Lattice-Point Mutually Guided Ground-to-Aerial Feature Matching for Urban Scene Images

Xianwei Zheng, Hongjie Li, Hanjiang Xiong, and Xiao Xie

***Abstract*—Ground-to-aerial feature matching bridges information from cross-view images, which enables optimized urban applications, e.g., pixel-level geolocating and complete urban 3-D reconstruction. However, urban ground and aerial images typically suffer from drastic changes in viewpoint, scale, and illumination, together with repetitive patterns. Thus, direct matching of local features between ground and aerial images is particularly difficult because of the low similarity of local descriptors and high ambiguity in true–false match discrimination. For this challenging task, we propose a novel lattice-point mutually guided matching (LPMG) method in this article. We specifically address two key issues: 1) reducing descriptor variance and 2) enhancing true–false match discriminability. The former is solved by recovering the geometry and appearance of the underlying image region in 3-D through automatic view rectification on ground and aerial images. The latter is circumvented by replacing the conventional mismatch removal with an LPMG strategy. In this strategy, the topology structure of repeated façade elements (i.e., lattice), and the high reliable point matching seeds, are first extracted from the rectified ground and aerial images. Then, the point matching seeds guide the self-similar lattice tiles from two views to be precisely aligned, thereby estimating an accurate transformation model from lattice tile correspondences. Finally, the estimated model powerfully supervises the differentiation of true and false matches from the entire putative match set. Extensive experiments conducted on several datasets show that our method can obtain a considerable number of nearly pure correct matches from urban ground and aerial images, significantly outperforming those existing methods.**

***Index Terms*—Aerial oblique imagery, feature matching, ground imagery, ground-to-aerial image matching, repetitive pattern.**

I. Introduction

With the rapid development of unmanned aerial vehicle (UAV)-borne sensors and the popularization of smartphones and consumer-level cameras, images from aerial and street views have become easily accessible. While aerial view images provide a larger and more complete range of top views, street view images delineate points of interest and details of urban façades. Establishing correspondence between street (ground) and aerial view images can facilitate a variety of urban applications that require supplementary information from both views, ranging from geolocating [1], [2], urban 3-D modeling [3]–[5], and autonomous driving [6] to ancient architecture preservation [7]. However, establishing such correspondence is a particularly challenging problem due to the drastic changes in viewpoint, scale, and illumination between ground and aerial images. Occlusion and repetitive patterns common in urban scenarios also complicate the problem [8].

Image matching aims to find correspondence between two or multiple overlapped images, which have been extensively studied under different baselines and scenarios [9]–[14]. The problem of image matching is typically solved by three steps: local feature extraction and description, putative match construction, and mismatch removal. Well localized feature points and strongly discriminative descriptors can bring good input to the subsequent matching stage. For this purpose, a variety of local appearance features and descriptors have been developed [15], [16]. A notable milestone is scale-invariant feature transform (SIFT) [17], which is designed to be invariant to scale and illumination. Its variants, such as speeded-up robust features (SURF) [18], affine SIFT (ASIFT) [19], and perspective SIFT (PSIFT) [20], also attempt to improve its speed or invariance to affine or perspective transformations. These local features can also be assembled to handle different image conditions, as like in MODS [21]. Due to the use of only local appearance for descriptors, putative matches constructed with local features are inevitably contaminated with mismatches.

Thus, mismatch removal is an essential step to discard false matches (outliers) for true matches (inliers), which is usually implemented by restricting matches to satisfy geometric constraints of an underlying image transformation. In the literature, numerous mismatch removal methods have been developed mainly on the basis of statistical regression [22], resampling [23], nonparametric interpolation [24], and graph matching [25]. Among them, resampling-based methods, such as RANSAC and its variants [26], are the most commonly used in engineering applications. Methods based on nonparametric interpolation or graph matching can generally achieve promising performance but usually have high computational complexity. Additionally, by incorporating piece-wise motion smoothness, bilateral function-based motion modeling (BF) [27] and grid-based motion statistics (GMS) [28] also yield good results in single-source wide-baseline matching. However, these typical matching systems are hardly effective on urban ground and aerial images. From a practical point of view, feature descriptors cannot be truly invariant under projective transformations [29].

Manuscript received November 10, 2020; revised January 25, 2021; accepted March 20, 2021. Date of publication March 26, 2021; date of current version May 24, 2021. This work was supported in part by the National Key Research, and Development Program of China under Grant 2018Y-FB0505401, in part by the National Natural Science Foundation of China Project under Grant 41871361 and 42071370. *(Corresponding author: Hanjiang Xiong.)*

Xianwei Zheng, Hongjie Li, and Hanjiang Xiong are with the State Key Lab. LIESMARS, Wuhan University, Wuhan 430079, China (e-mail: zhengxw@whu.edu.cn; lihongjie@whu.edu.cn; xionghanjiang@163.com).

Xiao Xie is with the School of Geodesy, and Geomatics, Wuhan University, Wuhan 430079, China (e-mail: xiexiao1229@163.com).

Digital Object Identifier 10.1109/JSTARS.2021.3069222

When abrupt viewpoint changes occur on two images, perspective deformations can severely undermine the local feature extraction and matching.

To allow view-dependent matching using local features, many studies have aimed to alleviate the difference in view angle. One strategy is to apply view rectification to transform images of arbitrary views into a standard view [30]–[33]. This task can be done by detecting the vertical and horizontal vanishing lines of urban façades to solve the camera rotation that unwarps the view [34]. More generally, *et al.* [29] performed view rectification by using transform invariant low-rank textures (TILT) formed by repetitive patterns, which needs neither 3-D geometry priors nor any intermediate low-level features (e.g., corners or edges). Although view rectification is helpful to reduce descriptor variance, other nuisance factors, such as large variation in scale and appearance, and repetitive patterns in urban ground and aerial images can still cause ambiguities to descriptors. As a result, the putative matches constructed from those cross-view images are usually highly noisy and chaotic. It is therefore particularly hard for existing mismatch removal methods to create a separability constraint for such a putative match set. A different way of matching local features between ground and aerial images is to warp one view into another by using camera pose and 3-D geometric information [3], [4], [10], and performing feature matching on synthetic images. Some of these methods have been successfully applied in practical applications. The disadvantage is that these methods need to reconstruct the 3-D geometry of scenes as prior, which may not be necessary in other lower-level applications.

