remote sensing

MDPI

Article

Leveraging Neural Radiance Fields for Large-Scale 3D Reconstruction from Aerial Imagery

Max Hermann [1,2,*], Hyovin Kwak [2], Boitumelo Ruf [2] and Martin Weinmann [1]

[1] Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany; martin.weinmann@kit.edu
[2] Fraunhofer IOSB, 76131 Karlsruhe, Germany; hyovin.kwak@iosb.fraunhofer.de (H.K.); boitumelo.ruf@iosb.fraunhofer.de (B.R.)
* Correspondence: max.hermann@kit.edu

Abstract: Since conventional photogrammetric approaches struggle with with low-texture, reflective, and transparent regions, this study explores the application of Neural Radiance Fields (NeRFs) for large-scale 3D reconstruction of outdoor scenes, since NeRF-based methods have recently shown very impressive results in these areas. We evaluate three approaches: Mega-NeRF, Block-NeRF, and Direct Voxel Grid Optimization, focusing on their accuracy and completeness compared to ground truth point clouds. In addition, we analyze the effects of using multiple sub-modules, estimating the visibility by an additional neural network and varying the density threshold for the extraction of the point cloud. For performance evaluation, we use benchmark datasets that correspond to the setting off standard flight campaigns and therefore typically have nadir camera perspective and relatively little image overlap, which can be challenging for NeRF-based approaches that are typically trained with significantly more images and varying camera angles. We show that despite lower quality compared to classic photogrammetric approaches, NeRF-based reconstructions provide visually convincing results in challenging areas. Furthermore, our study shows that in particular increasing the number of sub-modules and predicting the visibility using an additional neural network improves the quality of the resulting reconstructions significantly.

Keywords: Neural Radiance Fields; Multi-View Stereo; aerial imagery; 3D reconstruction; large-scale

check for updates

Citation: Hermann, M.; Kwak, H.; Ruf, B.; Weinmann, M. Leveraging Neural Radiance Fields for Large-Scale 3D Reconstruction from Aerial Imagery. *Remote Sens.* **2024**, *16*, 4655. https://doi.org/10.3390/rs16244655

Academic Editors: Francesco Nex and Henry Meißner

Received: 30 September 2024
Revised: 22 November 2024
Accepted: 29 November 2024
Published: 12 December 2024

1. Introduction

3D reconstruction from aerial imagery plays a crucial role in various remote sensing applications, including urban planning, environmental monitoring and disaster assessment. The ability to create accurate and detailed 3D models of large outdoor scenes can significantly improve decision-making and operational efficiency. The photogrammetric reconstruction of such scenes is facilitated by the widespread use of low-cost unmanned aerial vehicles (UAVs), while at the same time conventional photogrammetric techniques for 3D reconstruction from aerial imagery have made considerable progress over the years and are the de facto standard for reconstructions of large outdoor scenes [1–7]. These Multi-View Stereo (MVS) methods typically involve multiple overlapping images to estimate depth maps and reconstruct 3D models using hand-crafted features and parameters.

Despite these impressive results, challenges such as reflections, transparencies, low-textured regions, repeating patterns and changing lighting conditions complicate the reconstruction process, as no reliable pixel correspondences can be found between multiple images. This usually results in these regions being filtered out during the reconstruction process and leads to large gaps or can cause artifacts if the regions are not removed. Nevertheless, the pursuit of efficient and accurate large-scale 3D reconstruction continues, driven by the growing demand for photorealistic digital twins of real-world environments.

Recent advancements in photorealistic rendering, particularly the development of Neural Radiance Fields (NeRFs) [8], have opened new possibilities for 3D reconstruction [8–10]. NeRFs leverage deep learning techniques to create highly realistic models by representing scenes as a continuous 5D function that predicts radiance and density in each direction at every point in space. This approach enables the rendering of novel views with impressive accuracy and detail, capturing both geometry and appearance. Given the ability of NeRFs to overcome some of the limitations of traditional photogrammetric methods, there is a growing interest in exploring their potential for large-scale 3D reconstruction from aerial imagery.

Therefore, the focus of this work is to investigate the application of NeRF-based volume rendering for accurately reconstructing the geometry of large-scale outdoor scenes using aerial imagery and conventional aerial datasets intended for evaluating MVS techniques. These datasets often have a nadir camera perspective and a rather small image overlap. The quality of the reconstructions is evaluated by measuring the accuracy and completeness of the point clouds extracted from the implicit scene representations and compared with reference point clouds reconstructed by COLMAP [4,11].

In summary, our main contributions are as follows:

- We analyze the performance of three NeRF-based approaches on challenging large-scale aerial datasets and discuss both quantitative and qualitative results.
- We conduct a comprehensive ablative study to investigate the impact of different configurations.
- We compared the results with conventional MVS methods in the form of COLMAP.

1.1. Related Work

MVS has been a fundamental approach for reconstructing 3D models from 2D imagery. Traditional MVS techniques such as *PMVS/CMVS* [1,12] and COLMAP [4,11] have achieved significant success by leveraging multiple overlapping images for scene reconstruction. These methods typically involve structure-from-motion (SFM) using feature extraction and matching to reconstruct the camera trajectory followed by depth map estimation, and dense fusion to create dense 3D reconstructions. However, many conventional MVS approaches face major challenges when it comes to scenes with transparent, reflective or featureless surfaces. Methods like [5,7] try to approach low-textured regions by extrapolating the estimation from coarser image resolutions to finer scales, consequently filling those areas. Transparent or reflective surfaces, such as glass windows orwater bodies, however, do not provide reliable photometric consistency, leading most of the time to erroneous depth estimates and gaps in the resulting reconstructions.

With the advent of deep learning, several approaches have been proposed to enhance traditional MVS methods. First to learn the similarity of image patches [13–15] in combination with conventional approaches like Semi-Global Matching (SGM) [16]. Methods such as MVSNet [17] and DeepMVS [18] instead aim to learn the process of depth estimation end-to-end. While these approaches have demonstrated improved accuracy, they are computationally expensive due to the high memory and processing requirements of the employed 3D CNNs. To address this, subsequent methods like R-MVSNet [19] and CVP-MVSNet [20] have adopted 2D-CNNs and used either recurrent architectures or cost volume pyramids in addition. Other methods use transformer-based architectures [21] instead of CNNs to predict dense depth maps [22–24]. While learning-based methods are good at incorporating the general context and thus learning complex concepts, a major drawback of these methods is the large amount of training data required, which necessitates an additional sensor such as a stereo camera or LiDAR device to capture the ground truth depth data.

To mitigate this problem, self-supervised trained methods use novel view synthesis between multiple viewpoints as a proxy for the supervised training signal, as the reprojection error between the sampled images can only be minimized if a reasonable depth estimate is learned. This can either be performed by projecting images left to right in a calibrated stereo setup [25] or by using image sequences [26] together with camera ego motion. Most recently, a new category of methods has received attention that uses this concept of novel view synthesis to implicitly represent the entire scene. Mildenhall et al. [8] model the volume as a continuous 5D representation in order to be able to estimate radiance and density for each point in the volume depending on the viewing angle. Building on this, many approaches extend this model by either improving the quality [9,27,28], accelerating the execution speed [10,29,30], using unstructured images [31] or focusing on large-scale reconstructions [32,33].

Since the main focus of this work is the reconstruction of large-scale scenes from aerial imagery and there is very little training data for this scenario, classically supervised learned methods are not a suitable option. Since these scenes also contain photogrammetrically challenging areas such as reflections, areas with poor texture and fine structures such as vegetation, the question arises as to what extent NeRF-based methods represent an alternative to conventional approaches. Since NeRFs have shown their capabilities for view synthesis of small-scale areas such as single object reconstruction or indoor scenes, there have been several works in the field of NeRFs suitable for large-scale scene reconstruction, mainly involving the combination of multiple NeRFs, which are expected to recover the 3D geometry of large scenes in a scalable manner and allow expanding with additional NeRFs or updating sub-modules without retraining the entire scene. We focus in particular on the following approaches for neural rendering: Mega-NeRF [32] based on the NeRF architecture of Mildenhall et al. [8], Block-NeRF [33] based on Mip-NeRF [9] and then Direct Voxel Grid Optimization (DVGO) [10] for a better trade-off regarding execution speed.

1.2. Outline of This Work

This work is structured as follows: First, in Section 2, we provide an overview of the methodologies employed for large-scale reconstruction, including Mega-NeRF, Block-NeRF, and DVGO as voxel-based approach. In addition, we cover the pre-processing routines, point cloud extraction strategies, and the evaluation metrics employed to assess the quality of the reconstructed scenes, as well as describing the datasets used for evaluation, highlighting their unique challenges and characteristics. In Section 3, we describe our experiments on various datasets and compare the achieved results to those of traditional photogrammetric approaches using COLMAP. Additionally, the section includes a detailed ablative study analyzing the effects of different configurations, such as various numbers of NeRF sub-modules, density thresholds, and the use of a VisibilityNet [33] for predicting occlusions. In Section 4, we discuss the results as well as draw conclusions and present suggestions for future work. We conclude by summarizing the findings in Section 5.

