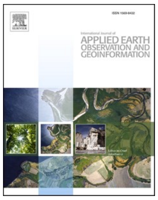
Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



NeRFOrtho: Orthographic Projection Images Generation based on Neural Radiance Fields

Dongdong Yue [a], Xinyi Liu [a,*], Yi Wan [a,*], Yongjun Zhang [a], Maoteng Zheng [b], Weiwei Fan [a], Jiachen Zhong [a]

[a] *School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, HB 430079, China*

[b] *National Engineering Research Center for Geographic Information System, China University of Geosciences (Wuhan), Wuhan, HB 430078 China*

ARTICLE INFO

Keywords:
Multi-view Images
Neural Radiation Fields
Orthographic Projection
True Orthophoto

ABSTRACT

The application value of orthographic projection images is substantial, especially in the field of remote sensing for True Digital Orthophoto Map (TDOM) generation. Existing methods for orthographic projection image generation primarily involve geometric correction or explicit projection of photogrammetric mesh models. However, the former suffers from projection differences and stitching lines, while the latter is plagued by poor model quality and high costs. This paper presents NeRFOrtho, a new method for generating orthographic projection images from neural radiance fields at arbitrary angles. By constructing Neural Radiance Fields from multi-view images with known viewpoints and positions, the projection method is altered to render orthographic projection images on a plane where projection rays are parallel to each other. In comparison to existing orthographic projection image generation methods, this approach produces orthographic projection images devoid of projection differences and distortions, while offering superior texture details and higher precision. We also show the applicative potential of the method when rendering TDOM and the texture of building façade.

1. Introduction

Generating high-quality orthographic projection images has been a major challenge in the field of remote sensing. Orthographic projection images can be applied not only in producing high-precision maps (Jauregui et al., 2002), but also bear profound significance in urban planning (Dornaika et al., 2016), land management (Szostak et al., 2014), building monitoring (Li et al., 2023; Qin et al., 2016), and environmental protection (Akbari et al., 2003). Orthographic projection image is a kind of data with high versatility, in which the True Digital Orthophoto Map (TDOM) can be regarded as a special case of orthographic projection image.

There are two main existing methods for generating orthographic projection images using multi-view image data, one is to geometrically correct and splice multiple images, which is mainly applied to the generation of TDOM. However, due to the influence of projection difference and stitching lines, the image generated by this method has the problem of perspective displacement or misalignment of stitching. Another method is to use multi-view images to construct an explicit 3D scene such as Mesh and project it to get the orthographic image. However, due to the large amount of data and discontinuous scene expression of the 3D explicit model, the orthographic image generated by this method has distortion problems (Qi et al., 2017). Traditional projection methods are subject to the limitations of viewing angle and geographical location, which limits their ability to generate images at multiple angles and locations, and restricts the diversity and practicality of remote sensing data.

In recent years, the introduction of the Neural Radiation Field (NeRF) has provided new possibilities for solving the above problems. NeRF is an implicit 3D scene modeling method based on neural networks, which can generate high-quality 3D reconstruction results from different viewpoints (Mildenhall et al., 2020). The core idea is to estimate the radiation intensity and color of each point in the 3D scene by training the neural network, making it possible to create new images from different viewpoints. The advantages of implicit expression of neural radiance fields include continuous expression, low cost, and low resource consumption (Zhang et al., 2020).

In this paper, we propose a neural radiation field rendering-based orthographic projection image generation method NeRFOrtho, which aims to achieve orthographic image generation at any angle, including

* Corresponding authors.
E-mail addresses: liuxy0319@whu.edu.cn (X. Liu), yi.wan@whu.edu.cn (Y. Wan).

https://doi.org/10.1016/j.jag.2025.104378
Received 25 October 2023; Received in revised form 30 October 2024; Accepted 11 January 2025
Available online 16 January 2025

TDOM as a special case. We discuss the technical details of NeRFOrtho in detail, including the training of the NeRF model and the projection process. The uniqueness of NeRFOrtho is that it can generate high-quality orthographic images that are not restricted by viewing angle and geographic location, thus extending the possibilities of remote sensing image generation.

Our experiments show that our Neural Radiation Fields rendering method NeRFOrtho for rendering orthographic images has excellent results and can achieve rendering of orthographic images at any views. Combined with the work of instant-ngp (Müller et al., 2022), NeRFOrtho can achieve real-time rendering, bringing a new possibility for orthographic image generation. Compared to traditional methods, NeRFOrtho also addresses the issue of self-obscuration caused by tall objects obstructing low objects in central projection (Xie and Zhou, 2010), as well as the problem of hierarchical occlusion between objects at different hierarchical levels (Wang et al., 2018). Fig. 1 shows these two occlusion problems. Moreover, we apply NeRFOrtho to two applications in the field of remote sensing: TDOM generation and texture mapping of building façades, achieving impressive results. By bringing NeRF into the field of orthographic projection image generation, we can better serve the diverse needs of various applications and provide new solutions for geographic information processing and urban planning.

The contributions of this work are outlined as follows:

1) This paper proposes a new method NeRFOrtho that uses the implicit three-dimensional expression of the Neural Radiance Fields to generate TDOM, the texture of building façade, and other orthographic images, providing a new data generation method for the field of remote sensing.
2) The proposed method generates high-quality orthographic projection image without stitching lines and expresses contour details better than traditional geometric correction and explicit modeling methods.
3) The proposed method can not only solve the self-occlusion problem of high objects occluding low objects caused by point projection but also directly solve the hierarchical occlusion problem between objects at different depth levels.

The second section provides a summary of the challenges involved in generating orthographic projection images and reviews the contributions made by previous authors in this field. In the third section, we introduce the techniques and methods of generating orthographic projection images from Neural Radiation Fields. The fourth section introduces the experiments and analyses the results of this study. Lastly, we summarize the work of this paper.

2. Related works

2.1. Traditional methods for generating orthographic projection images from multi-view images

There are two main ways to generate orthographic projection images using multi-view images. One is a 2D mosaic framework in which the multi-view images are corrected and stitched together. The other is an SfM framework that reconstructs a three-dimensional 3D explicit scene using multi-view images (Zhang et al., 2023). The former is mainly applied to the generation of TDOM in the field of remote sensing, which usually needs to refer to the digital surface model (DSM), orthorectify each pixel on the image by projection, and then stitch all the orthorectified images to obtain a large-scale TDOM. In the latter case, feature point extraction, feature matching, camera position estimation, 3D points cloud reconstruction, and other processes are performed on the multi-view image to recover the 3D explicit information of the scene. Then the orthographic projection images are generated by the digital differential correction model.

2.1.1. Orthographic projection images generation based on 2D mosaic framework

Generating orthographic projection images from 2D mosaic frames primarily addresses orthophoto production in remote sensing. This workflow generally includes three steps: ① Image Rectification: Utilizing a DSM to rectify central projection images, reducing projection errors from terrain undulations caused by tall objects. ② Occlusion Detection: Identifying shadowed areas influenced by substantial objects like buildings and trees. ③ Image Patching: Filling in shadows using information from other views. The rectification phase is crucial and can be achieved through polynomial (Tipdecho et al., 2001; Vassilopoulou et al., 2002), projective (Zhou et al., 2005), or differential rectification methods (Konecny, 1979). Each method relies on DSM accuracy, which may require additional modifications (Ebrahimikia and Hosseininaveh, 2022).

For occlusion detection, techniques based on Z-buffer, ray tracing, and polygonal methods have been proposed. Amhar et al. (1998) pioneered the Z-buffer algorithm for detecting occluded areas; Oliveira and

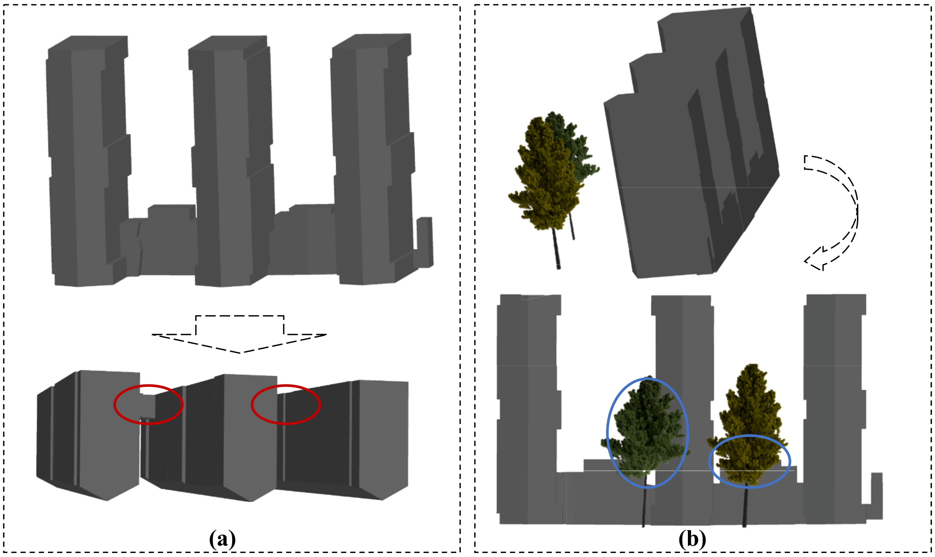


Fig. 1. Two common cases of occlusion in orthographic projection image generating. (a) shows the self-obscuration problem caused by heigh buildings; (b) shows the hierarchical occlusion problem caused by objects such as trees and buildings at different layers.

