Leveraging Photogrammetric Mesh Models for Aerial-Ground Feature Point Matching Toward Integrated 3D Reconstruction

Qing Zhu[a], Zhendong Wang[a], Han Hu[a,*], Linfu Xie[b], Xuming Ge[a], Yeting Zhang[c]

[a] *Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China*
[b] *Guangdong Key Laboratory of Urban Informatics & Shenzhen Key Laboratory of Spatial Smart Sensing and Services & Research Institute for Smart Cities, School of Architecture and Urban Planning, Shenzhen University, Shenzhen, China*
[c] *State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China*

Abstract

Integration of aerial and ground images has been proved as an efficient approach to enhance the surface reconstruction in urban environments. However, as the first step, the feature point matching between aerial and ground images is remarkably difficult, due to the large differences in viewpoint and illumination conditions. Previous studies based on geometry-aware image rectification have alleviated this problem, but the performance and convenience of this strategy is limited by several flaws, *e.g.* quadratic image pairs, segregated extraction of descriptors and occlusions. To address these problems, we propose a novel approach: leveraging photogrammetric mesh models for aerial-ground image matching. The methods of this proposed approach have linear time complexity with regard to the number of images, can explicitly handle low overlap using multi-view images and can be directly injected into off-the-shelf structure-from-motion (SfM) and multi-view stereo (MVS) solutions. First, aerial and ground images are reconstructed separately and initially co-registered through weak georeferencing data. Second, aerial models are rendered to the initial ground views, in which the color, depth and normal images are obtained. Then, the synthesized color images and the corresponding ground images are matched by comparing the descriptors, filtered by local geometrical information, and then propagated to the aerial views using depth images and patch-based matching. Experimental evaluations using various datasets confirm the superior performance of the proposed methods in aerial-ground image matching. In addition, incorporation of the existing SfM and MVS solutions into these methods enables more complete and accurate models to be directly obtained.

Keywords: Aerial-ground Integration, Feature Matching, 3D Reconstruction, Multi-View Stereo, Structure-from-Motion

1. Introduction

Aerial oblique images, specifically those obtained from penta-view camera systems (Lemmens, 2014), have become a major source of data for city-scale urban reconstruction. However, occlusion and viewpoint differences greatly perturb aerial imagery of the bottom parts of buildings, leading to holes in geometry and texture-blurring effects (Wu et al., 2018). Recent studies (Nex et al.,

*Corresponding Author: han.hu@swjtu.edu.cn

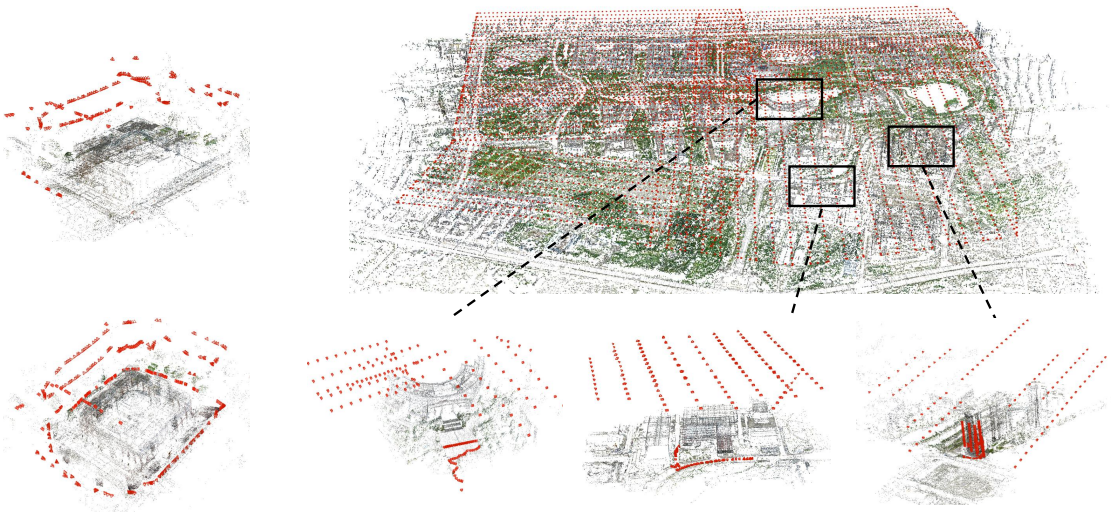


Figure 1: Aerial-ground reconstruction for the city hall of Dortmund, Germany (left) and three buildings of the Southwest Jiaotong University (SWJTU), Chengdu, China. The top row depicts the different structures of aerial image collections and the bottom row shows the reconstructed aerial and ground images. The images are rendered using Colmap (Schönberger and Frahm, 2016).

2015; Wu et al., 2018; Gao et al., 2018b) have confirmed that integration of aerial and ground images is a promising approach toward improved 3D reconstruction (see Figure 1).

The major obstacle to aerial-ground integration is the large viewpoint difference between the two sets of images, which means that it is impossible to find enough tie-points to conduct a successful bundle adjustment to integrate both datasets into the same coordinate frame. Scale invariant feature transform (SIFT) and SIFT-like features (Lowe, 2004; Arandjelović and Zisserman, 2012; Bursuc et al., 2015) are incapable of handling large perspective differences (Mikolajczyk et al., 2005), and learned features (Schonberger et al., 2017; Mishchuk et al., 2017) cannot greatly extend the classical approach (Arandjelović and Zisserman, 2012; Schonberger et al., 2017). Although some researchers have pioneered investigations in this area (Wu et al., 2018; Gao et al., 2018a,b), we argue that current approaches based on image warping (Hu et al., 2015) are not an ideal solution, and that some key problems remain unsolved.

1) Quadratically increased image rectifications. Warping all of the images to the ground (Hu et al., 2015) is a valid solution for the nadir and oblique views of aerial images, and the feature extraction has an $O(n)$ complexity with respect to the number of images. However, the *view-independent* ground structure is not applicable to the aerial-ground integration (Wu et al., 2018). In addition, the *view-dependent* homographic relationships and image rectifications are defined pairwise with respect to a virtual façade, leading to a feature extraction of $O(n^2)$, which is prohibitively high in practice. Further, such façade structures may be untenable.

2) Occlusion between pairwise aerial and ground image. Even if the aerial and ground images are rectified successfully, feature matching between the two images remains a non-trivial task. For instance, the overlapping region may be only a small part of the whole image, and this region may still be affected by occlusion, as seen in the work by Wu et al. (2018).

3) Mode of the data acquisition. An effective strategy to avoid the problem of aerial-ground feature matching is to systematically design the image acquisition for both datasets, i.e., to collect images with acceptable convergent angles around the objects of interest (such as shown in the Dortmund dataset (Nex et al., 2015) in Figure 1 left). However, in practice, regular

strip-flights are preferred even for regional applications, such as the campus dataset in Figure 1, and terrestrial data are captured to improve the quality of images of certain objects of interest. Therefore, it is inevitable that there will be perspective differences between aerial and ground images.

In this paper we leverage the photogrammetric meshes obtained from aerial images to solve the above problems. Accordingly, instead of rectifying the images pairwise, we directly render the textured meshes onto a virtual camera determined by the ground images. The rendered images also consist of depth values and normal vectors, and act as proxies between the ground and aerial images: *i.e.*, feature matches are conducted between the ground images and rendered images, and these are then propagated to the aerial images using depth and normal information via multi-photo geometrically constrained (MPGC) matching (Zhang, 2005) or patch-based matching propagation (Furukawa and Ponce, 2009). A single rendered image contains textural information from multiple aerial images, which are typically selected meticulously in the multi-view stereo (MVS) pipeline (Vu et al., 2011; Waechter et al., 2014); therefore, the proposed methods are explicitly occlusion-aware. Additional features are detected only from the rendered images and descriptor searching is conducted only on the pair of corresponding rendered and ground images; therefore, both feature extraction and feature matching have the complexity of $O(n)$ with respect to the number of ground images. To handle the illumination differences that lead to worsening descriptor performances, we add an additional filter prior to processing by a random sample consensus (RANSAC) model (Moisan et al., 2012) using a greedy approach.

In summary, our main contribution is a simple, fast, accurate and robust approach that solves the problem of aerial-ground feature point matching by performing off-screen rendering of the textured mesh models. The reminder of this paper is organized as follows. In Section 2 we briefly describe feature point matching between aerial and ground images. In Section 3 we elaborate on the two steps of the proposed methods, *i.e.* rendering and matching. Experimental evaluations for both the ISPRS datasets (Nex et al., 2015) and SWJTU datasets are demonstrated (Figure 1) in Section 4. Finally, concluding remarks are given.

