The Potential of Neural Radiance Fields and 3D Gaussian Splatting for 3D Reconstruction from Aerial Imagery

Dennis Haitz[1,2], Max Hermann[1,3], Aglaja Solana Roth[1], Michael Weinmann[2], Martin Weinmann[1]

[1] Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Karlsruhe, Germany
- (dennis.haitz, max.hermann, martin.weinmann)@kit.edu, aglaja.roth@student.kit.edu
[2] Intelligent Systems Department, Delft University of Technology, Delft, The Netherlands - m.weinmann@tudelft.nl
[3] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Karlsruhe, Germany
- max.hermann@iosb.fraunhofer.de

Keywords: Aerial Imagery, 3D Reconstruction, Multi-View Stereo, Neural Radiance Fields, 3D Gaussian Splatting, Sampling.

Abstract

In this paper, we focus on investigating the potential of advanced Neural Radiance Fields (NeRFs) and 3D Gaussian Splatting for 3D scene reconstruction from aerial imagery obtained via sensor platforms with an almost nadir-looking camera. Such a setting for image acquisition is convenient for capturing large-scale urban scenes, yet it poses particular challenges arising from imagery with large overlap, very short baselines, similar viewing direction and almost the same but large distance to the scene, and it therefore differs from the usual object-centric scene capture. We apply a traditional approach for image-based 3D reconstruction (COLMAP), a modern NeRF-based approach (Nerfacto) and a representative for the recently introduced 3D Gaussian Splatting approaches (Splatfacto), where the latter two are provided in the Nerfstudio framework. We analyze results achieved on the recently released UseGeo dataset both quantitatively and qualitatively. The achieved results reveal that the traditional COLMAP approach still outperforms Nerfacto and Splatfacto approaches for various scene characteristics, such as less-textured areas, areas with high vegetation, shadowed areas and areas observed from only very few views.

1. Introduction

3D reconstruction from multi-view imagery refers to the process of creating a 3D representation of an object or scene using multiple images captured from different viewpoints. This process leverages the information contained in multiple images to reconstruct the geometry and possibly also appearance of the considered object or scene in three dimensions. It is widely applied in fields such as computer vision, robotics, augmented reality, archaeology, architecture, and medical imaging.

In recent years, traditional approaches for image-based 3D reconstruction have been complemented by a series of learning-based approaches (Stathopoulou and Remondino, 2023) and, meanwhile, also approaches relying on a Neural Radiance Field (NeRF) (Mildenhall et al., 2020) and 3D Gaussian Splatting (Kerbl et al., 2023). The latter two hold great potential, since they also allow for impressively reconstructing reflectance properties beyond simple color in addition to geometry information. A comparative assessment of the potential of both traditional approaches and NeRF-based approaches has been addressed in previous work regarding the reconstruction of cultural heritage objects of various scales from either terrestrial or drone imagery (Croce et al., 2024) and regarding urban scene reconstruction from UAV-borne imagery (Nex et al., 2023). Whereas the former use case (Croce et al., 2024) focused on convergent image acquisition, the latter use case (Nex et al., 2023) included (mostly) imagery taken in almost nadir view.

In this paper, we direct our attention to the investigation of the potential of state-of-the-art scene representations like advanced Neural Radiance Fields (Tancik et al., 2023) and the recently introduced 3D Gaussian Splatting (Kerbl et al., 2023) for 3D scene reconstruction from aerial imagery obtained via sensor platforms with an almost nadir-looking camera (cf. Figure 1).

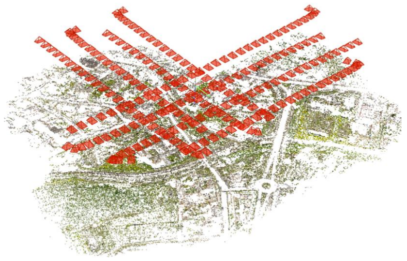

Figure 1. Illustration of camera poses given during data acquisition for a part of the UseGeo dataset (Nex et al., 2023).

Such a setting for image acquisition is convenient for capturing large-scale urban scenes, yet it poses particular challenges arising from imagery with large overlap, very short baselines, similar viewing direction and almost the same but large distance to the scene, and it therefore differs from the usual object-centric scene capture. We compare the modern NeRF-based approach Nerfacto as well as the 3D Gaussian Splatting implementation Splatfacto provided in the Nerfstudio framework (Tancik et al., 2023). Additionally, we investigate the use of different sampling strategies for the NeRF-based approach. Thereby, our work extends previous investigations (Nex et al., 2023) that demonstrated promising results for NeRF-based 3D reconstruction from high-resolution imagery with large overlap and very short baselines, while also showing that NeRF-based approaches cannot yet compete with traditional photogrammetric 3D reconstruction approaches in terms of point density and object/surface details regarding different scene characteristics. In summary, the main contributions of this paper are as follows:

- We investigate the potential of Nerfacto as a representative NeRF-based approach and Splatfacto as a representative approach for 3D Gaussian Splatting for the scenario of 3D reconstruction from almost nadir-looking aerial imagery.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-2-2024-97-2024 | © Author(s) 2024. CC BY 4.0 License.

- For the considered NeRF-based approach, we investigate different sampling strategies.
- We analyze results achieved on the recently released UseGeo dataset (Nex et al., 2023) quantitatively and qualitatively.

2. Related Work

In the following, we focus on 3D reconstruction from imagery and address related work in the form of traditional approaches (Section 2.1), learning-based approaches (Section 2.2), and approaches based on advanced scene representations (Section 2.3).

