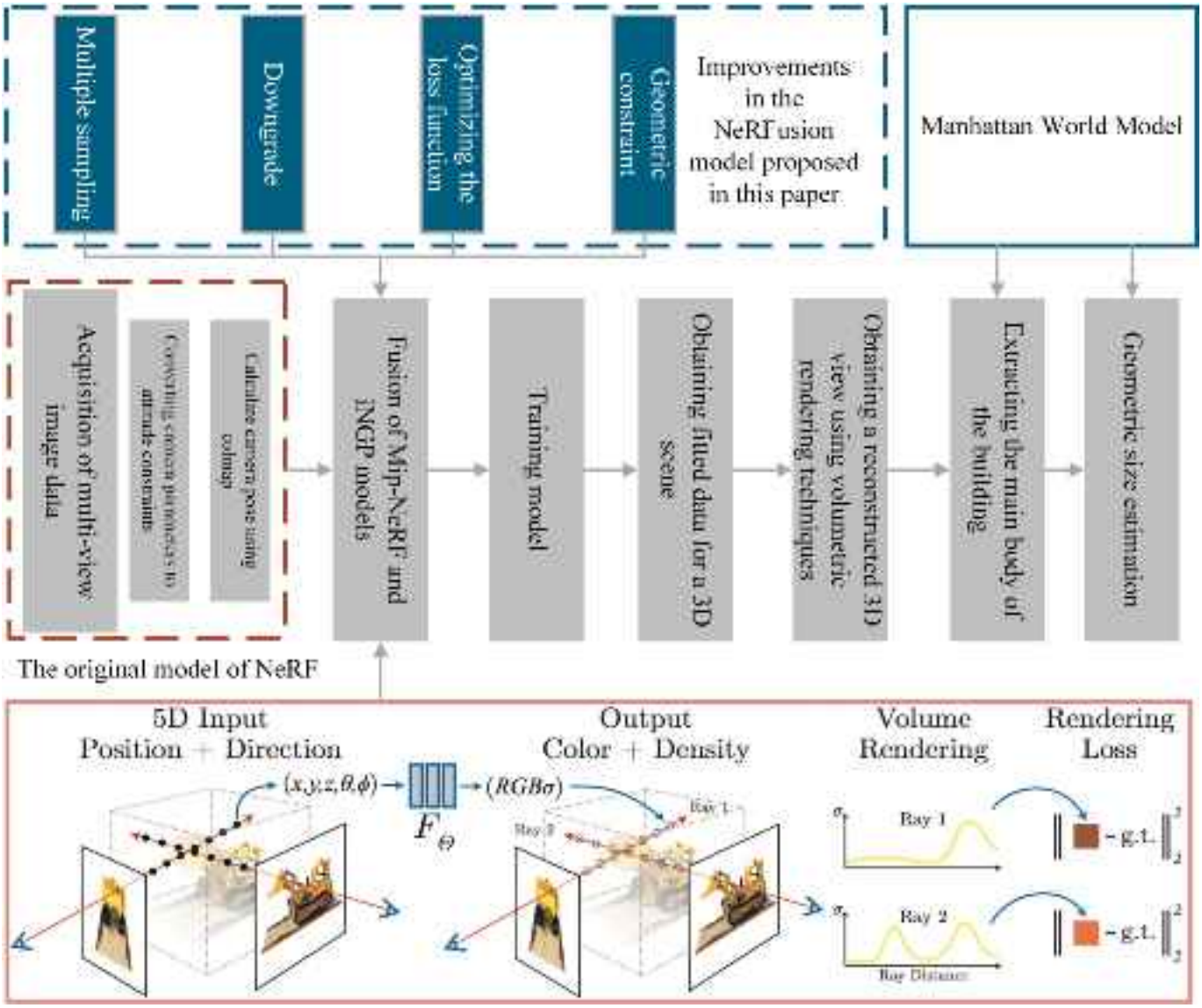


Fig. 2. Schematic diagram of NeRFusion modeling flow.

yielding surfaces that are smoother, more logical, and grounded in physical principles. The inclusion of these constraints filters out improbable outcomes, elevating the model's credibility and authenticity. Additionally, these constraints bolster the model's stability and structural integrity, preserving support relationships.

3.1. Multiple sampling

It is essential to acknowledge that both Mip-NeRF and iNGP, model variations closely aligned with NeRF, depend on 3D ray projection and featureization techniques for rendering. Mip-NeRF utilizes position coding and Gaussian approximation to represent rays, whereas iNGP employs a multi-level 3D mesh structure for feature vector generation. However, directly combining these methods introduces two distinct types of aliasing issues. A particular concern involves aliasing in spatial coordinates, evident from aliasing artifacts in the rendered outcomes due to iNGP's spatial coordinate aliasing. To mitigate this, the model integrates various sampling and weight reduction techniques, aiming to pre-filter iNGP features using multiple sub-Gaussian functions to effectively minimize aliasing. In essence, the core principle is as follows:

Following the Mip-NeRF methodology, it is postulated that every individual pixel can be associated with a cone of radius $\dot{r}t$, wherein t represents the distance traversed along the ray. $\dot{r}t$ collection of multiple samples is constructed to approximate the shape of the truncated cone within the interval along the ray $[t_0, t_1]$. The model utilizes a hexagonal pattern consisting of six points, with angle θ_j being a defining characteristic:

$$\theta = [0, 2\pi/3, 4\pi/3, 5\pi/3, \pi/3] \tag{5}$$

The angles in question represent the linear intervals of a complete revolution around a circle and are organized in such a way as to form a pair of triangles that are shifted by 60°. The measurement of the distance along the ray t_j (is)

$$t_j = t_0 + \frac{t_\delta\left(t_1^2 + 2t_\mu^2 + \frac{3}{\sqrt{7}}\left(\frac{2j}{5} - 1\right)\sqrt{\left(t_\delta^2 - t_\mu^2\right)^2 + 4t_\mu^4}\right)}{t_\delta^2 + 3t_\mu^2} \tag{6}$$

Where,

$$t_\mu = (t_0 + t_1)/2, t_\delta = (t_1 - t_0)/2 \tag{7}$$

The values in the set $[t_0, t_1]$ are arranged linearly, with a specific spacing between each value. These values are then adjusted by shifting and scaling them to concentrate the majority of the values near the base of the intercepted cone. This adjustment results in a higher density of points in the t-direction as it approaches the tip of the cone. The inclusion of points θ and t_j provides the coordinates that represent the multiple sampling points that express the cross-sectional shape of the final cone:

$$\left\{ \begin{bmatrix} \dot{r}t_j sin(\theta_j)/\sqrt{2} \\ \dot{r}t_j cos(\theta_j)/\sqrt{2} \\ \dot{r}t_j \end{bmatrix} \middle| j = 0, 1, \cdots, 5 \right\} \tag{8}$$

The 3D coordinates are multiplied by an orthogonal basis consisting of three vectors, with the third vector representing the ray's direction and the first two being arbitrary frames perpendicular to the ray. Subsequently, the coordinates are rotated to align with world coordinates and shifted relative to the ray's origin. Sample mean and variance, computed from multiple samples along and perpendicular to the ray direction, reflect the conic curve's characteristics. During training, patterns undergo random rotations and flips, whereas, in rendering, deterministic 30-degree rotations and flips are applied to each pattern, as visually depicted in Fig. 3.

The proposed model utilizes six multi-samples, denoted as $\{X_j\}$, which serve as the mean values for the isotropic Gaussians. Each

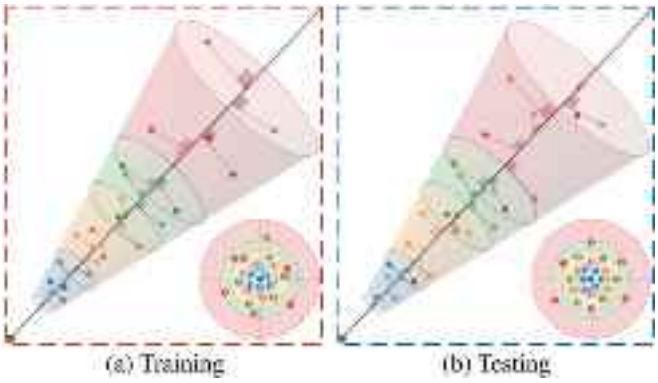


Fig. 3. Pattern rotation around ray and flip along ray: (a) randomly flipped during training, (b) fixed during rendering.

Gaussian distribution has a standard deviation of σ_j. The model establishes a relationship between variable σ_j and variable $\dot{r}t_j/\sqrt{2}$, and applies a scaling operation using a hyperparameter, which is consistently set to 0.5 in all experimental trials. The iNGP grid necessitates that the input coordinates be confined within a finite domain. Consequently, the contraction function is employed in the Mip-NeRF model.

3.2. Downweighting

The concept of multisampling enhances the sampling of continuous domain signals, aligning regional statistical characteristics and capturing low-frequency features to mitigate aliasing issues. Consequently, the iNGP grid representation layer adeptly represents spatial scales. However, challenges arise from high-frequency aliasing, a consequence of trilinear interpolation's inherent aliasing at each sampling point when relying solely on multisampling. Assigning weights to the interpolated values of grid features at each multisampling point acts as an additional prefiltering technique, reducing jagged artifacts in interpolation. This approach, referred to as down-weighting, serves as a secondary mechanism to improve the accuracy and reliability of interpolation. The synergistic combination of multisampling and down-weighting techniques enhances region information representation across various levels, facilitating smooth sampling and effective anti-aliasing within the grid structure.

This study elucidates the operational principle of down-weighting the interpolated outcomes of grid features for every multi-sampled point.

At a specified collection of multi-sampled points $\{X_j\}$, each individual point X_j is associated with a Gaussian distribution characterized by a mean and standard deviation σ_j. The collection of points X_j is associated with a Gaussian distribution characterized by a mean and standard deviation denoted as σ_j. In the iNGP, the grid size for layer l is denoted as n_l, while the range of values within this grid is represented by $[-0.5/n_l, 0.5/n_l]^3$. Calculate the degree of similarity between the Gaussian distribution of each X_j and the grid of layer l:

$$\omega_{j,l} = Erf\left(1/\sqrt{8}\cdot\sigma_j/n_l\right) \tag{9}$$

where Erf denotes the error function, which tends to 0 when $\sigma_j << n_l$, and $\omega_{j,l}$ tends to 1 when $\sigma_j >> n_l$.

The interpolated features for each X_j at the grid layer are subsequently acquired through the process of trilinear interpolation:

$$f_{j,l} = Trilerp(n_l \cdot X_j) \tag{10}$$

To obtain the weighted features, the original interpolated features are multiplied by the weights:

$$f_{j,l\_\omega} = \omega_{j,l} \cdot f_{j,l} \tag{11}$$

The fused features of the l-th layer of the grid are obtained by calculating the average of the weighted features across all sampled points:

$$f_l = mean\left(f_{j,l\_\omega}\right) \tag{12}$$

The aforementioned steps are iteratively performed for each grid layer, denoted as l, resulting in the aggregation of all f_l features in the channel dimension to yield the comprehensive grid features following weight reduction. Additionally, $\omega_{j,l}$ is encoded and concatenated to offer supplementary modeling of the sampling scale.