2. Related works

Here, we review only directly relevant studies on feature point matching methods in the context of large perspective differences. Specifically, three major strategies for image matching are considered, namely: 1) affine invariant features; 2) image rectification; and 3) 3D rendering. More detailed reviews and comparisons can be found in recent benchmark works (Schonberger et al., 2017).

1) Affine invariant features. Following the route of scale and rotation invariant SIFT features (Lowe, 2004), earlier researchers sought affine invariant regions for use in matching of image data comprising large viewpoint differences, which are generally represented as eclipses on the image (Mikolajczyk and Schmid, 2004; Matas et al., 2004; Ma et al., 2015). These affine invariant regions may also be detected by line structures (Chen and Shao, 2013). However, in practice, affine invariant detectors are more sensitive to image noise and their repeatability is inferior to that of difference of Gaussian (DoG) detectors (Lowe, 2004) and other corner detectors (Rublee et al., 2011; Rosten et al., 2010). Therefore, the overall performances of affine invariant detectors are generally worse than those based on SIFT-like features.

2) Image rectification. When no *a priori* geometry information is available, affine SIFT (ASIFT) (Morel and Yu, 2009) can be used to create a database of descriptors by synthesizing the image in a series of pre-defined affine transformations that cover a reasonable range of angles. A

similar approach is used in the database BRIEF (Calonder et al., 2012). However, ASIFT will significantly increase the number of features and therefore increase the search space, leading to longer runtimes and lower recall rate.

In most cases, we have access to the initial image poses, either from the global navigation satellite system (GNSS) and inertial measurements unit (IMU) information or from coarse registrations. This *a priori* geometry information can help us to determine the extent of the perspective deformation, and thus to rectify the images accordingly. For structured aerial oblique cameras, we can identify a *view-independent* structure for the rectification, *i.e.* the ground, and thereby the perspective deformation between the nadir and oblique views can be substantially alleviated by projecting all of the images to the ground through homographic transformations Hu et al. (2015). This strategy is also applicable to rectification of unmanned aerial vehicle (UAV images) (Jiang and Jiang, 2017) and of panoramas captured by mobile mapping systems (Jende et al., 2018; Javanmardi et al., 2017).

View-independent rectifications are convenient, as their feature extractions and feature matchings have the same time complexities, *e.g.*, $O(n)$, with respect to the original number of images. However, independent structures are not always available for rectification. Therefore, *view-dependent* rectifications have been proposed to remedy this problem. Wu et al. (2018) found virtual façade structures by fitting planes from the points inside the frustum of the camera, and rectified the resulting images by projecting both the aerial and ground images onto the planes. In addition, 3D structures can be rather more versatile than planar-only structures; as such, triangular meshes and even sparse point clouds also support the rectification. Gao et al. (2018a,b) projected ground images onto aerial views, in which the deformations were rectified using the triangular meshes. A similar strategy was also implemented using dense point clouds (Shan et al., 2014), by formulating a depth map corresponding to the ground image and warping the image to the aerial view in a pixelwise fashion.

However, *view-dependent* rectification also implies that at least the descriptor extraction must be conducted on the rectified images (which has quadratic time complexity), and also requires computation of the pairwise image rectifications. Such an onerous process is acceptable only for correlation-based feature matching in local windows, which is used to refine the position of known tie-points or expand these to neighboring regions, such as is required in a multi-photo geometrically constrained (MPGC) correlation (Zhang, 2005) or in patch-based dense image matching (Furukawa and Ponce, 2009; Wu et al., 2018).

3) 3D rendering. The above matching methods only use data from a pair of images, regardless of the methods used for image rectification. In the case of aerial-ground integration, the overlapping data may be quite sparse, limiting the recall rate of the descriptor searching. As an alternative, rendering 3D data onto the target view can explicitly utilize information from multiple images and also exploit the massively parallel power of the graphics computing unit (GPU) for efficient implementation. In this context, Untzelmann et al. (2013) rendered the sparse point clouds from SIFT matches using the splat representation (Sibbing et al., 2013) to generate synthesized views as proxies for wide baseline feature-matching. However, the sparsity of colored point clouds means that these are not ideal data sources for such rendering. However, recent solutions (Acute3D, 2019; Agisoft, 2019; Schönberger et al., 2016) can generate high resolution textured mesh models, which can be used as better proxies. In addition, this paper shows that the process depth and normal vectors of the meshes are preserved during rendering, which further supports the use of correlation-based local refinement (Zhang, 2005; Furukawa and Ponce, 2009) to correct textural defects.

3. Aerial-ground feature point matching by leveraging photogrammetric models

3.1. Overview of the approach

Integrated reconstruction from both aerial and ground images relies on the premise that the intrinsic and extrinsic orientation parameters are consistent in the same coordinate frame, which is achieved by a combined bundle adjustment of all of the images. The foundation of a successful bundle adjustment is the accurate and robust matching of tie-points, which necessitates solving the problem of large perspective deformation. In previous work (Wu et al., 2018; Gao et al., 2018b), view-dependent image rectifications have partially alleviated this problem, as these can obtain a few aerial-ground tie-points for the estimation of the rigid transformation. However, as the tie-points are too sparse, a general-purpose SfM cannot be used to solve the problem: the rigid transformation must be either directly used to merge the aerial and ground datasets at the image level (Gao et al., 2018b) or used as an *a priori* constraint in the bundle adjustment (Wu et al., 2018). Unfortunately, this *ad hoc* solution not only breaks the existing SfM and MVS pipeline, but also performs worse, due to the robustness of the estimation of the rigid transformation from sparse aerial-ground connections.

In this paper we surmount the problem of view-dependent rectification by using textured meshes; that is, we render textured meshes to ground images, and use these rendered images as delegates to establish feature matching of aerial and ground images. Figure 2 demonstrates the overall workflow of the proposed methods. Beginning with two separate datasets, we first reconstruct the sparse models via the existing SfM pipeline, and then conduct coarse registration to transform the ground models into the coordinate frame of the aerial datasets via methods similar to those used in previous works (Wu et al., 2018; Gao et al., 2018b), by either weak GNSS information or three interactively selected points. As our approach requires no planar structures (Wu et al., 2018), only an MVS pipeline is used for the aerial datasets required for tile-wise reconstruction of mesh models. The textured meshes are rendered using the camera defined by the ground images, and three rendered images are obtained, *i.e.*, color, depth and normal vectors. The synthesized color images are matched with the ground information, and are then propagated to the aerial views using the depth information. The insufficient geometric accuracy of the meshes and the blending problem of the texture (Waechter et al., 2014) in the MVS pipeline necessitate the use of the depth and normal vectors to determine a local patch from which to refine the feature matches from the synthesized images. Finally, the matches are directly injected into off-the-shelf SfM and MVS pipelines.

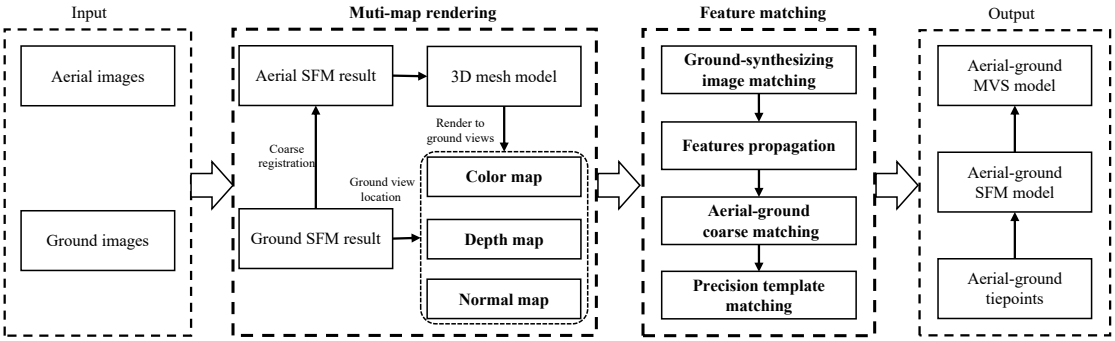


Figure 2: Workflow of the proposed methods.

3.2. View synthesizing the ground images by rendering of meshes

3.2.1. Definition of the camera models

To exploit the OpenGL graphics pipeline for the synthesis of the ground images from the textural information of the aerial meshes, the notation of the intrinsic and extrinsic orientation parameters from the SfM and camera matrices of the graphics pipeline must be converted.