Instead of attempting to establish accurate point correspondences, researchers have also focused on finding similarities between ground and aerial images to serve a specific downstream application. An active research field is geolocating, which aims to estimate the geolocation of a street-level image based on a database of referenced aerial view images. Considering the rich pattern information of urban façades, many methods have been developed for geolating by exploiting the near-regular and self-similar structures of urban façades for matching [35]–[37]. Bansal *et al.* [35] proposed a method for matching façades between ground and aerial images that relies on the calculation of statistical self-similarity between local patches on a façade. They further devised a scale-selective self-similarity (S_4) descriptor for the self-similar structure of façades for matching [36]. Repetitive façade elements can also be discovered and represented by lattice tiles or motifs [38] and used as a descriptor for façade matching [37], [39]. However, to exploit the regularities of urban façades, these methods usually require an additional segmentation step for accurate façade detection, which is complex and costly. Moreover, their descriptors created from repeated elements can only be assembled for façade-level matching, which are unable to determine a unique tile for matching. In recent years, deep learning-based methods [40], such as WhereCNN [41], CVM-Net [2], and OriCNN [1], have also made great process on ground-to-aerial matching for geolocating. Deep learning-based methods usually require large sample data for training and their performance could be degraded if the data distribution of the matching scenarios is different from those of sample datasets. Furthermore, the methods used in geolocating only achieve image- or façade-level matching, whereas the feature-level correspondences are still missed, which may boost geolating accuracy from an average of 5–20 m level to a pixel level (centimeter) [3].

In summary, direct matching local features between ground and aerial images without 3-D information as prior remains a problem. Basing correspondence solely on local appearance features for urban ground and aerial images is extremely intractable. The main difficulty lies on the inability to find a valid separability constraint from a highly noisy point match set. Lattices extracted from repeated elements of urban façades provide rich structural information that potentially forms a natural complement to local appearance features. Matching/aligning two lattices at tile level can build a compact relationship between two façades, thereby bringing strong structural constraints for reliable image transformation model estimation. In existing works, repetitive patterns are considered to be negative to accurate patch- or feature-level correspondences due to their inherent ambiguity [42]. For example, [36] and [37] exploited the repeated patterns of façades for ground-to-aerial image matching, but lack a solution to determine a unique tile or a local feature for matching. However, based on the spatial relationship between point features and lattice tiles, it is possible to use point features as matching indicators for lattice tiles. Although it is difficult to find all the exact feature matches from a highly noisy putative set at once, obtaining a few correct ones is feasible. Correct matches with feature points that overlap with or are adjacent to lattice tiles from two views are able to link two unique tiles together and therefore align two lattices. As a result, the local feature matching and the unique tile matching can be well coupled and benefit from each other.

To this end, we propose a lattice-point mutually guided feature (LMPG) method for urban ground and aerial images. Given a pair of ground and aerial view images, we first identify the rough building façades with an efficient local feature classification. By extracting and recovering the low-rank textures of urban facades, view rectification is performed on both view images to transform them into an orthorectified view, thereby making them more matchable. Then, the two key ingredients, lattices of building façades and point matching seeds, are obtained from the orthorectified images. To meet the quality requirement for subsequent matching, we incorporate the horizontal and vertical edge features from the orthorectified images to discover lattices with a good structure. Moreover, two simple yet effective constraints, i.e., a repeatability constraint (RC) and a triangular constraint (TC), are applied to ensure strict filtering on the putative match set, thereby generating point matching seeds with high reliability but are few in number. The mutually guided feature matching is accomplished after being equipped with well-structured lattices and point matching seeds. One point matching seed is first selected as an indicator to link two unique tiles, respectively, from two views, hence achieving a coarse alignment of two lattices. A refined alignment of two lattices is conducted by minimizing the total transformation error of all point matching seeds, which is computed by shifting one tile to the other pixel by pixel. During this process, the transformation model between two ground and aerial images are progressively estimated and optimized. Eventually, the resultant model imposes geometric constraints on the putative set to discriminate between all true and false matches.

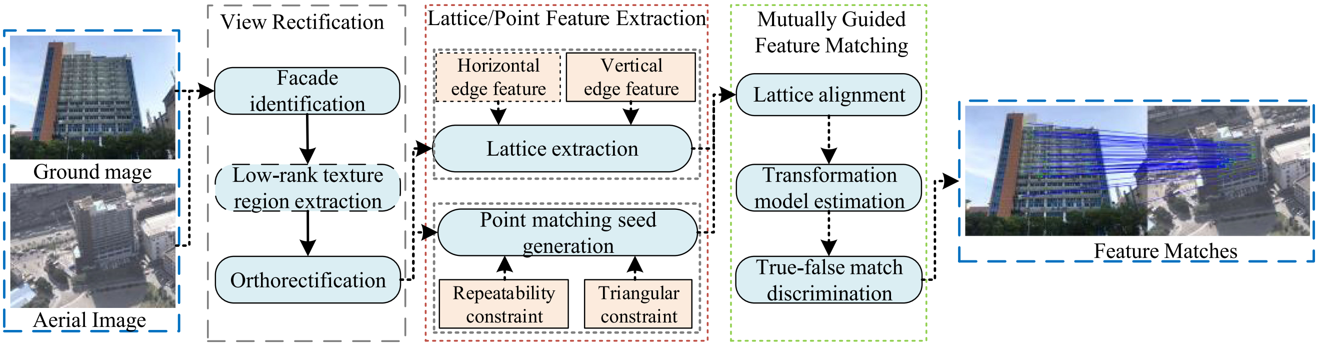


Fig. 1. Workflow of the proposed method.

The rest of this article is organized as follows. The methodology is described in Section II. Experimental results are presented in Section III, and discussion is provided in Section IV. Section V concludes this article.

II. Methodology

This study aims to exploit the rich pattern information of building facades to assist in the task of ground-to-aerial feature matching in urban areas. The main workflow of the proposed matching pipeline is shown in Fig. 1. The input consists of a pair of ground and aerial view images, and the output consists of the reliable feature matches directly obtained from raw images. First, view rectification is performed on both view images (see Section II-A). Second, the ortho-rectified images from both views are processed, which consist of the extraction of lattices from building facades (see Section II-B), and the generation of point matching seeds (see Section II-C). Finally, the lattices of building facades and point matching seeds coupled to undertake the mutually guided feature matching are described in Section II-D.

A. View Rectification

In our matching pipeline, view rectification serves two important purposes. One is to transform images of arbitrary views into an orthorectified view, thereby recovering the geometry and appearance of the underlying planar region in 3-D, which allows corresponding features from two views to have more similar descriptors. The other is to facilitate the use of horizontal and vertical edge features for well-structured lattice discovery from orthorectified views, thereby enabling the precise alignment of two ground and aerial tiles/lattices. Among all view rectification methods, the TILT-based implementation is highly robust to significant image deformation and corruption, which needs neither 3-D geometry priors or any intermediate low-level features but only a user-specified low-rank texture region as input. For further details, readers can refer to [29]. In this study, we first use an undirected feature graph to identify local features representing repetitive patterns of building facades. Then, we perform a refined clustering on the identified features to obtain such a low-rank texture region.

A key insight is that local features that correspond to the same repetitive pattern are spatially close and share a similar appearance and scale [43]. Based on this observation, we are able to group the local features representing different kinds of repetitive patterns. This task is performed by first constructing an undirected feature graph $G_f = (V, E_f)$ on local SIFT features extracted from a given ground or aerial image. In the graph, $V = \{(x_i, s_i, d_i)\}_{i=1}^{N}$ denotes N vertices of graph G_f. Each vertex v_i in V consists of feature location x_i, scale s_i and corresponding descriptor d_i, and E_f represents the edges. In the initial stage, E_f is fully connected. Then, the graph is segmented by applying three conditions to determine the connectivity on the edges, as illustrated in Fig. 2. In Fig. 2, conditions ① and ② constrain that two connected vertices v_i and v_j are spatially close and with similar scales, where c in condition ① is a constant set to 10 [43]. Conditions ③ constrains two connected vertices v_i and v_j share at least one visual word, where K_i, K_j are the top $K = 50$ nearest visual words assigned to d_i and d_j (descriptors of v_i and v_j) from a precomputed visual vocabulary of visual words provided in [43].

