Sensors and Materials, Vol. 36, No. 12 (2024) 5477–5489
MYU Tokyo

S & M 3880

Generating Digital Elevation Models from Satellite Imagery Using Neural Radiance Field

Tianjiao Wang, Junxing Yang, Tong Ye, and He Huang*

School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 102616, China

(Received August 26, 2024; accepted December 19, 2024)

Keywords: satellite 3D reconstruction, photogrammetry, neural radiance fields, DEM

As high-resolution satellite remote sensing images become essential tools for understanding geospatial information, the large-scale 3D reconstruction of Earth's surface using these images has emerged as a significant research area in computer vision, photogrammetry, and remote sensing. However, satellite-based 3D reconstruction is highly sensitive to image changes arising from the multitemporal acquisition of images. These changes are primarily caused by varying shadows, reflections, and transient objects (e.g., vegetation), which complicate accurate modeling. Neural radiance fields (NeRFs), utilizing differentiable rendering to learn implicit scene representations, offer a novel approach to generating 4D products from multiview images without requiring additional data, gaining considerable attention in 3D scene reconstruction and rendering. Building on this, we propose a method for generating digital elevation models (DEMs) from satellite images, leveraging NeRF to create 3D scenes from a set of images captured at different times while addressing the challenges posed by lighting variations and transient objects. Our experiments demonstrate that our approach can generate high-quality DEMs and corresponding mesh models, outperforming both traditional and recent methods in qualitative and quantitative evaluations.

1. Introduction

Reconstructing 3D models from multiview images is an important research direction in computer vision and photogrammetry.[1–3] This technique involves densely matching correspondences from images taken from different viewpoints to generate 3D point clouds and reconstruct the 3D surface of the model. Compared with laser imaging detection and ranging (LiDAR) technology, multiview image-based 3D reconstruction offers significant advantages such as lower cost, faster update, higher resolution, and broader mapping coverage.[4] It is suitable for large-scale 3D reconstruction and virtual reality applications. Digital elevation models (DEMs), as a crucial component of products derived from images, are widely used in scientific research and engineering applications.[5–8] They are an important source of data for the study and analysis of terrain, watershed, and object identification. Since DEM data can reflect

*Corresponding author: e-mail: huanghe@bucea.edu.cn
https://doi.org/10.18494/SAM5345

ISSN 0914-4935

local terrain features with a certain resolution, a large amount of surface morphological information can be extracted through DEMs. This information can be used to draw contour lines, slope maps, aspect maps, stereo perspective maps, and stereo landscape maps, and can be utilized to create orthophotos, stereo terrain models, and map revisions. However, existing DEM data often have lower resolution, making it difficult to accurately represent terrain features, particularly in high-precision areas. Low-altitude unmanned aerial vehicles (UAVs) provide an efficient and cost-effective remote sensing method, capable of quickly acquiring large-scale products for survey areas.[9] This technology provides real-time, accurate data for environmental safety and development applications, contributing to the green development of national economies. Despite the impressive performance of UAVs in high-precision 3D reconstruction, particularly in localized areas such as urban infrastructure and natural resource management, their application is still limited by factors such as airspace restrictions and high operational costs.

In contrast, satellite remote sensing can quickly and efficiently collect geographic spatial data across global scales, making it a vital tool for acquiring and understanding geospatial information. Compared with UAV images, satellite imagery can cover hundreds of kilometers of ground area in a single pass. With advancements in Earth observation technology, commercial satellites now achieve ground resolutions as high as 0.3 m, clearly displaying features such as buildings, bridges, and aircraft. Satellite imagery is increasingly becoming a crucial tool for Earth observation and offers new pathways for large-scale realistic 3D reconstruction. Its global coverage and long-term monitoring capabilities provide an economical and efficient solution for 3D reconstruction over extensive areas. However, since satellite remote sensing images capture dynamic scenes, variations in lighting, shadows, surface features, and seasonal changes add complexity to model processing. Traditional structure from motion or multiview stereo matching methods for 3D reconstruction involve recovering sparse point clouds from multiple images, followed by point cloud registration, dense point cloud recovery, outlier processing, and final point cloud reconstruction. This process is lengthy, with each step potentially affected by errors.[10]

