remote sensing

MDPI

Article

Automatic and Semantically-Aware 3D UAV Flight Planning for Image-Based 3D Reconstruction

Tobias Koch [1,*], Marco Körner [1] and Friedrich Fraundorfer [2,3]

[1] Chair of Remote Sensing Technology, Technical University of Munich, 80333 Munich, Germany
[2] Institute for Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria
[3] Remote Sensing Technology Institute, German Aerospace Center, 82234 Wessling, Germany
* Correspondence: tobias.koch@tum.de; Tel.: +49-89-289-22679

Received: 29 May 2019; Accepted: 24 June 2019; Published: 29 June 2019

Abstract: Small-scaled unmanned aerial vehicles (UAVs) emerge as ideal image acquisition platforms due to their high maneuverability even in complex and tightly built environments. The acquired images can be utilized to generate high-quality 3D models using current multi-view stereo approaches. However, the quality of the resulting 3D model highly depends on the preceding flight plan which still requires human expert knowledge, especially in complex urban and hazardous environments. In terms of safe flight plans, practical considerations often define prohibited and restricted airspaces to be accessed with the vehicle. We propose a 3D UAV path planning framework designed for detailed and complete small-scaled 3D reconstructions considering the semantic properties of the environment allowing for user-specified restrictions on the airspace. The generated trajectories account for the desired model resolution and the demands on a successful photogrammetric reconstruction. We exploit semantics from an initial flight to extract the target object and to define restricted and prohibited airspaces which have to be avoided during the path planning process to ensure a safe and short UAV path, while still aiming to maximize the object reconstruction quality. The path planning problem is formulated as an orienteering problem and solved via discrete optimization exploiting submodularity and photogrammetrical relevant heuristics. An evaluation of our method on a customized synthetic scene and on outdoor experiments suggests the real-world capability of our methodology by providing feasible, short and safe flight plans for the generation of detailed 3D reconstruction models.

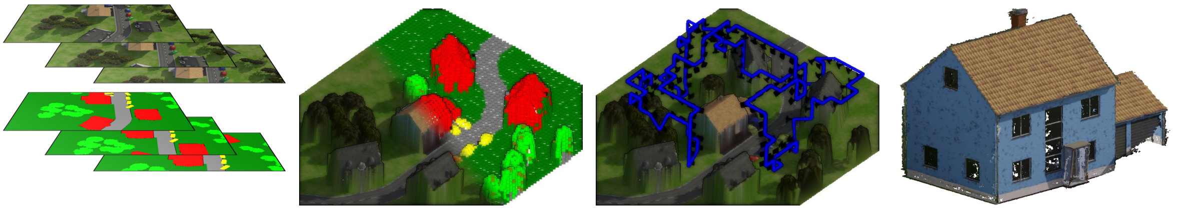
Keywords: UAV; trajectory optimization; path planning; discrete optimization; 3D reconstruction; semantics; urban mapping

1. Introduction

Unmanned aerial vehicles (UAVs) have attracted significant attention in the field of 3D modeling, as they are capable of carrying high-resolution cameras, combining advantages of both conventional airborne and terrestrial photogrammetry. The mobility and maneuverability of UAVs to freely move in three dimensions and simultaneously capture close-up images of an object with arbitrary viewing angles allow to generate high-resolution and photo-realistic 3D models with high accuracy by processing a series of overlapping images with current state-of-the-art structure from motion (SfM) and multi-view stereo (MVS) pipelines, such as Pix4D [1], Bundler [2], or Colmap [3]. These models are of high interest in various fields, such as the use of digitized building models for 3D city modeling [4], object inspection [5], or cultural heritage documentation [6]. However, the quality of resulting 3D models strongly relies on flight plans that satisfy the requirements of an image-based 3D modeling process which include the acquisition of multiple overlapping images, sufficient baselines between the camera viewpoints and the prevention of optical occlusions from surrounding obstacles. In terms of mapping mostly flat and spacious scenes, such as landscapes, flight planning can be easily executed in form of simple grid-like patterns or circular flights from the same altitude but can become exceedingly complex for densely built urban areas consisting of different kinds of human-made objects and vegetation. Planning a UAV trajectory in such areas involves considering the surrounding environment and keeping a safety distance toward any obstacle while ensuring that the entire object of interest is captured from close ranges and different perspectives.

The most common method to obtain aerial imagery in an automated fashion is to use an off-the-shelf flight planner, such as commercial flight planning software Pix4D [1], PrecisionHawk [7], DJI Flight Planner [8], or open-source based PixHawk Ardu Planner [9]. These easy-to-use planners can generate simple polygons, regular grids, or circular trajectories, however, some prior knowledge of the scene height must be known in advance for designing a collision-free flight plan. For more complex scenes, such as urban areas, standard path planning methods are usually insufficient to generate high-quality 3D models, as we will show later. Therefore, UAV flights in such complex scenarios still require manual operation by experienced pilots in case standard flight planners are not feasible or do not guarantee a sufficient reconstruction quality. From a practical or even legal point of view, it may be even necessary to adapt the flight plan with respect to the semantics of the environment, especially in densely built areas. Restricted airspaces may be defined in regions that are prohibited to be accessed by the UAV or which should be avoided in case of an unexpected malfunction of the vehicle. These restricted areas could include other buildings, train rails, water bodies, parked cars, highways or other heavily frequented roads. Flying UAVs in such environments is already challenging. If the resulting 3D model additionally demands certain photogrammetric properties, such as the desired ground sampling distance (GSD), the acquisition of highly overlapping close-up images covering the entire object could become infeasible in the presence of restricted or prohibited airspaces.

General research on path planning for UAV mapping has already been initiated in recent years focusing on automation of the generation of optimal flight plans. Automated flight planning methods can be classified either as model-free and model-based methods. The former performs an exploration task in unknown environments by iteratively updating the model with new measurements via selecting the next best view from a current view. These models do not require prior knowledge of the scene but usually, they do not guarantee full coverage of the object. Methods of the latter class, on the other hand, rely on a coarse proxy model of the scene and refine the model by an optimal subsequent flight which is globally optimized. The targets of these explore-and-exploit approaches are manyfold, such as maximizing the coverage of a target object [10,11] or minimizing the acquisition time [12] or energy consumption [13,14]. However, to the best of our knowledge, none of these works take into account the surrounding environment for generating safe UAV paths that additionally avoid or even restrict certain airspaces in the scene. With the tremendous advances in semantic image segmentation for aerial imagery by recent deep learning-based approaches [15], accurate and consistent dense semantic maps can be generated, which extend the purely geometric 3D scene representation, helping to generate safe UAV flights under the consideration of the real environment. Since we want to adapt the flight path to the semantic properties of the scenery, an initial semantically-enriched proxy model of the entire scene is required, which leads us to employ a model-based approach. An inspiration of our path planning method was given by the works of Roberts et al. [10] and Hepp et al. [11], formulating the path planning problem as a graph-based optimization for maximizing the information gain obtained from a UAV trajectory with a set of heuristics representing 3D modeling image acquisition practices. With this paper, we build on these works by introducing more interpretable heuristics which directly influence user-specified requirements of the reconstruction quality, as well as optimizing for a minimum path length. In addition, we show how to incorporate semantic information to safe path planning. Figure 1 illustrates the general idea of our path planning approach, consisting of a two-staged planning procedure, wherein a first nadir flight is used to generate a semantically-enriched proxy model of the entire environment which is further used to generate a set of viewpoint hypotheses in the free and accessible airspace. A discrete optimization among this camera graph is conducted to find a short and matchable path along the graph, which maximizes the reconstruction quality of the target object while considering restrictions on the airspace defined by the semantical cues.



(**a**) Unmanned aerial vehicle (UAV) images and semantic maps.

(**b**) Semantically-enriched initial model.

(**c**) Semantic-aware 3D UAV path.

(**d**) Final reconstructed 3D model.

Figure 1. Proposal of our UAV path planning methodology for generating 3D reconstructions of individual objects considering path restrictions based on semantics. A 3D proxy model is generated using a set of geo-referenced UAV images from a simple flight above the environment which is further enriched by transferring 2D segmentation labels into 3D space, defining free, conditionally accessible and prohibited airspaces (**b**). A graph-based optimization estimates a collision-free trajectory for image acquisition viewpoints considering restricted airspaces while minimizing the respective path length (**c**). The acquired images of the traversed trajectory are suitable to generate high-resolution 3D reconstruction models (**d**).

Particularly, our contributions are as follows:

(1) We propose a set of heuristics based on photogrammetric reconstruction parameters, leading to individual flight paths for arbitrary camera intrinsics that ensure the generation of 3D models in a user-specified resolution.
(2) We show how to exploit semantic segmentation of UAV imagery for extracting the target object and for generating a semantically-enriched initial 3D proxy model, which defines restricted and prohibited airspaces.
(3) We propose a model-based optimization scheme with respect to a semantic model that maximizes the object coverage while minimizing the corresponding path length and avoiding restricted airspaces.
(4) We propose a realistic synthetic 3D model suitable for a comprehensive evaluation of urban flight planning, including a highly detailed building model embedded in a realistic and interchangeable scenery.

2. Related Work

The rapid development of UAVs and sensors has contributed significantly to their popularity in many industries nowadays, such as urban mapping, object inspection, precision agriculture, and surveying tasks. Equipped with high-resolution cameras and the utilization of most recent SfM and MVS methods on image sequences, 3D models of the environment can be generated in a much greater level of detail compared to conventional manned aircraft. However, the quality of such reconstructions highly depends on the camera network configuration during the acquisition process. An exhaustive amount of work addressed the problem of selecting the best views from a large amount of different views hypotheses [2,16–19]. These works point out the crucial parameters which affect the reconstruction quality, such as parallax angles and baselines between views, as well as their observation angles and distances toward the object's surface and propose meaningful heuristics to model the reconstruction quality from different camera constellations.

An integration of these parameters is already used for automating the image acquisition process for large-scale areas [1,7–9], allowing the planning UAV flights as simple geometric patterns, such as regular grids or circular flights with respect to the desired GSD. These off-the-shelf planners are sufficient in case of spacious and flat terrains without obstacles [20], but are not suitable for use in uneven, densely built or heavily vegetated environments. Since no 3D model of the environment is taken into consideration, these trajectories either do not cover every part of the object of interest due to occlusions by obstacles or may even cause an accident with an adjacent obstacle in the environment.

More advanced path planning approaches aim to automatically map objects in either completely unknown environments or based on a very coarse prior model of the environment. Methods of the first group solve an exploration task by iteratively selecting the most promising view to refine the explored model based on a current view with new measurements. This incremental scene modeling and viewpoint planning is commonly known as next best view (NBV) planning, which is already a long-standing part of research in the field of Robotics. The methods alternately fuse incoming measurements from a new viewpoint into the reconstruction of the scene and estimate novel viewpoints in order to increase the information about the object. Classical sensors for these measurements include laser scanners [21], RGB-D sensors [22–28] and cameras [29–34]. Such methods are usually hard to implement utilizing cameras as selected sensors, as the generation of depth maps, which is necessary to derive new 3D information, requires significant onboard processing power or at least a wireless connection to the ground-station for data transmission, in order to merge incoming measurements with the current model. Additionally, selecting next best views in accordance to MVS requirements—in particular, maintaining sufficient baselines and parallax angles of adjacent views—on the fly is a challenging task since the actual mapped free airspace might be very limited.

In contrast to model-free exploration methods that focus on autonomy and real-time capability in unknown environments, model-based path planning algorithms rely on an available proxy model of the environment and focus on estimating a subsequent optimal path to maximize the coverage and accuracy of the object globally [10,11,35–38]. In contrary to active modeling, these explore-and-exploit methods do not receive any feedback from the acquired images during the exploitation flight, which demands high attention to the applied heuristics being used for generating the refinement path. The global optimization of coverage and accuracy, on the other hand, usually leads to larger completeness and smoother trajectories compared to model-free methods. Recent work has proposed to extend this procedure by iteratively refining the model from several subsequent flights, taking into account the remaining model uncertainty between each flight [38,39]. Furthermore, the execution of the optimized path is easy and fast for any kind of UAV by simply navigating alongside the optimized waypoints. The prior model can either be based on an existing map with height information [36] or is generated by photogrammetric reconstructions from a preceding manual flight at a safe altitude or via standard flight planning methods (e.g., regular grids or circular trajectories) [10,11] and is usually expressed by a set of discrete 3D points in a voxel space [10,11,37,40] or by volumetric surfaces, such as triangulated meshes [35,36,38,41]. In order to define appropriate views for the optimized trajectory, camera viewpoint hypotheses are either regularly sampled in the free 3D airspace [10,37] resulting in 3D camera graphs, or are sparsely sampled in a 2D view manifold [38] or in skeleton sets [42] around the object. Subsequently, an optimization is defined in order to find a connected subset of these viewpoint hypotheses to define a suitable path through the camera graph. Alternatively, the locations of the of regularly sampled viewpoint candidates can be continuously refined during the optimization [11]. As a means of assessing the suitability of camera viewpoints for the reconstruction, hand-crafted heuristics are usually defined considering the necessities for a successful SfM and MVS workflow. These include multi-view requirements [35,37,40], ground resolution [35,41], 3D uncertainty [43] and the coverage of the object [10,11,37]. Instead of using hand-crafted heuristics, several works used machine learning methods to learn heuristics that allow predicting the confidence in the output of a MVS without executing it [27,43,44].

Recently, efficient methodologies formulate the view planning problem as a discrete optimization task and exploit submodularity in the optimization process, standing for fast and reliable convergence, even for a large number of viewpoint hypotheses [10,11]. The main advantage of this idea is to jointly assess additional information gain of individual viewpoints for arbitrary viewpoint constellations in a global manner. This allows formulating the path planning task as an orienteering problem, which can be solved with simple greedy algorithms, by optimizing a path which collects as many information gains as possible for a specific path length. The results presented in previous work reveal notable trajectories for generating high-fidelity image-based 3D reconstructions. However, setting a suitable path length in the optimization may require expert knowledge and highly affects the trajectory estimation, since, due to the purely additive nature of orienteering problem, adding additional views will never decrease the objective function. This might lead to abundant redundant views for overestimated path lengths and incomplete reconstructions for underestimated path lengths. Although the presented heuristics follow best practices for MVS requirements, they do not respect user-specific demands on the resulting 3D model, such as the number of views and observations angles of the object surface or a required model resolution using arbitrary cameras. Additionally, prior work so far solely considers purely geometric cues for flight planning of both small-scale and large-scale areas. With the vast progress in semantic segmentation using deep learning-based approaches, the applicability of neural networks for semantic segmentation of aerial and UAV imagery was demonstrated in several works [45–47].