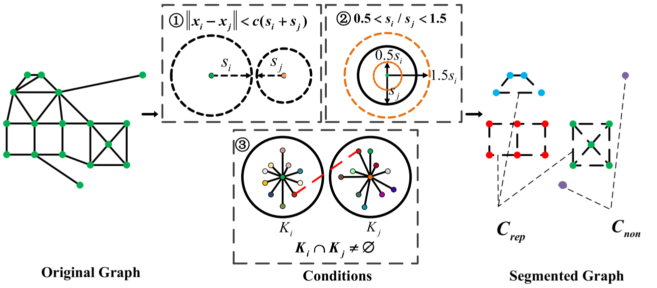


Fig. 2. Feature graph segmentation for different types of repetitive patterns.

After disconnecting the edges that do not satisfy these three conditions, all the vertices in G_f are segmented into a set of disjoint groups representing various types of repetitive patterns, and a number of isolated points representing nonrepetitive patterns. Generally, the dominant building façade in an image has a higher repetition intensity than other objects, and can receive more vertices in the segmented graph. Thus, for these disjoint groups, the one with the highest number of vertices can be identified as the repetitive pattern corresponding to a dominant facade, which is denoted as $G_{\text{rep},f}$. The rectangular region covered by $G_{\text{rep},f}$ is then defined as the rough facade region denoted as R_f. In practice, for the very high resolution images, objects with dense repeated elements, such as trees and grassy areas, may receive much more vertices than facades. In this case, when G_f is

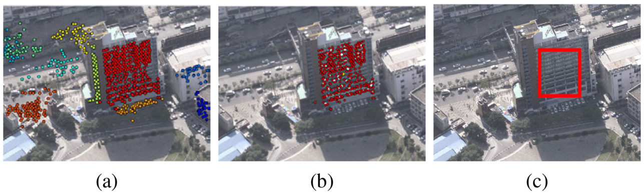


Fig. 3. Example of façade identification and low-rank texture region detection. (a) Repetitive patterns identified by local features grouped with different colors. (b) Mean-shift clustering on local features identified as façade (white points are obtained cluster centers). (c) Extracted candidate low-rank texture region.

presegmented with a small $K = 10$, features (vertices) obtained from facades cannot be grouped, whereas features (vertices) from objects with dense repeated elements can be grouped and thus removed in advance.

In general, R_f is too large to be used as the input window for TILT because it may contain considerable noise and background information that disturb the solving of an accurate transformation matrix for low-rank texture recovery. However, selecting a small region from R_f may lead to limited information, which is insufficient to calculate a transformation matrix representing the entire facade. It is also not feasible to use a predefined window with fixed size for facades with varied shapes. To select an appropriate candidate low-rank texture region that is adaptive to different facades, we first select a window W with its size as 2/3 width and length of R_f (This size helps to keep the selected window has a similar shape to building facade and contains enough information for TILT-based rectification). Then, we find a suitable location as the window center. As R_f only defines the rough building facade area, direct assigning its center point as the center of W may cause W to exceed the facade boundary. Thus, we perform a refined mean-shift-based clustering [44] on $G_{\text{rep},f}$ to find a location that is close to the center of the actual building façade. We use a default bandwidth of 20 to classify the vertices of $G_{\text{rep},f}$ into a set of smaller clusters, with their cluster centers denoted as $C = (c_1, c_2, \ldots, c_n)$. Then, we calculate a location P_c by averaging the coordinates of C as follows:

$$P_c = \left(\sum_{i=1}^{n} c_i\right) / n \tag{1}$$

where P_i is is assigned as the center of W, and we denote the image region covered by W as R_c. Using R_c as the input candidate low-rank texture region, TILT simultaneously recovers the intrinsic low-rank texture and the unknown image transformation, thereby transforming an arbitrary view into an ortho-rectified view. An example of rough facade identification and low-rank texture region detection is presented in Fig. 3.

B. Lattice Extraction

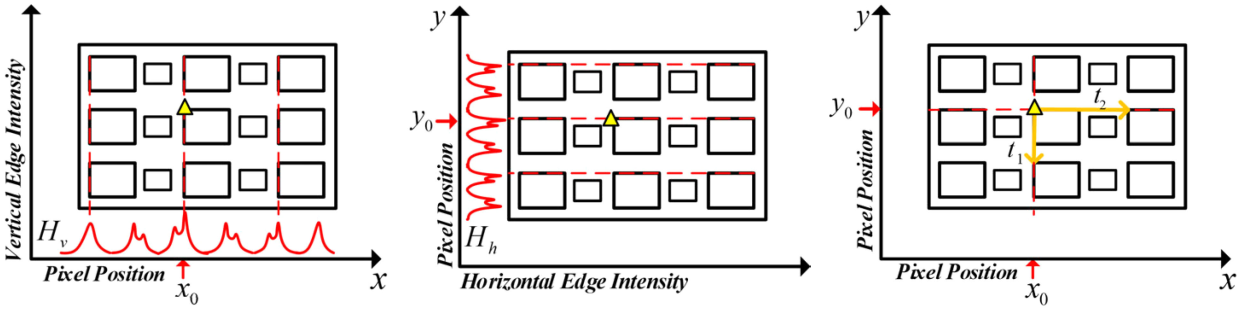
Building façades are commonly a type of near-regular and translationally symmetric texture that consists of multiple repeated pattern elements (e.g., windows). Previous work [45]–[47] has demonstrated that the structure of any of such texture can be generated by a pair of shortest vectors (t_1, t_2), saying that the orbits of (t_1, t_2) form a 2-D quadrilateral lattice representing the repeated texture elements and their topological structure. Their idea for lattice detection is mainly based on proposing an initial lattice by seeking t_1 and t_2 neighbors with point detector, and subsequently growing this initial lattice. However, using randomly distributed point features to search for t_1 and t_2 is prone to producing lattices with deformed structures and is unstable in the number of extracted tiles. In our task, the extracted lattice should correctly represent the structure and topology of repetitive pattern of a facade, otherwise it could be difficult to establish one-to-one tile correspondence between lattices from two views. Moreover, for large building facades, the extracted lattice should contain as many repeated elements as possible. This condition guarantees that lattices from two views can have overlapped tiles, thereby facilitating the correspondence model estimation.