2. Materials and Methods

In the following section, we will briefly introduce the applied methodology, in particular the investigated NeRF variants, the pre-processing routines and the point cloud extraction strategy. We then explain the metrics and datasets used to analyze the derived point clouds and image-based metrics to compare the NeRF approaches with each other. Figure 1 shows a flowchart of the individual steps of our approach. The input data on the left consists of images and their camera calibration. First we cluster the scenes in multiple sub-modules, which are then reconstructed using three different NeRF-based approaches. Add the end the dense point clouds are extracted using randomly generated rays from the camera point of view.

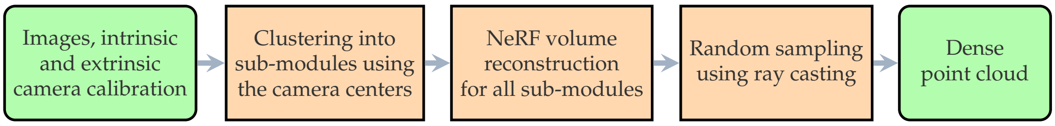
Images, intrinsic and extrinsic camera calibration
Clustering into sub-modules using the camera centers
NeRF volume reconstruction for all sub-modules
Random sampling using ray casting
Dense point cloud

Figure 1. Flowchart of our methodology. The input data consists of images and their camera calibration. Based on this, the images are clustered using the camera centers and assigned to individual sub-modules, which are then reconstructed as NeRF volumes. The point clouds are then extracted from the camera viewpoint using random sampling.

2.1. Scene Representation Using Neural Radiance Fields

The scene representation within a NeRF is a continuous 5D function called radiance field, which is capable of representing a static scene with complex geometry [8]. A scene is represented by a function that outputs the radiance emitted in each viewing direction $\mathbf{d} = (\phi, \psi)^T$ at each spatial location $\mathbf{p} = (x, y, z)^T$ and the density at each point which can be interpreted as the probability of a ray terminating at the location $\mathbf{p}$. This scene representation within a NeRF is represented by regressing from a single 5D coordinate to a volume density and view-dependent color:

$$MLP_{\Theta} \colon (\, \mathbf{p}, \mathbf{d}\,) \rightarrow (\mathbf{r}, \theta), \tag{1}$$

where $\mathbf{r}$ denotes the emitted color and θ the volume density. The approximation function MLP_{Θ} is a simple Multi-Layer Perceptron (MLP). The viewing direction $\mathbf{d}$ is expressed as a 3D Cartesian unit vector in practice. Within a NeRF, the direct optimization of the weights Θ is intended to allow mapping of each 5D input to its corresponding volume density and view-dependent emitted color. This mapping is designed to be multiview consistent by restricting the MLP to predict the volume density θ as a function of only the location $\mathbf{p}$ and to predict the color as a function of both the location and the viewing direction [8].

2.2. Scalable Reconstruction with Mega-NeRF

The first examined approach is Mega-NeRF, whose core contribution is the idea to train multiple NeRFs so that each of these sub-modules is specialized to a specific region of a scene [32]. The introduced initialization based on clustering splits the training images pixel-wise for parallel training on multiple GPUs. Based on the geometry-aware viewpoints estimated by SFM, Mega-NeRF decomposes a scene into N clusters with centroids $\bar{\mathbf{c}}^i = (\bar{c}^i_x, \bar{c}^i_y, \bar{c}^i_z)$ for $i = 1, ..., N$ and initializes each sub-module NeRF f^i_{Θ}. The continuous 5D scene representation is approximated with multiple modules:

$$MLP^i_{\Theta} \colon (\, \mathbf{p}, \mathbf{d}\,) \rightarrow (\mathbf{r}, \theta), \tag{2}$$

where $i = \arg\min_{i \in N} ||\bar{\mathbf{c}}^i - \mathbf{p}||^2_2$ denotes the closest sub-module index for each 3D point $\mathbf{p}$. Together with the viewing direction $\mathbf{d}$, radiance $\mathbf{r}$ and density θ can be predicted for this position. The construction of a large-scale NeRF is achieved through the training of individual sub-modules which may overlap. This enables the rendering of viewpoints with one or more sub-module NeRFs, which is beneficial when rays of a viewpoint fall within the boundaries of multiple modules. In such cases, the color and density of a point is determined by the relevant NeRFs in proportion to the distance to the cluster centroids. Mega-NeRF's disadvantage compared to the other methods is that it needs a parameter that indicates the lowest point to keep the ray point sampling of NeRF above ground level [32].

2.3. Scalable Reconstruction with Block-NeRF

Block-NeRF is another method that partitions a large scene into multiple regions so that individually trained NeRFs enable scalable reconstructions [33]. Its main difference from Mega-NeRF is that its core NeRF architecture called Mip-NeRF implements a multiscale representation for anti-aliasing [9]. Additionally, a small MLP for each NeRF is proposed for visibility prediction.

The core NeRF architecture of Block-NeRF, Mip-NeRF, is proposed to deal with multi-scale representation better than the classic NeRF architecture [9]. Rendering NeRF based on a single ray per pixel is known to work well when all training and test images are taken from approximately the same distance [8]. However, it is not applicable for training images at multiple resolutions, because the renderings may show artifacts in close-up views or aliasing in distant views. Mip-NeRF is inspired by the prefiltering-technique called mipmap that precomputes a signal at a set of different discrete down-sampling scales. The prefiltering strategy of Mip-NeRF proposes casting a 3D conical frustum defined by a camera pixel, of which rendering at intervals along the cone corresponds to rendering a prefiltered pixel.

In order to combine relevant NeRF sub-modules while handling occlusions, Block-NeRF proposes an additional MLP for each sub-module. This MLP is designed to predict whether a specific point in space is visible to its surrounding NeRF sub-modules. This enables the merging of only relevant sub-modules. For each sampled point along a ray, location and viewing direction information are fed to predict the corresponding transmittance of each point, which describes the probability that the ray travels without hitting any other particle before reaching its target. To render a camera perspective, the visibility of all rays for all neighboring NeRFs is predicted and those with visibility below a threshold are discarded. The estimates are merged based on the distance to each block's cluster centroid.

2.4. 3D Representation Using Direct Voxel Grid Optimization

Furthermore, with DVGO we are examining a voxel-based approach in addition to the above mentioned techniques. These methods have been demonstrated to produce high-quality and flexible view synthesis results that are comparable to those achieved by NeRF, while offering significant computational speed improvements [10]. This study assesses the effectiveness of volume rendering based on voxel grids in achieving a precise reconstruction of geometries. Building on the approach of developing multiple NeRFs for large-scale scene reconstruction, this work explores the potential of creating multiple voxel grids for scalable geometry reconstruction of large scenes. The voxel grid representation is capable of explicitly modeling the modalities of interest, such as density or color, in its grid cells. This enables efficient querying for any 3D position via interpolation. The direct optimization of the volume density represented in a dense voxel grid is conducted in two phases. Coarse geometry searching and fine detail reconstruction. It is advantageous to identify the coarse 3D areas and free space during the coarse geometry searching phase, prior to starting the fine detail reconstruction phase. Coarse geometry searching combines the density voxel grid with a color voxel grid in order to capture view-invariant color. Fine detail reconstruction employs a feature voxel grid with a shallow MLP in order to model view-dependent color [10]. The maximum resolution of the voxel grid is 15.6 million voxels, which corresponds to 2.5 million voxels per dimension.

2.5. Pre-Processing for NeRF-Based Training

To train the models, the data must first be pre-processed. For this, we assume that the camera intrinsic and extrinsic parameters are known, and all coordinates of the camera centers are normalized to the range $[-1, 1]$. This normalization is recommended for reliable training of the NeRF MLP, based on the positional encoding [8,34], and is applicable for the evaluated datasets, since the nadir camera perspective ensures the normalized reconstructed points are expected to lie within the expected range. The nadir camera perspective also allows the omission of the background NeRF [32,35] mentioned in Mega-NeRF, since pixels corresponding to objects that are very far away, such as the sky, are not visible. Regarding the partitioning into individual sub-modules for Mega-NeRF and Block-NeRF, we use K-Means clustering of the camera centers according to the work in [32]. An example of four clusters is provided in Figure 2, in which the camera frustums and their cluster assignments are color-coded.

