

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

Qing Zhu<sup>a</sup>, Zhendong Wang<sup>a</sup>, Han Hu<sup>a,\*</sup>, Linfu Xie<sup>b</sup>, Xuming Ge<sup>a</sup>, Yeting Zhang<sup>c</sup>

<sup>a</sup>Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China

<sup>b</sup>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

<sup>c</sup>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 are still 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 have linear time complexity with regard to the number of images. It explicitly handles low overlap using multi-view images. The proposed methods 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 color, depth and normal images are obtained. Then, feature matching between synthesized and ground images are conducted through descriptor searching and geometry-constrained outlier removal. Finally, oriented 3D patches are formulated using the synthesized depth and normal images and the correspondences are propagated to the aerial views through patch-based matching. **Experimental evaluations using five datasets reveal satisfactory performance of the proposed methods in aerial-ground image matching, which succeeds in all of the ten challenging pairs compared to only three for the second best. In addition, incorporation of existing SFM and MVS solutions enables more complete reconstruction results, with better internal stability.**

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

---

## <sup>1</sup> 1. Introduction

<sup>2</sup> Penta-view aerial oblique images (Lemmens, 2014) have become a major source of data for  
<sup>3</sup> city-scale urban reconstruction. However, occlusion and viewpoint differences greatly perturb  
<sup>4</sup> the bottom parts of buildings, leading to holes in geometry and texture-blurring effects (Wu

---

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

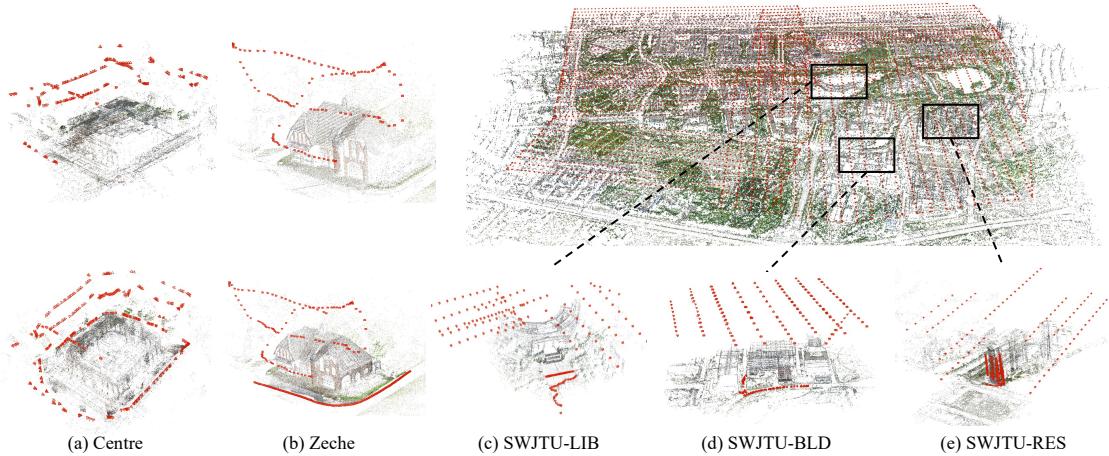


Figure 1: Aerial-ground reconstruction for the ISPRS benchmark (Nex et al., 2015) and three buildings of the Southwest Jiaotong University (SWJTU), Chengdu, China. The top row depicts the different structures of aerial image collections and the bottom row shows the reconstructed aerial and ground images. The images are rendered using Colmap (Schönberger and Frahm, 2016).

et al., 2018). Recent studies (Nex et al., 2015; Wu et al., 2018; Gao et al., 2018) 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. It is difficult to find enough tie-points to register both datasets into the same coordinate frame in a combined bundle adjustment. 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 (Mishchuk et al., 2017; Revaud et al., 2019; Dusmanu et al., 2019) 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., 2018), we argue that some key problems remain unfulfilled.

1) *Quadratically increased image rectifications.* Warping all of the images to 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 ground structure is not applicable in aerial-ground integration. Pairwise rectification is used to remedy this problem (Wu et al., 2018), by the adoption of virtual façades. But pairwise rectification leads to a feature extraction of  $O(n^2)$ , which is prohibitively high in practice. Furthermore, such façade structures may be untenable in certain scenarios.

2) *Problem of pairwise rectification.* Even if the aerial and ground images are rectified successfully, feature matching between them still remains a non-trivial task. For pairwise rectification, contents from only two images are involved, which will lead to some problems in feature matching. 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 (Molina et al., 2017). For instance, collecting images with acceptable convergent angles around the objects of interest is tenable for certain objects, such as the Centre and Zeche datasets (Nex et al., 2015) in Figure 1. However, in practice, flights with regular strips are preferred even for regional

34 applications, such as the campus of SWJTU in Figure 1. Terrestrial images are only captured to  
35 improve the quality of objects of interest. Therefore, the perspective deformation between aerial  
36 and ground images is inevitable.

37 In this paper, we leverage the photogrammetric meshes obtained from aerial images to solve  
38 the above problems. Accordingly, instead of rectifying the images pairwisely, we directly render  
39 the textured meshes onto a virtual camera determined by the ground images. The rendered  
40 images also consist of depth values and normal vectors, and act as proxies between the ground  
41 and aerial images. Feature matches are conducted between the ground and rendered images.  
42 The correspondences are then enriched with depth and normal information, which can formulate  
43 3D patches in the object space. The 3D patches are then propagated to the aerial images via  
44 multi-photo geometrically constrained (MPGC) matching (Zhang, 2005) or patch-based match-  
45 ing (Furukawa and Ponce, 2009). A single rendered image contains textural information from  
46 multiple aerial images, which are typically selected meticulously in the multi-view stereo (MVS)  
47 pipeline (Vu et al., 2011; Waechter et al., 2014); therefore, the proposed methods are explicitly  
48 occlusion-aware. Additional features are detected only from the rendered images and descriptor  
49 matchings are conducted only on the pairs of rendered and ground images; therefore, both fea-  
50 ture extraction and feature matching have time complexity of  $O(n)$ , with respect to the number  
51 of ground images. To handle the illumination differences that lead to degraded descriptor per-  
52 formances, we add an additional filter prior to random sample consensus (RANSAC) (Moisan  
53 et al., 2012) using geometry constraints.

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

## 61 2. Related works

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

67 1) *Affine invariant features.* Following the route of scale and rotation invariant SIFT features  
68 (Lowe, 2004), earlier researchers sought affine invariant regions to alleviate perspective deforma-  
69 tions. Affine invariant features are generally represented as ellipses on the image (Mikolajczyk  
70 and Schmid, 2004; Matas et al., 2004; Ma et al., 2015). These affine invariant regions may also be  
71 detected by line structures (Chen and Shao, 2013). However, in practice, affine invariant detec-  
72 tors are more sensitive to image noise and their repeatability is inferior to that of the difference  
73 of Gaussian (DoG) detectors (Lowe, 2004) or other corner detectors (Rublee et al., 2011; Rosten  
74 et al., 2010). Therefore, the overall performances of affine invariant detectors are generally worse  
75 than those based on SIFT-like features (Lowe, 2004).

76 2) *Image rectification.* When no *a priori* geometry information is available, affine SIFT (ASIFT)  
77 (Morel and Yu, 2009) can be used to create a database of descriptors by synthesizing the image in  
78 a series of pre-defined affine transformations. A similar approach is used in the database BRIEF

(Calonder et al., 2012), which retrieves BRIEF features on multiple scales and orientations. Roth et al. (2017) also synthesized a series of views using pairwise perspective transformation and the features are detected using similar sampling strategies as ASIFT (Morel and Yu, 2009). 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 of standard photogrammetric applications, we have access to the initial image poses, from either the global navigation satellite system (GNSS) or from coarse registrations (Wu et al., 2018; Gao et al., 2018). The *a priori* geometry information can help us to rectify the images. For aerial oblique images obtained with regular flight strips, we can identify a *view-independent* structure for the rectification, *i.e.* the ground. For *view-independent* rectification, the base plane for all the images is the same and the perspective deformation between the nadir and oblique views can be alleviated by projecting all the images onto the base plane (Hu et al., 2015). This strategy is also applicable to unmanned aerial vehicle (UAV) images (Jiang and Jiang, 2017) or panoramas captured by mobile mapping systems (Jende et al., 2018; Javanmardi et al., 2017).

*View-independent* rectifications (Hu et al., 2015; Jiang and Jiang, 2017) are convenient, as feature extractions and matchings have the same time complexity  $O(n)$ , with respect to the original number of images. However, it is not always possible to find a suitable base plane that all the images can be projected to. Therefore, *view-dependent* rectifications (Wu et al., 2018; Gao et al., 2018) have been proposed to remedy this problem, for which the surface for rectification is determined pair-wisely rather than unified for all the images. Wu et al. (2018) found virtual façade structures by fitting planes from the points inside the frustum of camera, and rectified images by projecting both the aerial and ground images onto the façade planes. The façade structures are also used by Fanta-Jende et al. (2019) for the co-registration of mobile mapping images and aerial oblique images. In addition, 3D structures can also be considered for pairwise rectification. Gao et al. (2018) projected ground images onto aerial views, using the triangular meshes as proxies. 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 aerial view in a pixelwise fashion.

However, *view-dependent* rectification also implies that the descriptor must be extracted 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 rather than the whole image. For instance, previous works have rectified local patches to refine the positions of known tie-points or expand them to neighboring regions, such as *e.g.* multi-photo geometrically constrained (MPGC) correlation (Zhang, 2005) and patch-based multi-view stereo (PMVS) (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 region of two images may be quite narrow, 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 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; Gao et al., 2018). However, the sparse point clouds from SFM are not ideal sources for such rendering.

Recent solutions (Acute3D, 2019; Agisoft, 2019; Schönberger et al., 2016) can generate high resolution textured mesh models, which can be used as better proxies for the feature matching. And learned MVS approaches (Yu and Gao, 2020; Yao et al., 2019) have demonstrated impressive performances on benchmark tests, which are promising alternatives for off-the-shelf MVS solu-

127 tions. Except for rendered color images, this paper shows that depth and normal information of  
128 the meshes can also be preserved during rendering, which further supports the correlation-based  
129 local refinement of matches (Zhang, 2005; Furukawa and Ponce, 2009).