In recent years, neural rendering technologies have made significant advancements. These technologies use multilayer perceptron (MLP) parameterized continuous volume functions to encode 3D scenes, representing space as implicit radiance fields and performing volumetric rendering from multiview images to regress density and color. For neural radiance fields (NeRFs), it is assumed that the scene is static in terms of geometry, material, and camera angle, i.e., the density and radiation field of the scene are static. Therefore, NeRFs require that two photos taken at the same position and orientation must be exactly the same. Since it is almost impossible to have satellite images with the same shadows, there are some problems with the original NeRF in generating DEMs based on satellite images. Many variants have been proposed to address this problem. NeRF-W gains robustness to radiometric variation and transient objects by separating transient phenomena from the static scene.[11–14] An extra head of fully connected layers is used to predict a transient color c_τ and volume density σ_τ for each input point, in addition to the usual c and σ. The transients are linearly combined with the static ones to render the scene. Shadow-NeRF (S-NeRF)[15] is the first attempt to apply NeRF to multiview satellite

photogrammetry. It uses solar angles to learn the amount of light reaching each point in the scene, enabling a more reliable modeling of shadow areas than with the model alone. Sat-NeRF[16] learns the transients (e.g., cars, etc.) present in each view with a similar approach to NeRF-W,[17] which introduces a coefficient. This coefficient predicts for each point whether that point corresponds to a transient object or not. These techniques have been successfully applied to satellite photogrammetry with impressive results, primarily focusing on novel view synthesis.

In this paper, we propose an enhanced method based on the Sat-NeRF model to generate 3D models from a set of multiview satellite images of a scene. The goal is to generate a DEM and the corresponding 3D mesh model of the surface. This enhanced method is expected to broaden the application scope of satellite remote sensing data in 3D reconstruction and provide a new approach to large-scale 3D reconstruction.

2. Methods

2.1 NeRF

When discussing 3D rendering technology, NeRF marks a significant leap. This innovative method leverages the mapping function f that transforms the 3D spatial location x and the viewing direction d into the volume density σ and the color value c, respectively. MLPs face an inherent challenge in effectively mapping low-frequency signals, prompting NeRF to employ positional encoding. The method for MLPs can be formulated as

$$\mathrm{enc}(x, L) = (\sin(2^0\pi x), \cos(2^0\pi x), \ldots, \sin(2^L\pi x), \cos(2^L\pi x)), \quad (1)$$

where f is a neural network comprising eight perceptron (MLP) layers parameterized by θ, denoted as f_θ: $(enc(x), enc(d)) \rightarrow (\sigma, c)$, where $enc()$ represents a positional encoding. The expected pixel color $\hat{C}(r)$ is obtained by casting a ray $r(t)=o+td$, where o is the origin of the ray, d is the direction of the ray, and t is the parameter along the ray. The ray is constrained by near and far bounds t_n and t_f, which define the valid range of t values along the ray. We evenly partition $[t_n, t_f]$ into N points $(t_1, t_2, \ldots, t_N)$ along a ray r and compute the expected pixel color $\hat{C}(r)$ as $\hat{C}(r) = \frac{1}{N} \sum_{i=1}^{N} C_i^*$. The weighted color C_i^* of a 3D point is computed as $C_i^* = w_i c_i$, where $w_i = T_i(1-\exp(-\sigma_i\delta_i))$, $T_i = \exp(-\sum_{j=1}^{i-1}\sigma i\delta i)$, and $\delta_i = t_i - t_{i-1}$. Therefore, the NeRF reconstruction loss can be formulated as

$$L_{NeRF} = \|(\hat{C}(r), C(r))\|^2, \quad (2)$$

where $\hat{C}(r)$ represents colors blended by N samples and $C(r)$ represents the ground-truth pixel color. We utilize coarse-to-fine sampling as discussed for the original NeRF. Here, we omit the detailed rendering for simplification.

2.2 Sat-NeRF

NeRF assumes that the scene is static in terms of geometry, material, and camera angle, i.e., the density and radiance field of the scene are static. Therefore, NeRF requires that two photos taken from the same position and orientation must be exactly the same. Since it is almost impossible to have satellite images with identical shadows, a more flexible model is needed. Unlike the original NeRF, Sat-NeRF revises the color dependence on viewing angles. By combining neural rendering with native satellite camera models, Sat-NeRF uses the shadow-aware irradiance model proposed for S-NeRF to compute the color of each point on the ray. To account for transient phenomena in the input image, Sat-NeRF adopts a method similar to NeRF-W to learn a transient embedding vector specific to each image by learning an uncertainty image as a network output, based on a latent time vector, to constrain the loss to focus on areas without transient objects.

The inputs are as follows: x is a 3D vector representing the spatial coordinates of points located in the volume. ω indicates the viewer's position with respect to the sun in the satellite metadata, represented as a 3D direction vector encoding the direction of solar rays. For each input image, ω is extracted from the azimuth and elevation (θ, ϕ) that indicate the position of the sun in the satellite image metadata. t_j is the $N(t)$-dimensional embedding vector, learned as a function of the image index j. The objective of t_j is to capture the transient elements in the j-th view that cannot be explained by the sun's given position. We manually set $N(t) = 4$, so that the volumetric function of Sat-NeRF can be expressed as $f_{\theta:}(x, \omega, t_j) \rightarrow (\sigma, c_a, \mathrm{S}, a, \beta)$, where the outputs are defined as follows:

σ: scalar encoding of the volume density at location x;
c_a: albedo RGB color, which depends exclusively on the geometry, i.e., the spatial coordinates x;
S: shadow-aware shading scalar, learned as a function of x and the solar ray direction vector ω;
a: ambient RGB color, independent of scene geometry, which defines a global hue bias based on the sun's position as given by ω;
β: uncertainty coefficient related to the probability that the color of x is explained by a transient object.

