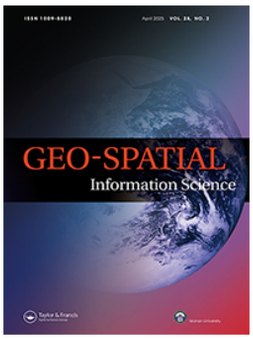

Geo-spatial Information Science

ISSN: 1009-5020 (Print) 1993-5153 (Online) Journal homepage: www.tandfonline.com/journals/tgsi20

Ortho-NeRF: generating a true digital orthophoto map using the neural radiance field from unmanned aerial vehicle images

Shihan Chen, Qingsong Yan, Yingjie Qu, Wang Gao, Junxing Yang & Fei Deng

To cite this article: Shihan Chen, Qingsong Yan, Yingjie Qu, Wang Gao, Junxing Yang & Fei Deng (2025) Ortho-NeRF: generating a true digital orthophoto map using the neural radiance field from unmanned aerial vehicle images, Geo-spatial Information Science, 28:2, 741-760, DOI: 10.1080/10095020.2023.2296014

To link to this article: https://doi.org/10.1080/10095020.2023.2296014

Published online: 08 Mar 2024.

GEO-SPATIAL INFORMATION SCIENCE
2025, VOL. 28, NO. 2, 741–760
https://doi.org/10.1080/10095020.2023.2296014

Ortho-NeRF: generating a true digital orthophoto map using the neural radiance field from unmanned aerial vehicle images

Shihan Chen [a], Qingsong Yan [a], Yingjie Qu [a], Wang Gao [b], Junxing Yang [c] and Fei Deng [a,d]

[a]School of Geodesy and Geomatics, Wuhan University, Wuhan, China; [b]Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, Beijing, China; [c]School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China; [d]Wuhan Tianjihang Information Technology Co. Ltd, Wuhan, China

ABSTRACT

True Digital Orthophoto Maps (TDOMs) have high geometric accuracy and rich image characteristics, making them essential geographic data for national economic and social development. Complex terrain and artificial structures, automatic distortion elimination and occluded area recovery in TDOM generation pose significant challenges. Hence, the need for further improvements in both mapping accuracy and automation is highlighted. In this paper, we present an approach for generating a TDOM based on a Neural Radiance Field (NeRF) without utilizing prior three-dimensional geometry information called an Ortho Neural Radiance Field (Ortho-NeRF). The Ortho-NeRF divides a large-scale scene into small tiles, implicitly reconstructing each tile by selecting pixels on posed images, and individually generate TDOMs of all tiles using a true-ortho-volume rendering before mosaicking. Additionally, the Ortho-NeRF uses a strategy to skip empty spaces and adaptively set the spatial resolution of a voxel grid, improving the generated TDOM quality with fewer computational resources. Many experiments showed that our approach outperforms ContextCapture, Metashape, Pix4DMapper, and Map2DFusion, especially in challenging areas. Owing to its global consistency and continuous nature, Ortho-NeRF was able to effectively reconstruct the geometry information and details, generating TDOMs without distortion or misalignment. Eight ground control points were randomly selected to evaluate the geometric accuracy of the TDOMs, with an average median error of 0.267 m. The length between two points on a plane was also measured for quantitative evaluation, with a mean absolute error of 0.08 m and a mean relative error of 0.14%. Compared with the NeRF efficiency, that of the Ortho-NeRF increased 104 times in training and about 1000 times in rendering.

ARTICLE HISTORY
Received 26 October 2022
Accepted 11 December 2023

KEYWORDS
Neural radiance field; neural implicit representations; True Digital Orthophoto Map (TDOM); Unmanned Aerial Vehicle (UAV); true-ortho-volume rendering

1. Introduction

Unmanned Aerial Vehicle (UAV) remote-sensing methods are flexible, convenient, low-cost, and safe, so they have been widely adopted to quickly obtain massive amounts of data from complex environments (Colomina and Molina 2014; Nex and Remondino 2014; Shao et al. 2021; Toth and Jóźków 2016). Various products can be generated from images containing flight metadata, such as the Digital Surface Model (DSM) (Bhandari et al. 2015; Haarbrink and Eisenbeiss 2008), Digital Elevation Model (DEM) (Ruzgiene et al. 2014; Uysal, Toprak, and Polat 2015), Digital Orthophoto Map (DOM) (Barazzetti et al. 2014; Popescu, Iordan, and Păunescu 2016; Yuan et al. 2023), and point clouds (Everaerts 2008; Harwin and Lucieer 2012; Yang, Haala, and Dong 2023). However, because images from UAVs are obtained with a perspective projection, causing a non-uniform scale between the center and edge of the image is produced. Thus, the use of a True Digital Orthophoto Map (TDOM) is recommended to extract precise geometric information.

Perspective projection and oblique capture cause geometric distortion in remote-sensing images, hindering the restoration of the actual information of the ground surface. DOMs are vertical views of a surface and only eliminate the projection deformation caused by terrain fluctuations and oblique photography. Hence, issues such as artificial ground object dislocation and occlusion still remain. A TDOM is a map produced by a vertical parallel projection of the Earth's surface, eliminating the projection differences of standard DOMs while retaining the geometric accuracy of the map and the visual characteristics of the image (Liu et al. 2018; Wolf, Dewitt, and Wilkinson 2014).

The main step of conventional methods for TDOM generation is detecting the masked areas and using additional images (for the occluded regions) and differential rectification (for geometric distortions) to repair the texture (Amhar, Jansa, and Ries 1998; Hu, Stanley, and Xin 2016; Oda et al. 2004; Schickier and Thorpe 1998; Zhou et al. 2022). The accuracy of this visibility analysis is the key factor affecting TDOM quality (Amhar, Jansa, and Ries 1998; Biasion,

CONTACT Fei Deng fdeng@sgg.whu.edu.cn

Dequal, and Lingua 2004). Three types of TDOM generation methods can be distinguished according to the analyzed data and selected approach, the first of which is DSM-based TDOM generation (Amhar, Jansa, and Ries 1998; de Oliveira et al. 2018; Habib, Kim, and Kim 2007; Liu et al. 2018; Oda et al. 2004; Qin et al. 2003; Rau, Chen, and Chen 2002; Sutherland, Sproull, and Schumacker 1974). Compared to digital terrain models (DTMs), DSM can represent geographic features above the ground, which ensures the height variation of artificial structures and vegetation are considered in the orthorectification process (Amhar, Jansa, and Ries 1998). However, some clear errors appear in true orthophotos when using this method, including double mapping, dislocation, and sawtooth effects (Gilani, Awrangjeb, and Lu 2016; Sheng 2007; Slonecker, Johnson, and McMahon 2009; Zhou et al. 2005). Double mapping is caused when only differential rectification is performed, without occlusion detection and texture filling; simultaneously, dislocation and sawtooth effects are the result of low DSM quality and georeferencing errors (Wang et al. 2018). The second type is Digital building model (DBM)-based generation (Chen et al. 2007; Cheng et al. 2010; Deng et al. 2017; Wang, Jiang, and Xie 2009; Kwak and Habib 2014; Schickier and Thorpe 1998; Zhang et al. 2012). DBM stores building boundaries, which helps detect multi-view image occlusion and compensate texture (Deng et al. 2017). Nevertheless, because DBM is obtained through manual interaction, the production costs are relatively high. Furthermore, DBM-based methods can only work over areas with buildings and heavily rely on the quality of DBM. The third method is deep learning-based TDOM generation (Ebrahimikia and Hosseininaveh 2022; Shin, Hyung, and Lee 2020; Shin, Lee, and Jung 2021). Only a few studies have focused on deep learning-based TDOM generation, but these have shown potential to produce higher-quality, detailed results exceeding those of traditional methods (Ebrahimikia and Hosseininaveh 2022; Shin, Lee, and Jung 2021). Even though superior results could be obtained in most areas, traditional methods cannot deal with complex scenes like thin structure and vegetation, while deep-learning-based methods have weak generalization (Shin, Lee, and Jung 2021). Therefore, there is an urgent need for a simple and efficient method that can generate high-quality TDOMs.

With the development of deep learning and computer vision, neural implicit representations have emerged for three-dimensional (3D) scenes (Chen and Zhang 2019; Mescheder et al. 2019; Park et al. 2019). Neural implicit representations use neural networks to map 3D coordinates to attributes, such as colors, radiance, or distance to object surface. Attributes in a scene cannot be displayed directly and can only be queried by coordinates and observation directions. In particular, Mildenhall et al. (2020) together with several studies (Martin-Brualla et al. 2021; Niemeyer and Geiger 2021; Schwarz et al. 2020; Tewari et al. 2022; Wang et al. 2021), demonstrated photorealistic novel viewpoint rendering that captures scene geometry as well as view-dependent effects using a Neural Radiance Field (NeRF). The NeRF represents a scene using a fully connected deep network, whose inputs are continuous spatial location and viewing direction and whose outputs are volume density and the view-dependent emitted radiance at that spatial location. Subsequently, volume rendering techniques map the output colors and densities to form an image, which is compared with the observed image to optimize theNeRF. However, the NeRF method requires a large number of computational resources for training and inference; some researchers have improved it by decreasing its time consumption (Garbin et al. 2021; Hu et al. 2022; Müller et al. 2022; Xu et al. 2022; Yu et al. 2021). Plenoxels (Fridovich-Keil et al. 2022) uses a voxels grid with Spherical Harmonic (SH) functions – a simple and efficient structure – to represent a 3D scene, which decreases the training and inference times by two orders of magnitude compared to NeRF. Our work builds on Plenoxels, introducing three enhancements to overcome the challenges associated with large-scale TDOM generation.

This study describes a novel TDOM-gemeration method based on neural implicit representation from UAV data. Our method does not inherit the framework of visibility analysis and padding, nor does it need prior of 3D geometry information. The Ortho Neural Radiance Field (Ortho-NeRF) developed here implicitly reconstructs the scene and subsequently generates a TDOM by true-ortho-volume rendering, avoiding the above disadvantages owing to the global consistency and continuous nature of NeRF. By performing several experiments, we also show that our method obtains better TDOMs than those generated using traditional methods and mainstream photogrammetric software. The main contributions of our study are as follows. First, we established a NeRF of the target scene with posed images, and a TDOM was rendered from a given set of ray poses, whose origin points were evenly distributed in the scene, with a vertically upward direction. Second, we expanded our method to a large scene by dividing it and reconstructing each tile to produce a TDOM; the TDOMs were arranged in a mosaic according to their geographic coordinates. Third, the spatial resolution of grids and the range of reconstruction was adapted according to the spatial resolution of images and sparse point clouds after image orientation.

2. Related work

A brief overview of TDOM-generation methods in previous research is given in this section, including DSM-based and DBM-based methods; deep learning-based methods are discussed independently. Subsequently, a novel view synthesis based on the NeRF is described.

2.1. Traditional TDOM generation

According to visibility analysis methods, the two main types of traditional TDOM-generation technology are DSM-based and DBM-based methods. In DSM-based TDOM generation, the Z-buffer method is a classical visibility analysis algorithm developed in the early days of computer graphics (Amhar, Jansa, and Ries 1998; Sutherland, Sproull, and Schumacker 1974) that is easy to implement and widely used. However, this method also has limitations, such as its sensitivity to the sampling interval of the DSM. The angle-based occlusion detection method analyses visibility by comparing the ray and object elevation angles (Gharibi and Habib 2018; Habib, Kim, and Kim 2007; Oda et al. 2004). However, angle-based occlusion analysis requires a frequent comparison of angle size, resulting in a low efficiency. The height-based approach is an alternative methodology for true orthophoto generation using satellite imagery, which determines whether the vision of a camera is occluded according to the elevation (Bang et al. 2007). The height-gradient-based method aims to prevent double mapping by analyzing the height gradient of objects (de Oliveira, Galo, and Poz 2015).

Conversely, in DBM-based TDOM-generation technology, an occlusion detection method (Wang, Jiang, and Xie 2009) computes the intersection points of the rays and Earth's surface with an iterating and rasterizing strategy. The Polygon Based Inversion Imaging (PBI) method (Zhong et al. 2010) projects a building polygon to an image space and inverses the status of objects and occlusions between objects. The Triangulated Irregular Network (TIN)-based occlusion detection method (Zhang et al. 2012) further solves problems regarding building tilt and ghost images. Moreover, the overall projection-based occlusion detection method (Deng et al. 2017) obtains occluded areas by subtracting the orthographic and perspective projections of buildings. Hyperspectral sensors (Wang and Gu 2022) have also been used, including Light Detection and Ranging (LiDAR) for obtaining elevation information lost in the imaging process, and the Global Navigation Satellite System (GNSS)/Inertial Navigation System (INS) unit for direct georeferencing to remove geometric distortions.