In this paper, we show how to incorporate semantic cues into UAV flight planning for generating safe trajectories for real-world 3D mapping applications, which allow to define inadmissible airspaces above user-defined object types. Additionally, we propose a set of heuristics for SfM and MVS image acquisition used in the optimization allowing for the maintenance of a pre-defined model resolution for the entire targeted object. Although the preferred task of photogrammetric 3D modeling is to maximize the reconstruction quality rather than minimizing the path length, we integrate a penalization for lengthy paths without a significant drop in the reconstruction quality.

3. Proposed Flight Planning Pipeline

Our flight planning methodology follows a two-staged explore-and-exploit approach, consisting of two subsequent flights, where a first safe exploration flight is used to generate an initial proxy model of the environment which is further refined by an optimized exploitation path in terms of full coverage, high-resolution and accuracy of the object to be reconstructed. Latter additionally respects restrictions of the airspace derived from semantic cues to avoid hazardousness and prohibited areas and to elude collisions with the surrounding environment. Our work is inspired by the works of Roberts et al. [10] and Hepp et al. [11] in the matter of estimating a closed trajectory from numerous viewpoint hypotheses by exploiting submodularity in the optimization procedure. An overview of our complete workflow is depicted in Figure 2. First, the acquired images of the exploration flight are processed to generate a semantically-enriched coarse proxy model which defines free and occupied airspace. The semantic cues help to extract the object of interest and, based on the proxy model, a set of viewpoint hypotheses is generated and evaluated according to their eligibility for reconstructing the target object with respect to our heuristics used for MVS image acquisition. Adjacent viewpoints are evaluated according to their matchability and connected to a camera graph. Finally, an exploitation flight is optimized by finding a closed and short path among the camera graph which maximizes the reconstruction quality and avoids prohibited and minimizes hazardousness airspaces defined by the semantics of the proxy model. Summarizing the objectives of the path planning problem, the following requirements need to be fulfilled by our methodology:

1. Coverage: every point on the object surface has to be visible in at least two images to be able to triangulate its position in 3D space from the images.
2. Safety: the estimated trajectory has to avoid collisions with obstacles and has to be aware of the semantics of the surrounding environment in terms of restricted and prohibited airspaces.
3. Path length: the estimated trajectory should be as short as possible and avoid redundant views, as several images taken from similar camera poses introduce local uncertainties in depth estimation by glancing intersections.

4. Heuristics: The estimated trajectory should facilitate complete reconstruction of the target object considering photogrammetric reconstruction criteria, such as GSD, observation angles, number of views, and sufficient overlap between adjacent views.
5. Quality assessment: the path planning method should return an approximation of the expected reconstruction quality before the execution of the flight, in order to adjust the path or plan another subsequent path.

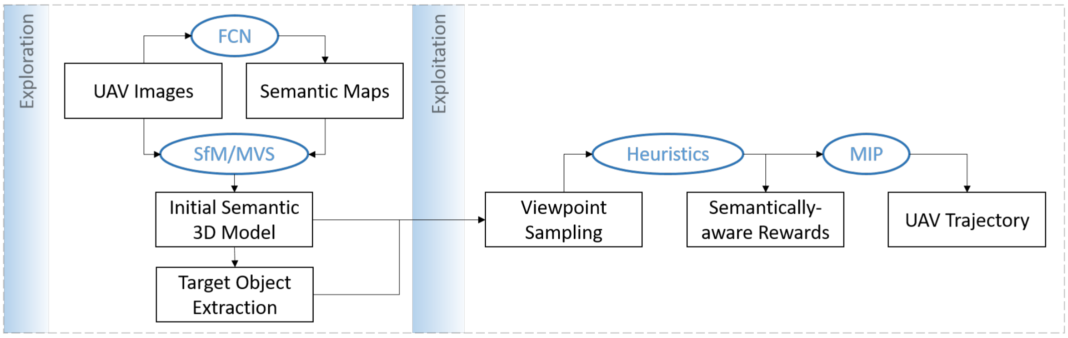


Figure 2. Overview of the proposed workflow of our UAV path planning approach. Based on a exploration flight at a safe altitude, the captured images are segmented and fused to an initial semantic 3D model. After selection of the target object, numerous camera viewpoints are sampled and assessed according to their eligibility for the reconstruction process, while the semantic information of the environment assigns restricted or prohibited airspaces for the UAV. Finally, a discrete graph-based optimization estimates the optimal semantically-aware trajectory which ensures a high-quality 3D model of the target object.

The following sections provide a detailed description of the proposed methodology, starting with the outline of the path planning problem and the definition of the optimization objective in Section 3.1. Details on the generation of the semantically-enriched proxy model from a set of nadir images and the extraction of the target object from the proxy model are provided in Section 3.2. Section 3.3 describes the generation of numerous viewpoint hypotheses, which are assessed with respect to our proposed heuristics explained in Section 3.4. Finally, Section 3.5 presents the semantic-aware optimization.

3.1. Notation and Definition of the Path Planning Problem

The objective of our path planning problem is to find a feasible UAV trajectory to acquire images of a target object such that the final 3D reconstruction model is of high quality. The object of interest is expressed as a sparse set of discrete surface points $\boldsymbol{s}_{j=1...J} = (\boldsymbol{x}_j, \boldsymbol{\eta}_j) \in \mathcal{S}$, comprised of 3D locations $\boldsymbol{x}_j \in \mathbb{R}^3$ and normal vectors $\boldsymbol{\eta}_j \in \mathbb{R}^3$ on the tangent plane of the object. We consider a discrete optimization scheme and represent our camera viewpoint hypotheses as an undirected weighted graph $G = (\mathcal{P}, \mathcal{E})$, composed of a set $\mathcal{P}$ of nodes as camera poses $\boldsymbol{p}_{i...I} = (\boldsymbol{c}_i, \boldsymbol{r}_i) \in \mathcal{P}$ consisting of 3D locations $\boldsymbol{c}_i \in \mathbb{R}^3$ and camera orientations $\boldsymbol{r}_i \in \mathbb{R}^3$ defined as roll, pitch and yaw angles. Adjacent and matchable viewpoints in the graph are connected through a set of edges $\mathcal{E} = \{\boldsymbol{e}_k = (\boldsymbol{p}_i, \boldsymbol{p}_j)\}$ with associated weights $\mathcal{W} = \{\boldsymbol{w}_k = (w_k^{\text{eucl}}, w_k^{\text{sem}})\}$, representing a Euclidean distance $w_k^{\text{eucl}} \in \mathbb{R}$ and a semantic label cost $w_k^{\text{sem}} \in \mathbb{R}$. We define a feasible trajectory $\mathcal{T} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_n\} \subset \mathcal{P}$ as a subset of connected camera poses in the camera graph G. The goal of the path planning problem is to find an optimal trajectory

$$\begin{aligned} \mathcal{T}^* &= \arg\max_{\mathcal{T}} R(\mathcal{T}) \\ \text{subject to } &\sum_{e \in \mathcal{E}} w^{\text{eucl}} \rightarrow \min, \\ &\sum_{e \in \mathcal{E}} w^{\text{sem}} < L^{\text{sem}} \end{aligned} \tag{1}$$

that maximizes the reconstructability $R : \mathcal{P} \to \mathbb{R}$ of the target object $\mathcal{S}$, while minimizing the corresponding path length and restricting the path not to exceed an accumulated label cost limit L^{sem}. The reconstructability $R(\mathcal{T}) = \sum_{\mathcal{T}} I(\boldsymbol{p}(\mathcal{T}), \mathcal{S})$ obtained from a trajectory $\mathcal{T}$ is defined as the accumulated information reward $I(\boldsymbol{p}(\mathcal{T}), \mathcal{S})$ of all camera poses $\boldsymbol{p}(\mathcal{T})$ of that trajectory. The computation of rewards I requires a set of heuristics, approximating the impact of an arbitrary camera pose $\boldsymbol{p}$ for the reconstruction quality of the object surface $\mathcal{S}$. Besides rating of the camera poses regarding the distance toward the object surface and the incidence angles of camera rays, the proposed heuristics also address the assessment of camera configurations of adjacent camera poses with respect to a successful multi-view stereo matching.

3.2. Semantically-Enriched Initial 3D Model

Given a series of nadir or oblique images encompassing the object of interest and its surrounding environment, a coarse proxy model of the entire scene is generated by processing the initial images with current state-of-the-art SfM and MVS pipelines, such as Pix4D [1], Colmap [3], or Bundler [2]. The initial flight can be realized either by a manual flight at a safe altitude or via commonly used predefined flight planning systems resulting in grid-like or circular patterns, which is feasible in most sceneries. In order to compute the subsequent trajectory in the same reference frame as the initial flight, we incorporate GNSS coordinates of the UAV or utilize ground control points (GCP) to the bundle adjustment. The model only requires a low resolution and can exhibit gaps in the reconstruction, such as missing façades, but should cover a large amount of the surrounding environment, which determines accessible and occupied air space for the viewpoint planning. The model itself can be either expressed as a dense point cloud with low point sampling density or by regularly sampled 3D points from the faces of a triangulated mesh. The initial proxy model generation can be computed fast even with off-the-shelf mobile computers. Alternatively, a coarse 3D model can be already generated on-board the UAV during the exploration flight [48].

At the same time, a pixel-wise dense semantic segmentation of the images is conducted using a fully convolutional network (FCN) [49]. An adjustment of the number of classes required for our task (building, lawn, tree, street, car, others) and the utilization of a diverse set of available and manually annotated UAV and aerial nadir images from different altitudes and various scenes was carried out for training the network. Available training data from [50,51] was extended with manually annotated UAV images from different scenes to achieve a total amount of 3069 images split into 60% training and 40% validation images. The quantitative evaluation after refining the pre-trained model for 50 epochs yield a global accuracy of 0.81 and a mean intersection over union (IoU) score of 0.52, indicating a reasonable segmentation performance for our task. We infer every single UAV image used for the initial 3D reconstruction to the segmentation network, in order to facilitate the redundancy of overlapping areas for reducing labeling uncertainty. To propagate the 2D semantic labels, we make use of the visibility information obtained from the 3D modeling process and back-project every single 3D point into every image in which it is visible and compute the point label by majority voting. Despite the rather small receptive field of the FCN-8s providing merely coarse segmentation boundaries as shown in Figure 3, an adequate semantic enrichment of the 3D model can be achieved for a relatively large grid spacing of adjacent viewpoints (3–4 m in our experiments) by exploiting the redundancy of overlapping images.

Since the initial 3D model is coarsely geo-referenced, it is possible to refine the segmentation results of the 3D scene for hardly distinguishable objects of the same semantic class with the use of open street map (OSM) information. For instance, the distinction of various types of roads, which can hardly be determined by 2D semantic segmentation methodologies, could be a crucial requirement for generating safe UAV paths. Heavily frequented road sections (e.g., parking lots, highways and trunk roads) should be highly avoided, while restrictions on side roads and driveways could be less strict. The already segmented road sections of the initial 3D model can therefore be extended with subtypes by automatically inferring the classes from OSM to the labeled 3D points. Since OSM provides numerous and detailed map features, this procedure can be extended for various land cover classes and facilitates user-defined restrictions, such as the differentiation of residential and industrial buildings.

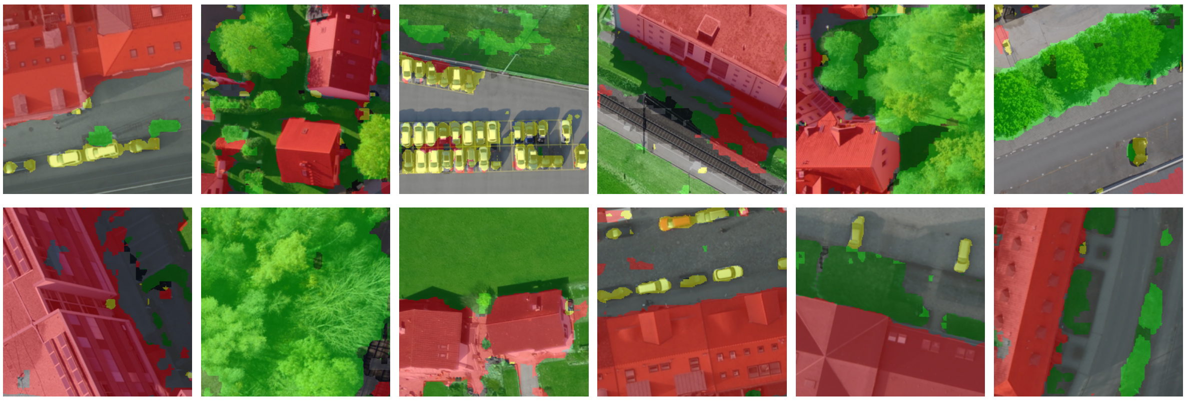


Figure 3. Example of semantic segmentation results on our validation images with a fine-tuned fully convolutional network (FCN) model [49] trained on UAV and aerial images. Visualization is color coded for buildings (🟥), streets (▨), low vegetation (🟩), high vegetation (🟩), cars (🟨), and others (⬛).

Given an approximate semantically-enriched 3D model of the environment, the target object to be finally reconstructed needs to be identified, extracted and completed in a semi-automatic manner. As the initial model could be incomplete during the reconstruction process and the usage of nadir-views results in gaps in the model, such as missing façades and other unseen object details, the target model needs to be completed to ensure camera poses pointing toward these missing details. With the assumption of simplified building models, we identify and extract the target object by a simple 3D region growing approach exploiting the semantic labels of the 3D points. A user input of one corresponding 3D point belonging to the object to be reconstructed serves as the seed for the region growing process. After isolation of the target object, we equally sample surface points $\boldsymbol{s} = (\boldsymbol{x}, \boldsymbol{\eta})$ of the object outline to the ground level and compute 3D point normals required for the proposed heuristics.

3.3. Camera Viewpoint Hypotheses Generation

The goal of the trajectory planning is to define a set of viewpoints allowing the triangulation of as many 3D points of the target object surface as possible according to photogrammetric necessities for a successful reconstruction. A large amount of evenly distributed viewpoint candidates $\boldsymbol{c}$ is sampled in the free airspace inside a bounding box around the extracted object, excluding camera viewpoints which are closer to any surrounding obstacle than a predefined safety buffer. This safety buffer can be adapted according to the corresponding semantic labels of the environment in order to increase the distance toward hazardous objects, such as trees, which often lack in completeness for photogrammetric reconstructions. For each viewpoint candidate, we also store a vector containing the semantic labels of all proxy 3D points located below the camera viewpoints.