3.3. Loss function optimization

The integration of Mip-NeRF and iNGP presents a unique aliasing challenge, notably ray direction aliasing during training. This arises due to the acceleration effect of iNGP when combined with Mip-NeRF's online distillation method, leading to "z-aliasing" - instability and disappearance of scene content with changes in the camera's viewpoint. Such instability stems from the proposal network used in Mip-NeRF for online knowledge distillation and iNGP's optimization, causing irregularities in geometric distributions and skipping of layers at certain distances. To counteract this, researchers have developed an optimization solution that enhances the loss function by pre-filtering the NeRF histogram and applying a smoothing technique along each ray's direction during the calculation. This adjustment ensures the loss function's continuous smoothness, contrasting with the segmented nature observed in Mip-NeRF, and prevents the proposal network from solely learning the scene content. This method effectively reduces z-axis aliasing and the omission of specific distance layers, ensuring a more stable and accurate model representation. The key principle is elucidated as follows:

In the provided NeRF histogram (s, w), the variable s represents the sampled endpoints, while the variable w represents the weighted sum of each endpoint. The histogram of the proposed network output is represented as $(\hat{s}, \hat{w})$, with $\hat{s}$ representing the endpoints that have been sampled by the proposed network. The objective is to calculate the discrepancy between the NeRF and the histogram of the proposed network. However, due to the dissimilarity in their sampled endpoints, a direct comparison between w and $\hat{w}$ is not feasible. To address this issue, the present study suggests an initial approach of smoothing and resampling the variable w in the NeRF model, which will be referred to as w^s. The resampled weights will be denoted as $\hat{s}$. The function w is treated as a step-order function and convolved with a rectangular filter, which is a rectangular pulse with an integral value of 1. This convolution process results in the creation of a smooth segmented linear distribution, denoted as $P(s)$. The convolution operation is implemented by performing two integrations on the discretized Delta function, as illustrated in Fig. 4. By performing integration with respect to $P(s)$, the result is the cumulative distribution function $F(s)$ of $P(s)$. This function can be described as a segmented quadratic function. At the specific location denoted as point $\hat{s}$, the function $F(s)$ undergoes a process of sampling through quadratic interpolation, resulting in the value $F(\hat{s}_i)$. The index i represents the endpoint in this context. w^s is calculated by computing the difference between neighboring $F(\hat{s}_i)$ after resampling:

$$w_i^s = F(\hat{s}_i) - F(\hat{s}_{i-1}) \tag{13}$$

Eventually, the loss function can be calculated based on w^s and $\hat{w}$:

$$L(s, w, \hat{s}, \hat{w}) = \sum_i \frac{1}{\hat{w}_i} max\left(0, \nabla(w_i^{\hat{s}}) - \hat{w}_{i\partial}\right)^2 \tag{14}$$

Where the symbol ∇ represents the gradient operator, which is used to calculate the gradient of a function. For instance, when applied to a scalar function $y = f(x)$, the gradient is denoted as $\nabla y = d_y/d_x$. Similarly, for a multivariate function $y = f(x_1, x_2, \cdots x_n)$, its gradient is represented as $\nabla y = \left[d_y/d_{x_1}, d_y/d_{x_2}, \cdots, d_y/d_{x_n}\right]$. The aforementioned loss function exhibits continuous smoothness concerning variations in the viewing angle, thereby effectively mitigating the sawtooth problem.

To further elucidate the process highlighted in the preceding discussion, Fig. 4 presents an overview of the process necessary to achieve the smoothing and resampling of the NeRF histogram (s, w), aligning it with the proposed histogram $(\hat{s}, \hat{w})$ by utilizing the same set of endpoints. The NeRF histogram weight, denoted as w, is depicted as a stepwise function over the sampled endpoints, denoted as s, as illustrated in Fig. 4(1). The utilization of a rectangular filter for the convolution of w, as depicted in Fig. 4(2), results in the generation of a segmented linear function P(s) that exhibits a smooth characteristic. The cumulative distribution function F(s) of the P(s) is obtained through integration, as depicted in Fig. 4(3). It is worth noting that F(s) is represented by a segmented quadratic function. The process of sampling the function F(s) involves employing quadratic interpolation on the designated network sampling endpoints $\hat{s}$, as depicted in Fig. 4(4), resulting in the value $F(\hat{s}_i)$. In conclusion, the computation of the difference between adjacent $F(\hat{s}_i)$ values, as illustrated in Fig. 4(5), yields the resampled $w^{\hat{s}}$. This resampled $w^{\hat{s}}$ aligns with $\hat{w}$.

This paper presents the NeRFusion model, which integrates the strengths of Mip-NeRF and iNGP models for rapid and precise 3D architectural scene reconstruction. NeRFusion enhances anti-aliasing capabilities and training efficiency by merging these models. Employing multisampling and downweighting, the technique fuses spatial scale information with the iNGP mesh for exceptional anti-aliasing effects. Further, it optimizes the loss function to reduce aliasing through NeRF histogram pre-filtering during the online distillation phase. This innovative approach effectively amalgamates the features of both models, significantly boosting training velocity and generating high-quality anti-aliased imagery.

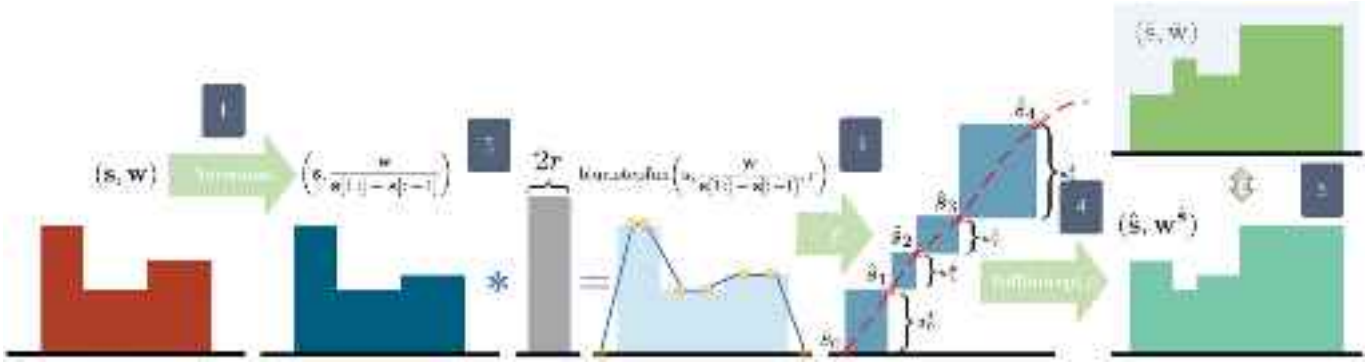


Fig. 4. Schematic diagram of anti-aliasing loss calculation process.

3.4. Geometric constraints

Despite the NeRFusion model's ability to create highly realistic scenes, challenges arise with specific structures, which adhere to strict geometric principles and constraints characterized by right angles and symmetry. Balancing image quality with geometric accuracy becomes complex for NeRF models, especially when precision in scene geometry representation is crucial. These models sometimes exhibit overfitting, compromising precision and geometric detail, which is particularly detrimental in applications like building structure modeling. This study introduces geometric constraints into the NeRFusion model to maintain geometric integrity and ensure adherence to established principles, combating overfitting and enhancing generalization capabilities for new and unseen data.

The NeRFusion model links 3D spatial points with their corresponding color and density attributes, generating realistic images. It incorporates geometric constraints through three supplementary branches responsible for predicting normals, plane parameters, and enforcing vertical/horizontal constraints, instrumental in scene generation. A comprehensive loss function, including reconstruction and geometric constraint losses, is tailored to specific task requirements. During training, the model minimizes this combined loss function, producing accurate images and a 3D model that adheres to normality, planarity, and alignment constraints. The normal consistency constraint ensures surface normals align with the true geometry, quantified by the angle between the generated and true normals, defining the loss measure:

$$L_{NC} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - sin\left(arccos\left(\frac{\nabla I_i \cdot \nabla I_{true}}{\|\nabla I_i\|^* \|\nabla I_{true}\|}\right)\right)\right) \tag{15}$$

Where N denote the quantity of points present in the scene. I_i represents the image intensity of the i-th point within the generated scene, while ∇I_i representing the gradient of the image intensity of the i-th point within the generated scene. I_{true} signifies the image intensity of the corresponding point in the real geometry, and ∇I_{true} denotes the gradient of the image intensity of the corresponding point in the real geometry.

The optimization process of the model involves minimizing the normals L_{NC} to ensure that the generated normals exhibit a high level of consistency with the true normals. The objective of the plane fitting constraint is to ensure that the surface generated by the model closely aligns with the plane. This task can be achieved by predicting the plane parameters and subsequently comparing them to the actual plane parameters.

$$L_{PP} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - sin\left(arccos\left(\frac{P_{gen,i} \cdot P_{real,i}}{\|P_{gen,i}\|^* \|P_{real,i}\|}\right)\right)\right) \tag{16}$$

Let $P_{gen,i}$ represent the predicted plane parameter for the i-th point in the generated scene, and $P_{real,i}$ represent the real plane parameter for the corresponding point in the real geometry. Additionally, let $\|P_{gen,i}\|$ and $\|P_{real,i}\|$ represent the number of paradigms of the plane parameters in the generated scene and the real geometry, respectively. The model is subsequently optimized through the minimization of the plane L_{PF}, aiming to maximize the consistency between the generated surface and the actual plane. The purpose of the vertical/horizontal constraints is to guarantee that the produced line segments conform to the intended vertical or horizontal alignment. The extent of the loss can be determined by computing the angular difference between the orientation of the produced line segment and the intended direction:

$$L_{VH} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - sin\left(arccos\left(\frac{|v_i \cdot u|}{\|v_i\|^* \|u\|}\right)\right)\right) \tag{17}$$

In the given context, N represents the total count of line segments present in the scene. v_i represents the direction vector of the i-th line segment in the generated scene, while u represents the desired direction vector. Additionally, $\|v_i\|$ and $\|u\|$ represents the paradigms of vectors v_i and u, respectively.

Hence, the integrated loss function incorporates the aforementioned three geometric constraints while taking into account the reconstruction loss, which can be mathematically represented as:

$$L_{All} = L(s, w, \widehat{s}, \widehat{w}) + \lambda_1 \times L_{NC} + \lambda_2 \times L_{PF} + \lambda_3 \times L_{VH} \tag{18}$$

The weight λ is essential in balancing each loss component. The inclusion of these three geometric constraints within the NeRFusion model significantly enhances the fidelity of the resulting 3D building structure model. This enhancement ensures better alignment with real-world characteristics, encompassing normals, plane shapes, and vertical/horizontal constraints. By integrating geometric considerations into the generation process, the model's geometric integrity is bolstered. This, in turn, leads to a more precise representation of the building's structural features in the generated image, amplifying the model's fidelity to the actual building.