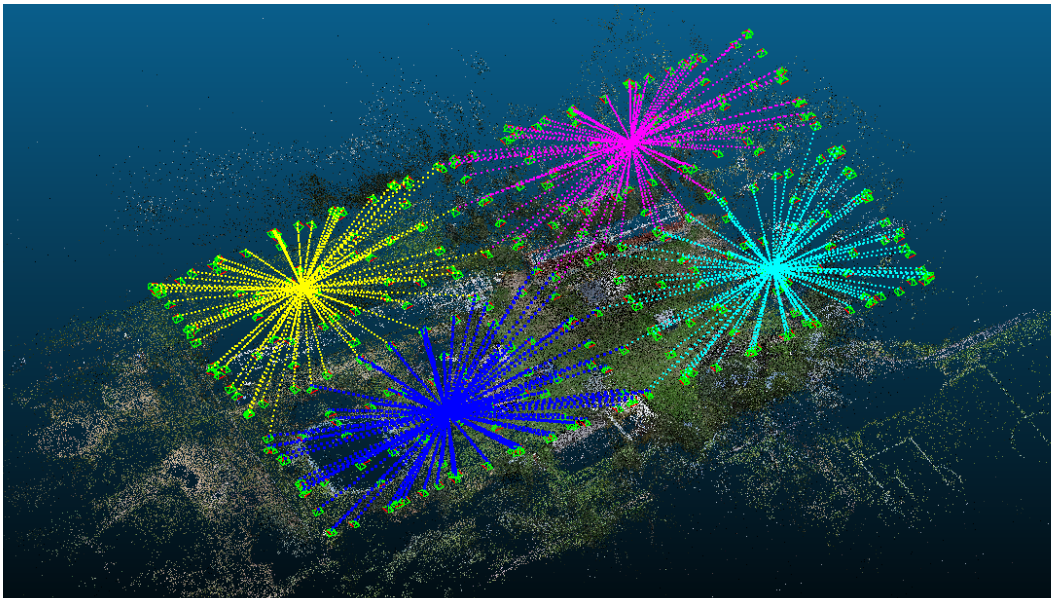

Figure 2. Visualization of the cluster pre-processing based on the position of the camera centers. The colors symbolize the four sub-modules.

2.6. Rendering Strategy to Extract Point Clouds from NeRFs

Assuming that the volume density in NeRF volume rendering can be interpreted as the probability of a ray terminating at a small particle, each point cloud is rendered by exporting the points with the maximum volume density value along each ray. This can be performed not only for each image pixel, but also for sub-pixels, which results in very dense point clouds. In order to ensure comparability between NeRF volume rendering-based reconstructions and COLMAP dense reconstructions, we sample approximately the same number of points in the reconstructed point clouds. To obtain these 3D points, random uniformly sampled rays are generated from the view of all cameras and the position and color along the ray are extracted at the location with the maximum density above a selected threshold value.

2.7. Evaluation Metrics

Since the focus of this work is the reconstruction of point clouds, we focus on computing the accuracy using the L1-absolute error and the completeness to measure the coverage of the ground truth point clouds following Seitz et al. [36]. The accuracy of a reconstructed model is determined by the distance of the points in the estimated model to the respective ground truth points, which is quantified by the absolute L1 distance. This is calculated as the mean distance from points in the reconstructed model to their nearest neighboring point in the ground truth model. The completeness of a reconstructed model is determined by the extent to which the ground truth model is represented by the estimated model, based on the distances between the two models. The completeness of a reconstructed model is defined as the ratio between the number of neighboring points from the ground truth model to the reconstructed model. The extent of the neighborhood is defined by a threshold parameter X_{th}. The L1-abs and completeness are formalized as follows:

$$\text{L1-abs}(\mathcal{P}_{est}, \mathcal{P}_{gt}) = \frac{1}{|\mathcal{P}_{est}|} \sum_{\mathrm{p} \in \mathcal{P}_{est}} |\mathcal{P}_{est}(\mathrm{p}) - \mathcal{P}_{gt}(\mathrm{p})| \tag{3}$$

where L1-abs$(\mathcal{P}_{est}, \mathcal{P}_{gt})$ denotes the accuracy of the reconstructed model $\mathcal{P}_{est}$ given the ground truth model $\mathcal{P}_{gt}$. The accuracy is normalized by the number of points in the point cloud $\mathcal{P}_{est}$. Following [36], we rank the points by their distance to the nearest ground truth point and select 90 % of the points with the shortest distance to exclude outliers and noise as much as possible.

$$\mathrm{Cpl}_{X_{th}}(\mathcal{P}_{est}, \mathcal{P}_{gt}) = \frac{1}{|\mathcal{P}_{gt}|} \sum_{\mathrm{p} \in \mathcal{P}_{gt}} [|\mathcal{P}_{est}(\mathrm{p}) - \mathcal{P}_{gt}(\mathrm{p})| < X_{th}] \tag{4}$$

$\mathrm{Cpl}_{X_{th}}(\mathcal{P}_{est}, \mathcal{P}_{gt})$ denotes the completeness of the reconstructed model $\mathcal{P}_{est}$ given the ground truth model $\mathcal{P}_{gt}$ and the threshold X_{th}. To calculate the completeness, we use a threshold value X_{th} of 0.2 m. Subsequently, all points from $\mathcal{P}_{gt}$ with at least one point from $\mathcal{P}_{est}$ in their neighborhood X_{th} are included. We use the in [37] published evaluation routines for the calculation of accuracy and completeness

Since the primary objective of NeRF-based approaches is to generate synthetic images through novel view synthesis, the loss function for the training routine is derived from the goal of learning to reconstruct a set of test images given the training data. The synthesized images are usually evaluated using specific metrics to measure image quality. In addition, to being used at the time of training, these metrics can also be employed to compare the methods with each other. In the following, we use three metrics: Structural Similarity Index (SSIM) [38], Peak Signal-to-Noise Ratio (PSNR) and Learned Perceptual Image Patch Similarity (LPIPS) [39]. While LPIPS is a learned metric based on deep feature maps of a neural network, SSIM and PSNR can be formalized as follows:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{5}$$

whereby μ_x and μ_y represent the pixel sample means of the original image x and the reconstructed image y respectively. σ_x^2 and σ_y^2 denote the variances of x and y, while σ_{xy} is the covariance of x and y. The constants c_1 and c_2 are introduced to stabilize the division in the equation, where $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$. Here, k_1 and k_2 are constants, and L represents the dynamic range of the pixel values.

$$PSNR(x, y) = 10 \cdot \log_{10}\left(\frac{MAX_I(x, y)^2}{MSE(x, y)}\right) \tag{6}$$

Given MAX_I of the original image x and the reconstructed image y as the maximum pixel value and MSE as the mean square error, which measures the mean square differences between the original and reconstructed image pixels.

2.8. Datasets Used for Evaluation

In order to evaluate the reconstruction methods, three different datasets are considered in this study: The two public benchmark datasets UseGeo [40] and Hessigheim 3D [41], which each have ground truth point clouds, and the non-public TMB dataset. These three datasets contain both urban settings with buildings and roads as well as areas that are particularly challenging for classical photogrammetric approaches, such as vegetation or water surfaces.

The UseGeo dataset consists of three sub-datasets, referred to as UseGeo1-3, which are independent of each other and each have ground truth camera extrinsic and intrinsic parameters as well as a LiDAR point cloud. In contrast, the Hessigheim 3D dataset also provides a LiDAR point cloud, but the images used were gathered at a slightly different time, which leads to inconsistencies in dynamic objects such as vehicles and vegetation, which limits a quantitative evaluation. In this case, the camera calibration was carried out using COLMAP. Due to the large number of very high-resolution images available in this case, these images were down-sampled by a factor of 4 for further processing. These images belong to the non-public part of the Hessigheim dataset. The TMB dataset was used to analyze the performance especially in rural areas as it contains large amounts of vegetation. However, as LiDAR data is not available, we reconstructed a reference point cloud using COLMAP. The UseGeo and Hessigheim 3D datasets also feature challenging terrain, as they include hills. All images were acquired by UAVs in the nadir camera perspective and feature an overlap typical for classical photogrammetric approaches. Table 1 summarizes the characteristics of the investigated datasets.

Table 1. Overview of the datasets investigated. The camera poses of the TMB and Hessigheim 3D datasets were estimated with COLMAP. In the case of the TMB dataset, the ground truth point cloud was also reconstructed using COLMAP. *Although the point cloud of the Hessigheim 3D dataset was recorded with a LiDAR scanner, it was acquired at a slightly different point in time, which leads to errors, especially with dynamic objects.

Name	Number of Images	Resolution	Camera Poses	Point Cloud	Altitude (m)
TMB	329	3956 × 2973	COLMAP	COLMAP	47
UseGeo1	225	7953 × 5279	Provided	LiDAR	102
UseGeo2	328	7954 × 5279	Provided	LiDAR	122
UseGeo3	278	7955 × 5279	Provided	LiDAR	113
Hessigheim 3D	2858	3551 × 2663	COLMAP	LiDAR*	97

3. Results

In the following section, we present the results of the three different methods and compare them to those achieved by classical photogrammetric approaches. In addition, we perform a detailed ablative study to analyze the effects of different configurations like using various numbers of NeRF sub-modules, varying the density threshold of the rendering strategy and the potential improvements using a VisibilityNet to predict occlusions.