Galo (2013) calculated the radial height gradient of the DSM to find occluding objects; Marsetic (2021) utilized incoming angle ray tracing to realize satellite-based TDOM generation; De Oliveira et al. (2018) introduced the surface-gradient-based method (SGBM) for occlusion detection; and Kuzmin et al. (2004) used rectangular calculations to determine occluded regions. Shadow patching, viewed as a texture restoration process, has seen methods like line matching optimization (Wang et al., 2018) and texture synthesis from UAV imagery (Barazzetti et al., 2014). The generation of TDOM can be categorized into three scenarios: simple structures, densely built areas, and regions with extensive occlusions (Amhar et al., 1998). However, traditional methods face challenges, including stitching problems in large-scale scenes and the inherent limitations of rectification processes (Li et al., 2020).

2.1.2. Orthographic projection images generation based on SfM framework

The Structure from Motion (SfM) algorithm is crucial for 3D reconstruction from multi-view images (Beardsley et al., 1996; Fitzgibbon and Zisserman, 1998). However, achieving high-quality scene models efficiently remains a challenge. Mainstream SfM approaches include incremental SfM (Schonberger and Frahm, 2016; Wu et al., 2015), global SfM (Sweeney et al., 2015; Wilson and Snavely, 2014), hierarchical SfM (Farenzena et al., 2009; Toldo et al., 2015), hybrid SfM (Cui et al., 2017; Zhu et al., 2017), and deep learning-based SfM methods (Vijayanarasimhan et al., 2017; Yao et al., 2018). However, each method has its limitations. Incremental SfM methods can suffer from trajectory drift, while global SfM methods may be sensitive to matching quality. Hierarchical SfM methods offer efficiency but may compromise accuracy, and hybrid SfM methods often inherit disadvantages. Although deep learning-based SfM methods have made great progress, the reconstruction accuracy of explicit models is only comparable to that of traditional SfM methods and is not robust (Zhang et al., 2023).

Once a 3D explicit model is generated, the next step is projection. For smaller scenes, the resolution of orthographic images is defined, allowing pixel sampling to extract corresponding color or grayscale values. This enables the creation of various orthographic projection images, including TDOM and the texture of building façade. For larger scenes, especially in generating TDOM, stitching techniques are often necessary to create panoramic true projective images, which can be time-consuming and memory-intensive. Well-known 3D reconstruction software like Pix4D (Pix4D, 2023), Photoscan (Agisoft, 2023), and Smart3D (Smart3D, 2023) integrate SfM algorithms for stable 3D scene recovery. However, generating orthographic images from 3D models involves complex computations and can struggle with intricate structures. Implicit 3D models offer better continuity and detail but have seen limited exploration in orthographic image generation. Exploring how to utilize implicit models for high-quality orthographic images, such as true projective images and building elevations, presents a promising research direction.

2.2. Neural Radiance Fields

2.2.1. Development of Neural Radiance Fields

The concept of Neural Radiance Fields (NeRF) was first proposed by Mildenhall et al. (2020) and used for view synthesis. NeRF represents a form of implicit three-dimensional representation that offers advantages over explicit representations like point clouds, meshes, and voxels. NeRF is an implicit three-dimensional expression. Compared with point cloud, mesh, voxel, and other explicit expressions, it has the advantages of small memory usage and continuous scene expression. Even for complex scenes, NeRF can generate a continuous three-dimensional radiation field based on images of known viewpoints and positions, rendering new views.

As NeRF becomes popular, researchers begin to enhance and optimize various aspects of NeRF to further elevate its performance and efficiency. Mip-NeRF, introduced by Barron et al. (2021), replaces the sparse point sampling along a ray with cones to address blurring and aliasing effects caused by multi-scales greatly improving the rendering of images by NeRF at different scales. The method of Marching Cubes (Lorensen and Cline, 1998) transforms NeRF's implicit 3D representation into a mesh. Subsequently, Wang et al. (2023) introduced NeuS, utilizing a signed distance function (SDF) to represent object surfaces, enabling higher-quality multi-view mesh reconstruction. In pursuit of reduced training and rendering times, the instant-ngp (Müller et al., 2022) of NVIDIA employs CUDA acceleration and hashing techniques to accelerate the entire process to seconds, providing a new direction for real-time rendering tasks.

Furthermore, researchers have begun exploring the application of NeRF in various downstream tasks, such as object detection (Hu et al., 2023), semantic segmentation (Zhi et al., 2021), pose estimation (Yen-Chen et al., 2021), and urban scene representation (Xiangli et al., 2022), etc., to expand its application scope in real scenes. These applications make NeRF more diverse and practical in the fields of computer vision and computer graphics.

2.2.2. Ray sampling strategy

The vanilla NeRF (Mildenhall et al. 2020) sampling process involves two stages: a coarse and a fine process. The coarse stage utilizes uniform sampling (as illustrated in Fig. 2(a)), while the fine sampling stage refines these samples based on the coarse outputs. However, this approach proves inefficient due to the sparsity of 3D space, where large portions remain empty and do not contribute to the final color calculations. Consequently, methods like instant-ngp (Müller et al., 2022), Plenoxels (Fridovich-Keil et al. 2022), and KiloNeRF (Reiser et al., 2021) introduce spatial skipping sampling to bypass these empty regions, thus improving efficiency, as shown in Fig. 2(b). These approaches dynamically identify and exclude regions with minimal contribution to the final rendered color, significantly reducing computational overhead.

In addition, the Probability Density Function (PDF) sampling strategy is adopted by several methods, including Mip-NeRF 360 (Barron et al. 2022). This technique selectively allocates more sampling points to regions with a higher likelihood of contributing to the final color while allocating fewer points in empty areas (as shown in Fig. 2(c)). This probabilistic approach further enhances the efficiency and accuracy of the sampling process.

Considering these advancements, the goal of NeRFOrtho is to achieve parallel projection by improving the positional relationship of rays. As illustrated in Fig. 2(d), NeRFOrtho samples along parallel rays to align with the concept of orthographic projection. This shift in the projection paradigm allows each ray's sampling strategy to leverage any of the existing techniques, such as spatial skipping sampling or PDF-based sampling. By aligning the sampling direction with parallel rays, NeRFOrtho achieves greater geometric accuracy in generating orthographic images, making it ideal for applications like architectural reconstruction and orthophoto generation.

This parallel ray-based approach maintains the efficiency improvements introduced by existing sampling optimizations while adapting them to an orthographic projection context. As a result, the generated orthographic projection images retain high-quality photorealistic details without perspective displacement, accurately capturing structural information. Moreover, this integration of sampling strategies enables NeRFOrtho to handle complex and large-scale scenes efficiently, offering an effective solution for orthophoto generation from implicit 3D representations.

3. Methodology

The pipeline of NeRFOrtho is shown in Fig. 3. First, use the central projection image to train the multilayer perceptron (MLP) network to construct the neural radiation field. Then construct an orthographic projection plane based on the positions and views to modify the rendering method. Unlike the central projection, the rays on the orthographic projection plane are neatly arranged and perpendicular to

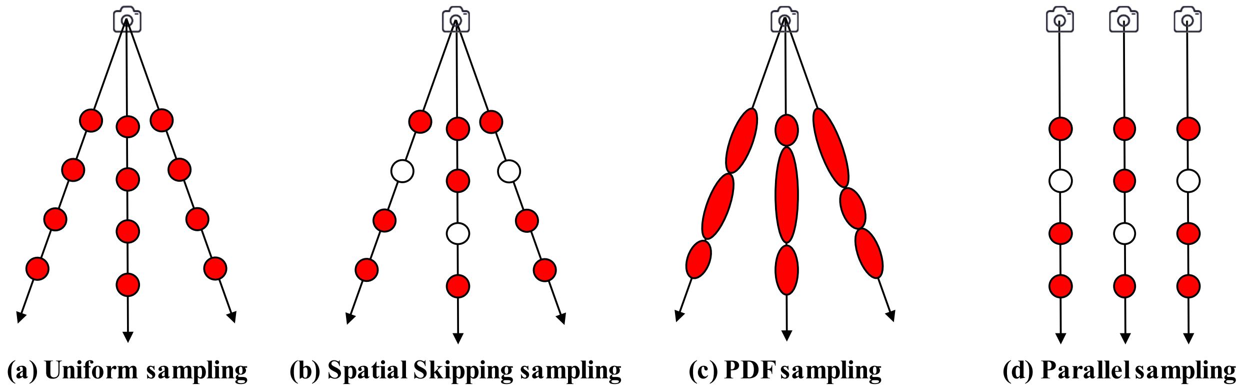


Fig. 2. Illustration of rays sampling. The red represents the spatial positions of ray sampling.

