remote sensing

MDPI

Article

3D Reconstruction of Ancient Buildings Using UAV Images and Neural Radiation Field with Depth Supervision

Yingwei Ge [1], Bingxuan Guo [1,*], Peishuai Zha [1], San Jiang [2], Ziyu Jiang [3] and Demin Li [1,4]

[1] The State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[2] School of Computer Science, China University of Geosciences, Wuhan 430074, China; jiangsan@cug.edu.cn
[3] School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China
[4] School of Artificial Intelligence, Zhejiang College of Security Technology, Wenzhou 325016, China
* Correspondence: b.guo@whu.edu.cn

Abstract: The 3D reconstruction of ancient buildings through inclined photogrammetry finds a wide range of applications in surveying, visualization and heritage conservation. Unlike indoor objects, reconstructing ancient buildings presents unique challenges, including the slow speed of 3D reconstruction using traditional methods, the complex textures of ancient structures and geometric issues caused by repeated textures. Additionally, there is a hash conflict problem when rendering outdoor scenes using neural radiation fields. To address these challenges, this paper proposes a 3D reconstruction method based on depth-supervised neural radiation fields. To enhance the representation of the geometric neural network, the addition of a truncated signed distance function (TSDF) supplements the existing signed distance function (SDF). Furthermore, the neural network's training is supervised using depth information, leading to improved geometric accuracy in the reconstruction model through depth data obtained from sparse point clouds. This study also introduces a progressive training strategy to mitigate hash conflicts, allowing the hash table to express important details more effectively while reducing feature overlap. The experimental results demonstrate that our method, under the same number of iterations, produces images with clearer structural details, resulting in an average 15% increase in the Peak Signal-to-Noise Ratio (PSNR) value and a 10% increase in the Structural Similarity Index Measure (SSIM) value. Moreover, our reconstruction model produces higher-quality surface models, enabling the fast and highly geometrically accurate 3D reconstruction of ancient buildings.

Keywords: 3D reconstruction; UAV images; neural radiation field; deep supervision; hash coding

Citation: Ge, Y.; Guo, B.; Zha, P.; Jiang, S.; Jiang, Z.; Li, D. 3D Reconstruction of Ancient Buildings Using UAV Images and Neural Radiation Field with Depth Supervision. *Remote Sens.* **2024**, *16*, 473. https://doi.org/10.3390/rs16030473

Academic Editor: Riccardo Roncella

Received: 14 November 2023
Revised: 20 January 2024
Accepted: 23 January 2024
Published: 25 January 2024

1. Introduction

The utilization of 3D reconstruction techniques not only facilitates the restoration of the original structure and color of ancient buildings but also enables the digital preservation of these historical treasures [1,2]. Through 3D reconstruction, meticulous digital replicas can be generated to safeguard and document these invaluable cultural legacies [3,4]. This paper employs the neural radiance fields (NeRF) technique [5] in the 3D reconstruction of ancient buildings, aiming to explore a swift and highly precise method for reconstructing buildings through neural rendering.

Unmanned Aerial Vehicles (UAVs) are known for their mobility, flexibility, speed and cost-effectiveness. Utilizing UAVs as aerial photography platforms enables the rapid acquisition of high-quality, high-resolution images, holding significant promise for the production of geographic mapping data [6,7]. With the advancement of tilt photogrammetry, techniques for dense point cloud generation and the construction of 3D triangular grid models from 2D images have matured, incorporating sparse reconstruction (Structure from Motion, SFM) [8] and dense reconstruction (Multiple View Stereo, MVS) [8,9]. This has

made 3D solid building reconstruction a reality. However, existing tilt photogrammetry-based 3D reconstruction methods are slow and entail substantial time overheads [10]. Dense reconstruction, which involves matching all or most of the pixels in multiple images, demands extensive data processing and often redundant computations, resulting in an overall low reconstruction efficiency. These limitations hinder its real-time applications [11]. Additionally, this method necessitates a complex process involving feature extraction, feature matching, depth fusion and Poisson reconstruction [12,13], which can introduce errors at various stages and lead to incomplete or flawed final results. This paper addresses the following issues that need to be resolved: (1) The conventional approach to reconstructing the surface model of ancient buildings is hampered by the slow processing speed. (2) The intricate surface textures found on ancient buildings, coupled with the presence of repetitive textures, can have a detrimental impact on the geometric accuracy of the model reconstruction.

In recent years, the NeRF technique, based on neural rendering, has gained extensive use in the field of 3D reconstruction. NeRF leverages neural implicit representation, employing neural networks to implicitly learn 3D scene features. It reconstructs triangular mesh models by combining these learned features with the Marching Cubes algorithm [14]. However, NeRF faces efficiency challenges due to the use of computationally intensive large Multilayer Perceptrons (MLPs), requiring hours or even days for training. Additionally, NeRF represents geometry by predicting the object density through neural networks, which lacks a strong physical foundation. This leads to the generation of triangular mesh models with rough surfaces, low geometric accuracy and suboptimal quality, limiting its applications [15]. Recent research has introduced new ideas based on NeRF, such as PlenOctrees [16] and Instant Neural Graphics Primitives (Instant-ngp) [17], aimed at accelerating NeRF network model training to minutes. However, these methods often compromise geometric accuracy, resulting in rough surface meshes that do not faithfully represent real-world physical structures. Subsequently, the Instant-NSR method [18] emerged, combining the approaches of Instant-ngp and NeuS [19], enhancing the model's geometrical structure. While this approach has improved the results, it may still exhibit depressions and uneven surface pits. Mip-NeRF [20] effectively resolves NeRF's challenges with high-frequency detail aliasing and distortion by refining the encoding of the sampling points, yet it still requires a considerable amount of time for network training. Neuralangelo [21] enhances the network architecture, but this advancement comes at the cost of increased computational demands and prolonged training periods due to additional sampling requirements. Meanwhile, 3D Gaussian splatting (3D GS) [22] introduces Gaussian functions for scene representation, offering increased adaptability in scene portrayal. However, its utility is somewhat constrained, as it struggles to accommodate images captured at varying scales.

In modern times, the 3D reconstruction of ancient buildings, achieved through the utilization of UAVs and various data collection methods, seeks to create more comprehensive models by integrating vast amounts of information. However, these data-rich approaches often lead to a significant computational burden in traditional 3D reconstruction, which places added strain on computers and prolongs the reconstruction process. Consequently, this paper proposes to improve the accuracy and training speed of reconstructions by combining the truncated signed distance function (TSDF) with sparse point cloud depth supervision, as well as implementing a progressive training strategy. This technique is introduced into the field of the three-dimensional reconstruction of ancient buildings to address the challenges of extensive computational demands and slow reconstruction speeds in traditional methods. This paper aims to enhance the geometric accuracy of NeuS-reconstructed models through two methods of geometric optimization. The primary contributions of this paper are as follows:

- Combined network training with the TSDF and depth supervision: Our approach combines the TSDF and depth supervision in network training. Integrating the TSDF into the signed distance function (SDF) neural network to improve geometric representation within the neural network. Simultaneously, this study utilizes sparse point

cloud depth information to supervise the training of the SDF neural network, further enhancing the geometric accuracy of three-dimensional mesh models.

- A progressive training method that gradually enhances the resolution of hash coding during the training process has been designed. This approach focuses on improving the characteristics of the scene and hash coding, effectively utilizing the feature hash table's capacity. By doing so, it mitigates hash conflicts within the mesh feature hash table under multi-resolution conditions. The ultimate goal is to produce rendered images with clear, detailed textures, enriching the visual quality.

This paper aims to enhance the accuracy of the NeuS-reconstructed geometric model through two geometric optimization methods. The first method involves the incorporation of the TSDF into the SDF neural network, which results in an improved geometric representation within the neural network. The second method utilizes depth information to supervise the neural network training, further enhancing the geometric accuracy of the reconstruction model using data from a sparse point cloud. In outdoor scenes, where large hash conflicts are common, this paper proposes a progressive training method based on multi-resolution hash coding technology to alleviate these conflicts and improve the expressive capabilities of the neural network.

2. Related Work

In a range of fields including mapping, remote sensing and computer vision, the NeRF technique has enabled the rendering and reconstruction of 3D scenes [23]. Despite its groundbreaking capabilities, NeRF still grapples with issues related to model generation efficiency, quality and scalability. One of the primary concerns is its computational intensity, both in terms of the number of sampling points and the time required for training, particularly due to the utilization of two large MLPs containing eight hidden layers [24]. Moreover, NeRF's reliance on straightforward volume rendering and direct density prediction through density MLP neural networks, lacking a robust physical foundation, often results in a rough surface and low geometric accuracy in the generated triangular mesh model [25]. In light of these challenges, researchers worldwide are dedicating efforts to improve and innovate the NeRF model.

In traditional geometric reconstruction, the literature [26–28] all focuses on the optimization of dense point clouds to enhance their quality. The literature [28] leverages images from multiple viewpoints, combines scene geometry constraints and estimates depths for sparse points to achieve high-quality dense reconstruction. The literature [26] proposes the sparse voxel DAGs method, efficiently reconstructing point clouds by establishing a sparse voxel data structure and employing dynamic adaptive mesh refinement and local region. The literature [27] presents a progressive 3D point set upsampling method based on localized blocks, gradually increasing the point density by utilizing the geometric and normal information among these blocks, thereby enhancing the point cloud details and resolution. However, due to the substantial memory requirements of these methods, they are more suited for small-scale reconstruction projects, where they tend to yield better results.