Specifically, for the camera models, we use the protocol of BlockExchange (Acute3D, 2019), in which a 3D point $\boldsymbol{X}$ is projected to the corresponding image coordinate $\boldsymbol{x}$ as,

$$\boldsymbol{x} = fD(\Pi(\mathbf{R}(\boldsymbol{X} - \boldsymbol{C}))) + \boldsymbol{x}_0, \tag{1}$$

where f and $\boldsymbol{x}_0$ are the principal distance and principal point measured in pixels, respectively; $D(\cdot)$ is the distortion mapping from an undistorted focal plane coordinate to the distorted position and the Brown model with five parameters $(k_1, k_2, k_3, p_1, p_2)$ is considered; $\Pi(\cdot) : \mathbb{R}^3 \mapsto \mathbb{R}^2$ is the projection function needed to obtain the homogeneous normalized coordinate; and $\mathbf{R}$ and $\boldsymbol{C}$ denote the extrinsic orientation for the rotation matrix and projection center, respectively. In addition, each image is enriched by three depth values recorded in the BlockExchange format, in terms of the nearest z_n, furthest z_f and median z_m depth; even without these values, it is trivial to estimate the depth information from the sparse point clouds or the bounding box of the region of interest.

3.2.2. Estimation of the rendering matrices for the view synthesis

In the graphics pipeline, the homogeneous coordinate $\tilde{\boldsymbol{X}} \in \mathbb{R}^4$ of the 3D point $\boldsymbol{X}$ is projected to the normalized screen space $\boldsymbol{m} \in \mathbb{R}^3$ (and the homogeneous coordinate $\tilde{\boldsymbol{m}} \in \mathbb{R}^4$) using the view $\mathbf{V} \in \mathbb{R}^{4\times4}$ and projection $\mathbf{P} \in \mathbb{R}^{4\times4}$ matrices as below,

$$\tilde{\boldsymbol{m}} = \mathbf{P}\mathbf{V}\tilde{\boldsymbol{X}}. \tag{2}$$

The view matrix $\mathbf{V}$ is defined with three parameters, i.e., eye $\boldsymbol{E}$, center $\boldsymbol{O}$ and up $\boldsymbol{U}$, with its physical significance determined using the *lookat* routine (GLM, 2019), which describes the position and orientation of the camera. The projection matrix $\mathbf{P}$ is defined by the *perspective* routine (GLM, 2019) using the field of view θ, aspect ratio ρ, nearest z_n and furthest z_f depths data, which describes the frustum of the camera. Although it is possible to consider the principal point offsets and distortion of the camera in the graphics pipeline by exploiting the program shaders, as the induced deformation is almost negligible, we (for simplicity) ignore the principal point and distortion during conversion.

To obtain the eye $\boldsymbol{E}$, center $\boldsymbol{O}$ and up vector $\boldsymbol{U}$ for the *lookat* function, the conversion is determined intuitively as:

$$\begin{aligned} \boldsymbol{E} &= \boldsymbol{C} \\ \boldsymbol{O} &= \boldsymbol{C} + z_m \mathbf{R}^T \boldsymbol{e}_z, \\ \boldsymbol{U} &= -\mathbf{R}^T \boldsymbol{e}_y \end{aligned} \tag{3}$$

where $\boldsymbol{e}$ denotes the unit vector along the corresponding axis and $\mathbf{R}^T$ transforms the axis in camera coordinate space to object coordinate space. With respect to the parameters in the *perspective* function, z_n and z_f are directly used for the depth range and the other two parameters are calculated as:

$$\begin{aligned} \theta &= 2 \arctan \frac{h}{2f}, \\ \rho &= \frac{w}{h} \end{aligned} \tag{4}$$

where w and h are the width and height of the images, respectively.

3.2.3. Rendering of the color, depth and normal images

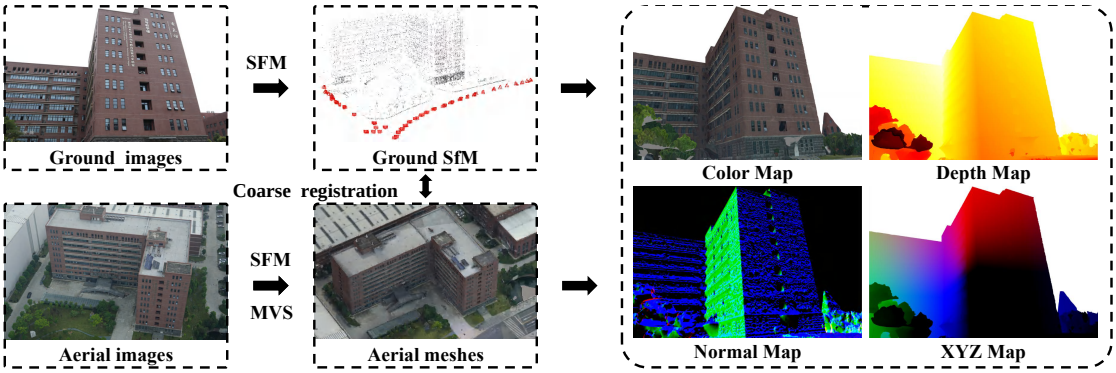


Figure 3: Illustration of the rendering of the meshes to various maps, comprising color images, depth images and normal images. The coordinates of each pixel in the rendered image can be obtained.

Another practical requirement for the rendering of the textured meshes is that the meshes are tiled on a tree structure, e.g., a quad-tree, octree or an adaptive KD-tree, and dynamically fed into the main memory for rendering. We use the OpenSceneGraph (Osfield and Burns, 2014) for model-paging and force-loading the best model in the frustum of the corresponding camera before rendering. For the rendering, we allocate three frame-buffer objects to store the color, depth and normal information (Figure 3), and the meshes are then directly rendered to the buffers rather than to the physical screen. The sizes of the frame-buffer objects are the same as those of the corresponding cameras, therefore reducing the differences scale and other geometric aspects.

Notably, the rendering of the meshes explicitly utilizes the massively parallel GPU and can be achieved almost in real time. In addition, use of the depth information means that any point in the color image is one-on-one mapped to 3D object space (*i.e.* the XYZ map in Figure 3); thus, by enriching a point with a normal vector, we obtain a locally oriented patch; this is similar to the concept used in previous work by Furukawa and Ponce (2009). This locally oriented patch is helpful for the process of multi-photo geometrically constrained matching needed to refine the tie-points.

3.3. Feature matching and refinement with the synthesized images

Figure 4 illustrates the two steps of the aerial-ground feature-point matching process. For coarse matching, we extract SIFT features (Lowe, 2004) on the synthesized images, and compare descriptors using the ratio check and filter outliers, using both the proposed geometrical constraints (subsection 3.3.1) and RANSAC (Fischler and Bolles, 1981). Specifically, we use a recent variant of RANSAC, the *a contrario* RANSAC (AC-RANSAC), due to its automatic threshold-tuning capability (Moisan et al., 2012). If the remaining number of pairwise matches between the synthesized and ground images is less than five, we consider the matching to be not stable and ignore the results for this pair.

For refined matching, the matches are propagated to aerial views (subsection 3.3.2) and refined using the original images (subsection 3.3.3). For each match, a template image I_g on the ground images is extracted, the size of which is determined by a correlation window w_c. Then, the coordinate $\boldsymbol{X}$ in 3D space is calculated from the corresponding depth value on the synthesized images using the reverse of Equation 2, using the depth value d in the object space and the corresponding ground sample distance $\delta = \frac{d}{f}$. We assign a relatively large search window

$w_s\delta$ centered on and tangential to the oriented points $(\boldsymbol{X}, \boldsymbol{n})$, which serve as delegates for the subsequent matching refinement. In the following section, we use the term $p = (\boldsymbol{X}, \boldsymbol{n}, w_s\delta)$ to denote the oriented patches in the object space, inspired by previous work (Furukawa and Ponce, 2009).

3.3.1. Local geometry constraints for outlier removal

Due to illumination differences between synthesized and ground images, the SIFT match may contain significantly more outliers after ratio checking, which leads to inferior RANSAC performance. However, because the geometrical differences between the ground and synthesized images are almost negligible, the discrepancies between the coordinates of correct matches should be small and follow consistent patterns in local regions. Based on these insights, we propose a greedy search algorithm to remove outliers prior to the RANSAC. Specifically, from a pair of matched images $p(x_p, y_p)$ and $q(x_q, y_q)$, a directed vector can be obtained as $m = p - q$. If the initial coarse registration is correct, $m = \mathbf{0}$ should be satisfied. However, due to alignment errors and uncompensated distortion, the discrepancies m should be consistent with the following three constraints (Figure 5), which are used sequentially to filter outliers.