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

131 *3.1. Overview of the approach*

132 Integrated reconstruction from both aerial and ground images relies on the premise that  
133 the intrinsic and extrinsic orientation parameters are consistent in the same coordinate frame,  
134 which is achieved by a combined bundle adjustment. The foundation of a successful bundle  
135 adjustment is accurate and robust matching of tie-points, which faces the problem of large  
136 perspective deformation between aerial and ground images. In previous works (Wu et al., 2018;  
137 Gao et al., 2018), pairwise image rectifications have partially alleviated this problem, for the  
138 estimation of rigid transformations. However, due to the amount and quality of inter-platform  
139 tiepoints, previous works need *ad hoc* strategies in the SfM and MVS pipeline. For instance, Gao  
140 et al. (2018) degraded SfM to a rigid transformation and simplified the MVS as fusion of point  
141 clouds from different platforms. Wu et al. (2018) co-registered images from different platforms  
142 by weighted bundle adjustment with parameters regularized by the rigid transformation and also  
143 only fused point clouds without a full MVS pipeline. In fact, the key problem still remained to  
144 be fulfilled, *i.e.* finding enough inter-platform tiepoints for both the SfM and MVS pipelines.

145 In this paper we surmount the problem of view-dependent rectification using textured meshes.  
146 We render textured meshes to ground images, and use these rendered images as delegates to  
147 establish feature matching between aerial and ground images. Figure 2 demonstrates the overall  
148 workflow of the proposed methods. Beginning with two separate datasets, we first reconstruct the  
149 sparse models via existing SfM pipeline. Coarse registration is conducted to fuse both aerial and  
150 ground models into the same coordinate frame, similar to previous works (Wu et al., 2018; Gao  
151 et al., 2018); the coarse registration can be achieved by either weak GNSS information or three  
152 interactively selected points. As our approach requires no planar structures (Wu et al., 2018),  
153 dense reconstruction using existing MVS pipeline is only required for the aerial datasets, from  
154 which tile-wise models are obtained. The textured meshes are rendered using the camera defined  
155 by the ground images; the results consist of color, depth and normal vectors. The synthesized  
156 color images are matched with the ground images, and correspondences are then propagated  
157 to the aerial views using the depth information. Due to insufficient geometric accuracy of the  
158 meshes and blending problems of the texture (Waechter et al., 2014) in the MVS pipeline, the  
159 correspondences have to be refined on the original aerial images. The refinement is achieved  
160 through the 3D local patches determined by the depth and normal vectors of the synthesized  
161 images. Finally, the matches are directly injected into off-the-shelf SfM and MVS pipelines for  
162 integrated reconstruction.

163 *3.2. View synthesizing the ground images by rendering of meshes*

164 *3.2.1. Definition of the camera models*

165 To exploit OpenGL graphics pipeline for the synthesis of ground images from textural in-  
166 formation of aerial meshes, the notations of intrinsic and extrinsic orientation parameters from  
167 SfM and camera matrices of graphics pipeline must be converted.

168 Specifically, for camera model, we use the protocol of BlockExchange (Bentley, 2019), in  
169 which a 3D point  $\mathbf{X}$  is projected to image  $\mathbf{x}$  as,

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

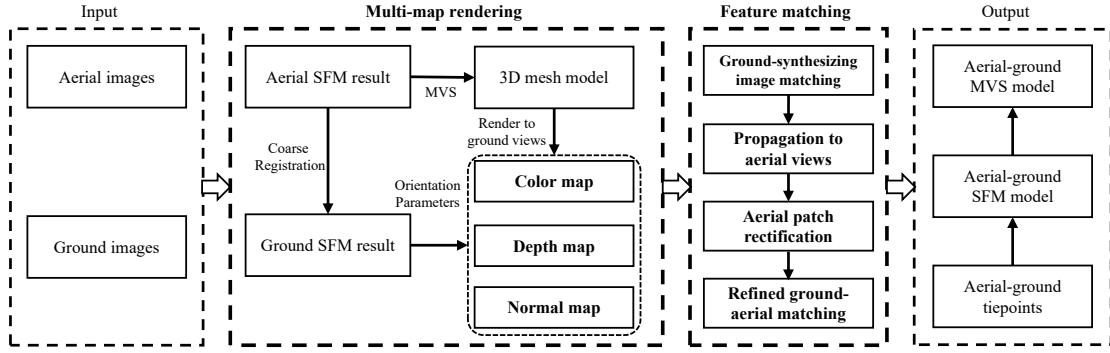


Figure 2: Workflow of the proposed method.

170 where  $f$  and  $\mathbf{x}_0$  are the principal distance and principal point measured in pixels, respectively;  
 171  $D(\cdot)$  is the distortion mapping from an undistorted focal plane coordinate to the distorted position  
 172 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  
 173 the projection function mapping the 3D point in camera space to the homogeneous normalized  
 174 coordinate; and  $\mathbf{R}$  and  $\mathbf{C}$  denote the extrinsic orientation for the rotation matrix and projection  
 175 center, respectively. In addition, each image is enriched by three depth values recorded in the  
 176 BlockExchange format, in terms of the nearest  $z_n$ , furthest  $z_f$  and median  $z_m$  depth; even without  
 177 these values, it is trivial to estimate the depth information from the sparse point clouds or the  
 178 bounding box of the region of interest.

179 *3.2.2. Estimation of the rendering matrices for the view synthesis*

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

$$\tilde{\mathbf{m}} = \mathbf{P} \mathbf{V} \tilde{\mathbf{X}}, \quad (2)$$

183 where the view matrix  $\mathbf{V}$  is defined with three parameters, *i.e.* eye  $\mathbf{E}$ , center  $\mathbf{O}$  and up  $\mathbf{U}$   
 184 vectors. The matrix is generally implemented in the *lookat* routine (GLM, 2019), which describes  
 185 the position and orientation of the camera. The projection matrix  $\mathbf{P}$  is defined by the *perspective*  
 186 routine (GLM, 2019) using the field of view  $\theta$ , aspect ratio  $\rho$ , nearest  $z_n$  and furthest  $z_f$  depth  
 187 values, which describes the frustum of the camera. Although it is possible to consider the  
 188 principal point offsets and distortion of the camera in the graphics pipeline by exploiting the  
 189 program shaders, we ignore them for two reasons: (1) the influences of them on perspective  
 190 deformation are almost negligible and (2) they only influence the intermediate coordinates on  
 191 the synthesized images, which will be eventually propagated to aerial views and refined.

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

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

194 where  $\mathbf{e}$  denotes the unit vector along the corresponding axis and  $\mathbf{R}^T$  transforms the axis in  
 195 camera coordinate space to object coordinate space. With respect to the parameters in the

196 perspective function,  $z_n$  and  $z_f$  are directly used for the depth range and the other two parameters  
 197 are calculated as:

$$\theta = 2 \arctan \frac{h}{2f},$$

$$\rho = \frac{w}{h}$$
(4)

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

### 199 3.2.3. Rendering of the color, depth and normal images

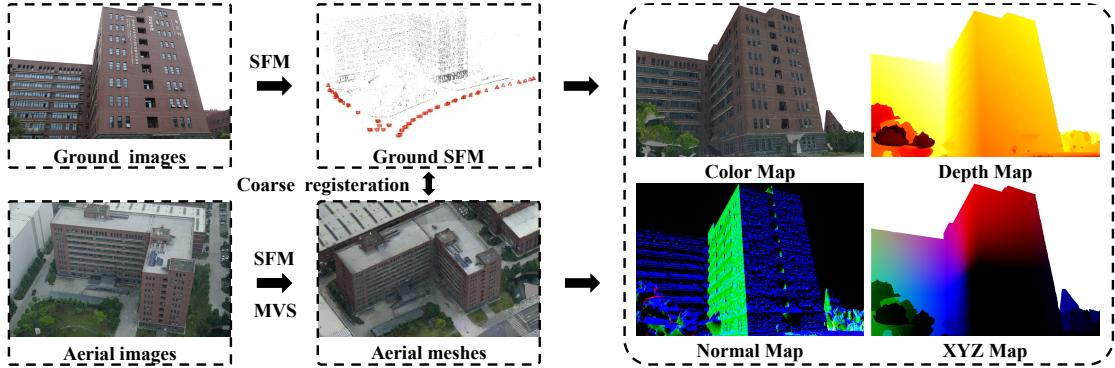


Figure 3: Illustration of the rendering of the meshes to various maps, comprising color images, depth images and normal images. The coordinates of each pixel in the rendered image can be obtained as the XYZ map.

200 Another practical issue for the rendering of the textured meshes is that the meshes are tiled on  
 201 a tree structure, e.g. quad-tree, octree or adaptive KD-tree. Even inside a single tile, the models  
 202 are still segmented into small pieces with different level-of-details to accelerate the loading of  
 203 files from disks. The render engine should use a scene graph to organize the dynamic loading (or  
 204 unloading) of the meshes that are inside (or outside, respectively) the frustum of current view.  
 205 This is non-trivial in implementation, but fortunately, OpenSceneGraph (Osfield and Burns,  
 206 2014) has already implemented an optimized database manager with its native data format. For  
 207 each frame, we wait for the database manager to fully load the load the finest level of detail of  
 208 model in the current view, before actually rendering the models. For the rendering, we allocate  
 209 three frame-buffer objects to store the color, depth and normal information (Figure 3), and the  
 210 meshes are then directly rendered to the buffers rather than to the physical screen. The sizes of  
 211 the frame-buffer objects are the same as those of the corresponding cameras, therefore reducing  
 212 the differences of scale and other geometric factors.

213 Notably, the rendering of the meshes explicitly utilizes the massively parallel GPU and can  
 214 be achieved almost in real time. In addition, any point in the color image is one-on-one mapped  
 215 to the 3D object space with the depth map (XYZ map in Figure 3). Therefore, by enriching  
 216 a point with a normal vector, we can obtain a locally oriented 3D patch; this is similar to the  
 217 concept of previous work (Furukawa and Ponce, 2009). The patch is helpful for the refinement  
 218 of correspondences between aerial and ground images.