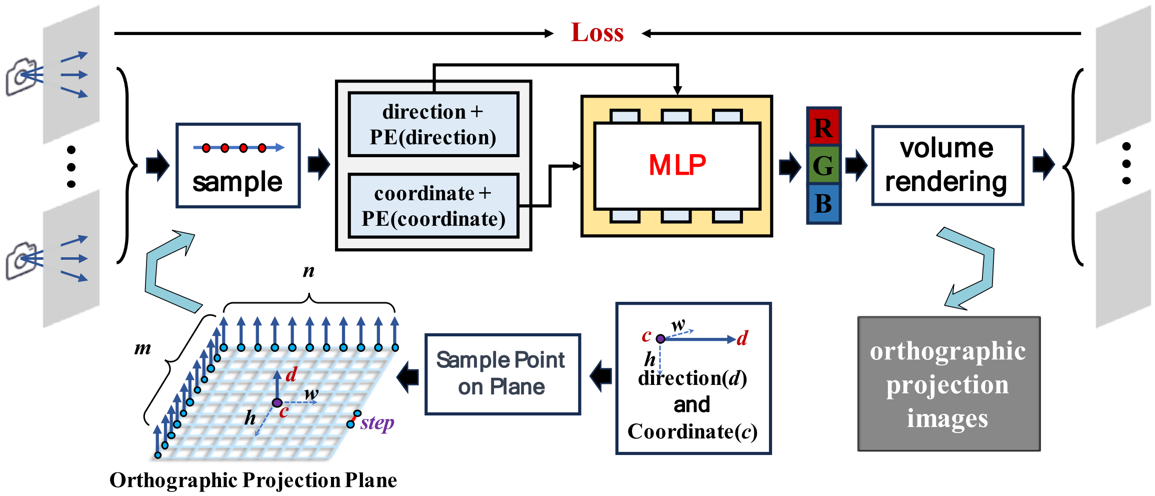

Loss
sample
direction + PE(direction)
coordinate + PE(coordinate)
MLP
R
G
B
volume rendering
orthographic projection images
Sample Point on Plane
c
w
d
h
direction(d) and Coordinate(c)
n
m
d
w
c
h
step
Orthographic Projection Plane

Fig. 3. The pipeline of NeRFOrtho. The upper portion represents the process of constructing Neural Radiance Fields, which can be replaced by variants of NeRF. The lower portion represents the process of constructing the orthographic projection plane.

the plane. Finally, the sampling points on the rays of the orthographic projection plane are fed into the previously trained MLP network, and the corresponding colors are output to render the orthographic projection images. More importantly, the use of NeRF in this paper can be vanilla NeRF or any of its variants, such as instant-ngp that can be rendered in real-time.

3.1. Construction of Neural radiation Fields

The process of NeRF is shown in the upper part of Fig. 3 which can be divided into four main parts. The first step involves sampling the central projected ray corresponding to the image captured by the camera, discretizing the continuous ray into a series of spatial 3D points, and calculating the coordinates and viewing angles of the sampled points based on the internal and external parameters of the camera. Each pixel corresponds to a ray $r(t) = \boldsymbol{o} + t\boldsymbol{d}$ emitted from the origin $\boldsymbol{o}$ and propagated along the direction $\boldsymbol{d}$, and t corresponds to the position of the ray.

The second step requires the positions and viewing angles of the 3D sampling points as inputs to the MLP network and outputs the color c and the volume density σ of the 3D sampling points. Position Encoding (PE) of the input $\boldsymbol{x}$ is required before passing into the MLP network:

$$PE(\boldsymbol{x}) = \left[sin\left(2^0\boldsymbol{x}\right), cos\left(2^0\boldsymbol{x}\right), \cdots, sin\left(2^{L-1}\boldsymbol{x}\right), cos\left(2^{L-1}\boldsymbol{x}\right)\right] \tag{1}$$

where L is a hyperparameter. The third step then integrates the colors of the 3D points output from the network to get the color corresponding to each pixel:

$$\widehat{C}(r) = \sum\nolimits_{i=1}^{N} T_i\alpha_i c_i,\ \text{with}\ T_i = exp\left(-\sum\nolimits_{j=1}^{i-1}\sigma_i(t_{i+1} - t_i)\right) \tag{2}$$

where r represents the ray; N is the number of sampled points; c_i is the color of the i-th sampled point; σ_i is the volume density of the i-th sampled point; t_i is the position of the i-th sampled point; t_{i+1} is the position of the $(i+1)$-th sampled point; α_i is the alpha value of the i-th sampled point, and is definded as follows:

$$\alpha_i = 1 - \exp(-\sigma_i(t_{i+1} - t_i)) \tag{3}$$

Finally, the computed color $\widehat{C}(r)$ is compared with the actual color C of that pixel to compute the loss L_c used in the MLP network. L_c is defined as follows:

$$L_c = \sum_{r\in R}\left[\|\widehat{C}(r) - C(r)\|_2^2\right] \tag{4}$$

where R represents all rays corresponding to the pixels.

To address the slow speed issue, the construction process of the Neural Radiance Fields can be replaced by instant-ngp. Instant-ngp uses multi-resolution hash encoding to improve the efficiency of the neural network and significantly reduce the training time.

3.2. Construction of the orthographic projection plane

In the training stage of NeRF, central projection is used to obtain projection rays and 3D points are sampled along these rays. The color corresponding to each pixel in the central projection image is calculated using Eq. (2), which corresponds to the color of each ray. In the inference stage of NeRFOrtho, the projection method is simply changed to orthographic projection, while still rendering colors along parallel rays according to Eq. (2), thus obtaining the color corresponding to each pixel in the orthographic projection image.

As shown in Fig. 4, due to the parallelism of the rays in orthographic projection, there is no unique camera origin in the imaging process, making it impossible to directly determine the emission point $\boldsymbol{o}$ and ray direction $\boldsymbol{d}$ for each ray based solely on the camera's intrinsic and extrinsic parameters. Therefore, solving how to obtain information for each parallel ray is a key challenge. To use the known conditions, we construct an orthographic projection plane to determine the information for each ray. As shown in Fig. 3, the orthographic projection plane consists of a series of spatial 3D points that are arranged neatly in the same plane. These points can be regarded as virtual camera positions. The directions of the rays on the orthographic projection plane are parallel to each other and perpendicular to the plane.

To construct an orthographic projection plane of size $m \times n$, several key elements are required, including the coordinates $\boldsymbol{c}$ of the camera located at the center of the plane, the direction $\boldsymbol{d}$ of the corresponding camera ray, the direction vectors $\boldsymbol{w}$ and $\boldsymbol{h}$ that define the two mutually perpendicular edges of the orthographic projection plane, and the interval *step* between virtual cameras. The *step* is a hyperparameter to adjust the resolution. Usually, the known conditions are the coordinates $\boldsymbol{c}_{camera}$ of the camera at the center of the plane in the camera coordinate system, the direction $\boldsymbol{d}_{camera}$ of the associated camera ray, and the extrinsic parameters for the camera: The transformation matrix from the world coordinate system to the camera coordinate system, including the rotation matrix $\boldsymbol{R}$ and the translation vector $\boldsymbol{T}$. In the camera coordinate system, the positional coordinates of the camera are the origin $[0,0,0]$, the direction is $[0,0,1]$, and the direction vectors on both sides are $[1,0,0]$ and $[0,1,0]$. Therefore, in the world coordinate system, the position $\boldsymbol{c}$ of the center point camera in the world coordinate system is:

$$c = \boldsymbol{R}^T \boldsymbol{c}_{camera} - \boldsymbol{R}^T T \tag{5}$$

The direction $\boldsymbol{d}$ of the camera at the center is as follows:

$$d = \boldsymbol{R}^T \boldsymbol{d}_{camera} \tag{6}$$

The direction vectors $\boldsymbol{w}$ and $\boldsymbol{h}$ of the two mutually perpendicular edges of the orthographic projection plane can be respectively computed as:

$$w = \boldsymbol{R}^T [1,0,0] \tag{7}$$

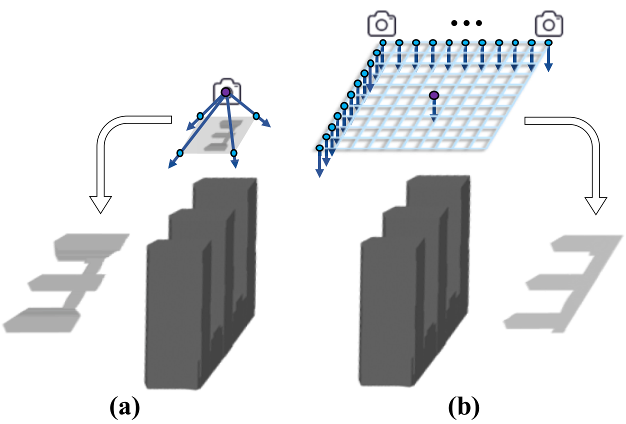


Fig. 4. The difference between central projection and orthographic projection. (a) is the imaging process of central projection, where projection rays converge at a single point. (b) is the imaging process of orthographic projection, where projection rays remain parallel to each other.

$$h = \boldsymbol{R}^T [0,1,0] \tag{8}$$

Therefore, the position c_{ij} of the virtual camera in the i-th row and j-th column on the orthographic projection plane can be computed as:

$$\begin{aligned} \boldsymbol{c}_{ij} = c + (i - 0.5{*}m){*}h{*}step + (j - 0.5{*}n){*}w{*}step, \\ \{i \in \mathbb{R}, j \in \mathbb{R}, 0 \le i < m, 0 \le j < n\} \end{aligned} \tag{9}$$

On the orthographic projection plane, the direction $\boldsymbol{d}_{ij}$ of the virtual camera in the i-th row and j-th column is consistent with the direction $\boldsymbol{d}$ of the central virtual camera. Then, the construction of the orthographic projection plane is completed. In fact, the formula $\boldsymbol{d} = \boldsymbol{R}^T \cdot \boldsymbol{d}_{camera} - \boldsymbol{R}^T \boldsymbol{T}$ is correct. However, the direction remains unaffected by translation and only depends on the rotation matrix. To maintain the scale consistency of the direction vector, the translation term $\boldsymbol{R}^T \boldsymbol{T}$ is excluded. This principle also applies to the calculations of other direction vectors.

As shown in Fig. 5, the constructed orthographic projection Plane. 1 is located on the left of the trees and buildings. Even though the challenge of self-occlusion by buildings has been addressed, the issue of hierarchical occlusion remains unresolved. Therefore, the depth of the orthographic projection plane is adjusted to Plane. 2 by adjusting the position of the orthographic projection plane to solve the hierarchical occlusion problem caused by other objects such as trees.