1) Length constraint. The length of the discrepancy vector $|m|$ is constrained by an upper limit τ_l, *i.e.* $|m| < \tau_l$. In practice, τ_l is chosen as 2% of the image extent.

2) Intersection constraint. First, we sort the matches by the discrepancy lengths of $|m|$ ascendingly. Then, we determine if each segment has an intersection with the K-nearest ($K = 5$) segments. If an intersection exists, the longest segment is marked as an outlier.

3) Direction constraint. First, we calculate the dominant direction for each segment with respect to the K-nearest $K = 5$ segments. Then, we remove segments that deviate from the dominant direction by an angle τ_a ($\tau_a = 90°$ is used).

3.3.2. Propagation of the matches to the aerial images

As the meshes are produced from the aerial images, the local patches p should be consistent with all of the aerial images. To propagate the local patch to aerial images, three criteria are considered: (1) *Containment*, meaning that the local patch should be inside the frustum of the aerial images; (2) *Consistency*, meaning that the orientation of the patch and the aerial image should be consistent, *i.e.* less than a threshold $\tau_n = 90°$; and (3) *Visibility*, meaning that the patch should not be occluded by the mesh itself. For occlusion detection, we used the optimized bounding volume hierarchy (BVH) of the triangular meshes implemented in Embree (Wald et al., 2014) for ray tracing. As these BVH structures have almost linear space complexity with regard to their number of triangles, we establish and cache the BVH structure in advance using the meshes that have the finest level of detail.

3.3.3. Matching refinement between aerial and ground images

Although the meshes used for rendering are obtained from the aerial images, the matches propagated to the aerial images may be inaccurate. The geometry of the meshes is noise-laden and the textural information is a blend of multiple images, as shown in Figure 6. Therefore, the coordinates on the ground images and projected aerial images must be further refined.

Inspired by the MPGC approach (Zhang, 2005) and our previous view-independent synthesis (Hu et al., 2015), we also project all of the patches to the same camera space using the homographic transformation $\mathbf{H}$ (Hartley and Zisserman, 2003) to induce the oriented patch p and the corresponding ground images:

$$\mathbf{H} = \mathbf{K}_g(\mathbf{R} + \boldsymbol{t}\boldsymbol{n}_d^T)\mathbf{K}_a^{-1}, \tag{5}$$

where $\mathbf{K}$ is the camera matrix; $\mathbf{R}$ and $\boldsymbol{t}$ are the relative orientation and translation parameters between the two images; $\boldsymbol{n}_d = \frac{\boldsymbol{n}}{d}$ is the scaled normal vector; and the subscripts g and a denote

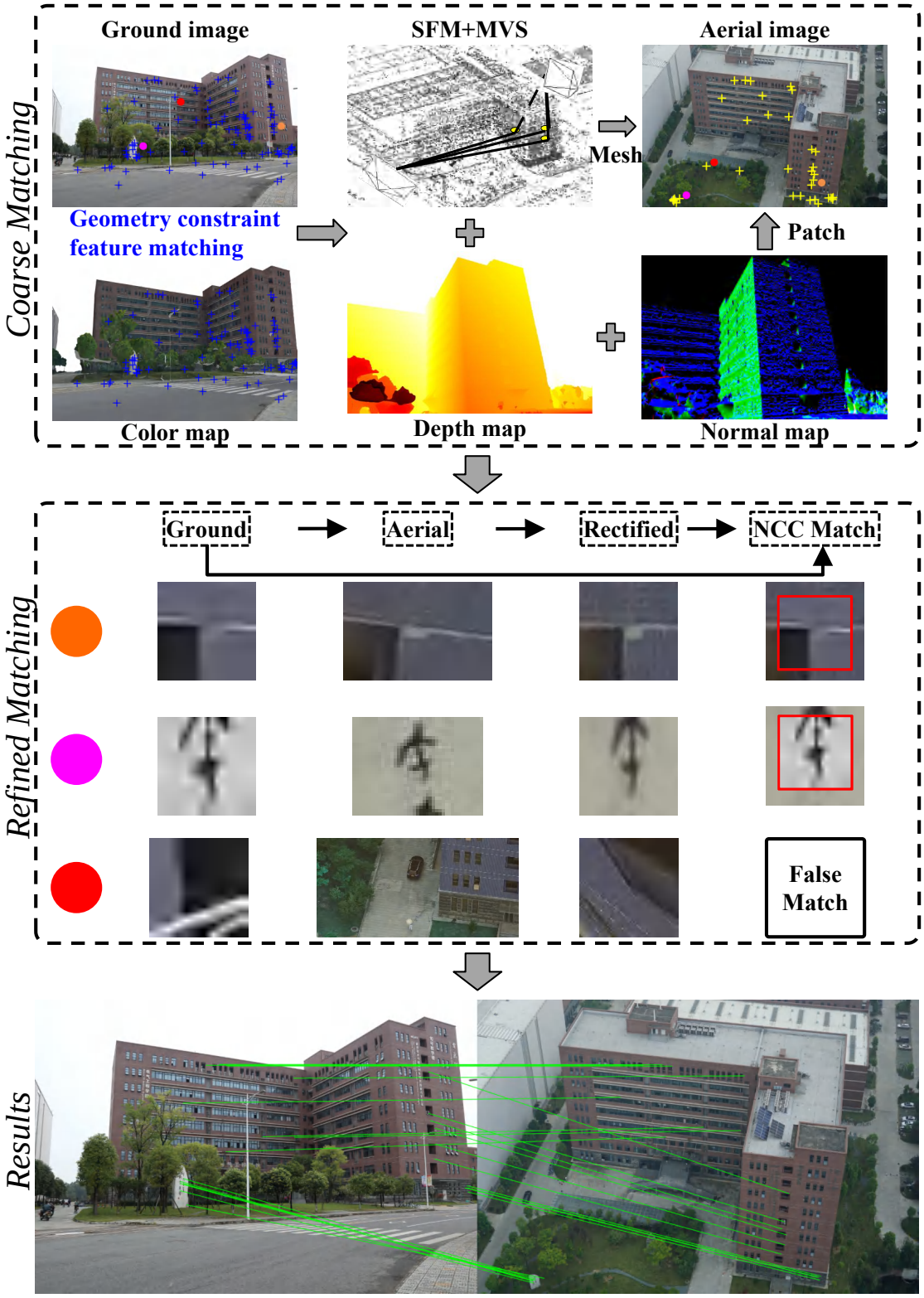
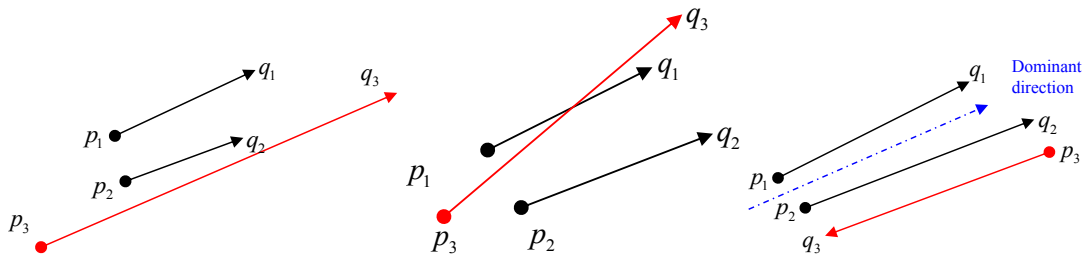


Figure 4: Overview of aerial-ground feature matching. The circles in the coarse-matching images denote the three patches in the refined matching.



(a) Length constraint (b) Intersection constraint (c) Direction constraint

Figure 5: Constraints for outlier filtering in the matching of ground and synthesized images. The points p and q denote the key points in the synthesized and ground images, respectively. Note that p is placed on the ground image. The red lines indicate matches that violate the constraints.

the ground and aerial images, respectively. Notably, unlike in our previous work (Hu et al., 2015), only the local patches surrounding the initial position are loaded and transformed, rather than the entire images.