### 219 3.3. Feature matching and refinement with the synthesized images

220 Figure 4 illustrates the two steps of the aerial-ground feature-point matching. For coarse  
 221 matching, we first extract SIFT features (Lowe, 2004) on the synthesized images, because SIFT is

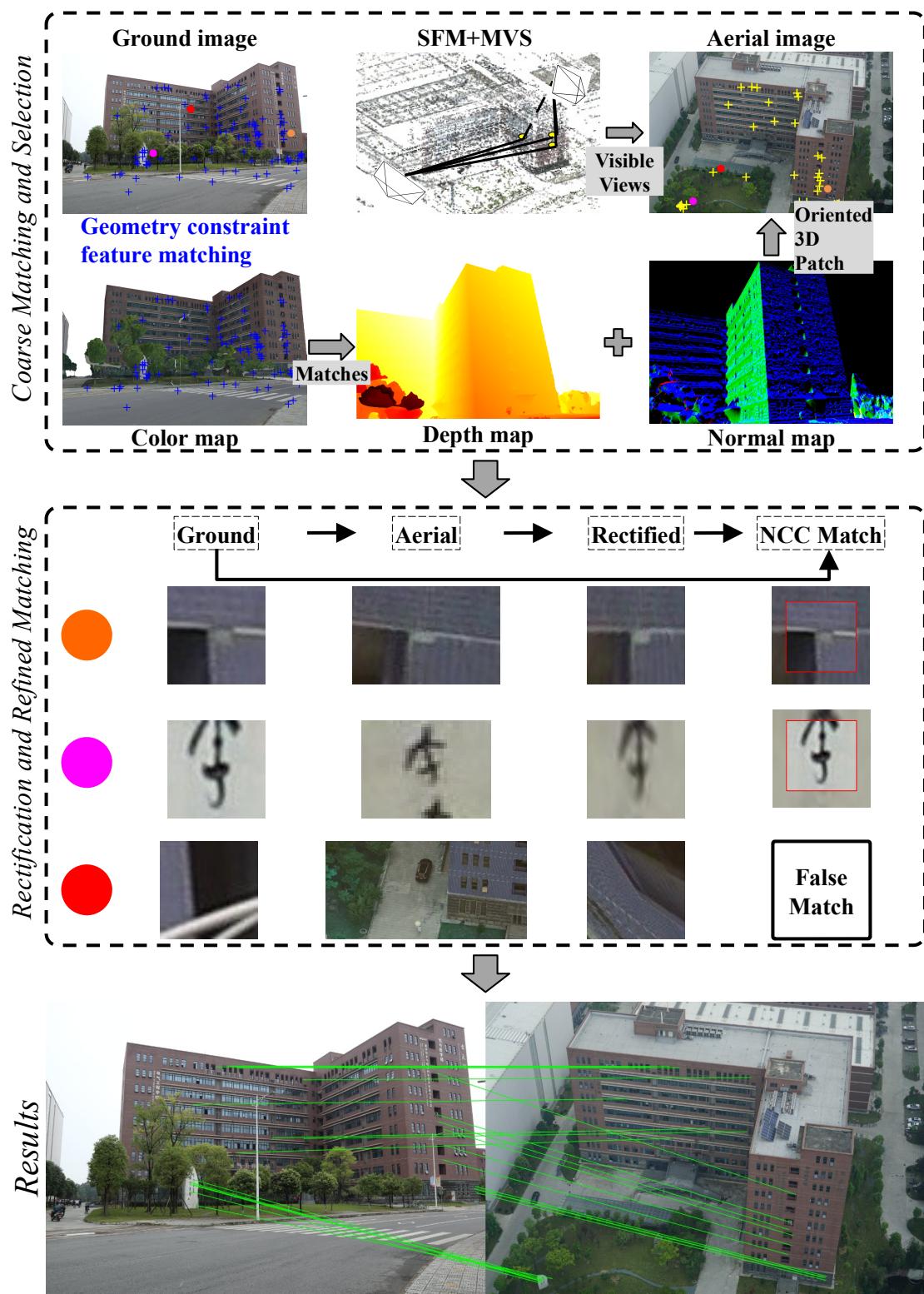


Figure 4: Overview of aerial-ground feature matching. The circles in the coarse-matching images denote the three patches in the refined matching.

222 still the default option in many solutions (Wu et al., 2011; Schönberger and Frahm, 2016). Then,  
 223 we compare descriptors between the ground and synthesized images, using the ratio check and  
 224 filter outliers, using both the proposed geometrical constraints (Subsection 3.3.1) and RANSAC  
 225 (Fischler and Bolles, 1981). Specifically, we use a recent variant of RANSAC, the *a contrario*  
 226 RANSAC (AC-RANSAC), which features automatic threshold tuning (Moisan et al., 2012). If  
 227 the remaining number of pairwise matches between the synthesized and ground images is less  
 228 than five, we consider the matching to be not stable and ignore the results for this pair.

229 3D patches are formulated using the depth and normal information from matches on the  
 230 synthesized images. The coordinate  $\mathbf{X}$  in 3D space is calculated from the corresponding depth  
 231 value using the reverse of Equation 2. The ground sample distance  $\delta = \frac{d}{f}$  is also estimated  
 232 from the depth value  $d$ . We assign a relatively large search window  $w_s\delta$  in the object space as  
 233 delegates, which is centered on and tangential to the oriented points  $(\mathbf{X}, \mathbf{n})$ . In the following  
 234 section, we use the term  $p = (\mathbf{X}, \mathbf{n}, w_s\delta)$  to denote the oriented patches in the object space,  
 235 inspired by previous work (Furukawa and Ponce, 2009). Suitable views of the aerial images are  
 236 selected (Subsection 3.3.2) for each local patch and then the patch is projected to aerial views  
 237 for subsequent refinement.

238 For refined matching (Subsection 3.3.3), a template  $I_g$  on the ground images is extracted, the  
 239 size of which is determined by a correlation window  $w_c$ . Then, correspondence image subsets of  
 240 aerial views  $I_a$  are also extracted and rectified, using the 3D patch and selected aerial views. The  
 241 rectified patches are matched against the template  $I_g$  using normalized correlation coefficient  
 242 (NCC) and least-squares matching (Gruen, 1985; Hu et al., 2016) to refine the aerial-ground  
 243 matches.

### 244 3.3.1. Local geometry constraints for ground-synthesized matching

245 Due to illumination differences between synthesized and ground images, the SIFT match  
 246 may contain significantly more outliers after ratio checking, which leads to inferior RANSAC  
 247 performance. However, because the geometrical differences between the ground and synthesized  
 248 images are almost negligible, the disparities of correct matches should be small and follow con-  
 249 sistent patterns in local regions. Based on these insights, we propose a greedy search algorithm  
 250 to remove outliers prior to RANSAC. Specifically, from a pair of matched points  $p(x_p, y_p)$  and  
 251  $q(x_q, y_q)$ , a directed vector can be obtained as  $m = p - q$ , which denotes the disparity of the  
 252 match. If the initial coarse registration is correct,  $m = \mathbf{0}$  should be satisfied. However, due to  
 253 alignment errors and uncompensated distortion, the disparities  $m$  is not exactly zero. But the  
 254 disparities should at least be consistent with the following three constraints (Figure 5), which  
 255 are used sequentially to filter outliers.

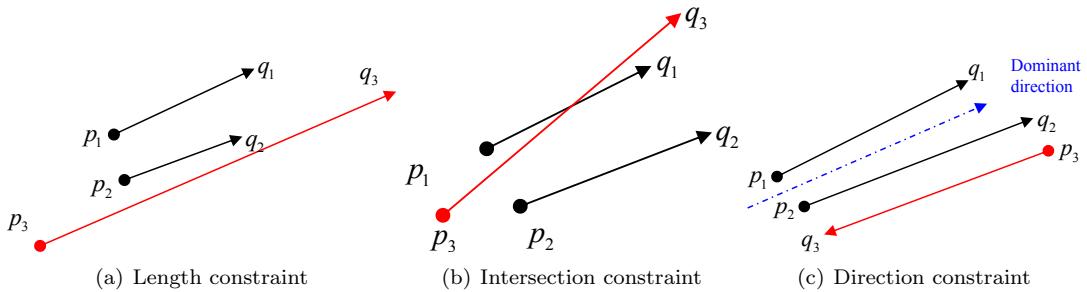


Figure 5: Constraints for outlier filtering in the matching of ground and synthesized images. The points  $p$  and  $q$  denote the key points in the synthesized and ground images, respectively. Note that  $p$  is placed on the ground image. The red lines indicate matches that violate the constraints.

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

258        2) *Intersection constraint.* First, we sort the matches by the lengths of  $|m|$  ascendingly. Then,  
259        we determine if each segment has an intersection with the  $K$ -nearest ( $K = 5$ ) segments. The  
260        segments are indexed using KD-tree. If an intersection exists, the longest segment is marked as  
261        outlier.

262        3) *Direction constraint.* First, we calculate the dominant direction for each segment with  
263        respect to the  $K$ -nearest ( $K = 5$ ) segments. Then, we remove segments that deviate from the  
264        dominant direction by an angle  $\tau_a$  ( $\tau_a = 90^\circ$  is used), similar to the motion consistency in the  
265        work by Jiang and Jiang (2018).

### 266        3.3.2. Propagation of the matches to the aerial images

267        As the meshes are produced from aerial images, the local patches  $p$  should be consistent with  
268        all of the aerial images. In theory, directly projecting the 3D point  $\mathbf{X}$  of the patch  $p$  to *suitable*  
269        aerial views will obtain correspondences between ground and aerial images. In this paper, three  
270        criteria are considered during the selection of *suitable* aerial views, as described below.

271        (1) *Containment*, the local patch should locate inside the frustum of the aerial images. This  
272        criterion is tested by projecting the four corners of the patch defined by the search window  $w_s\delta$   
273        onto all the aerial images.

274        (2) *Consistency*, the orientation of the patch  $\mathbf{n}$  and the direction of aerial image  $\mathbf{R}^T \mathbf{e}_z$  should  
275        be consistent, i.e. less than a threshold  $\tau_n = 90^\circ$ . This criterion is used because the subset of  
276        the aerial images will be rectified locally for the subsequent refinement; if the normal vector of  
277        the patch is inconsistent with the aerial view, the rectified image will be blurred due to large  
278        perspective deformation.

279        (3) *Visibility*, the patch should not be occluded by the mesh itself. For occlusion detection, the  
280        optimized bounding volume hierarchy (BVH) of the triangular meshes implemented in Embree  
281        (Wald et al., 2014) is used for ray tracing. As BVH structures have almost linear space complexity  
282        with regard to the number of triangles, we cache the BVH structure in advance using the meshes  
283        that have the finest level of detail. We use OpenSceneGraph (Osfield and Burns, 2014) to load  
284        the triangular meshes, which are segmented into small fragments. Then, Geogram (Lévy, 2015)  
285        is used to automatically clean the fragmented meshes, including welding close vertices and fixing  
286        miscellaneous topological defects.

### 287        3.3.3. Matching refinement between aerial and ground images

288        Although the meshes used for rendering are obtained from aerial images, the matches prop-  
289        agated to the aerial images may be inaccurate. The geometry of meshes is noise-laden and the  
290        textural information is blended and blurred, as shown in Figure 6. Therefore, the coordinates  
291        of the synthesized images and the corresponding depth value can not be used directly in the  
292        combined bundle adjustment. We add an additional step to solve this problem: propagating  
293        the matches to aerial images and directly matching the local patches between ground and aerial  
294        images. In this way, the matches on the original images will finally be used in the bundle  
295        adjustment.