For high-quality lattice extraction, we seek a pair of (t_1, t_2) lattice generating vectors from the ortho-rectified facades by incorporating the edge features distributed along horizontal and vertical directions, and generate the initial lattice by a periodic translation of this (t_1, t_2). As the accurate façade region is unknown, we use the recovered candidate low-rank region R'_c as an input for initial lattice generation. For simplicity, vectors t_1 and t_2 can be viewed as the vertical and horizontal edge of an initial lattice tile. The detailed procedure of initial lattice generation is as follows:

(1) Generating edge intensity histograms. Extract the binary edge map of R'_c with canny operator, and generate the edge intensity histograms along the vertical and horizontal directions. As illustrated in Fig. 4, the red curves in (a) and (b) are the generated vertical and horizontal edge intensity histograms, respectively, denoted as H_v and H_h. The H_v and H_h are actually two vectors and each of their element records the number of edge pixels at a specific x or y position.

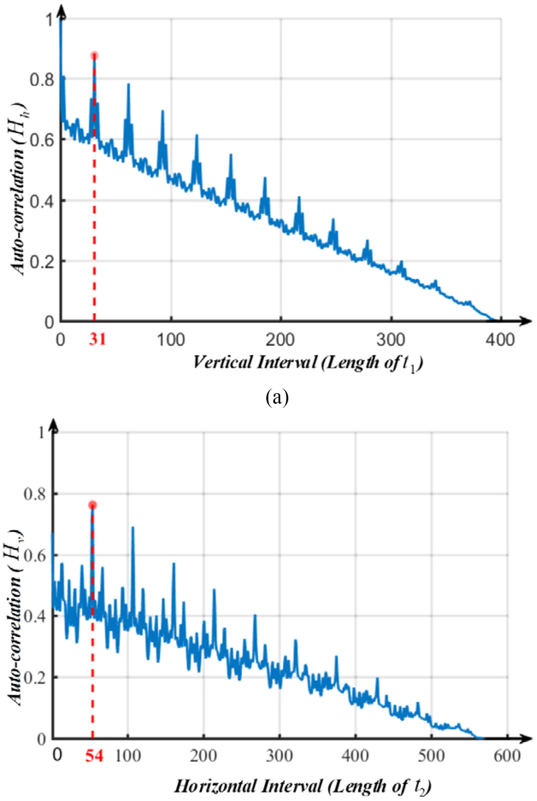
(2) Finding starting point of vectors t_1 and t_2. Find two peaks from the vertical and horizontal edge intensity histogram, each having a maximum intensity value in the corresponding histogram; use the positions of these two peaks to determine a point as the starting point of vectors $t_1$and t_2. As shown in Fig. 4(a) and (b), x_0 and y_0 are the positions of the two maximum peaks on H_v and H_h, respectively. Here, due to the disturbance of noise and occlusion, we consider the position where a peak with a maximum value localizes the boundary of a desired repeated element with the highest probability. Thus, two localized boundaries determine a corner point of that repeated element (their intersection point), which is used as the starting point of vectors t_1 and t_2, as point (x_0, y_0) shown in Fig. 4(c).

(3) Determining the vector lengths of t_1 and t_2. Generate the vertical and horizontal autocorrelation histograms for H_v and H_h (this can be simply implemented by inputting the two vectors into MATLAB autocorr function); find two smallest intervals with maxima autocorrelation from the two autocorrelation histograms, and assign the two intervals as the lengths of vectors t_1 and t_2. Note that multiple small repeated elements can form a larger repeated unit. Thus, repeated units with different sizes (lengths and widths) can have different repetition intervals along vertical and horizontal directions. As shown in Fig. 5, the local maxima points in vertical or horizontal autocorrelation histograms indicate the intervals of different repeated units. Among these local maxima points, the one with the smallest interval implies the interval of a smallest repeated unit, as the



(a) (b) (c)

Fig. 4. Finding starting point of t_1 and t_2 from edge intensity histograms: (a) and (b) are vertical and horizontal edge intensity histograms, and (c) starting point of vector t_1 and t_2 (yellow triangle).



(a)

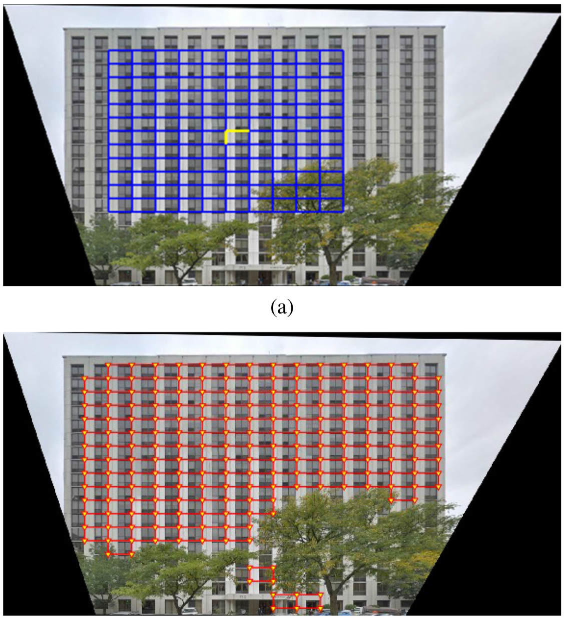
(b)

Fig. 5. Finding repetition intervals of vector t_1 and t_2 from edge autocorrelation histograms: (a) and (b) are autocorrelation histograms for vertical and horizontal edge intensity histograms, respectively.

red dot lines indicate in Fig. 5(a) and (b). These two intervals are then assigned as lengths of vectors t_1 and t_2.

(4) Constructing the initial lattice. After steps (1)–(3), a pair of initial (t_1, t_2) lattice generating vectors is obtained. Then, we use this (t_1, t_2) to produce an initial lattice tile, and periodically translate this tile to generate a lattice with m rows and n columns that cover low-rank texture region R_c'.

Once the initial lattice is generated, a heuristic mean-shift belief propagation [47] is used to simultaneously grow and refine the lattice tiles. An example of the initial lattice and the lattice after refinement and expansion is shown in Fig. 6, where the two perpendicular yellow lines indicate the initial (t_1, t_2) lattice lattice generating vector.



(a)

(b)

Fig. 6. Example of lattice extraction: (a) extracted initial lattice where the two yellow lines represent the initial (t_1, t_2) lattice-generating vectors and (b) lattice after refinement and expansion.

From Fig. 6, we can also observe that tiles in the initial lattice overlapped with trees are rejected because of the unaccepted appearance similarity between those tiles and the initial tile. It is necessary to exclude such tiles because the occluded façade region may not contain the same repeated elements. Within the lattices extracted from ground and aerial façades, we fine-tuned the rectified aerial image to force each pair of ground and aerial image to have the same length-to-width ratio. Suppose that (t_1^g, t_2^g) and (t_1^a, t_2^a) are lattice generating vectors of a pair of ground and aerial images, and we stretch the aerial image along the horizontal direction with a factor $(\|t_1^a\|\|t_2^g\|)/(\|t_2^a\|\|t_1^g\|)$, where $\|\cdot\|$computes the moduli of a vector. This operation can make the geometry and appearance of two ground and aerial

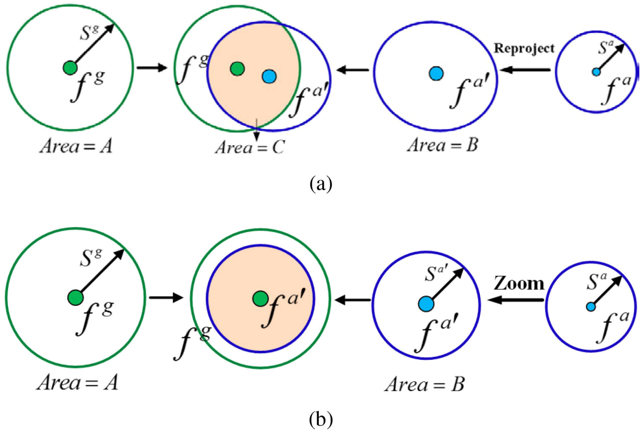


Fig. 7. Local feature repeatability calculation: (a) original repeatability calculation and (b) modified repeatability calculation.

facades more similar, thereby help to mitigate the descriptor difference between two correspondence features, except for scale difference.