3.1. Photogrammetric Reconstructions Using COLMAP

In order to compare the NeRF-based reconstructions with classical photogrammetric reconstructions, we used COLMAP to reconstruct the datasets UseGeo1-3 and Hessigheim. Each reconstruction is evaluated with respect to accuracy using a 90% threshold and calculating the completeness with a distance threshold of 0.2 m. As can be seen in the last three rows of Table 2, the accuracy on the UseGeo dataset for each sub-dataset is lower than 8 cm, with a completeness of just under 60%. The error on the Hessigheim 3D dataset is significantly higher, which is due to the time offset between the images and the LiDAR ground truth. For this reason, only a qualitative evaluation is feasible at this point. In the qualitative results in the following figures, COLMAP is listed as a reference in the first row. A complete LiDAR scan for the TMB dataset is not available, which prevents the evaluation of COLMAP in this context. Instead, COLMAP is utilized as a ground truth for the NeRF methods in order to assess their potential.

Table 2. Overview of NeRF volume rendering based reconstructions. View synthesis is evaluated with one central test image due to lack of images especially on the UseGeo datasets. Due to our implementation of DVGO, the image-based metrics were only calculated for the variant with one sub-module. The last three rows show the COLMAP dense reconstructions of the UseGeo sub-datasets as a reference.

No.	Dataset	Method	No. Clusters	PSNR↑	SSIM↑	LPIPS↓	*acc*↓	*comp*↑
1	TMB	Mega-NeRF	2	20.72	0.518	0.597	0.726	0.368
2	TMB	Block-NeRF	1	20.19	0.501	0.616	0.995	0.280
3	TMB	Block-NeRF	2	19.35	0.492	0.608	0.852	0.342
4	TMB	Block-NeRF	4	20.41	0.509	0.604	0.730	0.677
5	TMB	DVGO	1	20.03	0.491	0.609	1.884	0.053
6	TMB	DVGO	2	-	-	-	2.180	0.150
7	TMB	DVGO	4	-	-	-	2.461	0.189

Table 2. *Cont.*

No.	Dataset	Method	No. Clusters	PSNR↑	SSIM↑	LPIPS↓	*acc*↓	*comp*↑
8	UseGeo1	Mega-NeRF	1	20.75	0.463	0.684	1.089	0.172
9	UseGeo1	Block-NeRF	1	20.71	0.462	0.691	1.141	0.151
10	UseGeo1	Block-NeRF	2	20.53	0.461	0.688	0.909	0.183
11	UseGeo1	Block-NeRF	4	21.15	0.469	0.676	0.622	0.218
12	UseGeo1	DVGO	1	18.79	0.443	0.675	2.818	0.027
13	UseGeo1	DVGO	2	-	-	-	2.453	0.034
14	UseGeo1	DVGO	4	-	-	-	1.953	0.043
15	UseGeo2	Mega-NeRF	1	16.43	0.500	0.690	0.932	0.159
16	UseGeo2	Mega-NeRF	3	20.62	0.523	0.678	0.631	0.252
17	UseGeo2	Block-NeRF	1	15.61	0.491	0.692	1.027	0.153
18	UseGeo2	Block-NeRF	3	18.00	0.508	0.685	0.739	0.192
19	UseGeo2	DVGO	1	17.58	0.504	0.670	2.537	0.040
20	UseGeo2	DVGO	3	-	-	-	2.020	0.048
21	UseGeo3	Mega-NeRF	1	19.57	0.624	0.637	0.491	0.235
22	UseGeo3	Mega-NeRF	3	20.61	0.632	0.621	0.313	0.315
23	UseGeo3	Block-NeRF	1	20.61	0.627	0.640	0.461	0.240
24	UseGeo3	DVGO	1	18.49	0.614	0.636	2.742	0.036
25	UseGeo3	DVGO	3	-	-	-	2.036	0.053
26	UseGeo1	COLMAP	-	-	-	-	0.078	0.541
27	UseGeo2	COLMAP	-	-	-	-	0.069	0.596
28	UseGeo3	COLMAP	-	-	-	-	0.060	0.591

3.2. NeRF Volume Rendering Based Reconstructions

In the following section, we discuss the results based on NeRF approaches and present extensive quantitative and qualitative results. Table 2 provides an overview of the quantitative results broken down by dataset and metric. The same threshold values are used for the calculation of accuracy and completeness as for the COLMAP reconstructions. In addition, the image-based metrics PSNR, SSIM, and LPIPS [38,39] are listed in order to compare the methods with each other. Since the UseGeo dataset in particular has very few images, the image-based metrics are only calculated on one central test image. With regard to DVGO, due to limitations of our own implementation of multiple voxel grids, the image-based metrics were only calculated for the variant with one sub-module. The evaluation results show that the accuracy and completeness of the NeRF volume rendering based point cloud reconstructions are generally not competitive with the dense COLMAP reconstructions, shown in the last rows of Table 2.

However, except for the combination of TMB dataset and DVGO, all results show that increasing the number of clusters leads to improved accuracy and completeness, presumably because a larger number of sub-module NeRFs reduces the area each module has to cover, thus refining the sampling of the scene. An example of this is the increase to a completeness of 67% when four sub-modules are used on the TMB dataset with Block-NeRF. Point cloud reconstructions based on the voxel grid representation generally have lower accuracy and completeness. Although the reconstructions look visually acceptable, this is not reflected in the quantitative results, as the results do not compare favorably with Mega-NeRF or Block-NeRF reconstructions in terms of accuracy and completeness.

As can be seen from the qualitative results, the three methods manage to reconstruct the coarse context of the scenes. The first column of the Figures 3–5 in particular visualizes this well, as the point clouds are displayed directly from above, which is most similar to the nadir perspective of the training data. It can also be seen that the NeRF-based reconstructions capture significantly more area of the scene and produce much more visually complete results, especially in areas that are considered photogrammetrically challenging, such as dense vegetation and water surfaces. This is particularly visible in the TMB dataset in Figure 3, where the scene is enclosed by rows of trees and contains a small lake, and in the results of the Hessigheim 3D dataset in the second column of Figure 5, where a lot of water is visible.

However, as soon as the point cloud is viewed from an angle that deviates significantly from the nadir perspective, the visual quality suffers considerably. As can be seen in the figures, COLMAP in contrast can reconstruct significantly more details. Especially the experiments with DVGO show strong weaknesses with finer details. It can also be seen that Mega-NeRF and Block-NeRF have difficulty with the overall geometry of the scene, as both the UseGeo datasets and Hessigheim 3D dataset contain terrain of varying heights, which is only partially reflected in the results. As can be observed in the last columns of Figure 4, only DVGO is capable of reconstructing the rough geometry of the scene. With Mega-NeRF in particular, the question arises as to what extent the parameter for the ground level has a strong influence on the results, as many of the sampled 3D points are significantly above this defined ground level. This suggests that this specification of the lowest permitted point and the significant deviation from this in some cases may cause challenges.

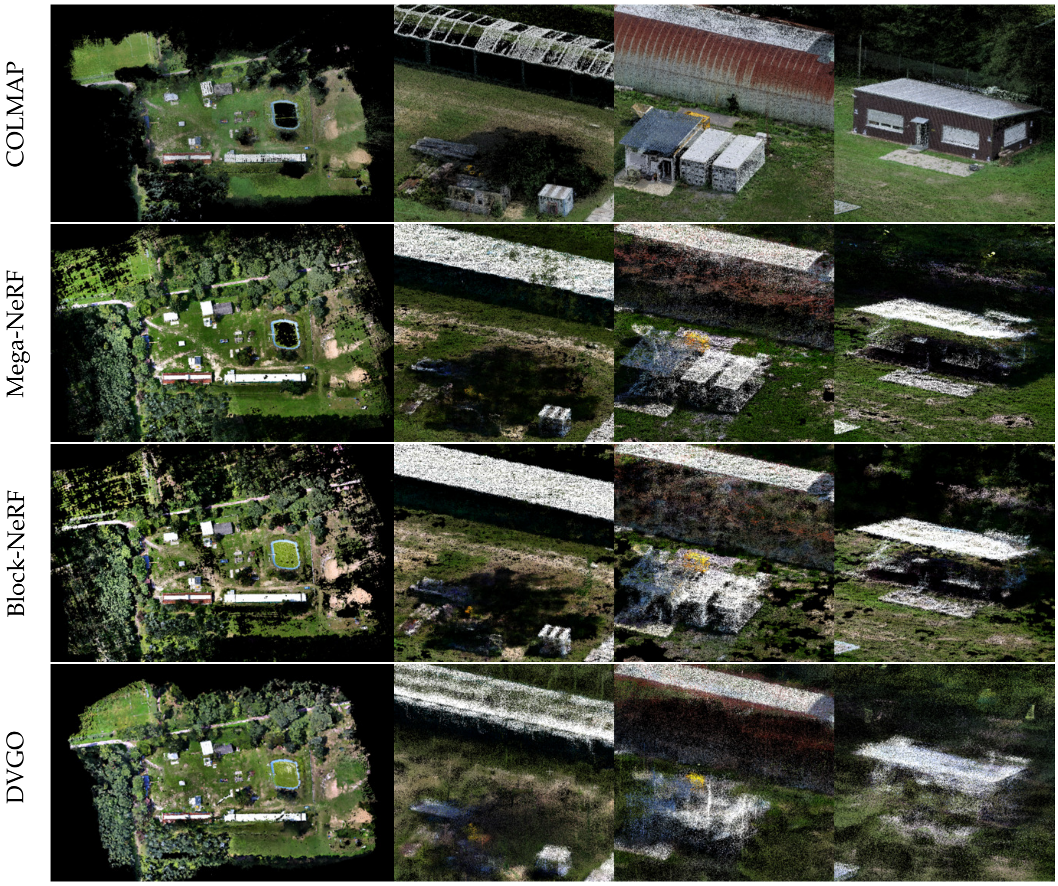

Figure 3. Point clouds derived for the TMB dataset. Each row shows examples of one approach, with COLMAP as a reference in the first row.