296        Inspired by the MPGC approach (Zhang, 2005) and our previous view-independent synthesis  
297        (Hu et al., 2015), we also project all of the patches to the same plane using the homographic  
298        transformation  $\mathbf{H}$  (Hartley and Zisserman, 2003):

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

299        where  $\mathbf{K}$  is the camera matrix;  $\mathbf{R}$  and  $\mathbf{t}$  are the relative orientation and translation parameters  
300        between the two images, which are deducted from the orientation parameters after coarse regis-



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

301 tration;  $\mathbf{n}_d = \frac{\mathbf{n}}{d}$  is the scaled normal vector of the patch, with  $\mathbf{n}$  the normal vector of the patch  
 302 and  $d$  the distance between patch and aerial view; and the subscripts  $g$  and  $a$  denote the ground  
 303 and aerial images, respectively. Notably, only the local patches surrounding the initial position  
 304 are loaded and transformed, rather than the entire images as our previous work (Hu et al., 2015).

305 After rectifying all of the patches, a classic NCC search is used to find the initial match,  
 306 followed by LSM to further improve the location. The patch extracted from the ground image  
 307 serves as the template for matching and all of the aerial images are aligned pairwisely. Any match  
 308 with a correlation smaller than a threshold  $\tau_c$  ( $\tau_c = 0.75$  is used) is pruned before LSM. After  
 309 LSM, reverse homographic transformation in Equation 2 is used to obtain the final coordinates  
 310 on the aerial images.

311 **4. Experimental evaluations**

312 *4.1. Dataset descriptions*

313 Five datasets (see Table 1 and Figure 1) are used to evaluate the proposed methods, which  
 314 comprise the ISPRS benchmark dataset collected at Centre of Dortmund and Zeche of Zurich  
 315 (Nex et al., 2015) and three datasets collected at the campus of SWJTU. Qualitative and quanti-  
 316 tative feature point matching experiments are conducted and compared with existing commercial

<sup>317</sup> solutions (Acute3D, 2019; Agisoft, 2019) and one of the most recent algorithm (Revaud et al.,  
<sup>318</sup> 2019). In addition, to further verify the capability of proposed method, 3D reconstruction results  
<sup>319</sup> are also presented and compared.

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

Dataset	Sensor		Resolution (cm)		#Images	
	Aerial	Ground	Aerial	Ground	Aerial	Ground
Centre	SONY Nex-7	SONY Nex-7	1.10	0.53	146	204
Zeche	SONY Nex-7	SONY Nex-7	0.56	0.28	172	147
SWJTU-LIB	SONY ICLE-5100	Cannon EOS M6	1.69	1.06	123	78
SWJTU-BLD	SONY ICLE-5100	Cannon EOS M6	1.93	1.33	207	88
SWJTU-RES	SONY ICLE-5100	DJI spark	1.97	2.56	92	192

<sup>320</sup> The Centre and Zeche datasets are collected by ISPRS in Dortmund and Zurich, respectively.  
<sup>321</sup> Both the aerial and ground images surrounding a typical building are captured using the same  
<sup>322</sup> sensor. The other three datasets were all collected in the campus of SWJTU, specifically at the  
<sup>323</sup> library (SWJTU-LIB), a general building (SWJTU-BLD) and residential areas (SWJTU-RES).  
<sup>324</sup> Unlike the ISPRS datasets, the aerial images of SWJTU datasets were collected in flights of  
<sup>325</sup> regular strips and the ground images were captured only for areas of interest. It should be noted  
<sup>326</sup> that the ground images of SWJTU-RES were not essentially obtained on the ground, which were  
<sup>327</sup> also captured by a low-cost UAV in a vertical uplift flight. However, because they share similar  
<sup>328</sup> characteristic of other ground images, we also deem them as ground in this study.

<sup>329</sup> SFM results of both the aerial and ground images are obtained prior to the processing pro-  
<sup>330</sup> posed in this paper. In addition, we assume that both image sets are registered roughly; the  
<sup>331</sup> coarse registration is conducted through either the weak GNSS data obtained in the EXIF header  
<sup>332</sup> of the images (for Center and Zeche) or three interactively selected tie-points when GNSS data  
<sup>333</sup> are not available (for the three SWJTU datasets).

#### <sup>334</sup> 4.2. Evaluation of feature matching

##### <sup>335</sup> 4.2.1. Evaluation of feature matching between ground and synthesized images

<sup>336</sup> Because the synthesized images are obtained using the orientation parameters after coarse  
<sup>337</sup> registration, the appearances between ground and synthesized images should be similar. In  
<sup>338</sup> addition, the disparities of the feature matches, *i.e.* the difference of image coordinates, should  
<sup>339</sup> be small. This is confirmed in Figure 7, in which the cyan arrows indicate the disparities drawn  
<sup>340</sup> on the ground images. In fact, the lengths of the disparities can also reflect the accuracies of  
<sup>341</sup> coarse registration. Another expected characteristic of the distribution of disparities is that they  
<sup>342</sup> are consistent in local regions, as shown in the enlarged subsets on the right of each subfigure.  
<sup>343</sup> This is, in fact, the rationale behind the proposed geometric constraints.

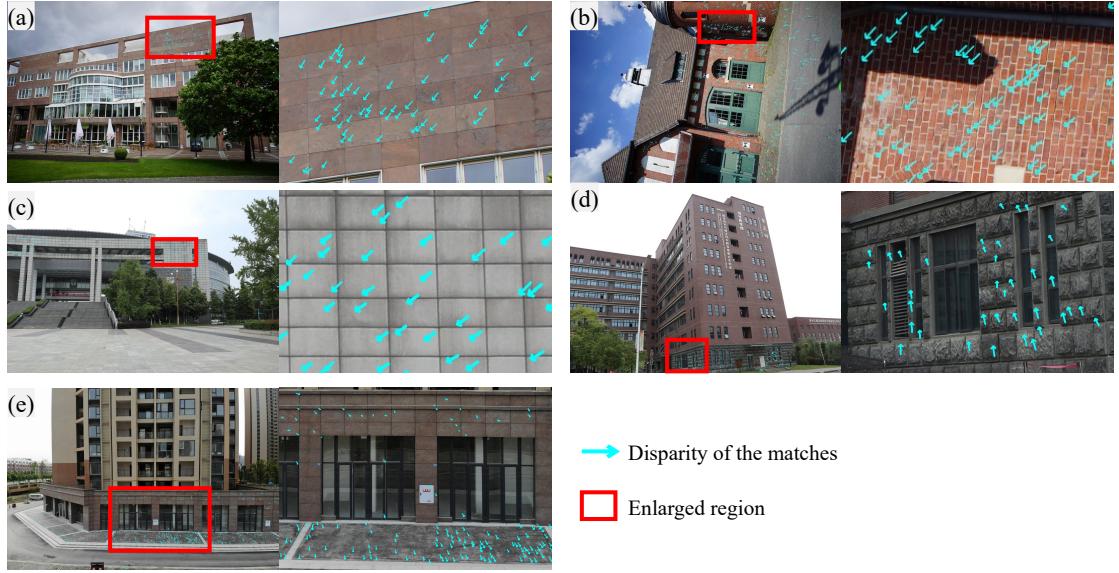


Figure 7: Disparities of the matches between ground and synthesized images drawn on the ground images. (a) to (e) represent results for Centre, Zeche, SWJTU-LIB, SWJTU-BLD and SWJTU-RES, respectively. The arrows are pointing from the coordinates of ground images to those of the synthesized images. Enlarged views indicated by the rectangles are shown on the right part of each subfigure.

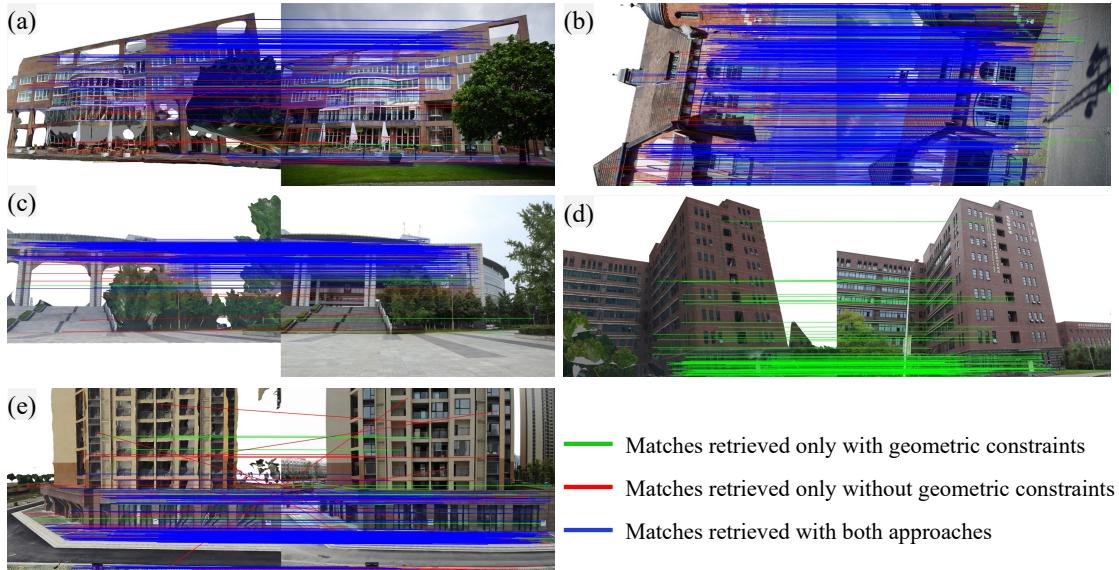


Figure 8: Comparisons of the matches between ground and synthesized images with and without the geometric constraints. (a) to (e) represent results for Centre, Zeche, SWJTU-LIB, SWJTU-BLD and SWJTU-RES, respectively. After ratio checks, the putative matches are categorized into three types: 1) green lines represent matches retrieved only with the geometric constraints; 2) red lines represent matches retrieved only without the geometric constraints; and 3) blue lines represent matches retrieved with both approaches.

344 To evaluate the performances of the proposed geometric constraints in the matching of synthesized  
 345 images, we compare feature matches with and without the proposed geometric

346 constraints. Figure 8 shows the matching results for the five datasets. With geometric con-  
 347 straints, the outlier filtering is more stable; we have succeeded in retrieving correct models for all  
 348 the five datasets, while the SWJTU-BLD is failed without geometric constraint as also demon-  
 349 strated in Table 2. Notably, even for datasets succeeded without geometric constraints, more  
 350 outliers are visible, such as Figure 8a and e.