3.5. Manhattan world model

This study employs the MWM to analyze point cloud data from the NeRFusion model, aiming to distill the core architectural structure of the reconstructed 3D scene. Predicated on the assumption that urban architectural elements primarily conform to vertical walls and horizontal planes within three orthogonal directions, the MWM efficiently processes 3D scenes. It methodically maps each point to the nearest plane or wall, thereby segmenting the primary structure. The effectiveness of the MWM hinges on a specific distance function, which is detailed further in the study. This approach aligns with the characteristic MWM assumption, enabling precise segmentation and analysis of complex urban structures. The MWM relies on the following distance function:

$$D(x, y, z) = \sum_{i=1}^{N} min(d_i(x, y, z), t_i) \tag{19}$$

where $D(x, y, z)$ signifies the minimum distance from a point to the nearest plane or wall within the MWM, N is the total count of such planes and walls, $d_i(x, y, z)$ calculates the distance to the i-th plane or wall, and t_i represents a threshold facilitating point classification.

The implementation of the MWM is executed in two primary stages, demonstrated in Fig. 5, utilizing the specified distance function. Initially, the input point cloud undergoes a classification process to discern local surface geometries, intended to remove irrelevant data points. This step enhances data quality by reducing noise and compensating for missing points. Points are classified into categories such as walls, edges, corners, or edge corners, determined by the proximity of points from surfaces with different orientations. The next stage focuses on characterizing the building volume through the integration of classified points and volume filling, ensuring a coherent segmentation of urban structures. This methodology highlights the model's component interconnectivity, with Fig. 5 providing a guide to the MWM's computational procedures, including initial classification and volume characterization. For further detail on the MWM and its role in 3D urban modeling, the seminal work by Vanegas et al. [54] is recommended.

Fig. 6 offers a systematic overview of the MWM applied to architectural reconstruction. The figure's left column showcases the input 3D model, highlighting the initial classification of the structure into walls, edges, and corners. Central to the depiction is the process of clustering these classified points, performing their triangulation, and creating adaptive bounding boxes to mirror the architectural spatial complexity accurately. The right column reveals the reconstruction phase, where volume rays extend to fully extrapolate the structure, ensuring a comprehensive model that accounts for occluded areas. This progression culminates in a texture-mapped reconstruction, affirming the model's geometric fidelity to the actual building, thereby illustrating the effective application of MWM in capturing and reconstructing architectural nuances.

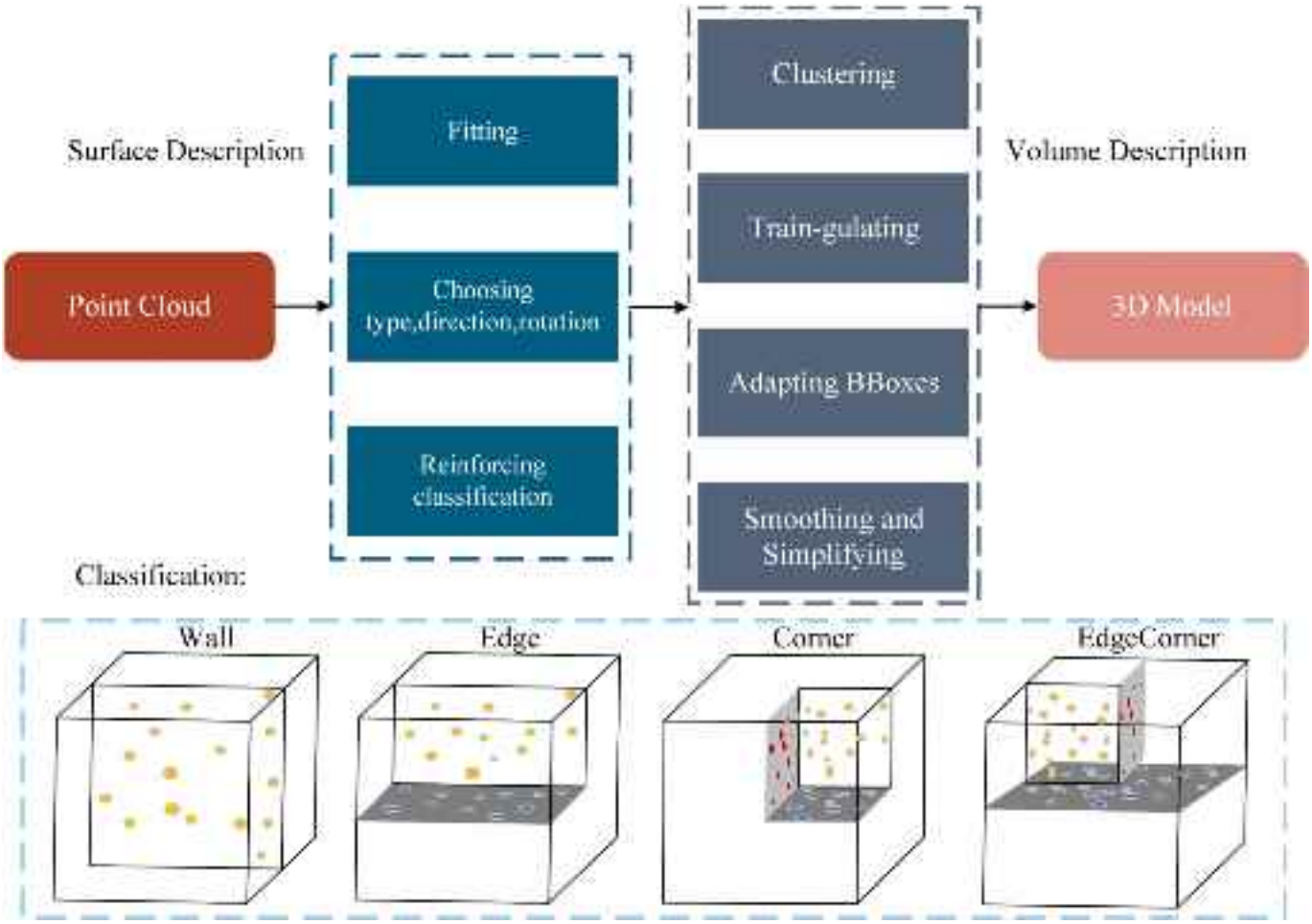


Fig. 5. Steps of MWM calculation.

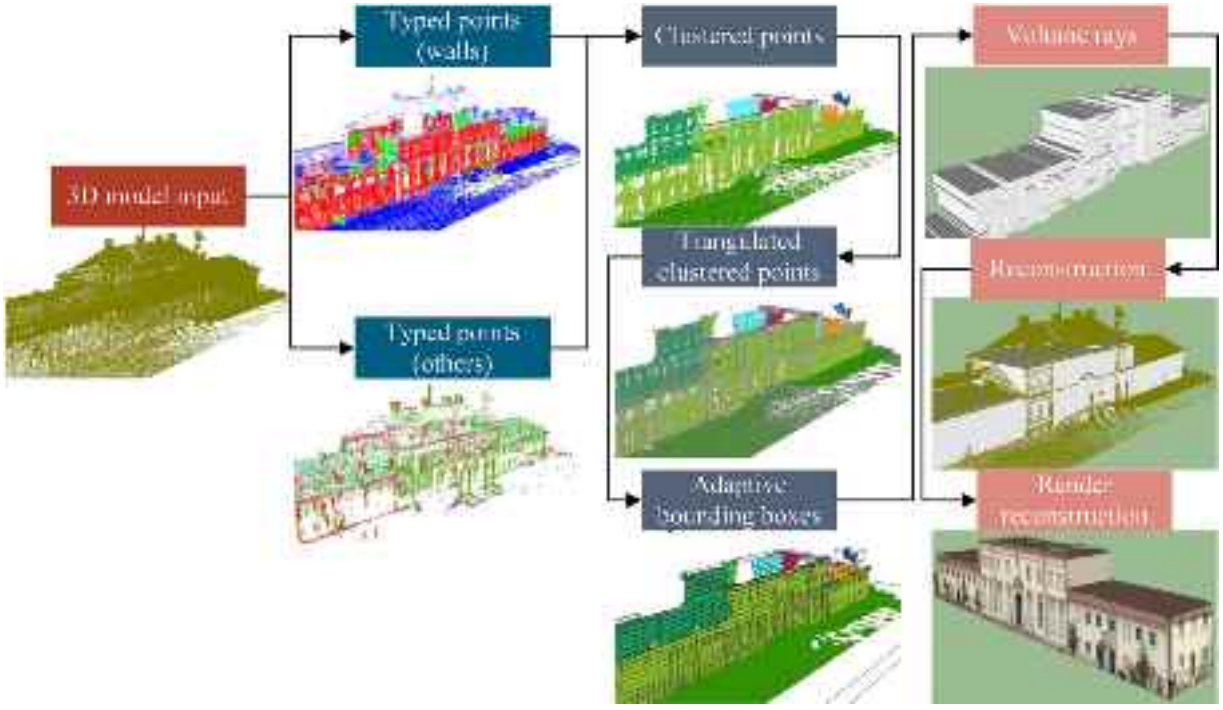


Fig. 6. MWM reconstruction workflow.

In this study, the MWM is utilized to establish a relationship between the continuous scene representation obtained through NeRFusion modeling and the Manhattan world concept. By applying a predefined threshold, points within a specified distance are classified as part of the primary structure. This process facilitates the identification of the primary structure's distinctive outlines and geometric features. Following the main building's extraction, the next step entails measuring its dimensions—height, width, and length. The Manhattan distance function can determine these dimensions. The vertical distance, in particular, aids in determining the primary building's height:

$$Height = max\left(\left|Z_{top} - Z_{bottom}\right| \cdot C\right) \tag{20}$$

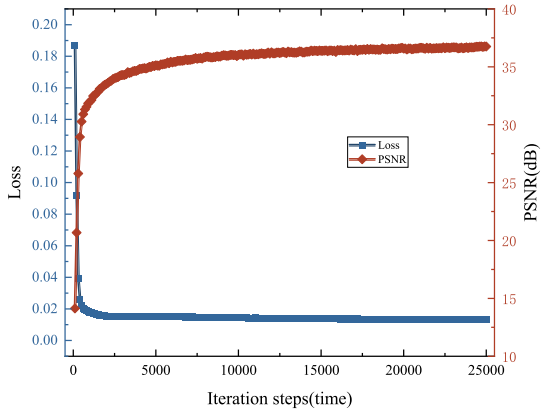
The variable Z_{top} represents the z-coordinate of the upper plane or wall, while the variable Z_{bottom} represents the z-coordinate of the lower plane or wall; The calibration coefficient C, which represents the conversion factor between pixel size and physical size. Similarly, the dimensions of width and length can be quantified.