Thus, Sat-NeRF effectively handles appearance changes caused by shadows and transient objects, achieving high-quality 3D models and view synthesis.

2.3 Loss function

The loss function of Sat-NeRF differs from that of the traditional NeRF and primarily consists of three components: uncertainty for transient objects, a solar correction term, and a depth supervision term.

The first component is uncertainty for transient objects. Transient objects refer to local features that vary between multiple images, such as vehicles. The position, number, and type of vehicles in the same area may change at different times of capture, leading to changes in image grayscale. These changes cannot be explained by the position of the sun or surface albedo, necessitating the introduction of an additional variable, β, representing the uncertainty of the

transient object. With the introduction of β, the loss function described in Eq. (2) becomes Eq. (3). In the equation $\beta(r) = \beta(r) + \beta(min)$, where $\beta(min)$ is set to 0.05, μ is set to 3. The logarithm in the formula is taken to prevent the algorithm from optimizing β to infinity. As described in Eq. (4), the uncertainty of the transient object along a ray r is calculated as the integral of the value at each point on the ray. During actual training, Eq. (2) is used to compute the loss in the first two rounds, and β is introduced from the third round onwards. This approach is taken to prevent the algorithm from mistakenly identifying shadows as transient objects, which would hinder accurate learning.

$$L_{RGB}(R) = \sum_{r \in R} \frac{\left\| C(r) - C_{gt}(r) \right\|}{2\beta'(r)^2} + \left(\frac{\log \beta'(r) + \mu}{2} \right) \tag{3}$$

$$\beta(r) = \sum_{1}^{N} T_i \alpha_i \beta\left(x_i t_j \right) \tag{4}$$

The second component is the solar correction term. For sunlight ω that is not represented in the training data, the algorithm may generate unrealistic shadow grayscale estimates. Therefore, in addition to the color loss described in Eqs. (2) and (3), the algorithm also incorporates a solar correction term, as shown in Eq. (5).

$$L_{SC}(R_{SC}) = \sum_{r \in R_{SC}} \sum_{i=1}^{N_{SC}} \left(T_i - s_i \right)^2 + 1 - \sum_{i=1}^{N_{SC}} T_i \alpha_i s_I \tag{5}$$

The third component is the depth supervision term. This term involves identifying key points using the scale-invariant feature transform algorithm and optimizing the rational polynomial coeffcients parameters through bundle adjustment at these points to enhance the scene rendering performance of the algorithm. Additionally, these key points are utilized to assist in training neural networks by providing true depth information. Previous studies have demonstrated that incorporating some 3D points with known coordinates as depth ground truth data can improve NeRF's performance. Consequently, in this paper, we introduce a depth supervision loss, as described in Eq. (6). In the equation, $x(r)$ represents a 3D point with known coordinates, $d(r)$ denotes the estimated depth value of a ray r, and $w(r)$ is the contribution weight of $x(r)$ to the depth supervision information, which is calculated on the basis of the reprojection error at point $x(r)$ during the adjustment process.

$$L_{DS}(R_{DS}) = \sum_{r \in R_{DS}} \omega(r) \left(d(r) - \left\| x(r) - o(r) \right\|_2 \right)^2 \tag{6}$$

Thus, the term of the Sat-NeRF loss function can be expressed as

$$L = L_{RGB}(R) + \lambda_{SC} L_{SC}(R_{SC}) + \lambda_{DS} L_{DS}(R_{DS}),$$

where λ_{SC} and λ_{DS} are weights assigned to each secondary term. In experiments, we chose λ_{SC} = 0.1/3 and λ_{DS} =1000/3 to provide good results, ensuring that the secondary terms are sufficiently relevant but remain below the magnitude of L_{RGB}. For depth supervision, we used approximately 2k–10k bundle adjustment points for each area of interest. R, R_{SC}, and R_{DS} have the same batch size.

2.4 Architecture

Our model is based on the Sat-NeRF architecture. In Sat-NeRF, the primary block consists of fully connected layers, each with h channels, dedicated to predicting the static properties of the scene: the volume density σ and the albedo color c_a. A secondary head, comprising fewer layers and half the number of channels per layer, estimates the shading scalar s on the basis of the direction of solar rays, ω, and the geometry-related features learned by the primary block. Additionally, two single-layer heads are employed to predict the uncertainty coefficient β and the ambient color a, on the basis of the transient embedding vector t_j and ω, respectively. For our implementation, we set h = 512.