After rectifying all of the patches, a classic normalized correlation coefficient (NCC) is used to find the initial match, followed by a least-squares matching (Gruen, 1985; Hu et al., 2016) step to enable sub-pixel accuracy. The patch extracted from the ground image serves as the template for matching and all of the aerial images are aligned pairwise. Any match with a correlation smaller than a threshold τ_c ($\tau_c = 0.75$ is used) is pruned before the least-squares matching step, and, after sub-pixel localization, reverse homographic transformation in Equation 2 is used to obtain the final coordinates on the aerial images.

In previous studies (Wu et al., 2018; Gao et al., 2018b) that obtained only a few tie-points connecting the aerial and ground images, a general-purpose SfM pipeline was not capable of robustly registering both datasets. This is due to the fact that the sparse and loose tie-points may not be reliable enough to formulate a strong connection to trigger the incremental SfM pipeline (Snavely et al., 2006; Schönberger and Frahm, 2016). Therefore, previous approaches depended on an *ad hoc* SfM pipeline that generally reduced to a rigid transformation (Gao et al., 2018b) or at least prioritized a rigid transformation (Wu et al., 2018). Our proposed synthesis and matching strategy can obtain enough long-track reliable feature matches, and we have found that the existing SfM and MVS pipeline (Acute3D, 2019) already suffices for the integrated reconstruction.

4. Experimental evaluations

4.1. Dataset descriptions

Four datasets (see Table 1 and Figure 1) are used to evaluate the proposed methods, which comprise an International Society for Photogrammetry and Remote Sensing (ISPRS) benchmark dataset collected at Dortmund (Nex et al., 2015) and three datasets collected at the campus of SWJTU. Qualitative and quantitative feature point matching experiments are conducted and compared with existing state-of-the-art solutions. In addition, to further verify the capability of the proposed method, 3D reconstruction results are also presented and compared.

The ISPRS dataset was collected at the center of Dortmund, with both the aerial and ground images surrounding the building being captured using the same sensor. In addition, the ground control points (GCPs) for accuracy validation were measured only from the ground images. The other three datasets were all collected in the campus of SWJTU, specifically at the library

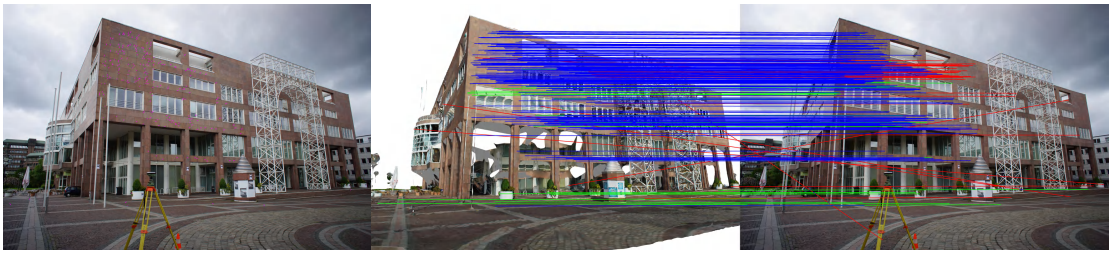
Figure 6: Aspects of the synthesized images that will cause non-negligible errors for aerial-ground matches.

(SWJTU-LIB), a general building (SWJTU-BLD) and the residential areas (SWJTU-RES). Unlike the Dortmund dataset, the SWJTU aerial images were collected in regular stripped flights and the ground images were captured only for areas of interest. SWJTU-RES ground images were also captured by a low-cost UAV in a vertical uplift flight, which took a sequence of images of the façade.

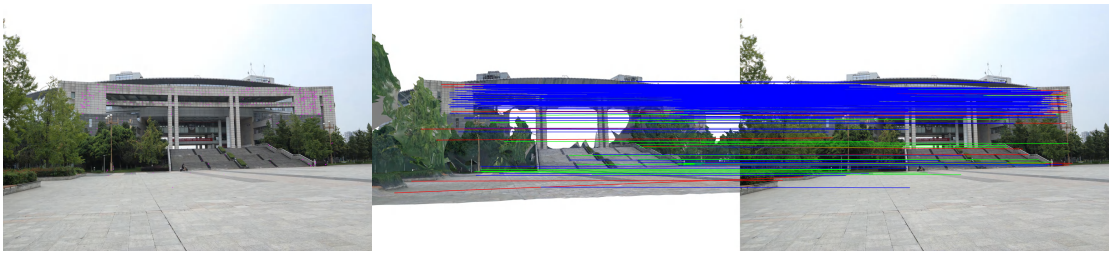
4.2. Evaluation of feature matching

4.2.1. Evaluation of feature matching between ground and synthesized images

To evaluate the performances of the proposed geometric constraints in the matching of synthesized and ground images, we compare feature matches with and without the additional outlier filter. Figure 7 shows the matching results for the four datasets. The blue, green and red lines indicate matches retrieved by both methods, by only the method with geometry constraints, and by only the method without geometry constraints, respectively. Notably, even for Dortmund and SWJTU-LIB, for which RANSAC succeeds in finding a correct model, there are more outliers remaining, as shown by the red lines in Figure 7a and b. Furthermore, RANSAC fails in the third and fourth dataset, either due to no model being found or too many outliers being present. Table 2 also confirms this finding.

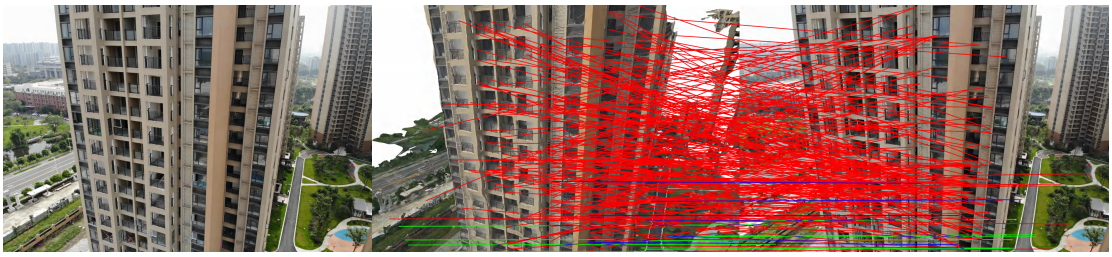


(a) Image DSC03002 in Dortmund dataset



(b) Image IMG1701 in SWJTU-LIB dataset

(c) Image IMG1899 in SWJTU-BLD dataset



(d) Image DGI0172 in SWJTU-RES dataset

Figure 7: Comparisons with and without the proposed geometry constraints in matching between ground and synthesized images. The left column of each subfigure is the ground image, for reference. Blue, green and red lines indicate the matches detected by both methods, i.e., matches detected only with geometry constraints and matches detected only without geometry constraints, respectively. It is notable that the application of some methods that lack geometry constraints leads to many more outliers and fewer correct matches.

Table 1: Detailed descriptions of the four datasets used for evaluations.

Dataset	Sensor		Resolution (cm)		#Images	
	Aerial	Ground	Aerial	Ground	Aerial	Ground
Dortmund	Sony Nex-7	SONY Nex-7	1.10	0.53	146	204
SWJTU-LIB	Sony ICLE-5100	Canon EOS M6	1.69	1.06	123	78
SWJTU-BLD	Sony ICLE-5100	Canon EOS M6	1.93	1.33	207	88
SWJTU-RES	Sony ICLE-5100	DJI Spark	1.97	2.56	92	192

Table 2: Comparison of the outlier filter with and without the proposed geometric constraints in the matching between ground and synthesized images. The number of SIFT matches is that subsequent to a ratio check. In fact, classical RANSAC fails in the SWJTU-BLD and SWJTU-RES datasets.

Dataset	Image	#SIFT	#RANSAC	#Proposed
Dortmund	DSC03002	2020	245	239
SWJTU-LIB	IMG1701	2057	344	356
SWJTU-BLD	IMG1899	2415	0	39
SWJTU-RES	DGI0172	2146	244	14

4.2.2. Evaluation of feature-matching between aerial and ground images

As the final results of feature matching are influenced by feature detection, descriptors and outlier removal, we directly compare the proposed method with complete solutions rather than with the feature descriptors themselves, such as SIFT (Lowe, 2004) and AKAZE (Alcantarilla and Solutions, 2011). Four solutions are considered, ours and one commercial solution, *i.e.* Agisoft MetaShape (Agisoft, 2019), and two freeware solutions, *i.e.* VisualSFM (Wu et al., 2011) and Colmap (Schönberger and Frahm, 2016; Schönberger et al., 2016). Eight pairs are randomly selected from the four datasets, with two pairs for each dataset. As it is possible that the matching results are noise-laden, we manually count the number of correct matches for the eight pairs. Figure 8 summarizes the results. Notably, the remaining three solutions often fail in these situations. Thus, although these solutions are quite robust for processing normal scenes or even Internet-scale datasets (Schönberger and Frahm, 2016; Wu et al., 2011), the large perspective deformation between aerial and ground images means that these are challenging targets for processing.