The methods mentioned here have shown satisfactory performances, but the following shortcomings still exist: 1) The detected accuracy is highly dependent on the DSM/DBM quality, and the pixel-by-pixel method has clear jaggedness, distortion, or blur at the edge of the detection area (Wang et al. 2018). 2) A vacancy in the masked area needs to be filled, and the compensated texture is always unnatural (Shin, Lee, and Jung 2021). 3) The detecting and filling algorithms are usually complex and inefficient, with low automation (Yuan et al. 2023; Zhou 2020). 4) Differences in color, geometry, and radiometric characteristics at the boundaries are common during mosaicking in large-scale TDOM generation (Chen et al. 2018; Yang et al. 2021).

2.2. Deep learning-based TDOM generation

With the development of deep learning, deep learning-based methods have been used for photogrammetric and remote sensing tasks (Ma et al. 2019; Zhang, Zhang, and Du 2016; Zhu et al. 2017), as well as for several TDOM-generation methods. Shin et al. (2021) proposed a method to generate TDOM using a Generative Adversarial Network (GAN) with a Pix2Pix model from LiDAR intensity data and a DSM. While true orthophotos can be directly generated from LiDAR data, their performance will be highly dependent on the quality of the LiDAR data, and the use of models may be limited to scenes with similar properties to those of the training scenes. Ebrahimikia and Hosseininaveh (2022) improved edge detection from two-dimensional (2D) images and 3D edge graph generation for generating high-quality true orthophotos. They used a deep learning approach identify the edges of artificial features and refined DSM by adding the estimated 3D edges to point clouds, thereby, improving the quality of the generated true orthophotos. This approach can successfully remove edge area twists but still uses a visibility analysis and texture compensation framework that is complex and fails in certain scenes with low-quality 3D point clouds.

2.3. Neural radiance field

Recently, studies on the NeRF have achieved impressive results regarding novel view rendering (Mildenhall et al. 2020). However, owing to invalid sampling along rays and the inquiry of the entire Multi-Layer Perceptron (MLP), completing the training of a scene takes from hours to days. Several studies have been carried out to improve its efficiency (Fridovich-Keil et al. 2022; Garbin et al. 2021; Hu et al. 2022; Karnewar et al. 2022; Kerbl et al. 2023; Müller et al. 2022; Reiser et al. 2023; Sun, Sun, and Chen 2022; Xu et al. 2022; Yu et al. 2021), while some methods have applied practical tricks

to reduce redundancy. For example, FastNeRF (Garbin et al. 2021) introduces a caching mechanism that can reduce computation in the rendering process. Point-NeRF (Xu et al. 2022) uses a neural 3D point cloud by combining the MLP and point cloud, skipping empty spaces in rendering to accelerate the process. EfficientNeRF (Hu et al. 2022) calculates the density distribution of a scene to prune sample points on rays and use a new data structure to cache during inference. Furthermore, 3D Gaussian Splatting (Kerbl et al. 2023) utilizes Gaussians to represent scenes, thus preserving the continuous properties of radiance fields while avoiding unnecessary computation in empty space.

Voxel grids are another efficient 3D representation for NeRFs (Fridovich-Keil et al. 2022; Müller et al. 2022; Sun, Sun, and Chen 2022; Yu et al. 2021). In them, the features of a scene are stored in voxel grid vertices, and the attributes of sampled points are obtained from interpolation, combining the editability of explicit representation and the continuity of NeRF. PlenOctrees (Yu et al. 2021) extracts a sparse voxel grid from the neural radiance field with SH coefficients to represent the view-dependent color, accelerating the rendering speed. Plenoxels (Fridovich-Keil et al. 2022) extend PlenOctrees by using density parameters in voxels, trilinearly interpolating the color and density of sample points along the ray from grid vertices instead of predicting them with a neural network. Finally, instant neural graphics primitives (Instant-ngp) (Müller et al. 2022) encode the inputs of the NeRF using adaptive and efficient multi-resolution hash encoding before the MLP. Subsequently, the encoding value and MLP are simultaneously optimized, which enables the use of smaller, more efficient MLPs to decrease the training and inference times.

3. Methods

In this section, we first describe the deep learning-based approach for novel view rendering used in this work and the vertical upward projection for TDOM generation, denoted as true-ortho-volume rendering. Then, we detail how to handle a large-scale scene and automatically calculate the scope for reconstruction and grid spatial resolution. Figure 1 displays the overview of our Ortho-NeRF model.

3.1. Basic framework of NeRF and plenoxels

NeRF (Mildenhall et al. 2020) is a continuous MLP function that maps from the coordinate x and observation direction d to the volume density σ and color c of a point. It is difficult for an MLP to map a low-

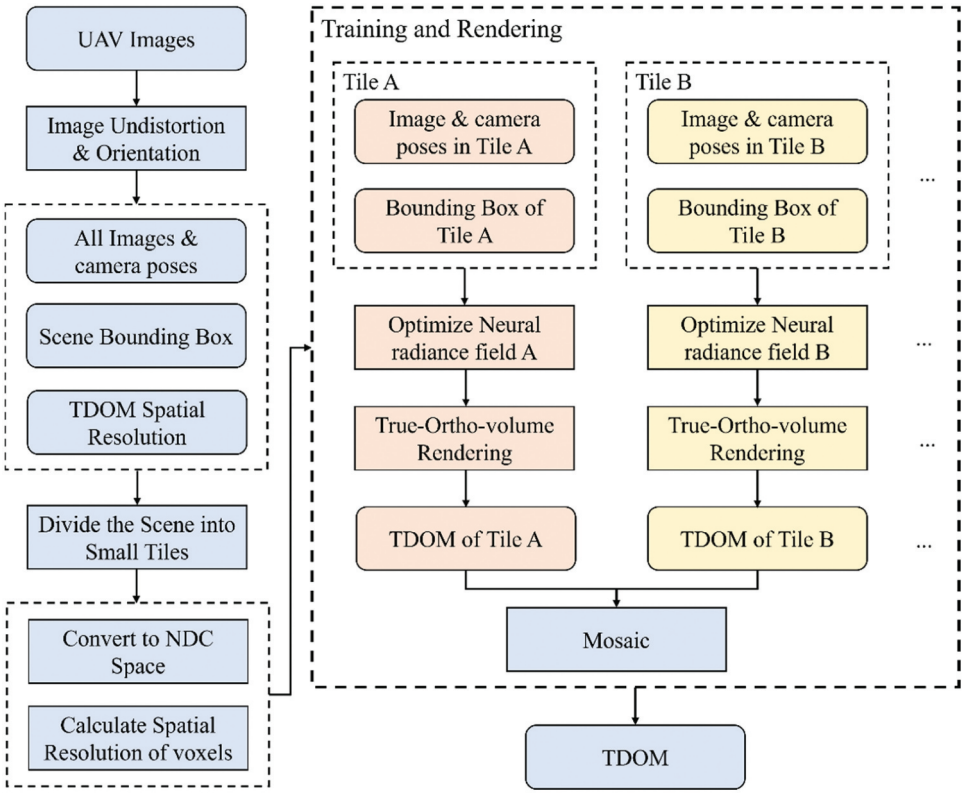

Figure 1. Flowchart of Ortho-NeRF. First, we obtain the poses of images and extract sparse point clouds by structure from motion (SfM). The spatial resolution of the target TDOM and the bounding box of the scene are calculated. We then divide the scene into small tiles and select the images and pixels for reconstructing every tile. The normalized device coordinate (NDC) space of Ortho-NeRF is initialized according to the tile's boundary, and the spatial resolution of the voxels grid is determined according to the target TDOM. Finally, we separately model each tile by plenoptic voxels. Every small TDOM is generated using georeferenced true ortho-volume rendering and mosaicked into a complete map.

frequency signal at a high frequency (Sitzmann et al. 2020). Owing to the high-frequency volume density and color in the field, the inputs x and d were first mapped to the Normalized Device Coordinate (NDC) space following Equation (1) and subsequently to the high-frequency space by sine-cosine transform as shown in Equation (2).

$$x_{NDC} = scale \times (x - x_{center}) \tag{1}$$

$$r(t, L) = (\sin(2^0 t\pi), \cos(2^0 t\pi), \ldots, \sin(2^L t\pi), \cos(2^L t\pi)) \tag{2}$$

where x_{NDC} represents the position in the NDC, and x_{center} represents the coordinates of the center point of the scene. *scale* represents the scaling factor of the scene, t is the combined vector of x_{NDC} and d, wherein d is the direction to observe the point, and L is the output frequency. The NeRF f is shown in Equation (3):

$$(\sigma, c) = f(r(x), r(d)) \tag{3}$$

The color of each pixel $\hat{C}$ is approximated by integrating samples queried along rays to render an image of a given view as shown in Equation (4).

$$\begin{aligned} \hat{C}(\mathrm{r}) &= \sum_{i=1}^{N} T_i w_i c_i, \\ \text{where } w_i &= \left(1 - e^{(-\sigma_i \delta_i)}\right), \\ T_i &= e^{-\sum_{j=1}^{i-1} \sigma_j \delta_j} \end{aligned} \tag{4}$$

where T_i and w_i denote the accumulated transmittance and opacity of point i, respectively, and δ_i is the distance between adjacent samples $i+1$ and i. To train the NeRF f, the loss (L_{rgb}) between observed and rendered images, which is calculated following Equation (5), should be minimized. After optimization, novel views can be synthesized by querying each ray/pixel.

$$L_{rgb} = \sum \|C(r) - \hat{C}(r)\|_2^2 \tag{5}$$

When rendering, NeRFs need to query all rays, which requires extensive computational time for training and rendering. Fridovich-Keil et al. (2022) replaced the MLP with sparse 3D grids with volume density and SH coefficients at each voxel, referred to as plenoptic voxels. In this model, the view-dependent color c at coordinate x was determined by the SH function $Y(d)$, as displayed in Equation (6):

$$c(d) = S(Y(d)) \tag{6}$$

where S is the sigmoid function for normalizing the colors. To render an image, it computed the volume density and color at each sample point along rays by a trilinear interpolation of the density and SH function stored at the nearest eight voxels. Additionally, the differentiable model for volume rendering in NeRF was the same as that used in Plenoxels; similar to NeRF, the difference between rendered and observed images should be minimized to optimize Plenoxels. After training, it rendered the image faster with a known view. Without neural networks, the training decreased sharply by two orders of magnitude, while rendering time decreased by three orders of magnitude, maintaining NeRF quality. Therefore, Ortho-NeRF is based on plenoptic voxels.

3.2. True-ortho-volume rendering

The NeRF shows great promise for the novel image synthesis of complex scenes, as does TDOM generation with parallel projection from plenoptic voxels. Considering this, we propose here a new method to generate TDOMs by setting each ray to project vertically upwards from NDC – the space to reconstruct plenoptic voxels – as illustrated in Figure 2.

Maps, including TDOMs, should have geographic coordinates to represent the location of geographic entities. Therefore, the position and direction of each rendered ray was defined in the world coordinate system, transformed into NDC space using Equation (1), and projected to obtain a TDOM. The position of each ray in the world coordinate system was determined, according to the spatial resolution of the TDOM and scene boundary, with the direction being vertically upward. Subsequently, all pixels were rendered with the transformed ray position and direction.