To address the issues of clarity and realism in NeRF technology, numerous researchers have conducted in-depth explorations into various aspects of the technology process, achieving significant improvements. To enable NeRF to handle a wider range of image situations and reduce its requirements for image sources, the literature [29] addresses the issue of NeRF producing poorer results with low-quality images by simulating the blurring process to synthesize blurred views, thereby improving NeRF's robustness to blurred input images. The literature [20] introduces Mip-NeRF, which transforms the original NeRF point sampling method into cone sampling, enriching the details of the sampling and considering the changes in the scale of the observation distance in ray sampling. The NeRF++ [30] model divides the scene into foreground and background parts. The foreground sampling method is consistent with NeRF, but background sampling involves projecting light onto a unit sphere, thus controlling the depth of light within a defined

range. Similarly, we have adopted this method in ancient architectural scenes, specializing in the encoding of foreground targets. The literature [31] integrates NeRF++ and Mip-NeRF concepts, ensuring positional relevance is maintained as sampling points extend to infinity. The literature [30,31] extends NeRF to large scene domains, but the increase in sampling information adds to the network training burden. To tackle the challenges of rough 3D models and noise low-fidelity geometric approximations, researchers both domestically and internationally have integrated deeper physical foundations into the geometric expression of neural networks to improve the accuracy. The literature [32] introduces UNISURF, using an occupancy network to represent implicit surfaces, assigning each sampling point as 0 or 1 to indicate the presence of a surface. The literature [33] presents Plenoxels, emphasizing the critical role of micro-voxel renderers in the evolution of NeRF technology. Plenoxels depart from using neural networks, focusing instead on optimizing the density and color parameters of voxel grid vertices through derivative-based solutions. This method achieves a training speed 3000 times faster than traditional NeRF. The literature [19] discusses NeuS, providing a mathematical explanation for NeRF's low geometric accuracy and employing SDF values to create an unbiased density function, thereby rectifying inherent biases in volumetric rendering formulas. To accelerate network training and reduce memory usage, the literature [34] presents NSVF, a strategy that manages scene data through a sparse voxel octree, selectively excluding irrelevant voxels during light sampling to speed up the process and minimize data overheads. The literature [17] proposes Instant-ngp, using a multi-resolution hash encoding (MHE) model [35] to encode the spatial information of 3D points, allowing for smaller MLP networks in training and rendering, marking a considerable advance in the NeRF training speed, reducing it from hours to just a few seconds. However, the need for pre-allocating fixed memory for data storage could lead to conflicts and impact the quality of results when training data volumes increase. To enhance the training efficacy, some researchers have integrated supervisory mechanisms during training. Point-NeRF [36] merges traditional MVS methods with NeRF, introducing a point cloud-based NeRF. The literature [37] uses MVS-generated depth maps to supervise SDF network training. Nerfing MVS [38] uses depth information from the NeRF network to train depth networks, then creates predicted depth maps to inversely guide NeRF network training. These methods, however, are time-consuming in generating depth information, leading to longer overall process times. Our approach, in contrast, does not use depth maps but instead employs sparse point clouds to gather depth information, considerably shortening the total process duration.

Despite the ongoing advancements in neural radiation field research, there remain certain unresolved issues: (1) The accuracy of neural radiation field reconstruction surfaces is not yet at a desirable level. (2) The training speed of the NeRF model remains relatively slow. To address these challenges, this study introduces a novel approach for surface representation based on multi-resolution hash coding using symbolic distance functions. Additionally, it also replaces the SDF with the TSDF to enhance model stability and employs sparse point cloud supervision to improve the depth expression within the model. Furthermore, this study advocates for the adoption of incremental training, aiming to significantly improve both the accuracy of the model reconstruction and training speed overall.

3. Methods

This paper integrates the multi-resolution hash position coding method and NeuS with the concept of a signed distance function into the NeRF framework for volume rendering. The optimization of the TSDF neural network, combined with sparse point cloud depth supervision, is utilized to reconstruct models of ancient buildings in outdoor environments from UAV images. The technology roadmap is depicted in Figure 1.

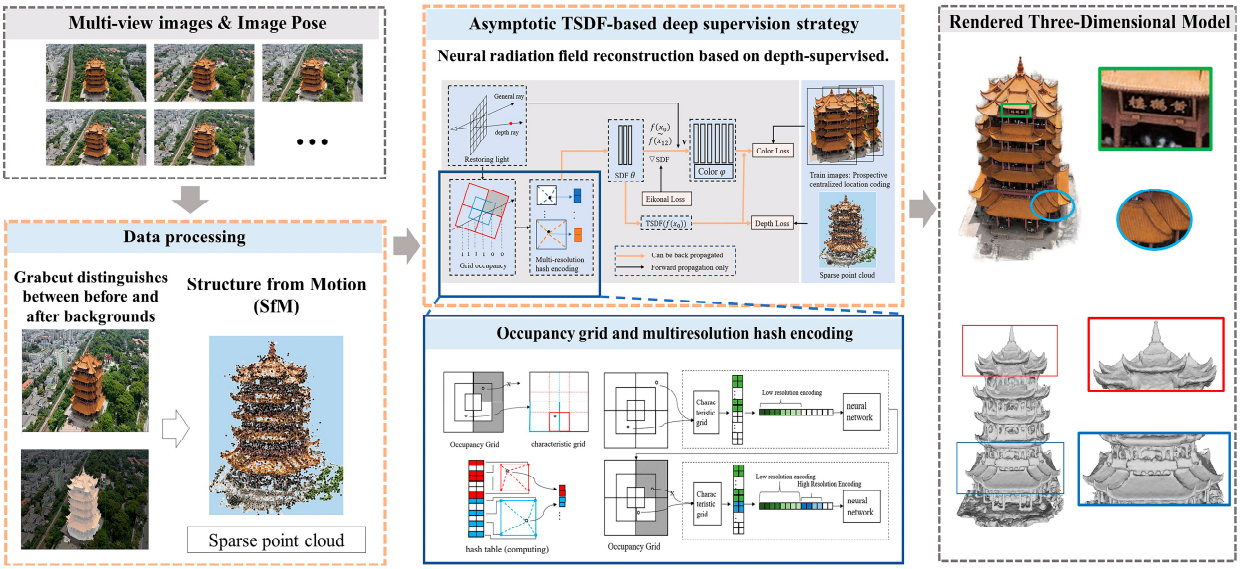

Figure 1. Flowchart of the algorithm of neural radiation field reconstruction based on depth supervision.

In this paper, the method is outlined as follows: starting from a pixel in an image and the light is recovered. The light passes through a multi-resolution hash grid and the internal hash features of the grid can be obtained using interpolation methods. These hash features are then combined with their positions in an SDF network. The SDF network provides SDF values and geometric features. These values, along with the viewing direction, are input into a color network to generate RGB values. The network is optimized by minimizing the difference between the output RGB values and the actual image pixel values. For pixels corresponding to sparse point clouds, the point cloud depth information is computed to supervise the optimization of the 3D model structure by weighting the pixel depths obtained from the TSDF values.

3.1. Data Processing

The fast retrieval feature of hash feature coding, as demonstrated in reference [17], has significantly reduced the training time of NeRF networks from hours to seconds. While multi-resolution hash coding provides computational efficiency by trading a larger memory footprint, the constraint is the finite memory and hash table size. This study introduces two methods to minimize conflicts when dealing with limited hash tables: (1) foreground centralized positional coding and (2) progressive multi-resolution hash coding, which will be detailed in Section 3.2.

Foreground centralized positional coding tackles the issue of growing scene content that exceeds the limited and fixed storage capacity of the 3D feature mesh. This overage results in severe hash conflicts in position encoding, which surpass the neural network's capacity to resolve. The surrounding environmental data can cause training neglect and result in image blurring.

In the wrap-around tilt photography approach, the scene is divided into foreground and background, as depicted in Figure 2; the foreground is our target object, while the background is the surrounding scene environment. The application of Grabcut [39] enables the distinction between the foreground (comprising the target building and the central region of interest) and the background (encompassing non-target scene elements along the image periphery). To enhance the neural network's grasp of vital target information, this paper primarily feeds the network with foreground information, while diminishing the influence of background data at the image edges. This approach curtails the feature overlap between critical information and edge information in the hash table, thereby reinforcing the network's attentional mechanism.

Figure 2. Grabcut distinguishes between before and after backgrounds.

Our depth supervision information is derived from a sparse point cloud, obtained through sparse reconstruction. Sparse reconstruction, also known as SfM, involves feature extraction from the input multi-view images, followed by feature matching to obtain homonymous image points between the images. Based on these homonymous image points, SfM can estimate the internal and external orientation elements of each image more accurately via methods such as forward rendezvous and backward rendezvous and obtain the sparse point cloud in the object-side space and use the depth information of the corresponding pixels of the point cloud as the a priori information for depth supervision.

3.2. Progressive Multi-Resolution Hash Coding

This paper employs progressive multi-resolution hash coding, as depicted in Figure 3, where blue represents low-resolution encoding grids, used for extracting low-resolution features, while pink represents high-resolution encoding grids, used for extracting high-resolution features. Hash coding can lead to data volume and hash conflict challenges. Progressive multi-resolution hash coding is adopted in this study, allowing low-resolution mesh features to capture scene or object outlines and similarities, while high-resolution mesh features prioritize detailed scene or object information.

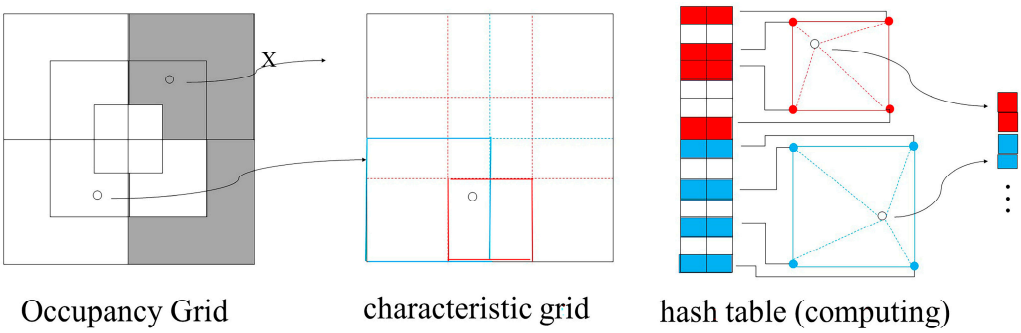

Figure 3. Occupancy grid and multi-resolution hash encoding.