C. Point Matching Seed Generation

As matching lattice tiles from two views is inherently ambiguous due to their self-similar structure, we generate a small number of highly reliable point matching seeds from the putative set, and use them as a guidance for corresponding lattice tiles. Given a pair of orthorectified ground and aerial images, we distill such matching seeds by first constructing a putative match set, *e.g.,* ASIFT, with a nearest-neighbor matching on extracted local features, and then strictly filtering the putative matches (after ratio test) with two simple yet effective constraints, that are, a RC and a TC.

RC. Given a pair of local features f^g and f^a from two overlapped images, as illustrated in Fig. 7(a), each consisting of a keypoint location and a circle region with radius equal to its scale, the overlap error is $O_e = 1 - C/A \cup B$. Here, A, B, C, represent the area of f^g, $f^{a'}$ (reprojected f^a) and the overlapped region of f^g and $f^{a'}$.

According to [48], local feature repeatability informs that two features of a correct match should have an overlap error less than 0.4 when reprojecting one feature from an image onto an matched image. Mishkin *et al.* [49] argued that local feature repeatability is too strict a condition for successful matching because the affine adaption procedure can cause the drop in the number of correct matches. However, this condition is suitable for our task, which aims to strictly filter the putative matches to obtain the high reliable ones. The problem is that the overlap error calculation requires a known transformation model for reprojection operation, which is to be solved subsequently. To address this problem, we adjust the condition of local feature repeatability as a simple RC. In Fig. 7, we can observe that the keypoint location and shape/area of the reprojected feature actually determines this overlap error. In our case, each pair of ground and aerial images is transformed into an orthorectified view and their façades are stretched to maintain the same length-to-width ratio. Thus, reprojecting a feature from an aerial image into a ground image does not change much in the shape of its descriptor support region, and vice versa. As shown in Fig. 7(b), with two features f^g and f^a of a match, the adjusted overlap error is calculated as

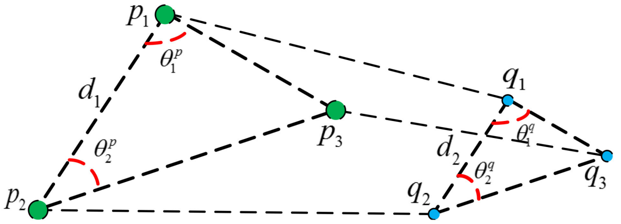


Fig. 8. Triangular constraint.

$$O_e = 1 - B/A \cup B = 1 - \frac{\pi(s^{a'})^2}{\pi(s^g)^2} \tag{2}$$

where $S^{a'} = S^a\|t_1^g\|/\|t_1^a\|$, S^g, S^a are feature scales of f^g, f^a, and t_1^g, t_1^a are t_1 lattice generating vector sought from a given pair of ground and aerial images. Although RC does not consider the real location of the reprojected features, it is able to incorporate feature area requirements early on to reject a large number of unwanted matches.

TC. To obtain the desired point matching seeds, we apply a TC to refine the matches filtered by RC, which is partially based on feature motion consistency theory [50]. The creation of an exact motion model for correspondence between ground and aerial images is complicated. Here, we instead use a joint topological relation of a group of three matches to impose constraints on feature motion smoothness. As illustrated in Fig. 8, we suppose that (p_1, q_1), (p_2, q_2), and (p_3, q_3) constitute a group of three matches after filtering by RC, and p_1, p_2, p_3 and q_1, q_2, q_3 are the corresponding feature points from ground and aerial images, respectively. Thus, TC restricts that the three feature points from the same image are not all inside the same tile and noncollinear. The constructed triangles $T(p_1, p_2, p_3)$, $T(q_1, q_2, q_3)$ should satisfy the following two conditions:

$$\left\lfloor \frac{d_1}{d_2} \right\rfloor = \left\lfloor \frac{\|t_1^g\|}{\|t_1^a\|} \right\rfloor ; \lfloor \theta_1^p \rfloor = \lfloor \theta_1^q \rfloor , \lfloor \theta_2^p \rfloor = \lfloor \theta_2^q \rfloor \tag{3}$$

where d_1 and d_2 represent the length of edge p_1p_2 and q_1q_2, and $\|t_1^g\|$, $\|t_1^a\|$ compute the moduli of vector t_1 sought in the step of lattice extraction (i.e., the length of a lattice tile) for the ground and aerial images, respectively. $\theta_1^p, \theta_1^q, \theta_2^p, \theta_2^q$ are the angles at the corresponding vertices. $\lfloor g \rfloor$ is the floor function. The number of outliers is greatly reduced after filtering by conditions in (3). In this manner, RANSAC is capable of creating a geometrical constraint for discarding outliers while retaining inliers. Therefore, TC also involves a RANSAC to finally determine the point matching seeds.

D. Mutually Guided Feature Matching

Within the obtained lattices and point matching seeds, the LMPG can be performed. This condition is started by establishing the tile correspondence under the guidance of point matching seeds, followed by progressively estimating the optimized transformation model from tile correspondences, and finished with

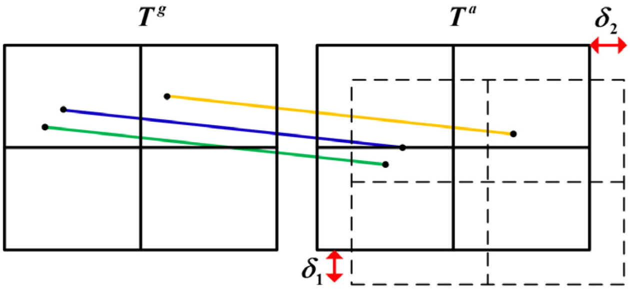


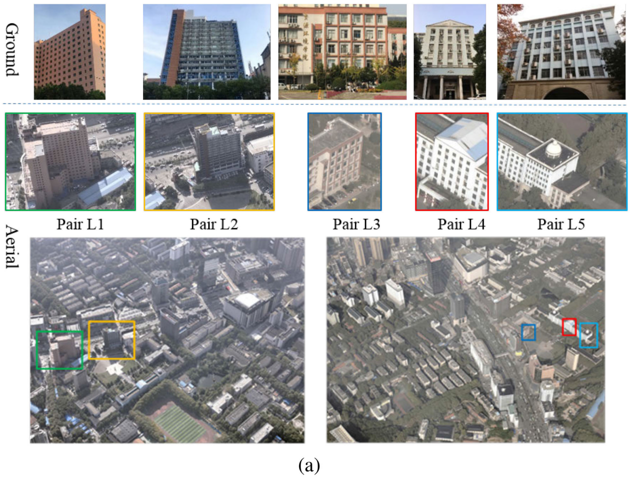
Fig. 9. Shift error between two ground and aerial lattices after coarse alignment.

identifying all true matches with the supervision of the estimated model.