2.1 Traditional Approaches

Traditional approaches for 3D reconstruction from imagery rely on hand-crafted features and matching metrics to achieve multi-view consistency. Among these approaches, PhotoTourism / Bundler (Snavely et al., 2006) was one of the first open-source pipelines for image-based 3D reconstruction and point cloud generation from a set of internet photos without prior knowledge of the scene or camera geometry. However, the 3D reconstruction achieved by such a Structure-from-Motion (SfM) approach is rather sparse, since only point-like image features between images are matched and triangulated. To derive a dense 3D reconstruction, Multi-View Stereo (MVS) approaches have been integrated with PMVS/CMVS (Furukawa and Ponce, 2009) and numerous follow-up developments (Stathopoulou and Remondino, 2023). Several of these approaches have meanwhile been provided in the form of open-source image-based 3D reconstruction pipelines (Stathopoulou et al., 2019, Ruano and Smolic, 2021). In particular, COLMAP (Schönberger and Frahm, 2016, Schönberger et al., 2016) is often used, since it represents an end-to-end SfM and MVS pipeline that is actively maintained and easy-to-use for both expert and non-expert users.

Besides the standard offline approaches, that focus on accurate 3D reconstruction with typically high computational burden, there are also approaches for online 3D reconstruction from imagery (Rothermel et al., 2012, Sinha et al., 2014, Kern et al., 2020, Hermann et al., 2021). Due to the focus on real-time capabilities, however, a significantly lower accuracy can usually be observed regarding 3D reconstruction.

Traditional approaches heavily rely on photometric consistency, and they encounter challenges in case of weakly-textured or reflective surfaces (Huang et al., 2018). Furthermore, such approaches do typically not involve other cues like illumination, shadows and/or semantics, since it is non-trivial to address such information within a hand-crafted objective function.

2.2 Learning-based Approaches

Learning-based approaches for 3D reconstruction from imagery are often categorized with respect to the given scene representation. The two predominant categories comprise voxel-based approaches and depth-map-based approaches (Wang et al., 2021b, Stathopoulou and Remondino, 2023). Voxel-based approaches rely on directly predicting globally coherent voxel-occupancy grids from the input imagery; they are constrained by the predefined resolution of a reconstructed object or scene and thus not suitable for use cases focusing on large-scale 3D reconstruction. In contrast, depth-map-based approaches focus on predicting the depth map corresponding to each view and subsequently fusing all these depth maps into a coherent 3D representation; this decoupling makes such approaches much more efficient and also suitable for use cases focusing on large-scale 3D reconstruction.

Among the depth-map-based approaches, deep networks for Multi-View Stereo reconstruction are represented by DeepMVS (Huang et al., 2018) and MVSNet (Yao et al., 2018) and numerous variants thereof, mainly focusing on optimization based on 3D cost volumes. For instance, UniMVSNet (Peng et al., 2022) represents a coarse-to-fine framework directly constraining the cost volume like classification methods, but also realizing sub-pixel depth prediction like regression methods. GeoMVSNet (Zhang et al., 2023) focuses on integrating coarse geometric structures into finer depth estimations for improved geometry awareness. The most recent learning-based approaches also comprise transformer-based approaches like TransMVSNet (Ding et al., 2022), MVSFormer (Cao et al., 2022) and MVSFormer++ (Cao et al., 2024).

2.3 Approaches based on Advanced Scene Representations

Recently, implicit neural scene representations have gained a lot of attention. The underlying idea is to represent a scene in terms of the weights of a neural network that is optimized in a supervised manner to match its predicted scene appearance under certain views to the respectively observed input photographs for the corresponding view configurations. This is based on the assumption that scene characteristics have been accurately captured in the neural scene model if the deviations to the input reference views are small.

In particular, Neural Radiance Fields (NeRFs) (Mildenhall et al., 2020) and their many extensions (Tewari et al., 2022) have been demonstrated to offer a high potential for accurate scene representation. The involved network predicts – for each point in the volume – density and view-dependent color information and involves volume rendering techniques to synthesize images from the volumetric scene information. Particularly relevant to the task of capturing accurate geometric models similar to the dense matching of standard MVS pipelines are the NeRF extensions towards improving rendering quality and, hence, also model quality by reducing aliasing (Barron et al., 2021, Wang et al., 2022, Barron et al., 2022, Barron et al., 2023) as well as the acceleration of the training of the underlying network (Müller et al., 2022, Chen et al., 2022), allowing the inference of scene models in the order of seconds (Müller et al., 2022). Further developments include NeRF variants designed to handle photo collections taken *in-the-wild* (Martin-Brualla et al., 2021). In addition, several approaches focused on investigating the suitability of NeRFs and respective extensions for large-scale scenarios (Tancik et al., 2022, Turki et al., 2022, Xiangli et al., 2022, Mi and Xu, 2023, Xie et al., 2023, Xu et al., 2024).

Instead of relying on a network to predict volumetric fields for density and view-dependent color information, several works focused on representing scenes in terms of implicit surfaces (Wang et al., 2021a, Wang et al., 2023, Ge et al., 2023) or explicit representations in terms of meshes (Munkberg et al., 2022) or based on 3D Gaussians (Kerbl et al., 2023).

The assessment of the potential of different neural scene representations with respect to conventional photogrammetric tools requires comparisons on benchmark datasets, such as (Knapitsch et al., 2017, Nex et al., 2023, Yan et al., 2023). Comparisons of the geometric accuracy provided by different methods can then be conducted based on point cloud representations derived from NeRF methods or 3D Gaussian Splatting.

Whereas NeRFs and respective variants have been demonstrated to outperform conventional photogrammetric approaches for challenging scenarios (Condorelli et al., 2021, Balloni et al., 2023, Pepe et al., 2023, Llull et al., 2023, Remondino et al., 2023) including texture-less, metallic, highly reflective, and transparent objects, conventional techniques still performed better in case of well-textured and partially textured objects (Remondino et al., 2023). However, all of these studies focused on the comparison of the respective techniques in terms of the quality achieved for 3D object reconstruction from input views all around the object. Instead, the scenario we investigate in this work differs in the following characteristics: We focus on high-resolution aerial imagery obtained from an almost constant flight altitude of about 80 m, where the camera is facing downward.

3. Methodology

In this work, we consider a traditional approach for 3D reconstruction (Section 3.1), a state-of-the-art NeRF-based approach (Section 3.2) and a state-of-the-art learning-based explicit representation of 3D Gaussian Splatting (Section 3.3), and we compare them against a reference in terms of LiDAR data.