Besides the location of camera viewpoints, orientations $\boldsymbol{r}$ need to be assigned pointing toward the target object while avoiding occlusions with obstacles. Although the subsequent reconstruction process favors fronto-parallel views toward the target surface to ensure a high-quality reconstruction, adjacent viewpoints also require smooth transitions with high overlap. Since viewpoint orientations pointing toward the closest surface point results in fronto-parallel views, the matchability at edges of the object might be insufficient due to large orientation changes. On the other hand, viewpoint orientations which always point toward the center of the object result in large overlap but slanted views toward the object surface in case of elongated or other complex object structures. In comparison to other approaches, which either assign orientations pointing toward the center of the object [12] or include the orientation estimation in the optimization [10,11], we perform a visibility assessment of each viewpoint to identify 3D surface points which are visible from each specific viewpoint location considering the surrounding environment. A fast visibility computation approach [52] is utilized and visibilities for all viewpoints and surface points are stored in an indicator matrix $\boldsymbol{U} \in \mathbb{R}^{I \times J}$. In particular, a look-at-vector $\boldsymbol{n}_i \in \mathbb{R}^3$ for each viewpoint $\boldsymbol{c}_i$ is computed and directed toward the weighted mean of all visible 3D points $\mathcal{S}_{c_i} \subset \mathcal{S}$ from the corresponding 3D location of the viewpoint. In order to prioritize object points that are closer to the camera, $\boldsymbol{n}_i$ is further weighted by the distance toward all visible 3D points. We begin with computing weighting coefficients τ_j for each visible surface point $\boldsymbol{x}_j \in \mathcal{S}_{c_i}$ from a camera view $\boldsymbol{c}_i$ utilizing the normalized distances from all visible surface points toward the camera location by

$$\tau_j = 1 - \sqrt[k]{\frac{\|\boldsymbol{x}_j - \boldsymbol{c}_i\| - \min\left(\{\|\boldsymbol{x} - \boldsymbol{c}_i\|\}\right)}{\max\left(\{\|\boldsymbol{x} - \boldsymbol{c}_i\|\}\right) - \min\left(\{\|\boldsymbol{x} - \boldsymbol{c}_i\|\}\right)}}, \tag{2}$$

where k controls the strength of favoring closer surface points toward the camera viewpoints. The weighting coefficients $\left\{\tau_j \in \mathbb{R} : 0 \leq \tau_j \leq 1\right\}$ reflect the influence of each surface point based on its distance to the camera viewpoint, resulting in large values for closer surface points and decreasing values for farther points. In combination with the normalized direction vectors toward each visible surface point, the weighted look-at vector $\boldsymbol{n}_i$ is computed by

$$\boldsymbol{n}_i = \frac{1}{\sum_j \tau_j} \sum_{\boldsymbol{x}_j \in \mathcal{S}_{c_i}} \tau_j \cdot \frac{\boldsymbol{x}_j - \boldsymbol{c}_i}{\|\boldsymbol{x}_j - \boldsymbol{c}_i\|}. \tag{3}$$

We found that this simple procedure results in suitable viewpoint orientations applicable for different object outlines, avoids occluded views and results in almost fronto-parallel views with smooth transitions at object boundaries allowing a large image overlap needed for a successful image registration. Additionally, the orientation estimation does not have to be included in the optimization, which would increase the complexity of the optimization. Finally, the look-at vectors $\boldsymbol{n}_i$ are converted into pose orientations $\boldsymbol{r}_i$, composed of three Euler angles $\varphi_i = 0$, $\theta_i = \sin^{-1}(-n_{i,y})$ and $\psi_i = \tan^{-1}(\frac{n_{i,x}}{n_{i,z}})$ representing roll, pitch and yaw angles, whereas roll angles are fixed to zero, as we assume axis aligned camera views. After updating the visibility matrix $\boldsymbol{U}$ with respect to the camera intrinsics and assigned orientations, theoretical overlaps between views are computed. Nodes of adjacent camera viewpoints which satisfy a specific overlap constraint (e.g., 75%), are connected via edges $\boldsymbol{e}$ in the graph G, comprised of the Euclidean distance w^{eucl} between the corresponding nodes and the semantic label costs w^{sem}, defined as the mean distribution of assigned labels of ground points between both nodes.

3.4. Path Planning Heuristics

In terms of the optimization defined in Equation (1), the abundant viewpoint hypotheses $\boldsymbol{c}_i$ have to be assessed with respect to their eligibility for reconstructing the object. Following best practices on image acquisition for photogrammetric 3D reconstruction, a set of heuristics is defined which reflect the requirements of the subsequent steps of image registration and dense matching. There is an extensive amount of relevant literature on the principles of photogrammetric 3D modeling [53–55] pointing out decisive aspects for achieving high-quality reconstructions from a set of images:

(1) Distance: the distance between camera viewpoints and object surface defines the resulting model resolution and depends on the desired point density and the camera intrinsics.
(2) Observation angle: shallow observation angles between the camera views and surface normals are favored in MVS approaches.
(3) Multiple views: every part of the scene has to be observed from at least two views from different perspectives with sufficient overlap between the views. The identification of corresponding points in overlapping images is the requirement for robustly estimating camera poses and for triangulating 3D object points.
(4) Parallax angle: shallow parallax angles increase the triangulation error and therefore affect the model quality, while too large angles decrease the matchability between the views due to a lack of image similarity between the views which could result in a failure of the image registration step or in gaps in the reconstructed 3D model.

The heuristics are used to predict the eligibility of the viewpoints for the reconstruction and ensures that the target object can be sufficiently reconstructed using the estimated viewpoints from the trajectory. Requirements (1) and (2) can be formulated independently for all viewpoints, while (3) and (4) depend on a pairwise or even multi-view assessment. We define information rewards

$$I(\boldsymbol{p}_i, \mathcal{S}_{\boldsymbol{p}_i}) = \sum_{\boldsymbol{x}_j \in \mathcal{S}_{\boldsymbol{p}_i}} I_{\mathrm{d}}(\boldsymbol{c}_i, \boldsymbol{x}_j) I_{\mathrm{a}}(\boldsymbol{n}_i, \boldsymbol{\eta}_j), \tag{4}$$

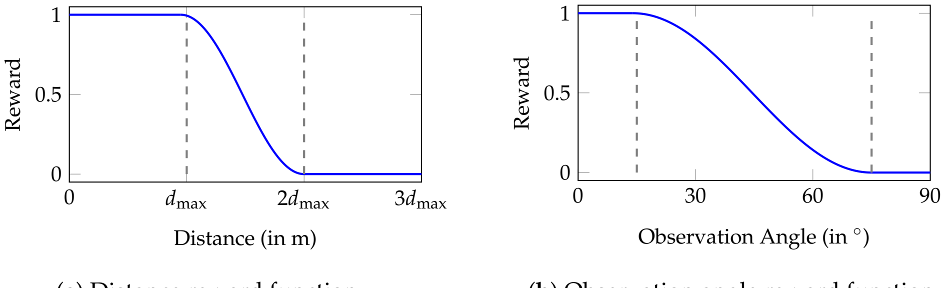
for all viewpoints combining requirements 1) and 2) as a distance-based and observation angle-based reward $I_{\mathrm{d}}(\boldsymbol{c}_i, \boldsymbol{x}_j)$ and $I_{\mathrm{a}}(\boldsymbol{n}_i, \boldsymbol{\eta}_j)$.

3.4.1. Distance

The resolution of the reconstruction depends on the camera intrinsics and the acquisition distances toward the object surface and is usually defined as the GSD or point density after the dense matching reconstruction step. High-resolution models are of high interest for modeling, monitoring and inspecting objects and can be realized by the use of high-resolution cameras or capturing close-up views of the object. Since the goal of the path planning is to provide an equal point density for every part of the object, regardless of its shape and height, we define a maximum distance threshold $d_{\max}$ between a viewpoint and the observed surface points in order to achieve a user-specified model resolution. The maximum tolerable distance to obtain the required GSD also depends on the camera intrinsics and is given by $d_{\max} = \frac{\mathrm{GSD} \cdot f}{\text{pixel size}}$ with a focal length f. We define a smooth symmetrical function $I_{\mathrm{d}}(d)$ for the distance $d = \|\boldsymbol{c}_i - \boldsymbol{x}_j\|$ between a camera viewpoint $\boldsymbol{p}_i$ and an object surface point $\boldsymbol{x}_j$, that assigns maximum reward for a distance less than $d_{\max}$ and decreasing returns up to a multitude of $d_{\max}$. The distance-based reward function $I_{\mathrm{d}}(d)$ is defined as

$$I_{\mathrm{d}}(d) = \begin{cases} 1, & \text{if } d < d_{\max}, \\ 0, & \text{if } d > 2d_{\max}, \\ \frac{1}{2}\left(1 - \cos\left(\frac{d\pi}{d_{\max}}\right)\right), & \text{otherwise,} \end{cases} \tag{5}$$

where no rewards are returned for distances larger than twice of $d_{\max}$. A visualization of $I_{\mathrm{d}}(d)$ is shown in Figure 4a.



(**a**) Distance reward function (**b**) Observation angle reward function

Figure 4. Heuristics for individual viewpoint candidates considering distances and observation angles toward an object point. (**a**) Rewards considering the distance between viewpoint and object points regarding the maximum distance d required to achieve a user-specified ground sampling distance (GSD). (**b**) Rewards based on observation angles defining maximum rewards for fronto-parallel views (here: up to 15°) and zero rewards for more than 75°.

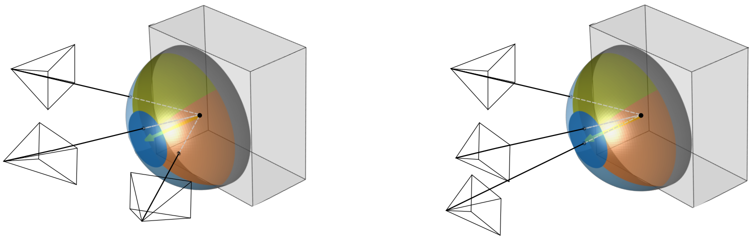
3.4.2. Observation Angle

Besides the importance of distances between camera viewpoints and surface points, the observation angles of the camera rays toward the surface normals are also of particular relevance for the quality of the reconstruction. It is commonly known that fronto-parallel views toward a planar surface result in a higher reconstruction quality, due to minor distortions of the objects appearance in the image, which leads to a more robust and reliable matching result [19]. Although viewpoint orientations are already computed and favoring fronto-parallel views, the abundance of viewpoint hypotheses still have to be evaluated according to their observation angles. Hence, we adapt our reward function for observation angles $\alpha = \cos^{-1}\left(\boldsymbol{n}_i^\top \cdot \boldsymbol{\eta}_j\right)$ and define two thresholds α_{min} and α_{max}, where the first is used for maximum rewards for low observation angles and second represents the maximum tolerable observation angle for returning rewards. We define the observation angle-based reward $I_{\text{a}}(\alpha)$ as

$$I_{\text{a}}(\alpha) = \begin{cases} 1, & \text{if } \alpha < \alpha_{\text{min}}, \\ 0, & \text{if } \alpha > \alpha_{\text{max}}, \\ \frac{1}{2}\left(1 + \cos\left(\frac{\pi(\alpha - \alpha_{\text{min}})}{\alpha_{\text{max}} - \alpha_{\text{min}}}\right)\right), & \text{otherwise.} \end{cases} \tag{6}$$

Since our experiments focus on the reconstruction of buildings, we follow the proposal of Furukawa and Hernández [19] pointing out that observation angles up to 15° yield best reconstruction results for planar surfaces, such as building façades. This suggestion is in accordance with the extensive study about the impacts of the acquisition geometry for dense matching algorithms by Wenzel et al. [56]. Therefore we set $\alpha_{\text{min}} = 15°$, while the upper threshold—indicating a failure of MVS algorithms due to large object distortions in the image— was empirically determined to $\alpha_{\text{max}} = 75°$ and approved by the study in [56]. A visualization of the reward function for these thresholds is depicted in Figure 4b. Note that these values are optimal for the reconstruction of objects mainly composed of flat surfaces, while more complex objects with curved or tilted surfaces would require stricter thresholds.

During the computation of observation angles, we also store observation directions due to the requirement of large parallax angles, as stated in requirement (4). The impact of different parallax angles for the reconstruction quality has already been largely investigated in several works [55–57]. In particular, a hemisphere is constructed for each surface point $\boldsymbol{x}_j$ directed along its corresponding normal vector $\boldsymbol{\eta}_j$ and discretized into six distinct segments in order to distinguish between different observation directions. A visualization of the hemispheres for potential camera constellations is shown in Figure 5. Aside from a segment for frontal views occupying an area of a unit circle with an opening angle of α_{min} and a segment for discarded observation angles above α_{max}, the remaining segments occupy equal areas on the surface of the hemisphere. Each viewpoint ray toward an object surface point intersects the hemisphere in a specific segment $\text{seg}(\boldsymbol{p}_i, \boldsymbol{s}_j) \in \{0, 1, 2, 3, 4\}$, which is stored for all visible viewpoint and surface point pairs. We exploit this result in order to find suitable viewpoints intersecting the hemispheres in as many segments as possible and avoiding similar intersection segments, leading to shallow parallax angles and glancing intersections.



(**a**) Good camera constellation (**b**) Bad camera constellation

Figure 5. Observation angle segments for two camera constellations. A hemisphere is generated along the objects point normal vector and divided into different segments. Each camera ray intersects the hemisphere in a specific segment. Good camera constellations intersect the hemisphere in different segments (**a**), while camera pairs with ego-motion and small baselines intersect in the same segment (**b**).