4. Experiments and results

This section describes the datasets used for the experiments and the details of the experimental setup. Firstly, we verify the feasibility of the Neural Radiance Fields to generate orthographic projection images in the publicly available synthetic dataset, then conduct experiments on our own synthetic data, the artificially constructed scene simulation datasets, and the real-world scene datasets. We also compare the generated TDOM and the texture of building façade with the products of other professional software, such as Pix4D, Photoscan, and Smart3D.

4.1. Dataset and corresponding experimental Settings

Synthetic Datasets: Since physically rendered synthetic data is not subject to environmental factors of the image, we first use the Lego synthetic data from the vanilla NeRF to explore the effect of Neural Radiance Fields in generating orthographic projection images. This

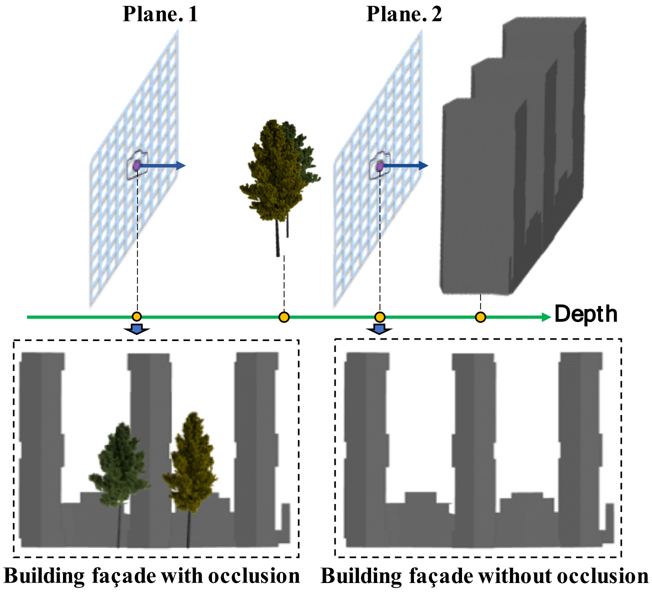

Plane. 1
Plane. 2
Depth
Building façade with occlusion
Building façade without occlusion

Fig. 5. Schematic diagram of using orthographic projection plane depth to address hierarchical occlusion issues.

dataset has 400 central projection images of size 800 × 800. In this dataset, we use the vanilla NeRF to construct the 3D scene, using 200 images as a training set and 200 for comparison with the generated orthographic projection images. To ensure the consistency of the view between the central projection images and our generated orthographic projection images, we set the coordinates of the center point of the orthographic projection plane and the viewpoint as the center point of the central projection images, as shown in Fig. 6. In addition, we make synthetic data of three single buildings around with trees for the generation of building elevations, each with 125 images and an image size of 800 × 800.

Artificially Constructed Scene Simulation Dataset: We make an urban environment with building models of different shapes, sizes, and textures. The different shapes of each building model bring great difficulties to the scene expression, especially the parapet on the top of the building. This dataset has a total of 54 artificially taken images, each of which is 800 × 800 in size. This dataset is used to generate TDOM and the texture of the building façade.

Real-World Scene Datasets: We use the UAV to take two different types of real datasets. One is a UAV aerial imagery dataset generated for the TDOM. The whole survey area includes complex scenes such as mountains, buildings, lakes, and roads. There are 57 images in total, with a size of 1993 × 1326, and all of them are UAV down-view shots. The other dataset includes a single-building UAV dataset generated for the texture of building façade. This data includes 73 images obtained using a UAV flying around a single building, and the size of each image is 1368 × 912.

The experimental setup employs Ubuntu 22.04 with NVIDIA RTX 3090 GPU and 128 GB RAM. The environment used is pytorch 1.13.1. The instant-ngp framework is used to build the neural radiation field, and the code implementation uses the Python version ngp_pl (ngp_pl, 2023) of instant-ngp. During the training process, the learning rate of training is 0.01, and iterates for 30 epochs.

4.2. Generation of TDOM

Fig. 7 shows the comparison results of the central and orthographic projection effects of real-world scene datasets within the Neural Radiation Fields. In the central projection image on the left, the high buildings can easily leak out the wall, and there are perspective displacement and self-occlusion problems compared with the orthographic image on the right. To further explore the effect of the TDOM generated by the neural radiative field, we compare the results with three major software: Pix4D, Photoscan, and Smart3D. Pix4D and Photoscan use traditional geometric correction and stitching techniques to generate TDOM, while Smart3D produces TDOM after explicit modeling. As shown in Fig. 8, in the red box, TDOM generated by Pix4D and Photoscan still exhibits the outer walls of certain buildings, displaying pronounced perspective displacement. In the yellow box, the generation of Smart3D lacks details at building edges and corners, a consequence of the explicit model's inadequate representation of fine details. In the blue box, TDOM generated through geometric correction methods shows stitching lines, particularly noticeable in the case of Pix4D software.

To explore the expressive ability of TDOM generated by Neural Radiative Fields to represent details, we explicitly model scenes using Pix4D, Photoscan, and Smart3D software and compare them to our method. All experiments are conducted on simulated data to reduce the cost and time of explicit modeling. Fig. 9 shows the comparison of the TDOM generated by the Neural Radiative Fields and the three software in the artificially constructed scene simulation dataset. Pix4D and Photoscan exhibit poor model quality in this dataset, with pronounced deformation and significant blurring along the edges of buildings. Although the overall quality of the TDOM generated by Smart3D is much better than that of Pix4D and Photoscan, when the building is closed, it is easy to have adhesion and deformation compared with the TDOM generated by the neural radiation field. Because the neural radiation field is a method that can achieve continuous implicit expression, the TDOM generated by the neural radiation field can better express the details of the building model.

4.3. Generation of texture of building Façade

To determine the orthographic projection planes for the texture building façade, we employ a method based on the surfaces of the building's 3D model. First, we align the projection planes with key

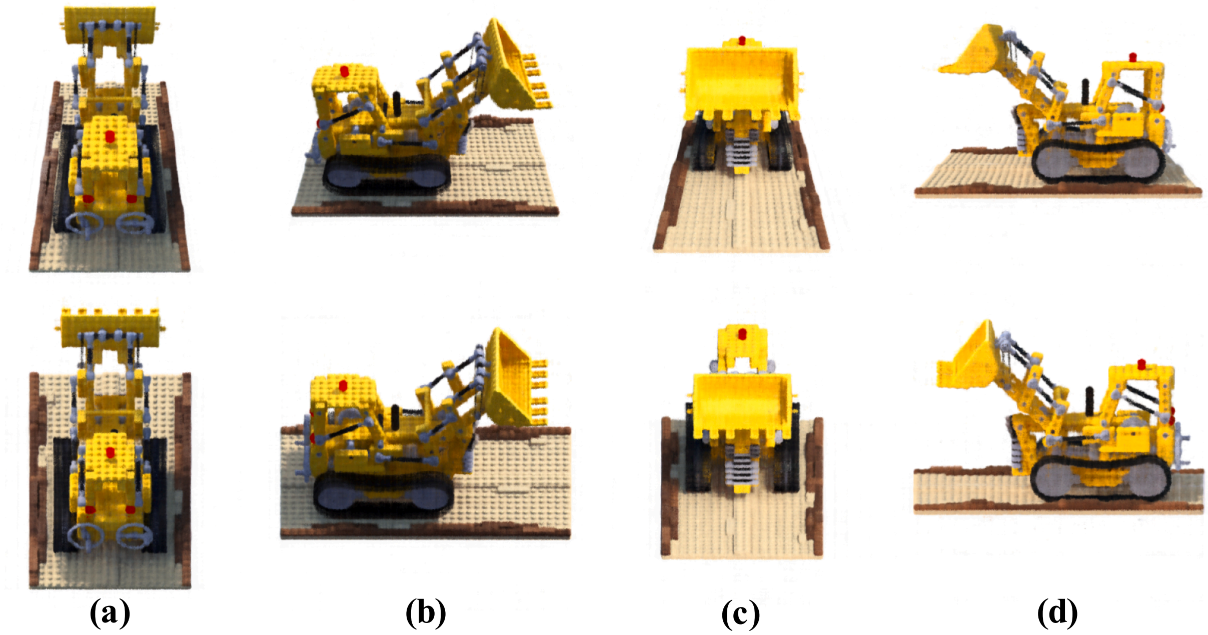
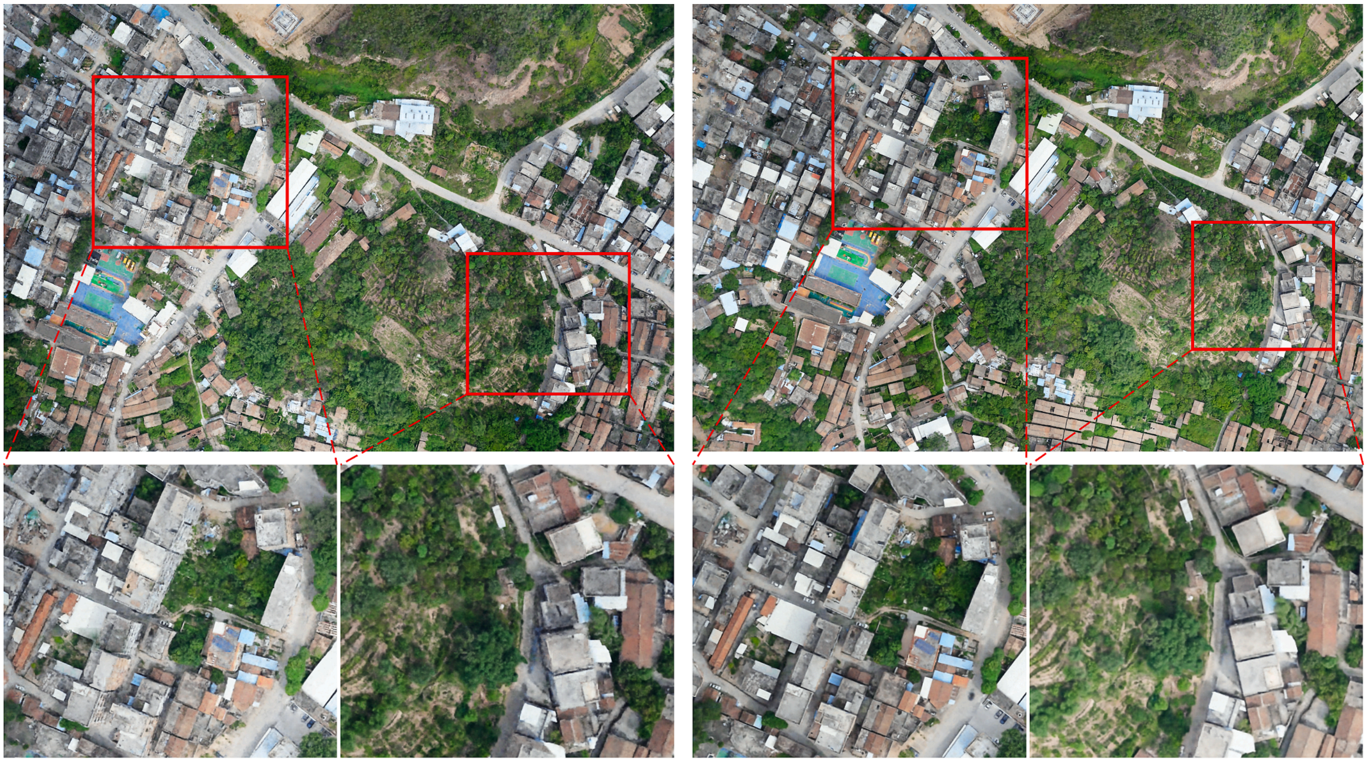