3.3. Divide and mosaic

The greater the number of voxels in a grid, the more realistic the rendered novel view images. To reconstruct a large scene accurately, so many voxels are needed that they overwhelm the memory of a Graphics Processing Unit (GPU). Therefore, Ortho-NeRF divides the scene into small tiles and individually restores them. This dividing process can be seen in Figure 3; the scene was divided along both the X and Y directions, and the pixels that belong to the tile were selected. We first calculated the bounding box of the scene from the sparse point cloud generated by structure from motion (SfM), and then divided the scene regularly using a rectangular grid, as displayed in Figure 3(a). The size of the grid was calculated based on the spatial resolution of voxels and memory size of the GPU (see Section 4.2 for details). The boundary of each tile was determined after expanding the side length of each side as shown by the red area in Figure 3(b). Because the reconstruction quality at the scene's boundary was degraded, we expanded the reconstruction range and discard the low-quality parts after generating the TDOM. Then, we projected the boundary of each tile onto every image and counted the pixels contained by this bounding box, as shown in Figure 3(c). If that

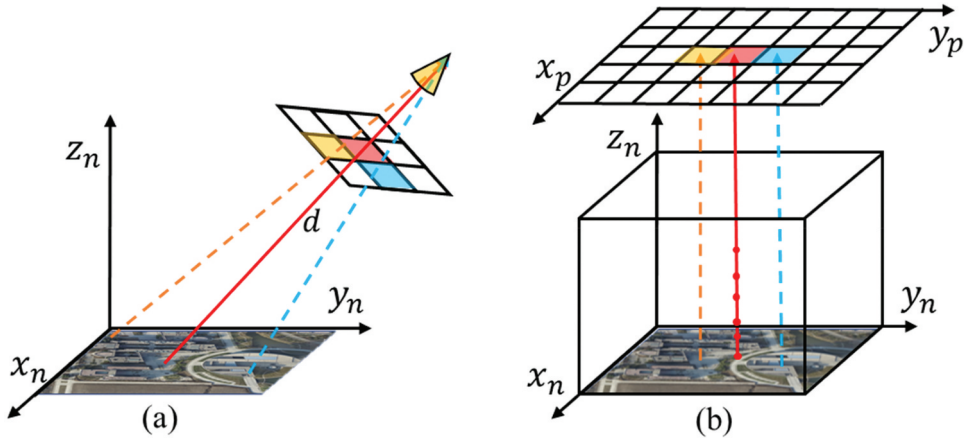

Figure 2. Figure of true-ortho-volume rendering. To render a TDOM from plenoptic voxels, unlike the central projection (a), we evenly generate multiple vertically upward rays from NDC and project them in parallel to a virtual plane (b).

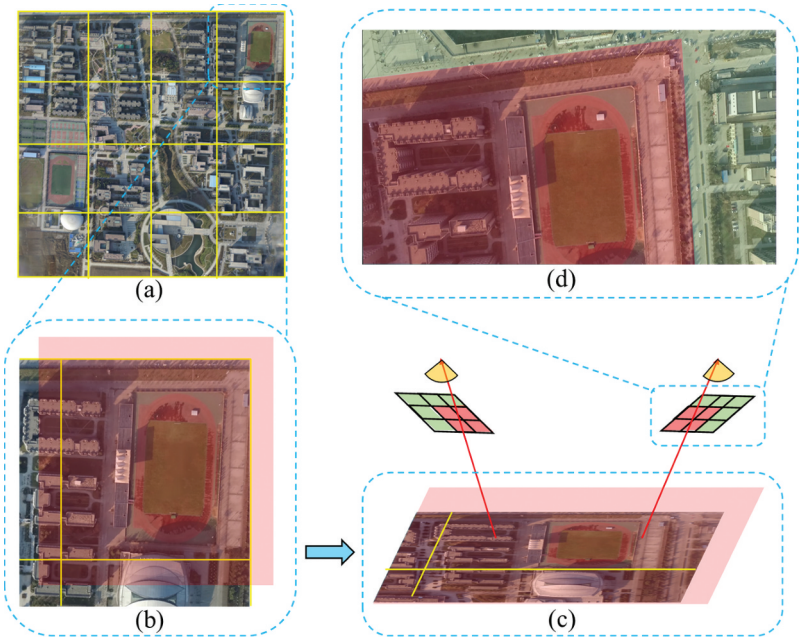

Figure 3. The process of dividing a large-scale scene and selecting only part of an image for training. We regularly divide the scene by rectangular grid (a) Before expanding the side length of each tile. (b) Then, the images (c) and pixels (d) that observe this tile are selected for reconstruction.

number exceeds 5% of the total pixels in an image, the pixels in this range, similarly to the red areas of Figure 3(d) were selected for restoring, whereas the other areas, such as the green areas in Figure 3(d), were removed. Selection considering the full image would have resulted in observations that do not belong to the tile being included, which represent extra noise and reduce the quality of reconstruction for each tile. Finally, the bounding box was calculated using the sparse points in the tile as input.

After reconstructing and rendering the TDOM using each tile, the map was mosaicked according to geographic coordinates. To ensure a certain degree of overlap and avoid mosaic gaps, we expanded the range of each tile by 1% during TDOM generation. Thus, the TDOMs were placed in the correct positions, and the later mosaicked TDOMs overwrote the previous ones.

3.4. Adaptive reconstructed scope and grid spatial resolution

To reconstruct a NeRF, the target scene needs to be centralized and normalized to NDC space, which our approach does after division. Prior research has often used camera poses to centralize and normalize the 3D space, resulting in the scene not filling the cube where the x_n, y_n, and z_n components range from −1 to 1. In other words, many spaces that do not contain any information are also reconstructed by the radiance field as illustrated by the black cube in Figure 4. These spaces have to be recorded as transparent, which makes optimization more difficult and requires extra memory and training time. To avoid these issues, we started with the scene's bounding box in Equation (1) and proportionally compressed it so that the

longest side of the bounding box ranged from −1 to 1. This reconstruction only took place inside the normalized bounding box, represented by the red cube in Figure 4.

The reconstruction accuracy improves with an increased number of voxels. Our methods employ a coarse-to-fine strategy to achieve high-quality reconstruction and rendered images. Figure 4(c) shows this process, which begins with a sparse grid at a low resolution. Unnecessary voxels are removed from the grid, while the remaining ones are subdivided, iteratively continuing the optimization process. In addition, we aim to generate high-quality TDOMs, focusing on reconstruction on the horizontal plane. To improve the TDOM quality, we decreased the spatial resolution of voxels on the vertical axis (z-axis), squeezing out the GPU memory to increase the spatial resolution of voxels on the horizontal plane as illustrated in Figure 4(b). Specifically, in the coarse-to-fine process, voxels in the z-axis direction were not subdivided, but their resolution was only increased in the horizontal plane. To cover all scenes, we set the minimum number of voxels on the z-axis to 128, which remained unchanged during training. Section 4.4 shows the influence of that number on TDOM generation.

A key challenge when choosing the spatial resolution of voxels is reaching a balance between quality and range. The spatial resolution should be high to reconstruct 3D scenes with high quality, which requires a small range of ground area imaged for a pixel. However, as the spatial resolution increases, the reconstructed range drops. Considering the limitation in GPU memory, the number of grids has an upper limit. Each reconstructed range is determined by the number of grids and spatial resolution of voxels ($range = number \times spatialresolution$). Because of the constant number of grids, a linear relationship exists between the spatial resolution and reconstruction range. If the spatial resolution is too high, the reconstruction range will be extremely small. Therefore, we reduced the final voxel's spatial resolution to 1.5 times that of the TDOM (see more in Section 4.4), maintaining the reconstruction quality.

4. Experiments

4.1. Dataset

This study utilized the NPU DroneMap Dataset and a self-made dataset called WHULab DroneMap to evaluate the proposed approach. The study area covers several regions in the Shanxi, Henan, and Hubei provinces in China, with various land use/land cover types, including bungalows, high-rise buildings, vegetation, water bodies, and roads. The NPU DroneMap Dataset was created by Bu et al. (2016) and recorded with a constructed hexacopter equipped with a GoPro

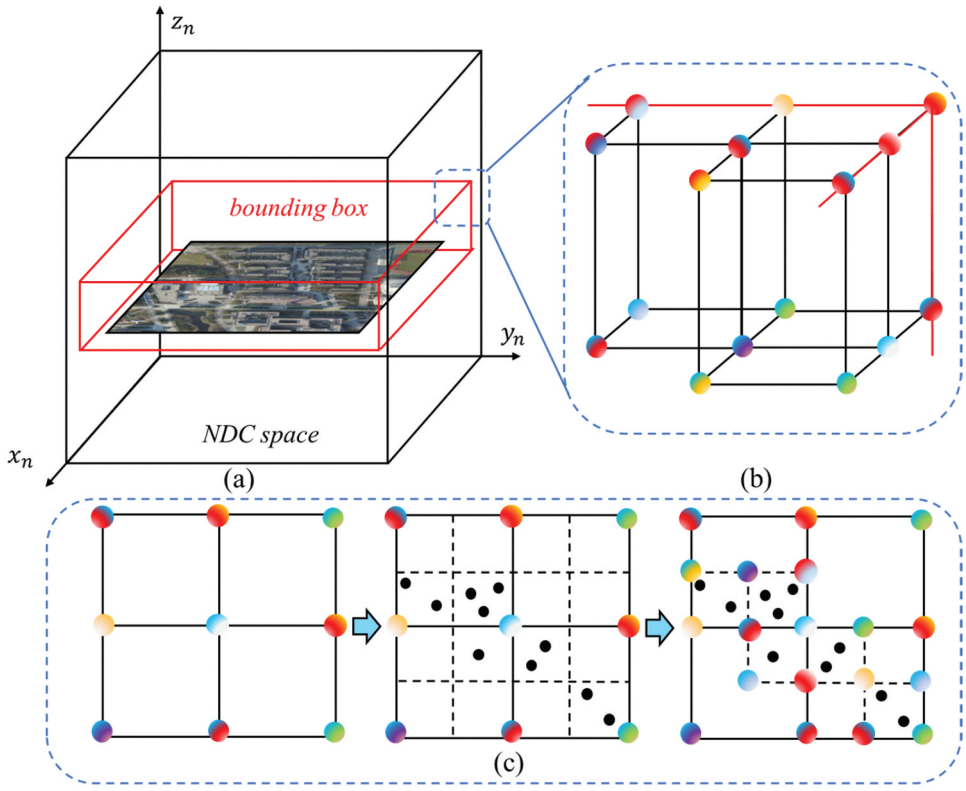

Figure 4. Figure of reconstructed scope and grid spatial resolution. We narrow the scope of reconstruction according to the bounding box of the scene (the red cube in a). The resolution in the *z*-axis direction is reduced to squeeze out the GPU memory and increase the spatial resolution of voxels on the horizontal plane (b). The coarse-to-fine strategy is illustrated in a 2D plane (c).

Table 1. Summary of the dataset used in the experiments.

Dataset	Scene	Location	Spatial resolution	H Max (m)	Number of Images	Area (km^2)
NPU DroneMap	gopro-saplings	Xi'an, Shanxi	0.10	129.2	482	0.455
	gopro-npu	Xi'an, Shanxi	0.34	376.8	337	2.739
	gopro-monticules	Xi'an, Shanxi	0.13	147.2	482	0.571
	phantom3-factory	–	0.15	198.72	402	0.912
	phantom3-freeway	–	0.20	258.3	415	1.457
	highflower	–	0.13	165.6	285	0.526
	lowflower	–	0.05	65	589	0.144
	phantom3-huangqi	Hengdong, Hunan	0.18	222.3	393	1.313
	phantom3-ieu	Zhenzhou, Henan	0.21	282.3	467	1.524
	phantom3-npu	Xi'an, Shanxi	0.20	254.5	457	1.598
	phantom3-village	Hengdong, Hunan	0.15	196.6	406	0.932
WHULab DroneMap	WHULab-0001	Wuhan, Hubei	0.04	161.65	216	0.129
	WHULab-0002	Wuhan, Hubei	0.04	161.95	186	0.141
	WHULab-0003	Wuhan, Hubei	0.04	164.4	136	0.231

–' represents missing location information.

Hero3+ camera and a DJI Phantom3. The dataset includes scenes captured at different flight heights, over differing land covers and terrains. The original data consist of video, flight trajectory, and camera calibration data; Ground Control Points (GCPs) are available for gopro-npu and phantom3-npu. All images in the dataset consist of keyframes extracted from captured videos, ensuring a high overlap, and were predominantly taken in a downward orientation, with a 1920×1080 resolution and a Joint Photographic Experts Group (JPG) format. In addition, we created a dataset called WHULab DroneMap by Phantom4 RTK with higher accuracy and resolution. The images in this last dataset were also acquired in a downward orientation and JPG format, with 90% forward overlap and 70% side overlap, providing a 5472×3648 resolution, as shown in Table 1. All the images were orientated and undistorted using ContextCapture (Bentley 2022), a prominent commercial photogrammetry software. The validation of our approach was conducted across 14 scenes using either a coarse flight trajectory or GCPs.