3.5. Submodular Trajectory Optimization

Recent works have shown that the task of path planning for MVS image acquisition can be efficiently addressed by employing submodularity to the candidate view selection [10,11] enabling approximation guarantees on the solution using greedy methods. With respect to our notation in Section 3.1, submodularity is a property of a set function $f : 2^{|\mathcal{P}|} \rightarrow \mathbb{R}$ that assigns each subset $\mathcal{T} \subseteq \mathcal{P}$ a value $f(\mathcal{T})$. $f(\cdot)$ is submodular if for every $\mathcal{T}_1 \subseteq \mathcal{T}_2 \subseteq \mathcal{P}$ and an element $\boldsymbol{p} \in \mathcal{P} \setminus \mathcal{T}_2$ it holds that $\Delta(\boldsymbol{p}|\mathcal{T}_1) \geq \Delta(\boldsymbol{p}|\mathcal{T}_2)$. An equivalent and more commonly used definition of submodularity for $\mathcal{T}_1, \mathcal{T}_2 \subseteq \mathcal{P}$ is given by $f(\mathcal{T}_1 \cup \mathcal{T}_2) + f(\mathcal{T}_1 \cap \mathcal{T}_2) \leq f(\mathcal{T}_1) + f(\mathcal{T}_2)$. In other words, submodularity implies that adding an element to a small subset results in large rewards while adding the same element to a larger subset leads to diminishing returns. Speaking of our path planning problem, as we increase more viewpoint candidates to our trajectory, the marginal benefit of adding another viewpoint with large overlap to the set decreases. Adding the same viewpoint to a smaller set with limited coverage, on the other hand, leads to larger rewards. This property hinders explicit modeling of stereo-matching, as already pointed out by Hepp et al. [11]. Adding a viewpoint to a smaller subset $\mathcal{T}_1$ which does not allow a stereo matching yields less reward (zero, as it is not matchable) than adding it to a larger set $\mathcal{T}_2$ to which it is matchable and therefore it violates the submodularity condition. For that reason, a submodular function $f(\cdot)$ has to be defined which approximates stereo matching in terms of contributions from single views for 3D modeling. This requires $f(\cdot)$ to be both monotone and non-decreasing stated as monotonicity, which means that adding more elements to the set cannot decrease its value. The marginal gain of a viewpoint candidate $\boldsymbol{p}$ toward a trajectory $\mathcal{T}$ is given by $\Delta(\boldsymbol{p}|\mathcal{T}) := f(\mathcal{T} \cup \boldsymbol{p}) - f(\mathcal{T})$. It has been shown that a simple greedy algorithm can be considered for providing a solution of the NP-hard maximization of submodular functions with a reasonable approximation guarantee [58]. Similar to Hepp et al. [11], we constrain our submodular objective function

$$f\left(\boldsymbol{s}_j, \mathcal{T}\right) = \min\left(1, \sum_{\boldsymbol{p}_i \in \mathcal{T}} \frac{1}{v} I(\boldsymbol{p}_i, \boldsymbol{s}_j)\right) \tag{7}$$

to limit the maximum reward for each surface point to 1, where v reduces the obtained reward from a single view in order to enforce at least v different views capturing the same surface point $\boldsymbol{s}_j$. Since this objective function is both monotone and non-decreasing, we can transform the individual information rewards $I(\boldsymbol{p}_i, \mathcal{S}_{\boldsymbol{p}_i})$ from Equation (4) for all viewpoint candidates to tightly additive information rewards I_i^{add} utilizing a simple greedy algorithm given in Algorithm 1. Note that the submodular function in Equation (7) on its own does not explicitly incorporate stereo matching, as it only considers single contributions based on the distance and observation angles from single viewpoints toward the object surface. However, a stereo matching approximation is firstly given by the matchability graph, ensuring paths along the graph for which viewpoints exhibit large overlap toward preceding viewpoints. Secondly, the greedy algorithm incorporates the observation angle segments by penalizing information rewards for camera viewpoints which intersect already seen surface points in the same observation angle segments. This helps to decrease the additive information rewards for cameras with only little parallax angles and therefore avoids ego-motions in the optimized path which are obstructive for stereo matching.

The greedy method iteratively computes the marginal rewards of each viewpoint for the current reconstructability of each surface point and adds the viewpoint with the highest additive information reward I_i^{add} toward the output set. After each iteration, the reconstructability of all surface points is updated according to the previously selected viewpoint rewards. The marginal reward of remaining viewpoints with similar intersection segments of already considered viewpoints is reduced and therefore these are less likely to be chosen in the next iteration. This procedure is repeated until the marginal rewards of all viewpoints have been considered and assigned to the output set. After executing the greedy method, each viewpoint candidate $\boldsymbol{p}_i$ is coupled with a marginal information reward I_i^{add} representing its value for the reconstructability of the object. Roberts et al. [10] presented an efficient way to transform additive rewards into a standard additive orienteering problem, formulated as a mixed-integer programming (MIP) problem, which can be solved with off-the-shelf solvers.

Algorithm 1 The greedy method for maximizing a monotone submodular function.

```
function GREEDY($\mathcal{P}, \mathcal{S}, I(\cdot)$)
    $I \leftarrow \forall \boldsymbol{p} \in \mathcal{P}$ : compute $I(\boldsymbol{p}, \mathcal{S})$    ▷ Compute individual rewards for all viewpoints
    Seg $\leftarrow \forall \boldsymbol{p} \in \mathcal{P}$ : compute seg$(\boldsymbol{p}, \mathcal{S})$    ▷ Compute intersection segments for all viewpoints
    $R \leftarrow \emptyset$    ▷ Initialize reconstructability of object $\mathcal{S}$
    $H \leftarrow \emptyset$    ▷ Initialize observation directions of object $\mathcal{S}$
    for $m \leftarrow 0$ to $|I|$ do
        $i^{\text{add}} \leftarrow \arg\max_{i \in I} f(R \cup i) - f(R) - |H \cap \text{Seg}_i|$
        $R \leftarrow R \cup i^{\text{add}}$
        $H \leftarrow H \cup \text{Seg}_{i^{\text{add}}}$
        $I \leftarrow I \setminus \{i^{\text{add}}\}$
    end for
    return $R, i^{\text{add}}$
end function
```

An orienteering problem can be considered as a combination of a traveling salesman problem and knapsack problem. In other words, the optimization needs to find a closed path that maximizes the collected rewards under a time or travel budget constraint. However, the choice of a suitable travel budget is hard to predict for some scenes and the optimization will almost always fulfil the full path constraint due to the pure additive nature of the rewards which always increases the full coverage of the model. Given an overestimated path length L^{eucl}, a similar amount of total rewards can be obtained with a shorter trajectory by penalizing lengthy paths with a regularization factor λ. With respect to the semantic restriction on the airspace, the optimized trajectory must not exceed a user-defined path length L^{sem} above restricted objects. Summarized, the optimization objective can be formulated as

$$\begin{aligned} \mathcal{T}^* = \arg\max_{\mathcal{T}} & \sum_{\boldsymbol{p}_i \in \mathcal{T}} I_i^{\text{add}} - \lambda \sum_{\boldsymbol{e}_k \in \mathcal{E}} w_k^{\text{eucl}} \\ \text{subject to} & \sum_{\boldsymbol{e}_k \in \mathcal{E}} w_k^{\text{eucl}} < L^{\text{eucl}}, \\ & \sum_{\boldsymbol{e}_k \in \mathcal{E}} w_k^{\text{sem}} < L^{\text{sem}}, \end{aligned} \tag{8}$$

where I_i^{add} defines the additive rewards of the nodes along a path $\mathcal{T}$ with traversed Euclidean distances $\sum_{\boldsymbol{e}_k \in \mathcal{E}} w_k^{\text{eucl}}$ and traversed distances above semantic restricted airspaces $\sum_{\boldsymbol{e}_k \in \mathcal{E}} w_k^{\text{sem}}$. The regularization forces to reduce the maximum path length L^{eucl} for similar optimization results in shorter paths. The second constraint allows the optimization to select nodes in restricted but not prohibited airspaces but, however, encourages to find the most efficient and shortest path through these conditionally accessible airspaces.

4. Experiments

We evaluated the proposed path planning approach both qualitatively and quantitatively with a series of different experiments using synthetic and real-world data. To provide a more profound analysis of the influence of semantic restrictions, the following evaluation consists of two components. First, we needed to evaluate the reconstruction results using our pipeline without semantic constraints in order to validate the general path planning itself. Secondly, comparing these baseline results with the reconstruction results of using paths which consider the semantic constraints. In the optimal case, paths which follow the semantic restrictions should return similar reconstruction results but avoid flyovers of certain objects. Since the complete pipeline from image acquisition to the final

3D model consists of several different tasks, including SfM and MVS, the reconstruction results are highly influenced by the performance of these algorithms. For this reason, we decided to use a state-of-the-art 3D reconstruction pipeline for all experiments with the same settings to analyze the differences in the 3D models as a result of the different acquisition plans. Since Pix4D [1] is a well-known photogrammetric mapping software which integrates state-of-the-art processing steps for both SfM and MVS steps and therefore is often used for processing UAV images, we decided to choose this software for our experiments. However, the results after processing the images with comparable software (e.g., Colmap [3]) do not substantially differ from the results of Pix4D. For the sake of simplicity we therefore only report the results using Pix4D. In order to investigate whether adjacent image viewpoints can be successfully matched, only temporally neighboring images were matched instead of an exhaustive image matching strategy.

Finding suitable data for a comprehensive analysis is hard to realize, as minor modifications of the parameters can end up with different trajectories, for which all of them need to be executed in individual flights. Moreover, comparing the reconstruction results lack the availability of ground truth data on a large scale. For this reason, we generated a synthetic dataset, composed of various objects arranged to a realistic and interchangeable scenery, which allows comparing the reconstruction results derived from different trajectories with exact ground truth. Additionally, we also show the real-world applicability with two real sceneries consisting of different building shapes with a diverse complexity of the surroundings.

4.1. Synthetic Scene

We introduced a new customized synthetic scene (the synthetic scene is freely available at https://www.bgu.tum.de/lmf/synbuil/) which was generated with the open-source computer graphics software Blender [59]. The main object of interest is a conventional living house located at the center of the scene, consisting of a balcony as overhang, an inset doorway and an adjacent garage. The buildings façades are textured with dirt allowing for a dense reconstruction without severe gaps from homogeneous areas. The roof consists of individual 3-dimensional roof tiles allowing for investigations of detailed structures. The building is surrounded by obstacles, such as trees and adjacent buildings placed beside a main road crossing the building. Additionally, a couple of cars are located on the roadside and in front of the building, which are later used to further restrict the airspace. Figure 6 shows an overview of the synthetic scene while properties of the scene are listed in Table 1.

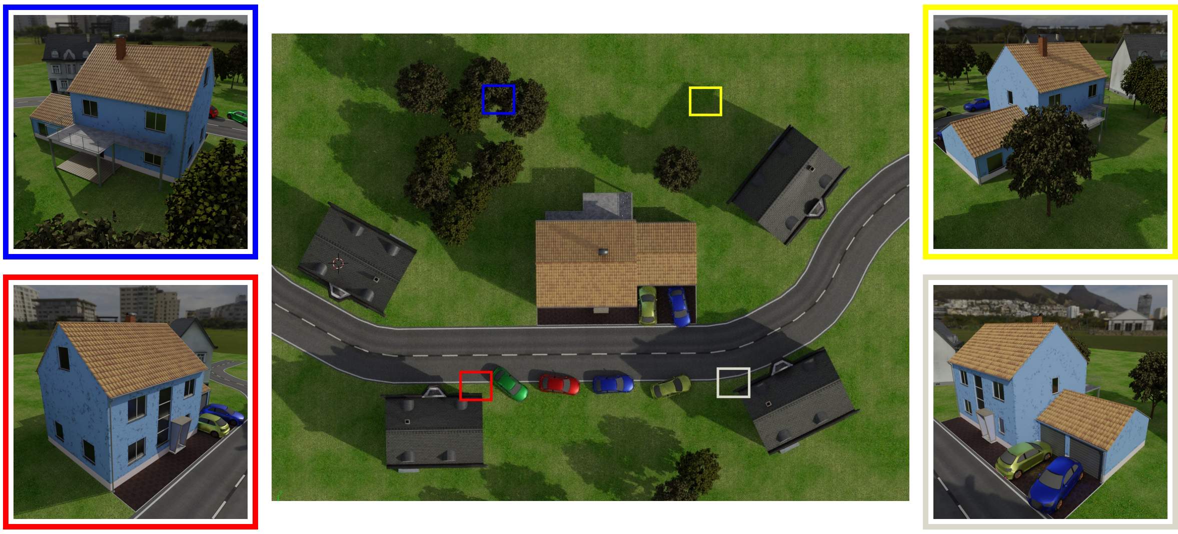


Figure 6. Overview of our synthetic scene used in our experiments (**mid**). Sample views of the building for the highlighted areas are shown in the left and right column.

Table 1. Statistics of the datasets used in our experiments.

Dataset	Data Type	Extent of Building (in m)	Extent of Scene (in m)	Nr. of Nodes	Grid Spacing (in m)	Required GSD (in cm)
House	Synthetic	16 × 8 × 12	50 × 50 × 30	2643	3	2.0
Silo	Real	25 × 22 × 25	93 × 85 × 30	2328	4	2.0
Farm	Real	60 × 16 × 9	110 × 65 × 30	1716	5	1.5

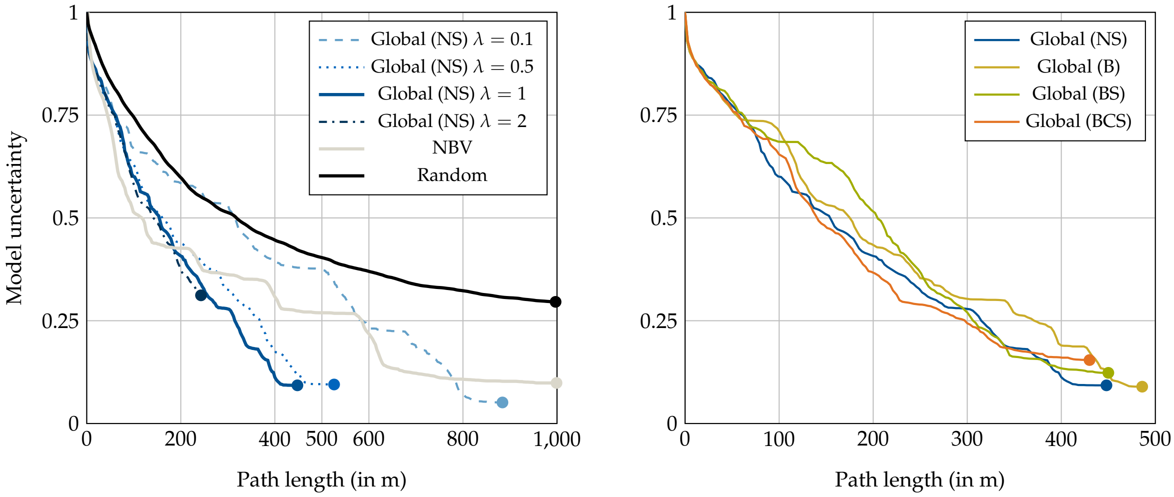
The 3D proxy model, which was considered as input for all methods which were investigated, was created from rendered RGB images of 10 nadir-directed viewpoints at a safe altitude of 70 m encompassing the whole scenery. All images were rendered with a resolution of 750 × 500 px for a virtual camera with a sensor size of $22.2 \times 14.6\,\text{mm}^2$ and a focal length of 30 mm. Since our semantic segmentation network was trained on real UAV images, the generalization on synthetic data is rather poor. For that reason, we additionally rendered semantic maps for all nadir views directly from Blender. The 3D proxy model was derived by feeding the rendered RGB images into Pix4D for generating a dense 3D point cloud, which was further enriched with the semantic maps following the strategy in Section 3.2. Evenly distributed viewpoint hypotheses were sampled in the free airspace from a regular 3D grid with a spacing of 3 m, while keeping a safe distance of 3 m toward all obstacles. The camera orientations for all viewpoints were assigned with the strategy explained in Section 3.3.