Fig. 6. Comparison of central and orthographic projection images of Lego dataset rendered in the Neural Radiation Fields. The first row is the central projection image and the second is the orthographic projection image.



(a) Images Rendered via Central Projection in Neural Radiation Fields

(b) Images Rendered via Parallel Projection in Neural Radiation Fields

Fig. 7. Comparison of the central and orthographic projection images of real UAV data rendered in the Neural Radiation Fields. The walls of many buildings can be seen in (a), while the orthographic projection image generated by the Neural Radiation Fields in (b) can only see the roofs of the buildings.

surfaces of the 3D model, establishing their initial position and direction. Then, NeRFOrtho renders depth information to refine the alignment, ensuring the planes are positioned at the correct depth to the building façades. This process generates high-quality orthographic projection images of the façades, which are used for texture mapping onto the 3D model. By combining the 3D model’s geometric information with NeRFOrtho’s depth rendering, this approach maintains geometric accuracy and effectively supports the texture mapping of complex façades.

Figs. 10, 11, and 12 show the comparison results of building façade texture generation by Pix4D, Photoscan, Smart3D, and NeRFOrtho in the synthetic building dataset, artificially constructed scene simulation dataset, and real-world scene dataset, respectively. Figs. 11 and 12 clearly show that Pix4D and Photoscan have challenges in effectively representing scenes and reconstructing models. Even with the synthetic building dataset, both tools struggle to produce coherent models from the multi-view images. In contrast, the experimental results demonstrate that building façade texture generated using Neural Radiance Fields offers notable advantages over explicit representation methods in several aspects:

1) Recovery of Obscured Building Walls: Neural Radiance Fields demonstrate a strong ability to recover building walls that are obscured by elements like trees. As shown in Fig. 10(c), the façade texture generated by Smart3D blends with the tree texture, making it challenging to separate them. In contrast, Neural Radiance Fields allow for refinement by adjusting the camera’s depth to resolve hierarchical occlusion. This is primarily because explicit mesh models struggle to accurately represent trees, while the implicit modeling of Neural Radiance Fields excels in maintaining spatial continuity.
2) Texture Synthesis in Regions with Limited Viewpoint Coverage: Neural Radiance Fields can also synthesize textures in areas with limited image coverage. As shown in Fig. 10(a) and (d), even at angles with poor data coverage, the texture of facades generated using the neural radiation field still has texture details instead of holes.
3) Minimal Structural Deformation: Building façade texture generated using Neural Radiance Fields exhibits minimal structural distortion. As shown in Fig. 10(b) and (f), the windows on the generated texture of façades maintain a regular rectangular shape with almost no distortion. Additionally, Figs. 11 and 12 show that the edges of the façades remain straight and undistorted, preserving the architectural structure.
4) More Accurate Detail Representation: Neural Radiance Fields capture intricate details with better accuracy. For example, in Fig. 10(b), the façade texture generated by NeRFOrtho displays a more refined and orderly texture on the balcony. Furthermore, as demonstrated in Fig. 12(c) and (d), the texture of façades generated using Neural Radiance Fields effectively preserves fine architectural features like needle-like elements.

NeRFOrtho can generate unobstructed orthographic projection images by adjusting the depth of the orthographic projection plane. As shown in Fig. 13, NeRFOrtho effectively resolves hierarchical occlusion issues, producing building façade textures that are not obstructed by trees or other objects.

4.4. Accuracy Evaluation

To effectively evaluate the accuracy of orthographic projection image generation using Neural Radiance Fields, we compare the lengths of building edges in the generated orthographic images with their actual lengths. We calculate the Root Mean Square Error (RMSE) for each building to quantify this accuracy. First, the lengths of building edges in

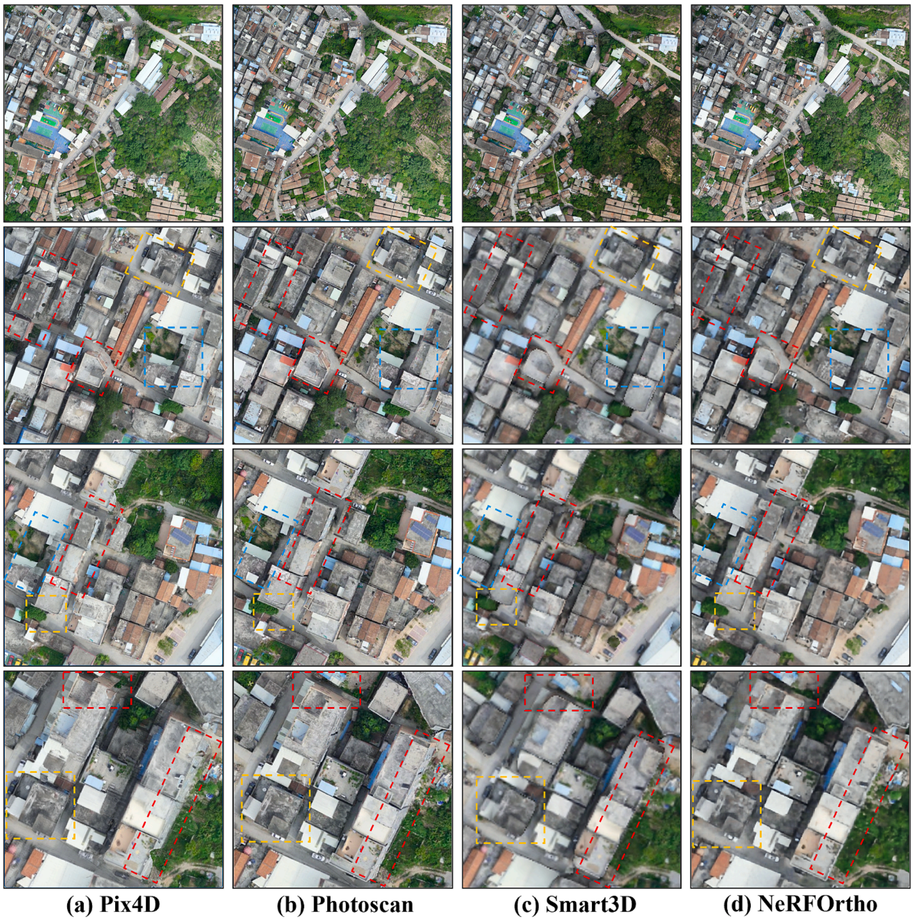


Fig. 8. Comparison of TDOM generated with software Pix4D, Photoscan, and Smart3D with NeRFOrtho in real UAV data.