4.2. Implementation

We divided the scene into many tiles before optimizing the NeRF. The way a scene is partitioned has a significant effect on the quality of reconstruction. After SfM, the bounding box's side length $L = (L_x, L_y, L_z)$ of the sparse point cloud and target spatial resolution S_p of the TDOM were calculated. The spatial resolution of the voxel grid S_v was 1.5 times that of S_p, while the total number of voxels on the horizontal plane was calculated using $(n_x, n_y) = (L_x, L_y)/S_v$, and that on the z-axis was always 128, as mentioned in Section 3.4. A radiance field can only contain about 3.2×10^8 voxels when using a single NVIDIA GeForce RTX3090 GPU; therefore, we set fine voxels as $1600 \times 1600 \times 128$. The number of tiles on each side was estimated from $(t_x, t_y) = [(n_x, n_y)/1600]$, where [] indicates the result was rounded up. Thus, each side length of the tile was given by $l_x = L_x/t_x$, $l_y = L_y/t_y$. However, the edge extended 0.25 times the total length to both ends after division and reconstruction, indicating that each edge of the scene was enlarged by 1.5 times. If the number of voxels did not change, its spatial resolution would have decreased, so we needed to further reduce the edge length to 2/3 of the original.

In the experiments, 5% of the images were randomly selected to validate the reconstruction quality. All experiments were performed on an RTX3090 GPU. To fully utilize the GPU memory, we set the batch size to 5000. We further used an epoch size of 102,400 and 8 epochs for each training, which required approximately 12–18 min to optimize for a single tile.

4.3. Comparison

To evaluate our method, we compared it with ContextCapture (Bentley 2022), Metashape (Agisoft 2022), Pix4DMapper (PIX4D 2022), and Map2DFusion (Bu et al. 2016). These methods can only generate DSM/DTM-based orthomosaic photos or georeferenced orthomosaic photos via a traditional processing workflow. All methods were tested against the NPU DroneMap dataset (Bu et al. 2016). However, Map2DFusion could not run on the WHULab DroneMap dataset owing to its high input requirement of over 95% forward overlap. Camera poses were calculated using ContextCapture for all generation methods. GNSS coordinates of the UAV and GCPs were used as the initial values and constraints, respectively, if they existed. The input views and configurations of our methods were the same for all the scenes.

The boundary areas of scenes lacked sufficient views, resulting in the cropping of the edge portion of the final result, which was necessary to meet the requirements of our method for utilizing multiple

views. The overviews of all scenes are shown in Figure 5; Ortho-NeRF successfully generated TDOMs for all scenes, indicating a good generalization. The results show that our approach achieved a satisfactory quality in different environments – such as vegetation, buildings, and water areas – which is difficult for traditional workflow in addition to an excellent performance with the WHULab DroneMap dataset. Owing to the high spatial resolution of the WHULab DroneMap dataset images, the resolution of the generated TDOM and plenoptic voxels was also high, as discussed in Section 3.4. As a result, the range of each tile was smaller, and fewer images were chosen for reconstruction, which is more challenging for TDOMs generations. However, our method minimized the influence of views by ideally using all the pixels that recorded information about each tile to restore the scene. Moreover, although our method required mosaicking, there were almost no color differences between adjacent tiles.

4.3.1. Qualitative evaluation

A comparison of certain characteristics from all the approaches are illustrated in Figure 6, while the TDOMs of the WHULab DroneMap dataset are compared with those of the other methods except for Map2DFusion Figure 7. The results from these approaches showed a similar quality to that of our approach in the most uniform elevation areas, whereas our approach achieved better results in the following circumstances:

4.3.1.1. Buildings. Our approach is performed better others over areas with buildings. As shown in Figures 6

Figure 5. The TDOM results of all scenes using Ortho-NeRF. Our method successfully handles all the scenes with high robustness and quality in different environments.

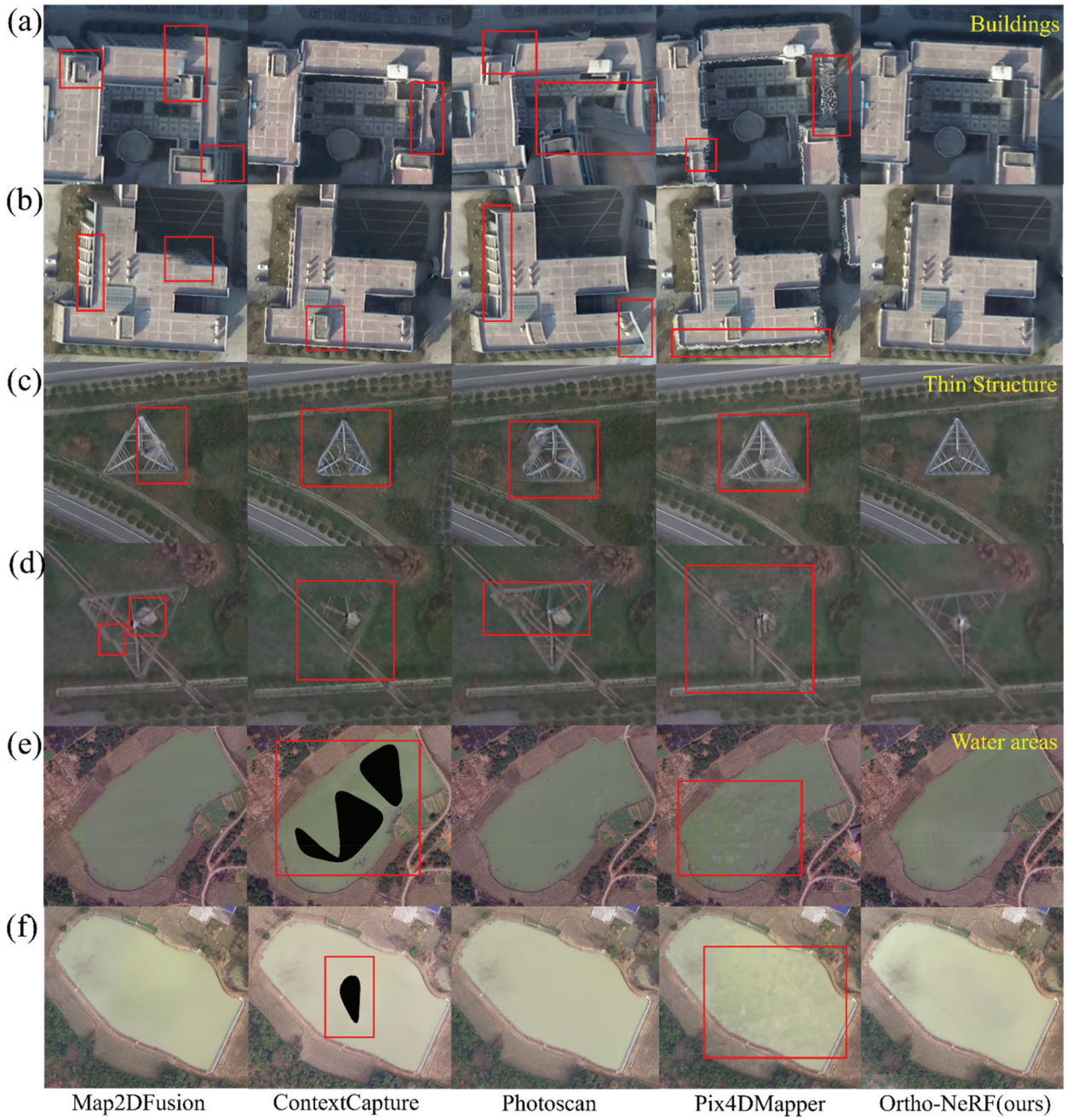

Figure 6. Comparison of TDOMs generated by commercial photogrammetry software, Map2DFusion (Bu et al. 2016) and our methods from NPU DroneMap. Our approach is able to recover better details than other methods in both texture and geometry over building areas (a,b), thin structures (c,d), and weakly textured areas (e,f). The red boxes highlight areas with incorrect or low-quality details.

(a,b) and 7(a,b), Ortho-NeRF clearly indicated borders and their occlusion relationships. However, Map2DFusion simply mosaicked orthophotos based on the geographic location, so building facades are still visible, and some parts are missing in Figure 6(a,b). The other three software showed varying degrees of distortion at building borders. In particular, maps from Metashape were twisty, and those from Pix4DMapper were misaligned, as shown in Figures 6(a,b) and 7(a,b).

4.3.1.2. Thin structures. In Figures 6(c,d) and 7(c,d), the geometric information of the power tower and monitor was only recovered well using our method. The results from Map2DFusion showed misalignment over the power tower and over the roads as shown by the red boxes in Figure 6(c). ContextCapture and Pix4DMapper failed to capture the power tower properly, with overly smooth images in Figures 6(d) and 7(d). All other results showed varying degrees of distortion, deformation, discontinuity, and noise. Thin structures have always been difficult for classical approaches to reconstruct, but Ortho-NeRF correctly restored the geometric relationships even if Figure 6(d) was slightly blurred. This is attributed to a small color difference between the power tower and the background, which hinders restoration.

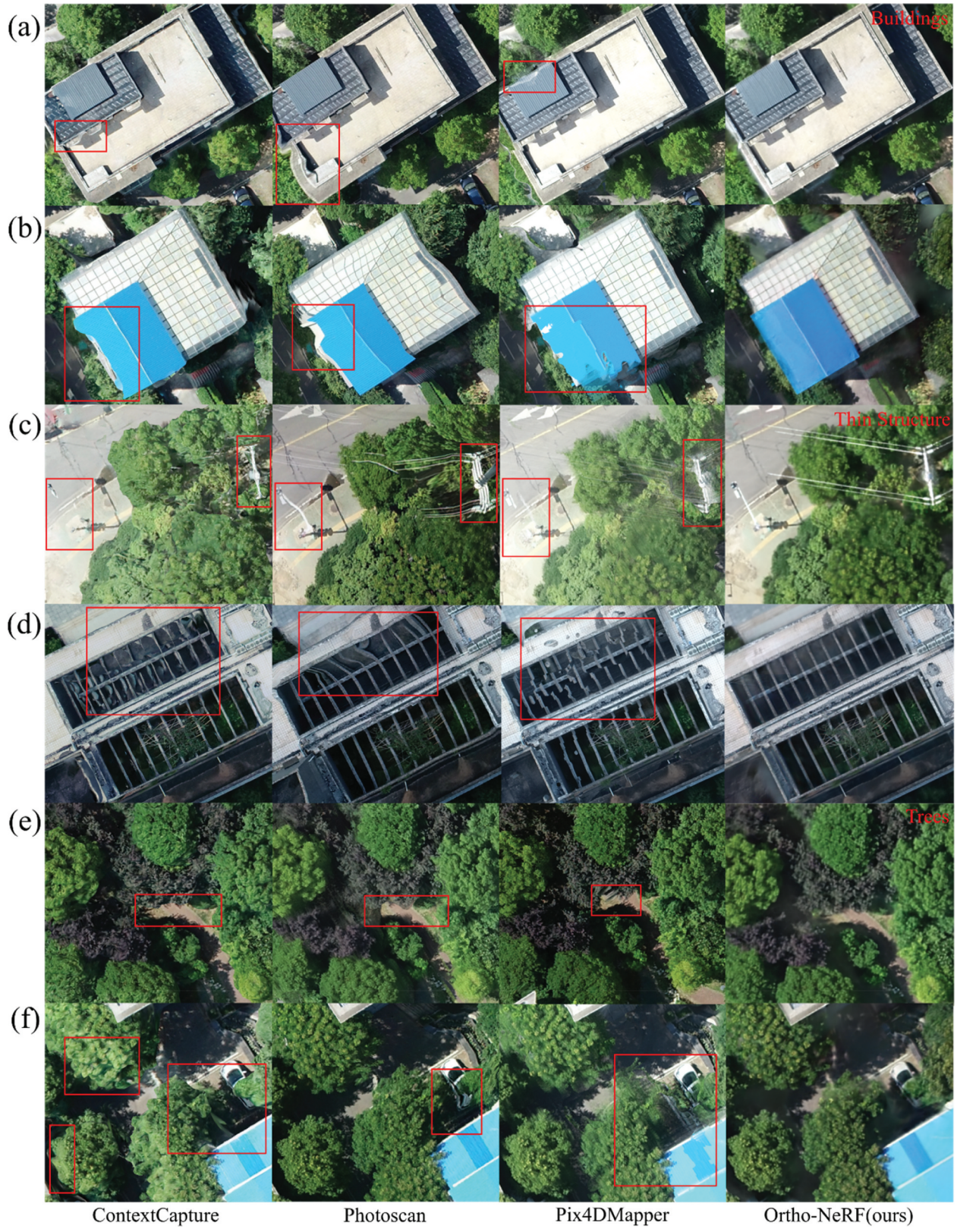

Figure 7. Comparison of TDOMs generated by commercial photogrammetry software and our methods from WHULab DroneMap. TDOMs generated by Ortho-NeRF can represent geometry more consistently and naturally than other results, as shown in building areas (a,b), thin structures (c,d), and trees (e,f). The red boxes highlight areas with incorrect or low-quality details.