3.1 Traditional 3D Reconstruction

As a representative technique for 3D reconstruction based on conventional SfM and MVS, we use COLMAP (Schönberger and Frahm, 2016, Schönberger et al., 2016).

3.2 NeRF-based 3D Reconstruction

Before discussing the selected method for this category of approaches, we briefly provide background information on the underlying concept of Neural Radiance Fields.

3.2.1 Preliminaries on Neural Radiance Fields: Neural Radiance Fields (NeRFs) (Mildenhall et al., 2020) rely on inputs in terms of N given training images with corresponding camera parameters (i.e., camera intrinsics and pose). Based on the concept of using a feed-forward network to predict view-dependent radiance $\mathbf{c}(\mathbf{x}, \mathbf{d}) \in \mathbb{R}^3$ and volume density $\sigma(\mathbf{x}) \in \mathbb{R}$ for a given spatial 3D location $\mathbf{x} \in \mathbb{R}^3$ and the view direction $\mathbf{d} \in \mathbb{R}^3$, the color observed in a particular pixel in the image is obtained by integrating along the respective viewing ray $\mathbf{r}(t) = \mathbf{o} + t\,\mathbf{d}$ in the volume. Here, $\mathbf{o} \in \mathbb{R}^3$ and $\mathbf{d} \in \mathbb{R}^3$ denote the origin and direction of the ray, respectively. To compute this integral, the standard NeRF approach leverages the sampling of $K \in \mathbb{N}$ positions $t_1, ..., t_K \in \mathbb{R}$ along the ray (Max, 1995), i.e., the observed color is obtained according to

$$C(\mathbf{r}) = \sum_{k=1}^{K} T_k\, \alpha_k\, \mathbf{c}_k, \quad \text{with } \alpha_k = (1 - e^{-\sigma_k \delta_k}), \tag{1}$$

where $T_k = \exp\left(-\sum_{j=1}^{k-1} \sigma_j\, \delta_j\right)$ represents the transmittance along the ray up to t_k, $\mathbf{c}_k = \mathbf{c}(\mathbf{r}(t_k))$ denotes the radiance, $\sigma_k = \sigma(\mathbf{r}(t_k))$ denotes the density at t_k, and $\delta_k = t_{k+1} - t_k$ denotes the distance of adjacent samples. The positions t_k along the ray are computed in a hierarchical manner. After an initial stratified sampling, a refinement according to the density distribution along each ray is conducted (Mildenhall et al., 2020). Finally, an optimization loss given in terms of the mean squared error between the predicted color $\hat{C}(\mathbf{r})$ and the corresponding observed color from the input image $C(\mathbf{r})$ for a batch of camera rays R according to

$$\mathcal{L} = \mathcal{L}_I = \frac{1}{|R|} \sum_{\mathbf{r} \in R} \| \hat{C}(\mathbf{r}) - C(\mathbf{r}) \|_2^2 \tag{2}$$

is used to train the network.

3.2.2 Nerfacto: As a current representative NeRF variant, we employ Nerfacto (Tancik et al., 2023), a method that has been released in the scope of the Python-based framework Nerfstudio (Tancik et al., 2023) and serves as the default model for static scenes. Nerfacto is a combination of several components from recent papers (i.e., camera pose refinement, per-image appearance conditioning, proposal sampling, scene contraction, and hash encoding) in order to achieve a trade-off between fast training and high reconstruction quality, while allowing to correct for inaccuracies in camera parameters and also remaining flexible towards further modifications (Tancik et al., 2023).

Efficiency and reconstruction trade-off. Utilizing fully-fused neural networks (Müller et al., 2021) and multi-resolution hash-encoding proposed in the Instant-NGP (iNGP) approach (Müller et al., 2022), fast training can be achieved. Regarding training time, iNGP usually performs better than Nerfacto due to its binary occupancy grid, which aims at efficient skipping of empty space within the scene. This occupancy grid approach, however, can lead to an inferior density-induced geometric scene representation. In order to tackle various challenges exposed by the binary occupancy grid approach, Nerfacto includes the proposal network sampling proposed in (Barron et al., 2022). While this method increases training time compared to the occupancy grid, the inherent surface-focused approach is especially of interest regarding 3D reconstruction. Originally, the implementation consists of a proposal network which acts as a density predictor on the ray, but does not include color. Thus, ground truth data is not accessible for training the proposal network. A distillation approach guided by the NeRF network is therefore implemented. In turn, the objective of the proposal network is then to learn a guidance for the NeRF ray sampling near the surface of objects within the scene. By default, the Nerfacto method first samples 256 uniformly distributed positions on a ray, performed by a piece-wise sampling. From those samples, a first refinement iteration is applied through proposal sampling with 96 samples, and in a second iteration with 48 samples. Those sampling parameter values can be set before training. The extended sampling strategy is especially of interest for our investigations, because for remote sensing imagery, there is a large distance between the projection center and the scene surface. In contrast, the height variation within the surface structure is usually rather low. For that matter, we want to employ a high piece-wise uniform sampling for an adequate initial surface approximation. In order to capture the rather low height variation of the surface, we also employ high sampling values for the proposal and NeRF sampler. We refer to Section 4.3 for details.

Network parameters. NeRFs store scene content within the parameters of neural networks, and, in case of the multiresolution hash-encoding, also in large feature tables, represented as hash tables. To enable encoding and storing more scene content during the training process, we aim at extending the number of neurons of each layer. The objective is that with a higher number of network parameters, especially areas with high variation in height (e.g., vegetation) can be adequately approximated. The fully-fused neural network is a hardware-optimized neural network implementation, which is usually used with layers of 64 neurons for optimized hardware saturation. (Müller et al., 2021) state that, with this number of neurons, the network performs four to five times faster than an analogous implementation in Tensorflow. However, this speed-up effect begins to vanish with an increasing number of neurons. To keep training time still as low as possible while increasing the number of network parameters, 128 neurons are used for each layer.