Instant-ngp combines low-resolution and high-resolution feature encoding for all scene points, which results in hash conflicts and partial blurring of image details. Progressive multi-resolution hash coding, depicted in Figure 4, aims to prevent non-critical points from affecting high-resolution mesh features. This approach enhances the speed and accuracy of 3D building reconstruction for neural rendering.

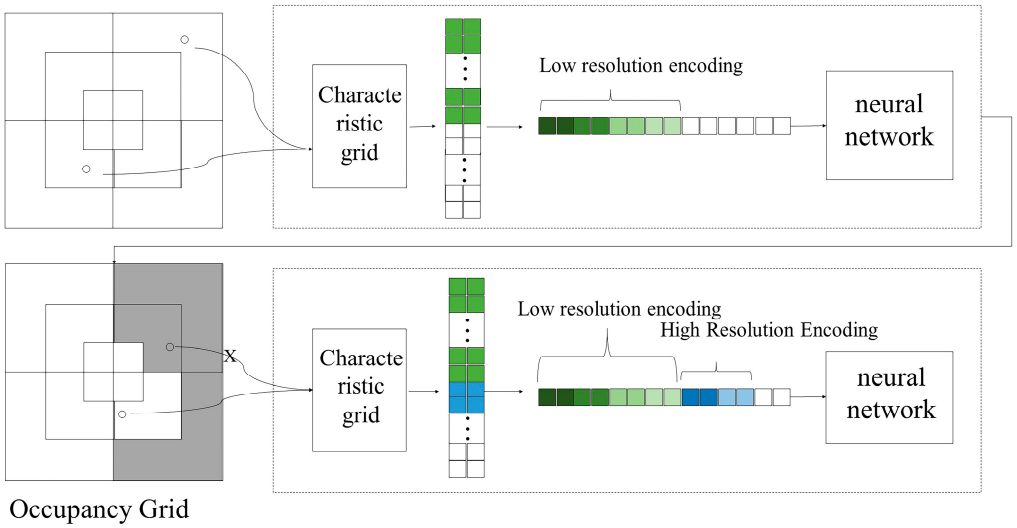

Figure 4. Asymptotic multi-resolution hash coding technology roadmap.

The proposed coding method follows a "from coarse to fine" principle. Initially, during network pre-training, high-resolution feature coding information is masked, while low-resolution hash feature coding is preserved to represent the model's general outline and location. Additionally, the low-resolution feature information is utilized to eliminate empty grid cells, speeding up light sampling and reducing interference from blank areas. As training progresses, the masking of high-resolution feature-encoding information is gradually reduced to enhance the model's surface representation. This encoding approach maximizes the utilization of the high-resolution hash feature table, mitigating hash conflicts to some extent. As a result, it leads to enhanced clarity in image rendering and a significant improvement in the detail of the geometric model.

3.3. Asymptotic TSDF-Based Deep Supervision Strategy

NeuS has exposed inherent errors in NeRF's volume rendering formulation, specifically related to the polar inconsistency of the density and weight values, which results in low geometric accuracy in the neural radiation field. This paper incorporates the concept of the SDF constraint network from NeuS and introduces the TSDF, a form of three-dimensional implicit expression. The TSDF represents an enhancement of the SDF concept, introducing truncation to create values within the range of $[-1, 1]$. The formula for the TSDF is depicted in Figure 5.

$$tsdf_i(x) = \max(-1, \min(1, \frac{tsdf_i(x)}{t})) \quad (1)$$

where t denotes the truncation distance and the TSDF will truncate to 1 or -1 when the absolute value of the SDF is greater than t. The TSDF reduces the variance between the data, increases the stability and makes it easier for the loss to converge in network training, while removing voxels that are farther away from the surface, reducing spurious airborne floats and decreasing the memory size of the reconstructed mesh.

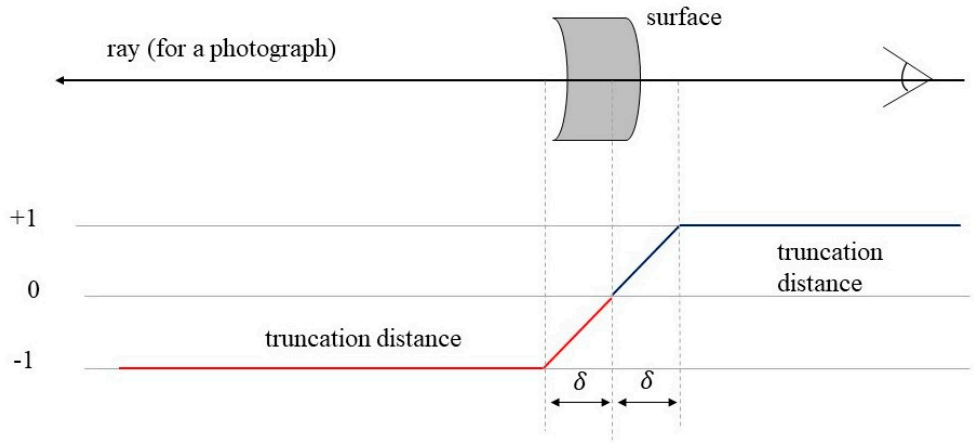

Figure 5. Truncated symbol distance function.

The TSDF is not differentiable at its truncation points, which makes it less suitable for neural network learning. In this paper, the Tanh function is introduced as an approximation of the TSDF. The computational formula is given in Equation (2), where "S" is a trainable hyperparameter and "Z" represents the value of the symbolic distance function. This function bears a resemblance to the TSDF, as both are monotonically increasing odd functions with a value range of $[-1, 1]$. During network training, the value of "S" is initially set to a smaller value, retaining the volume density of points farther from the surface. As the training progresses and the network's scene perception improves, "S" gradually increases, reducing the TSDF truncation distance, thereby focusing on preserving the volume density of points in closer proximity to the surface, which is critical for effective volume rendering.

$$TSDF = \frac{e^{SZ} - e^{-sz}}{e^{SZ} + e^{-sz}} \tag{2}$$

The TSDF neural network is established based on the SDF neural network, as depicted in the optimization flow chart in Figure 6, where the TSDF is introduced for truncation after the network outputs the SDF values, converting them into density values. Light-sampled spatial points are first filtered through the occupancy grid to retain points with high occupancy probabilities. These selected points undergo multi-resolution hash coding. The result of this coding is then fed into the SDF neural network, which produces a multidimensional feature vector where the first dimension represents the SDF value. The color neural network takes this feature vector along with additional information, such as the direction and normal vectors of the points output by the SDF neural network, and it outputs the RGB values. Each valid sampled point is assigned a density value, synthesized by the TSDF value and an RGB value. Points along the same ray are grouped together and their colors are combined according to an unbiased volume rendering formula to obtain the pixel's color value. During training, this paper employs network supervision for the RGB truth values, while the TSDF values are used to update the occupancy of the occupancy mesh. This explicit adjustment brings the voxels of the occupancy mesh close to the object's surface, effectively sieving out points that are far from the reconstructed surface or have no impact on the surface, thus enhancing the light sampling efficiency.

NeRF inputs are only image data and corresponding bitmap information. The rendering and reconstruction of the 3D scene are achieved solely based on the pixel values as supervision, which leads to a significantly constrained geometric representation within the neural network. On the one hand, there is an inherent error in the volume density values obtained by NeRF due to biased volume rendering formulas. On the other hand, there is a lack of supervision regarding the 3D information. In response to this situation, this paper introduces sparse depth information to supervise network training, aiming to enhance the neural network's capability to represent geometric structures.

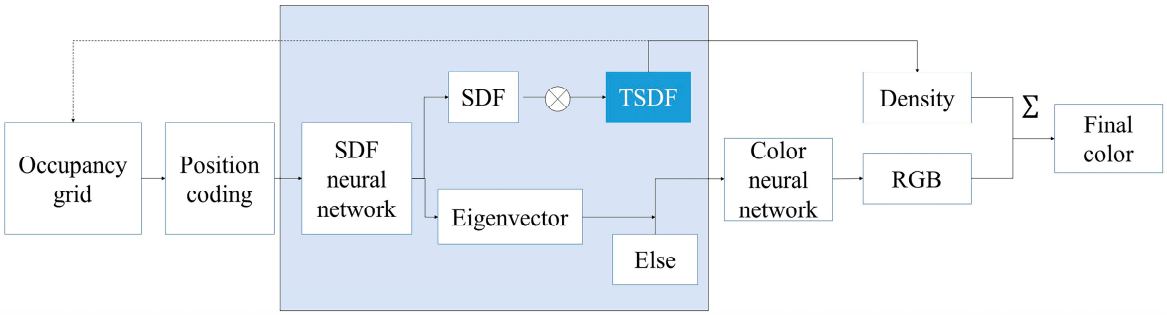

Figure 6. Optimization flow of TSDF neural network framework.

The sparse point cloud used in this paper is not for all pixels of all images, so the training of the deep supervised network is not for all rays. During the training process of the deep supervised network, this paper divides the training rays into two categories, which are ordinary rays and depth rays. As shown in Figure 7, ordinary rays are randomly extracted from all training images, while depth rays are extracted from the pixels corresponding to the sparse point cloud.

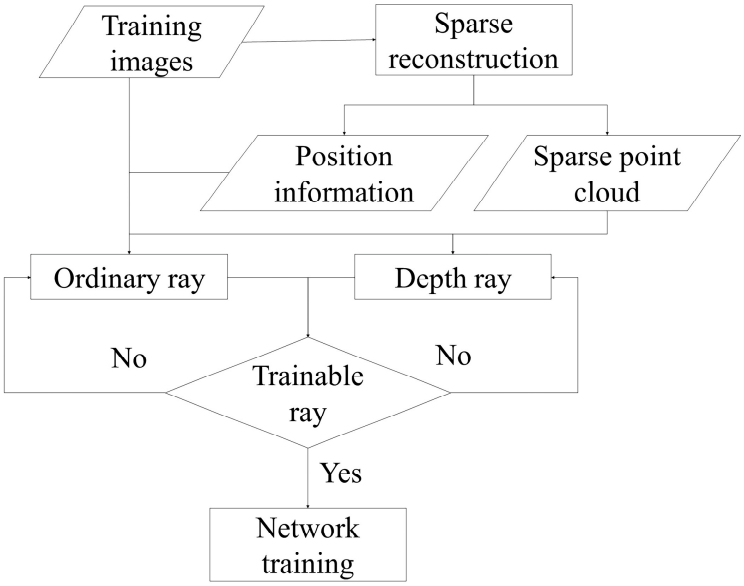

Figure 7. Recovery of normal and deep light training flowchart.