4.1.1. Optimization Evaluation

An analysis of the overall performance of the proposed methodology, as well as the influence of the path length regularization term was conducted with the introduced synthetic scene. A series of estimated trajectories with different regularization parameters were compared toward both automated and manual baseline trajectories. Since the objective function in Equation (1), in combination with the proposed heuristics, serves as a measure of the expected certainty of the object's reconstructability R, individual results for the reconstructability can be derived for arbitrary subsets $\mathcal{T}$ of the camera graph without the need of acquiring and processing the images. A study concerning the influence of the regularization for jointly optimizing the reconstructability and the path length was conducted by first optimizing a solely geometrical trajectory without semantic constraints with different regularization parameters λ and an overestimated path length $L_{\text{eucl}} = 1000\,\text{m}$.

Figure 7a depicts the expected model uncertainty with respect to the path length for different values of λ. As expected, low values (e.g., $\lambda = 0.1$) increased the optimized path lengths but also yielded a higher degree of certainty, while large values ($\lambda = 2$) resulted in shorter paths, but reduced certainties of the reconstructability. A reasonable compromise of short path lengths and high model certainty can be realized for regularization parameters in the range of $\lambda = 1$. Compared to the optimized path with $\lambda = 0.1$, a minor loss of 4.2% of the model certainty was recognizable for $\lambda = 1$ whereas the path length had been reduced by half.

We compared the global optimization against two other automated path planning baselines, specifically a random trajectory and an online-capable NBV approach for which both make use of the same camera graph, heuristics, and objective function. Regarding the random trajectory, subsequent views are randomly sampled from the camera graph, while the shortest paths between the selected nodes in our graph are computed until a total path length of 1000 m is reached. Each visited node between two sampled nodes is considered as an acquisition viewpoint. This procedure was repeated for 50 times and the averaged model uncertainty for all obtained trajectories are shown in Figure 7a. Due to the random sampling, highly redundant views from similar positions above the building and only a few views capturing the buildings façades result in a larger degree of model uncertainty compared to the globally optimized trajectories. The path planning strategy of the NBV method starts—similar to the global method—from the viewpoint with maximum reward and greedily selects the next best view from the neighboring nodes according to their marginal rewards. Again, this strategy was repeated until a path length of 1000 m was reached. Comparing toward to the global optimization, NBV rapidly decreases the model uncertainty, as it traverses along the largest gradients of the marginal rewards. However, due to the local search strategy and the highly non-linear nature of the objective function, the NBV approach can get stuck in a local minimum in already seen areas which results in diminishing marginal rewards. This characteristic property is clearly evident in the plateaus of Figure 7a. Summarizing, the NBV method can be effective for fast exploration of the object due to the gradient-based optimization, but, however, does not guarantee to recover all local details of the object. The global approach, on the other hand, exploits submodularity which contributes to the selection of suitable viewpoints covering all parts of the object, while the global optimization refines all viewpoints of the trajectory simultaneously, leading to less redundant acquisition views. It is evident, that the global approach is superior toward the baselines in terms of shorter flight paths and higher model certainty. In addition, the globally optimized trajectories have also been proven to be superior in terms of the quality of the generated 3D models, as presented in Section 4.1.3.



(**a**) Comparison of the reconstructability for different optimization approaches as a function of path length

(**b**) Comparison of semantically-aware optimization for $\lambda = 1$

Figure 7. Comparison of different optimization methods in terms of the expected model uncertainty for different path lengths assuming the same objective function. The effects of various regularization parameters are shown in blue and the performances of baseline approaches are depicted in gray and black (**a**). Note that $\lambda = 1$ leads to a balanced trade-off between short path lengths and high model certainty. Comparison of the sematically-aware global optimization with $\lambda = 1$ for different restrictions on the airspace (**b**).

4.1.2. Semantically-Aware Optimization Evaluation

Following our study in Section 4.1.1, a regularization parameter of $\lambda = 1$ allows for a reasonable trade-off between short path lengths and high model certainty and is therefore kept for further experiments investigating the semantic constraints of the airspace. Since path optimizations in Section 4.1.1 only consider purely geometric constraints resulting in a collision-free and matchable viewpoint path in the camera graph, we additionally restrict and prohibit certain airspaces according to the semantics of the underlying proxy model. Depending on the application, restrictions can be defined in two ways: a hard restriction eliminates nodes and their corresponding edges above a certain semantic cue in the camera graph, while soft restrictions limit the path length to a maximum tolerable distance L^{sem} above specified semantic cues. The latter is realized by the secondary condition in Equation (1).

Precisely, we optimized three semantically constrained trajectories with the following restrictions:

- No semantics (NS): this path from Section 4.1.1 serves as a baseline and only considers geometric constraints.

- Building (B): hard restriction for airspaces above other buildings than the target building.
- Building and Street (BS): in addition to (B), airspaces above streets are softly restricted to maximum path length of $L = 12\,\mathrm{m}$, approximately twice the width of a regular street.
- Building, Car and Street (BCS): in addition to (BS), hard restrictions above cars are imposed.

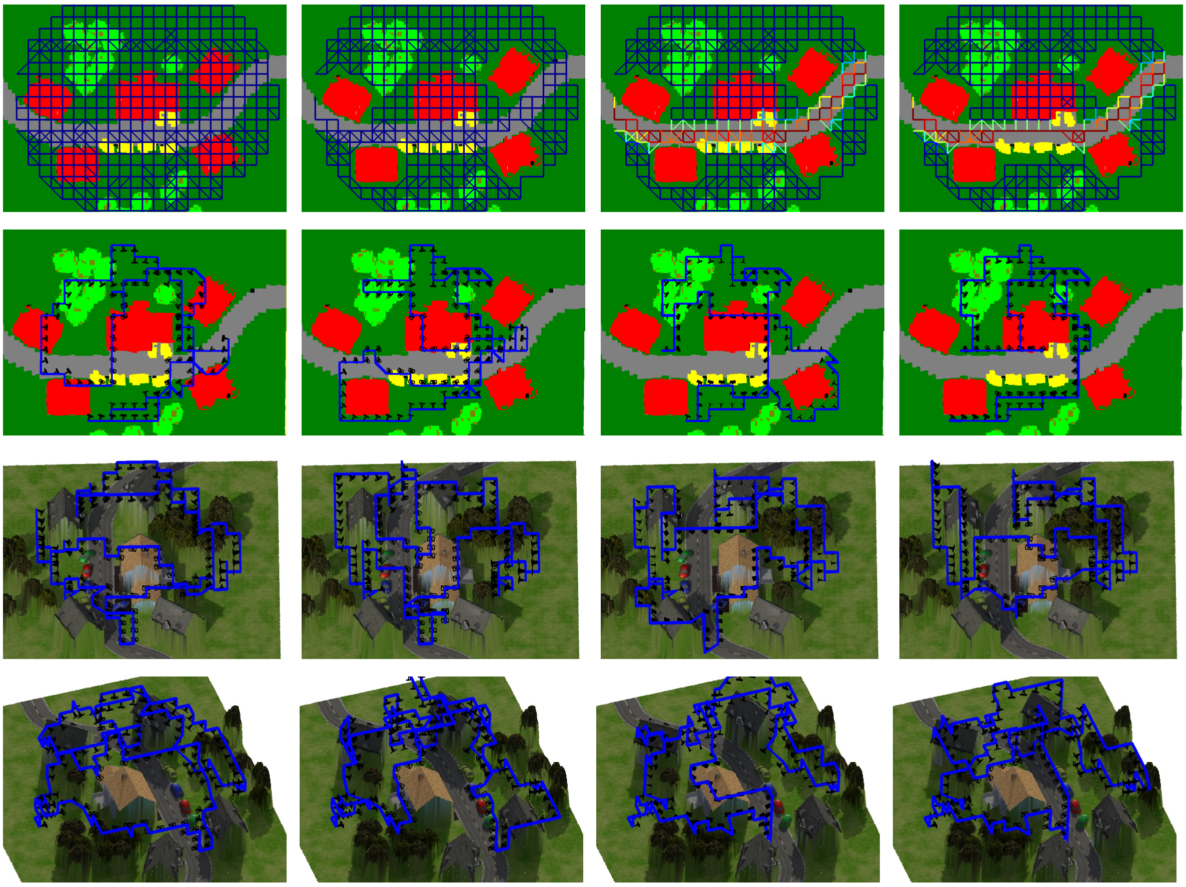
The semantic constraints affect the generation of the camera graph, resulting in a limited number of accessible nodes (hard restrictions) and only conditionally accessible nodes (soft restrictions). Statistics of the affected nodes and edges for the synthetic scene are listed in Table 2. The optimization for the semantically-aware path plans was conducted in the same fashion as in Section 4.1.1, except for the additional side constraint for the soft restrictions above streets. A comparison of the optimized paths with respect to the path length and model uncertainty is shown in Figure 7b. From this figure it can be seen that, despite further restrictions in the airspace, only slight losses in the model certainty have to be expected from the optimized paths, indicating that satisfactory reconstruction results can be achieved from these restricted trajectories with a similar path length.

Table 2. Effects of semantic constraints on the graph generation for the synthetic scene. The free airspace is further restricted for various semantical constraints affecting the number of accessible and conditionally accessible nodes.

Constraint	**Nodes**		**Edges**	
	Free	**Restricted**	**Free**	**Restricted**
No semantics (NS)	2643	0	13,634	0
Building (B)	2333 (88%)	0	11,836 (87%)	0
Building and Street (BS)	2333 (88%)	459 (20%)	11,836 (87%)	2555 (21%)
Building, Car and Street (BCS)	2208 (83%)	388 (17%)	11,084 (81%)	2164 (20%)

As follows from the visualization of the optimized paths in Figure 8, the increase of restrictions on the airspace has a substantial influence on the estimated path along the camera graph, yet yielding reasonable trajectories encompassing the entire building while avoiding prohibited objects. It is worth noting that the soft constraint on streets for (BS) and (BCS) result in trajectories which simply cross the street twice in a direct way at suitable locations. In terms of flight safety, these trajectories were by far more desirable than the unconstrained path, since risky long-term periods above hazardousness roads were mostly avoided. As outlined in Figure 9, the validity of the semantical restrictions can also be expressed as histograms of traversed semantic labels of the proxy model for the optimized paths. While viewpoints above streets were favored for (NS) and (B), they were highly avoided for (BS) and (BCS).

Since the heuristics were already computed for all potential viewpoints, the expected reconstruction quality can be assessed for only subsets obtained by the estimated trajectories. This allows for investigating whether the photogrammetric requirements are met for each surface point by analyzing the observation distances and observations angles between surface points and selected viewpoints, as well as their multi-view configuration. Distributions of the individual photogrammetrical properties of each surface point for different semantically-aware trajectories are shown in Figure 10. It is apparent that around 75% of the surface points were mapped from at least three different perspectives according to the observation direction segments when considering the non-semantically restricted path (NS). Changes in the semantic-based restrictions on the free airspace only affected 4.4% of the surface points for the utmost restriction on the airspace (BCS). Regarding the observation angles, up to 64% of the surface points were seen within 15° observation angle and 85% with less than 30°, indicating the compliance of fronto-parallel views. Similar to the multi-view assessment, further restrictions on the airspace had only a minor effect on the observation angles. The maximum distance for achieving a GSD below 2 cm with the virtual camera is $d_{\max} = 20\,\mathrm{m}$ which was met for 70% of the surface points. It is worth mentioning, that for an increasing restriction on the airspace even closer views were selected. Reason for this finding is that viewpoints with an optimal distance toward the object could be restricted and eluded to closer views for gaining at least an equal amount of rewards instead of more distant views with fewer rewards.



(**a**) Global (NS) (**b**) Global (B) (**c**) Global (BS) (**d**) Global (BCS)

Figure 8. Visualization of the optimized paths for different semantical restrictions on the airspace for the synthetic scene. The first row shows a nadir view of the entire camera graph as accessible and traversable UAV viewpoints. Color-coded edges represent associated semantical costs w_k^{sem} for the corresponding restrictions. The second row visualizes the optimized camera paths together with the acquisition viewpoints as black camera symbols. Different perspectives with the RGB proxy model are shown in the third and fourth row.

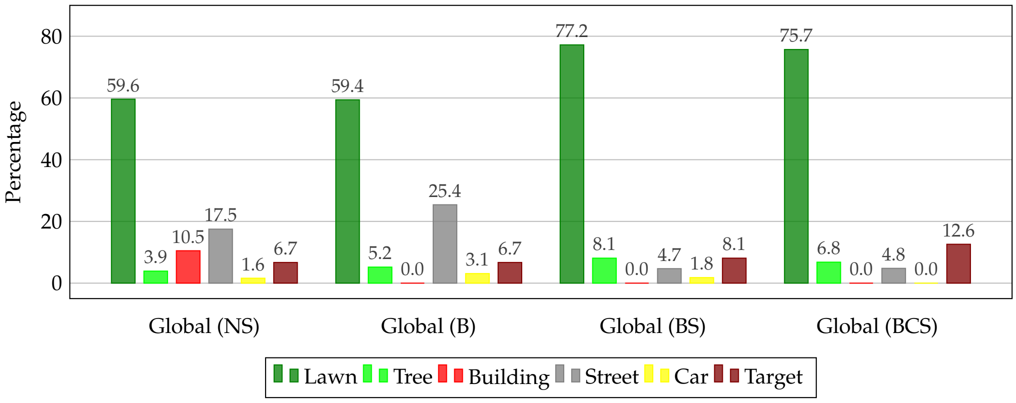
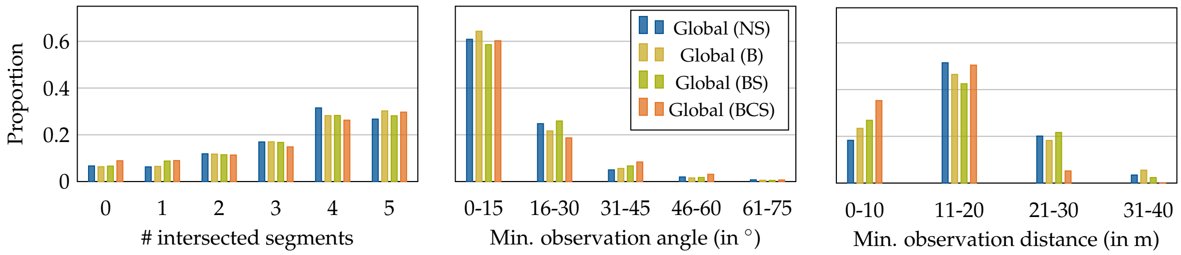


Figure 9. Proportions of traversed ground labels for different semantically constrained flight paths.