Camera pose refinement. Due to remaining inaccuracies in the camera calibration data (poses and intrinsics) computed with COLMAP, we have to account for the flexibility to further refine the position and viewing configuration of each pose. For this purpose, Nerfacto involves the pose and intrinsics refinement strategy proposed by (Wang et al., 2021c). The pose and intrinsic parameters are stored as GPU-processable parameters, fed to an Adam optimizer and updated based on the photometric backward loss-gradient. Because of the model-specific optimization process, we aim to further enhance the geometric representation of the surface.

Derivation of point cloud representation. To generate point clouds from the trained NeRF model, we render depth maps per training pose. Utilizing camera pose and intrinsics, the depth values per pixel are projected into the NeRF coordinate space. Per-pixel depth is calculated by accessing the median density response on the sampled ray.

3.3 3D Gaussian Splatting

Training the involved neural network as well as image synthesis based on volume rendering makes NeRF-based approaches computationally costly, particularly when aiming for high visual quality. Instead of trading off training time and visual quality, 3D Gaussian Splatting (Kerbl et al., 2023) has been introduced as a scene representation with 3D Gaussians that maintain the favorable characteristics of continuous volumetric radiance fields for optimizing the scene whilst circumventing superfluous computations in unoccupied space.

A 3D Gaussian consists of a mean vector μ, covariance matrix Σ, opacity α and spherical harmonics coefficients as optimizable parameters. After an initialization based on the sparse point cloud derived in the camera calibration process as provided by COLMAP (Schönberger and Frahm, 2016), an interleaved optimization and density control of the 3D Gaussians is conducted to optimize the placement of the Gaussians and their anisotropic covariance. Density control is executed after every n iterations, where n is a hyperparameter that can be set accordingly. It includes the operations of *removing*, *splitting* and *duplicating* Gaussians in order to densify the scene, especially in over- and under-reconstructed areas. The use of a respectively designed, fast visibility-aware rendering algorithm that supports anisotropic splatting allows efficient training and real-time rendering. Due to the explicit nature of 3D Gaussian Splatting, the mean vectors μ_G can be interpreted as scene points and exported as a point cloud. The iterative execution of density control leads to a usually perpetually growing number of 3D Gaussians over training time, resulting in a denser point cloud than the initial sparse SfM point cloud.

4. Experimental Results

After briefly describing the used dataset (Section 4.1), we explain the used evaluation metrics (Section 4.2) and implementation details (Section 4.3). Finally, we focus on the presentation of achieved results (Section 4.4).

4.1 Dataset

The UseGeo dataset (Nex et al., 2023) has been acquired with a UAV equipped with a SONY ILCE-7RM3 camera and a RIEGL miniVUX-3UAV scanner. In total, three flights have been performed from an average flight altitude of about 80 m with an image overlap of about 80 % (forward) to 60 % (side). Provided data comprise aerial imagery (with corresponding camera poses and camera intrinsics) and LiDAR point clouds as well as photogrammetric MVS point clouds. The provided imagery (829 images with a size of 7952×5304 pixels) has a Ground Sampling Distance (GSD), i.e. image resolution in object space, of approximately 2 cm, while the corresponding LiDAR point cloud has a point density of about 50 points/m^2. The data has been released in the form of three datasets which are referred to as Dataset-1, Dataset-2, and Dataset-3, representing three different urban and peri-urban areas to address scenarios of different complexity, respectively. Dataset-1 comprises 224 images, Dataset-2 327 images and Dataset-3 277 images.

4.2 Evaluation Procedure

The quality of derived 3D reconstructions may generally be derived by comparison to reference data on the level of depth maps, point clouds, or triangle meshes. In our work, we focus on evaluation on the level of point clouds.

Co-registration with reference. The evaluation on the level of point clouds requires a co-registration of the derived 3D point cloud $\mathcal{P}$ and the corresponding reference point cloud $\mathcal{P}_{\text{ref}}$. For this purpose, a coarse manual alignment of $\mathcal{P}$ and $\mathcal{P}_{\text{ref}}$ is performed. This is followed by a 2-step refinement, where the first step comprises a standard point cloud registration based on the Iterative-Closest-Point (ICP) algorithm (Besl and McKay, 1992). Subsequently, an extension of ICP to similarity transformations (including scale) is applied to refine the registration of the dense point clouds (Knapitsch et al., 2017).

Evaluation metrics. Once $\mathcal{P}$ and $\mathcal{P}_{\text{ref}}$ have been co-registered, all errors are calculated at a completeness level of 90 % to assign less significance to outliers. A threshold of 0.2 m is used to determine completeness, indicating that all points in the ground truth point cloud within this distance of an estimated point are considered when calculating the completeness metric. On this basis, standard evaluation metrics like the absolute error (L1-abs) and the root-mean-square error (RMSE) can be calculated from the distances between 3D points of the derived 3D reconstruction to the corresponding closest points in the reference point cloud, normalized by the number of points in $\mathcal{P}$. Furthermore, precision and recall metrics can be derived for a given distance threshold as proposed with a standard benchmark for image-based 3D reconstruction (Knapitsch et al., 2017). Following this benchmark, the precision quantifies the accuracy of the derived 3D reconstruction (i.e., it indicates how closely the reconstructed points lie to the ground truth), whereas the recall / completeness quantifies the completeness of the derived 3D reconstruction (i.e., it indicates to which extent all the ground truth points are covered).

For our work, we use L1-abs, RMSE, and recall / completeness (Cpl) as evaluation metrics. The reference point cloud is represented by the LiDAR point cloud available for the respective dataset. Furthermore, we take into account the number of 3D points in the derived 3D reconstructions (N_{pts}).

4.3 Implementation Details

In the following, we describe the used software configurations as well as the used hardware.