Based on the topology of lattice structure, only a pair of tiles being matched is enough to derive the complete tile correspondences between two lattices. To find such a tile match, we apply a coarse-to-fine alignment strategy. Here, for convenience, we denote the collection of the point matching seeds as M_{seed}, and denote the feature points of M_{seed} from a ground image I^g and an aerial image I^a as F^g and F^a, respectively. We first search for one point match from M_{seed}, with their feature points overlapped with two tiles, respectively, from two views. We denote this pair of tiles as (T_i^g, T_i^a). Here, in the extreme condition, if the point matching seeds do not overlap with any of the extracted lattice tiles, we can perform a nearest neighbor search to obtain such a pair of tiles. Due to the large-scale difference between the ground and aerial images, the lattices extracted from I^g and I^a usually exhibit a certain shift on tile boundaries along the vertical and horizontal directions, as indicated by the errors δ_1 and δ_2 in Fig. 9.

Therefore, when the point match from M_{seed} is directly used as the matching indicator, a tile may be wrongly matched with a neighbor of its correct match. To solve this problem, we use a tile distance index (TDI) to allow a pair of tiles to be matched with high probability. This TDI is defined as the total transformation error of all point matching seeds when using a homographgy model computed from corner point correspondences of (T_i^g, T_j^a). The smaller the TDI, the higher is the probability that the two tiles are matched. Given a tile T_i^g, it is able to find a tile T_x^a from T_j^a and its eight neighbors that has a minimum TDI with T_i^g. We regard this T_x^a as a correct match of T_i^g.

Note that identifying (T_i^g, T_x^a) only builds a coarse alignment between two lattices because the deviations between corresponding corners or boundaries of T_i^g and T_x^a could still exist. Thus, based on minimizing the transformation error of the point matching seeds, a refined alignment is further applied on the two tiles T_i^g and T_x^a. This is done by shifting T_x^a along the vertical and horizontal directions pixel by pixel, and finding a location that derives a minimum TDI for T_i^g and T_x^a. We find that the TDI of T_i^g and T_x^a decreases with an increase toward shifting along a certain direction in the range of half-length or width of T_x^a. Thus, we can easily find a location that achieves a minimum TDI for T_i^g and T_x^a. During the process of coarse-to-fine tile alignment, the transformation model between two façades is actually progressively estimated and optimized until a minimum TDI is obtained. As the point matching seeds are extremely few and the number of searched tiles in coarse matching is only nine, the entire coarse-to-fine tile alignment is highly efficient. Transformation model computed from lattice tile correspondences is generally highly reliable because of the strong restriction imposed by the lattice structure. We denote the resultant model as $H_{a\to g}$ and use it to filter the entire putative set for true matches. For matches that satisfy $H_{a\to g}$ with a transformation error, no more than three pixels [49] are classified as true matches. We have to emphasize that when two lattices are well aligned, the corner points of their overlapped tiles are also established correspondences. Thus, correspondence from the lattice corner points and local feature points together determine the final point correspondences for a pair of ground and aerial images.



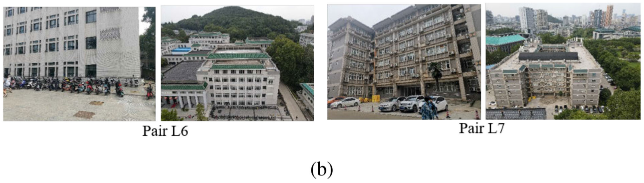


Fig. 10. Local dataset. (a) Five ground and aerial image pairs with different architecture types and visual appearances of buildings from local dataset I. The first row shows street-level building images collected by a mobile phone, and the second row shows the respective aerial oblique images. In the third row, rectangles in different colors indicate the locations of the aerial images. (b) Two image pairs from local dataset II.

III. Experimental Results

In this section, we provide the experimental results and comparisons with other sophisticated methods to evaluate the effectiveness of the proposed LPMG. The datasets are first described, and then the results for the different datasets are presented.

A. Datasets

We first employed two challenging local datasets (termed as local dataset I and II) obtained from Wuhan University in the experiments, as shown in Fig. 10. The local dataset I contains a set of oblique images captured by UAVs and a number of street-level view images collected by a mobile phone. For the purpose of detailed evaluation, five ground images (pair L1–L5), which contain buildings with different heights, architectural

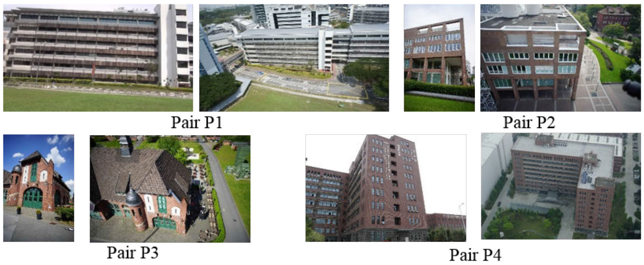


Fig. 11. Four ground and aerial image pairs with diverse buildings from public dataset.

styles, and degrees of occlusion are selected in our experiment. The corresponding aerial image regions are manually cropped from airborne oblique images. Fig. 10(a) shows the selected image regions, in which the drastic difference between ground and aerial images, in terms of image resolution, viewpoint, and illumination, are clearly visible. Furthermore, the temporal and sensing disparities make those ground and aerial image pairs visually different, bringing extra difficulties in matching with local appearance features. The local dataset II contains two image pairs that are captured in two architectural scenes of Wuhan University, as shown in Fig. 10(b). The aerial images are captured at a relatively close range. For both datasets, the repeated windows of building façade are obvious.

To further evaluate the generalization ability of the proposed LMPG in different scenes, we also included those public datasets in experiments. As mentioned in [10], few large-scale publicly available datasets have ground and aerial images for the same architectural scene. Many released datasets are specifically meant for the task of urban 3-D reconstruction, and each dataset only contains complete ground and aerial image pairs for one building. To include diverse buildings for evaluation, we collected four image pairs from four public datasets, as shown in Fig. 11. The image pair P1–P4 is from BF [27], Centre of Dortmund [51], Zeche of Zurich [51], and SWJTU-BLD [4], respectively. In Fig. 11, the drastic variation in viewpoint between ground and aerial images is visible. However, for the purpose of optimized 3-D reconstruction, those images are captured at a relatively close range; thus, both the ground and aerial images have a high resolution.

B. Results on Local Dataset

In this section, we present some relevant results of LMPG conducted on the local dataset, which include some intermediate results (from local datast I) for different steps of LMPG and the final ground-to-aerial feature matching results and comparisons.