4. Model training and testing

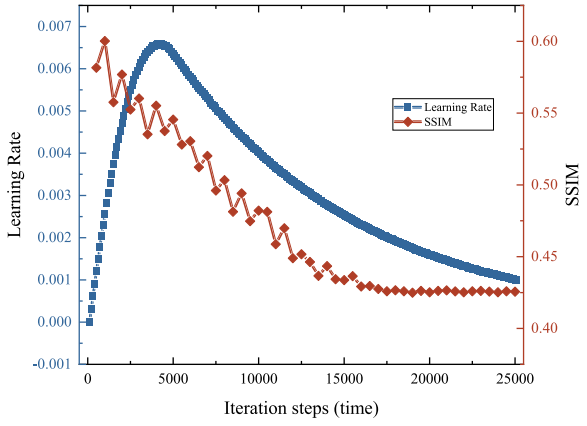
4.1. Training process

Model training parameters were carefully set: two proposal sampling rounds with 64 samples each, followed by a final NeRFusion sampling round of 32 samples. Anti-aliasing layer loss, with rectangle pulse widths of 0.03 and 0.003 for the initial and second rounds, respectively, and a loss multiplier of 0.01, was implemented. Feature codes underwent normalized weight decay with a loss multiplier of 0.1. The learning rate was incrementally adjusted from 10^{-8} to 1 over the first 5 k iterations. The model utilized 10 grid scales from 16 to 8192, each with 4 channels, within the iNGP grid and hash hierarchy, across 25 k iterations with a batch size of 216. Optimization was achieved using the Adam optimizer, parameters set to $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\varepsilon = 10^{-15}$, with a learning rate logarithmically decaying from 10^{-2} to 10^{-3}.

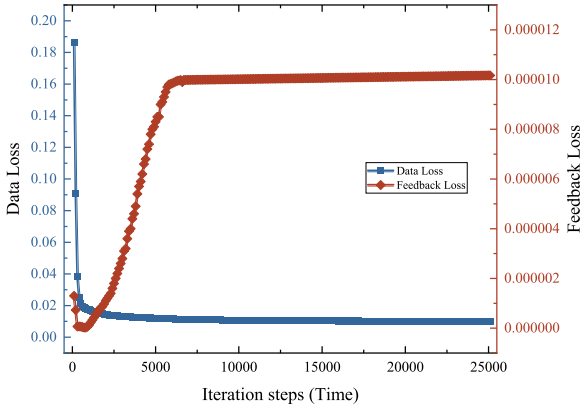
The NeRFusion model is trained using image data and camera parameters estimated by colmap. Several metrics assess training quality: Loss, reflecting overall training loss, indicates model fit to the training data, with lower values denoting better performance. PSNR measures the generated and original image similarity, with higher values indicating better image quality. Learning Rate controls the model's weight update rate per iteration. Data Loss quantifies the model-generated and training data image difference, with lower values indicating a better fit. Feedback Loss, used in backpropagation, guides weight updates to minimize overall loss. Distance Loss gauges the disparity between the model-generated depth map and the true depth map. Hash Loss assesses the similarity between the model-generated hash or encoding and the true hash. The NeRFusion model utilizes a two-stage sampling strategy and integrates an anti-aliasing mechanism during training. Two-stage sampling accelerates training, while the iNGP module facilitates rapid multi-scale mesh feature extraction. The anti-aliasing mechanism reduces sawtooth artifacts through multi-stage sampling and feature weighting. The training process for 3D reconstruction of building



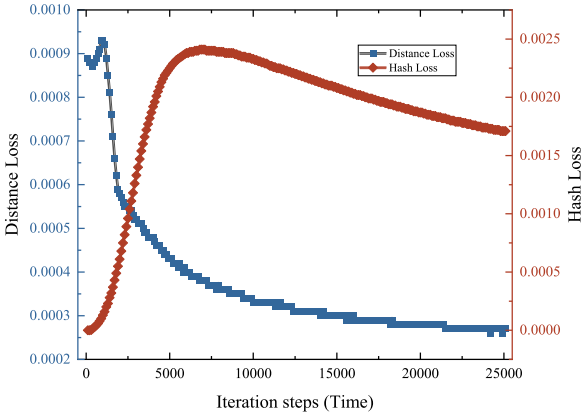
(a) Variation of PSNR and Loss with training steps



(b) Variation of LR and SSIM with training steps



(c) Evolution of Data Loss and Feedback Loss with Training Steps



(d) Variation of Distance Loss and Hash Loss with Training Steps

Fig. 7. Parameter variation with iteration steps during training process of model.

structures, illustrated in Fig. 7, exhibits favorable trends. PSNR, indicating image quality, significantly improves from 14.25 dB to 36.23 dB. Concurrently, Loss decreases from 0.19 to 0.019, signifying progressive scene depiction refinement.

Throughout the training process, as shown in Fig. 7, the learning rate is systematically reduced to ensure stable model convergence. In contrast, the SSIM exhibits a fluctuating pattern but shows an overall declining trend, suggesting the model's advancements in generating both the structure and finer details of the images. Additionally, data loss declines in the early training stages and later stabilizes, indicating the model's improved alignment with the training data. The back-propagation loss, termed feedback loss, initially decreases rapidly, followed by gradual stabilization. This pattern indicates progressive smoothing in the weight update process, contributing to training stability. Distance Loss exhibits a slight initial increase followed by a rapid decrease, indicating improved generation of accurate depth maps. Hash Loss initially increases rapidly then decreases gradually, suggesting challenges in effective hash encoding. However, over time, the model demonstrates gradual improvements in this aspect. The observed trends in these metrics indicate progressive enhancement in the model's image quality and precision during training, ultimately enhancing 3D architectural reconstruction.

Using the same computer hardware configuration, the PSNR values of the models proposed in this study exceed the highest values of Mip-NeRF and iNGP in Table 1 (33.09 dB and 30.05 dB, respectively). Training time is subject to variation based on factors such as image resolution, data volume, and computational performance. Specifically, in the same computing environment utilizing the NeRF datasetSynthetic, the modeling time of the proposed model is approximately 50 min, approximately 15 times faster than that of the original NeRF model.

4.2. Test background description

The test collected 98 images of a library building captured from multiple perspectives. Aerial photography was conducted using a DJI Mavic 3 Pro drone, equipped with a 4/3-in. CMOS sensor and a 20-megapixel main camera, providing an image resolution of 5280 × 3956. The data gathering, employing oblique photography mode, was conducted in August 2023 at a university library located in Yunnan, China. Subsequent experiments were uniformly conducted on a computing system equipped with an Nvidia RTX 3090 graphics card and 64 GB of RAM (Random Access Memory). Image data preprocessing entailed the estimation of parameters pertinent to the camera's internal and external characteristics, utilizing colmap software [55,56] for this purpose.

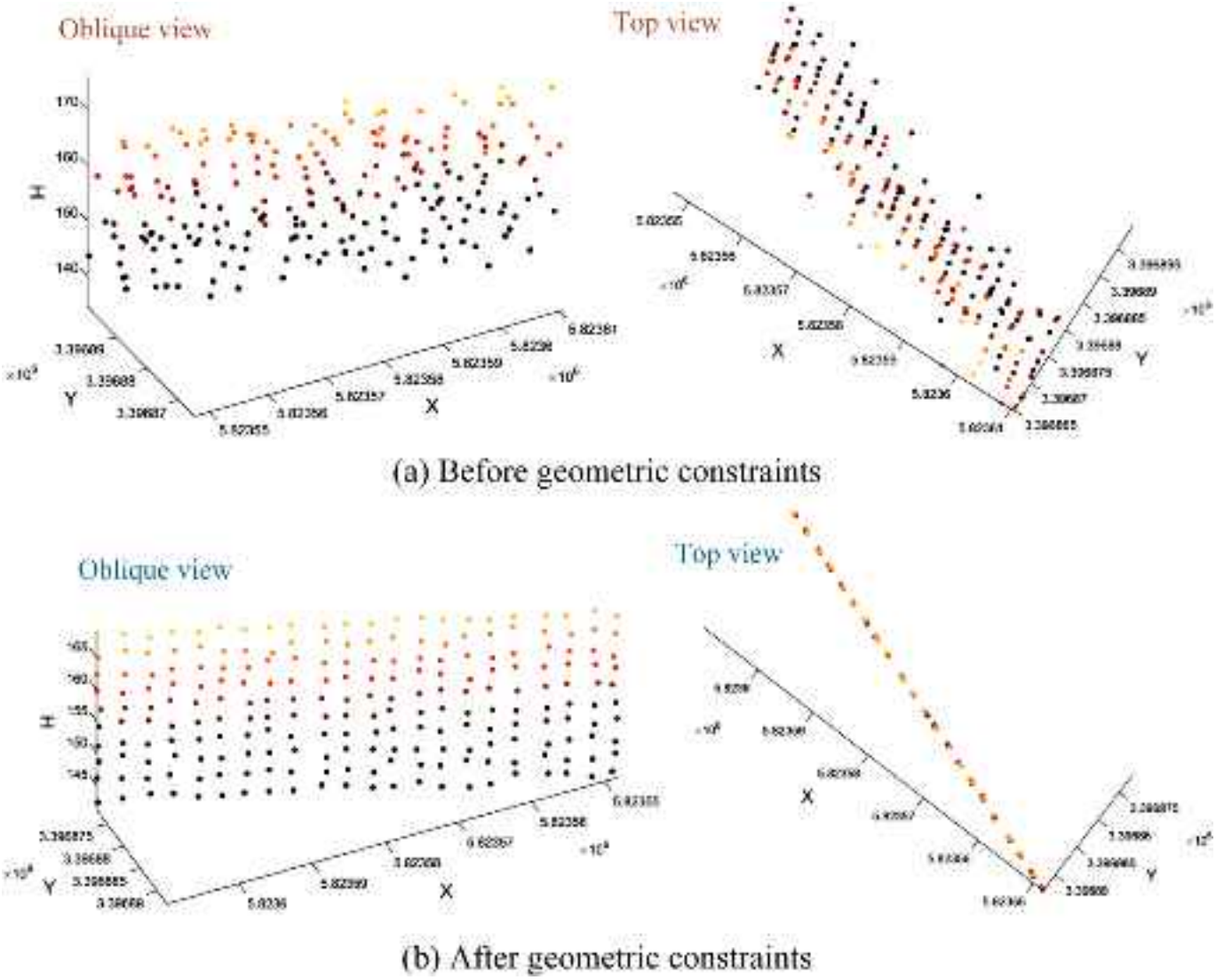


Fig. 8. Comparison before and after adding geometric constraints for planar reconstruction of model.