(**a**) Multi-view assessment. (**b**) Observation angle. (**c**) Distance.

Figure 10. Evaluation of heuristics for optimized semantically-aware trajectories per surface point s. Number of intersected surface hemispheres (**a**), minimum observed observation angles (**b**) and minimum observation distance (**c**).

4.1.3. Reconstruction Performance

The use of the synthetic model allows for a revealing quantitative and qualitative evaluation of the reconstruction quality from arbitrary viewpoints. RGB images from the viewpoints of the globally optimized paths and baseline paths were rendered in Blender and subsequently processed in Pix4D, including the registration of the images and the generation of a densified point cloud, which was further assessed with respect to the ground truth model. Following the evaluation protocol of related works [10,11,37,60], the quality of the reconstructed point clouds can be quantitatively assessed by comparing them toward the ground truth model using the quantities of precision, completeness, and F-score. Precision quantifies how many reconstructed points are located close to the ground truth model with a distance equal or less than an investigated threshold d. Completeness is defined as vice versa and quantifies how many ground truth points are located in an equal or less distance toward the reconstructed points than d. We analyzed the results for two different thresholds $d_1 = 5\,\text{cm}$ and $d_2 = 10\,\text{cm}$. Furthermore, an assessment of the point density, which was required to be consistent along the entire object surface, was conducted by computing geometrical distances between neighboring reconstructed points.

Additionally, we compared the optimized paths against commonly used flight planning baselines, precisely we generated circular flights at two different altitudes and radii (30 m altitude with 30 m radius and 20 m altitude with 37 m radius) with oblique views pointing toward the center of the building. A quantitative evaluation regarding the reconstruction errors and point density error are listed in Table 3 and a visualization of the spatial occurrences of these errors are shown in Figure 11. While circular baseline paths revealed unsatisfying reconstruction results in terms of a low point density and gaps in the reconstruction due to occlusions from overhangs of the roof and balcony, the unconstrained global optimization (NS) yielded best reconstruction quality for all investigated errors. The distance-based heuristics led to close-up views, resulting in a high global point density of more than 97% for all reconstructed points of the building. Comparing the completeness error, lower circular flights yielded less optical occlusions, which, however, is limited by the surrounding environment. Paths considering the proxy model generally performed better in terms of completeness, since low altitude viewpoints can be selected from the free airspace, however globally optimized paths revealed significantly better completeness, especially for occluded areas. The last few percentage points are generally hard to achieve since the building consists of different materials, such as windows, which are generally difficult to reconstruct. It is worth noting that, according to Section 4.1.1, all globally optimized paths did not exceed a path length of 490 m acquiring a maximum amount of 162 images for (B), while both random and NBV paths were limited to 1000 m resulting in 321 and 323 viewpoints, respectively. The visualizations in Figure 11 reveal local inaccuracies for the random and NBV paths, whereas all globally optimized paths show decent results for all parts of the building. The most difficult part concerns the façade beneath the balcony and the occluded façade of the garage caused by the single tree, whereby former resulted from a low contrast of the weakly textured and illuminated façade and latter from a hardly observable area.

Table 3. Quantitative evaluation of the reconstruction results for the synthetic scene obtained from different path planning methods. We report the point density as the percentage of reconstructed points that have a shorter distance towards their nearest neighbor than the demanded GSD = 2 cm, as well as one and a half times the distance (1.5 · GSD = 3 cm). The reconstruction errors are stated for $d_1 = 5$ cm and $d_2 = 10$ cm. The proposed globally optimized paths are superior to the baseline methods, while featuring a shorter path. The severely limited free airspaces due to different semantic restrictions lead only to a slight drop in the reconstruction quality.

Method	**Images**	**Density (%) ↑**		**Precision (%) ↑**		**Completeness (%) ↑**		**F-Score (%) ↑**	
		GSD	**1.5·GSD**	d_1	d_2	d_1	d_2	d_1	d_2
Circle 30 m	100	46.9	73.3	88.8	96.3	79.2	91.8	83.7	94.0
Circle 20 m	100	29.7	60.3	89.7	95.8	84.0	93.9	86.7	94.8
Random	321	94.1	98.9	96.4	98.6	83.3	91.0	89.4	94.6
Greedy NBV	323	96.9	99.8	96.8	98.7	86.5	92.7	91.4	95.6
Global (NS)	148	97.6	99.9	96.7	98.9	91.1	95.7	93.8	97.2
Global (B)	162	97.3	99.8	96.2	98.7	88.3	95.5	92.1	97.1
Global (BC)	148	97.6	99.8	96.4	98.7	89.4	94.8	92.8	96.7
Global (BCS)	152	97.3	99.8	96.5	98.8	87.7	95.1	91.9	96.9

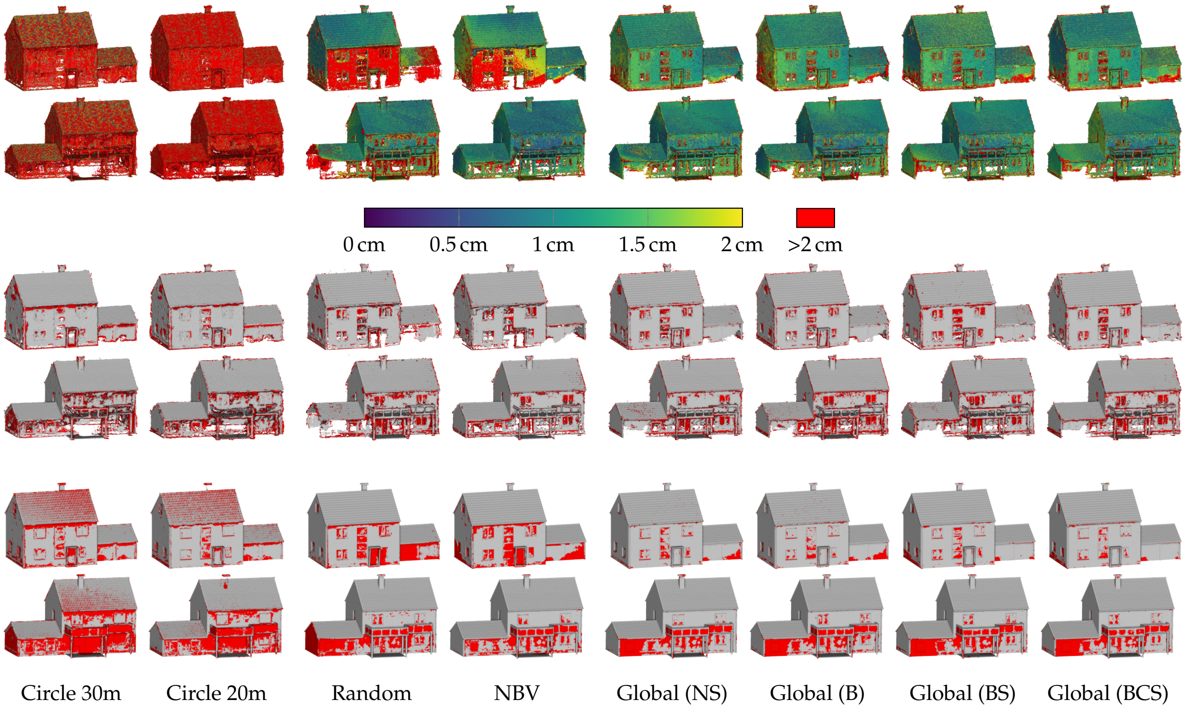


Figure 11. Qualitative comparison of the reconstruction results on a dense point cloud for the synthetic scene using different methodologies (columns). The first two rows show the point density as colored distances towards adjacent points on the front and rear side of the building, while red points indicate distances above the required GSD of 2 cm. The reconstruction errors of precision and completeness for $d = 5$ cm are visualized in rows three and four, and rows five and six, respectively, wherein red points indicate erroneous areas.

Comparing the results of different semantic restrictions on the airspace, only a minor decrease in terms of completeness is notable, which matches the expected model uncertainty in Figure 7b. Regarding the precision of the reconstruction—a quality measure according to the noise of the reconstruction depending on the camera constellations—it can be noted that all paths considering viewpoints from our camera graph achieved comparable good values, which proves the suitability of the proposed viewpoint generation process in Section 3.3.

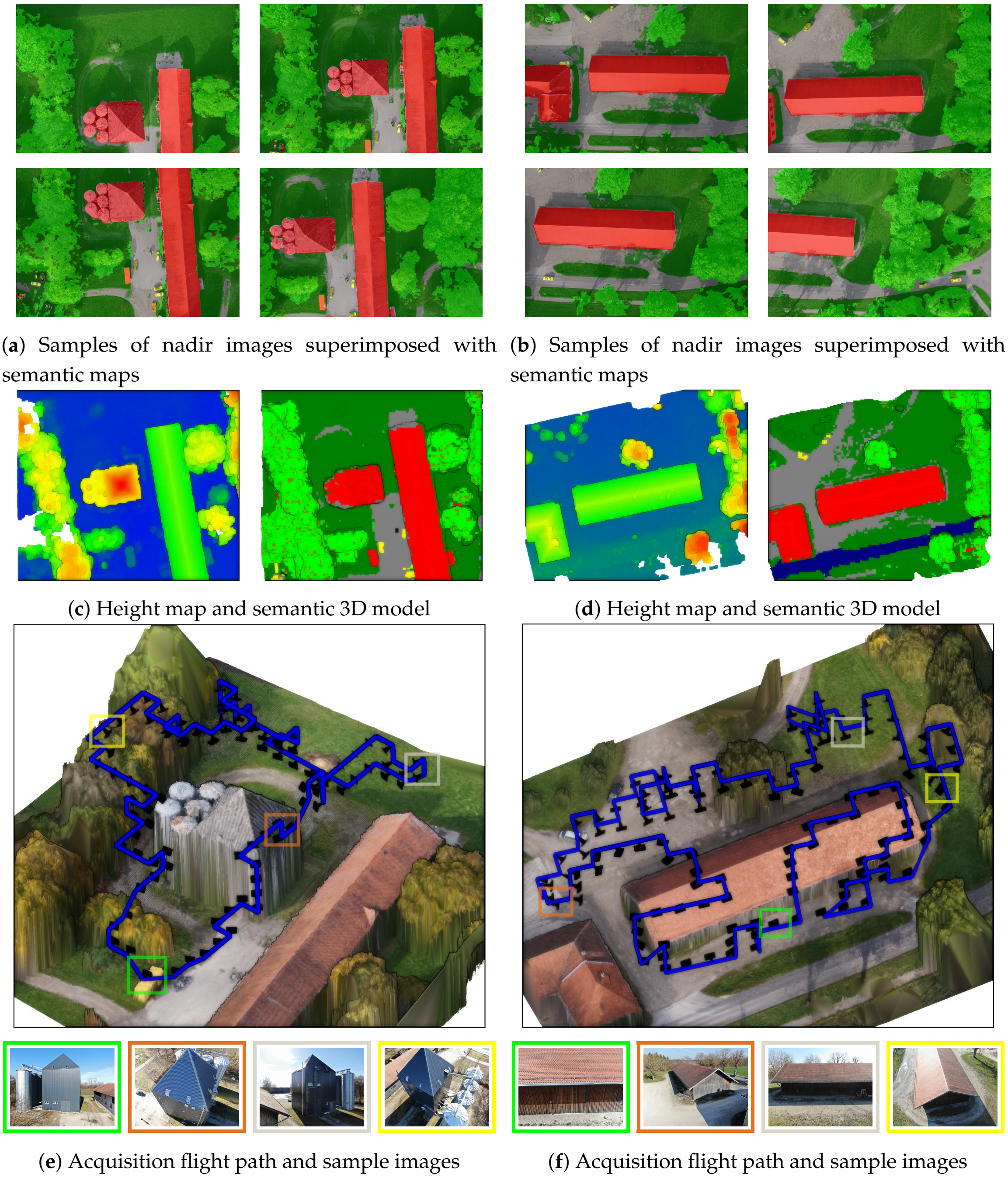
4.2. Real-World Performance

We showed the real-world applicability of our methodology by planning and executing safe flight paths for high-fidelity reconstructions of two buildings. Our experimental site consisted of a silo and a farm building, which define the objects to be finally reconstructed from our estimated flight paths with a user-specified GSD for the entire object surface w.r.t. known camera intrinsics. The buildings differed in their shapes, while the surrounding environment featured hazardous obstacles—such as high vegetation, buildings, cars and a trunk road—which were considered during the flight planning. An overview of the real-world scenes are depicted in Figure 12 and statistics of the scene extent are shown in Table 1. We evaluated and qualitatively compared the reconstructed models generated with acquired images from the estimated trajectories against regular baseline flight paths prepared in accordance with established flight planning practices. A DJI Mavic Pro 2 was used for both experiments, equipped with a 12 Mpx Hasselblad camera with a focal length of 24 mm. The parameters of the camera intrinsics were included in our heuristic computation. The estimated flight plans were finally executed by uploading the waypoints to the UAV, followed by an autonomous acquisition flight without human intervention. Similar to the reconstruction process in Section 4.1.3, the acquired images were processed in Pix4D for generating a dense point cloud and a triangulated mesh, which served as our final reconstruction model.

4.2.1. Silo

The first object of interest was a high-rise granary, which featured large planar façades. In order to generate a high fidelity reconstruction, the flight path required low flight altitudes for capturing frontal images of the façades as well as glimpses of an occluded façade by contiguous pipes. The surroundings of the granary impeded the execution of a simple circular low altitude flight by high vegetation and another building. We generated an optimized path, avoiding the high-grown trees and restricting fly-overs above the adjacent building.