1) Results of Different Components of LMPG: As a necessary step, view rectification is first performed on both the ground and aerial images. Some results are shown in Fig. 12, where the first column is the different repetitive patterns obtained with SIFT feature classification; the second column shows the extracted candidate low-rank texture region (red window) and recovered texture region returned by TILT (green window); and the third column displays the orthorectified images after view rectification. In Fig. 12, by extracting the reasonable candidate low-rank texture regions from rough building façades, TILT can successfully recover the underlying image transformation and rectify the images of arbitrary views into orthorectified views. Thus, the large difference in viewpoint between ground and aerial images is corrected.

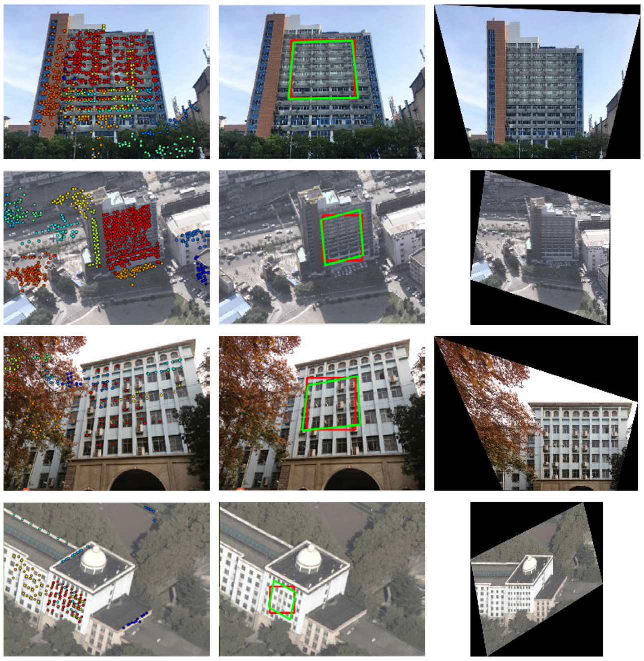


Fig. 12. Results of view rectification. The first column is the repetitive patterns identified by local features with different colors; the second column displays the original (in red) and recovered low-rank texture regions (in green); and the third column shows the orthorectified images.

After view rectification, lattices are extracted from the orthorectified views of the ground and aerial images. To evaluate the effectiveness and necessity of the proposed solution for lattice extraction, the method developed by [47] as a representative is also selected for comparison. The extracted lattices were projected back into the original view to allow good inspection. Some examples are presented in Fig. 13. As shown in the first row of Fig. 13, lattices extracted by [47] cannot meet the demand for the subsequent ground-to-aerial feature matching. For example, the two lattices in the first and second columns only cover different small parts of the same building façade, which have no overlapped tiles for establishing correspondences. Lattices in the third and fourth columns are seriously distorted, and are also difficult to align finely. The results from [47] are not surprising because their method is not specially designed for urban building façades; it does not consider the pattern regularities of building façades but only uses the scattered point features to seek t_1 and t_2 neighbors, which are unstable in delivering well-structured lattice tiles. Taking advantage of the good initial region for lattice discovery and the exploiting pattern information of building façades, our solution can extract high-quality lattices with a more reasonable structural representation of repeated façade elements.

In our study, point matching seeds are adopted as matching indicators for aligning lattice tiles from two views. As a result, it is crucial to extract a few but highly credible matching seeds from noisy putative matches. To verify the effectiveness of the two constraints, i.e., RC and TC, for point matching seed generation, some visual effects of filtering putative matches with RC and TC

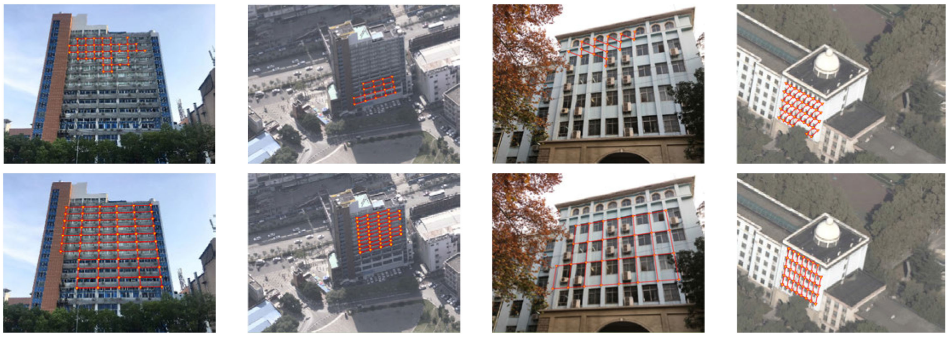


Fig. 13. Qualitative comparison results of lattices extracted by different methods. The first and second rows are lattices derived by [47] and the proposed method, respectively.

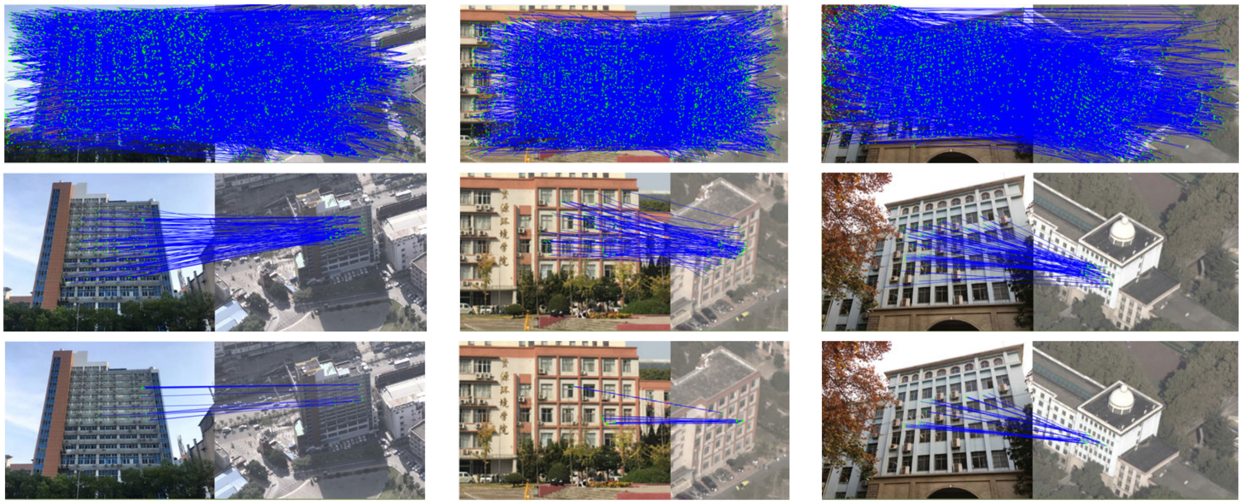


Fig. 14. Analysis of RC and TC for point matching seed generation. From the first to last row are initial putative matches, matches after RC and TC filtering, respectively.

are presented in Fig. 14. In all our experiments, we use the ASIFT feature for putative match construction. Acting on highly noisy putative matches, RC helps to filter out most significant outliers, while TC further eliminates matches that have an inconsistent motion with inliers. Considering the results after TC, we can observe that although the finally obtained point matching seeds are few in number, they are highly reliable for use as matching indicators for lattice tiles.