4.3.1.3. Weakly textured areas. For weakly textured areas such as water bodies, the TDOMs generated by ContextCapture have holes, as shown in Figure 6(e,f), owing to failed reconstruction, whereas those generated by Pix4DMapper showed color inconsistencies. However, our learning-based method was able to avoid these problems because the sparse voxels grid filled the entire scene, and the NeRF was continuous and smooth, thereby, filtering out noise while preserving detail.

4.3.1.4. Trees. In urban areas are always challenging to reconstruct because, many thin structures make up their crowns of trees, and they trees are not rigid, becoming deformed with the wind, which does not meet the assumptions of traditional reconstruction. In Figure 7(e,f), results of using different methods along with the WHULab DroneMap dataset for tree reconstruction are shown. Our method yielded more natural images, especially where the trees and the background met. The other methods showed sudden changes in color, varying degrees of distortion, and stretching at the junctions.

Our method successfully restored the geometry information of the scenes with little distortion owing to the advantages of the implicit NeRF. Simultaneously, Ortho-NeRF had no offset when mosaicking, which differs from orthomosaic methods that may cause apparent seams. We also reduced the chromatic aberration between TDOMs by expanding the tile range, and our method demonstrated high robustness and successfully handled all the scenarios, obtaining results without holes.

4.3.2. Quantitative evaluation

The geometric accuracy of TDOMs are usually an essential indicator in quality evaluation. Previously, researchers have generally compared the coordinates of GCPs in the world against the corresponding location in the TDOM. In this study, the coordinates of eight random GCPs were random measured in the scene of the WHULab DroneMap dataset and were used as true values. The accuracy of Ortho-NeRF was evaluated by calculating the horizontal position error (Δx and Δy in Table 2) and median error (Δxy in Table 2) for corresponding points on the TDOM. These results were compared with the outcomes obtained using ContextCapture, Metashape, and Pix4Dmapper. The findings indicate that our approach outperformed the baseline methods, exhibiting the lowest mean absolute horizontal position errors (0.095 m in the x direction and 0.233 m in the y direction) and a mean absolute median error of 0.267 m.

Furthermore, as the distance between two points on the same plane using ContextCapture are considered close to the true value, we compared the distances on the same plane obtained using ContextCapture with those using Ortho-NeRF. The Phantom3-ieu and gopro-npu datasets were chosen to be evaluated using the different sensors, and 10 separate distances were randomly compared, which are shown in Table 3. The maximum absolute and relative errors did not exceed 0.2 m and 0.5%, respectively in these two datasets. The mean absolute error was 0.08 m for Phantom3-ieu and 0.06 m for gopro-npu, while the mean relative error was 0.21% for Phantom3-ieu and 0.06% for gopro-npu, indicating a good geometric performance of our approach for most areas.

4.3.3. Efficiency

We further compared the consumption times of Ortho-NeRF and NeRF in training and rendering. For NeRF, we employed 300k batches to ensure the network reached full convergence. After training, we recorded the consumption time of rendered TDOMs with a spatial resolution of 0.2 m and image resolution of 2826×1911. Table 4 displays the training and rendering times of Ortho-NeRF and NeRF for each tile of the Phantom3-npu, consisting of a total of 6 tiles. Ortho-NeRF achieved an average training time of 12.67 min per tile and a rendering time of 0.31 s. In contrast, NeRF required an average of 18.1 h (1327.2 min) for training and 5.63 mins (337.4 s) for rendering on each tile. This demonstrates a remarkable increase in efficiency – approximately 104 times faster for training and about 1000 times faster for rendering – making it applicable to industrial production. The consumption time is directly proportional to the number of tiles reconstructed. That is, the time complexity of this method is $O(n)$, where n is the number of tiles, and the space complexity is $O(1)$, indicating the utilization of only one GPU computing resource. However, employing multiple computing resources, such as multiple GPUs, may reduce the time consumed.

Table 2. Comparison of horizontal position errors (Δx and Δy) and median error Δxy between Ortho-NeRF, ContextCapture (Bentley 2022), Metashape (Agisoft 2022), and Pix4Dmapper (PIX4D 2022).

	ContextCapture (m)			Metashape (m)			Pix4Dmapper (m)			Ortho-NeRF (m)		
GCP number	Δx	Δy	Δxy	Δx	Δy	Δxy	Δx	Δy	Δxy	Δx	Δy	Δxy
1	0.009	0.012	**0.015**	−0.883	−1.112	1.420	0.010	−0.792	0.792	0.010	−0.026	0.028
2	0.107	0.097	0.144	−1.879	0.097	1.882	−0.279	−0.203	0.345	0.067	0.108	**0.127**
3	−0.610	−0.017	0.610	−0.739	−0.926	1.184	0.386	0.956	1.031	0.052	0.016	**0.055**
4	−0.422	−0.788	0.894	1.766	1.257	2.168	−0.258	0.518	**0.579**	−0.455	−0.696	0.832
5	−0.153	0.916	0.929	1.772	−0.125	1.776	0.578	−0.587	**0.824**	0.026	0.898	0.898
6	−0.108	0.085	0.138	−0.317	1.422	1.457	0.167	0.144	0.221	−0.086	0.086	**0.122**
7	0.055	−0.029	0.062	−0.035	−1.896	1.896	−1.456	−2.301	2.723	0.034	−0.021	**0.040**
8	0.032	0.019	0.037	−1.618	1.313	2.084	−0.285	−0.013	0.285	−0.028	0.013	**0.031**
MAE	0.187	0.245	0.354	1.126	1.018	1.733	0.427	0.689	0.850	**0.095**	**0.233**	**0.267**

'GCP number' represents the ground control point's number, and 'MAE' indicates the mean absolute error. Bold values highlight the top-performing method for each metric (lower values are better).

Table 3. Distances on the same plane obtained using ContextCapture and Ortho-NeRF, along with their absolute errors and relative errors.

Scene	ContextCapture	Ortho-NeRF	Absolute error (m)	Relative error (%)
Phantom3-ieu	30.7119	30.5778	0.1342	0.4368
	26.5001	26.5979	−0.0979	0.3693
	35.6854	35.7172	−0.0318	0.089
	105.9119	105.9759	−0.0640	0.0604
	68.5417	68.3634	0.1783	0.2602
	30.1305	30.1362	−0.0057	0.0191
	108.0387	107.9964	0.0423	0.0392
	31.2796	31.3956	−0.1160	0.3709
	22.1137	22.1504	−0.0367	0.1659
	37.1925	37.3034	−0.1109	0.2982
Mean	–	–	0.0818	0.2109
Gopro-npu	66.4656	66.4639	0.0017	0.00252
	160.7820	160.5890	0.1930	0.1200
	49.2569	49.2499	0.0070	0.0142
	62.7652	62.8335	−0.0684	0.1089
	43.1427	43.0867	0.0560	0.1299
	66.0069	66.0474	−0.0405	0.0613
	94.6102	94.7379	−0.1277	0.1349
	54.1610	54.0841	0.0769	0.1420
	64.3367	64.3324	0.0043	0.0067
	117.8323	117.8655	−0.0332	0.0282
Mean	–	–	0.0649	0.0609

Table 4. Training and rendering time comparison between Ortho-NeRF and NeRF.

		Ortho-NeRF		NeRF	
Scene	Tile Number	Training (min)	Rendering (s)	Training (h)	Rendering (min)
Phantom3-npu	1	14.2	1.4	17.8	2.1
	2	13.8	1.3	17.9	2.2
	3	13.6	1.3	18.1	2.1
	4	14.5	1.3	18.3	2.3
	5	13.9	1.3	18.2	2.2
	6	14.0	1.3	17.9	2.1

Two primary reasons account for the significant reduction in consumption time achieved by our method. First, because Ortho-NeRF is voxel-based, it only requires voxels near sample points for density and color during ray marching. In contrast, NeRF infers the properties of sample points from the entire MLP. Similarly, during gradient backpropagation, NeRF modifies the weights and biases of all MLPs, whereas Ortho-NeRF only modifies the SH coefficients and volume density scalars of adjacent voxels. This local optimization process substantially reduces computational requirements while keeping the quality of reconstruction in comparison to global processes. Second, the ray marching process of Ortho-NeRF is implemented in C++, whereas NeRF is entirely written in Python. The inherent high-efficiency advantages of C++ confer additional efficiency improvement to Ortho-NeRF (Fridovich-Keil et al. 2022; Yu et al. 2021).

4.4. Ablations and limitations

4.4.1. Ratio

To evaluate the effect of the ratio of the voxel's spatial resolution to the TDOM's spatial resolution we tested different ratios – from 1 to 3.5–on the phantom3-npu dataset. The peak signal-to-noise ratio (PSNR) is the ratio between the maximum possible power of a signal and the power of noise, which is calculated between ground truth and rendered images to estimate the novel view and reconstruction qualities. We randomly selected 5% of the images as the ground truth, which were not put in training, to calculate and validate the PSNR, as indicated in Figure 8(a). The training PSNR can be observed in Figure 8(b), while some details of the generated TDOM can be found in Figure 8(c). Figure 8(a) shows that when the ratios were 1 and 1.5, the training PSNR were 30.04 and 29.96, respectively, while the validating PSNR were 28.30 and 28.22, respectively. While these values are similar, they decreased sharply when the ratio exceeded 1.5. This indicates a reduction in the quality of novel view synthesis. Similarly, the TDOM maintained a high quality under ratios of 1 and 1.5, but lost details under ratios greater than 1.5. Therefore, in our work, the ratio of the voxel's spatial resolution to the TDOM's spatial resolution was 1.5, balancing performance and efficiency.