Software. We use Nerfstudio v1.0.0 (Tancik et al., 2023) for training NeRF and 3D Gaussian Splatting models as well as generating point clouds from the trained models. Our NeRF models are based on the Nerfacto architecture (Tancik et al., 2023), and they are trained with two different sets of hyperparameters for each dataset, resulting in the *Nerfacto-default* and *Nerfacto-big* models. The parameter sets particularly differ in the number of neurons per layer, the sampling of the proposal network (Barron et al., 2022) and the sampling of the NeRF network (Table 1). Note that there are two hyperparameters for the proposal network. The first one corresponds to a piece-wise surface approximation sampling, while the second one represents a resampling thereof as a first refinement iteration. A second iteration is executed by the NeRF network itself through a PDF sampler. To further saturate GPU capabilities and accelerate training time, fully-fused neural networks (Müller et al., 2021) are utilized for all NeRF models.

Configuration	Nerfacto-default	Nerfacto-big
Neurons per Layer	64	128
Sampling PN	256 / 96	512 / 256
Sampling NN	48	96

Table 1. Hyperparameter values for the two defined Nerfacto models (PN: proposal network; NN: NeRF network)

The Nerfacto-default and Nerfacto-big models are trained for 100k iterations as (Tancik et al., 2023) propose 70k or more iterations for state-of-the-art-competitive image reconstruction quality. In order to refine the camera poses, the $SE(3)$ optimization for position and direction is used for all models. Training data consists of camera poses with intrinsics and corresponding ground truth images. Out of each dataset, 10 % are used as a validation set at training time.

3D Gaussian Splatting is also integrated in Nerfstudio with an approach referred to as *Splatfacto*. The CUDA-based backbone of Splatfacto is the *gsplat* (Ye et al., 2024) library as part of the Nerfstudio github project, which is a re-implementation of the original technique (Kerbl et al., 2023). We utilized the default training configuration with the exception, that the images are not resampled for training regarding the COLMAP data parser of Nerfstudio. The default configuration sets the number of training iterations to 30k. Densification is executed every 100 iterations after the warm-up process, which is active for the first 500 iterations. Splatfacto does not support camera pose refinement currently.

Hardware. All experiments regarding the NeRF and 3D Gaussian Splatting training as well as generation of point clouds are performed on a desktop PC with an Intel i9 10950K CPU with 32 GB of RAM and an Nvidia RTX 3090 GPU.

4.4 Results

The 3D reconstructions derived via COLMAP, Nerfacto-default and Nerfacto-big are shown in Figures 2, 3 and 4 for Dataset-1, Dataset-2 and Dataset-3, respectively. Results achieved for Splatfacto are shown in Figure 5. The corresponding quantitative results are summarized in Table 2 for Dataset-1, Dataset-2 and Dataset-3, respectively. Zoomed-in versions and more detailed visualization of failure cases will be released at the website of this project (`https://github.com/UseGeoEvaluation/DepthEstimationAnd3DReconstruction`).

Dataset-1	L1-abs [m]	RMSE [m]	Cpl	N_{pts}
COLMAP	0.0609	0.0700	0.5911	13.8M
Nerfacto-default	0.1686	0.2227	0.3712	5.0M
Nerfacto-big	0.2024	0.2698	0.3127	5.0M
Splatfacto	0.2926	0.4004	0.2333	3.6M

Dataset-2	L1-abs [m]	RMSE [m]	Cpl	N_{pts}
COLMAP	0.0690	0.0812	0.5965	20.3M
Nerfacto-default	0.2448	0.3145	0.3221	15.0M
Nerfacto-big	0.2643	0.3388	0.3238	15.0M
Splatfacto	0.2060	0.2753	0.2686	5.5M

Dataset-3	L1-abs [m]	RMSE [m]	Cpl	N_{pts}
COLMAP	0.0782	0.0921	0.5418	17.2M
Nerfacto-default	0.2776	0.3599	0.3130	15.0M
Nerfacto-big	0.3052	0.3990	0.3317	15.0M
Splatfacto	0.2742	0.3687	0.2425	5.4M

Table 2. Results achieved for Dataset-1, Dataset-2 and Dataset-3, respectively

5. Discussion

The L1-abs and RMSE scores (cf. Table 2) reveal that the level of accuracy and completeness achieved by COLMAP cannot be reached with the Nerfacto and Splatfacto approaches. This also becomes visible in Figures 2-4 which indicate rather accurate 3D reconstructions achieved by COLMAP, whereas the two Nerfacto approaches reveal a lower accuracy of the achieved 3D reconstructions and seem to contain systematic deviations. The Nerfacto results seem best for scene parts that are contained in many overlapping images, while accuracy decreases significantly towards the scene boundaries. In contrast, Splatfacto does not show such systematic deviations besides a decrease in accuracy towards the scene boundaries (cf. Figure 5), but reveals lower completeness scores compared to the Nerfacto approaches (cf. Table 2), indicating that less points of the LiDAR reference point cloud are covered by the achieved 3D reconstruction. Both Nerfacto and Splatfacto approaches exhibit larger errors in less-textured areas, areas with high vegetation, shadowed areas and areas observed from only very few views.

6. Conclusions

In this paper, we focused on investigating the potential of advanced Neural Radiance Fields and 3D Gaussian Splatting for 3D scene reconstruction from aerial imagery obtained via sensor platforms with an almost nadir-looking camera. Such a setting for image acquisition is convenient for capturing large-scale urban scenes, yet it poses particular challenges arising from imagery with large overlap, very short baselines, similar viewing direction and almost the same but large distance to the scene. The achieved results reveal that a traditional approach for image-based 3D reconstruction represented by COLMAP still outperforms approaches based on NeRF and 3D Gaussian Splatting for various scene characteristics, for which we applied Nerfacto and Splatfacto from the Nerfstudio v1.0.0 framework. Thereby, our investigation differs from the often conducted studies regarding 3D object reconstruction in the object-centered acquisition scenario.

Future work will address further ablation studies regarding different design choices of Nerfacto and Splatfacto in terms of hyperparameter settings and extensions via dedicated modules for improved 3D reconstruction. Furthermore, more recent approaches like Zip-NeRF (Barron et al., 2023), mip-NeRF 360 (Barron et al., 2022) or multi-tiling NeRF (Xu et al., 2024) and multi-tiling 3D Gaussian Splatting to better exploit the capacity of the representation to accurately capture local scene structures will be addressed.