We also select one pair from each dataset and compare the matching results visually against the results afforded by the second-best processing system, VisualSFM, in Figures 9 to 12. As can be seen, we successfully obtain enough correct matches for all of the pairs, in comparison to the only two correct matches that are obtained by VisualSFM. These enlarged regions demonstrate the accuracies of the proposed methods even when applied to these challenging scenes. In fact, the number of original matches from the synthesized and ground images is even higher, as some of these are filtered out in the propagation step.

4.3. Evaluations of the integrated reconstruction

We develop an add-on solution to the integrated reconstruction using the feature matches, based on ContextCapture (Acute3D, 2019) and by fusing the feature matches and separated aerial and ground images into the same block. In addition, we also compare three other solutions: the vanilla ContextCapture (Acute3D, 2019), MetaShape (Agisoft, 2019) and Colmap (Schönberger and Frahm, 2016). Both sparse and dense reconstructions are evaluated in the following section.

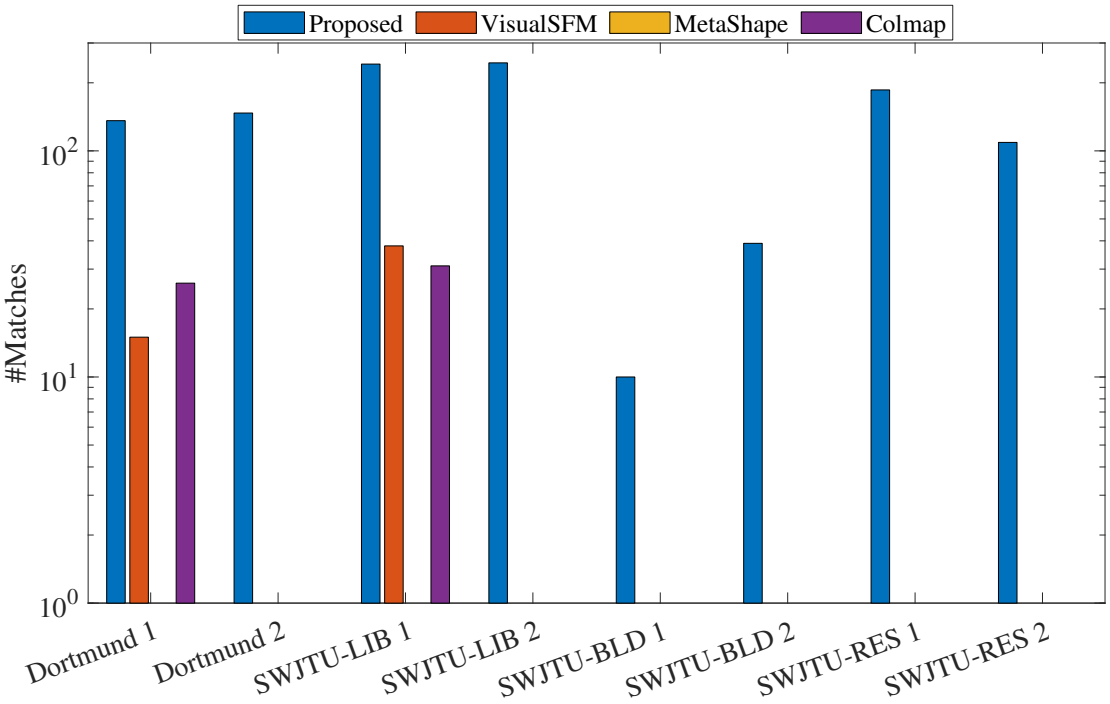


Figure 8: Comparison of the number of aerial-ground feature matches on randomly selected sets of eight pairs.

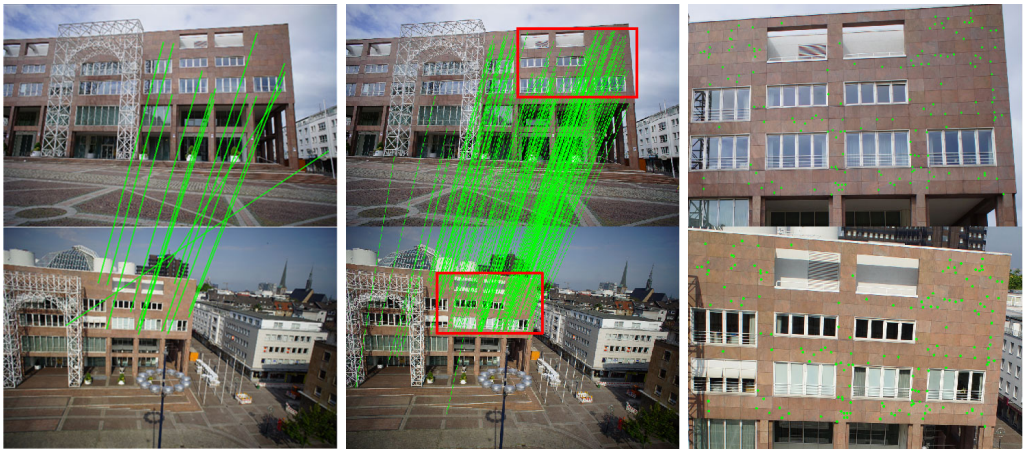
4.3.1. Evaluation of integrated sparse reconstruction

First, we demonstrate the SfM results by comparing the final numbers of reconstructed images. As some solutions can automatically separate the images into several clusters, only the largest cluster is considered. In addition, we report the number of tie-points that connect aerial and ground images, as these points are the most crucial aspects for the integrated reconstruction. Table 3 shows the status of the SfM performances, and it can be seen that the proposed solutions succeed in each case, while the SfM pipeline in vanilla ContextCapture fails in each case. MetaShape and Colmap occasionally succeed for the Dortmund and SWJTU-LIB datasets, however, these methods generate far fewer aerial-ground tie-points than the proposed solution and therefore their performance is arguably less robust.

To further evaluate the precision and accuracy of the proposed methods, the position uncertainties from the aerial triangulation report and the root-mean-square error (RMSE) of the checkpoints are used. The former metric denotes the internal stability of the SfM results and the latter denotes performance against those of external control networks. Table 4 summarizes the results. As different datasets have different accuracies, we also report the results generated using only aerial images as a baseline. Notably, the integrated reconstruction achieves satisfactory precision and accuracy, which are similar to those obtained using only aerial images are used.

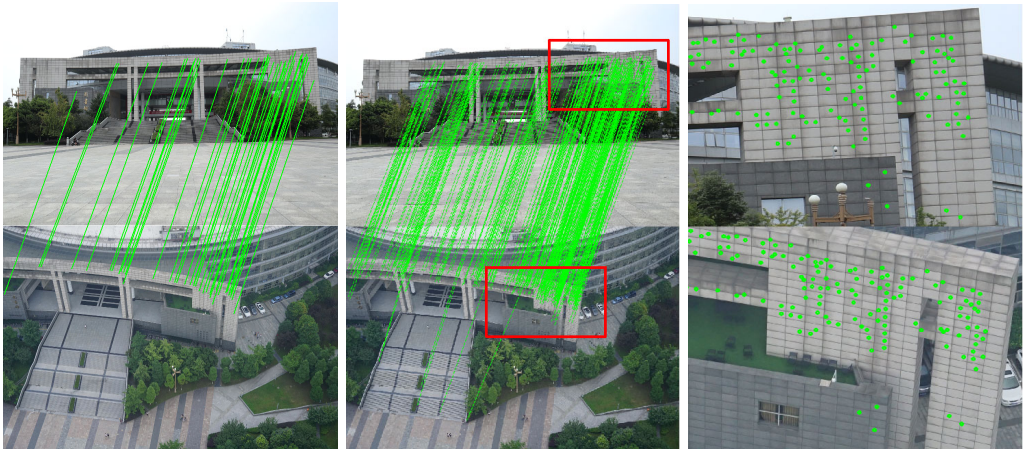
4.3.2. Evaluation of integrated dense reconstruction