After constructing the Sat-NeRF, we convert the ray data and depth information into geographic coordinates. These geographic coordinates are subsequently transformed into the Universal Transverse Mercator (UTM) coordinate system, specifically into the easting coordinate.

We then integrate the UTM coordinates with elevation data to generate a point cloud, with each point representing a location on the ground. To ensure the accuracy of the point cloud model, we filter out ground points by setting an appropriate elevation threshold; points with elevations below this threshold are classified as ground points. Subsequently, we read parameters from a file (such as offset, size, and resolution) to define the extent of the DEM. Finally, we apply interpolation and smoothing to the filtered point cloud model to produce the DEM and the mesh model with the specified resolution and size. See Fig. 1 for details.

3. Experimental Datasets and Analysis of Results

All experiments were conducted on an RTX A5000 GPU, with a batch size of 4096 rays to optimize GPU memory usage. The learning rate started at 0.001 and was adjusted dynamically during the training process on the basis of the model's convergence behavior. The loss function is described in Sect. 2.3. In all experiments, a single NeRF model was used, trained with the Adam optimizer starting with a learning rate of 5×10^{-4}, which is decreased at every epoch by a factor of γ = 0.9 according to a step scheduler. The batch size is 1024 rays, and each ray r is discretized into 64 uniformly distributed 3D points. The loss function used is consistent with that implemented in the Sat-NeRF framework.

3.1 Dataset preparation

The experiments were conducted using the Data Fusion Contest (DFC2019) dataset, which includes features of Jacksonville (JAX) obtained from satellite images captured by WorldView-3, with a resolution of 0.3 m per pixel, over the course of a year.

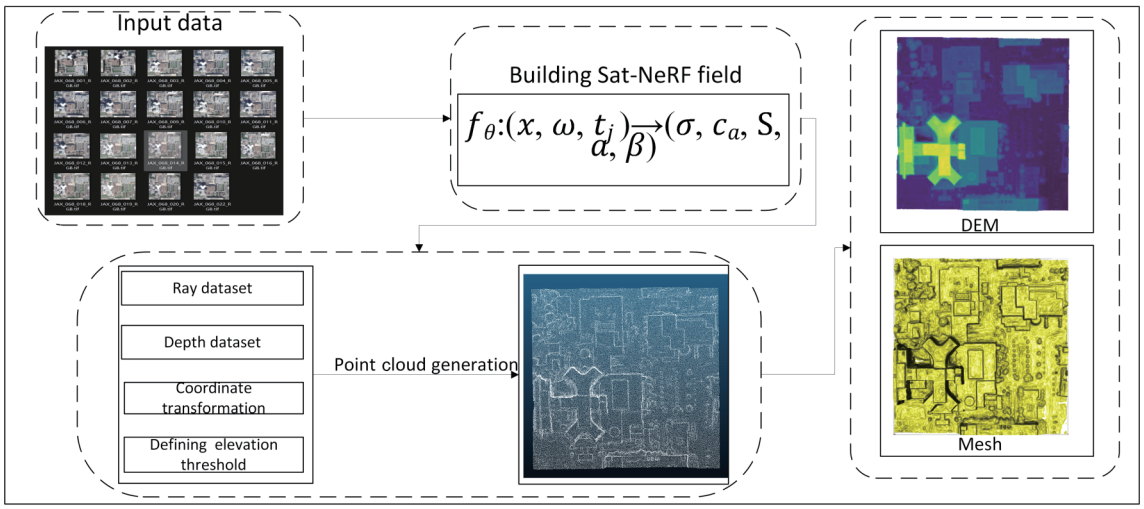

Fig. 1. (Color online) The flowchart of our method is as follows. First, we construct the Sat-NeRF field on the basis of satellite imagery. Then, we convert the ray data and depth information coordinates into geographic coordinates. After setting an appropriate threshold, we generate a point cloud. Finally, we apply interpolation and smoothing to the filtered point cloud model to produce the DEM and the mesh model with the specified resolution and size.

For the qualitative analysis, we compared our method with traditional photogrammetry and the latest research techniques, such as NeRF and S-NeRF. Traditional photogrammetry involves several steps to generate a 3D model or DEM. Initially, multiple images with a specific degree of overlap are imported and preprocessed. The software automatically extracts feature points from the images, estimates the camera poses by matching these points, and generates a sparse point cloud. By utilizing depth information from the multiview images, the software then creates a dense point cloud and reconstructs the surface to produce a 3D mesh model. Given that traditional photogrammetry is a well-established technique, we selected the mainstream commercial software Metashape (formerly known as PhotoScan) for comparison.[18]

For the quantitative analysis, we utilized LiDAR data with a ground sampling distance of 0.5 m as the reference for the 3D reconstruction. To assess the accuracy of the generated DEM, we interpolated the digital surface model (DSM) to the DEM and evaluated the associated errors across various DEM generation methods. In all instances, we ensured that the input views and configurations were consistent across all methodologies to facilitate a fair comparison.