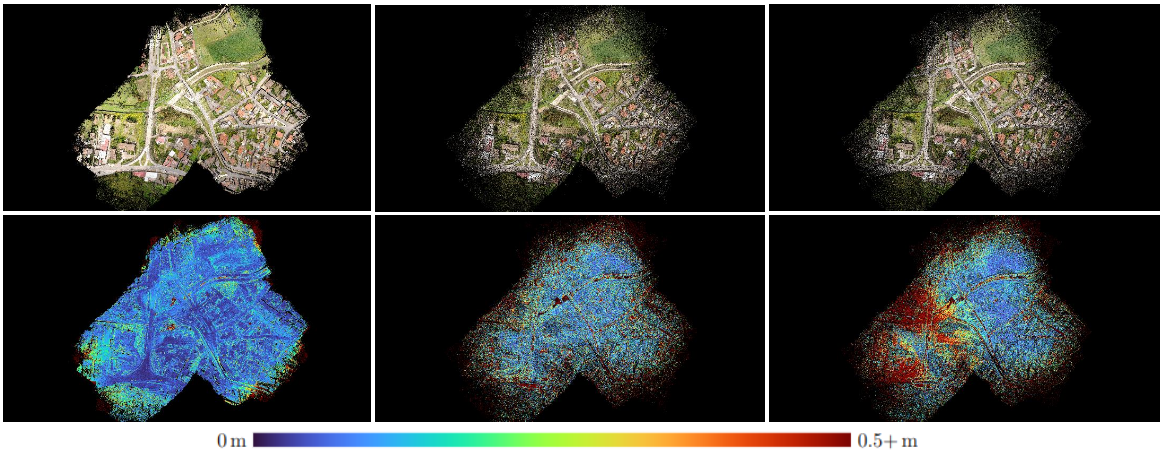

Figure 2. Colored point clouds (top) and their deviation from the LiDAR point cloud (bottom) for the 3D reconstructions achieved by COLMAP (left), Nerfacto-default (center) and Nerfacto-big (right) for Dataset-1.

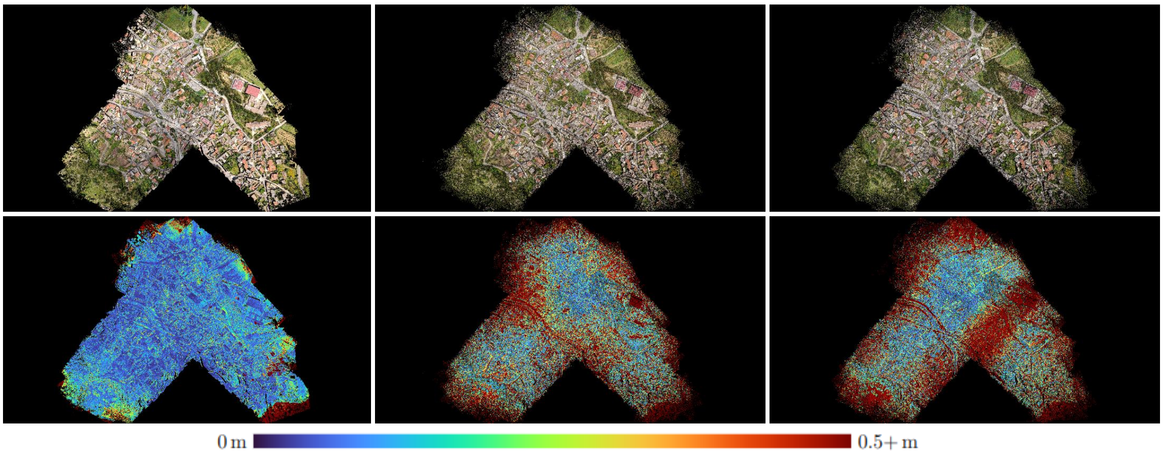

Figure 3. Colored point clouds (top) and their deviation from the LiDAR point cloud (bottom) for the 3D reconstructions achieved by COLMAP (left), Nerfacto-default (center) and Nerfacto-big (right) for Dataset-2.

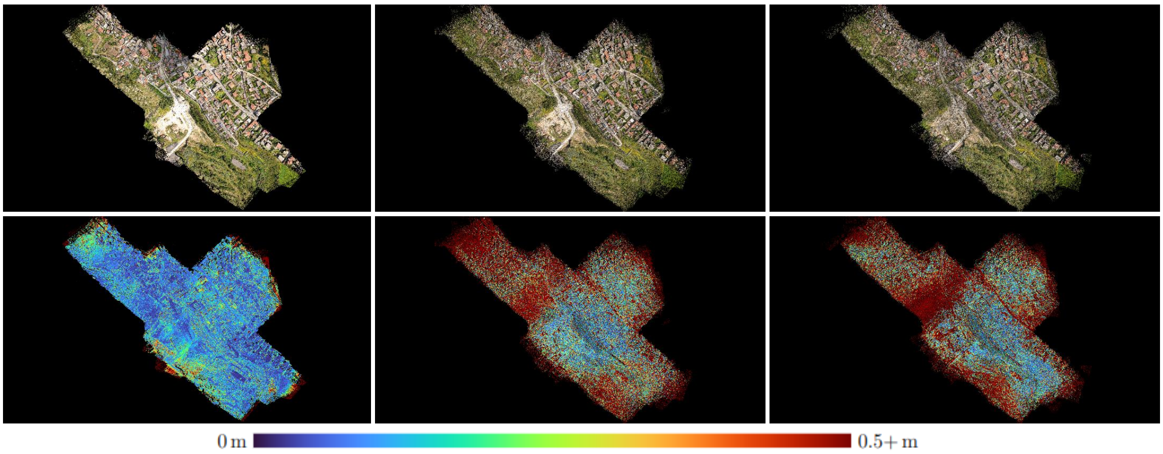

Figure 4. Colored point clouds (top) and their deviation from the LiDAR point cloud (bottom) for the 3D reconstructions achieved by COLMAP (left), Nerfacto-default (center) and Nerfacto-big (right) for Dataset-3.