The initial model was generated from eight nadir images acquired in a grid-like pattern at 100 m altitude encompassing the entire surrounding area. Since the reconstruction of branches and leaves is often incomplete due to their small size and the disturbance of wind affect the consistency of matches across multiple images, we sampled evenly distributed points from a coarse mesh, filling up gaps in the reconstruction model. After inferring the images into the FCN model, we assigned each point in the model with a semantic label leading to the semantic initial 3D model used for our trajectory planning. Visualizations of the semantic input images and the respective proxy model are shown in Figure 12a,c. A total amount of 2328 viewpoint hypotheses in the accessible airspace was sampled with a grid spacing of 4 m, while keeping a safety buffer of 10 m toward high vegetation and 5 m toward other obstacles. The viewpoints were evaluated in terms of a required GSD of 2.0 cm (for half image resolution) and connected in the camera graph when a mandatory overlap of adjacent views of at least 75% was met. The optimization was conducted with $\lambda = 1$, yielding to a trajectory with a path length of 405 m with 98 different views. A visualization of the optimized trajectory and its viewpoints is shown in Figure 12e. The path features both oblique views covering the roof of the silo and close-up fronto-parallel views of the façades, while it avoids the surrounding trees and passes through the narrow gap between the two buildings without crossing the adjacent building. We compared the reconstruction results using the acquired images of the optimized trajectory against a baseline of a circular flight at a safe altitude of 40 m with 90 acquired images pointing toward the center of the silo. The reconstructed models of both paths are shown in Figure 13, while Table 4 compares the trajectories and reconstruction results for the baseline and optimized trajectory. It was evident that our optimized path recovered a higher amount of details than the baseline path, as well as a more complete model, even for hardly observable parts of the silo, such as the highly occluded façade and the façade towards the restricted airspace above the adjacent building. The triangulated mesh exhibited planar façades but still preserved local details, such as sharp edges and almost completely reconstructed pipes with a high level of detail. The visualization of the closest distances for the reconstructed points shows that the desired GSD was achieved for almost every part of the silo, except for the occluded façades. The point density of the baseline model, on the other hand, decreases toward the ground part of the building, due to a fixed flight altitude. Moreover, the baseline path was not able to recover the occluded façade, exhibits distortions in the planar façades, and lost the preservation of local details.



(**a**) Samples of nadir images superimposed with semantic maps

(**b**) Samples of nadir images superimposed with semantic maps

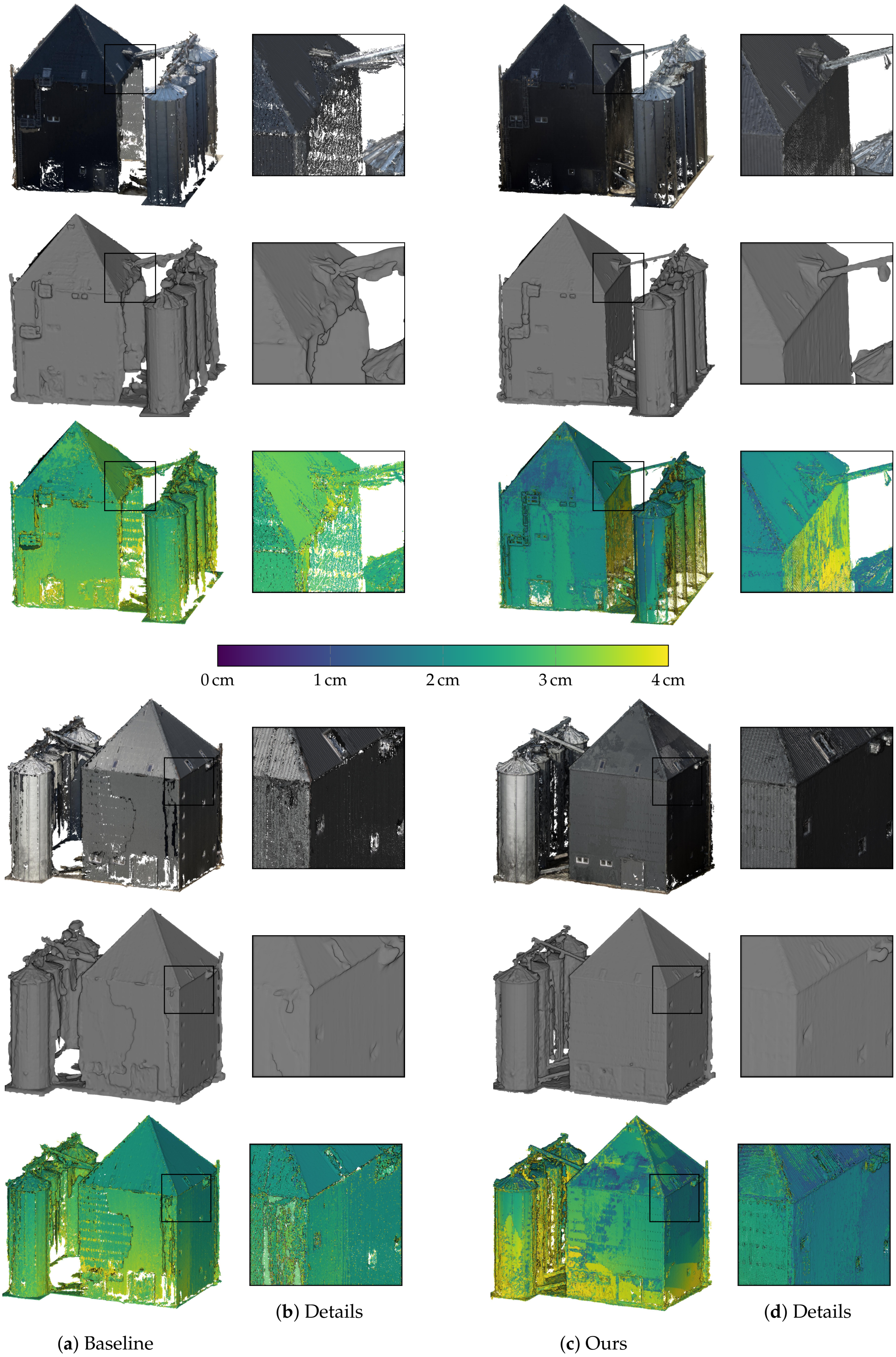
(**c**) Height map and semantic 3D model

(**d**) Height map and semantic 3D model

(**e**) Acquisition flight path and sample images

(**f**) Acquisition flight path and sample images

Figure 12. Real world experiments for the silo scene (**left**) and farm scene (**right**). A set of segmented nadir images (**a**,**b**) is used to generate a semantically enriched 3D proxy model of the entire scene (**c**,**d**). The final trajectories (blue lines) and discrete image acquisition viewpoints (black cameras) are visualized in (**e**,**f**) including sample images for the highlighted viewpoints. The restricted areas include adjacent buildings for the silo scene and adjacent buildings, as well as trunk roads for the farm scene.



(**b**) Details

(**d**) Details

(**a**) Baseline

(**c**) Ours

Figure 13. Qualitative reconstruction results for the silo scene using a baseline UAV path (**a**) and our optimized path (**c**). Rows show the densified point cloud (**top**), a triangulated mesh (**mid**) and the point density (**bottom**). Two different viewpoints are visualized which are separated by the colormap of the point density, showing the closest distances between adjacent 3D points.

Table 4. Comparing automatic and semantically-aware trajectories toward established baseline trajectories for the real world experiments including the number of acquired images, the number of acquisition viewpoints above restricted areas, the path length, and the average distance and standard deviation of adjacent 3D points after generating a dense 3D point cloud from the acquired images. More details about the generation of the baseline trajectories are given in Sections 4.2.1 and 4.2.2.

Dataset	**Baseline**				**Optimized**			
	Images	**Restricted Viewpoints**	**Path Length (m)**	**Density (cm)**	**Images**	**Restricted Viewpoints**	**Path Length (m)**	**Density (cm)**
Silo	90	24	184	2.2 ± 1.2	98	0	405	2.0 ± 0.9
Farm	89	23	732	2.3 ± 0.8	131	0	677	0.9 ± 0.4

4.2.2. Farm

The second object is an elongated farm building of low height and with large overhangs from the roof toward the buildings façades. In order to recover the entire building, it is, therefore, necessary to capture images from very low altitudes facing the buildings façades. However, the surrounding environment, as shown in Figure 12b,d, impeded established flight planning due to high-grown adjacent trees, buildings, and a crossing trunk road. In particular, the latter should avoid being overflown, especially at very low altitudes. For that reason, the semantic proxy model was further enriched by the use of OSM data by converting already as road labeled 3D points into restricted areas. Similar to the silo scene, further restrictions were imposed on flights above other buildings.

The parameters for the optimization were set in the same way as in Section 4.2.1, yielding a trajectory with a path length of 677 m and 131 unique image acquisition viewpoints, as shown in Figure 12e. The trajectory strictly follows the boundary toward the trunk road, avoids the adjacent building, and evades the single tree in front of the buildings façade. Besides oblique images covering the roof of the building, the buildings façades were captured from fronto-parallel viewpoints at very low altitudes up to 5 m. A comparison of the reconstruction result from the optimized trajectory was conducted against a baseline of a grid-like acquisition pattern at 40 m with 89 images pointing toward the center of the building. Table 4 summarizes both trajectory statistics and quantitative results of the obtained 3D models from the baseline and optimized trajectories, while visualizations of the 3D models are shown in Figure 14. Due to the large overhangs of the building's roof and the high altitude of the baseline trajectory, the façades are mostly occluded and therefore hardly reconstructed. In contrast, the reconstruction of the buildings façades in the model of the optimized trajectory is vastly improved, with the exception of a partial gap at one side, caused by a technical malfunction of the gimbal of the UAV for some images. It is worth noting that the optimized model features a similar point density for both the roof and the façades of the building in the range of the required GSD. Comparing the point densities of both models, 95.2%, 99.7% and 99.9% of the reconstructed 3D points derived from the optimized trajectory have an equal or less distance toward adjacent neighboring points for different distance thresholds ($d_1 = 1.50$ cm, $d_2 = 2.25$ cm, $d_3 = 3.00$ cm), while the baseline only achieved 17.1%, 37.1% and 91.8%, respectively.

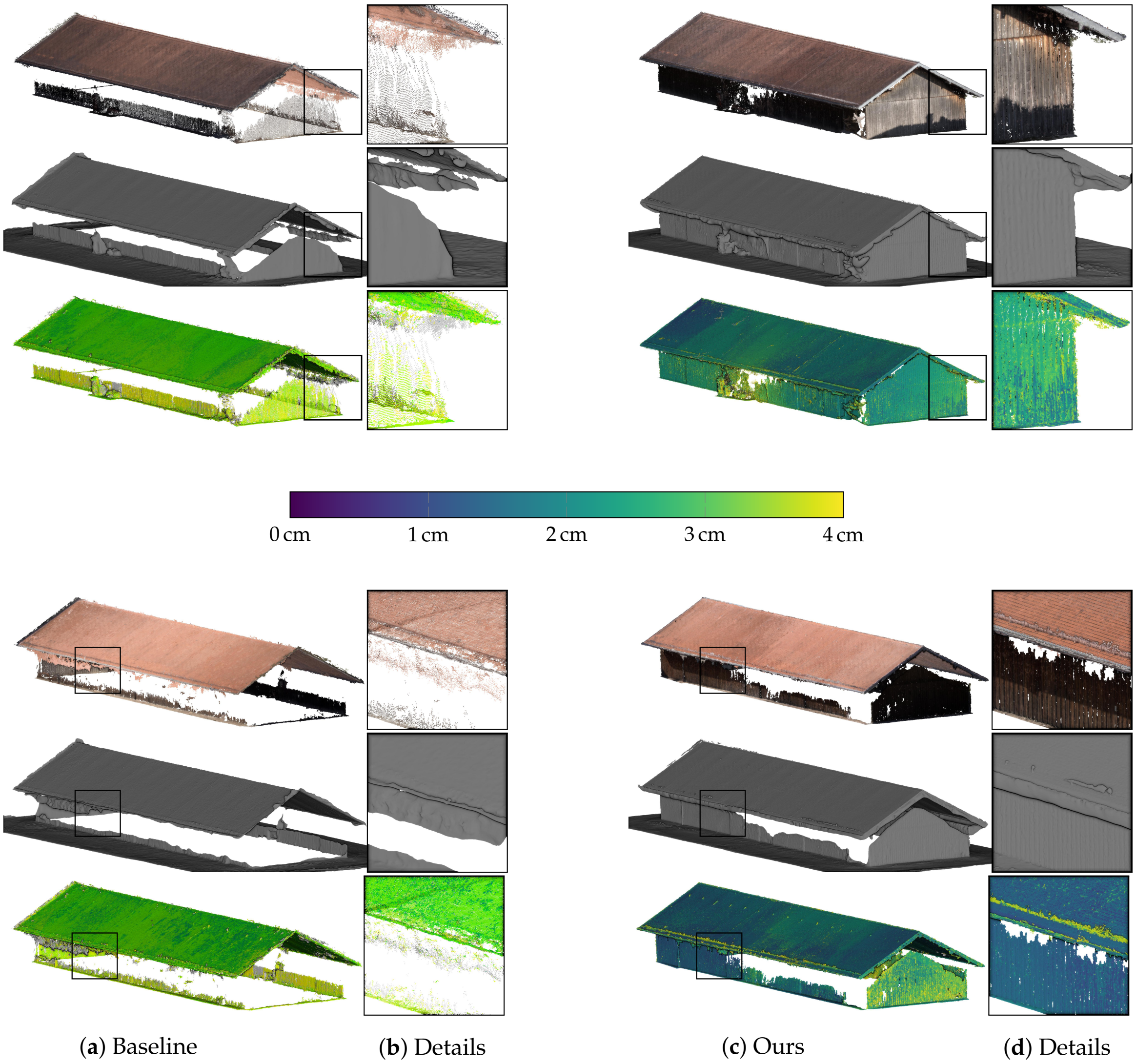


Figure 14. Qualitative reconstruction results for the farm scene using a baseline UAV path (**a**) and our optimized path (**c**). Rows show the densified point cloud (**top**), a triangulated mesh (**mid**) and the point density (**bottom**). Two different viewpoints are visualized which are separated by the colormap of the point density, showing the closest distances between adjacent 3D points.

5. Conclusions

We proposed a semantically-aware 3D UAV path planning pipeline for acquiring images to generate high-fidelity 3D models. Our framework is based on a semantically-enriched proxy model of the environment from a set of safely acquired images, which is used to restrict and prohibit accessible airspaces for the UAV, allowing for safe acquisition paths in complex and densely built environments. An optimized subsequent refinement path allows for acquiring a sequence of close-up images with respect to a user-defined model resolution and fulfils the requirements of SfM and MVS image acquisition, considering the surrounding geometric and semantic environment. We proposed a set of meaningful heuristics and exploit submodularity for formulating the path planning problem as a discrete graph-based optimization. The optimization follows an orienteering problem and maximizes the reconstructability of the object while minimizing the corresponding path length. Additionally, it includes the avoidance of prohibited airspaces and respects conditionally restricted airspaces, such as traversing highly frequented roads.