4.4.2. Voxels in the z-axis

To ensure that setting the 128 voxels in the z-axis did not blur the generated TDOM, Figure 9 displays the quality of the TDOM under different numbers of voxels in the z-axis, ranging from 32 to 256. Even

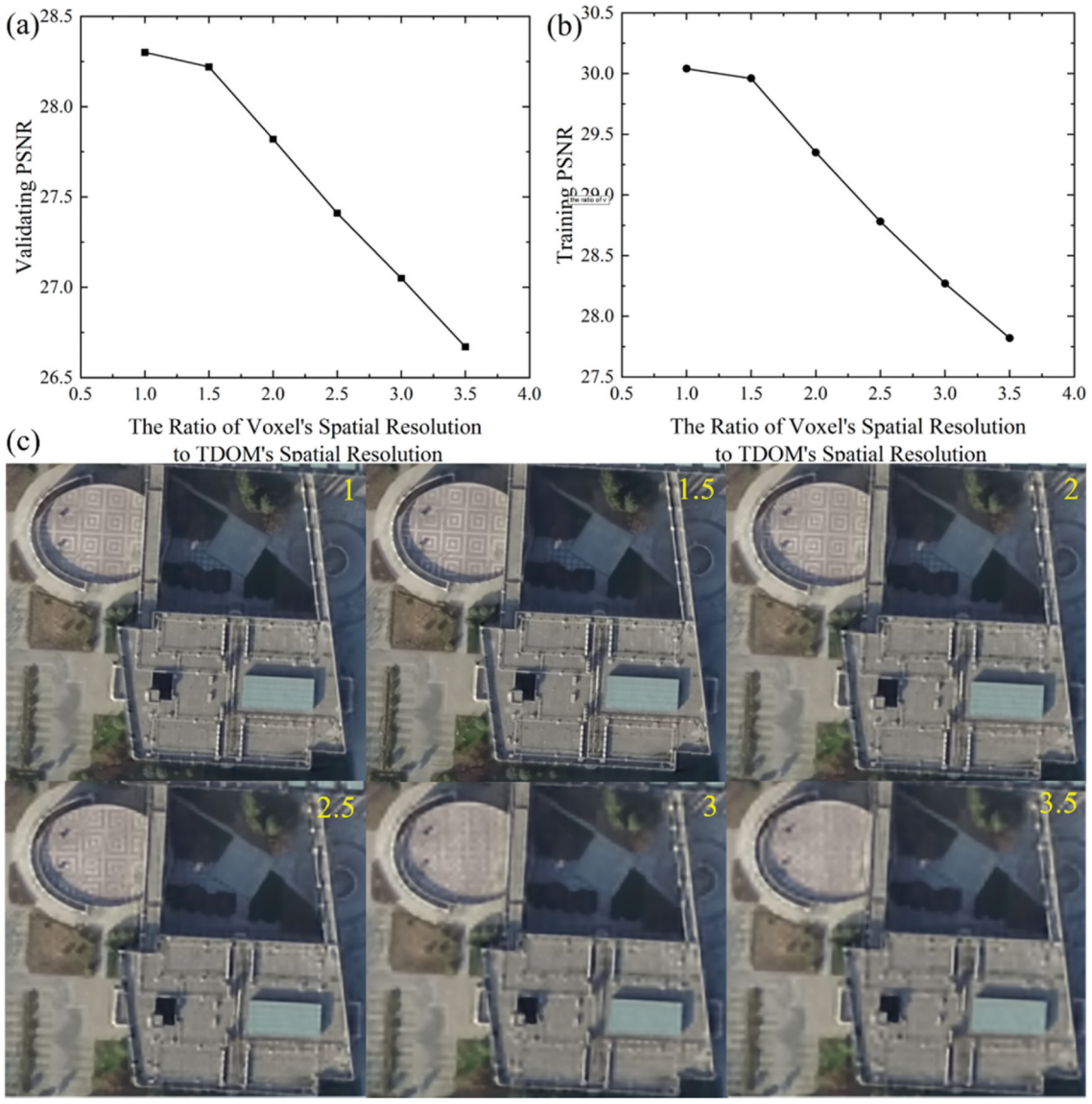

Figure 8. Comparison under different ratios of voxel spatial resolution to TDOM spatial resolution. The validating and training PSNR are shown in (a,b), and the details are sorted by ratios from smallest to largest (c).

with 256 voxels on the *z*-axis, the validating and training PSNR (28.36 and 30.00, respectively) were similar to those under 128 voxels (28.22 and 29.96, respectively). However, these values dropped sharply under fewer voxels, as shown in Figure 9(a,b), with chromatic aberrations appearing in the red box areas of the TDOM, as illustrated in Figure 9(c). This is because the z-axis voxels only recorded information in the vertical direction, influencing the reconstructed quality of information at different elevations, such as building facades. Properly reducing voxels in the z-axis does not drop the quality of the TDOM, but very few voxels cannot cover the scene, leading to a reconstruction failure.

4.4.3. Overlap

In NeRF-based or Plenoxels-based methods, the high overlap of input images is significant. Previous research has often used many images to reconstruct a scene, with an overlap exceeding 90%. Therefore, we explored the effect of overlap on TDOM generation. The number of views is used to express the percentage of forward overlap. For example, 10 views indicate that a position in the world appears in 10 images of a flight line, with a forward overlap of approximately 90%, indicating a large number of views. In Figure 10(a), the validating PSNR peaks at 12 views (28.65) and decreases sharply with the reduction of the number of views, but the training PSNR is increasing in Figure 10(b). This is a manifestation of overfitting, because fewer views in the input mean fewer constraints on the scene, resulting in the model overfitting these views. Too many views also result in poor reconstruction, because even with distortion correction and bundling adjustments, images still have errors compared to a strictly central projection, and too many views add errors. In Figure 10(c), the quality of TDOM can be observed to not obviously decrease until the

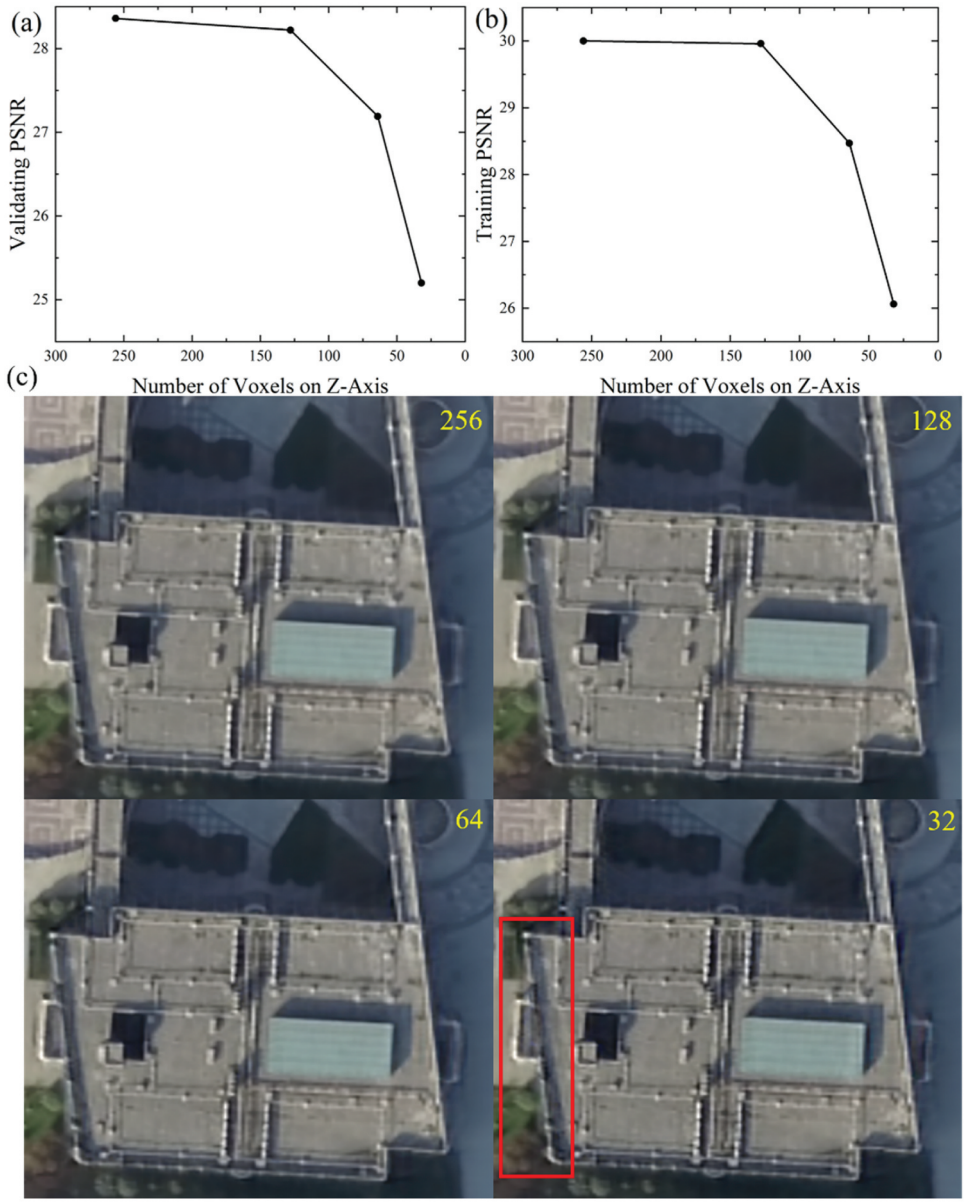

Figure 9. Comparison under a different number of voxels in the z-axis. The validating and training PSNR are shown in (a,b), and the details are sorted by numbers from largest to smallest. The red box highlights area with incorrect or low-quality details.

number of views is eight, at which point radiant ghosting appears around some highlights. To guarantee the quality of the generated TDOM, we recommend forward and side overlaps higher than 90% and 70%, respectively.

Even though Ortho-NeRF can represent high-frequency scenes and accurate geometries for high-quality TDOM rendering as displayed in Figures 6 and 7, this model requires inputs with highly overlapping and similar rendering and training viewing directions. Because our method based on rendering losses is only supervised at known poses, it requires many input views as constraints. In the future, we will improve our work to apply it to oblique photogrammetry and lower overlap datasets. Additionally, the processing time of our method is related to the tile number of a scene. Generating the TDOM of a single tile takes about 12–18 min and approximately 22 G of GPU memory. The computational complexity of our system grows linearly with the addition of tiles. While the efficiency of this approach is considerably improved compared to that of the NeRF, it does not provide a competitive advantage over efficient commercial software owing to the training required for scene reconstruction, which consumes the most time. Even though the training strategy can be adjusted to significantly decrease the processing time while achieving similar accuracy, we still intend to accelerate the training of Ortho-NeRF in a future work.

5. Conclusions

In this study, we presented Ortho-NeRF, a simple learning-based yet expressive method, to generate large-scale, high-quality TDOM from UAV data without prior 3D information. This method splits the scene

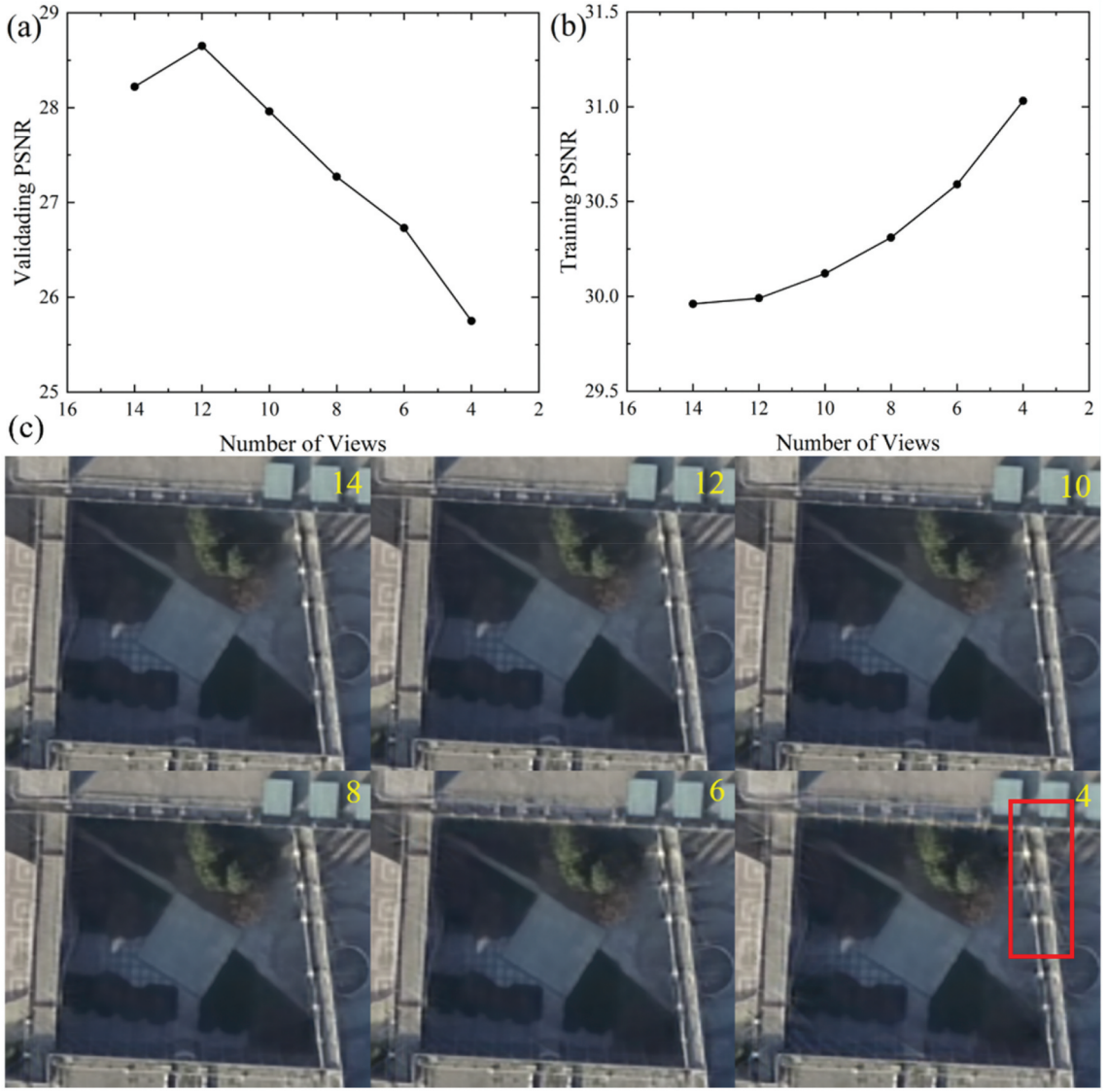

Figure 10. Comparison under different numbers of views. The validating and training PSNR are shown in (a,b), and the details are sorted by views from largest to smallest. The red box highlights area with incorrect or low-quality details.