3.3. Impact of the Density Threshold

A common side effect of using filtering techniques in classical MVS approaches to reduce outliers and noise is that the accuracy is increased by removing incorrect points, but the completeness decreases, as some accurate points are also incorrectly removed. Regarding NeRF-based methods, it is of interest whether a similar effect occurs when the density threshold is varied. As described in Section 2.6, the individual points of the point cloud are selected along randomly generated rays from the camera viewpoints using the density above a pre-defined threshold. In the following, we investigate the effects of different density thresholds by analyzing the changes in completeness and accuracy. Figure 6 visualizes different configurations by altering the internal density threshold between 0 and 200. A clear decrease in completeness can be seen, particularly in areas like canals and roads.

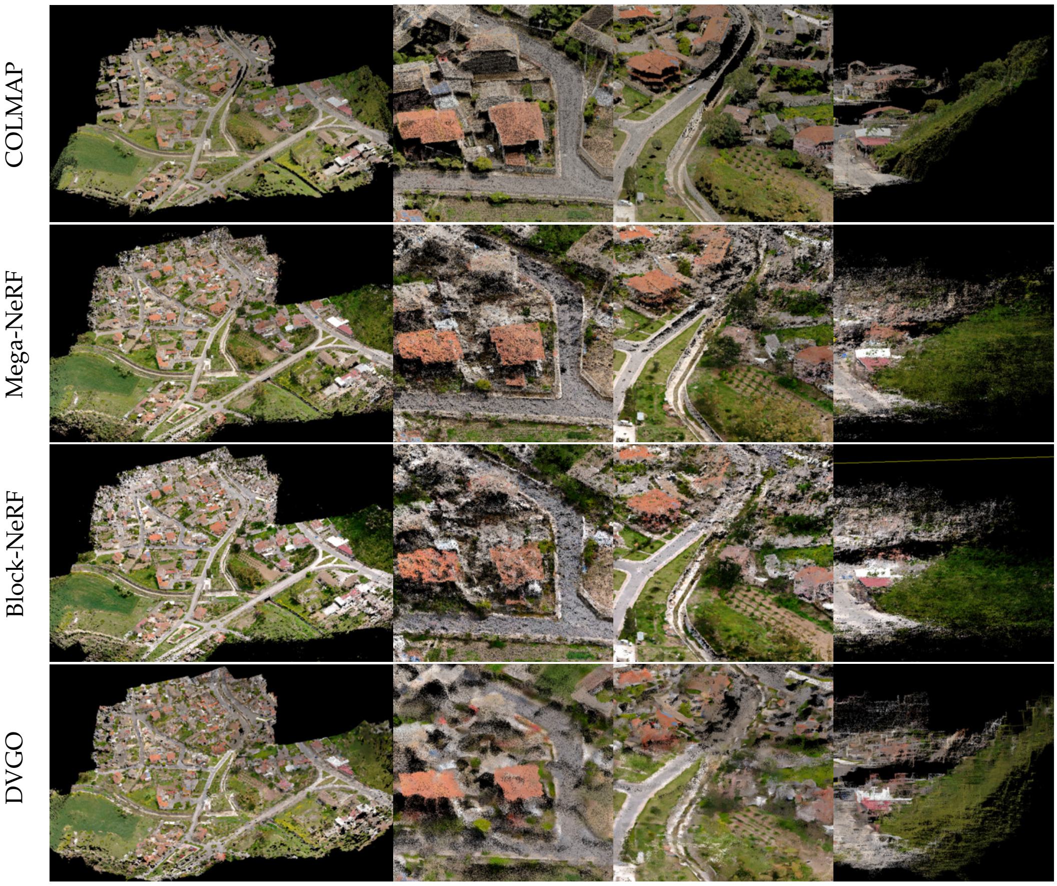

Figure 4. Point clouds derived for the UseGeo3 dataset. Each row shows examples of one approach, with COLMAP as a reference in the first row.

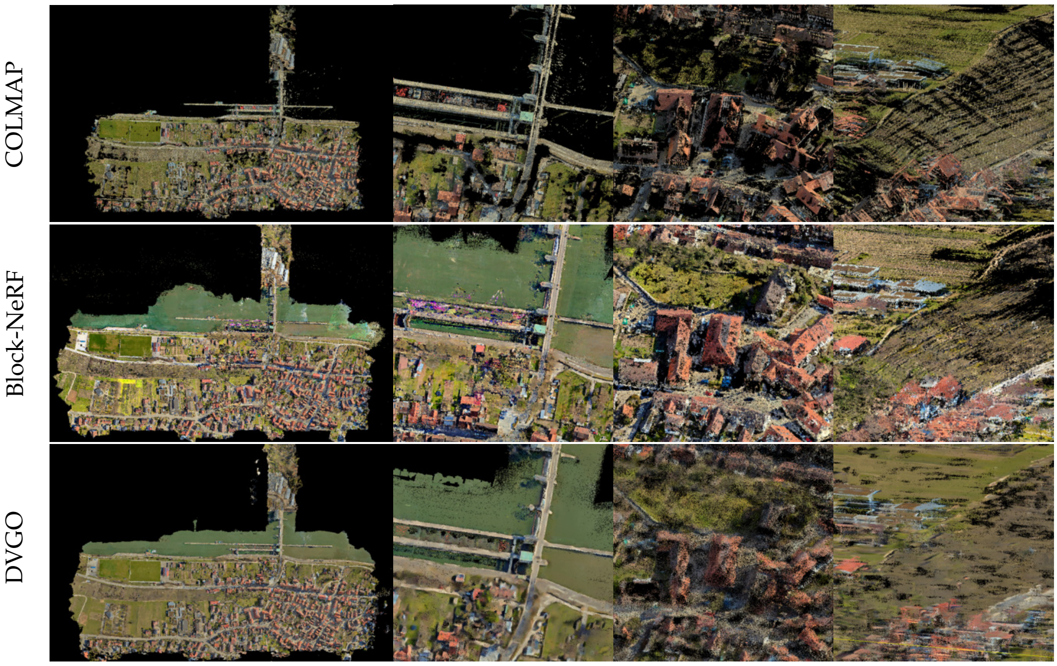

Figure 5. Point clouds derived for the Hessigheim 3D dataset. Each row shows examples of one approach, with COLMAP as a reference in the first row.

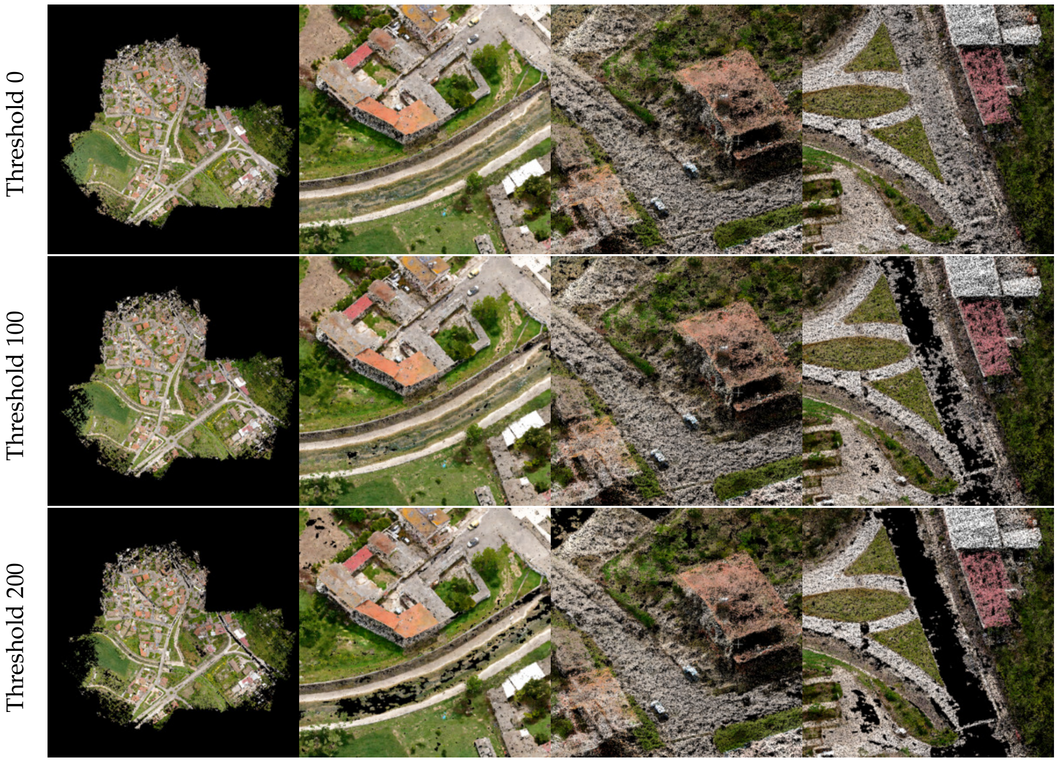

Figure 6. Qualitative evaluation of different density thresholds using Mega-NeRF. As the density threshold increases, roads and canals in particular show significantly more holes.