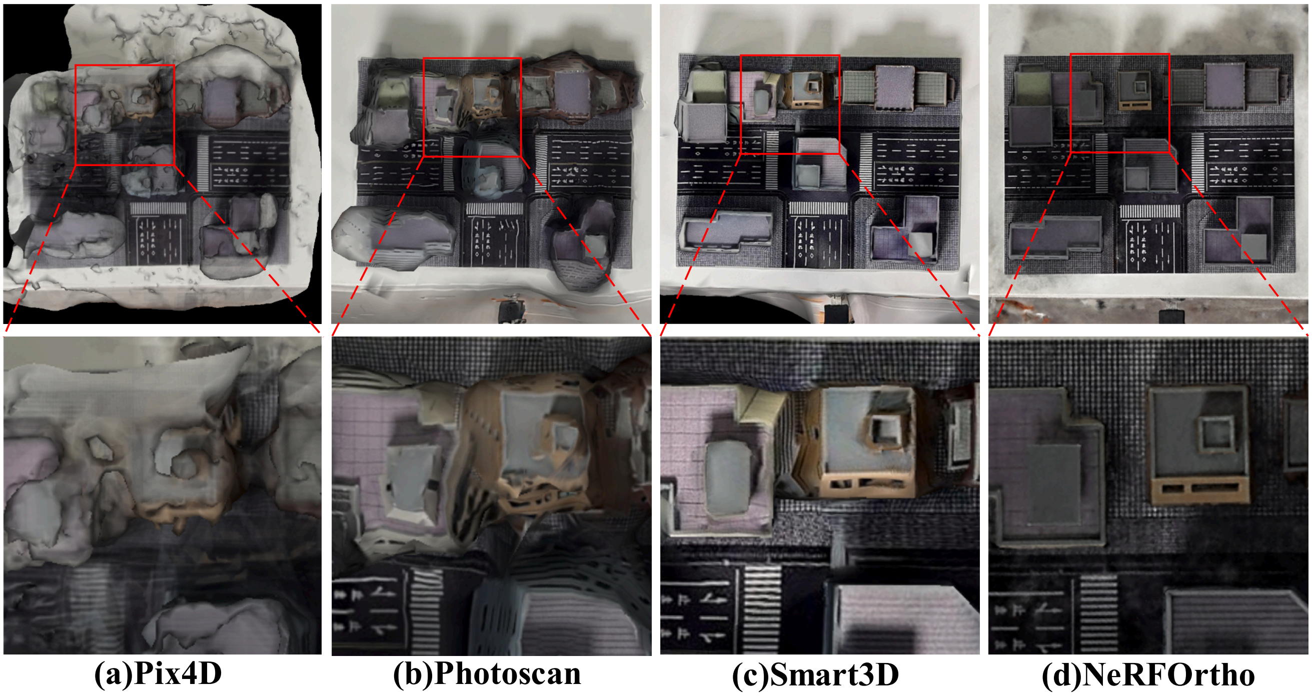


Fig. 9. Comparison of TDOM generated by Pix4D, Photoscan, Smart3D, and NeRFOrtho in artificially constructed scene simulation datasets.

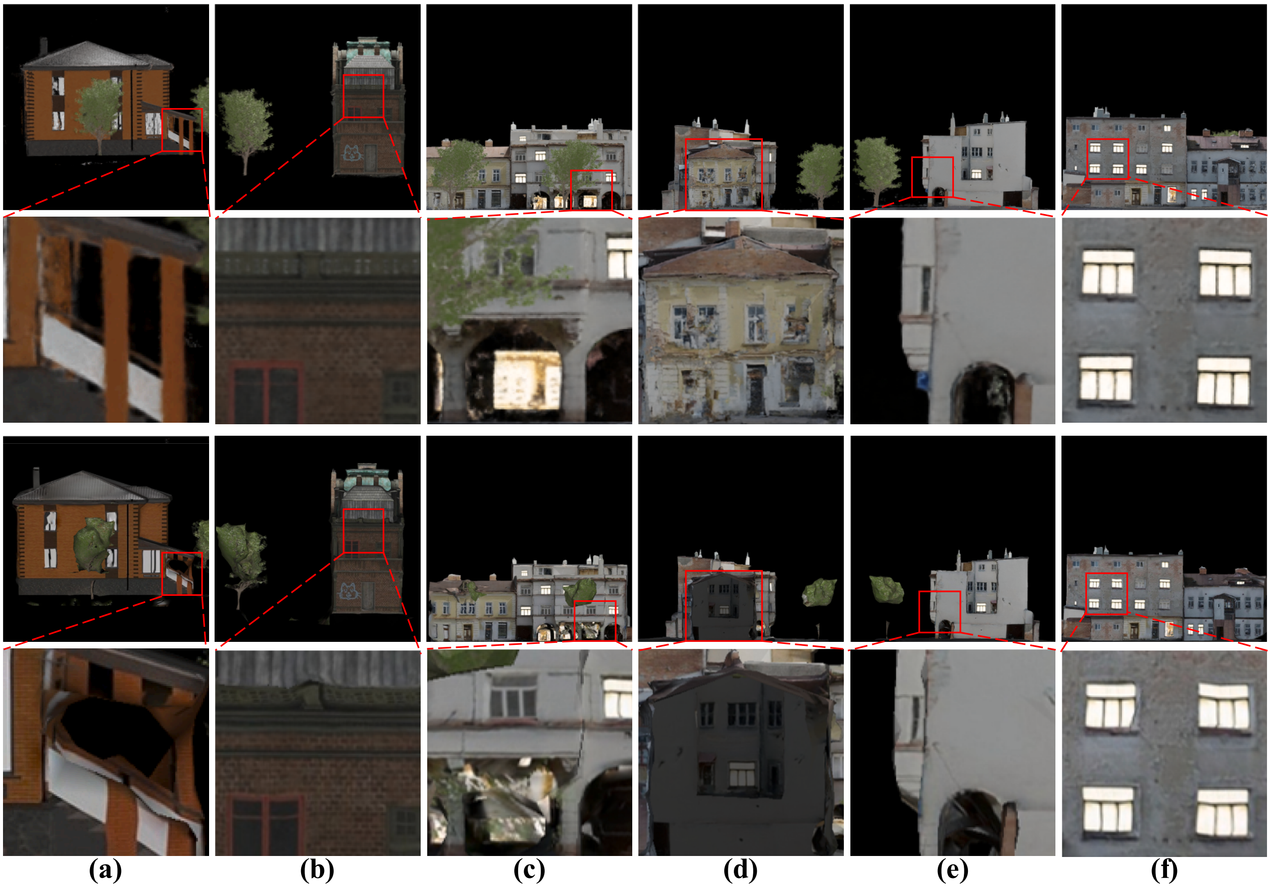


Fig. 10. Comparison of building façade texture generated by Smart3D and NeRFOrtho within the synthetic datasets. The first two rows are the results of NeRFOrtho, while the last two rows are the results of Smart3D.

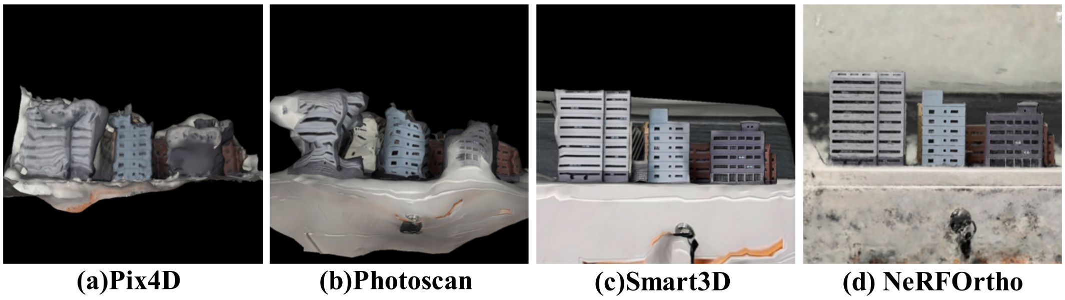


Fig. 11. Comparison of building façade texture generated by Pix4D, Photoscan, Smart3D, and NeRFOrtho in an artificially constructed scene simulation dataset.

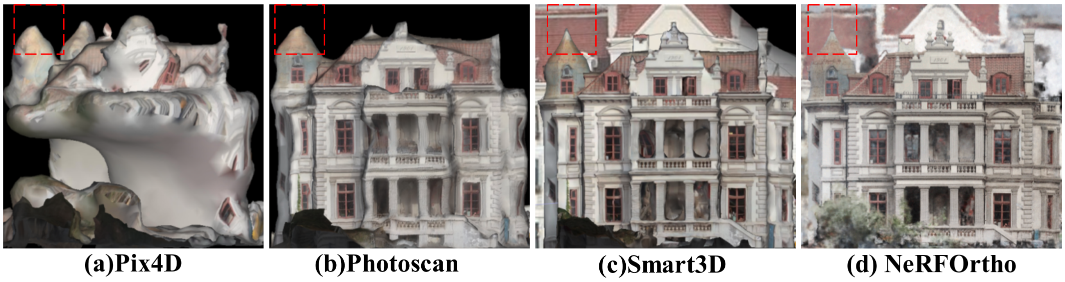


Fig. 12. Comparison of building façade texture generated by Pix4D, Photoscan, Smart3D, and NeRFOrtho in a real-world scene dataset.

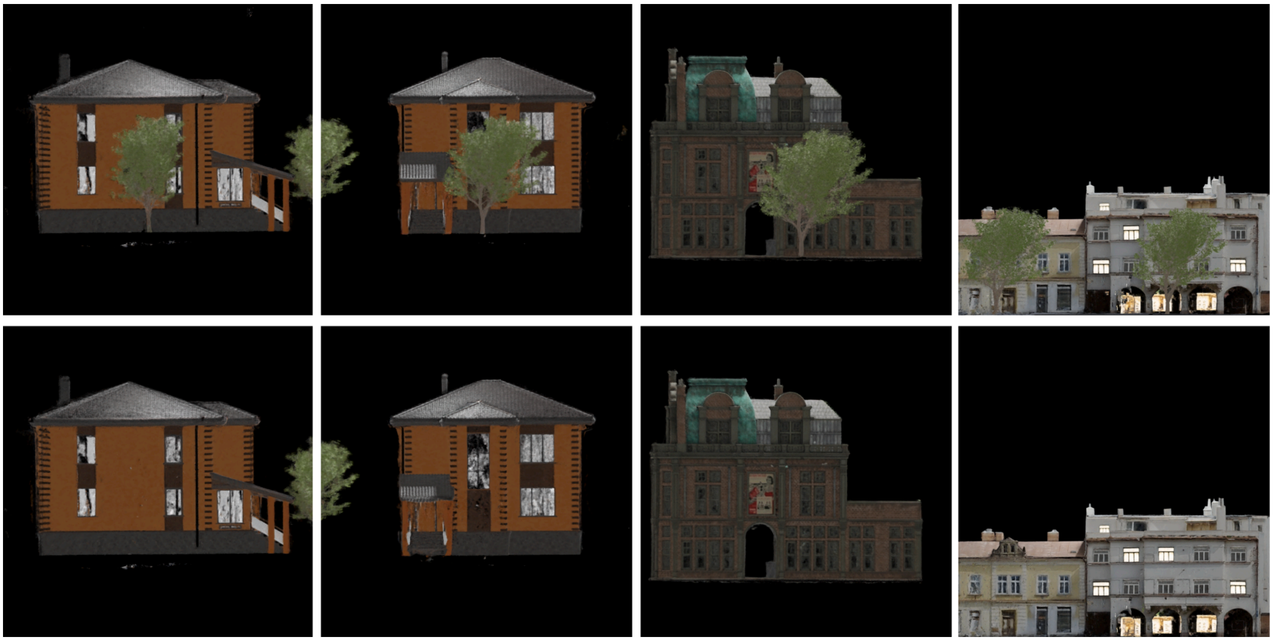


Fig. 13. Rendering building façades images without occlusion by adjusting the depth of the orthographic projection plane. The top row illustrates results obtained with the orthographic projection plane is placed farther away, showing building façade texture with hierarchical occlusions. The bottom row illustrates unobstructed building façade texture achieved by placing the orthographic projection plane closer.