Table 2: Comparisons of the outlier filter with and without the proposed geometric constraints in the matching between ground and synthesized images. The values for SIFT are putative matches after ratio checks. The values for the fourth and fifth columns are correct matches after outlier filter.

Dataset	Image	#SIFT	#Without Constraint	#With Constraint
Centre	DSC02820	1863	180	152
Zeche	DSC04664	2685	530	525
SWJTU-LIB	DSC01726	2152	385	316
SWJTU-BLD	IMG1919	2111	0	84
SWJTU-RES	DJI0137	2098	266	263

#### 351 4.2.2. Evaluation of feature-matching between aerial and ground images

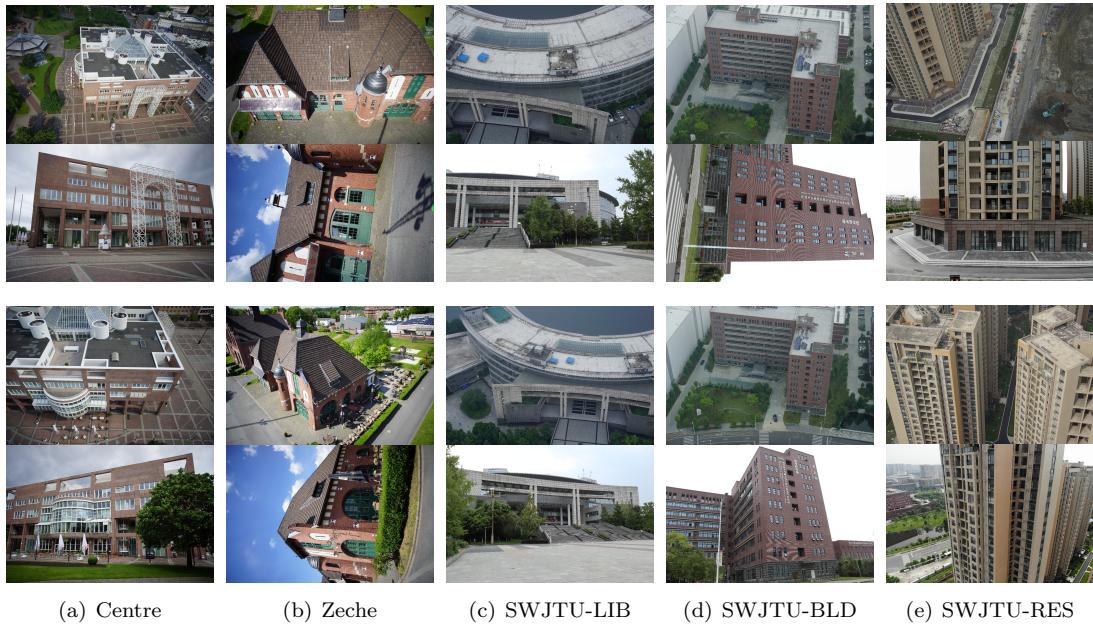


Figure 9: The selected 10 image pairs from the five test datasets. The odd and even rows show images from aerial and ground sets, respectively.

352 We compare the feature matching results against both SFM pipelines and *ad hoc* features.  
 353 Five solutions are considered, including the proposed approach, one commercial solution, *i.e.*  
 354 Agisoft MetaShape (Agisoft, 2019), two freeware solutions, *i.e.* VisualSfM (Wu et al., 2011)  
 355 and Colmap (Schönberger and Frahm, 2016; Schönberger et al., 2016) and a recent feature based  
 356 on deep learning, *i.e.* R2D2 (Revaud et al., 2019). Ten pairs are selected from the five datasets,  
 357 with two pairs for each dataset (Figure 9). We prefer pairs with large convergent angles as

long as the selected pairs have enough overlaps. As it is possible that the matching results are noise-laden, we manually count the number of correct matches for the ten pairs; the correctness is validated only roughly, such as the same tile of the wall.

Table 3 summarizes the results. Notably, the other four 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 are still not solved by them. On the contrary, the proposed methods succeeds in all the cases, with convergent angle up to 75°

Table 3: Comparisons of the numbers of matches for 10 pairs of images between aerial and ground datasets, in which two pairs are selected for each dataset. The convergent angles for the image pairs are shown in the second row.

Dataset Angle (°)	Centre		Zeché		SWJTU-LIB		SWJTU-BLD		SWJTU-RES	
	50.8	61.9	40.9	51.5	54.6	61.2	59.6	70.2	34.6	75.1
Proposed	243	114	188	304	91	161	24	5	72	94
VisualSFM	0	12	0	0	12	0	0	0	6	0
Photoscan	0	0	0	0	0	0	0	0	0	0
Colmap	0	17	0	0	29	0	0	0	0	0
R2D2	17	15	0	0	0	0	0	0	0	0

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 10 through 14. During these comparisons, the pair with larger convergent angle in Table 3 is chosen. The proposed methods succeeds in obtaining a certain amount of correct matches for all the five pairs; and VisualSFM only manages to obtain some correct matches for the Centre dataset only, with noticeably higher outlier ratio.

We also highlight some interesting and appealing characteristics of the proposed methods in the enlarged regions. 1) *Repeated pattern*, the walls of Centre, Zeche and SWJTU-LIB all demonstrate clear repeated patterns and the proposed approach achieves satisfactory performances in this scenario. 2) *Perspective deformation*, the proposed method is agnostic to perspective deformation as seen in the deformed wall tiles of Centre and SWJTU-LIB; this is because the descriptor searching is only conducted between the ground and synthesized images and template matching and least-squares matching are conducted after rectification guided by the local patch. 3) *Different deformation models*, pairwise rectification based on a common plane (Wu et al., 2018; Gao et al., 2018) can only alleviate perspective deformation on a single plane, but the proposed method can obtain matches on both the ground and façades at the same time, as seen in Centre, Zeche and SWJTU-RES. 4) *Glassy objects*, it is arguably that glassy objects are the most challenging cases for image matching, which also causes problem for the proposed approaches; however, we still obtain several correct matches for the SWJTU-BLD dataset; in fact, tens of matches are obtained between ground and synthesized images and five remains after propagating to the aerial view.

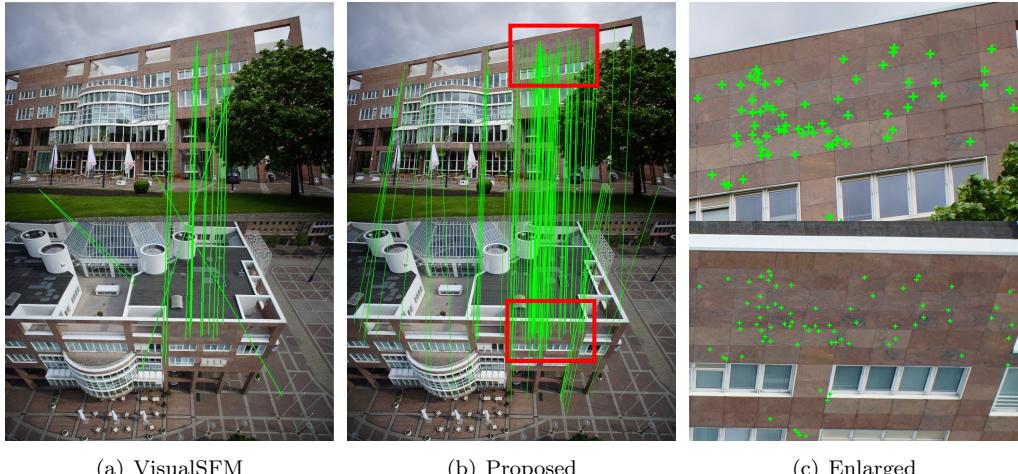


Figure 10: Aerial-ground matching results for the DSC02820-DSC07379 pair from the Dortmund dataset. The red rectangles denote the enlarged areas. The convergent angle between the two images is  $61.9^\circ$ .

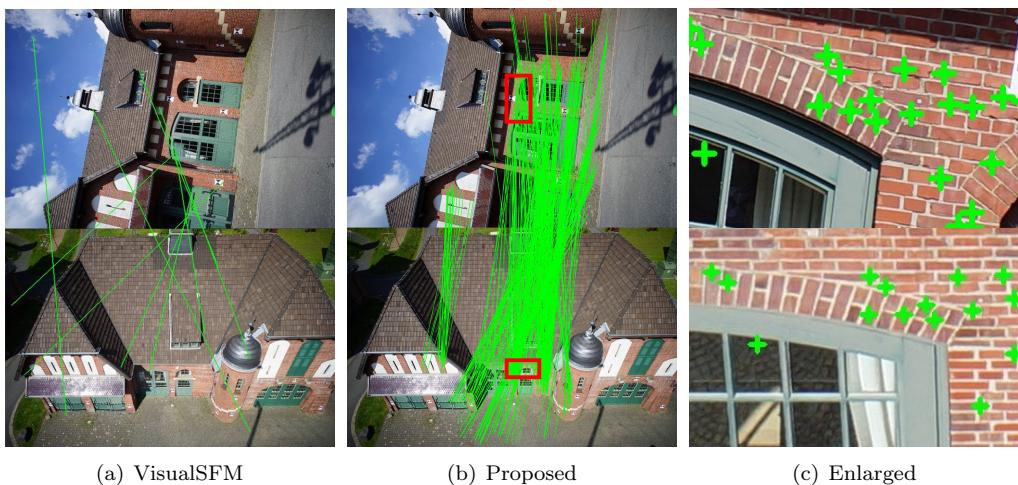


Figure 11: Aerial-ground matching results for the DSC04664-DSC06239 pair from the Zeche dataset. The red rectangles denote the enlarged areas. The convergent angle between the two images is  $51.5^\circ$  and the enlarged view for the ground image is rotated  $90^\circ$  clock-wisely for better visualization.