The visualization in Figure 7 shows the relationship between accuracy, completeness and the density threshold. Six experiments, each with different density thresholds, are color-coded. The higher the opacity of the color, the higher the density threshold for the point cloud extraction process. The letters M, B and D correspond to the methods used: Mega-NeRF, Block-NeRF and DVGO. The number in front corresponds to the position in Table 2. The number after it corresponds to the number of sub-modules. The black arrow indicates the approximate direction of the values over the increasing threshold. As illustrated in Figure 7, the overall completeness of all experiments declines as the density threshold value increases. However, with the exception of experiment 10B2, the accuracy of all experiments improves. This is similar to the effect of filtering as used in many methods, including conventional photogrammetry. A strong filtering of points with higher uncertainty leads to higher accuracy at the expense of less complete results.

3.4. Effect of Filtering by Visibility

Especially in urban areas buildings and street canyons can lead to a lot of occlusions, making it difficult to estimate which point is best seen from which camera angle. To predict this visiblity during training Tancik et al. [33] propose a second model in addition to the NeRF model which predicts the visibility of a ray. This involves a small MLP to regress the transmittance of each point, which indicates the probability that the ray travels without hitting any object. If the value is above a previously defined threshold, the point in space is considered visible from this sub-module. This visibility prediction during training has inspired the investigation of whether this network also can improve the quality during extraction of the point cloud. This is based on the idea that a NeRF should produce more accurate geometries for clearly visible points. In reconstructions with the VisibilityNet, each point with the maximum density value along the ray is examined to ensure that its visibility value, the output of the integrated position encoding of 3D coordinates from the VisibilityNet, is above 50%.

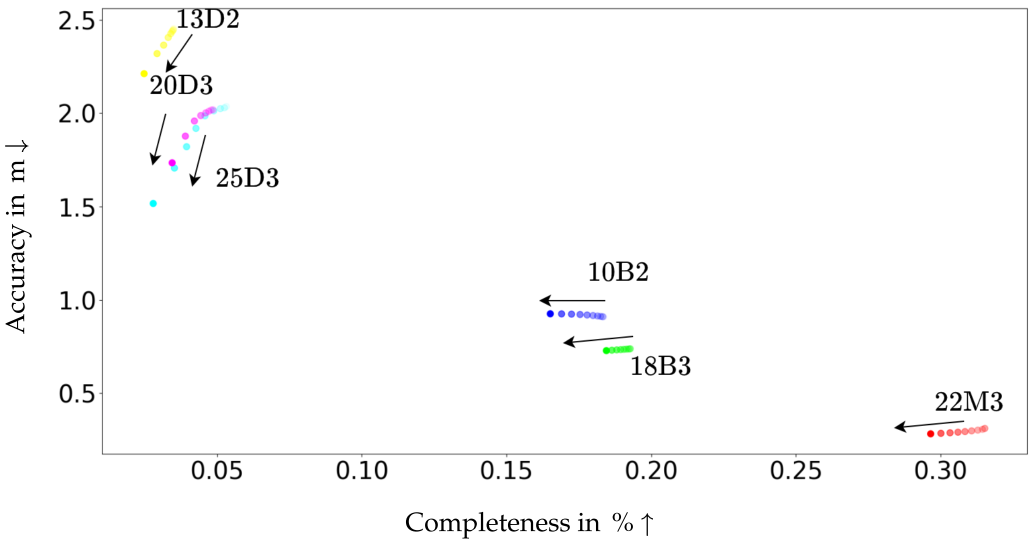

Figure 7. Relationship between accuracy, completeness and the density threshold. The letters represent the three methods analyzed: Mega-NeRF (M), Block-NeRF (B) and DVGO (D). The numbers in front of them refer to their position in Table 2, whereas the number after the letter indicates the number of sub-modules. The black arrows indicate the approximate direction of the values over the increasing threshold.

As can be seen in Table 3, this brings great added value, especially on the UseGeo dataset. Not only does the accuracy of the point clouds increase, but also the completeness. This could be due to the fact that we decided to extract the same number of points for all experiments, comparable to the number of points reconstructed by COLMAP. Because the visibility filtering is applied beforehand, the same number of points is still extracted as in all other experiments, but the probability to sample a better point may increase as a result. Figure 8 shows examples of qualitative results with and without the visibility prediction when extracting the point clouds. It can be seen here that the point clouds are significantly less noisy.

Table 3. Evaluation results of point cloud reconstructions extracted with and without VisibilityNet. The quality improves with VisibilityNet for all experiments and shows the potential of visibility filtering in addition to density filtering.

			acc↓		*comp*↑	
No.	**Dataset**	**No. Clusters**	**Without VisNet**	**With VisNet**	**Without VisNet**	**B**
2	TMB	1	0.995	0.963	0.280	0.282
3	TMB	2	0.852	0.806	0.342	0.344
9	UseGeo1	1	1.141	0.442	0.151	0.191
10	UseGeo1	2	0.909	0.398	0.183	0.215
11	UseGeo1	4	0.622	0.332	0.218	0.242
17	UseGeo2	1	1.027	0.677	0.153	0.182
18	UseGeo2	3	0.739	0.505	0.192	0.220
23	UseGeo3	1	0.461	0.385	0.240	0.230