into small tiles, encodes implicitly from the data of each tile, and renders TDOMs by parallelly projecting them upward. The reconstruction method represents a scene as a 3D grid with SH with five-dimensional (5D) inputs (3D observed locations and 2D viewing directions) and four-dimensional (4D) outputs (volume density and emitted radiance of RGB). As such, our method can quickly restore implicit multi-view consistent scenes used for novel view rendering. The reconstructed geometry was correctly represented, without twisting in all terrain, including weak texture areas like water areas. Some details of the TDOMs demonstrated that our approach can generate high-quality TDOMs. Additionally, we quantitatively demonstrated that our approach achieves precise geometric features. The absolute error of eight GCPs outperformed three mainstream commercial photogrammetry software – ContextCapture, Metashape, and Pix4DMapper – with a mean absolute median error of 0.267 m, a mean absolute horizontal position error of 0.095 m in the x direction and 0.233 m in the y direction The length difference on the same plain between the results of ContextCapture and those of Ortho-NeRF was small. The mean absolute errors were 0.08 m and 0.07 m on phantom3-ieu and gopro-npu, respectively, while the mean relative errors were 0.21% and 0.06%, respectively. Compared with NeRF, the time consumed by training Ortho-NeRF was reduced by 99.05%, and that consumed by rendering was reduced by 99.91%. However, highly overlapping inputs and large training time requirements limit the application of our method. In the future, we will improve Ortho-NeRF to increase its reconstruction efficiency and generalize it to more settings, such as oblique photography and fewer overlapping data.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was funded by the Department of Science and Technology of Hubei Province People's Government [grant number 2022BAA035].

Notes on contributors

Shihan Chen received a master's degree from Wuhan University and is working toward a Ph.D. degree at The Hong Kong Polytechnic University majoring in photogrammetry and remote sensing. His research interests include neural implicit reconstruction, photogrammetry and orthoimage generation.

Qingsong Yan is currently a Ph.D. student at the School of Geodesy and Geomatics, Wuhan University. He received the B.Eng. degree and MSc degree from Wuhan University in 2016 and 2019, respectively. He is majoring in omnidirectional computer vision, 3D reconstruction, multiview-stereo, and neural implicit representations.

Yingjie Qu is studying at Wuhan University as a Ph.D. candidate. His research interests are photogrammetry, remote sensing, and 3D reconstruction from satellite images.

Wang Gao received his master's degree from The Third Research Institute of China Aerospace Science and Industry Corporation in 2017. His research interests are scene matching and visual navigation.

Junxing Yang received his Ph.D. degree from Wuhan University and is currently a lecturer at Beijing University of Civil Engineering and Architecture. He is focused on photogrammetry and remote sensing, 3D reconstruction, image stitching, and scene understanding.

Fei Deng received a Ph.D. degree in geodesy and surveying engineering from Wuhan University in 2006. At present, he is a professor at the School of Geodesy and Geomatics, at Wuhan University. He has been committed to photogrammetry and remote sensing, 3D reconstruction, and scene understanding research.

ORCID

Shihan Chen http://orcid.org/0000-0003-0487-3426
Qingsong Yan http://orcid.org/0000-0002-7095-5844
Yingjie Qu http://orcid.org/0000-0001-5527-6299
Wang Gao http://orcid.org/0000-0002-3827-2111
Junxing Yang http://orcid.org/0000-0003-1893-3274
Fei Deng http://orcid.org/0000-0003-0886-4324

Data availability statement

The NPU DroneMap dataset used in this study is available online (https://pan.baidu.com/s/1bW-4qtNzJzdQAo8QdOGKA?pwd=vaxv; Password: vaxv). Other data cannot be shared at this time, as the data also form part of an ongoing study.

References

Agisoft. 2022. "Discover Intelligent Photogrammetry with Metashape." Accessed September 1, 2022. https://www.agisoft.com/.

Amhar, F., J. Jansa, and C. Ries. 1998. "The Generation of True Orthophotos Using a 3D Building Model in Conjunction with a Conventional DTM." *International Archives of Photogrammetry and Remote Sensing* 32 (4): 16–22.

Bang, K., A. F. Habib, S. Shin, and K. Kim. 2007. "Comparative Analysis of Alternative Methodologies for True Ortho-Photo Generation from High Resolution Satellite Imagery." Paper presented at American Society for Photogrammetry and Remote Sensing Annual Conference, Tampa, Florida, USA, May 7–11.

Barazzetti, L., R. Brumana, D. Oreni, M. Previtali, and F. Roncoroni. 2014. "UAV-Based Orthophoto Generation in Urban Area: The Basilica of Santa Maria Di Collemaggio in L'Aquila." Paper presented at Proceedings of the International Conference on Computational Science and Its Applications. Guimarães, Portugal, June 30–July 3. https://doi.org/10.1007/978-3-319-09147-1_1.

Bentley. 2022. "ContextCapture Viewer." Accessed September 1, 2022. https://www.bentley.com/software/contextcapture-viewer/.

Bhandari, B., U. Oli, U. Pudasaini, and N. Panta. 2015. "Generation of High Resolution DSM Using UAV Images." Paper presented at FIG Working Week—From the Wisdom of the Ages to the Challenges of the Modern World, Sofia, Bulgaria, May 17–21.

Biasion, A., S. Dequal, and A. Lingua. 2004. "A New Procedure for the Automatic Production of True Orthophotos." *The International Archives of Photogrammetry, Remote Sensing & Spatial Information Sciences* 35:1682–1777.

Bu, S., Y. Zhao, G. Wan, and Z. Liu. 2016. "Map2DFusion: Real-Time Incremental UAV Image Mosaicing Based on Monocular SLAM." Paper presented at 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2016, 4564–4571. Daejeon, South Korea, October 9–14. https://doi.org/10.1109/IROS.2016.7759672.

Chen, G., S. Chen, X. Li, P. Zhou, and Z. Zhou. 2018. "Optimal Seamline Detection for Orthoimage Mosaicking Based on DSM and Improved JPS Algorithm." *Remote Sensing* 10 (6): 821. https://doi.org/10.3390/rs10060821.

Chen, L., T. Teo, J. Wen, and J. Rau. 2007. "Occlusion-Compensated True Orthorectification for High-Resolution Satellite Images." *Photogrammetric Record* 22 (117): 39–52. https://doi.org/10.1111/j.1477-9730.2007.00416.x.

Chen, Z., and H. Zhang. 2019. "Learning Implicit Fields for Generative Shape Modeling." Paper presented at 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5932–5941. Long Beach, CA, USA. https://doi.org/10.1109/LGRS.2015.2459671.

Cheng, Z., X. Huang, D. Li, and H. Li. 2010. "Polygon Based Inversion Imaging for Occlusion Detection in True Orthophoto Generation." *Journal of Geodesy and Geoinformation Science* 39 (1): 59–64.

Colomina, I., and P. Molina. 2014. "Unmanned Aerial Systems for Photogrammetry and Remote Sensing: A Review." *ISPRS Journal of Photogrammetry and Remote Sensing* 92:79–97. https://doi.org/10.1016/j.isprsjprs.2014.02.013.

Deng, F., P. Li, Y. Kan, J. Kang, and F. Wan. 2017. "Overall Projection of DBM for Occlusion Detection in True Orthophoto Generation." *Geomatics and Information Science of Wuhan University* 42 (1): 97–102. https://doi.org/10.13203/j.whugis20140660.

de Oliveira, H. C., M. Galo, and A. P. D. Poz. 2015. "Height-Gradient-Based Method for Occlusion Detection in True Orthophoto Generation." *IEEE Geoscience & Remote Sensing Letters* 12 (11): 2222–2226. https://doi.org/10.1109/LGRS.2015.2459671.

de Oliveira, H. C., A. P. D. Poz, M. Galo, and A. F. Habib. 2018. "Surface Gradient Approach for Occlusion Detection Based on Triangulated Irregular Network for True Orthophoto Generation." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11 (2): 443–457. https://doi.org/10.1109/JSTARS.2017.2786162.

Ebrahimikia, M., and A. Hosseininaveh. 2022. "True Orthophoto Generation Based on Unmanned Aerial Vehicle Images Using Reconstructed Edge Points." *Photogrammetric Record* 37 (178): 161–184. https://doi.org/10.1111/phor.12409.

Everaerts, J. 2008. "The Use of Unmanned Aerial Vehicles (UAVs) for Remote Sensing and Mapping." *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 37 (B1): 1187–1192.

Fridovich-Keil, S., A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa. 2022. "Plenoxels: Radiance Fields without Neural Networks." Paper presented at Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5501–5510. New Orleans, LA, USA, June 18–24. https://doi.org/10.1109/CVPR52688.2022.00542.

Garbin, S. J., M. Kowalski, M. Johnson, J. Shotton, and J. Valentin. 2021. "Fastnerf: High-Fidelity Neural Rendering at 200fps." Paper presented at Proceedings of the IEEE/CVF International Conference on Computer Vision, 14346–14355. Virtual, October 11–17 https://doi.org/10.1109/ICCV48922.2021.01408.

Gharibi, H., and A. Habib. 2018. "True Orthophoto Generation from Aerial Frame Images and LiDar Data: An Update." *Remote Sensing* 10 (4): 581. https://doi.org/10.3390/rs10040581.

Gilani, S. A. N., M. Awrangjeb, and G. Lu. 2016. "An Automatic Building Extraction and Regularisation Technique Using Lidar Point Cloud Data and Orthoimage." *Remote Sensing* 8 (3): 258. https://doi.org/10.3390/rs8030258.

Haarbrink, R. B., and H. Eisenbeiss. 2008. "Accurate DSM Production from Unmanned Helicopter Systems." *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 37 (B1): 1259–1264. https://doi.org/10.3929/ethz-b-000011976.

Habib, A. F., E. Kim, and C. Kim. 2007. "New Methodologies for True Orthophoto Generation." *Photogrammetric Engineering & Remote Sensing* 73 (1): 25–36. https://doi.org/10.14358/PERS.73.1.25.

Harwin, S., and A. Lucieer. 2012. "Assessing the Accuracy of Georeferenced Point Clouds Produced via Multi-View Stereopsis from Unmanned Aerial Vehicle (UAV) Imagery." *Remote Sensing* 4 (6): 1573–1599. https://doi.org/10.3390/rs4061573.

Hu, T., S. Liu, Y. Chen, T. Shen, and J. Jia. 2022. "EfficientNeRF Efficient Neural Radiance Fields." Paper presented at Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12902–12911. New Orleans, LA, USA, June 19–24. https://doi.org/10.1109/CVPR52688.2022.01256.

Hu, Y., D. Stanley, and Y. Xin. 2016. "True Ortho Generation of Urban Area Using High Resolution Aerial Photos." *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences* 3 (4): 3–10. https://doi.org/10.5194/isprs-annals-III-4-3-2016.

Karnewar, A., T. Ritschel, O. Wang, and N. Mitra. 2022. "ReLu Fields: The Little Non-Linearity That Could." Paper presented at ACM SIGGRAPH 2022 Conference Proceedings, 1–9. Vancouver, BC, Canada, August 7–11. https://doi.org/10.1145/3528233.3530707.

Kerbl, B., G. Kopanas, T. Leimk¨uhler, and G. Drettakis. 2023. "3D Gaussian Splatting for Real-Time Radiance Field Rendering." *ACM Transactions on Graphics* 42 (4): 1–14. https://doi.org/10.1145/3592433.