Figure 13 compares the textured mesh models between aerial-only images and the aerial-ground images. The top row of each subfigure is obtained using only aerial images and the bottom row is obtained using all of the images. We also highlight some parts of the models on the right of each subfigure. Using the integrated solution, the textures on the façades are clearer,



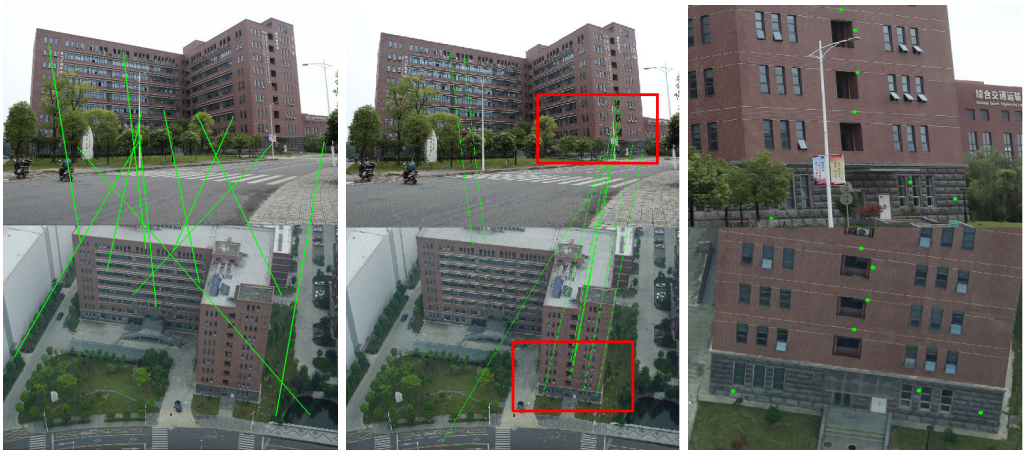
(a) VisualSFM (b) Proposed (c) Enlarged

Figure 9: Aerial-ground matching results for the DSC02315-DSC07055 pair from the Dortmund dataset. The red rectangles denote the enlarged areas.



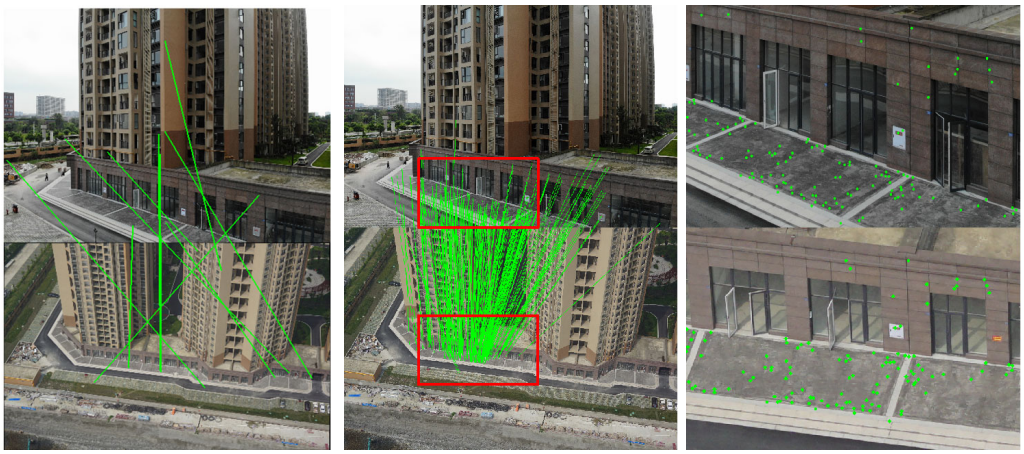
(a) VisualSFM (b) Proposed (c) Enlarged

Figure 10: Aerial-ground matching results for the IMG711-W0760 pair from the SWJTU-LIB dataset. The red rectangles denote the enlarged areas.



(a) VisualSFM (b) Proposed (c) Enlarged

Figure 11: Aerial-ground matching results for the IMG1901-X0651 pair from the SWJTU-BLD dataset. The red rectangles denote the enlarged areas.



(a) VisualSFM (b) Proposed (c) Enlarged

Figure 12: Aerial-ground matching results for the DJI0117-W0405 pair of the SWJTU-RES dataset. The red rectangles denote the enlarged areas.

Table 3: Comparisons of different solutions for the four datasets on the sparse reconstruction. The number of reconstructed images and the number of aerial-ground tie-points are presented.

Dataset	Method	#Reconstructed Images		#Aerial-ground tiepoints	Status
		Ground	Aerial		
Dortmund	Proposed+ContextCapture	203/204	146/146	23648	Succeeded
	ContextCapture	204/204	0/146	0	Failed
	MetaShape	203/204	146/146	1237	Succeeded
	Colmap	168/204	0/146	0	Failed
SWJTU-LIB	Proposed+ContextCapture	78/78	123/123	11399	Succeeded
	ContextCapture	0/78	123/123	0	Failed
	MetaShape	78/78	123/123	3740	Succeeded
	Colmap	78/78	123/123	1374	Succeeded
SWJTU-BLD	Proposed+ContextCapture	88/88	207/207	1706	Succeeded
	ContextCapture	0/88	0/206	0	Failed
	MetaShape	0/88	0/206	0	Failed
	Colmap	38/88	0/206	0	Failed
SWJTU-RES	Proposed+ContextCapture	192/192	88/92	779	Succeeded
	ContextCapture	192/192	0/92	0	Failed
	MetaShape	192/192	91/92	0	Failed
	Colmap	0/192	16/92	0	Failed

Table 4: Precision and accuracy of the integrated sparse reconstruction. For reference, we also report the results generated using only the aerial images as baseline. The checkpoints are not captured by the aerial images for the Dortmund dataset.

Dataset	Aerial Only (cm)						Aerial-ground Integration (cm)					
	Position Uncertainty			Checkpoint RMSE			Position Uncertainty			Checkpoint RMSE		
	x	y	z	x	y	z	x	y	z	x	y	z
Dortmund	0.10	0.10	0.10				0.07	0.07	0.07	2.6	2.0	2.2
SWJTU-LIB	0.53	0.46	0.58	1.0	1.1	32.1	0.32	0.30	0.32	2.4	3.3	15.5
SWJTU-BLD	0.58	0.56	0.45	1.6	1.0	4.9	0.71	0.77	0.59	3.4	9.9	12.1
SWJTU-RES	3.59	7.89	7.26	4.7	0.9	12.7	2.65	1.06	3.33	2.7	0.7	14.5

as shown in Figure 13a, c and d. In addition, the geometries are obviously better, as can be seen by detailing of the holes in Figure 13b and the small objects in Figure 13c.

4.4. Discussions and limitations

Based on the above evaluations for feature matching and integrated reconstruction, we next summarize some characteristics and limitations of the proposed methods.

1) Integration with existing SfM and MVS pipeline. Although previous solutions (Wu et al., 2018; Gao et al., 2018b) can satisfactorily incorporate aerial and ground images into the same framework, the tie-points connecting the two sets of images are too sparse, and break the existing SfM pipeline. As most state-of-the-art MVS approaches (Furukawa and Ponce, 2009; Vu et al., 2011) are also based on the use of sparse reconstruction for robustness, these tie-points are also important for subsequent dense reconstruction. Our proposed method effectively retrieves sufficient connections between aerial and ground images, and can be directly used as add-ons to existing solutions.

2) Efficiency and accuracy. Our proposed pipeline is also fast and accurate. We do not need to exhaust pairs between aerial and ground images, as feature matching is only required between ground and synthesized images, and is propagated to the aerial views. This is important,

(a) Dortmund (b) SWJTU-LIB

(c) SWJTU-BLD (d) SWJTU-RES

Figure 13: Comparison of the textured mesh models generated from only aerial images (top row), and those generated from aerial-ground images (bottom row). The right column of each subfigure is an enlargement of the regions highlighted by the rectangles.

because if large viewpoint differences exist, we cannot rely on descriptor-based image retrieval to reduce the numbers of matching pairs. In addition, the propagation can also achieve sub-pixel performance.

3) Limitations. A limitation of our proposed approach is the same as in a previous approach (Wu et al., 2018; Gao et al., 2018b), namely that dense reconstruction is required prior to the SfM pipeline. In addition, our method also requires textural information. Although only regions of interest need to be retouched (Acute3D, 2019) in our proposed pipeline and the runtime overhead may be ignored, the quality of the textured mesh models will unavoidably influence the performance of our approach.