In this paper, the TSDF values obtained from network training are converted into weight values. This weight value can not only synthesize the color, but also the depth. Knowing the position and step spacing of all sampling points on the ray, it is easy to obtain the distance of each point from the origin, which is the depth value. By performing a weighted sum using the depth value and its corresponding weight value, the depth value for this specific ray can be accurately determined. As depicted in Figure 8, the neural network consists of two fully connected MLP networks: the SDF neural network and the color neural network. The SDF neural network comprises one hidden layer, while the color neural network comprises three hidden layers.

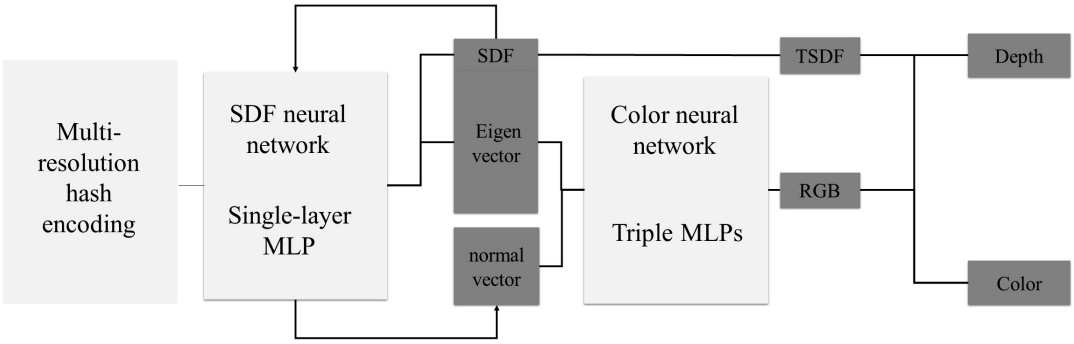

Figure 8. Flowchart of forward propagation of deeply supervised neural radiation field.

The inputs and outputs of the two networks are different. The input of the SDF neural network comprises three-dimensional point coordinates (x, y, z), which are encoded utilizing a multi-resolution hash position encoding methodology. The output from the SDF neural network is a feature vector of 13 dimensions. The foremost dimension of this vector signifies the SDF value, which can be further convertible into the TSDF value. The inputs of the color neural network are the 13-dimensional feature vectors, including the direction vector and the normal vector information of the point, where the normal vector can be obtained by finding the gradient of the SDF function or approximated by Equation (3). The output produced by the color neural network is a tri-dimensional vector, representing the RGB components.

$$\vec{n} = \begin{bmatrix} f(x+\varepsilon, y, z) - f(x-\varepsilon, y, z) \\ f(x, y+\varepsilon, z) - f(x, y-\varepsilon, z) \\ f(x, y, z+\varepsilon) - f(x, y, z-\varepsilon) \end{bmatrix} \tag{3}$$

To train the neural network, three loss functions are constructed in this paper, which are the color loss, SDF loss and depth loss. The color loss is calculated as follows:

$$\mathcal{L}_{color} = \frac{1}{m}\sum_k \mathcal{R}\left(\hat{C}_k, C_k\right) + \frac{1}{m}\sum_k MSE\left(\hat{C}_k, C_k\right) \tag{4}$$

where m denotes the number of rays per batch, $\mathcal{R}$ denotes the L1 loss, MSE denotes the mean square error loss and $\hat{C}_k$ and C_k denote the predicted and true color values.

The SDF loss is the Eikonal loss, which is used to constrain the symbolic distance function and is calculated as follows:

$$\mathcal{L}_{Eikonal} = \frac{1}{nm}\sum_{k,i}\left(\left\|\vee f\left(\hat{P}_{k,i}\right)\right\|_2 - 1\right)^2 \tag{5}$$

where n denotes the number of all sampling points, m denotes the number of rays per batch and $\vee f\left(\hat{P}_{k,i}\right)$ denotes the derivative of the SDF function, which can also be interpreted as the normal vector of the sampling points.

The depth loss is used to supervise the depth value of a depth ray and the depth loss of a general ray is calculated as follows:

$$\mathcal{L}_{depth} = \frac{1}{m}\sum_k MSE\left(\hat{D}_k, D_k\right) \tag{6}$$

where MSE denotes the mean square error loss, and $\hat{D}_k$ and D_k denote the predicted depth value and the true depth value.

4. Experiments

4.1. Experimental Data

In order to verify the effectiveness of the algorithm, three sets of DTU building datasets are used for the experiments in this paper; each set of data contain image data, mask data,

empty three-file data, etc., and the description of the datasets is shown in Table 1. When collecting the DTU data, the position of the camera is placed on a sphere with a radius of 50 cm and the camera is roughly 35 cm from the surface of the object.

Table 1. Description of the DTU dataset.

Dataset	Numbers of Image	Data Content
DTU15	49	Resolution (of a photo) 1600 × 1200 Camera parameters Mask data Point cloud data
DTU24	49	Resolution (of a photo) 1600 × 1200 Camera parameters Mask data Point cloud data
DTU40	49	Resolution (of a photo) 1600 × 1200 Camera parameters Mask data Point cloud data

The other set of experimental data are the UAV-acquired building image data, one set of Pix4d sample data and one set of self-collected data from the Yellow Crane Building, as shown in Table 2; the two sets of data are acquired by flying in a circular manner around the building. The third set of data are from Huayan Temple, consisting of five camera shots, with the shooting angle being from above the Huayan Temple tower.

Table 2. Drone image data.

Dataset	Number of Images	Image Size
Pix4d sample Data	36	4592 × 3056
Yellow Crane Data	60	3965 × 2230
Huayan Temple Data	40	6000 × 4000

4.2. Evaluation Indicators

The Peak Signal-to-Noise Ratio (PSNR), which can be used to measure the difference between two images, is calculated as shown in Equation (7).

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{MAX_G^2}{MSE}\right) \tag{7}$$

where MAX_G^2 is the maximum pixel value appearing in the truth image. Usually, if the pixel value is represented by B-bit binary, then $MAX_G = 2^B - 1$. MSE is the mean square error between the true value image G and the rendered image R of the same size. This paper uses color images, so it is necessary to calculate the PSNR of the three channels of RGB separately and take the average, as the final PSNR value. The higher the PSNR value, it means that the image is closer to the original image.

The Structural Similarity Index Measure [40] (SSIM) is a full-reference image quality evaluation index, which can better reflect the subjective perception of the human eye. The calculation is relatively complex, respectively, from the brightness L, contrast C and structure S, which are three aspects of the measure of image similarity. The formulas for the three functions are as follows:

$$\mathrm{L(x,y)} = \frac{2\mu_x\mu_y + C_1}{u_x^2 + \mu_y^2 + C_1} \tag{8}$$

$$C(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{9}$$

$$S(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_1} \tag{10}$$

where μ denotes the mean, σ denotes the variance and C_1, C_2 and C_3 denote the constants used to keep the formula stable; the $\sigma_x\sigma_y$ in the above formula is calculated as follows:

$$\sigma_x\sigma_y = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) \tag{11}$$

SSIM combines the three functions, and the final formula is as follows:

$$SSIM(x,y) = [L(x,y)]^{\alpha}\cdot[C(x,y)]^{\beta}\cdot[S(x,y)]^{\gamma} \tag{12}$$

where $\alpha > 0$, $\beta > 0$ and $\gamma > 0$ denote the weight values of each metric, which are generally equal weights.

$SSIM \in [0,1]$, the larger the SSIM value, the smaller the image distortion and closer to the original image it is. In practical applications, the image can be chunked using sliding windows so that the total number of chunks is N. Considering the influence of the window shape on the chunks, Gaussian weighting is used to compute the mean, variance and covariance of each window and then the structural similarity of the corresponding chunks is computed as the SSIM and, finally, the mean value is used as the structural similarity measure of the two images, i.e., the average SSIM.

4.3. Hash Coding Experiment

The experimental platform was an ubuntu system with 32 G of RAM, a GeForce RTX 3080Ti graphics card with 12 G of video memory and a 12th Gen Intel@CoreTM i7-12700KF × 20 processor. The number of network training iterations for Instant-ngp, NeuS and the method in this paper were 100,000, 50,000 and 50,000, respectively.

4.3.1. Qualitative Experimental Analysis

This paper employs progressive multi-resolution hash coding and primarily focuses on comparing and analyzing the results of two methods, Instant-ngp and NeuS. Instant-ngp utilizes multi-resolution hash coding, while NeuS employs frequency coding in NeRF. Figure 9 illustrates the comparison of the rendering results for the three algorithms on DTU15, DTU24 and DTU40, respectively.