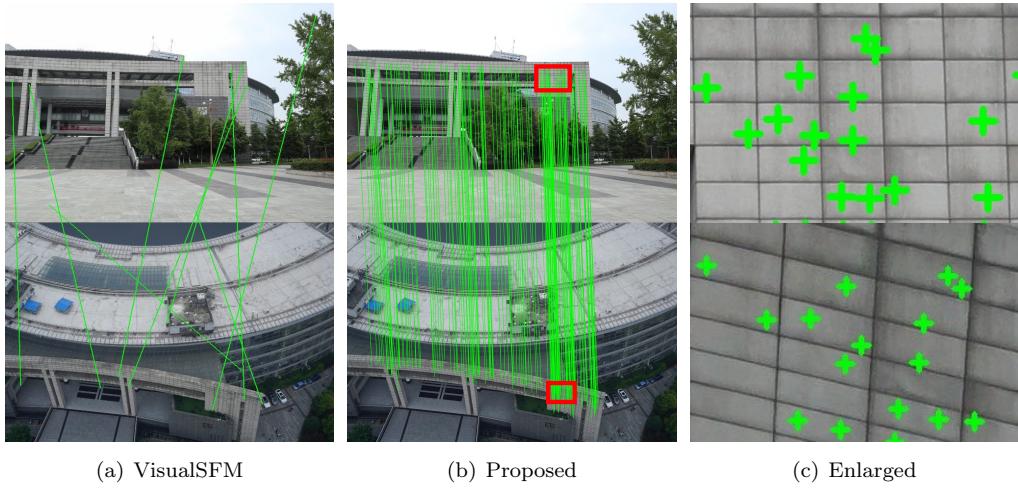


Figure 12: Aerial-ground matching results for the IMG1726-W0762 pair from the SWJTU-LIB dataset. The red rectangles denote the enlarged areas. The convergent angle between the two images is  $61.2^\circ$ .

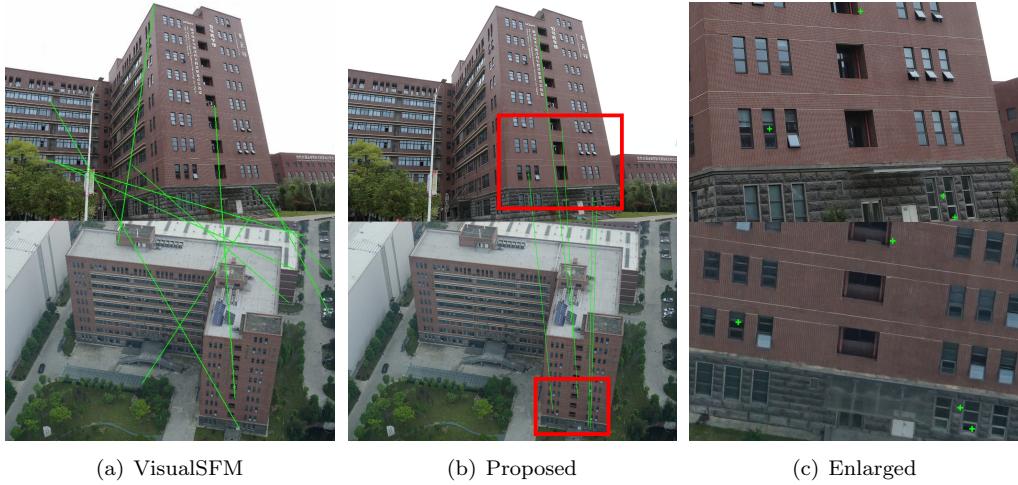


Figure 13: Aerial-ground matching results for the IMG1919-X0650 pair from the SWJTU-BLD dataset. The red rectangles denote the enlarged areas. The convergent angle between the two images is  $70.2^\circ$ .

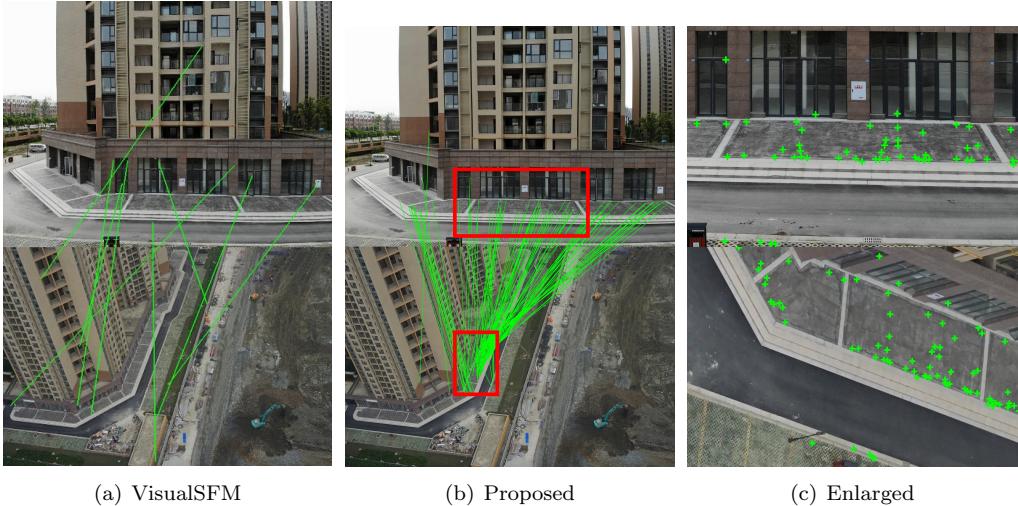


Figure 14: Aerial-ground matching results for the DJI0312-D0605 pair of the SWJTU-RES dataset. The red rectangles denote the enlarged areas. The convergent angle between the two images is  $75.1^\circ$  and the enlarged view for the aerial image is rotated  $90^\circ$  clock-wisely for better visualization.

#### 387 4.2.3. Evaluation of efficiencies of the feature matching

388 The time complexity of the feature matching strategy to connect aerial and ground sets of  
 389 images is  $O(n)$ , with respect to the number of ground images. On the contrary, simply enumerating  
 390 all the pairs has time complexity of  $O(n^2)$ . Considering the large appearant differences  
 391 between aerial and ground images, the image retrieval technique that achieves time complexity  
 392 of  $O(Kn)$  may not be quite helpful, in which  $K$  is a constant for the most similar  $K$  images.

393 However, the runtime of a single pair is absolutely longer due to the additional steps involved.  
 394 Therefore, we separate the feature matching for a single ground image into three stages: 1)  
 395 rendering, which consists of loading the mesh models and retrieving all the synthesized images; 2)  
 396 pairwise matching, which consists of detecting features, descriptor searching and outlier removal  
 397 and this is a common step involved in almost all feature matching methods; and 3) propagation,  
 398 which collects visible aerial views, loads the local patches from disks and refines the matches.  
 399 As shown in Table 4, the costs of additional stages, *e.g.* rendering and propagation, are always  
 400 *on par* with pairwise matching. The ratios between additional steps and pairwise matching are  
 401 in the range of (1,2), which indicates that the proposed approach has a linear time complexity,  
 402 with respect to the number of ground images.

Table 4: Comparisons of different stages of the proposed matching strategy for a ground image. The number of matches are also recorded in the second row and the runtime of last stage is dependent on this number.

Dataset	Centre		Zeche		SWJTU-LIB		SWJTU-BLD		SWJTU-RES	
#Matches	277	152	349	525	352	316	74	61	151	263
Rendering (s)	2.8	8.5	2.9	3.3	6.3	5.2	2.2	5.8	1.1	0.7
Pairwise Match (s)	4.0	5.6	2.5	4.2	6.2	5.7	4.2	5.9	3.6	2.7
Propagation (s)	1.5	3.6	5.0	8.8	11.5	4.1	1.7	0.9	1.3	1.5

403    *4.3. Evaluations of the integrated reconstruction*

404    We develop an add-on solution for integrated reconstruction, based on ContextCapture  
405    ([Acute3D, 2019](#)). In addition, we also compare three other solutions: the vanilla ContextCapture  
406    ([Acute3D, 2019](#)), MetaShape ([Agisoft, 2019](#)) and Colmap ([Schönberger and Frahm, 2016](#)). Both  
407    sparse and dense reconstructions are evaluated in the following subsections.

408    *4.3.1. Evaluation of integrated sparse reconstruction*

409    First, we demonstrate the SFM results by comparing the final numbers of reconstructed  
410    images. As some solutions can automatically separate the images into several clusters, only the  
411    largest cluster is considered. In addition, we report the number of tie-points that connect aerial  
412    and ground images, as these points are the most crucial for the integrated reconstruction. In our  
413    experiments, without interactively selected tie-points, most other solutions will not converge to  
414    a reasonable results in the SFM procedure. To make a fair comparison, we take about an hour  
415    of labor work to add user-input tie-points in the solutions of ContextCapture and MetaShape,  
416    for each dataset.

417    Table 5 shows the SFM performances, and it can be seen that the proposed add-on solution  
418    for ContextCapture succeeds in all the cases, while the vanilla ContextCapture fails in most of  
419    them even with interactively selected tie-points. With user-input tie-points, MetaShape manages  
420    to register four out of the five datasets, but the number of tie-points connecting images between  
421    aerial and ground sets are fewer than the proposed methods. It is also interesting to see that  
422    Colmap succeeds in two datasets even without human interventions using SIFT features; this is  
423    probably due to the reliable incremental SFM pipeline ([Schönberger and Frahm, 2016](#)). However,  
424    we argue that enough tie-points are also important, considering that the proposed approach out-  
425    performs other solutions even with a relatively weak SFM solution bundled in ContextCapture.  
426    In the Zeche dataset, 31 aerial images are not reconstructed, this is because that the original  
427    aerial-only SFM result from ContextCapture does not contain them.

428    To further evaluate the precision and accuracy of the proposed methods, the position un-  
429    certainties from the aerial triangulation report and the root-mean-square error (RMSE) of the  
430    check points are used. The former (Table 6) metric denotes the internal stability of the SFM  
431    results, which is estimated from the covariance matrix ([Agarwal et al., 2012](#)) of the least-squares  
432    solver and taken from the report of ContextCapture. The latter (Table 7) denotes performance  
433    against external control networks. As different datasets have different accuracies, we also report  
434    the results generated using only aerial images as baseline.

435    For the uncertainties of image positions (Table 6), almost all the results from aerial-ground  
436    integrated approach are better than that of only UAV images, except for SWJTU-BLD; this  
437    is probably due to better convergent geometries formed by both aerial and ground images; for  
438    SWJTU-BLD, the reason is that the feature matching performances are less robust due to the  
439    glassy objects.

440    For the accuracies of the check points, the results from MetaShape are also compared on the  
441    four datasets that MetaShape successfully registered. For each dataset, three or four control  
442    points are used in the bundle adjustment, and five to eight check points are used for evaluations.  
443    Both control and check points are manually marked at least on three images. Compared to the  
444    reference results using UAV images only, both the proposed solution and MetaShape achieved  
445    satisfactory results. The proposed solution using ContextCapture as the backend for SFM gen-  
446    erally has slightly better horizontal accuracies and MetaShape has better vertical accuracies.

Table 5: Comparisons of different solutions for the five datasets on the sparse reconstruction. The numbers of reconstructed images proportional to the total image numbers are reported in the third and fourth columns. In addition, the numbers of aerial-ground tie-points are presented in the fifth column.