3.2 DEM generation experiment

Figure 2 illustrates the DEM results for the AOI area generated using our proposed method alongside comparative approaches. In these visualizations, darker colors represent higher elevations. We analyzed our algorithm across three key dimensions.

Overall, our method is better than NeRF-based, S-NeRF-based, and traditional photogrammetry approaches. The NeRF-based method exhibits the poorest performance, particularly struggling with DEM generation in JAX_004 and JAX_260. This problem is due to the satellite's high altitude, which leads to blank areas in traditional sampling methods and

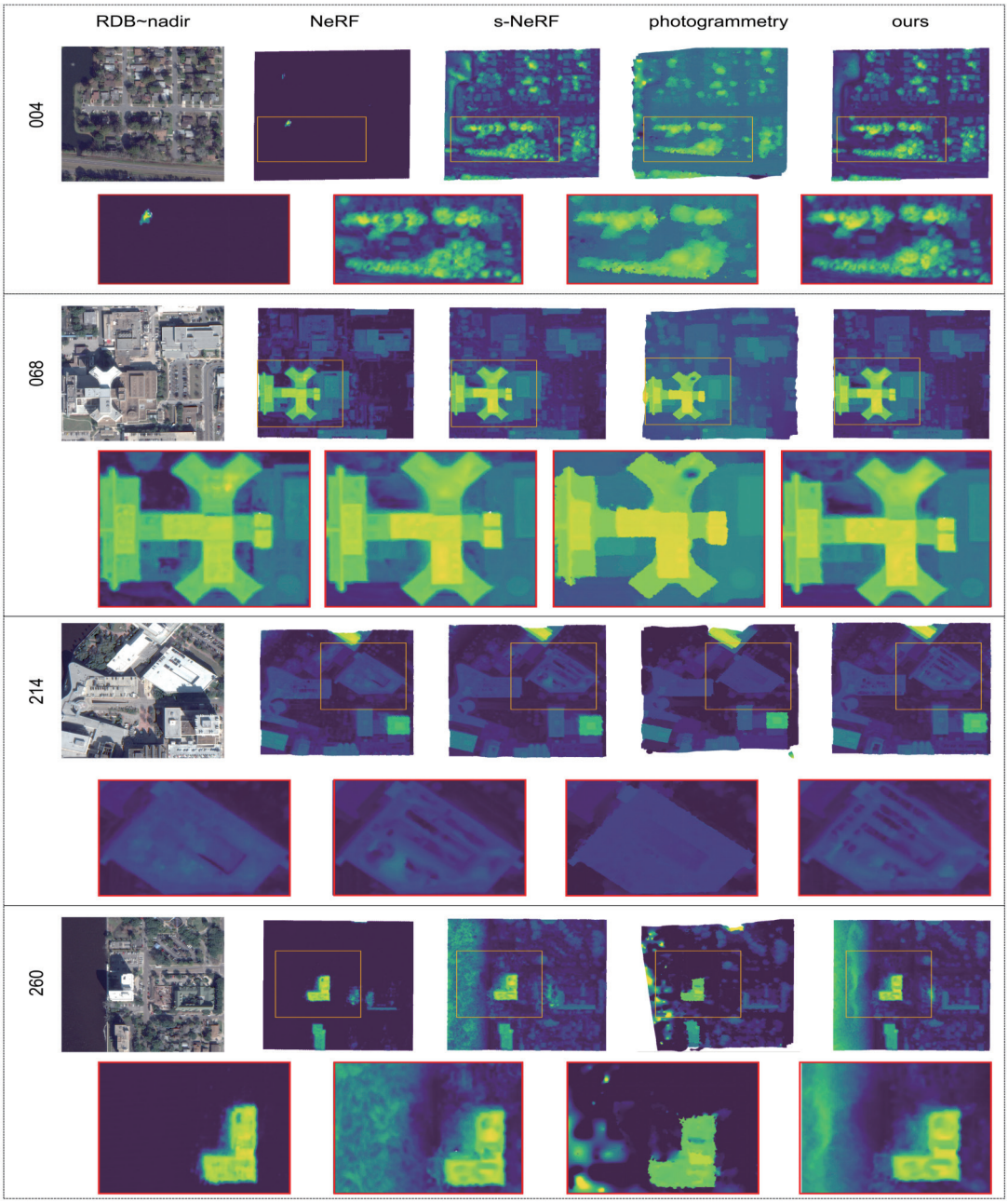

Fig. 2. (Color online) Left to right: Ground truth RGB, NeRF DEM, S-NeRF DEM, photogrammetry DEM, and our method. Red areas represents the clear areas, and yellow areas are enlarged examples of the areas.

generates noise in other regions. Although traditional photogrammetry can preserve details, it tends to produce overly smooth transitions, resulting in detail loss in specific areas, such as JAX_214. The S-NeRF-based approach, although superior to NeRF, falls short of our method in terms of detail preservation. For instance, in JAX_068, traditional photogrammetry introduces holes near the white building area owing to its inability to handle weak textures. The NeRF-based method also has artifacts caused by shadow misinterpretations, leading to incorrect

mappings. Although S-NeRF addresses shadow-related issues, it fails to account for moving objects, leading to erroneous mappings.