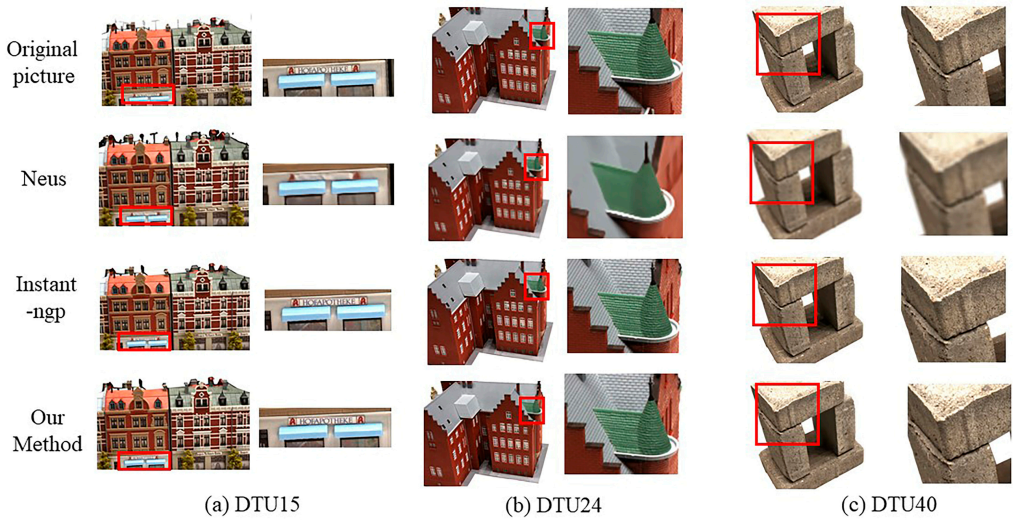

Figure 9. Rendering results of different methods. (**a**) Shows the DTU dataset scene 15 rendering results; (**b**) shows scene 24; (**c**) shows scene 40.

As a whole, NeuS has the most iterations, but has the worst rendering quality and cannot render the image clearly; both Instant-ngp and this paper's method can synthesize the viewpoints better and the image obtained via this paper's method is clearer in comparison between the two. In the DTU15 dataset, the method proposed in this paper is clearer and more realistic than the Instant-ngp method, particularly evident in the billboard letters shown in Figure 9a, which is closer to the original image. In the roof surface part of the DTU24 dataset, the results of this paper's method are clearer than the Instant-ngp texture structure, more granular and three-dimensional. In the DTU40 dataset, there is no significant difference between the results of Instant-ngp and this paper's method, but it is clearer than NeuS.

4.3.2. Quantitative Experimental Analysis

This subsection evaluates Instant-ngp, NeuS and the method of this paper using two metrics, the PSNR and SSIM. After the network is trained to a certain extent, this paper randomly selects a number of images from the image dataset to be used for testing and obtains the corresponding rendered images. Then, the PSNR value and SSIM value between the rendered image and the original image are calculated and the average is taken as the final evaluation value. Table 3 shows the comparison of the PSNR value of the rendered images of the three methods, and six rendered images and the original image are randomly selected from each method for comparison. It can be observed that for the rendered images of the three datasets, the NeuS method exhibits the lowest PSNR values, which are 20.9014, 22.0228 and 27.8526, indicating a lower proximity to the original image and a large amount of blurring. In contrast, the average PSNR values of the method proposed in this paper are 22.2156, 24.3423 and 28.7186, respectively. These values are notably higher than those achieved via the Instant-ngp method, exceeding Instant-ngp's PSNR values by more than 25%. This suggests that the application of low-conflict progressive multi-resolution hash coding can enhance the detail expression capability of the neural network, leading to rendered images that, consequently, are clearer and more closely resemble the original image.

Table 3. PSNR evaluation table of the rendered image results of the three methods.

		Instant-ngp	**NeuS**	**Ours**
DTU15	1	21.5906	17.8316	24.5007
	2	22.8636	16.7975	21.3661
	3	20.4145	16.8967	23.5825
	4	20.2797	18.2014	21.3009
	5	19.2271	16.1140	20.8408
	6	21.0331	18.9674	21.7025
	Average	20.9014	17.4681	22.2156
DTU24	1	23.8592	19.6505	24.0335
	2	19.9375	19.8496	21.9429
	3	23.5673	21.7333	24.5284
	4	25.3783	17.9581	29.2147
	5	21.3128	18.5247	22.9470
	6	18.0817	18.3397	23.3875
	Average	22.0228	19.3427	24.3423
DTU40	1	26.8330	21.1750	29.2166
	2	26.9306	20.4683	29.2910
	3	27.3707	21.6330	28.7746
	4	27.7076	19.8074	28.3549
	5	28.7993	19.5547	28.1579
	6	29.4745	21.3349	28.5163
	Average	27.8526	20.6622	28.7186

Table 4 shows the comparison of the SSIM values of the rendered images of the three different methods. From the table, it can be seen that the NeuS method shows a relatively low image structure similarity, with values around 0.7, which suggests that the images produced using NeuS are not adequately trained, leading to an incomplete expression of detailed structures. However, the method discussed in this paper exhibits the highest structural similarity value for the rendered images. Following closely is Instant-ngp and both these methods achieve SSIM values generally in the range of 0.9, which is significantly higher compared to NeuS. This comparison further demonstrates the effectiveness of multi-resolution hash coding in the fine-grained representation of structures.

Table 4. Evaluation table of SSIM values of rendered image results for the three methods.

		Instant-ngp	NeuS	Ours
DTU15	1	0.8540	0.7951	0.8883
	2	0.8975	0.5711	0.9267
	3	0.9107	0.6188	0.9142
	4	0.8301	0.7983	0.8395
	5	0.9002	0.9002	0.9076
	6	0.8497	0.8497	0.8666
	Average	0.8450	0.6953	0.8809
DTU24	1	0.9313	0.7350	0.8795
	2	0.6199	0.7510	0.9290
	3	0.9090	0.7978	0.9299
	4	0.9164	0.6847	0.9471
	5	0.8806	0.7028	0.9176
	6	0.8055	0.8079	0.7687
	Average	0.8438	0.7465	0.8953
DTU40	1	0.9193	0.7186	0.9246
	2	0.9210	0.6985	0.9228
	3	0.9179	0.6346	0.9020
	4	0.9119	0.7381	0.9193
	5	0.9041	0.7309	0.9324
	6	0.9025	0.7215	0.9437
	Average	0.9128	0.7070	0.9275

Table 5 shows the training efficiency comparison between the NeuS method represented by frequency position coding and Instant-ngp represented by multi-resolution hash coding. It is obvious from the table that multi-resolution hash coding has an absolute advantage in time and Instant-ngp is almost 50 times faster than NeuS. For the rendered images obtained via different methods, NeuS needs at least 8 h to obtain the corresponding rendering results, but the rendered image has a large gap with the original image and the clarity is not high, while Instant-ngp only needs about 10 min to obtain the rendered image with relatively good quality.

Table 5. Evaluation table of training time for the three methods.

	Ours Method/min	Instant-ngp/min	NeuS/min
DTU15	10.1	10	497
DTU24	10.3	10	501
DTU40	10.2	10	494

The method in this paper is based on multi-resolution hash coding and the training time is similar to Instant-ngp for the same number of iterations. The training efficiency is also significantly improved compared to the NeuS method.

4.4. Depth-Supervised Ablation Experiments on Ancient Buildings

The Instant-ngp, NeuS and Colmap methods are compared in this section of experiments. Among them, the number of NeuS iterations is 100,000 times and the number of

Instant-ngp and the method in this paper is 50,000 times. The experimental platform is the ubuntu system with 32 G of RAM, GeForce RTX 3080Ti with 12 G of video memory and 12th Gen Intel@CoreTM i7-12700KF × 20 processor.

4.4.1. Qualitative Experimental Analysis

The qualitative experiment is divided into two parts, a comparison of the rendering quality of the methods and a comparison of the reconstruction models between the methods. (1) Rendering quality comparison. The three columns in Figure 10, respectively, show the rendered images and local magnification effects of NeuS, Instant-ngp and the method presented in this paper. As a whole, NeuS can only render the general structure and outline of the model and cannot capture the detail information, which is due to the insufficient network expression of NeuS and the need for a longer training time; Instant-ngp and the method in this paper have better rendering results and both of them have the ability to express detail.

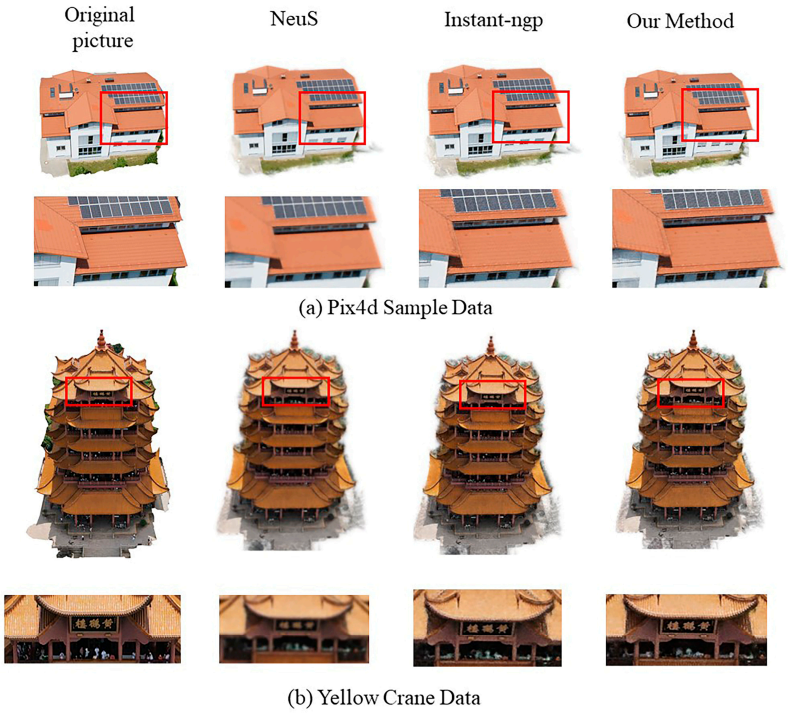

Figure 10. Comparison of rendering results of different methods. (**a**) Shows Pix4d sample data rendering results; (**b**) shows the Yellow Crane Tower data rendering results.

For the Pix4d sample data, the rendering result of NeuS can only vaguely express the shape and appearance of the building and fails to adequately render the detailed structure, such as the tile structure on the roof, three rows of solar panels, etc. Instant-ngp and the method described in this paper are both capable of quickly rendering the detailed structure of the building in a short time. However, the method presented in this paper outperforms Instant-ngp by producing a clearer rendering and more pronounced texture, resulting in a rendered image with enhanced clarity and a more distinct structural representation.