4.3. Test results

Insights from Fig. 8 underscore the critical role of geometric constraints—including normal consistency, plane fitting, and vertical/horizontal constraints—in significantly enhancing the NeRFusion model's framework. This integration promises substantial improvements in the overall quality of the reconstructed results. Key findings are summarized as follows:

1) Geometric constraints effectively direct the model to generate results aligning closely with the scene's structure, ensuring consistent normal vectors for walls, columns, etc., and establishing horizontal planes on the roof, among other elements.
2) Including these constraints results in notable improvements in outcome realism and accuracy. Consequently, interactions between scene elements demonstrate enhanced cohesion.
3) Furthermore, applying these constraints enhances stability in the generated outcomes. This stability manifests in consistent scene structure rendering from various viewpoints, mitigating quality fluctuations.

Incorporating these constraints enhances stability and fidelity of the generated results, ensuring consistent scene structure representation from various perspectives and minimizing quality fluctuations. Fig. 8 demonstrates a comparison of modeling planar structural objects. Without geometric constraints, the model's estimated points exhibit a random distribution on either side of the plane, attributable to randomness in matching points and associated errors. However, integrating the NeRFusion model with geometric constraints notably promotes estimation of points with high coplanarity. The robust influence of geometric constraints is evident.

The NeRFusion model's performance is exemplified through the transformation process depicted in Fig. 9, demonstrating significant improvements in image quality across training iterations. Initially, the model displays limited sensitivity to geometric constraints, resulting in outputs marred by noise, artifacts, and structural inaccuracies. However, as training progresses, these issues are mitigated through the integration of geometric constraints, leading to a notable enhancement in rendering fidelity. This progression is quantitatively supported by an increase in the PSNR index from 18.51 dB to 36.23 dB by 25,000 steps, indicating a marked improvement in image quality and alignment with actual scene structures.

Further analysis reveals that early in the training, network parameters are not optimally configured, contributing to high-frequency noise and underfitting. With continued training, the incorporation of geometric constraints refines the model's output, enhancing depth estimation and reducing artifacts, notably ghosting phenomena. This refinement leads to significant improvements in color and detail reproduction, attributed to the model's multi-scale feature extraction mechanism. Consequently, the model achieves realistic reconstructions of primary geometric structures and intricate details. The introduction of geometric constraints not only aids in network convergence, enhancing stability and efficiency but also serves as a form of regularization against overfitting, thereby optimizing the training process. These findings underscore the critical role of geometric constraints in enhancing the NeRFusion model's rendering quality, as evidenced by the sustained improvement observed throughout the training duration.

The NeRFusion model effectively captures the 3D structural characteristics of the building scene. As demonstrated by Fig. 10's frontal rendering, the model accurately reconstructs the building's primary structure, including window arrangements, roof configurations, and other elements. Lateral renderings offer a detailed view of the building's side elevation, showcasing features like window sills and the wall's grid structure with greater clarity, highlighting accurately represented fine details. Moreover, posterior renderings show the model's precise grasp of the building's 3D morphology, effectively revealing concealed rear details. Overall, the model effectively captures and encodes a 3D representation of the scene, facilitating high-quality detail rendering from various perspectives. The model's consistent representation across viewpoints indicates global coherence and minimal quality loss. This suggests that including geometric constraints positively impacts result stability. The visual representations show high realism and authenticity, with architectural elements displaying coherence without distortions. These observations support the claim that the model captures the scene's intrinsic physical characteristics. This achievement is attributed to including geometric constraints like normal vectors and plane constraints. Therefore, the tests demonstrate the NeRFusion model's robust modeling capabilities in intricate architectural scenes. The system acquires a profound understanding of a building's intricate 3D architecture and generates accurate visual depictions, incorporating fine details from any perspective, while respecting physical constraints.

This study utilizes the MWM to outline the primary structure in a

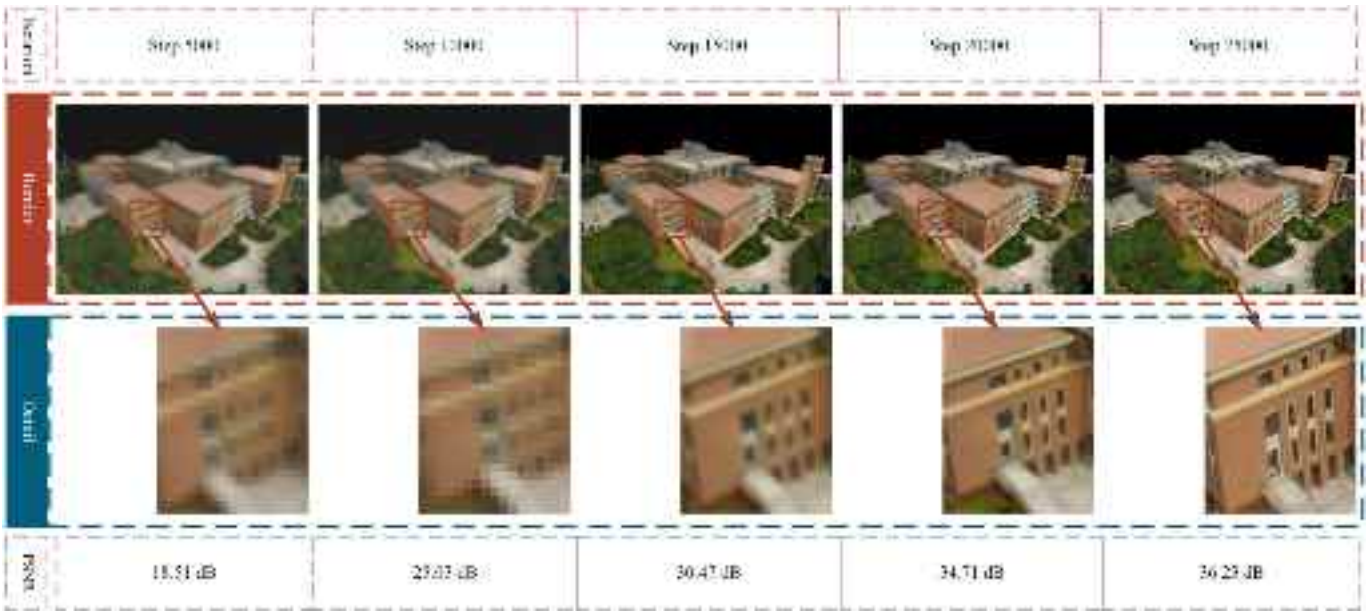


Fig. 9. Rendering of model with steps.

Fig. 10. Reconstructed renderings of large buildings.

complex architectural environment, aiming at detailed modeling. Initially, employing prior knowledge of the MWM, it computes the 3D masks of the scene's primary planar components. As shown in Fig. 11, the model effectively identifies three primary planar clusters in the scene, each perpendicular to the others, reflecting standard architectural features. Subsequently, the computed 3D mask segments the primary building structures from the rendered scene. Fig. 11 demonstrates that the contours of the primary structure, featuring well-defined wall boundaries and window layouts, are precisely extracted. The consistency of primary building segmentation across perspectives confirms the planar assumptions' validity employed by the MWM. In this study, major architectural elements in a complex scene were analyzed and extracted using the a priori knowledge of MWM and 3D mask calculations. This process establishes the foundation for detailed modeling and rendering of the primary structure. The work highlights the effectiveness of integrating a priori structural knowledge into neural networks, enhancing outcome congruence with real-world physical properties.

This study utilizes the MWM to determine the primary building contour and dimensions in three orientations: length, width, and height. The model's estimation precision for the building's front length and

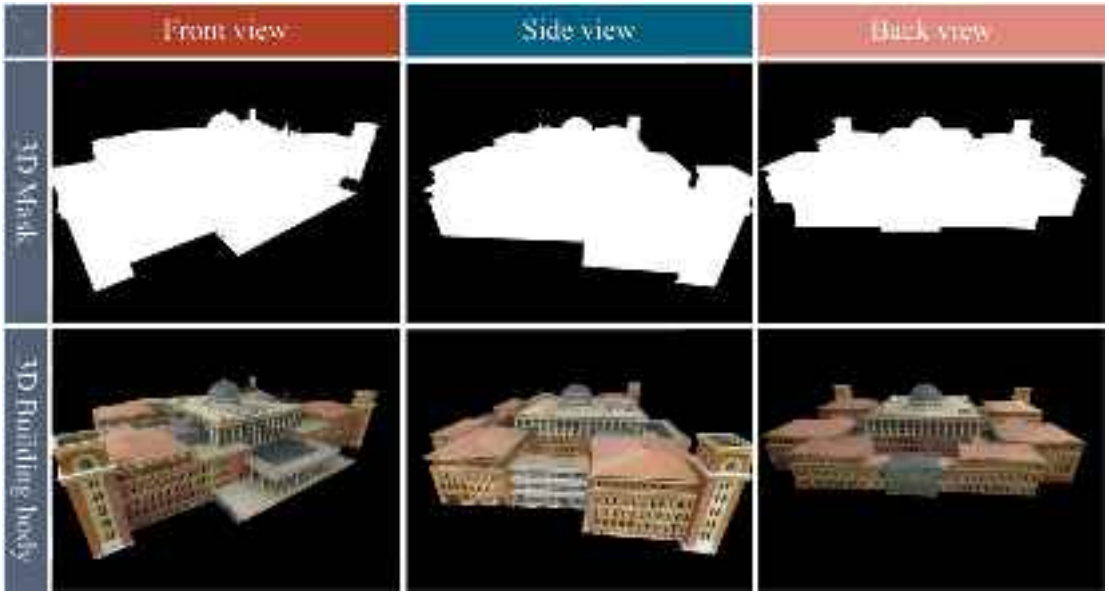


Fig. 11. Building body extracted using MWM.

width is evident in Fig. 12's contours and quantitative results. Length and width are calculated at 157.5 m and 152.5 m, respectively, precisely defining the building's ground plan. Height is determined at 23.6 m, aligning with expectations. For the rear perspective, contour maps and calculations show a length of 157.2 m, a width of 152.8 m, and a height of 23.5 m, demonstrating close to the calculations for the front perspective. This consistency highlights the model's reliability in determining building dimensions from various viewpoints. The side view profile indicates the building's side elevation extent, with length, width, and height measured at 157.3 m, 152.6 m, and 23.5 m, respectively. Summarized in Table 2, the building's 3D dimensions underscore the model's robust understanding of the 3D scene. These findings indicate the model's excellence in predicting the primary structure's 3D measurements, serving as a crucial reference for detailed modeling. Moreover, this paper emphasizes the MWM's efficacy and its ability to accurately capture real-world geometric properties.

Table 2
Building profile dimensions calculated by model.