Moreover, in JAX_260, traditional photogrammetry struggles with sunlight reflections from water bodies, leading to significant mapping errors. Although S-NeRF addresses this issue, it also introduces discontinuities. In contrast, our algorithm effectively handles sunlight reflections from water bodies, producing better results with more details.

In summary, our method not only preserves good details but also delivers the best overall performance compared with the other algorithms. Figure 2 shows the DEM results of the AOI area obtained using the proposed and comparative methods. These results are analyzed from three perspectives, as follows. Overall, our method is better than the methods based on NeRF, S-NeRF, and traditional photogrammetry. The NeRF-based and traditional photogrammetry methods exhibit significant noise, likely due to insufficient geometric regularization. The S-NeRF-based method does not achieve the same level of detail as our method. As shown in JAX_068, near the white building area, NeRF and traditional photogrammetry methods show holes and artifacts in the details. This is because these two methods fail to adequately address shadows, leading to inaccurate mapping. Although S-NeRF can mitigate shadow effects, it introduces white noise in the details owing to its inability to account for the physics of motion over time. As shown in JAX_004 the section of Fig. 2, our algorithm can effectively eliminate the problem of moving objects, and particularly objects such as trees and other natural features, obtaining better results and smoother details. Furthermore, for the building details in JAX_214 and JAX_260 our method shows better details and performs better overall.

To further illustrate the advantages of our method in 3D modeling, Fig. 3 shows the mesh model reconstructed by our approach alongside those produced by other comparison methods.

The figure also explains the DEM results from Fig. 3 for reference. For instance, in JAX_068, the mesh model generated by the traditional photogrammetry method exhibits holes near the white building area, a consequence of its inability to handle weak textures. The NeRF-based method introduces artifacts, whereas S-NeRF struggles with moving objects, resulting in incorrect mappings in roof details. Similar issues are observed in JAX_260, where the NeRF-based method fails to accurately recognize the ground, and traditional photogrammetry cannot resolve sunlight reflections from water bodies, leading to mapping errors. Although S-NeRF mitigates the reflection problem, it introduces discontinuities. In summary, our method not only preserves better details but also delivers the best overall performance among the algorithms compared.

3.3 Accuracy evaluation

To evaluate the accuracy of the generated DEM, in this study, we interpolate the DSM to create the DEM and calculate the elevation error using various DEM generation methods. Standard quantitative evaluation metrics, including the mean absolute error (*MAE*), median absolute error (*MED*), and root mean squared error (*RMSE*), are employed to assess the accuracy of the different methods. The formulas for these metrics are as follows:

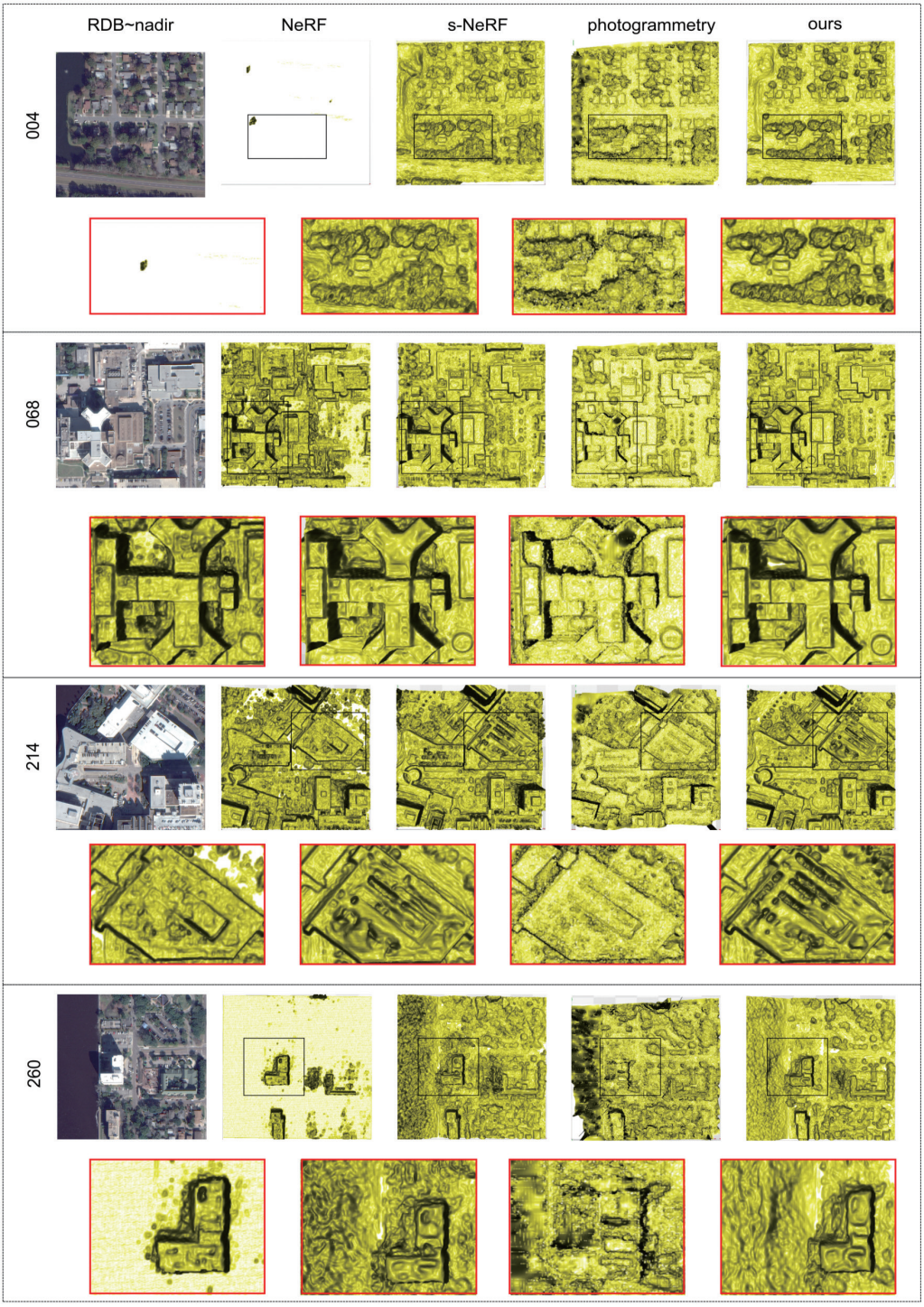

Fig. 3. (Color online) Left to right: Ground truth RGB, NeRF, S-NeRF, photogrammetry, and our method for generating meshes. Black areas represent clear areas, and red areas are enlarged examples of the areas.

Table 1
Quantitative results of the DFC2019 dataset. Arrows pointing upwards indicate higher precision for higher values, whereas arrows pointing downwards indicate higher precision for lower values (in meters).

		NeRF	S-NeRF	Photogrammetry	Ours
JAX_004	*MAE*↓	3.327	1.830	1.531	1.288
	MED↓	2.800	1.230	0.932	0.8
	RMSE↓	4.500	2.300	1.900	1.500
JAX_068	*MAE*↓	3.644	1.496	1.301	1.249
	MED↓	2.794	0.860	0.794	0.66
	RMSE↓	4.900	2.000	1.600	1.500
JAX_214	*MAE*↓	3.687	2.691	2.402	2.009
	MED↓	2.590	2.041	1.732	1.030
	RMSE↓	4.900	3.400	3.100	2.400
JAX_260	*MAE*↓	3.257	3.245	2.051	1.864
	MED↓	2.942	2.590	1.760	1.530
	RMSE↓	4.400	4.300	2.700	2.300

$$
\begin{aligned}
MAE &= \frac{1}{n}\sum_{i=1}^{n}\left|h_{recon,i} - h_{lidar,i}\right|, \\
MED &= median\left(\left|h_{recon,i} - h_{lidar,i}\right|\right) i = 1,2,\ldots,n, \\
RMSE &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(h_{recon,i} - h_{lidar,i}\right)^2},
\end{aligned}
\tag{7}
$$

where $h_{recon,i}$ and $h_{lidar,i}$ represent the elevation values of the i-th pixel in the reconstructed DEM and the reference DEM, respectively. n represents the number of all pixels in the DEM.

Table 1 presents the quantitative comparison results for the DFC2019 dataset, which provides only the DSM as ground truth. *MAE* and *MED* measure the absolute elevation error, offering a straightforward evaluation of overall and median error levels, respectively. On the other hand, *RMSE* emphasizes larger errors by squaring the differences, making it particularly effective for highlighting extreme errors and assessing the robustness of the methods. Our approach obtains good results compared with the other methods across every metric and scenario.