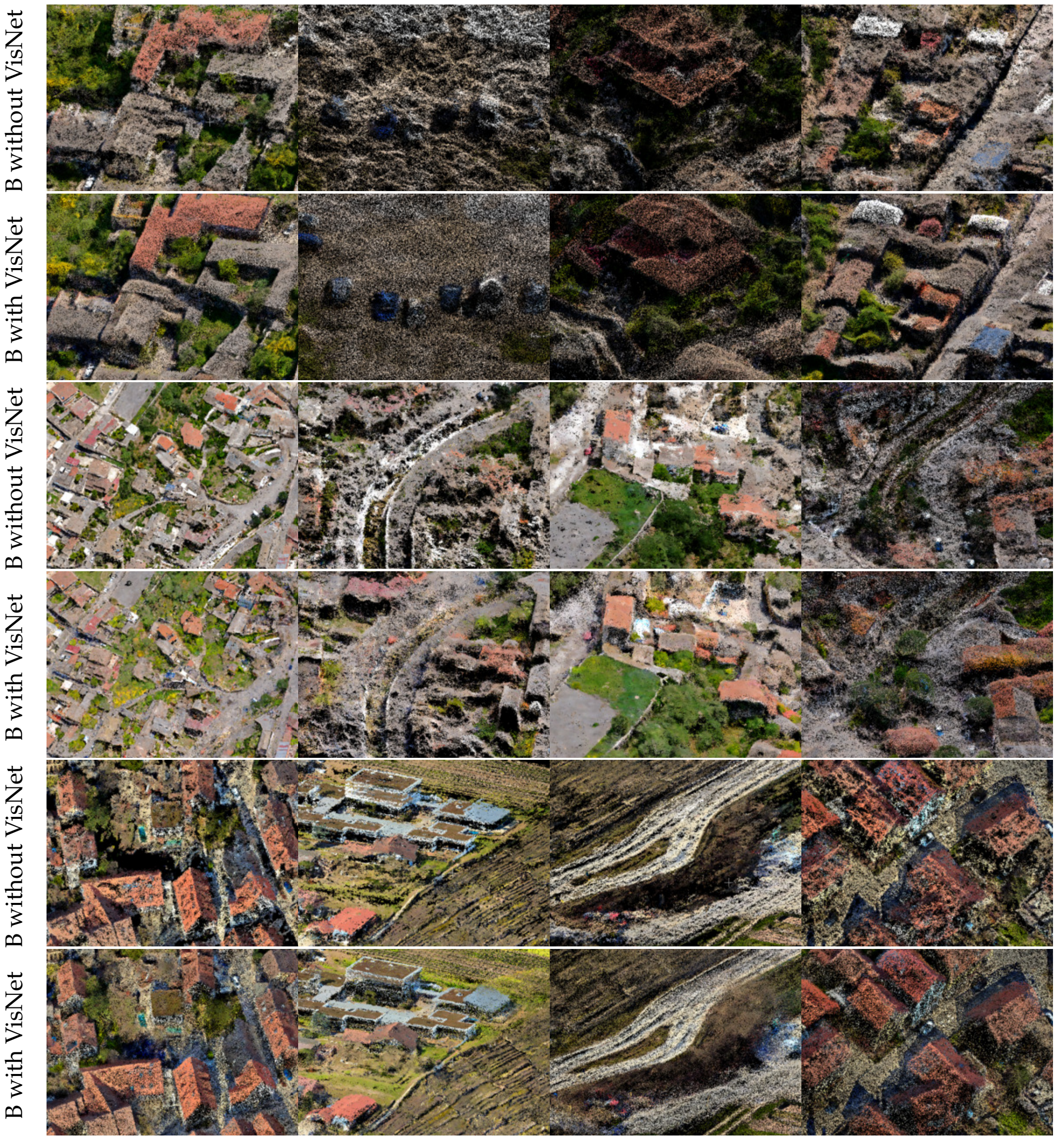

Figure 8. Reconstructions with and without VisibilityNet derived for the UseGeo and Hessigheim 3D dataset. B in the first column denotes Block-NeRF. Point clouds extracted using the VisibilityNet show a better level of detail in general.

4. Discussion

The results of our investigation indicate that the performance of NeRF-based methods for the reconstruction of dense point clouds is not yet at a level comparable to that of classic photogrammetric approaches, such as COLMAP. Both qualitatively and quantitatively, the COLMAP models appear to be of higher quality. However, it can be seen in Figures 3 and 5 that challenging regions such as fine vegetation and reflective water as well as textureless regions can be represented in a visually convincing way using NeRF models. One explanation for the inferior performance is most likely the image acquisition of the employed datasets, with nadir perspective and low image overlap. This represents the conventional setup for classical photogrammetry, which is therefore very different from the experiments shown for Mega-NeRF or Block-NeRF in [32,33], as they use several thousand images with different camera angles. Often these datasets are created synthetically, as described in [8], or extracted from videos or dense image sequences, as in the case of the Tanks & Temples dataset [42]. It is also notable that the investigated methods were originally developed with the primary objective of rendering photo-realistic scenes using novel view synthesis. The utilization for extracting the point cloud is a secondary objective, in which the density representation is exploited. It is therefore reasonable to suggest that a combination of both classic and NeRF approaches would be beneficial to achieve an accurate reconstruction of the regions, while also utilizing NeRF-based techniques to fill in the remaining areas.

The conducted experiments suggest that splitting the scene into individual sub-modules can be a promising strategy. We observed that both accuracy and completeness improve as the number of sub-modules increases. Similarly, the use of a neural network to predict visibility of 3D points leads to an improvement in both metrics when it comes to extracting the point cloud. In contrast, adjusting the density threshold to extract the 3D points improves accuracy, but the completeness suffers considerably. This can be seen in Figure 6, where areas such as roads are particularly affected, indicating an inadequate representation in the NeRF model. Instead of using only the maximum density along a ray, one could also extract all points above the threshold, but our experiments showed that sometimes there are very dense points inside of objects, although there is no information in the images for these points, leading to errors in the point cloud. In contrast, another strategy is to select the first point along the ray that is above the threshold, but in our experiments this led to more floating artifacts. Perhaps an adaptive threshold value would be more suitable here, for example in combination with predicting the visibility or exploiting available semantic information.

To account for the specific acquisition conditions of our datasets, which consist of only nadir imagery with low overlap, we also briefly investigated the impact of ambiguous positional encoding. Mildenhall et al. [8] have shown in their work that transforming the training data into a higher dimensional space by representing it as a periodic signal is a very important pre-processing step. As shown in Figure 9 on the left, the nadir perspective can result in a degenerated point cloud rendered in multiple layers, as presumably local minima during training lead to a sub-optimal solution. One widely used solution is to concatenate the vector of 3D coordinates with the transformed periodic signal, which can help to resolve this positional ambiguity. While this identity function is not explicitly described by Mildenhall et al. [8], it can be retrieved from their open source repository. It has been shown that this approach significantly reduces the occurrence of these artifacts in the datasets examined, as can also be seen in Figure 9 on the right.

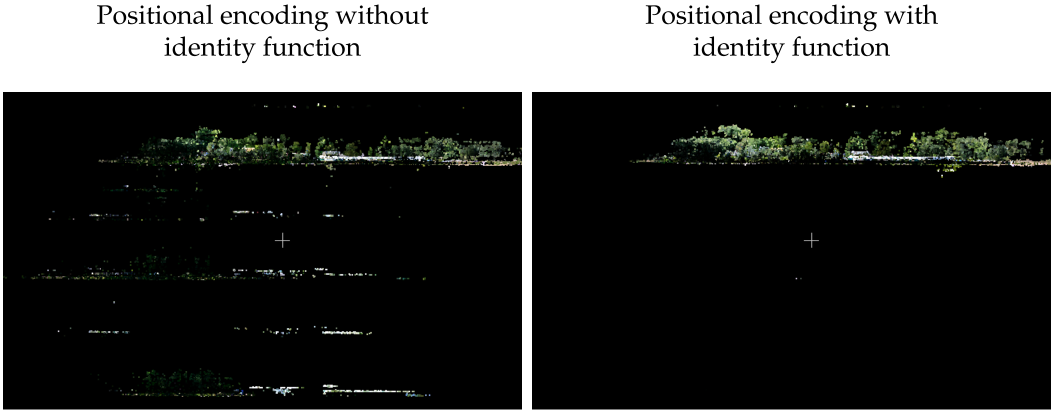
Positional encoding without identity function
Positional encoding with identity function

Figure 9. Mip-NeRF based point cloud reconstruction without and with an identity function for the positional encoding, which means additionally concatenating the original input coordinates. Positional encoded features without an identity function lead to points that are scattered over several periodic layers, whereas with an identity function these artifacts disappear.

The comparison of the NeRF methods reveals that Mega-NeRF consistently produces better results across all experiments, despite Block-NeRF being based on a newer architecture. This may be attributed to the manually defined parameter for the ground plane in Mega-NeRF, which could help during training. In addition, the image acquisition setting using the nadir perspective might be better suited for Mega-NeRF, since the images always have the same resolution and have a very similar distance from the scene. This should be well suited to the basic NeRF configuration and the improvements introduced in Mip-NeRF for non-equidistant images and different resolutions are probably of little importance. In the future, it would be interesting to see to what extent the use of a VisibilityNet during training and point cloud extraction can further improve the results of Mega-NeRF.

Although the accuracy and completeness of point cloud reconstructions based on DVGO are not competitive to the results of Mega-NeRF or Block-NeRF in this work, the reconstructions based on the voxel grid representation are at least capable of correctly capturing coarse geometries of scenes. Analogous to the other two approaches, our adaptation of DVGO to use multiple voxel grids leads to better results than the use of a single voxel grid, showing that splitting the scene into multiple independent modules and then combining them can be a viable approach in general. One big advantage is that the training time is much shorter, with DVGO being up to 20 times faster on average than the other two methods, equivalent to about 30 min per sub-module on these datasets using a Tesla V100 GPU.

One way to evaluate the quality of a trained model before extracting the point cloud is to measure the image-based metrics. However, in contrast to the often expected changes in accuracy and completeness across experiments, the image-based metrics do not always show the same results. Even the addition of a new sub-module, which significantly improves the quality of the extracted point clouds, does not necessarily lead to better results in terms novel view synthesis performance. This can be seen in the experiments on the TMB and UseGeo1 datasets in Table 2. Here, PSNR and SSIM sometimes decrease when an additional sub-module is added while accuracy and completeness improve. Only the values for LPIPS change consistently the point cloud metrics across all experiments, suggesting that this metric provides a better approximation.

One aspect that was not investigated in this work, but should have a major impact, is the intrinsic and extrinsic calibration of the camera. Especially for datasets consisting of aerial imagery these are often not given or not accurate. An investigation of the effects of sub-optimal SFM on the reconstruction quality and a further refinement of the camera calibration parameters, for example with PixSFM [43], could be of interest in the future, as the worse performance on the TMB dataset may also be explained by inadequate camera calibration.

5. Conclusions

In summary, we evaluated different NeRF based approaches on challenging aerial datasets and compared them with conventional photogrammetric methods utilizing COLMAP. Although the proposed NeRF methods are typically used for novel view synthesis, a dense point cloud can be extracted by exploiting the internal density representation.