Dataset	Method	#Images		#Aerial-ground tie-points	Status
		Ground	Aerial		
Center	Proposed+ContextCapture	203/204	146/146	23648	Succeeded
	ContextCapture	204/204	0/146	0	Failed
	Photoscan	203/204	146/146	10428	Succeeded
	Colmap	168/204	0/146	0	Failed
Zeché	Proposed+ContextCapture	172/172	116/147	38796	Succeeded
	ContextCapture	172/172	116/147	817	Succeeded
	Photoscan	172/172	147/147	23201	Succeeded
	Colmap	172/172	147/147	3171	Succeeded
SWJTU-LIB	Proposed+ContextCapture	78/78	123/123	11399	Succeeded
	ContextCapture	78/78	123/123	20	Succeeded
	Photoscan	78/78	123/123	1614	Succeeded
	Colmap	78/78	123/123	1374	Succeeded
SWJTU-BLD	Proposed+ContextCapture	88/88	207/207	1706	Succeeded
	ContextCapture	0/88	205/207	0	Failed
	Photoscan	0/88	207/207	0	Failed
	Colmap	38/88	0/207	0	Failed
SWJTU-RES	Proposed+ContextCapture	192/192	88/92	779	Succeeded
	ContextCapture	192/192	0/92	0	Failed
	Photoscan	192/192	91/92	323	Succeeded
	Colmap	0/192	16/92	0	Failed

Table 6: Evaluation of the position uncertainties for each images after bundle adjustment. The values are taken from the report of ContextCapture. For reference, the results from only the aerial images are also demonstrated.

Dataset	UAV only (cm)			Integrated (cm)		
	X	Y	Z	X	Y	Z
Centre	0.10	0.10	0.10	0.07	0.07	0.07
Zeché	0.04	0.04	0.04	0.03	0.03	0.03
SWJTU-LIB	0.53	0.46	0.58	0.32	0.30	0.32
SWJTU-BLD	0.58	0.56	0.45	0.71	0.77	0.59
SWJTU-RES	3.59	7.89	7.26	2.65	1.06	3.33

Table 7: Comparisons of the accuracies of check points for the integrated reconstruction. For reference, we also report the results generated using only the aerial images as baseline. The symbol “-” indicates missed results due to either lack of check points or failure of the SFM pipeline.

Dataset	UAV Only (cm)			Proposed (cm)			MetaShape (cm)		
	X	Y	Z	X	Y	Z	X	Y	Z
Centre	-	-	-	2.6	-2.0	2.2	8.3	5.9	4.8
Zeche	1.2	2.3	1.4	1.3	1.9	1.6	2.2	2.2	0.7
SWJTU-LIB	1.0	1.1	32.1	2.4	3.3	15.5	7.8	7.5	8.8
SWJTU-BLD	1.6	1.0	4.9	3.4	9.9	12.1	-	-	-
SWJTU-RES	4.7	0.9	12.7	2.7	0.7	14.5	9.7	6.6	6.5

#### 447 4.3.2. Evaluation of integrated dense reconstruction

448 Figure 15 compares the textured mesh models obtained using only aerial images (top row)  
449 and integrated solutions (bottom row). We also highlight some parts of the models on the right  
450 of each subfigure. Using the integrated solution, the textures on the façades are clearer, as shown  
451 in Figure 15a, c and d. In addition, the reconstructed models are obviously better and more  
452 complete, as can be seen in Figure 15c and the small objects in Figure 15d. The quality of  
453 texture is also improved, such as the blurred areas under the eaves in Figure 15b.

#### 454 4.4. Discussions and limitations

455 Based on the above evaluations for feature matching and integrated reconstruction, we sum-  
456marize some characteristics and limitations of the proposed methods.

457 1) *Integration with existing SFM and MVS pipeline.* Although previous solutions (Wu et al.,  
458 2018; Gao et al., 2018) can satisfactorily incorporate aerial and ground images into the same  
459 framework, they break existing SFM pipeline and require *ad hoc* bundle adjustment approaches.  
460 In fact, the tie-points in the sparse reconstruction are also important for subsequent MVS  
461 pipeline, which are used as initial surfaces or constraints, such as the patch-based expanding  
462 (Furukawa and Ponce, 2009), variational refinement (Vu et al., 2011; Yu and Gao, 2020) or De-  
463 naulay triangulation constraints (Wu et al., 2012). Instead, the proposed method can be directly  
464 used as add-on to existing SFM and MVS pipelines (Acute3D, 2019).

465 2) *Efficiency and accuracy.* The proposed pipeline is also fast and accurate. We do not  
466 need to enumerate all the pairs between aerial and ground images, which has quadratic time  
467 complexity. Instead, feature matching is only required between ground and synthesized images  
468 and is propagated to the aerial views, which has linear time complexity. This is important,  
469 because if large viewpoint differences exist, we cannot rely on descriptor-based image retrieval  
470 to reduce the numbers of matching pairs. In addition, an additional refinement step is adopted  
471 to improve the location of aerial-ground matches.

472 3) *Limitations.* A limitation of the proposed approach is shared by previous works (Wu et al.,  
473 2018; Gao et al., 2018), namely that dense reconstruction is required prior to the SFM pipeline.  
474 Although our method also requires an additional step, *i.e.* texture mapping, all the above steps  
475 are generally bundled in an unified MVS pipeline. In addition, only regions of interest need to be  
476 retouched (Acute3D, 2019) and the runtime overhead may be ignored. Nonetheless, the quality  
477 of the textured mesh models will inevitably influence the performance of our approach.

## 478 5. Conclusion

479 In this paper, we address the problem of feature matching between aerial and ground images,  
480 which currently suffers from severe perspective deformation resulting from viewpoint differences.

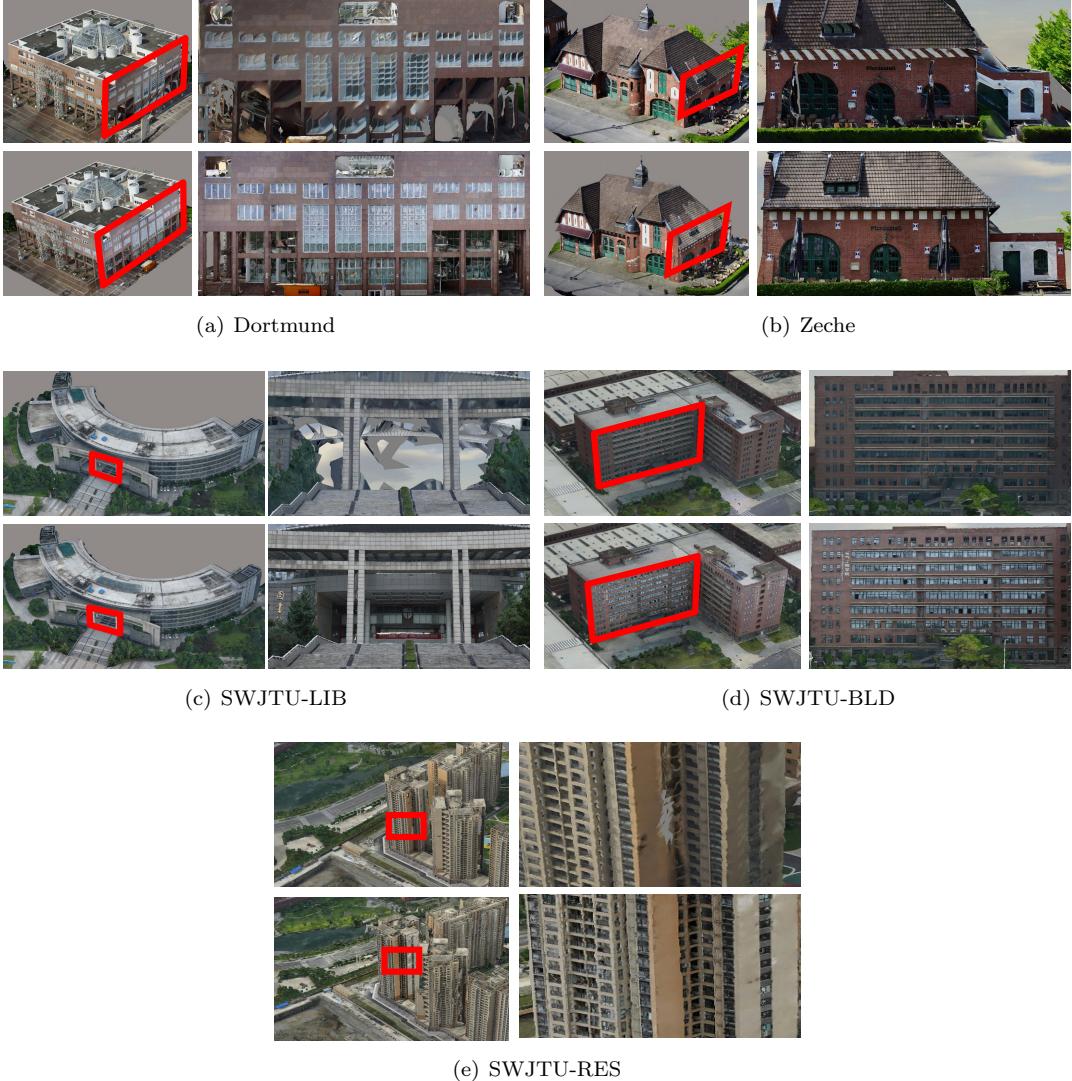


Figure 15: Comparison of the textured mesh models generated from only aerial images (top row), and those generated from aerial-ground images (bottom row). The right column of each subfigure is an enlargement of the regions highlighted by the rectangles.

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 proposed method is featuring four appealing characteristics: 1) simplicity, the proposed method can be used as add-on solution to existing SfM and MVS pipelines, which simplifies the integration; 2) efficiency, the proposed strategy has linear time complexity rather than quadratic for pairwise rectification (Wu et al., 2018; Gao et al., 2018); 3) accurate, the matches are refined locally between aerial and ground images; and 4) robust, the proposed approach is agnostic to the convergent angle between aerial and ground images. Future works may be devoted to further

490 exploiting the possibility of integrating light detection and ranging (LiDAR) point clouds and  
491 panoramas collected by the ground mobile-mapping systems into aerial datasets. Code and the  
492 SWJTU datasets will be made publicly available at <https://github.com/saedrna/RenderMatch>.