For the Yellow Crane Tower data, the difference in the rendering quality between the three different methods is even more obvious. From the perspective of the plaque of the Yellow Crane Tower, NeuS does not render the shape and content of the plaque because of insufficient training and the complexity of the structure of the Yellow Crane Tower itself; Instant-ngp and this paper's method can directly render the shape of the plaque and the three words "Yellow Crane Tower" and the two methods have a significant improvement in rendering quality compared with NeuS. Both of them have a significantly improved rendering quality compared with NeuS. Compared with Instant-ngp, this paper shows that

under the same resolution and the same number of training times, the method in this paper renders the "Yellow Crane Tower" with a higher clarity. Similarly, the image obtained via this method is more detailed and can significantly represent the arrangement of the tiles. (2) Reconstructing geometric contrasts. This paper proposes two geometric optimization methods: one is TSDF optimization and the other is the introduction of a depth supervision method based on TSDF optimization. This paper compares the Instant-ngp, NeuS and Colmap methods and analyzes the differences between the reconstruction models of each method.

Figure 11 shows the comparison of the reconstructed models of the Instant-ngp, NeuS, TSDF and Colmap methods. The geometric reconstruction quality of Instant-ngp is lower and cannot reconstruct the surface well; NeuS and the TSDF method in this paper can reconstruct the closed watertight model, but the surface of the TSDF optimization method in this paper is flatter and the reconstruction effect is slightly better.

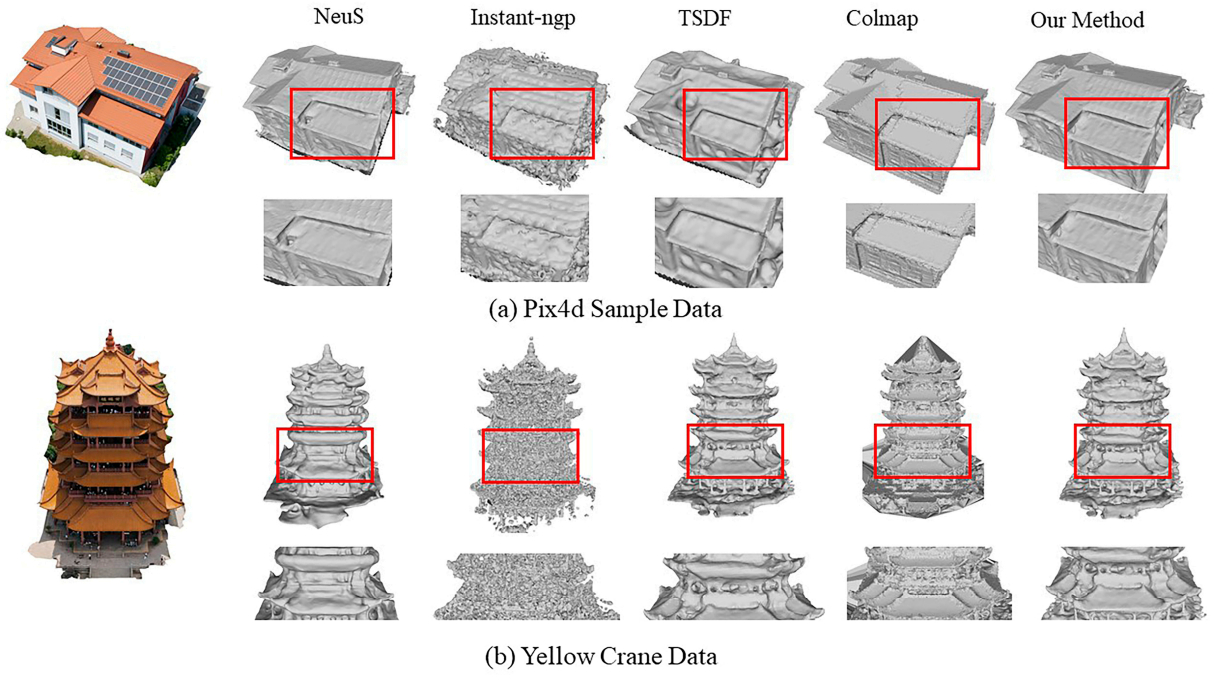

Figure 11. Comparison of TSDF optimization and reconstruction effect of each method. (**a**) Shows the reconstruction results for Pix4d sample data; (**b**) shows the reconstruction results for the Yellow Crane Tower data.

As shown in Figure 11, the Instant-ngp method results in a relatively sparse and fragmented reconstructed model for both the Pix4d sample data and the Yellow Crane Building data, failing to form a satisfactory surface model. While the NeuS method is capable of reconstructing the surface, it falls short in adequately expressing the geometric structure of the building over a certain period, leading to structural errors or imperfections in some areas, such as sunken roofs and uneven solar panels, etc. The TSDF method presented in this paper offers a more comprehensive reconstruction than both Instant-ngp and NeuS, particularly for buildings with simpler structures like those in Pix4d. For complex structures, such as the Yellow Crane Tower, the results are superior to other methods, but the visualization still does not meet the criteria for high precision.

Figure 11 shows the reconstruction model and local method effects of the TSDF method, Colmap method and the addition of the depth supervision method in this paper. For the complex structure of the Yellow Crane Tower data, the surface refinement achieved via the TSDF method is inadequate. However, the reconstruction quality significantly improves after adding the depth supervision on the basis of the TSDF optimization method. The eave edges of the Yellow Crane Tower exhibit a fine and even structure, with sharp protruding

edges and a flat, smooth eave surface. Compared with the Colmap reconstruction model, the surface of the model of this paper's method is smooth, avoiding the problem of surface noise and the detailed parts are also more prominent, such as the corridors, columns and other structures of the Yellow Crane Tower in the local zoomed-in image.

For the Pix4d building, the model after adding depth supervision can show the staggered feeling of the roof tile structure. This effect is attributed to a portion of the sparse point cloud on the roof, which constrains the geometric representation in the neural network. However, the solar panels appear uneven due to the intense light reflection on their surfaces, leading to deviations in the point cloud position and thus the unevenness of the reconstructed surface. The surface of the model of the Colmap method is too smooth and many structures are not fully expressed, such as the eaves of the tiles and their appendage structures, etc.

As shown in Figure 12, for the complex Huayan Temple data, using the SDF method did not achieve sufficient surface refinement. Adding depth supervision to the TSDF method significantly improved the reconstruction, resulting in finely detailed roof edges, sharp and prominent edge parts and a smooth eave surface. Compared to Colmap and NeuS, our method produced a model with a smoother surface, avoiding noise issues and more pronounced details.

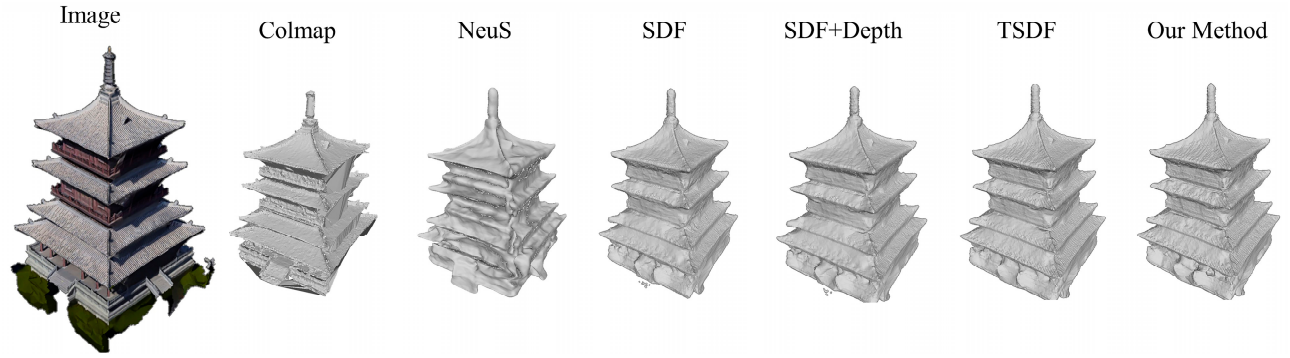

Figure 12. Comparison of the reconstruction effect. The SDF, SDF + Depth supervision, TSDF and the method in this paper are the results of ablation experiments; Colmap and NeuS methods are the results of comparison experiments.

4.4.2. Quantitative Experimental Analysis

This part of the quantitative analysis focuses on the quality analysis of the rendered images and the overall modeling efficiency analysis. The quality of the rendered image represents the expressive ability of the neural network and, to a certain extent, it can also indicate the geometric effect of the reconstruction. Table 6 shows the comparison of the PSNR indexes of the rendered images of Instant-ngp, NeuS and the method in this paper.

It can be seen from Table 6 that the NeuS method renders the worst image quality, with the average PSNR values for the two datasets being 21.2128 and 22.0479, respectively. Although NeuS demonstrates superior geometric expression capabilities, its training efficiency is suboptimal, resulting in inadequately rendered images over a short period. Compared to the rendering quality of the Instant-ngp method, the PSNR values of this paper's method are higher at 24.0229 and 25.5023.

Table 7 shows the comparison of the structural similarity index of the results of each method. From the data in the table, it can be seen that the rendered image of this paper's method has a higher degree of restoration and a clearer texture structure.

Table 6. PSNR evaluation of rendered images via different methods.

		Instant-ngp	NeuS	Our Method
Pix4d	1	25.4210	21.5347	25.8437
	2	24.4684	22.8885	25.4387
	3	24.8765	21.9155	25.8641
	4	24.7463	22.7518	26.5812
	5	24.7451	21.5997	24.8237
	6	24.1549	21.7302	24.4624
	Average	24.7353	22.0701	25.5023
Yellow Crane	1	22.0467	20.5486	23.9559
	2	22.3473	19.5063	24.2902
	3	21.7972	22.2307	23.7182
	4	21.6883	22.6365	23.7127
	5	22.3321	20.8539	24.3762
	6	22.0755	21.5008	24.0845
	Average	22.0479	21.2128	24.0229

Table 7. SSIM evaluation of different methods for rendering images.