	Length (m)	Width (m)	Height (m)
Front View	157.5	152.5	23.6
Back View	157.2	152.8	23.5
Side View	157.3	152.6	23.5

5. Comparative analysis of cluster building reconstructions

5.1. Comparative analysis backgrounder

This study aims to evaluate and compare the fused NeRF model's effectiveness with Mip-NeRF and iNGP for 3D reconstruction of clustered buildings near a train station. The study area covers approximately 2000 acres, featuring residential structures, railway infrastructure, and industrial facilities. Data collection was conducted in August 2023 using a DJI Mavic 3 Pro drone, equipped with a 4/3-in. CMOS sensor and a 20-megapixel main camera. Imagery was captured near a railway station in oblique photography mode. A total of 243 high-resolution images were collected, each with a ground sample distance (GSD) of 2.6 cm per pixel. The camera's internal and external parameters were computed with colmap software and used in the reconstruction process alongside the original images.

5.2. Model comparison and analysis

Note that the MWM has limitations in reconstructing 3D structures of clustered buildings. This model assumes buildings have distinct vertical and horizontal lines, suited for those with regular geometries. However, clustered buildings may deviate from this regularity due to varied design and foundation norms. Consequently, extracting key features of clustered buildings with the MWM may be inaccurate or incomplete. Therefore, this paper solely compares the models' 3D scene reconstruction capabilities.

Findings in Fig. 13 compare the three models' 3D reconstruction capabilities for clustered buildings. The model proposed exhibits exceptional performance across critical dimensions. This model excels in generating high-quality outputs and capturing intricate local details. The model has a strong ability to replicate the cluster buildings' external and geometric characteristics accurately. This includes accurately replicating geometric elements such as walls, doors, windows, and roofs. Moreover, the model accurately reproduces buildings' intricate textures, enhancing visual fidelity. The performance advantage can partly be attributed to incorporating geometric constraints. Incorporating geometric constraints enhances the model's architectural understanding, improving reconstruction precision. Consequently, this modeling approach enhances rendering outcomes and addresses localized intricacies.

A thorough comparison and evaluation of the reconstruction results from the three models were conducted. Initial findings reveal suboptimal performance by the Mip-NeRF model in reconstruction, exhibiting noise, voids, and blurriness in rendering. Additionally, it struggled to accurately capture architectural structures and intricate textural details. Furthermore, its capabilities were limited to basic object discrimination, lacking in complex analysis. On the other hand, the iNGP model yielded the least satisfactory reconstruction outcomes. Despite commendable reconstruction speed, its learning ability was deficient, resulting in significant noise and limited accuracy in resolving structures.

Table 3 provides an insightful comparison of model training metrics, illustrating the proposed model's efficacy in handling high-resolution images of cluster buildings, demanding significant computational effort. Despite a slightly longer training duration of 2.26 h, the proposed model significantly outperforms Mip-NeRF and iNGP across several metrics. Specifically, it reduces training time by nearly 50.00% compared to Mip-NeRF and shows improvements in image quality metrics like SSIM, PSNR, and LIPIS. The SSIM score is 13.41% higher than Mip-NeRF's and 50.00% higher than iNGP's, indicating superior image structural integrity. The PSNR value exceeds Mip-NeRF's by 10.45% and iNGP's by 25.10%, indicating enhanced image fidelity. Moreover, the LIPIS score is over three times higher than Mip-NeRF's and eight times higher than iNGP's, highlighting exceptional perceptual detail replication. This analysis underlines the model's suitability for engineering applications, particularly in modeling buildings and clusters, achieving superior rendering quality and structural precision.

In summary, the NeRFusion model outperforms Mip-NeRF and iNGP across critical performance metrics. The remarkable reconstruction quality offers a practical solution for 3D reconstruction in complex architectural environments.

6. Discussion

This paper introduces the NeRFusion model, advancing 3D architectural reconstruction by combining Mip-NeRF's multi-scale capabilities with iNGP's rapid rendering. The model integrates geometric constraints to improve scene consistency and detail accuracy, effectively addressing challenges in complex architectural scenes. The use of the MWM enhances accuracy in depicting primary building structures. Experimental comparisons show NeRFusion's advantages over Mip-NeRF and iNGP in metrics like PSNR and SSIM. However, NeRFusion's

Fig. 12. Main building outline and dimensions.

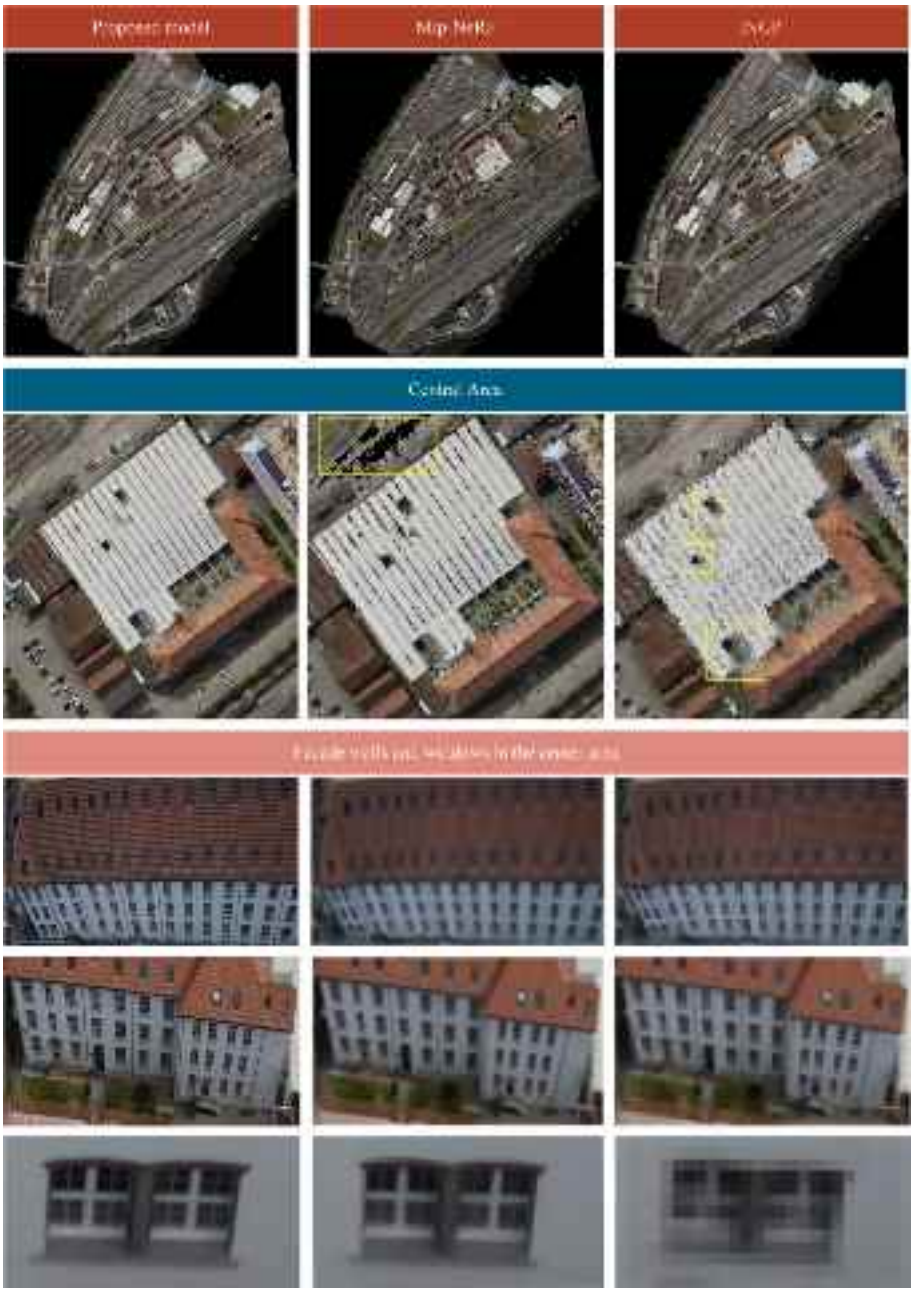


Fig. 13. Comparison of three model reconstructions based on cluster buildings.

Table 3
Comparison of three model training metrics.

	Proposed model	Mip-NeRF	iNGP
Training time (H)	~2.26	~4.55	~0.84
SSIM	0.93	0.82	0.62
PSNR(dB)	35.84	32.45	28.65
LIPIS	0.26	0.08	0.03

reliance on MWM restricts its effectiveness in clustered buildings due to MWM's assumptions of vertical and horizontal lines. Moreover, while NeRFusion improves efficiency and detail in reconstructing clustered buildings, its training time does not consistently outperform iNGP.

While the NeRFusion model demonstrates exceptional capabilities across numerous aspects, it is not without limitations. A notable example is the reconstruction of a stadium using 217 high-resolution images (5280 pixels × 3956 pixels) with a GSD of 0.8 cm per pixel. Although the model's overall reconstruction is impressive, as shown in Fig. 14, it encounters classic challenges. For instance, modeling glass curtain walls leads to inaccuracies due to numerous artifacts, primarily because the model struggles to capture the complex interactions of light with reflective and transparent surfaces accurately. Additionally, the performance in rendering fine details has not achieved real-world accuracy, mainly due to the NeRFusion model's limited capability to handle high-frequency details and complex textures during the multi-layer sampling and detail interpolation of input images. In some occluded areas captured from a limited number of viewpoints, the model's reconstruction, inference, and generalization are also inaccurate. This is primarily because NeRFusion relies on voxel-based rendering, which requires dense sampling to achieve high-quality reconstruction. These challenges highlight broader obstacles in the modeling domain, indicating areas for future improvement.