493 **Acknowledgments**

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

498 **References**

- 499 Acute3D, 2019. Context capture. <https://www.acute3d.com/>.
- 500 Agarwal, S., Mierle, K., et al., 2012. Ceres solver. <http://ceres-solver.org/>.
- 501 Agisoft, 2019. Agisoft metashape. <https://www.agisoft.com/>.
- 502 Arandjelović, R., Zisserman, A., 2012. Three things everyone should know to improve object  
503 retrieval, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp.  
504 2911–2918.
- 505 Bentley, 2019. Contextcapture camera model. [https://docs.bentley.com/LiveContent/web/ContentCapture\\_Help-v9/en/index.html](https://docs.bentley.com/LiveContent/web/ContentCapture_Help-v9/en/index.html).
- 506 Bursuc, A., Tolias, G., Jégou, H., 2015. Kernel local descriptors with implicit rotation matching,  
507 in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM.  
508 pp. 595–598.
- 509 Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., Fua, P., 2012. Brief: Computing  
510 a local binary descriptor very fast. IEEE Transactions on Pattern Analysis and Machine  
511 Intelligence 34, 1281–1298.
- 512 Chen, M., Shao, Z., 2013. Robust affine-invariant line matching for high resolution remote sensing  
513 images. Photogrammetric Engineering and Remote Sensing 79, 753–760.
- 514 Dusmanu, M., Rocco, I., Pajdla, T., Pollefeyns, M., Sivic, J., Torii, A., Sattler, T., 2019. D2-  
515 net: A trainable cnn for joint detection and description of local features. arXiv preprint  
516 arXiv:1905.03561 .
- 517 Fanta-Jende, P., Nex, F., Vosselman, G., Gerke, M., 2019. Co-registration of panoramic mobile  
518 mapping images and oblique aerial images. The Photogrammetric Record 34, 148–173.
- 519 Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with  
520 applications to image analysis and automated cartography. Communications of the ACM 24,  
521 381–395.
- 522 Furukawa, Y., Ponce, J., 2009. Accurate, dense, and robust multiview stereopsis. IEEE transactions  
523 on pattern analysis and machine intelligence 32, 1362–1376.
- 524 Gao, X., Shen, S., Zhou, Y., Cui, H., Zhu, L., Hu, Z., 2018. Ancient chinese architecture 3d  
525 preservation by merging ground and aerial point clouds. ISPRS Journal of Photogrammetry  
526 and Remote Sensing doi:<https://doi.org/10.1016/j.isprsjprs.2018.04.023>.

- 528 GLM, 2019. Opengl mathematics. <https://glm.g-truc.net/>.
- 529 Gruen, A., 1985. Adaptive least squares correlation: a powerful image matching technique. South  
530 African Journal of Photogrammetry, Remote Sensing and Cartography 14, 175–187.
- 531 Hartley, R., Zisserman, A., 2003. Multiple view geometry in computer vision. Cambridge uni-  
532 versity press.
- 533 Hu, H., Ding, Y., Zhu, Q., Wu, B., Xie, L., Chen, M., 2016. Stable least-squares matching  
534 for oblique images using bound constrained optimization and a robust loss function. ISPRS  
535 journal of photogrammetry and remote sensing 118, 53–67.
- 536 Hu, H., Zhu, Q., Du, Z., Zhang, Y., Ding, Y., 2015. Reliable spatial relationship constrained  
537 feature point matching of oblique aerial images. Photogrammetric Engineering & Remote  
538 Sensing 81, 49–58.
- 539 Javanmardi, M., Javanmardi, E., Gu, Y., Kamijo, S., 2017. Towards high-definition 3d urban  
540 mapping: Road feature-based registration of mobile mapping systems and aerial imagery.  
541 Remote Sensing 9, 975.
- 542 Jende, P., Nex, F., Gerke, M., Vosselman, G., 2018. A fully automatic approach to register  
543 mobile mapping and airborne imagery to support the correction of platform trajectories in  
544 gnss-denied urban areas. ISPRS Journal of Photogrammetry and Remote Sensing 141, 86–99.  
545 doi:<https://doi.org/10.1016/j.isprsjprs.2018.04.017>.
- 546 Jiang, S., Jiang, W., 2017. On-board gnss/imu assisted feature extraction and matching for  
547 oblique uav images. Remote Sensing 9, 813.
- 548 Jiang, S., Jiang, W., 2018. Hierarchical motion consistency constraint for efficient geometrical  
549 verification in uav stereo image matching. ISPRS Journal of Photogrammetry and Remote  
550 Sensing 142, 222–242.
- 551 Lemmens, M., 2014. Oblique imagery: the standard for mapping. GIM International 28, 14–17.
- 552 Lévy, B., 2015. Geogram. <http://alice.loria.fr/software/geogram/doc/html/index.html>.
- 553 Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International Jour-  
554 nal of Computer Vision 60, 91–110.
- 555 Ma, X., Liu, D., Zhang, J., Xin, J., 2015. A fast affine-invariant features for image stitching under  
556 large viewpoint changes. Neurocomputing 151, 1430–1438. doi:[10.1016/j.neucom.2014.10.045](https://doi.org/10.1016/j.neucom.2014.10.045).
- 558 Matas, J., Chum, O., Urban, M., Pajdla, T., 2004. Robust wide-baseline stereo from maximally  
559 stable extremal regions. Image and Vision Computing 22, 761–767.
- 560 Mikolajczyk, K., Schmid, C., 2004. Scale and affine invariant interest point detectors. Interna-  
561 tional Journal of Computer Vision 60, 63–86.
- 562 Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir,  
563 T., Van Gool, L., 2005. A comparison of affine region detectors. International Journal of  
564 Computer Vision 65, 43–72.
- 565 Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J., 2017. Working hard to know your neigh-  
566 bor's margins: Local descriptor learning loss, in: Advances in Neural Information Processing  
567 Systems, pp. 4826–4837.

- 568 Moisan, L., Moulon, P., Monasse, P., 2012. Automatic homographic registration of a pair of  
569 images, with a contrario elimination of outliers. *Image Processing On Line* 2, 56–73.
- 570 Molina, P., Blázquez, M., Cucci, D.A., Colomina, I., 2017. First results of a tandem terrestrial-  
571 unmanned aerial mapkite system with kinematic ground control points for corridor mapping.  
572 *Remote Sensing* 9, 60.
- 573 Morel, J.M., Yu, G., 2009. Asift: A new framework for fully affine invariant image comparison.  
574 *SIAM Journal on Imaging Sciences* 2, 438–469.
- 575 Nex, F., Remondino, F., Gerke, M., Przybilla, H.J., Bäumker, M., Zurhorst, A., 2015. Isprs  
576 benchmark for multi-platform photogrammetry. *ISPRS Annals of Photogrammetry, Remote*  
577 *Sensing & Spatial Information Sciences* 2.
- 578 Osfield, R., Burns, D., 2014. Open scene graph. <http://www.openscenegraph.org>.
- 579 Revaud, J., De Souza, C., Humenberger, M., Weinzaepfel, P., 2019. R2d2: Reliable and re-  
580 peatable detector and descriptor, in: *Advances in Neural Information Processing Systems*, pp.  
581 12405–12415.
- 582 Rosten, E., Porter, R., Drummond, T., 2010. Faster and better: A machine learning approach  
583 to corner detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32,  
584 105–119.
- 585 Roth, L., Kuhn, A., Mayer, H., 2017. Wide-baseline image matching with projective view  
586 synthesis and calibrated geometric verification. *PFG-Journal of Photogrammetry, Remote*  
587 *Sensing and GeoInformation Science* 85, 85–95.
- 588 Rublee, E., Rabaud, V., Konolige, K., Bradski, G.R., 2011. Orb: An efficient alternative to sift  
589 or surf., in: *ICCV*, Citeseer. p. 2.
- 590 Schönberger, J.L., Frahm, J.M., 2016. Structure-from-motion revisited, in: *Conference on Com-  
591 puter Vision and Pattern Recognition (CVPR)*.
- 592 Schönberger, J.L., Hardmeier, H., Sattler, T., Pollefeys, M., 2017. Comparative evaluation of  
593 hand-crafted and learned local features, in: *Proceedings of the IEEE Conference on Computer  
594 Vision and Pattern Recognition*, pp. 1482–1491.
- 595 Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M., 2016. Pixelwise view selection for  
596 unstructured multi-view stereo, in: *European Conference on Computer Vision (ECCV)*.
- 597 Shan, Q., Wu, C., Curless, B., Furukawa, Y., Hernandez, C., Seitz, S.M., 2014. Accurate geo-  
598 registration by ground-to-aerial image matching, in: *2014 2nd International Conference on 3D  
599 Vision*, IEEE. pp. 525–532.
- 600 Sibbing, D., Sattler, T., Leibe, B., Kobbelt, L., 2013. Sift-realistic rendering, in: *2013 Interna-  
601 tional Conference on 3D Vision-3DV 2013*, IEEE. pp. 56–63.
- 602 Untzelmann, O., Sattler, T., Middelberg, S., Kobbelt, L., 2013. A scalable collaborative on-  
603 line system for city reconstruction, in: *Proceedings of the IEEE International Conference on  
604 Computer Vision Workshops*, pp. 644–651.
- 605 Vu, H.H., Labatut, P., Pons, J.P., Keriven, R., 2011. High accuracy and visibility-consistent  
606 dense multiview stereo. *IEEE transactions on pattern analysis and machine intelligence* 34,  
607 889–901.

- 608 Waechter, M., Moehrle, N., Goesele, M., 2014. Let there be color! large-scale texturing of 3d  
609 reconstructions, in: European Conference on Computer Vision, Springer, Zurich, Switzerland.  
610 pp. 836–850.
- 611 Wald, I., Woop, S., Benthin, C., Johnson, G.S., Ernst, M., 2014. Embree: a kernel framework  
612 for efficient cpu ray tracing. ACM Transactions on Graphics (TOG) 33, 143.
- 613 Wu, B., Xie, L., Hu, H., Zhu, Q., Yau, E., 2018. Integration of aerial oblique imagery and terres-  
614 trial imagery for optimized 3d modeling in urban areas. ISPRS Journal of Photogrammetry and  
615 Remote Sensing 139, 119–132. doi:<https://doi.org/10.1016/j.isprsjprs.2018.03.004>.
- 616 Wu, B., Zhang, Y., Zhu, Q., 2012. Integrated point and edge matching on poor textural images  
617 constrained by self-adaptive triangulations. ISPRS journal of photogrammetry and remote  
618 sensing 68, 40–55.
- 619 Wu, C., et al., 2011. Visualsfm: A visual structure from motion system .
- 620 Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L., 2019. Recurrent mvsnet for high-resolution  
621 multi-view stereo depth inference, in: Proceedings of the IEEE Conference on Computer Vision  
622 and Pattern Recognition, pp. 5525–5534.
- 623 Yu, Z., Gao, S., 2020. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation  
624 and gauss-newton refinement. arXiv preprint arXiv:2003.13017 .
- 625 Zhang, L., 2005. Automatic digital surface model (DSM) generation from linear array images.  
626 ETH Zurich.