		Instant-ngp	NeuS	Our Method
Pix4d	1	0.9469	0.9024	0.9470
	2	0.9517	0.9100	0.9518
	3	0.9437	0.9052	0.9535
	4	0.9465	0.9127	0.9559
	5	0.9404	0.9082	0.9493
	6	0.9395	0.9070	0.9563
	Average	0.9448	0.9076	0.9523
Yellow Crane	1	0.9276	0.8943	0.9406
	2	0.9278	0.8982	0.9397
	3	0.9255	0.8867	0.9394
	4	0.9264	0.8999	0.9389
	5	0.9288	0.8922	0.9427
	6	0.9274	0.8972	0.9416
	Average	0.9273	0.8948	0.9405

The average SSIM values of the two datasets of this paper's method are 0.9405 and 0.9523, respectively. In contrast, the rendered images of the NeuS method are more blurred and lack detail in parts, resulting in the lowest quality scores of 0.8948 and 0.9076. The SSIM values of the rendered images using the Instant-ngp method are 0.9273 and 0.9448, in which the structural similarity of the Pix4d data is quite close to that of the method proposed in this paper, because the structure and texture of the building are relatively simple, thus minimizing the differences. However, from the data of the Yellow Crane Building, we can see that this paper's method demonstrates superior rendering capabilities in more complex scenes.

Table 8 shows the comparison of the training time for NeuS, Instant-ngp, Colmap and the method in this paper.

Table 8. Training schedule for different methods.

Dataset	Instant-ngp/min	NeuS/min	Colmap/min	Our Method/min
Pix4d	9	504	41	16
Yellow Crane	10	517	44	16

The data presented in the table indicate that the NeuS method exhibits the longest reconstruction time, with training durations exceeding 8 h. Despite 100,000 iterations of learning, the neural network's expressive capability remains suboptimal. Followed by Colmap, the reconstruction time is 40 min to 50 min. The method in this paper, while marginally longer

in training duration compared to Instant-ngp, significantly enhances both the rendering quality and the geometric precision of the reconstruction. Consequently, the training time for the method delineated in this paper is considered within an acceptable threshold. The PSNR and SSIM in the ablation experiments are shown in Tables 9 and 10, respectively:

Table 9. PSNR evaluation of rendered images via different methods.

		NeuS	SDF	SDF + Depth	TSDF	Our Method
Huayan temple	1	20.0790	19.7807	18.6804	22.0038	21.4941
	2	21.4474	21.8201	19.8980	20.3897	23.0139
	3	20.1258	20.5362	21.3039	20.1366	21.4459
	4	19.4199	19.5288	20.8696	21.6823	22.2220
	5	19.3783	19.7992	18.2688	19.7121	20.1800
	6	18.2056	21.7851	20.9672	22.3610	21.0538
	Average	19.7760	20.5417	19.9980	21.0476	21.5683

Table 10. SSIM evaluation of different methods for rendering images.

		NeuS	SDF	SDF + Depth	TSDF	Our Method
Huayan Temple	1	0.8131	0.8327	0.8915	0.9105	0.9012
	2	0.8512	0.7858	0.8854	0.8654	0.8733
	3	0.8859	0.8069	0.7965	0.8421	0.9102
	4	0.7964	0.7934	0.8701	0.8369	0.9171
	5	0.7842	0.8610	0.8531	0.8554	0.8760
	6	0.8701	0.8714	0.8068	0.9024	0.8821
	Average	0.8335	0.8252	0.8506	0.8514	0.8933

The comparison of the training as well as reconstruction durations is shown in Table 11.

Table 11. Training schedule for different methods.

Dataset	Colmap/min	NeuS/min	SDF/min	SDF + Depth/min	TSDF/min	Our Method/min
Huayan Temple	35	311	23	24	22	23

Based on Tables 9 and 10, it can be observed that the average PSNR and SSIM metrics in this paper are superior to those of other experiments. However, the difference is not very significant, mainly due to issues with the aerial perspective and the presence of certain occlusions. The effect is not as good as surround shooting. Nevertheless, through ablation experiments using the method employed in this paper, it can be seen that the accuracy is still better than other algorithms.

From Table 11, it can be deduced that the NeuS method has the longest reconstruction time, exceeding 5 h of training time. After 100,000 iterations, the neural network's expressive capability is insufficient. Next is Colmap, with a reconstruction time of 35 min. When compared to the ablation experiments, the rendering quality of the method in this paper has significantly improved. This paper's method is on par with the SDF, SDF depth supervision, TSDF and it outperforms Colmap in terms of rendering speed.

5. Discussion

This study proposes a deep-learning-based method for the 3D reconstruction of ancient buildings from UAV-captured images. The method comprises three main steps: processing sampling points using multi-resolution hash coding, introducing the TSDF for threshold truncation during training and integrating depth information for supervised training. The innovations and characteristics of this research can be summarized as follows: (1) Progressive multi-resolution hash coding: This study focuses on target objects in large scenes, implementing centralized foreground position coding and adopting a "coarse-to-fine" progressive multi-resolution hash coding strategy. In the initial phase

of network training, high-resolution feature-encoding information is masked, retaining only the low-resolution hash feature encoding. As the training progresses, the masking of high-resolution feature-encoding information is gradually reduced, thereby optimizing feature expression. (2) Progressive TSDF-based depth supervision strategy: The Tanh function is used instead of the traditional piecewise distance function in the TSDF and the truncation distance of the TSDF is set to decrease progressively with the training time. Additionally, depth information from sparse point clouds generated by SfM is introduced as prior knowledge, enhancing the network's capability to express 3D geometric structures.

This paper utilizes a dataset of building images collected by UAVs conducting a comparative analysis with several classical neural radiance field technology-based methods to validate the practicality of the proposed algorithm. From Figure 9, it is evident that, compared to classical neural radiance field methods, the rendered images from this paper's method exhibit enhanced detail richness and superior texture clarity. In comparison with NeuS, the improved method in this paper not only ensures the quality of the rendered images but also significantly enhances the network training time. When contrasted with Instant-ngp, the rendered image details in this paper's method are more distinct. Furthermore, as seen in Figures 10 and 11, the 3D implicit reconstruction method in this paper demonstrates a higher accuracy compared to other methods. Finally, as shown in Table 8, compared to Instant-ngp and Colmap, this method is capable of reconstructing high-quality 3D models more swiftly compared to Instant-ngp and Colmap. Despite taking slightly longer than Instant-ngp for reconstruction, it is within an acceptable range.

The main reasons for the improvements in the rendered image quality, model geometric structure and network training efficiency of the proposed method are analyzed as follows:

(1) Reasons for improvement in rendered image quality: In this study, the images were preprocessed during the model training phase, employing a strategy of masking the background area to reduce the interference from background noise. Additionally, the adoption of progressive multi-resolution hash coding combined with occupying a three-dimensional grid fully exploits the high-resolution feature space in the hash table. Such a strategy allows the high-resolution grid to more accurately and intensively represent the detailed structure of the scene. This not only effectively resolves hash conflicts but also substantially improves the quality of the rendered images, leading to a more precise and detailed visual output.

(2) Reasons for improvement in model geometric structure: The integration of the TSDF values in this method ensures that the voxels in the occupied grid more closely adhere to the object's surface. This mechanism effectively filters out key points that significantly impact the reconstructed surface while eliminating points with little or no effect. Furthermore, the incorporation of depth supervision information enhances the model's depth representation capability, significantly improving the geometric structure of the generated model.

(3) Reasons for improvement in network training efficiency: At the initial stage of training, this study employed progressive multi-resolution hash coding, accelerating the ray sampling process by eliminating ineffective grids in the occupied grid. As the training progresses, the strategic application of the TSDF values for the threshold truncation continuously updates the occupancy of the grid, further speeding up the ray sampling efficiency. Moreover, integrating depth supervision information into the training regimen significantly hastens the model's convergence towards high-quality outcomes, ensuring the rapid attainment of superior results.

Therefore, the method proposed in this study is suitable for processing 3D ancient buildings data reconstruction, especially in scenarios requiring rapid and high-precision reconstruction. Not only can this method quickly reconstruct high-quality 3D models, but it also excels in maintaining the clarity of details and textures in rendered images.

6. Conclusions

This paper introduces a low-conflict multi-resolution hash feature location coding method that alleviates hash conflicts through background masking and progressive training. The initial step involves masking the background region in the scene, followed by a "from coarse to fine" approach where low-dimensional position encoding is applied prior to high-dimensional position encoding. This reduction in hash conflicts within high-dimensional features and the mitigation of aliasing in high-dimensional features not only enhances the quality of neural radiance field rendering but also ensures efficient network training, thereby facilitating subsequent geometric optimization. This paper tackles two main issues: (1) The development of a TSDF representation for surface reconstruction and model training supervision through the use of sparse point clouds. This approach serves to stabilize model training and enhance the model's depth representation, thereby significantly enhancing the overall model accuracy. (2) The introduction of an asymptotic training strategy based on multi-resolution hash grids. This strategy gradually refines the details of the reconstructed model, boosting model convergence and expediting the model training process.

Furthermore, this paper introduces an advanced geometric optimization technique for TSDF networks. The native NeRF relies on a biased volume rendering formulation that synthesizes colors solely through density and color, resulting in noisy reconstructed surfaces and low geometric accuracy. To address this, the SDF value is introduced as a weight for color synthesis instead of the original density value. The SDF is asymptotically truncated to obtain the TSDF using the SDF-MLP network, thereby enhancing the geometric constraints of the network and improving the geometric accuracy and detail expression in the reconstructed model. Additionally, a geometric optimization method is employed for deep-information supervised neural networks. Sparse reconstruction estimates the bitmap information from the input image and acquires a sparse point cloud for the depth information. In this approach, training rays are divided into depth rays and ordinary rays, both of which are input into the neural network simultaneously. The depth rays are supervised by depth information during training, enhancing the network's geometric expression capabilities. This method fully utilizes the depth information from sparse reconstruction, facilitating the accurate reconstruction of intricate architectural structures. Through experimental comparisons, this method outperforms the Colmap 3D reconstruction method in terms of reconstruction efficiency and quality.