Kwak, E., and A. Habib. 2014. "Automatic Representation and Reconstruction of DBM from LiDar Data Using Recursive Minimum Bounding Rectangle." *ISPRS Journal of Photogrammetry and Remote Sensing* 93:171–191. https://doi.org/10.1016/j.isprsjprs.2013.10.003.

Liu, Y., X. Zheng, G. Ai, Y. Zhang, and Y. Zuo. 2018. "Generating a High-Precision True Digital Orthophoto Map Based on UAV Images." *ISPRS International Journal of Geo-Information* 7 (9): 333. https://doi.org/10.3390/ijgi7090333.

Ma, L., Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson. 2019. "Deep Learning in Remote Sensing Applications: A Meta-Analysis and Review." *ISPRS Journal of Photogrammetry and Remote Sensing* 152:166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015.

Martin-Brualla, R., N. Radwan, M. S. M. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth. 2021. "Nerf in the Wild: Neural Radiance Fields for Unconstrained Photo Collections." Paper presented at Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7210–7219. Nashville, TN, USA, June 20–25. https://doi.org/10.1109/CVPR46437.2021.00713.

Mescheder, L., M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. 2019. "Occupancy Networks: Learning 3d Reconstruction in Function Space." Paper presented at Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4460–4470. Long Beach, CA, USA, June 15–20. https://doi.org/10.1109/CVPR.2019.00459.

Mildenhall, B., P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. 2020. "Nerf: Representing Scenes as Neural Radiance Fields for View Synthesis." Paper presented at European Conference on Computer Vision, 405–421. Glasgow, UK, August 23–28. https://doi.org/10.1007/978-3-030-58452-8_24.

Müller, T., A. Evans, C. Schied, and A. Keller. 2022. "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding." *ACM Transactions on Graphics* 41 (4): 102. https://doi.org/10.1145/3528223.3530127.

Nex, F., and F. Remondino. 2014. "UAV for 3D Mapping Applications: A Review." *Applied Geomatics* 6 (1): 1–15. https://doi.org/10.1007/s12518-013-0120-x.

Niemeyer, M., and A. Geiger. 2021. "Giraffe: Representing Scenes as Compositional Generative Neural Feature Fields." Paper presented at Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11453–11464. Nashville, TN, USA, June 20–25. https://doi.org/10.1109/CVPR46437.2021.01129.

Oda, K., W. Lu, O. Uchida, and T. Doihara. 2004. "Triangle-Based Visibility Analysis and True Orthoimage Generation." *The International Archives of Photogrammetry, Remote Sensing & Spatial Information Sciences* 35 (Part 3): 623–628.

Park, J. J., P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. 2019. "Deepsdf: Learning Continuous

Signed Distance Functions for Shape Representation." Paper presented at Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 165–174. Long Beach, CA, USA, June 15–20. https://doi.org/10.1109/CVPR.2019.00025.

PIX4D. 2022. "PIX4Dmapper." Accessed September 1, 2022. https://www.pix4d.com/product/pix4dmapper-photogrammetry-software/.

Popescu, G., D. Iordan, and V. Păunescu. 2016. "The Resultant Positional Accuracy for the Orthophotos Obtained with Unmanned Aerial Vehicles (UAVs)." *Agriculture and Agricultural Science Procedia* 10:458–464. https://doi.org/10.1016/j.aaspro.2016.09.016 .

Qin, Z., W. Li, M. Li, Z. Chen, and G. Zhou. 2003. "A Methodology for True Orthorectification of Large-Scale Urban Aerial Images and Automatic Detection of Building Occlusions Using Digital Surface Model." Paper presented at IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium Proceedings, 729–731. Toulouse, France, July 21–25. https://doi.org/10.1109/IGARSS.2003.1293898.

Rau, J.-Y., N.-Y. Chen, and L.-C. Chen. 2002. "True Orthophoto Generation of Built-Up Areas Using Multi-View Images." *Photogrammetric Engineering & Remote Sensing* 68 (6): 581–588.

Reiser, C., R. Szeliski, D. Verbin, P. Srinivasan, B. Mildenhall, A. Geiger, J. Barron, and P. Hedman. 2023. "MERF: Memory-Efficient Radiance Fields for Real-Time View Synthesis in Unbounded Scenes." *ACM Transactions on Graphics* 42 (4): 89. https://doi.org/10.1145/3592426.

Ruzgiene, B., T. Berteska, S. Gecyte, E. Jakubauskiene, and V. Aksamitauskas. 2014. "Photogrammetric Processing of UAV Imagery: Checking DTM." Paper presented at ICEE Proceedings of the International Conference on Environmental Engineering. Vilnius, Lithuania, May 22–23. https://doi.org/10.3846/enviro.2014.242.

Schickier, W., and A. Thorpe. 1998. "Operational Procedure for Automatic True Orthophoto Generation." *International Archives of Photogrammetry and Remote Sensing* 32 (4): 527–532.

Schwarz, K., Y. Liao, M. Niemeyer, and A. Geiger. 2020. "Graf: Generative Radiance Fields for 3d-Aware Image Synthesis." *Advances in Neural Information Processing Systems* 33:20154–20166. https://doi.org/10.48550/arXiv.2007.02442 .

Shao, Z., G. Cheng, D. Li, X. Huang, Z. Lu, and J. Liu. 2021. "Spatio-Temporal-Spectral-Angular Observation Model That Integrates Observations from UAV and Mobile Mapping Vehicle for Better Urban Mapping." *Geo-Spatial Information Science* 24 (4): 615–629. https://doi.org/10.1080/10095020.2021.1961567.

Sheng, Y. 2007. "Minimising Algorithm-Induced Artefacts in True Ortho-Image Generation: A Direct Method Implemented in the Vector Domain." *Photogrammetric Record* 22 (118): 151–163. https://doi.org/10.1111/j.1477-9730.2007.00425.x.

Shin, Y. H., S. W. Hyung, and D. Lee. 2020. "True Orthoimage Generation from LiDar Intensity Using Deep Learning." *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography* 38 (4): 363–373.

Shin, Y. H., D. Lee, and H.-S. Jung. 2021. "True Orthoimage Generation Using Airborne Lidar Data with Generative Adversarial Network-Based Deep Learning Model." *Journal of Sensors* 2021:1–25. https://doi.org/10.1155/2021/4304548.

Sitzmann, V., J. N. P. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein. 2020. "Implicit Neural Representations with Periodic Activation Functions." *Advances in Neural Information Processing Systems* 33:7462–7473. https://doi.org/10.48550/arXiv.2006.09661.

Slonecker, E. T., B. Johnson, and J. McMahon. 2009. "Automated Imagery Orthorectification Pilot." *Journal of Applied Remote Sensing* 3 (1): 033552. https://doi.org/10.1117/1.3255042.

Sun, C., M. Sun, and H. Chen. 2022. "Direct Voxel Grid Optimization: Super-Fast Convergence for Radiance Fields Reconstruction." Paper presented at Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5459–5469. New Orleans, LA, USA, June 18–24. https://doi.org/10.1109/CVPR52688.2022.00538.

Sutherland, I. E., R. F. Sproull, and R. A. Schumacker. 1974. "A Characterization of Ten Hidden-Surface Algorithms." *ACM Computing Surveys* 6 (1): 1–55. https://doi.org/10.1145/356625.356626.

Tewari, A., J. Thies, B. Mildenhall, P. Srinivasan, E. Tretschk, W. Yifan, C. Lassner, et al. 2022. "Advances in Neural Rendering." *Paper Presented at Computer Graphics Forum* 41 (2): 703–735. https://doi.org/10.1111/cgf.14507.

Toth, C., and G. Jóźków. 2016. "Remote Sensing Platforms and Sensors: A Survey." *ISPRS Journal of Photogrammetry and Remote Sensing* 115:22–36. https://doi.org/10.1016/j.isprsjprs.2015.10.004.

Uysal, M., A. S. Toprak, and N. Polat. 2015. "DEM Generation with UAV Photogrammetry and Accuracy Analysis in Sahitler Hill." *Measurement* 73:539–543. https://doi.org/10.1016/j.measurement.2015.06.010.

Wang, C., and Y. Gu. 2022. "A Method for Generating True Digital Orthophoto Map of UAV Platform Push-Broom Hyperspectral Scanners Assisted by Lidar." IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, 7523–7526. Kuala Lumpur, Malaysia, July 17–22. https://doi.org/10.1109/IGARSS46834.2022.9884838.

Wang, Q., L. Yan, Y. Sun, X. Cui, H. Mortimer, and Y. Li. 2018. "True Orthophoto Generation Using Line Segment Matches." *Photogrammetric Record* 33 (161): 113–130. https://doi.org/10.1111/phor.12229.

Wang, X., W. Jiang, and J. Xie. 2009. "A New Method for True Orthophoto Generation." *Geomatics and Information Science of Wuhan University* 34 (10): 1250–1254.

Wang, Z., S. Wu, W. Xie, M. Chen, and V. A. Prisacariu. 2021. "NeRf–: Neural Radiance Fields without Known Camera Parameters." *ArXiv Preprint. https://doi.org/10.48550/arXiv.2102.07064*.

Wolf, P. R., B. A. Dewitt, and B. E. Wilkinson. 2014. *Elements of Photogrammetry with Applications in GIS*. New York: McGraw-Hill Education.

Xu, Q., Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, and U. Neumann. 2022. "Point-Nerf: Point-Based Neural Radiance Fields." Paper presented at Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5438–5448. New Orleans, LA, USA, June 18–24. https://doi.org/10.1109/CVPR52688.2022.00536.

Yang, B., N. Haala, and Z. Dong. 2023. "Progress and Perspectives of Point Cloud Intelligence." *Geo-Spatial Information Science* 26 (2): 189–205. https://doi.org/10.1080/10095020.2023.2175478.

Yang, J., L. Liu, J. Xu, Y. Wang, and F. Deng. 2021. "Efficient Global Color Correction for Large-Scale Multiple-View Images in Three-Dimensional Reconstruction." *ISPRS*

Journal of Photogrammetry and Remote Sensing 173:209–220. https://doi.org/10.1016/j.isprsjprs.2020.12.011.

Yu, A., R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa. 2021. "PlenOctrees for Real-Time Rendering of Neural Radiance Fields." Paper presented at Proceedings of the IEEE/CVF International Conference on Computer Vision, 5752–5761. Montreal, QC, Canada, October 10–17. https://doi.org/10.1109/ICCV48922.2021.00570.

Yuan, W., X. Yuan, Y. Cai, and R. Shibasaki. 2023. "Fully Automatic DOM Generation Method Based on Optical Flow Field Dense Image Matching." *Geo-Spatial Information Science* 26 (2): 242–256. https://doi.org/10.1080/10095020.2022.2159886.

Zhang, J., B. Xu, M. Sun, and Y. Zhang. 2012. "True Orthoimage Generation Based on Occlusion Detection with TIN." *Geomatics and Information Science of Wuhan University* 37 (3): 326–329. https://doi.org/10.13203/j.whugis2012.03.020.

Zhang, L., L. Zhang, and B. Du. 2016. "Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art." *IEEE Geoscience and Remote Sensing Magazine* 4 (2): 22–40. https://doi.org/10.1109/MGRS.2016.2540798.

Zhong, C., X. Huang, D. Li, and H. Li. 2010. "Polygon Based Inversion Imaging for Occlusion Detection in True Orthophoto Generation." *Journal of Geodesy and Geoinformation Science* 39 (1): 59–64.

Zhou, G. 2020. *Urban High-Resolution Remote Sensing: Algorithms and Modeling*. Boca Raton: CRC Press. https://doi.org/10.1201/9781003082439.

Zhou, G., W. Chen, J. A. Kelmelis, and D. Zhang. 2005. "A Comprehensive Study on Urban True Orthorectification." *IEEE Transactions on Geoscience and Remote Sensing* 43 (9): 2138–2147. https://doi.org/10.1109/TGRS.2005.848417.

Zhou, G., Q. Wang, Y. Huang, J. Tian, H. Li, and Y. Wang. 2022. "True2 Orthoimage Map Generation." *Remote Sensing* 14 (17): 4396. https://doi.org/10.3390/rs14174396.

Zhu, X., D. Tuia, L. Mou, G. Xia, L. Zhang, F. Xu, and F. Fraundorfer. 2017. "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources." *IEEE Geoscience and Remote Sensing Magazine* 5 (4): 8–36. https://doi.org/10.1109/MGRS.2017.2762307.