References

Balloni, E., Gorgoglione, L., Paolanti, M., Mancini, A., Pierdicca, R., 2023. Few shot photogrametry: A comparison between NeRF and MVS-SfM for the documentation of cultural heritage. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-M-2-2023, 155–162.

Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P. P., 2021. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5835–5844.

Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., Hedman, P., 2022. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5460–5469.

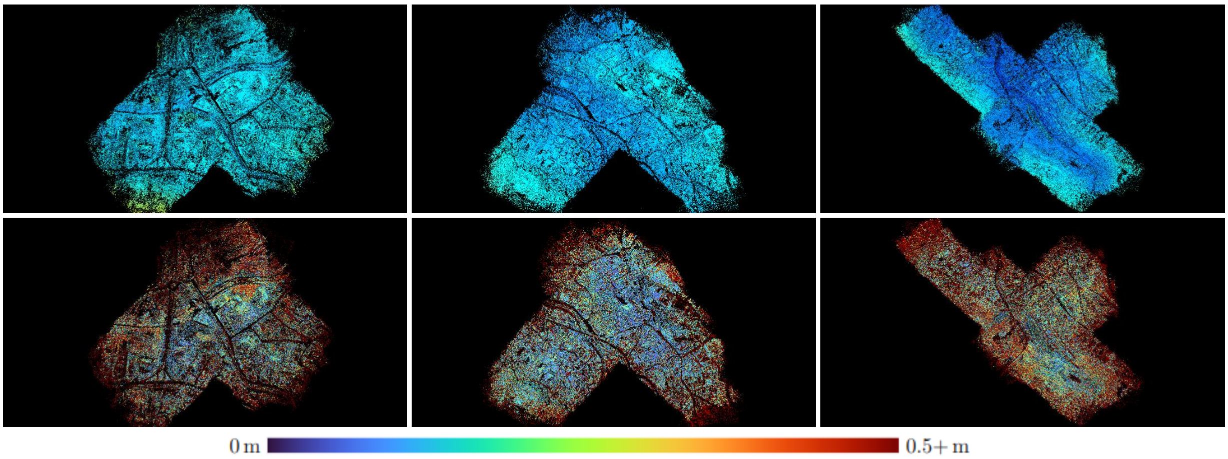

Figure 5. Point clouds colored by height (top) and their deviation from the LiDAR point cloud (bottom) for the 3D reconstructions achieved by Splatfacto for Dataset-1 (left), Dataset-2 (center) and Dataset-3 (right), respectively.

Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., Hedman, P., 2023. Zip-NeRF: Anti-aliased grid-based neural radiance fields. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19697–19705.

Besl, P., McKay, N. D., 1992. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(2), 239–256.

Cao, C., Ren, X., Fu, Y., 2022. MVSFormer: Multi-view stereo by learning robust image features and temperature-based depth. *Trans. Mach. Learn. Res.* https://openreview.net/forum?id=2VWR6JfwNo.

Cao, C., Ren, X., Fu, Y., 2024. MVSFormer++: Revealing the devil in transformer's details for multi-view stereo. arXiv preprint arXiv:2401.11673.

Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H., 2022. TensoRF: Tensorial radiance fields. *Proceedings of the European Conference on Computer Vision*, 333–350.

Condorelli, F., Rinaudo, F., Salvadore, F., Tagliaventi, S., 2021. A comparison between 3d reconstruction using NeRF neural networks and MVS algorithms on cultural heritage images. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIII-B2-2021, 565–570.

Croce, V., Billi, D., Caroti, G., Piemonte, A., De Luca, L., Véron, P., 2024. Comparative assessment of neural radiance fields and photogrammetry in digital heritage: Impact of varying image conditions on 3d reconstruction. *Remote Sens.*, 16(2), 301:1–301:19.

Ding, Y., Yuan, W., Zhu, Q., Zhang, H., Liu, X., Wang, Y., Liu, X., 2022. TransMVSNet: Global context-aware multi-view stereo network with transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8575–8584.

Furukawa, Y., Ponce, J., 2009. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(8), 1362–1376.

Ge, W., Hu, T., Zhao, H., Liu, S., Chen, Y.-C., 2023. Ref-NeuS: Ambiguity-reduced neural implicit surface learning for multi-view reconstruction with reflection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4228–4237.

Hermann, M., Ruf, B., Weinmann, M., 2021. Real-time dense 3d reconstruction from monocular video data captured by low-cost UAVs. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIII-B2-2021, 361–368.

Huang, P.-H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.-B., 2018. DeepMVS: Learning multi-view stereopsis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2821–2830.

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3d Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 139:1–139:14.

Kern, A., Bobbe, M., Khedar, Y., Bestmann, U., 2020. Real-time dense 3d reconstruction from monocular video data captured by low-cost UAVs. *Proceedings of the International Conference on Unmanned Aircraft Systems*, 902–911.

Knapitsch, A., Park, J., Zhou, Q.-Y., Koltun, V., 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graph.*, 37(4), 78:1–78:13.

Llull, C., Baloian, N., Bustos, B., Kupczik, K., Sipiran, I., Baloian, A., 2023. Evaluation of 3d reconstruction for cultural heritage applications. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 1642–1651.

Martin-Brualla, R., Radwan, N., Sajjadi, M. S. M., Barron, J. T., Dosovitskiy, A., Duckworth, D., 2021. NeRF in the Wild: Neural radiance fields for unconstrained photo collections. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7206–215.

Max, N., 1995. Optical models for direct volume rendering. *IEEE Trans. Vis. Comput. Graph.*, 1(2), 99–108.