Extensive testing has shown that NeRFusion performs best with high-

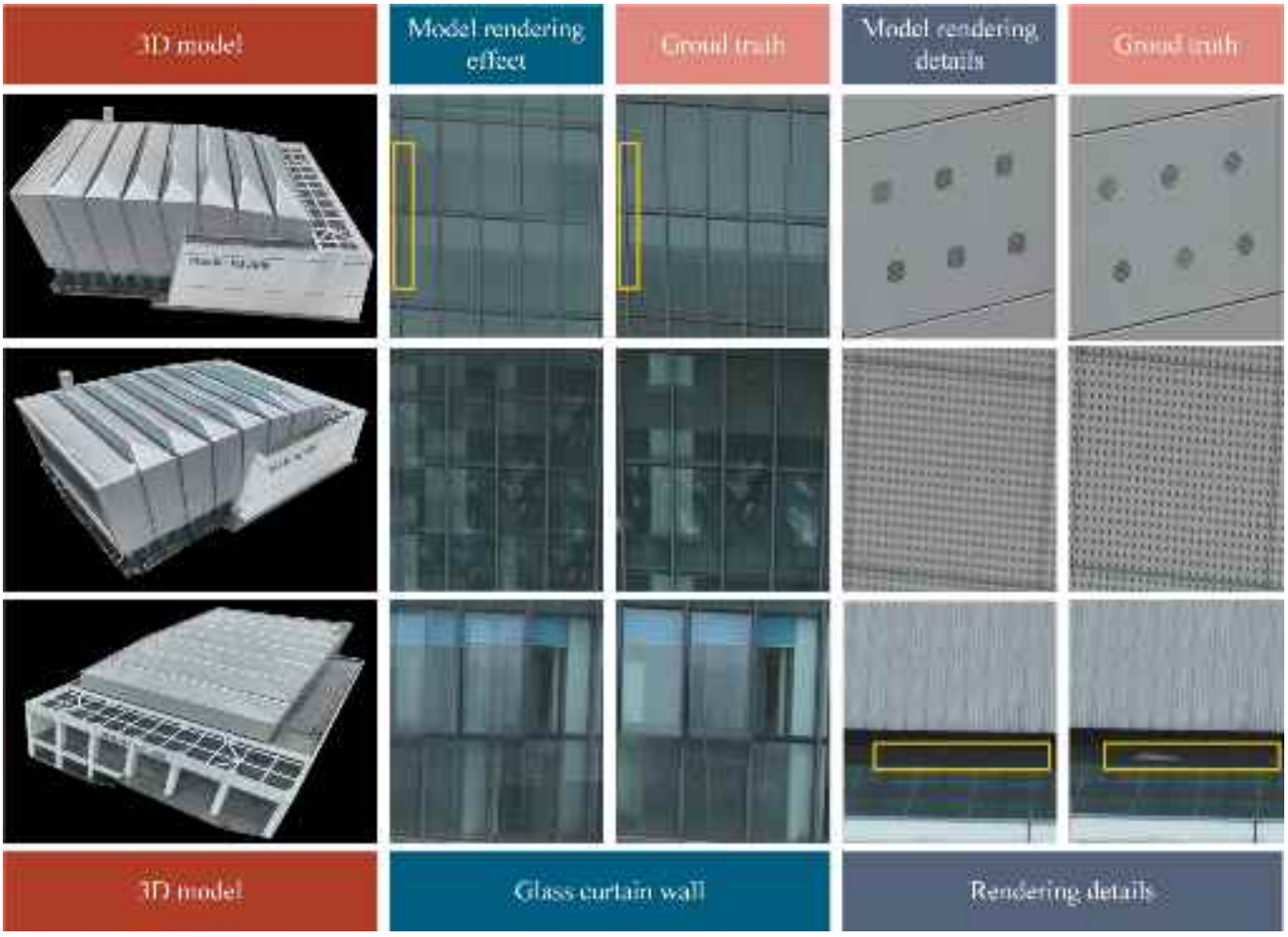


Fig. 14. Modeling the stadium using the NeRFusion model.

resolution images (above 2 K) and datasets ranging from 50 to 500 images. It excels at reconstructing regular structures; however, its accuracy decreases when handling complex spatial frameworks such as truss structures or intricate details in large scenes, like railway power lines. Accuracy improves when details are captured from multiple angles and scales. Additionally, NeRFusion is limited to static scenes and cannot be extended to dynamic ones. It also struggles with physical simulations, making it difficult to output explicit structures and edit them, as seen in traditional 3D reconstructions.

7. Conclusions

This paper presents the NeRFusion model to provide an efficient and accurate 3D reconstruction technique for architectural structures with advantages in visual fidelity, detail preservation, and training efficiency for single large buildings and building complexes. In conclusion, the main findings of this paper can be summarized as follows:

1) This paper amalgamates two advanced NeRF models, Mip-NeRF and iNGP, formulating the NeRFusion model for intricate 3D reconstruction of expansive architectural scenes. The present model exhibits enhanced training efficiency through techniques such as multisampling, weight reduction, and loss function optimization. It mitigates aliasing issues, resulting in substantial enhancement of rendering quality. Furthermore, the research incorporates geometric constraints—normal consistency, plane fitting, and vertical/horizontal constraints—into the NeRFusion framework. The primary objective is to elevate realism in generated building structures and enhance the quality and precision of the reconstruction process.
2) This paper employs the MWM to extract the principal component of the 3D architectural scene and accurately determine the outlines and measurements of the primary building. The empirical findings validate the effectiveness of this approach, although it is constrained by the fact that the MWM is specifically designed for individual structures, not clusters of buildings.
3) Ultimately, the NeRFusion model significantly outperforms existing models like Mip-NeRF and iNGP in reconstructing large structures, achieving a PSNR of 36.23 dB and reducing training time by 15 times. It excels in rendering details of buildings with high fidelity, showing a 13.41% and 50.00% improvement in SSIM over Mip-NeRF and iNGP, and a 10.45% and 25.10% increase in PSNR, respectively. These results showcase NeRFusion's superior rendering quality and efficiency in scene modeling.

Despite the NeRFusion model achieving notable success in 3D architectural reconstruction, demonstrating strengths in visual realism, detail retention, and training efficiency, its primary applicability remains with standard architectural shapes. Faced with unique designs such as stadiums, truss bridges, and ancient towers, the model encounters challenges in detail expression and adaptability, highlighting the limitations of the existing framework. Future efforts will focus on enhancing the model's capability for multi-source data fusion to improve detail accuracy and scene depiction. By combining physical illumination models with deep learning and using multi-source data

fusion techniques, it is hoped to enhance the rendering of optical properties of complex materials like glass curtain walls and address limitations in capturing depth information. Additionally, improving the NeRFusion model's ability to measure structural geometric dimensions, edit the model online, and support future physical simulations based on reverse engineering is necessary.

CRediT authorship contribution statement

Depeng Cui: Writing – review & editing, Writing – original draft, Software, Methodology, Data curation. **Weidong Wang:** Supervision, Project administration, Methodology, Funding acquisition. **Wenbo Hu:** Visualization, Software, Data curation. **Jun Peng:** Conceptualization, Data curation. **Yida Zhao:** Visualization, Software. **Yukun Zhang:** Formal analysis, Data curation. **Jin Wang:** Software, Project administration, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this study.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [grant numbers 52178442, U1734208, GJGS-2020-ZX-076], and in part by the High Performance Computing Center of Central South University.