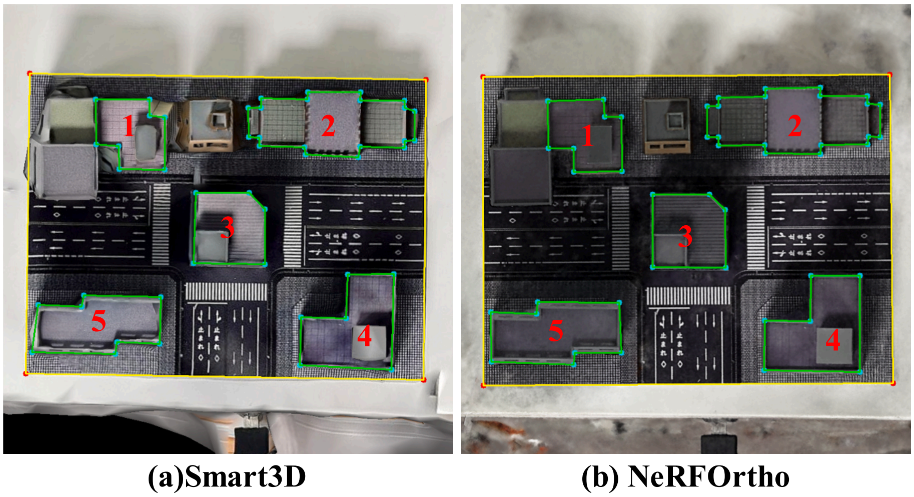
real space are measured and denoted as $l_1, \cdots, l_N$. Then, the corresponding lengths of edges in the orthographic projection images generated using Neural Radiance Fields and explicit modeling methods are measured and denoted as $l_1{'}, \cdots, l_N{'}$, where N represents the number of measured edges. If there is no deformation of the building in the generated orthographic projection images, then the length of each side of the building in the orthographic projection images should be equal to the actual side length of the building, otherwise, there is an error. The smaller the error, the higher the accuracy of the image.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(l_i - l_i^{'})^2} \tag{10}$$

Since only Smart3D works better among the methods for generating orthographic projection images using multi-view image explicit modeling, we compare the RMSE of Smart3D and NeRFOrtho. As shown in Fig. 14, we measure the edge lengths of five buildings located in the upper-left, upper-right, middle, lower-left, and lower-right corners of the image. The lengths collected for each building are no less than 5 sides, and the measurements are repeated 5 times to take the average value to offset the measurement error. Fig. 15 shows the RMSE of Smart3D and the proposed method in this paper, and it shows that the RMSE of the five buildings is better than those of the explicit modeling method of generating orthographic projection images, which is consistent with the visual results.

5. Experiments and results

This paper introduces a new method for generating orthographic projection images from multi-view images based on neural radiance fields, called NeRFOrtho. This method can render orthogonal projection images from any viewpoint and has important applications in generating DOM and building facade texture from remote sensing data. Compared



(a)Smart3D (b) NeRFOrtho

Fig. 14. Schematic representation of selected building edges in the generated orthoimage, including five regions: top-left, top-right, center, bottom-left, and bottom-right of the image.

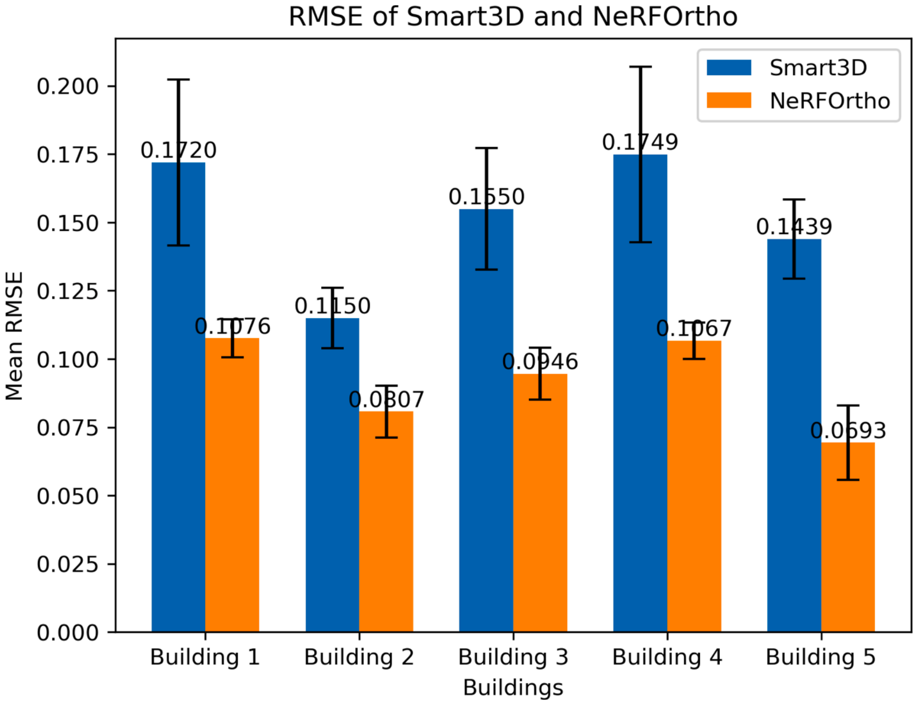


Fig. 15. RMSE comparison between Smart3D and NeRFOrtho. The smaller the RMSE value, the higher the accuracy.

with traditional methods for constructing 3D models from multi-view images, the proposed NeRFOrtho avoids the stitching line problem and preserves building details with excellent quality, performing well in complex scenes. NeRFOrtho effectively solves the self-occlusion of high buildings and the hierarchical occlusion of objects such as trees.

In addition, our research delves into the application of Neural Radiance Fields in remote sensing, potentially offering new directions for data representation paradigms in remote sensing imagery and geographic information. However, the construction of Neural Radiance Fields for large-scale, high-resolution scenes remains in its early stages. Therefore, future research will focus on developing methods to efficiently construct Neural Radiance Fields in large-scale scenes, exploring the integration of NeRFOrtho with advanced rendering techniques to enhance image quality, investigating applications in various remote sensing scenes, such as urban planning and environmental monitoring, and incorporating scene understanding to improve the semantic analysis of complex environments. Such explorations may pave the way for advancements in remote sensing applications and geographic information systems.

CRediT authorship contribution statement

Dongdong Yue: Writing – original draft, Conceptualization, Formal analysis, Investigation, Methodology. **Xinyi Liu:** Methodology, Writing – review & editing. **Yi Wan:** Methodology, Writing – review & editing. **Yongjun Zhang:** Supervision, Writing – review & editing. **Maoteng Zheng:** Writing – review & editing. **Weiwei Fan:** Writing – review & editing. **Jiachen Zhong:** Writing – review & editing.

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 paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 42201474) and Open Fund of National Engineering Research Center of Geographic Information System (Grant No. 2022KFJJ10).

Data availability

Data will be made available on request.

References

Agisoft, 2023. Agisoft PhotoScan: Agisoft PhotoScan [WWW Document]. URL https://www.agisoft.cn/.

Akbari, H., Shea Rose, L., Taha, H., 2003. Analyzing the land cover of an urban environment using high-resolution orthophotos. Landsc Urban Plan. 63, 1–14. https://doi.org/10.1016/S0169-2046(02)00165-2.

Amhar, F., Jansa, J., Ries, C., 1998. The generation of true orthophotos using a 3D building model in conjunction with a conventional DTM. ISPRS Arch. 32, 16–22.

Barazzetti, L., Brumana, R., Oreni, D., Previtali, M., Roncoroni, F., 2014. True-orthophoto generation from UAV images: Implementation of a combined photogrammetric and computer vision approach. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. 2, 57–63. https://doi.org/10.5194/isprsannals-II-5-57-2014.

Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P., 2021. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In: In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5835–5844. https://doi.org/10.1109/ICCV48922.2021.00580.

Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P., 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5470–5479.

Beardsley, P., Torr, P., Zisserman, A., 1996. 3D model acquisition from extended image sequences. In: Computer Vision – ECCV 1996. pp. 683–695. Doi: 10.1007/3-540-61123-1_181.

Cui, H., Gao, X., Shen, S., Hu, Z., 2017. HSfM: Hybrid Structure-from-Motion. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1212–1221. https://doi.org/10.1109/CVPR.2017.257.