5. Conclusion

In this paper, we address the problem of feature matching between aerial and ground images, which currently suffers from severe perspective deformation resulting from viewpoint differences. We elegantly solve the problem by leveraging textured mesh models, which are rendered to the virtual cameras of the ground images. In addition, robust geometric constraints and patch-based matching refinement are used to improve the robustness and quality of the matches. The abundance of tie-points obtained by our proposed methods mean that we can directly utilize existing SfM and MVS pipelines for improved surface reconstruction. Future works may be devoted to further exploiting the possibility of integrating light detection and ranging (LiDAR) point clouds and panoramas collected by the ground mobile-mapping systems into aerial datasets. Code and the SWJTU datasets will be made publicly available at https://github.com/saedrna/RenderMatch.

Acknowledgments

The authors gratefully acknowledge the provision of the datasets by ISPRS and EuroSDR, which were released in conjunction with the ISPRS Scientific Initiatives 2014 and 2015, led by ISPRS ICWG I/Vb. In addition, this work was supported by the National Natural Science Foundation of China (Projects No.: 41631174, 41871291, 41871314).

Reference

References

Acute3D, 2019. Context capture. https://www.acute3d.com/.

Agisoft, 2019. Agisoft metashape. https://www.agisoft.com/.

Alcantarilla, P.F., Solutions, T., 2011. Fast explicit diffusion for accelerated features in nonlinear scale spaces. IEEE Trans. Patt. Anal. Mach. Intell 34, 1281–1298.

Arandjelović, R., Zisserman, A., 2012. Three things everyone should know to improve object retrieval, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 2911–2918.

Bursuc, A., Tolias, G., Jégou, H., 2015. Kernel local descriptors with implicit rotation matching, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM. pp. 595–598.

Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., Fua, P., 2012. Brief: Computing a local binary descriptor very fast. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 1281–1298.

Chen, M., Shao, Z., 2013. Robust affine-invariant line matching for high resolution remote sensing images. Photogrammetric Engineering and Remote Sensing 79, 753–760.

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24, 381–395.

Furukawa, Y., Ponce, J., 2009. Accurate, dense, and robust multiview stereopsis. IEEE transactions on pattern analysis and machine intelligence 32, 1362–1376.

Gao, X., Hu, L., Cui, H., Shen, S., Hu, Z., 2018a. Accurate and efficient ground-to-aerial model alignment. Pattern Recognition 76, 288–302.

Gao, X., Shen, S., Zhou, Y., Cui, H., Zhu, L., Hu, Z., 2018b. Ancient chinese architecture 3d preservation by merging ground and aerial point clouds. ISPRS Journal of Photogrammetry and Remote Sensing doi:https://doi.org/10.1016/j.isprsjprs.2018.04.023.

GLM, 2019. Opengl mathematics. https://glm.g-truc.net/.

Gruen, A., 1985. Adaptive least squares correlation: a powerful image matching technique. South African Journal of Photogrammetry, Remote Sensing and Cartography 14, 175–187.

Hartley, R., Zisserman, A., 2003. Multiple view geometry in computer vision. Cambridge university press.

Hu, H., Ding, Y., Zhu, Q., Wu, B., Xie, L., Chen, M., 2016. Stable least-squares matching for oblique images using bound constrained optimization and a robust loss function. ISPRS journal of photogrammetry and remote sensing 118, 53–67.

Hu, H., Zhu, Q., Du, Z., Zhang, Y., Ding, Y., 2015. Reliable spatial relationship constrained feature point matching of oblique aerial images. Photogrammetric Engineering & Remote Sensing 81, 49–58.

Javanmardi, M., Javanmardi, E., Gu, Y., Kamijo, S., 2017. Towards high-definition 3d urban mapping: Road feature-based registration of mobile mapping systems and aerial imagery. Remote Sensing 9, 975.

Jende, P., Nex, F., Gerke, M., Vosselman, G., 2018. A fully automatic approach to register mobile mapping and airborne imagery to support the correction of platform trajectories in gnss-denied urban areas. ISPRS Journal of Photogrammetry and Remote Sensing 141, 86–99. doi:https://doi.org/10.1016/j.isprsjprs.2018.04.017.

Jiang, S., Jiang, W., 2017. On-board gnss/imu assisted feature extraction and matching for oblique uav images. Remote Sensing 9, 813.

Lemmens, M., 2014. Oblique imagery: the standard for mapping. GIM International 28, 14–17.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110.

Ma, X., Liu, D., Zhang, J., Xin, J., 2015. A fast affine-invariant features for image stitching under large viewpoint changes. Neurocomputing 151, 1430–1438. doi:10.1016/j.neucom.2014.10.045.

Matas, J., Chum, O., Urban, M., Pajdla, T., 2004. Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing 22, 761–767.

Mikolajczyk, K., Schmid, C., 2004. Scale and affine invariant interest point detectors. International Journal of Computer Vision 60, 63–86.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L., 2005. A comparison of affine region detectors. International Journal of Computer Vision 65, 43–72.

Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J., 2017. Working hard to know your neighbor's margins: Local descriptor learning loss, in: Advances in Neural Information Processing Systems, pp. 4826–4837.

Moisan, L., Moulon, P., Monasse, P., 2012. Automatic homographic registration of a pair of images, with a contrario elimination of outliers. Image Processing On Line 2, 56–73.

Morel, J.M., Yu, G., 2009. Asift: A new framework for fully affine invariant image comparison. SIAM Journal on Imaging Sciences 2, 438–469.

Nex, F., Remondino, F., Gerke, M., Przybilla, H.J., Bäumker, M., Zurhorst, A., 2015. Isprs benchmark for multi-platform photogrammetry. ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences 2.

Osfield, R., Burns, D., 2014. Open scene graph. http://www.openscenegraph.org.

Rosten, E., Porter, R., Drummond, T., 2010. Faster and better: A machine learning approach to corner detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 105–119.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G.R., 2011. Orb: An efficient alternative to sift or surf., in: ICCV, Citeseer. p. 2.

Schönberger, J.L., Frahm, J.M., 2016. Structure-from-motion revisited, in: Conference on Computer Vision and Pattern Recognition (CVPR).

Schonberger, J.L., Hardmeier, H., Sattler, T., Pollefeys, M., 2017. Comparative evaluation of hand-crafted and learned local features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1482–1491.

Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M., 2016. Pixelwise view selection for unstructured multi-view stereo, in: European Conference on Computer Vision (ECCV).

Shan, Q., Wu, C., Curless, B., Furukawa, Y., Hernandez, C., Seitz, S.M., 2014. Accurate geo-registration by ground-to-aerial image matching, in: 2014 2nd International Conference on 3D Vision, IEEE. pp. 525–532.

Sibbing, D., Sattler, T., Leibe, B., Kobbelt, L., 2013. Sift-realistic rendering, in: 2013 International Conference on 3D Vision-3DV 2013, IEEE. pp. 56–63.

Snavely, N., Seitz, S.M., Szeliski, R., 2006. Photo tourism: exploring photo collections in 3d, in: ACM transactions on graphics (TOG), ACM. pp. 835–846.

Untzelmann, O., Sattler, T., Middelberg, S., Kobbelt, L., 2013. A scalable collaborative online system for city reconstruction, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 644–651.

Vu, H.H., Labatut, P., Pons, J.P., Keriven, R., 2011. High accuracy and visibility-consistent dense multiview stereo. IEEE transactions on pattern analysis and machine intelligence 34, 889–901.

Waechter, M., Moehrle, N., Goesele, M., 2014. Let there be color! large-scale texturing of 3d reconstructions, in: European Conference on Computer Vision, Springer, Zurich, Switzerland. pp. 836–850.

Wald, I., Woop, S., Benthin, C., Johnson, G.S., Ernst, M., 2014. Embree: a kernel framework for efficient cpu ray tracing. ACM Transactions on Graphics (TOG) 33, 143.

Wu, B., Xie, L., Hu, H., Zhu, Q., Yau, E., 2018. Integration of aerial oblique imagery and terrestrial imagery for optimized 3d modeling in urban areas. ISPRS Journal of Photogrammetry and Remote Sensing 139, 119–132. doi:https://doi.org/10.1016/j.isprsjprs.2018.03.004.

Wu, C., et al., 2011. Visualsfm: A visual structure from motion system .

Zhang, L., 2005. Automatic digital surface model (DSM) generation from linear array images. ETH Zurich.