The results of this study demonstrate that NeRF-based approaches are capable of reconstructing large-scale scenes from an aerial perspective in 3D even on challenging datasets. In particular, the partitioning of the scene into several sub-modules has been identified as a crucial strategy. Furthermore, we have demonstrated the effect of changing the density threshold on the accuracy and completeness of the extracted point cloud. A higher threshold value is accompanied by improved accuracy, but also a notable reduction in completeness, particularly in areas lacking texture. On the other hand, using a small additional MLP to estimate visibility as described in [33], leads to an improvement in both metrics.

Nevertheless, the quantitative results presented are lower compared to classical photogrammetric methods, as demonstrated using COLMAP as a reference. This is particularly evident in the lower level of noise and the sharper edges of the reconstructed models. At the same time, the datasets used contain some regions that are difficult for classical photogrammetry, such as water surfaces or vegetation, which could not be reconstructed with COLMAP. Here, the NeRF models provide a visually appealing result, which suggests that a combination of both approaches would compensate for their respective limitations.

Author Contributions: Conceptualization, M.H.; Methodology, M.H. and H.K.; Software, H.K.; Investigation, H.K.; Data curation, H.K.; Writing—original draft, M.H.; Writing—review & editing, B.R. and M.W.; Supervision, B.R. and M.W. All authors have read and agreed to the published version of the manuscript.

Funding: There was no external funding for the conducted work.

Data Availability Statement: Publicly available datasets were used in this study. The used datasets are available at https://usegeo.fbk.eu/ (UseGeo dataset), access on 1 December 2024, and https://ifpwww.ifp.uni-stuttgart.de/benchmark/hessigheim/default.aspx (Hessigheim 3D dataset), access on 1 December 2024, respectively.

Acknowledgments: We would like to thank the IFP of the University of Stuttgart for the access to the images of the Hessigheim 3D dataset. We acknowledge support by the KIT-Publication Fund of the Karlsruhe Institute of Technology.

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

References

1. Furukawa, Y.; Ponce, J. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1362–1376. [CrossRef]
2. Galliani, S.; Lasinger, K.; Schindler, K. Massively Parallel Multiview Stereopsis by Surface Normal Diffusion. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 873–881.
3. Goesele, M.; Snavely, N.; Curless, B.; Hoppe, H.; Seitz, S.M. Multi-View Stereo for Community Photo Collections. In Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8. [CrossRef]
4. Schönberger, J.L.; Zheng, E.; Pollefeys, M.; Frahm, J.M. Pixelwise View Selection for Unstructured Multi-View Stereo. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 501–518.
5. Xu, Q.; Tao, W. Multi-Scale Geometric Consistency Guided Multi-View Stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5483–5492.
6. Xu, Q.; Tao, W. Planar Prior Assisted PatchMatch Multi-View Stereo. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12516–12523.
7. Xu, Q.; Kong, W.; Tao, W.; Pollefeys, M. Multi-Scale Geometric Consistency Guided and Planar Prior Assisted Multi-View Stereo. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4945–4963. [CrossRef] [PubMed]
8. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **2021**, *65*, 99–106. [CrossRef]
9. Barron, J.T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; Srinivasan, P.P. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 5835–5844. [CrossRef]
10. Sun, C.; Sun, M.; Chen, H.T. Direct Voxel Grid Optimization: Super-fast Convergence for Radiance Fields Reconstruction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5449–5459. [CrossRef]
11. Schönberger, J.L.; Frahm, J.M. Structure-from-Motion Revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4104–4113. [CrossRef]
12. Furukawa, Y.; Curless, B.; Seitz, S.M.; Szeliski, R. Towards internet-scale multi-view stereo. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1434–1441.
13. Han, X.; Leung, T.; Jia, Y.; Sukthankar, R.; Berg, A.C. MatchNet: Unifying feature and metric learning for patch-based matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3279–3286.
14. Zbontar, J.; LeCun, Y. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.* **2016**, *17*, 2287–2318.
15. Hartmann, W.; Galliani, S.; Havlena, M.; Van Gool, L.; Schindler, K. Learned multi-patch similarity. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1586–1594.
16. Hirschmueller, H. Stereo processing by semi-global matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 328–341. [CrossRef] [PubMed]
17. Yao, Y.; Luo, Z.; Li, S.; Fang, T.; Quan, L. MVSNet: Depth inference for unstructured multi-view stereo. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 767–783.
18. Huang, P.H.; Matzen, K.; Kopf, J.; Ahuja, N.; Huang, J.B. DeepMVS: Learning multi-view stereopsis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2821–2830.
19. Yao, Y.; Luo, Z.; Li, S.; Shen, T.; Fang, T.; Quan, L. Recurrent MVSNet for High-Resolution Multi-View Stereo Depth Inference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5520–5529.
20. Yang, J.; Mao, W.; Alvarez, J.M.; Liu, M. Cost Volume Pyramid Based Depth Inference for Multi-View Stereo. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 4748–4760. [CrossRef]
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.

22. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision transformers for dense prediction. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 12179–12188.
23. Cao, C.; Ren, X.; Fu, Y. MVSFormer: Multi-View Stereo by Learning Robust Image Features and Temperature-based Depth. *Trans. Mach. Learn. Res.* **2023**.
24. Cao, C.; Ren, X.; Fu, Y. MVSFormer++: Revealing the Devil in Transformer's Details for Multi-View Stereo. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
25. Xie, J.; Girshick, R.; Farhadi, A. Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 842–857.
26. Flynn, J.; Neulander, I.; Philbin, J.; Snavely, N. DeepStereo: Learning to predict new views from the world's imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5515–5524.
27. Wang, Z.; Li, L.; Shen, Z.; Shen, L.; Bo, L. 4K-NeRF: High Fidelity Neural Radiance Fields at Ultra High Resolutions. *arXiv* **2023**, arXiv:2212.04701.
28. Xiangli, Y.; Xu, L.; Pan, X.; Zhao, N.; Rao, A.; Theobalt, C.; Dai, B.; Lin, D. BungeeNeRF: Progressive Neural Radiance Field for Extreme Multi-scale Scene Rendering. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 106–122. [CrossRef]
29. Müller, T.; Evans, A.; Schied, C.; Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.* **2022**, *41*, 1–15. [CrossRef]
30. Chen, A.; Xu, Z.; Geiger, A.; Yu, J.; Su, H. TensoRF: Tensorial Radiance Fields. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 333–350.
31. Martin-Brualla, R.; Radwan, N.; Sajjadi, M.S.M.; Barron, J.T.; Dosovitskiy, A.; Duckworth, D. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7206–7215. [CrossRef]
32. Turki, H.; Ramanan, D.; Satyanarayanan, M. Mega-NERF: Scalable Construction of Large-Scale NeRFs for Virtual Fly-Throughs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12922–12931.
33. Tancik, M.; Casser, V.; Yan, X.; Pradhan, S.; Mildenhall, B.P.; Srinivasan, P.; Barron, J.T.; Kretzschmar, H. Block-NeRF: Scalable Large Scene Neural View Synthesis. In Proceedings of the Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8238–8248. [CrossRef]
34. Jayalakshmi, T.; A., S. Statistical normalization and back propagation for classification. *Int. J. Comput. Theory Eng.* **2011**, *3*, 89–93. [CrossRef]
35. Zhang, K.; Riegler, G.; Snavely, N.; Koltun, V. NeRF++: Analyzing and Improving Neural Radiance Fields. *arXiv* **2020**, arXiv:2010.07492.
36. Seitz, S.M.; Curless, B.; Diebel, J.; Scharstein, D.; Szeliski, R. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; Volume 1, pp. 519–528. [CrossRef]
37. Hermann, M.; Weinmann, M.; Nex, F.; Stathopoulou, E.; Remondino, F.; Jutzi, B.; Ruf, B. Depth estimation and 3D reconstruction from UAV-borne imagery: Evaluation on the UseGeo dataset. *ISPRS Open J. Photogramm. Remote Sens.* **2024**, *13*, 100065. [CrossRef]
38. Wang, Z. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
39. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
40. Nex, F.; Stathopoulou, E.; Remondino, F.; Yang, M.; Madhuanand, L.; Yogender, Y.; Alsadik, B.; Weinmann, M.; Jutzi, B.; Qin, R. UseGeo—A UAV-based multi-sensor dataset for geospatial research. *ISPRS Open J. Photogramm. Remote Sens.* **2024**, *13*, 100070. [CrossRef]
41. Kölle, M.; Laupheimer, D.; Schmohl, S.; Haala, N.; Rottensteiner, F.; Wegner, J.D.; Ledoux, H. The Hessigheim 3D (H3D) benchmark on semantic segmentation of high-resolution 3D point clouds and textured meshes from UAV LiDAR and Multi-View-Stereo. *ISPRS Open J. Photogramm. Remote Sens.* **2021**, *1*, 11. [CrossRef]
42. Knapitsch, A.; Park, J.; Zhou, Q.Y.; Koltun, V. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graph.* **2017**, *36*, 1–13. [CrossRef]
43. Lindenberger, P.; Sarlin, P.E.; Larsson, V.; Pollefeys, M. Pixel-perfect structure-from-motion with featuremetric refinement. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 5987–5997.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.