De Oliveira, H.C., Poz, A.P.D., Galo, M., Habib, A.F., 2018. Surface Gradient Approach for Occlusion Detection Based on Triangulated Irregular Network for True Orthophoto Generation. IEEE J-STARS. 11, 443–457. https://doi.org/10.1109/JSTARS.2017.2786162.

Dornaika, F., Moujahid, A., El Merabet, Y., Ruichek, Y., 2016. Building detection from orthophotos using a machine learning approach: An empirical study on image segmentation and descriptors. Expert Syst. Appl. 58, 130–142. https://doi.org/10.1016/j.eswa.2016.03.024.

Ebrahimikia, M., Hosseininaveh, A., 2022. True orthophoto generation based on unmanned aerial vehicle images using reconstructed edge points. Photogramm. Rec. 37 (178), 161–184. https://doi.org/10.1111/phor.12409.

Farenzena, M., Fusiello, A., Gherardi, R., 2009. Structure-and-motion pipeline on a hierarchical cluster tree. In: In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1489–1496. https://doi.org/10.1109/ICCVW.2009.5457435.

Fitzgibbon, A.W., Zisserman, A., 1998. Automatic camera recovery for closed or open image sequences. In: Computer Vision – ECCV 1998. pp. 311-326. Doi: 10.1007/BFb0055675.

Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A., 2022. Plenoxels: Radiance fields without neural networks. In: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5501–5510.

Hu, B., Huang, J., Liu, Y., Tai, Y.W., Tang, C.K., 2023. NeRF-RPN: A General Framework for Object Detection in NeRFs. In: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23528–23538.

Jauregui, M., Vílchez, J., Chacón, L., 2002. A procedure for map updating using digital mono-plotting. Comput Geosci. 28, 513–523. https://doi.org/10.1016/S0098-3004(01)00068-1.

Konecny, G., 1979. Methods and Possibilities for Digital Differential Rectification. Photogramm Eng Remote Sensing. 45 (6), 727–734.

Kuzmin, Y.P., Korytnik, S.A., Long, O., 2004. Polygon-based true orthophoto generation. In ISPRS Congress Proceedings. 12–23.

Li, J., Bosché, F., Lu, C.X., Wilson, L., 2023. Occlusion-free Orthophoto Generation for Building Roofs Using UAV Photogrammetric Reconstruction and Digital Twin Data. In ISARC 2023, 371–378.

Li, T., Jiang, C., Bian, Z., Wang, M., Niu, X., 2020. A Review of True Orthophoto Rectification Algorithms. IOP Conf. Ser.: Mater. Sci. Eng. 780, 022035. https://doi.org/10.1088/1757-899X/780/2/022035.

Lorensen, W., Cline, H., 1998. Marching cubes: A high resolution 3D surface construction algorithm. In Seminal Graphics: Pioneering Efforts That Shaped the Field. 347–353. https://doi.org/10.1145/37402.37422.

Marsetic, A., 2021. Robust Automatic Generation of True Orthoimages from Very High-Resolution Panchromatic Satellite Imagery Based on Image Incidence Angle for Occlusion Detection. IEEE J-STARS. 14, 3733–3749. https://doi.org/10.1109/JSTARS.2021.3067457.

Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R., 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In Computer Vision – ECCV 2020. Doi: 10.1007/978-3-030-58452-8_24.

Müller, T., Evans, A., Schied, C., Keller, A., 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. ACM Trans. Graph. 41, 1–15. https://doi.org/10.1145/3528223.3530127.

Ngp_pl, 2023, Instant-ngp in pytorch+cuda trained with pytorch-lightning [WWW Document]. URL https://github.com/kwea123/ngp_pl.

Oliveira, H.C., Galo, M., 2013. Occlusion Detection by Height Gradient for True Orthophoto Generation, using LiDAR Data. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XL-1/W1, 275–280. https://doi.org/10.5194/isprsarchives-XL-1-W1-275-2013.

Pix4D, 2023. Professional photogrammetry and drone mapping software | Pix4D [WWW Document]. URL https://www.pix4d.com/.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. Adv. Neural Inf. Process, Syst, p. 30.

Qin, R., Tian, J., Reinartz, P., 2016. 3D change detection – Approaches and applications. ISPRS J. Photogramm. Remote Sens. 122, 41–56. https://doi.org/10.1016/j.isprsjprs.2016.09.013.

Reiser, C., Peng, S., Liao, Y., Geiger, A., 2021. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In: In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14335–14345.

Schonberger, J.L., Frahm, J.-M., 2016. Structure-from-Motion Revisited. In: In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4104–4113.

Smart3D, 2023. Smart3D [WWW Document]. URL https://www.smart3d.cn/.

Sweeney, C., Sattler, T., Hollerer, T., Turk, M., Pollefeys, M., 2015. Optimizing the Viewing Graph for Structure-from-Motion. In: In: Proceedings of the IEEE International Conference on Computer Vision, pp. 801–809. https://doi.org/10.1109/ICCV.2015.98.

Szostak, M., Wezyk, P., Tompalski, P., 2014. Aerial Orthophoto and Airborne Laser Scanning as Monitoring Tools for Land Cover Dynamics: A Case Study from the Milicz Forest District (Poland). Pure Appl. Geophys. 171, 857–866. https://doi.org/10.1007/s00024-013-0668-8.

Tipdecho, T., Chen, X., Tokunaga, M., Phien, H.N., 2001. Transformation Based Polynomial Model: In Case of Generating Orthophoto. ANN GIS. 7, 90–98. https://doi.org/10.1080/10824000109480559.

Toldo, R., Gherardi, R., Farenzena, M., Fusiello, A., 2015. Hierarchical structure-and-motion recovery from uncalibrated images. Comput vis Image Underst. 140, 127–143. https://doi.org/10.1016/j.cviu.2015.05.011.

Vassilopoulou, S., Hurni, L., Dietrich, V., Baltsavias, E., Pateraki, M., Lagios, E., Parcharidis, I., 2002. Orthophoto generation using IKONOS imagery and high-resolution DEM: a case study on volcanic hazard monitoring of Nisyros Island (Greece). ISPRS J. Photogramm. Remote Sens. 57, 24–38. https://doi.org/10.1016/S0924-2716(02)00126-0.

Vijayanarasimhan, S., Ricco, S., Schmid, C., Sukthankar, R., Fragkiadaki, K., 2017. SfM-Net: Learning of Structure and Motion from Video. arXiv preprint arXiv:1704.07804.

Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W., 2023. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. arXiv preprint arXiv:2106.10689.

Wang, Q., Yan, L., Sun, Y., Cui, X., Mortimer, H., Li, Y., 2018a. True Orthophoto Generation using Line Segment Matches. Photogram Rec. 33, 113–130. https://doi.org/10.1111/phor.12229.

Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C., 2018b. Repulsion Loss: Detecting Pedestrians in a Crowd. In: In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7774–7783. https://doi.org/10.1109/CVPR.2018.00811.

Wilson, K., Snavely, N., 2014. Robust Global Translations with 1DSfM, In Computer Vision – ECCV 2014. pp. 61–75. Doi: 10.1007/978-3-319-10578-9_5.

Wu, Z., Song, S., Khosla, A., Fisher Yu, Linguang Zhang, Xiaoou Tang, Xiao, J., 2015. 3D ShapeNets: A deep representation for volumetric shapes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1912–1920. Doi: 10.1109/CVPR.2015.7298801.

Xiangli, Y., Xu, L., Pan, X., Zhao, N., Rao, A., Theobalt, C., Dai, B., Lin, D., 2022. BungeeNeRF: Progressive Neural Radiance Field for Extreme Multi-scale Scene Rendering. In: Computer Vision – ECCV 2022, Lecture Notes in Computer Science. pp. 106–122. Doi: 10.1007/978-3-031-19824-3_7.

Xie, W., Zhou, G., 2010. Occlusion and Shadow Detection of Large-scale True Orthophoto in Urban Area. Acta Geod. Et Cartogr. Sin. 39, 52–58.

Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L., 2018. MVSNet: Depth Inference for Unstructured Multi-view Stereo. In: Computer Vision – ECCV 2018. pp. 785–801. Doi: 10.1007/978-3-030-01237-3_47.

Yen-Chen, L., Florence, P., Barron, J.T., Rodriguez, A., Isola, P., Lin, T.Y., 2021. INeRF: Inverting Neural Radiance Fields for Pose Estimation. In: In 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1323–1330. https://doi.org/10.1109/IROS51168.2021.9636708.

Zhang, J., Xu, S., Zhao, Y., Sun, J., Xu, S., Zhang, X., 2023. Aerial orthoimage generation for UAV remote sensing: Review. Inf. Fusion. 89, 91–120. https://doi.org/10.1016/j.inffus.2022.08.007.

Zhang, K., Riegler, G., Snavely, N., Koltun, V., 2020. NeRF++: Analyzing and Improving Neural Radiance Fields. arXiv preprint arXiv:2010.07492.

Zhi, S., Laidlow, T., Leutenegger, S., Davison, A.J., 2021. In-Place Scene Labelling and Understanding with Implicit Scene Representation. In: In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15818–15827. https://doi.org/10.1109/ICCV48922.2021.01554.

Zhu, S., Shen, T., Zhou, L., Zhang, R., Wang, J., Quan, L., Fang, T., 2017. Accurate, Scalable and Parallel Structure from Motion. (Doctoral dissertation, Hong Kong University of Science and Technology). Doi: 10.14711/thesis-991012532269103412.

Zhou, G., Chen, W., Kelmelis, J.A., Zhang, D., 2005. A comprehensive study on urban true orthorectification. IEEE Trans. Geosci. Remote Sensing. 43, 2138–2147. https://doi.org/10.1109/TGRS.2005.848417.