Experiments on synthetic and real-world scenes have demonstrated the applicability of our proposed method requiring only minimal human interaction for complicated scenes, for which established flight plans yield insufficient reconstruction results and highly experienced pilots are demanded for manual operation of the vehicle. We have shown that the optimized trajectories are safe in terms of user-specified restrictions and prohibitions on the accessible airspace but are still capable of generating high-fidelity reconstruction models with respect to the desired model resolution. The model-based approach and the proposed heuristics furthermore allow for retrieving information about the expected reconstruction quality before the actual execution. This allows for further adaptations of the flight path or even the localization of suitable image acquisition viewpoints on the ground level for capturing images of hardly observable parts of the object with a hand-held camera. It is worth noting, that the proposed framework is not limited to building reconstruction tasks, but will work for any 3D object of interest.

The discrete nature of the regularly sampled viewpoints leads to a multitude of images, which are necessary for the registration of adjacent views but do not enhance MVS processing, for which similar results could be achieved with only a subset of the acquired images. This limits the potential placement of viewpoints and leads to over- and undersampled areas. A more flexible viewpoint placement strategy could lead to even more sophisticated viewpoints with a reduced number of hypotheses, thus a reduction of the optimization complexity. Furthermore, it would be conceivable to include additional costs to the optimization for the gimbal motion needed between adjacent viewpoint perspectives in order to minimize the required gimbal operations for the entire flight. Although the optimization would account for estimating a single trajectory for reconstructing several isolated target objects at the same time, the viewpoint orientations are currently assigned toward a single target object. However, an extension of the optimization could incorporate multiple orientations for each camera viewpoint toward several target objects. Hepp et al. [11] and Roberts et al. [10] have shown that viewpoint orientations can be integrated in the optimization as well, however, the complexity of the optimization exceedingly increases. A reduction to only few meaningful orientation hypotheses for each viewpoint would be favorable for the optimization, which selects the best perspective for each viewpoint for maximizing the total reconstructabililty of all target objects. Besides leveraging semantics for restricting the airspace, an extension of the trajectory optimization could include respecting the material of individual object parts, such as windows, roofs, and façades, which require customized acquisition requirements. In terms of safe automated flight planning, further research should include keeping the UAV in sight with the pilot at any time during the autonomous acquisition flight, by, for instance, constraining the UAV trajectory optimization to the visible airspace of the pilot's path.

Author Contributions: T.K. contributed to the conceptualization of the work; T.K. and F.F. proposed the methodology and designed the experiments; T.K. performed the experiments and analyzed the data; T.K. prepared the original draft and all authors contributed to reviewing and editing the manuscript. This project is under the supervision of F.F. and M.K. and under the administration of M.K.

Funding: This work was supported by the German Research Foundation (DFG) and the Technical University of Munich (TUM) in the framework of the Open Access Publishing Program.

Acknowledgments: We thank our student Markus Spitzner for generating the synthetic scene using Blender.

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

References

1. Pix4D: Pix4Dcapture. Available online: https://pix4d.com/product/pix4dcapture/ (accessed on 28 May 2019).
2. Snavely, N.; Seitz, S.M.; Szeliski, R. Photo Tourism: Exploring Photo Collections in 3D. *ACM Trans. Graph.* **2006**, *25*, 835–846. [CrossRef]
3. Schönberger, J.L.; Frahm, J.M. Structure-from-Motion Revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.

4. Vacanas, Y.; Themistocleous, K.; Agapiou, A.; Hadjimitsis, D. Building Information Modelling (BIM) and Unmanned Aerial Vehicle (UAV) Technologies in Infrastructure Construction Project Management and Delay and Disruption Analysis. In Proceedings of the International Conference on Remote Sensing and Geoinformation of the Environment, Paphos, Cyprus, 16–19 March 2015.
5. Hallermann, N.; Morgenthal, G. Visual Inspection Strategies for Large Bridges using Unmanned Aerial Vehicles (UAV). In Proceedings of the 7th International Conference on Bridge Maintenance, Safety and Management (IABMAS), Shanghai, China, 7–11 July 2014; pp. 661–667.
6. Mostegel, C.; Prettenthaler, R.; Fraundorfer, F.; Bischof, H. Scalable Surface Reconstruction from Point Clouds with Extreme Scale and Density Diversity. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 904–913.
7. Precisionhawk: Precision Flight. Available online: https://www.precisionhawk.com/precisionflight/ (accessed on 28 May 2019).
8. DJI: Flight Planner. Available online: https://www.djiflightplanner.com/ (accessed on 28 May 2019).
9. Ardupilot: Mission Planner. Available online: http://ardupilot.org/planner/ (accessed on 28 May 2019).
10. Roberts, M.; Dey, D.; Truong, A.; Sinha, S.; Shah, S.; Kapoor, A.; Hanrahan, P.; Joshi, N. Submodular Trajectory Optimization for Aerial 3D Scanning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5324–5333.
11. Hepp, B.; Nießner, M.; Hilliges, O. Plan3D: Viewpoint and Trajectory Optimization for Aerial Multi-View Stereo Reconstruction. *ACM Trans. Graph.* **2018**, *38*, 4. [CrossRef]
12. Cheng, P.; Keller, J.; Kumar, V. Time-optimal UAV Trajectory Planning for 3D Urban Structure Coverage. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Nice, France, 22–26 September 2008; pp. 2750–2757.
13. Chakrabarty, A.; Langelaan, J. Energy Maps for Long-range Path Planning for Small-and Micro-UAVs. In Proceedings of the AIAA Guidance, Navigation, and Control Conference (GNC), Chicago, IL, USA, 10–13 August 2009; p. 6113.
14. Di Franco, C.; Buttazzo, G. Coverage Path Planning for UAVs Photogrammetry with Energy and Resolution Constraints. *J. Intell. Robot. Syst.* **2016**, *83*, 445–462. [CrossRef]
15. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
16. 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 (ICCV), Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
17. Furukawa, Y.; Curless, B.; Seitz, S.M.; Szeliski, R. Towards Internet-scale Multi-view Stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 1434–1441.
18. Rumpler, M.; Irschara, A.; Bischof, H. Multi-view Stereo: Redundancy Benefits for 3D Reconstruction. In Proceedings of the 35th Workshop of the Austrian Association for Pattern Recognition (AAPR), Graz, Austria, 26–27 May 2011.
19. Furukawa, Y.; Hernández, C. Multi-view Stereo: A Tutorial. *Found. Trends Comput. Graph. Vis.* **2015**, *9*, 1–148. [CrossRef]
20. Nex, F.; Remondino, F. UAV for 3D Mapping Applications: A Review. *Appl. Geomat.* **2014**, *6*, 1–15. [CrossRef]
21. Kriegel, S.; Rink, C.; Bodenmüller, T.; Suppa, M. Efficient Next-best-scan Planning for Autonomous 3D Surface Reconstruction of Unknown Objects. *J. Real-Time Image Process.* **2015**, *10*, 611–631. [CrossRef]
22. Heng, L.; Lee, G.H.; Fraundorfer, F.; Pollefeys, M. Real-time Photo-realistic 3D Mapping for Micro Aerial Vehicles. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), San Francisco, CA, USA, 25–30 September 2011; pp. 4012–4019.
23. Sturm, J.; Bylow, E.; Kerl, C.; Kahl, F.; Cremers, D. Dense Tracking and Mapping with a Quadrocopter. In Proceedings of the ISPRS—International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XL-1/W2, Rostock, Germany, 4–6 September 2013; pp. 395–400.
24. Loianno, G.; Thomas, J.; Kumar, V. Cooperative Localization and Mapping of MAVs using RGB-D Sensors. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 4021–4028.

25. Michael, N.; Shen, S.; Mohta, K.; Kumar, V.; Nagatani, K.; Okada, Y.; Kiribayashi, S.; Otake, K.; Yoshida, K.; Ohno, K.; et al. Collaborative Mapping of an Earthquake Damaged Building via Ground and Aerial Robots. *J. Field Robot.* **2012**, *29*, 832–841. [CrossRef]
26. Fan, X.; Zhang, L.; Brown, B.; Rusinkiewicz, S. Automated View and Path Planning for Scalable Multi-object 3D Scanning. *ACM Trans. Graph.* **2016**, *35*, 239. [CrossRef]
27. Hepp, B.; Dey, D.; Sinha, S.N.; Kapoor, A.; Joshi, N.; Hilliges, O. Learn-to-Score: Efficient 3D Scene Exploration by Predicting View Utility. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 437–452.
28. Meng, Z.; Qin, H.; Chen, Z.; Chen, X.; Sun, H.; Lin, F.; Ang, M.H., Jr. A Two-Stage Optimized Next-View Planning Framework for 3-D Unknown Environment Exploration, and Structural Reconstruction. *IEEE Robot. Autom. Lett.* **2017**, *2*, 1680–1687. [CrossRef]
29. Dunn, E.; Frahm, J.M. Next Best View Planning for Active Model Improvement. In Proceedings of the British Machine Vision Conference (BMVC), London, UK, 7–10 September 2009; pp. 1–11.
30. von Stumberg, L.; Usenko, V.; Engel, J.; Stückler, J.; Cremers, D. Autonomous Exploration with a Low-Cost Quadrocopter Using Semi-Dense Monocular SLAM. *arXiv* **2016**, arXiv:1609.07835.
31. Mendez, O.; Hadfield, S.; Pugeault, N.; Bowden, R. Taking the Scenic Route to 3D: Optimising Reconstruction from Moving Cameras. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; Volume 3.
32. Palazzolo, E.; Stachniss, C. Effective Exploration for MAVs Based on the Expected Information Gain. *Drones* **2018**, *2*, 9. [CrossRef]
33. Kumar Ramakrishnan, S.; Grauman, K. Sidekick Policy Learning for Active Visual Exploration. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 413–430.
34. Border, R.; Gammell, J.D.; Newman, P. Surface Edge Explorer (SEE): Planning Next Best Views Directly from 3D Observations. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 1–8.
35. Hoppe, C.; Wendel, A.; Zollmann, S.; Pirker, K.; Irschara, A.; Bischof, H.; Kluckner, S. Photogrammetric Camera Network Design for Micro Aerial Vehicles. In Proceedings of the Computer Vision Winter Workshop (CVWW), Hernstein, Austria, 4–6 February 2012; Volume 8, pp. 1–3.
36. Jing, W.; Polden, J.; Tao, P.Y.; Lin, W.; Shimada, K. View Planning for 3D Shape Reconstruction of Buildings with Unmanned Aerial Vehicles. In Proceedings of the IEEE International Conference on Control, Automation, Robotics and Vision (ICARCV), Phuket, Thailand, 13–15 November 2016; pp. 1–6.
37. Smith, N.; Moehrle, N.; Goesele, M.; Heidrich, W. Aerial Path Planning for Urban Scene Reconstruction: A Continuous Optimization Method and Benchmark. In Proceedings of the ACM SIGGRAPH Asia, Tokyo, Japan, 4–7 December 2018; p. 183.
38. Peng, C.; Isler, V. Adaptive View Planning for Aerial 3D Reconstruction of Complex Scenes. *arXiv* **2018**, arXiv:1805.00506.
39. Huang, R.; Zou, D.; Vaughan, R.; Tan, P. Active Image-based Modeling with a Toy Drone. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 1–8.
40. Alsadik, B.; Gerke, M.; Vosselman, G. Automated Camera Network Design for 3D Modeling of Cultural Heritage Objects. *J. Cult. Herit.* **2013**, *14*, 515–526. [CrossRef]
41. Bircher, A.; Kamel, M.; Alexis, K.; Burri, M.; Oettershagen, P.; Omari, S.; Mantel, T.; Siegwart, R. Three-dimensional Coverage Path Planning via Viewpoint Resampling and Tour Optimization for Aerial Robots. *Auton. Robot.* **2016**, *40*, 1059–1078. [CrossRef]
42. Snavely, N.; Seitz, S.M.; Szeliski, R. Skeletal Graphs for Efficient Structure from Motion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 23–28 June 2008; Volume 1, pp. 1–8.
43. Mostegel, C.; Rumpler, M.; Fraundorfer, F.; Bischof, H. UAV-based Autonomous Image Acquisition with Multi-view Stereo Quality Assurance by Confidence Prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-WS), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1–10.

44. Devrim Kaba, M.; Gokhan Uzunbas, M.; Nam Lim, S. A Reinforcement Learning Approach to the View Planning Problem. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6933–6941.
45. Marmanis, D.; Wegner, J.D.; Galliani, S.; Schindler, K.; Datcu, M.; Stilla, U. Semantic Segmentation of Aerial Images with an Ensemble of CNNs. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 473. [CrossRef]
46. Kaiser, P.; Wegner, J.D.; Lucchi, A.; Jaggi, M.; Hofmann, T.; Schindler, K. Learning Aerial Image Segmentation from Online Maps. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 6054–6068. [CrossRef]
47. Chen, K.; Fu, K.; Yan, M.; Gao, X.; Sun, X.; Wei, X. Semantic segmentation of aerial images with shuffling convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 173–177. [CrossRef]
48. Wendel, A.; Maurer, M.; Graber, G.; Pock, T.; Bischof, H. Dense Reconstruction on-the-fly. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 1450–1457.
49. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2015; pp. 3431–3440.
50. Semantic Drone Dataset. Available online: http://dronedataset.icg.tugraz.at (accessed on 28 May 2019).
51. ISPRS 2D Semantic Labelling Contest—Potsdam. Available online: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html (accessed on 28 May 2019).
52. Katz, S.; Tal, A.; Basri, R. Direct Visibility of Point Sets. *ACM Trans. Graph.* **2007**, *26*, 24. [CrossRef]
53. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*; Cambridge University Press: Cambridge, UK, 2003.
54. Luhmann, T.; Robson, S.; Kyle, S.; Boehm, J. *Close-Range Photogrammetry and 3D Imaging*; Walter de Gruyter: Berlin, Germany, 2013.
55. Förstner, W.; Wrobel, B.P. *Photogrammetric Computer Vision*; Springer: Berlin, Germany, 2016.
56. Wenzel, K.; Rothermel, M.; Fritsch, D.; Haala, N. Image acquisition and model selection for multi-view stereo. *ISPRS Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2013**, *40*, 251–258. [CrossRef]
57. Kraus, K. *Photogrammetry: Geometry from Images and Laser Scans*; Walter de Gruyter: Berlin, Germany, 2011.
58. Krause, A.; Golovin, D. *Submodular Function Maximization*; Cambridge University Press: Cambridge, UK, 2014.
59. Blender Online Community. *Blender—A 3D Modelling and Rendering Package*; Blender Foundation, Blender Institute: Amsterdam, The Netherlands, 2018
60. Knapitsch, A.; Park, J.; Zhou, Q.Y.; Koltun, V. Tanks and Temples: Benchmarking Large-scale Scene Reconstruction. *ACM Trans. Graph.* **2017**, *36*, 78. [CrossRef]