This paper introduces an improved neural radiance field technique into the field of the 3D reconstruction of ancient architecture, capable of performing centralized multi-resolution hash coding for large-scale ancient architectural scenes captured by UAVs. This method effectively eliminates irrelevant background information, minimizing redundant data encoding, thus significantly enhancing the rendering quality of ancient architectural images. Additionally, this paper proposes a progressive TSDF depth supervision network, providing robust support for the geometric optimization of ancient buildings. Compared to traditional NeRF methods, which may suffer from surface noise and insufficient geometric accuracy in processing ancient buildings, our proposed approach can reconstruct the geometric structure and surface details of ancient architecture more precisely, greatly improving the accuracy in the preservation and restoration of cultural relics. Through this advanced 3D reconstruction technology, a new perspective and methodology are offered for the digital preservation and study of ancient buildings, aiding in the better conservation and heritage of these precious cultural assets.

The 3D reconstruction of ancient architecture using NeRF with depth map supervision is a method that utilizes neural networks and deep-learning techniques. Despite achieving certain effects, there are still limitations in data quality: the reconstruction quality heavily relies on the quality of the input data. If the resolution of the depth map data is low, contains a significant amount of noise or lacks diversity, it may result in the model being unable to accurately capture the details of the building. Subsequent measures, such as using UAVs and ground-level supplementary captures, can be employed to achieve a more refined 3D reconstruction.

Author Contributions: B.G. and Y.G. conceived and designed the whole procedure of this paper. S.J. and Y.G. contributed to the introduction, system model sections and manuscript writing. P.Z. performed and analyzed the computer simulation results and drew partial figures. Z.J. and D.L. reviewed and amended writing. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Research Program of Wuhan University-Huawei Geoinformatics Innovation Laboratory [grant No. K22-4201-011], the Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources [grant No. KF-2022-07-003] and the CRSRI Open Research Program [grant No. CKWV20231167/KF].

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Acknowledgments: The Research Program of Wuhan University-Huawei Geoinformatics Innovation Laboratory and The Project Supported by the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Klimkowska, A.; Cavazzi, S.; Leach, R.; Grebby, S. Detailed three-dimensional building façade reconstruction: A review on applications, data and technologies. *Remote Sens.* **2022**, *14*, 2579. [CrossRef]
2. Geiger, A.; Ziegler, J.; Stiller, C. Stereoscan: Dense 3d reconstruction in real-time. In Proceedings of the 2011 IEEE Intelligent Vehicles Symposium (IV), Baden, Germany, 5–9 June 2011; pp. 963–968.
3. Wang, T.; Zhao, L. Virtual reality-based digital restoration methods and applications for ancient buildings. *J. Math.* **2022**, *2022*, 2305463. [CrossRef]
4. Qu, Y.; Huang, J.; Zhang, X. Rapid 3D reconstruction for image sequence acquired from UAV camera. *Sensors* **2018**, *18*, 225. [CrossRef]
5. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **2021**, *65*, 99–106. [CrossRef]
6. Shang, L.; Wang, C. Three-Dimensional Reconstruction and Protection of Mining Heritage Based on Lidar Remote Sensing and Deep Learning. *Mob. Inf. Syst.* **2022**, *2022*, 2412394. [CrossRef]
7. Pepe, M.; Alfio, V.S.; Costantino, D.; Scaringi, D. Data for 3D reconstruction and point cloud classification using machine learning in cultural heritage environment. *Data Brief* **2022**, *42*, 108250. [CrossRef] [PubMed]
8. Schonberger, J.L.; Frahm, J.-M. Structure-from-motion revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4104–4113.
9. Pepe, M.; Alfio, V.S.; Costantino, D. UAV platforms and the SfM-MVS approach in the 3D surveys and modelling: A review in the cultural heritage field. *Appl. Sci.* **2022**, *12*, 12886. [CrossRef]
10. Pei, S.; Yang, R.; Liu, Y.; Xu, W.; Zhang, G. Research on 3D reconstruction technology of large-scale substation equipment based on NeRF. *IET Sci. Meas. Technol.* **2023**, *17*, 71–83. [CrossRef]
11. Lee, J.Y.; DeGol, J.; Zou, C.; Hoiem, D. Patchmatch-rl: Deep mvs with pixelwise depth, normal, and visibility. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11 October 2021; pp. 6158–6167.
12. Schönberger, J.L.; Price, T.; Sattler, T.; Frahm, J.-M.; Pollefeys, M. A vote-and-verify strategy for fast spatial verification in image retrieval. In Proceedings of the Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; Revised Selected Papers, Part I 13, 2017; pp. 321–337.
13. Dang, W.; Xiang, L.; Liu, S.; Yang, B.; Liu, M.; Yin, Z.; Yin, L.; Zheng, W. A Feature Matching Method based on the Convolutional Neural Network. *J. Imaging Sci. Technol.* **2023**, *67*, 030402. [CrossRef]
14. Cubes, M. A high resolution 3d surface construction algorithm/william e. *Lorensen Harvey E. Cline–SIG* **1987**, *87*, 76.
15. Tolstikhin, I.O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J. Mlp-mixer: An all-mlp architecture for vision. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24261–24272.
16. Yu, A.; Li, R.; Tancik, M.; Li, H.; Ng, R.; Kanazawa, A. Plenoctrees for real-time rendering of neural radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11 October 2021; pp. 5752–5761.
17. Müller, T.; Evans, A.; Schied, C.; Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph. (ToG)* **2022**, *41*, 102. [CrossRef]
18. Zhao, F.; Jiang, Y.; Yao, K.; Zhang, J.; Wang, L.; Dai, H.; Zhong, Y.; Zhang, Y.; Wu, M.; Xu, L. Human performance modeling and rendering via neural animated mesh. *ACM Trans. Graph. (TOG)* **2022**, *41*, 1–17. [CrossRef]
19. Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; Wang, W. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv* **2021**, arXiv:2106.10689.

20. Barron, J.T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; Srinivasan, P.P. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11 October 2021; pp. 5855–5864.
21. Li, Z.; Müller, T.; Evans, A.; Taylor, R.H.; Unberath, M.; Liu, M.-Y.; Lin, C.-H. Neuralangelo: High-Fidelity Neural Surface Reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 8456–8465.
22. Kerbl, B.; Kopanas, G.; Leimkühler, T.; Drettakis, G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* **2023**, *42*, 1–14. [CrossRef]
23. Condorelli, F.; Rinaudo, F.; Salvadore, F.; Tagliaventi, S. A comparison between 3D reconstruction using nerf neural networks and mvs algorithms on cultural heritage images. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2021**, *43*, 565–570. [CrossRef]
24. Lehtola, V.V.; Koeva, M.; Elberink, S.O.; Raposo, P.; Virtanen, J.-P.; Vahdatikhaki, F.; Borsci, S. Digital twin of a city: Review of technology serving city needs. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *114*, 102915. [CrossRef]
25. Gao, K.; Gao, Y.; He, H.; Lu, D.; Xu, L.; Li, J. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv* **2022**, arXiv:2210.00379.
26. Villanueva, A.J.; Marton, F.; Gobbetti, E. SSVDAGs: Symmetry-aware sparse voxel DAGs. In Proceedings of the 20th ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games, Redmond, WA, USA, 27–28 February 2016; pp. 7–14.
27. Verbin, D.; Hedman, P.; Mildenhall, B.; Zickler, T.; Barron, J.T.; Srinivasan, P.P. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5481–5490.
28. Newcombe, R.A.; Izadi, S.; Hilliges, O.; Molyneaux, D.; Kim, D.; Davison, A.J.; Kohi, P.; Shotton, J.; Hodges, S.; Fitzgibbon, A. Kinectfusion: Real-time dense surface mapping and tracking. In Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality, Basel, Switzerland, 26–29 October 2011; pp. 127–136.
29. Ma, L.; Li, X.; Liao, J.; Zhang, Q.; Wang, X.; Wang, J.; Sander, P.V. Deblur-nerf: Neural radiance fields from blurry images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12861–12870.
30. Zhang, K.; Riegler, G.; Snavely, N.; Koltun, V. Nerf++: Analyzing and improving neural radiance fields. *arXiv* **2020**, arXiv:2010.07492.
31. Barron, J.T.; Mildenhall, B.; Verbin, D.; Srinivasan, P.P.; Hedman, P. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5470–5479.
32. Oechsle, M.; Peng, S.; Geiger, A. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11 October 2021; pp. 5589–5599.
33. Fridovich-Keil, S.; Yu, A.; Tancik, M.; Chen, Q.; Recht, B.; Kanazawa, A. Plenoxels: Radiance fields without neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5501–5510.
34. Liu, L.; Gu, J.; Zaw Lin, K.; Chua, T.-S.; Theobalt, C. Neural sparse voxel fields. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 15651–15663.
35. Huang, X.; Alkhalifah, T. Efficient physics-informed neural networks using hash encoding. *arXiv* **2023**, arXiv:2302.13397. [CrossRef]
36. Xu, Q.; Xu, Z.; Philip, J.; Bi, S.; Shu, Z.; Sunkavalli, K.; Neumann, U. Point-nerf: Point-based neural radiance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5438–5448.
37. Zhang, J.; Yao, Y.; Quan, L. Learning signed distance field for multi-view surface reconstruction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11 October 2021; pp. 6525–6534.
38. Wei, Y.; Liu, S.; Rao, Y.; Zhao, W.; Lu, J.; Zhou, J. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11 October 2021; pp. 5610–5619.
39. Rother, C.; Kolmogorov, V.; Blake, A. "GrabCut" interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph. (TOG)* **2004**, *23*, 309–314. [CrossRef]
40. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]