4. Conclusions

In this study, we implemented the NeRF model and developed a data processing method to generate DEMs with mesh structures from satellite images. We integrated a recent approach based on the Sat-NeRF model, which directly accounts for lighting conditions in conjunction with radiance modeling. Our approach constructs 3D scenes from a series of satellite images captured at different times, effectively addressing challenges such as lighting variations and transient objects. Compared with existing methods, our experiments demonstrate that our approach can produce high-quality DEMs and corresponding mesh models. Furthermore, it outperforms both traditional and more recent techniques in qualitative and quantitative evaluations. Our exploration was focused on the NeRF, S-NeRF, and Sat-NeRF implementations, which require a significant amount of training time. We will further study time acceleration in the future.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant no. 42201483), the China Postdoctoral Science Foundation (Grant no. 2022M710332), and Funding for Postdoctoral Research Activities in Beijing (Grant no. 2023-zz-140).

References

1 Q. Zhao, L. Yu, Z. Du, D. Peng, P. Hao, Y. Zhang, and P. Gong: Remote Sens. **14** (2022) 1863. https://doi.org/10.3390/rs14081863
2 R. D. G. Loyola: Neural Networks **19** (2006) 168.
3 S. Salcedo-Sanz, P. Ghamisi, M. Piles, M. Werner, L. Cuadra, A. Moreno-Martínez, E. Izquierdo-Verdiguier, J. Muñoz-Marí, A. Mosavi, and G. Camps-Valls: Inf. Fusion **63** (2020) 256. https://doi.org/10.1016/j.inffus.2020.07.004
4 M. Rothermel, K. Gong, D. Fritsch, K. Schindler, and N. Haala: ISPRS J. Photogramm. Remote Sens. **166** (2020) 52. https://doi.org/10.1016/j.isprsjprs.2020.05.001
5 P. Migon, M. Kasprzak, and A. Traczyk: Landform Anal. **22** (2013) 89. https://doi.org/10.12657/landfana.022.007
6 M. Wieczorek and P. Migoń: Geomorphology **206** (2014) 133. https://doi.org/10.1016/j.geomorph.2013.10.005
7 J. Iwahashi, and R. J. Pike: Geomorphology **86** (2007) 409. https://doi.org/10.1016/j.geomorph.2006.09.012
8 L. Dragut and C. Eisank: Geomorphology **129** (2011) 83. https://doi.org/10.1016/j.geomorph.2011.03.003
9 H. B. Makineci, H. Karabörk, and A. Durdu: Türkiye Uzaktan Algılama Dergisi **2** (2020) 58.
10 M. Pepe, V. S. Alfio, and D. Costantino: Appl. Sci. **12** (2022) 2886. https://doi.org/10.3390/app122412886
11 B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng: Commun. ACM **65** (2022) 99. https://doi.org/10.1145/3503250
12 K. Gao, Y. Gao, H. He, D. Lu, L. Xu, and J. Li: arXiv (2022). https://doi.org/10.48550/arxiv.2210.00379
13 V. Croce, G. Caroti, L. De Luca, A. Piemonte, and P. Véron: Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. **XLVIII-M-2-2023** (2023) 453. https://doi.org/10.5194/isprs-archives-XLVIII-M-2-2023-453-2023
14 F. Remondino, A. Karami, Z. Yan, G. Mazzacca, S. Rigon, and R. Qin: Remote Sens. **15** (2023) 585. https://doi.org/10.3390/rs15143585
15 D. Derksen and D. Izzo: arXiv (2021). https://doi.org/10.48550/arxiv.2104.09877
16 R. Mari, G. Facciolo, and T. Ehret: Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW) (IEEE, 2022) 1310–1320. https://doi.org/10.48550/arxiv.2203.08896
17 R. Martin-Brualla, N. Radwan, M. S. M. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth: Proc. 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR) (IEEE, 2021) 7206–7215. https://doi.org/10.1109/CVPR46437.2021.00713
18 J. R. Over: Open-File Report 2021–1039 (2021). https://doi.org/10.3133/ofr20211039

About the Authors

Tianjiao Wang is currently pursuing his master's degree at the School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture. His research interests include 3D reconstruction, photogrammetry, and orthophoto generation.
(2108160122006@stu.bucea.edu.cn)

Junxing Yang received his Ph.D. degree from Wuhan University, focusing on photogrammetry and remote sensing. His research pursuits extend to diverse domains such as 3D reconstruction, image stitching, and scene understanding. (yangjunxing@bucea.edu.cn)

Tong Ye is currently pursuing his master's degree at the School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture. His research interests cover 3D reconstruction and building facade image generation. (1523080488@qq.com)

He Huang received his B.S. degree from Wuhan University, China, in 2000, and pursued further education abroad, earning both his M.S. and Ph.D. degrees from Sungkyunkwan University, South Korea, in 2004 and 2010, respectively. Since 2010, he has held positions as a lecturer and associate professor at Beijing University of Civil Engineering and Architecture, China. His research focuses on areas such as autonomous driving, high-precision navigation maps, and visual navigation and positioning.
(huanghe@bucea.edu.cn)