Mi, Z., Xu, D., 2023. Switch-NeRF: Learning scene decomposition with mixture of experts for large-scale neural radiance fields. *Proceedings of the 11th International Conference on Learning Representations*, 1–15.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2020. NeRF: Representing scenes as neural radiance fields for view synthesis. *Proceedings of the European Conference on Computer Vision*, 405–421.

Müller, T., Evans, A., Schied, C., Keller, A., 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4), 102:1–102:15.

Müller, T., Rousselle, F., Novák, J., Keller, A., 2021. Real-time neural radiance caching for path tracing. *ACM Trans. Graph.*, 40(4), 36:1–36:16.

Munkberg, J., Hasselgren, J., Shen, T., Gao, J., Chen, W., Evans, A., Müller, T., Fidler, S., 2022. Extracting triangular 3d models, materials, and lighting from images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8270–8280.

Nex, F., Zhang, N., Remondino, F., Farella, E. M., Qin, R., Zhang, C., 2023. Benchmarking the extraction of 3d geometry from UAV images with deep learning methods. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-1/W3-2023, 123–130.

Peng, R., Wang, R., Wang, Z., Lai, Y., Wang, R., 2022. Rethinking depth estimation for multi-view stereo: A unified representation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8635–8644.

Pepe, M., Alfio, V. S., Costantino, D., 2023. Assessment of 3d model for photogrammetric purposes using AI tools based on NeRF algorithm. *Heritage*, 6(8), 5719–5731.

Remondino, F., Karami, A., Yan, Z., Mazzacca, G., Rigon, S., Qin, R., 2023. A critical analysis of NeRF-based 3d reconstruction. *Remote Sens.*, 15(14), 3585:1–3585:22.

Rothermel, M., Wenzel, K., Fritsch, D., Haala, N., 2012. SURE: Photogrammetric surface reconstruction from imagery. *Proceedings of the LowCost3D Workshop*.

Ruano, S., Smolic, A., 2021. A benchmark for 3d reconstruction from aerial imagery in an urban environment. *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 5 (VISAPP), 732–741.

Schönberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113.

Schönberger, J. L., Zheng, E., Pollefeys, M., Frahm, J.-M., 2016. Pixelwise view selection for unstructured multi-view stereo. *Proceedings of the European Conference on Computer Vision*, 501–518.

Sinha, S. N., Scharstein, D., Szeliski, R., 2014. Efficient high-resolution stereo matching using local plane sweeps. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1582–1589.

Snavely, N., Seitz, S. M., Szeliski, R., 2006. Photo tourism: Exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3), 835–846.

Stathopoulou, E. K., Remondino, F., 2023. A survey on conventional and learning-based methods for multi-view stereo. *Photogramm. Rec.*, 38(138), 374–407.

Stathopoulou, E.-K., Welponer, M., Remondino, F., 2019. Open-source image-based 3d reconstruction pipelines: Review, comparison and evaluation. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W17, 331–338.

Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B. P., Srinivasan, P., Barron, J. T., Kretzschmar, H., 2022. Block-NeRF: Scalable large scene neural view synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8238–8248.

Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Kerr, J., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., McAllister, D., Kanazawa, A., 2023. Nerfstudio: A modular framework for neural radiance field development. *SIGGRAPH'23: ACM SIGGRAPH 2023 Conference Proceedings*, 72:1–72:12.

Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P., Tretschk, E., Yifan, W., Lassner, C., Sitzmann, V., Martin-Brualla, R., Lombardi, S., Simon, T., Theobalt, C., Nießner, M., Barron, J. T., Wetzstein, G., Zollhöfer, M., Golyanik, V., 2022. Advances in neural rendering. *Comput. Graph. Forum*, 41(2), 703–735.

Turki, H., Ramanan, D., Satyanarayanan, M., 2022. Mega-NeRF: Scalable construction of large-scale NeRFs for virtual fly-throughs. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12912–12921.

Wang, C., Wu, X., Guo, Y.-C., Zhang, S.-H., Tai, Y.-W., Hu, S.-M., 2022. NeRF-SR: High quality neural radiance fields using supersampling. *Proceedings of the 30th ACM International Conference on Multimedia*, 6445–6454.

Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W., 2021a. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Adv. Neural Inf. Proc. Syst.*, 354, 27171–27183.

Wang, X., Wang, C., Liu, B., Zhou, X., Zhang, L., Zheng, J., Bai, X., 2021b. Multi-view stereo in the deep learning era: A comprehensive review. *Displays*, 70, 102102:1–102102:12.

Wang, Y., Han, Q., Habermann, M., Daniilidis, K., Theobalt, C., Liu, L., 2023. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3272–3283.

Wang, Z., Wu, S., Xie, W., Chen, M., Prisacariu, V. A., 2021c. NeRF–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*.

Xiangli, Y., Xu, L., Pan, X., Zhao, N., Rao, A., Theobalt, C., Dai, B., Lin, D., 2022. BungeeNeRF: Progressive neural radiance field for extreme multi-scale scene rendering. *Proceedings of the European Conference on Computer Vision*, 106–122.

Xie, S., Zhang, L., Jeon, G., Yang, X., 2023. Remote sensing neural radiance fields for multi-view satellite photogrammetry. *Remote Sens.*, 15(15), 3808:1–3808:17.

Xu, N., Qin, R., Huang, D., Remondino, F., 2024. Multi-tiling neural radiance field (NeRF) – Geometric assessment on large-scale aerial datasets. arXiv preprint arXiv:2310.00530v3.

Yan, Z., Mazzacca, G., Rigon, S., Farella, E. M., Trybala, P., Remondino, F., 2023. Nerfbk: A holistic dataset for benchmarking NeRF-based 3d reconstruction. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-1/W3-2023, 219–226.

Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L., 2018. MVSNet: Depth inference for unstructured multi-view stereo. *Proceedings of the European Conference on Computer Vision*, 785–801.

Ye, V., Turkulainen, M., the Nerfstudio team, 2024. gsplat.

Zhang, Z., Peng, R., Hu, Y., Wang, R., 2023. GeoMVSNet: Learning multi-view stereo with geometry perception. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21508–21518.