References

[1] Peiran Ren, Wang Jiaping, Gong Minmin, Lin Stephen, Tong Xin, Guo Baining, Global illumination with radiance regression functions, ACM Trans. Graph. 32 (4) (2013) 1–12, https://doi.org/10.1145/2461912.2462009.
[2] Jonathan T. Barron, Jitendra Malik, Shape, illumination, and reflectance from shading, IEEE Trans. Pattern Anal. Mach. Intell. 37 (8) (2014) 1670–1687, https://doi.org/10.1109/TPAMI.2014.2377712.
[3] Andreas Humpe, Bridge inspection with an off-the-shelf 360° camera drone, Drones 4 (4) (2020) 67, https://doi.org/10.3390/drones4040067.
[4] Zipeng Qi, Zou Zhengxia, Chen Hao, Shi Zhenwei, Remote-sensing image segmentation based on implicit 3-D scene representation, IEEE Geosci. Remote Sens. Lett. 19 (2022) 1–5, https://doi.org/10.1109/LGRS.2022.3227392.
[5] Junjie Wang, Lei Ying, Yang Xiongjun, Zhang Fubo, A refinement network embedded with attention mechanism for computer vision based post-earthquake inspections of railway viaduct, Eng. Struct. 279 (2023), https://doi.org/10.1016/j.engstruct.2022.115572.
[6] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, Rohit Pandey, Sean Fanello, Gordon Wetzstein, Jun-Yan Zhu, Christian Theobalt, Maneesh Agrawala, Eli Shechtman, Dan B. Goldman, Michael Zollhöfer, State of the art on neural rendering, Computer Graphics Forum. 39 (2) (2020) 701–727, https://doi.org/10.48550/arXiv.2004.03805.
[7] Frank Dellaert, Yen-Chen Lin, Neural Volume Rendering: Nerf and Beyond, arXiv Preprint arXiv:2101.05204, 2020, https://doi.org/10.48550/arXiv.2101.05204.
[8] Yonas Zewdu Ayele, Mostafa Aliyari, David Griffiths, Enrique Lopez Droguett, Automatic crack segmentation for UAV-assisted bridge inspection, Energies 13 (23) (2020) 6250, https://doi.org/10.3390/en13236250.
[9] Tatsuro Yamane, Chun Pang-jo, Ji Dang, Honda Riki, Recording of bridge damage areas by 3D integration of multiple images and reduction of the variability in detected results, Comput. Aided Civ. Inf. Eng. 38 (2023) 2391–2407, https://doi.org/10.1111/mice.12971.
[10] Lingkun Chen, Haifeng Li, Mengmeng Huang, Convolutional neural networks (CNNs)-based multi-category damage detection and recognition of high-speed rail (HSR) reinforced concrete (RC) bridges using test images, Eng. Struct. 276 (2023) 115306, https://doi.org/10.1016/j.engstruct.2022.115306.
[11] Chi-Yun Liu, Jui-Sheng Chou, Bayesian-optimized deep learning model to segment deterioration patterns underneath bridge decks photographed by unmanned aerial vehicle, Autom. Constr. 146 (2023) 104666, https://doi.org/10.1016/j.autcon.2022.104666.
[12] Koubouratou Idjaton, Janvier Romain, Balawi Malek, Desquesnes Xavier, Brunetaud Xavier, Treuillet Sylvie, Detection of limestone spalling in 3D survey images using deep learning, Autom. Constr. 152 (2023) 104919, https://doi.org/10.1016/j.autcon.2023.104919.
[13] Hyunjun Kim, Yasutaka Narazaki, Billie F. Spencer Jr, Automated bridge component recognition using close-range images from unmanned aerial vehicles, Eng. Struct. 274 (2023), https://doi.org/10.1016/j.engstruct.2022.115184.
[14] Hyojoo Son, Changmin Kim, Changwan Kim, 3D reconstruction of as-built industrial instrumentation models from laser-scan data and a 3D CAD database based on prior knowledge, Autom. Constr. 49 (2015) 193–200, https://doi.org/10.1016/j.autcon.2014.08.007.
[15] Klaus Thoeni, A. Giacomini, R. Murtagh, E. Kniest, A comparison of multi-view 3D reconstruction of a Rock Wall using several cameras and a laser scanner, Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. 40 (2014) 573–580, https://doi.org/10.5194/isprsarchives-XL-5-573-2014.
[16] Joohyuk Lee, Son Hyojoo, Changmin Kim, Changwan Kim, Skeleton-based 3D reconstruction of as-built pipelines from laser-scan data, Autom. Constr. 35 (2013) 199–207, https://doi.org/10.1016/j.autcon.2013.05.009.
[17] Albitar, Chadi, Pierre Graebling, and Christophe Doignon. Robust Structured Light Coding for 3D Reconstruction. 2007 IEEE 11th international conference on computer vision. IEEE, (2007): pp. 1–6, doi: https://doi.org/10.1109/ICCV.2007.4408982.
[18] Song Zhang, High-speed 3D shape measurement with structured light methods: a review, Opt. Lasers Eng. 106 (2018) 119–131, https://doi.org/10.1016/j.optlaseng.2018.02.017.
[19] Ulrich Boesl, Time-of-flight mass spectrometry: introduction to the basics, Mass Spectrom. Rev. 36 (1) (2017) 86–109, https://doi.org/10.1002/mas.21520.
[20] Yunshan Jiang, Sebastian Karpf, Bahram Jalali, Time-stretch LiDAR as a spectrally scanned time-of-flight ranging camera, Nat. Photonics 14 (1) (2020) 14–18, https://doi.org/10.1038/s41566-019-0548-6.
[21] Siyuan Chen, Debra F. Laefer, M.ASCE, Eleni Mangina, S.M. Iman Zolanvari, Jonathan Byrne, UAV bridge inspection through evaluated 3D reconstructions, J. Bridg. Eng. 24 (4) (2019) 05019001, https://doi.org/10.1061/(ASCE)BE.1943-5592.0001343.
[22] Kwasi Nyarko Poku-Agyemang, Alexander Reiterer, 3D reconstruction from 2D plans exemplified by bridge structures, Remote Sens. 15 (3) (2023) 677, https://doi.org/10.3390/rs15030677.
[23] Yue Pan, Dong Yiqing, Wang Dalei, Chen Airong, Ye Zhen, Three-dimensional reconstruction of structural surface model of heritage bridges using UAV-based photogrammetric point clouds, Remote Sens. 11 (10) (2019) 1204, https://doi.org/10.3390/rs11101204.
[24] Silvio Savarese, Andreetto Marco, Rushmeier Holly, Bernardini Fausto, Perona Pietro, 3D reconstruction by shadow carving: theory and practical evaluation, Int. J. Comput. Vis. 71 (2007) 305–336, https://doi.org/10.1007/s11263-006-8323-9.
[25] Yuxuan Wang, Heng-Da Cheng, Juan Shan, Detecting Shadows of Moving Vehicles Based on HMM, in: 2008 19th international conference on pattern recognition, IEEE, 2008, pp. 1–4, https://doi.org/10.1109/ICPR.2008.4761498.
[26] Fangqiao Hu, Zhao Jin, Huang Yong, Li Hui, Structure-aware 3D reconstruction for cable-stayed bridges: a learning-based method, Comput. Aided Civ. Inf. Eng. 36 (1) (2021) 89–108, https://doi.org/10.1111/mice.12568.
[27] Ben Mildenhall, Pratul P. Srinivasan, Tancik Matthew, Jonathan T. Barron, Ramamoorthi Ravi, Ren Ng, Nerf: representing scenes as neural radiance fields for view synthesis, Commun. ACM 65 (1) (2021) 99–106, https://doi.org/10.1145/3503250.
[28] Marc Levoy, Pat Hanrahan, Light Field Rendering, Seminal Graphics Papers: Pushing the Boundaries. 2 (2023) 441–452, https://doi.org/10.1145/3596711.3596759.
[29] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, Michael Cohen, Unstructured lumigraph rendering. Seminal Graphics Papers: Pushing the Boundaries, Volume 2, Association for Computing Machinery. 2, 2023, pp. 497–504, https://doi.org/10.1145/3596711.3596764.
[30] Stephen Lombardi, Tomas Simon, Saragih Jason, Schwartz Gabriel, Lehrmann Andreas, Sheikh Yaser, Neural Volumes: Learning Dynamic Renderable Volumes from Images, arXiv Preprint arXiv:1906.07751 38 (4) (2019) 1–14, https://doi.org/10.1145/3306346.3323020.
[31] Guo Kaiwen, Lincoln Peter, Davidson Philip, Yu Busch Jay, Whalen Matt Xueming, Harvey Geoff, Orts-Escolano Sergio, Pandey Rohit, Dourgarian Jason, Tang Danhang, Tkach Anastasia, Kowdle Adarsh, Cooper Emily, Dou Mingsong, Fanello Sean, Fyffe Graham, Rhemann Christoph, Taylor Jonathan, Debevec Paul, Izadi Shahram, The Relightables: volumetric performance capture of humans with realistic relighting, ACM Transactions on Graphics (ToG). 38 (6) (2019) 1–19, https://doi.org/10.1145/3355089.3356571.
[32] Jonathan T. Barron, Ben Mildenhall, Tancik Matthew, Hedman Peter, Martin-Brualla Ricardo, P. Srinivasan Pratul, Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5835–5844, https://doi.org/10.48550/arXiv.2103.13415.
[33] Thomas Müller, Alex Evans, Christoph Schied, Alexander Keller, Instant neural graphics primitives with a multiresolution hash encoding, ACM Transactions on Graphics (ToG) 41 (4) (2022) 1–15, https://doi.org/10.1145/3528223.3530127.
[34] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, Andreas Geiger, Occupancy Networks: Learning 3D Reconstruction in Function Space, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4455–4465, https://doi.org/10.1109/CVPR.2019.00459.

[35] Zhiqin Chen, Hao Zhang, Learning Implicit Fields for Generative Shape Modeling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5932–5941, https://doi.org/10.1109/CVPR.2019.00609.
[36] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, Steven Lovegrove, DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 165–174, https://doi.org/10.1109/CVPR.2019.00025.
[37] Shunsuke Saito, Huang Zeng, Natsume Ryota, Morishima Shigeo, Kanazawa Angjoo, Li Hao, Pifu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2304–2314, https://doi.org/10.1109/ICCV.2019.00239.
[38] Vincent Sitzmann, Michael Zollhöfer, Gordon Wetzstein, Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations. Advances in Neural Information Processing Systems, ArXiv abs/1906.01618, 2019, https://doi.org/10.48550/arXiv.1906.01618.
[39] Michael Niemeyer, Lars Mescheder, Michael Oechsle, Andreas Geiger, Differentiable Volumetric Rendering: Learning Implicit 3D Representations Without 3D Supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3501–3512, https://doi.org/10.1109/CVPR42600.2020.00356.
[40] Lior Yariv, Yoni Kasten, Moran Dror, Galun Meirav, Atzmon Matan, Basri Ronen, Lipman Yaron, Multiview neural surface reconstruction by disentangling geometry and appearance, Adv. Neural Inf. Proces. Syst. 33 (2020) 2492–2502, https://doi.org/10.48550/arXiv.2003.09852.
[41] Dor Verbin, Peter Hedman, Ben Mildenhall, Zickler Todd, Jonathan T. Barron, Pratul P. Srinivasan, Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5481–5490, https://doi.org/10.1109/CVPR52688.2022.00541.
[42] Garbin, Stephan J. and Kowalski, Marek and Johnson, Matthew and Shotton, Jamie and Valentin, Julien, FastNeRF: High-Fidelity Neural Rendering at 200fps, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14326–14335, https://doi.org/10.1109/ICCV48922.2021.01408.
[43] Alex Yu, Vickie Ye, Matthew Tancik, Kanazawa Angjoo, PixelNeRF: Neural Radiance Fields from One or Few Images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4576–4585, https://doi.org/10.1109/CVPR46437.2021.00455.
[44] Ajay Jain, Matthew Tancik, Pieter Abbeel, Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 5865–5874, https://doi.org/10.48550/arXiv.2104.00677.
[45] Anpei Chen, Zexiang Xu, Zhao Fuqiang, Zhang Xiaoshuai, Xiang Fanbo, Jingyi Yu, Su Hao, MVSNeRF: Fast Generalizable Radiance Field Reconstruction from Multi-View Stereo, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14104–14113, https://doi.org/10.1109/ICCV48922.2021.01386.
[46] Alex Yu, Li Ruilong, Tancik Matthew, Li Hao, Ng Ren, Kanazawa Angjoo PlenOctrees for Real-Time Rendering of Neural Radiance Fields, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 5732–5741, https://doi.org/10.1109/ICCV48922.2021.00570.
[47] Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, Paul E. Debevec, Baking Neural Radiance Fields for Real-Time View Synthesis, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 5855–5864, https://doi.org/10.48550/arXiv.2103.14645.
[48] Wei Yi, Liu Shaohui, Rao Yongming, Zhao Wang, Jiwen Lu, Zhou Jie, NerfingMVS: Guided Optimization of Neural Radiance Fields for Indoor Multi-View Stereo, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 5590–5599, https://doi.org/10.48550/arXiv.2109.01129.
[49] Kangle Deng, Andrew Liu, Jun-Yan Zhu, Ramanan Deva, Depth-Supervised NeRF: Fewer Views and Faster Training for Free, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12872–12881, https://doi.org/10.1109/CVPR52688.2022.01254.
[50] Jian Zhang, Zhang Yuanqing, Fu Huan, Zhou Xiaowei, Cai Bowen, Huang Jinchi, Jia Rongfei, Zhao Binqiang, Tang Xing, Ray Priors through Reprojection: Improving Neural Radiance Fields for Novel View Extrapolation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18355–18365, https://doi.org/10.1109/CVPR52688.2022.01783.
[51] Xu, Qiangeng, Lifen Song, Sanae Oubeidillah, et al. Point-NeRF: Point-Based Neural Radiance Fields. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2022): pp. 5428–5438, doi: https://doi.org/10.1109/CVPR52688.2022.00536.
[52] Gang-Joon Yoon, Song Jinjoo, Hong Yu-Jin, Yoon Sang Min, Single image based three-dimensional scene reconstruction using semantic and geometric priors, Neural. Process. Lett. 54 (5) (2022) 3679–3694, https://doi.org/10.1007/s11063-022-10780-2.
[53] Han Dong, Jiao Zekun, Zhou Liangjiang, Ding Chibiao, Wu Yirong, Geometric constraints based 3D reconstruction method of tomographic SAR for buildings, SCIENCE CHINA Inf. Sci. 66 (1) (2023) 112301, https://doi.org/10.1007/s11432-022-3521-0.
[54] Carlos A. Vanegas, Daniel G. Aliaga, Bedrich Benes, Automatic extraction of Manhattan-world building masses from 3D laser range scans, IEEE Trans. Vis. Comput. Graph. 18 (10) (2012) 1627–1637, https://doi.org/10.1109/TVCG.2012.30.
[55] Mark William Smith, Jonathan L. Carrivick, Duncan J. Quincey, Structure from motion photogrammetry in physical geography, Prog. Phys. Geogr. 40 (2) (2016) 247–275, https://doi.org/10.1177/0309133315615805.
[56] Johannes L. Schönberger, Enliang Zheng, Marc Pollefeys, Pixelwise View Selection for Unstructured Multi-View Stereo. Computer Vision–ECCV 2016: 14th European conference, Amsterdam, the Netherlands, October 11-14, 2016, proceedings, part III 14, Springer International Publishing, 2016, pp. 501–518, https://doi.org/10.1007/978-3-319-46487-9_31.