2) Matching Results on Local Dataset I: Based on the extracted lattices and point matching seeds, the final mutually guided feature matching is conducted on each pair of the selected ground and aerial images. The qualitative results are presented in Fig. 15, which show that the proposed LMPG can be successfully applied to urban ground and aerial images, and achieved promising matching performance consistently on image pairs with different conditions.

We also tested a considerable body of existing matching methods on the selected datasets. However, most of them almost completely failed. To help evaluate the proposed method, we selected for comparison the sophisticated matching methods GMS [28], SparseVFC [24], and BF [27], which have available implementations. To ensure fairness, all methods use the ASIFT feature for matching, and the common ASIFT matcher that uses RANSAC for mismatch removal is also included for comparison. For those compared methods, a ratio test at standard 0.66 ASIFT threshold is conducted after the nearest-neighbor matching. The comparison results are listed in Table I, where correct matches are manually checked.

TABLE I
QUANTITATIVE COMPARISON RESULTS OF DIFFERENT METHODS

Method / Image	ASIFT	GMS	SparseVFC	BF	LPMG(ours)
	N_c/N_t	N_c/N_t	N_c/N_t	N_c/N_t	N_c/N_t
Pair L1	0/21	0/112	7/63	42/349	304/305
Pair L2	0/20	0/169	0/54	0/81	113/114
Pair L3	17/23	0/10	0/47	2/170	70/72
Pair L4	0/26	0/64	1/103	10/393	78/78
Pair L5	0/24	0/0	0/34	15/48	154/157

N_c denotes the number of correct matches, and N_t represents the total number of obtained feature matches.

The results in Table I show that the ASIFT matcher obtains no correct matches in most cases except in image pair L3, demonstrating that basing correspondence solely on local features for urban ground and aerial images is intractable. Similar to ASIFT, GMS achieves no correct matches in

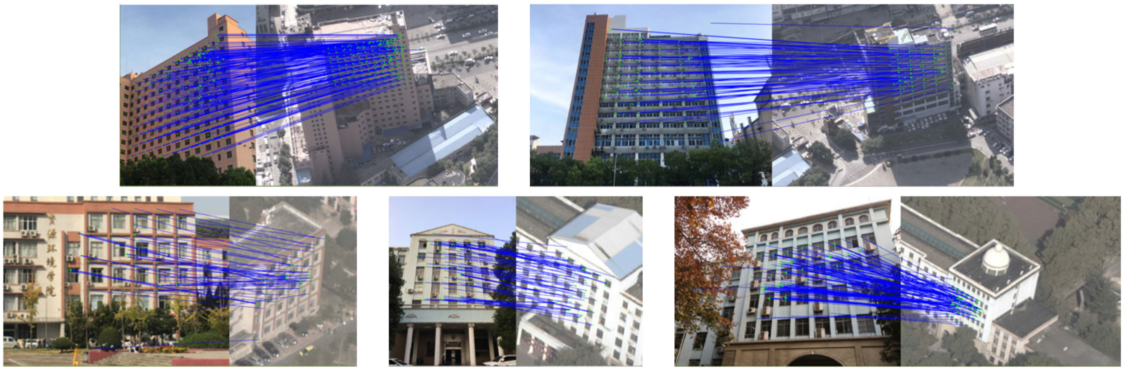


Fig. 15. Matching results obtained by the proposed LMPG on local dataset I.

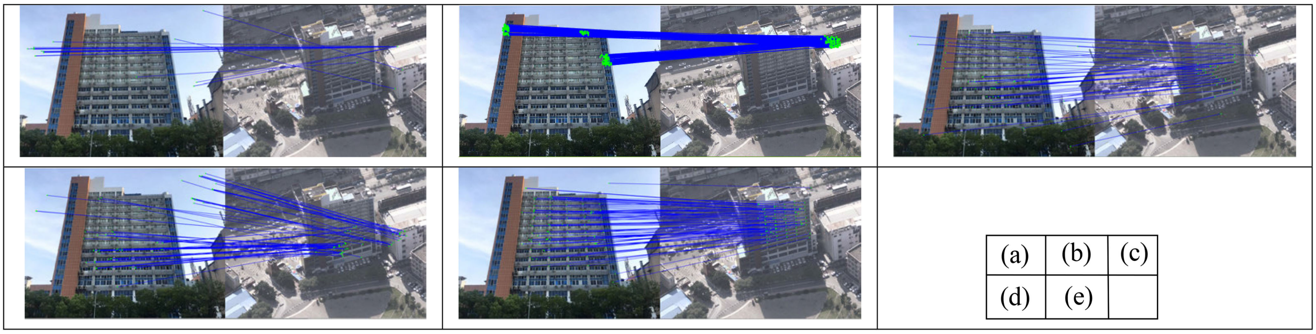


Fig. 16. Qualitative comparison results for image pair L2 under original view: (a) ASIFT, (b) GMS, (c) SparseVFC, (d) BF, and (e) proposed LPMG.

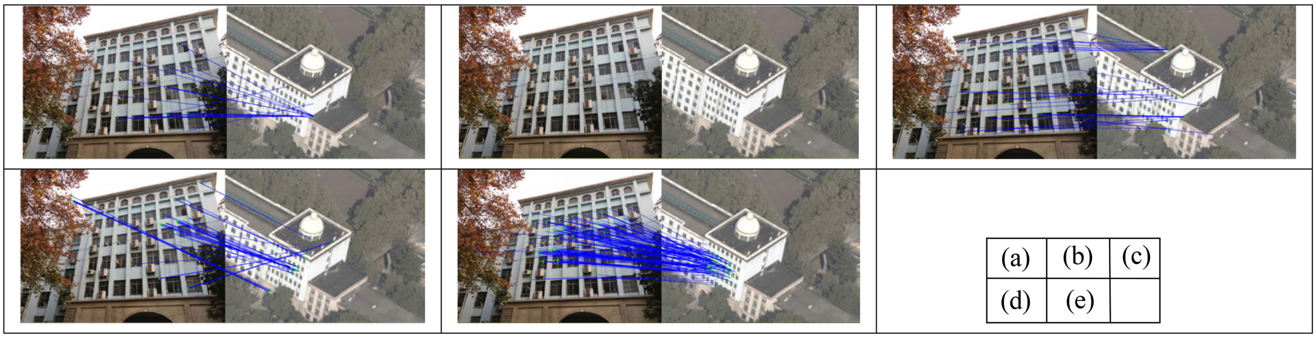


Fig. 17. Qualitative comparison results for image pair L5 under original views: (a) ASIFT, (b) GMS, (c) SparseVFC, (d) BF, and (e) proposed LPMG.

all cases probably because the local grid-based motion smoothness is miscomputed in the presence of large perspective deformations and repetitive patterns. The SparseVFC that bases correspondence on the estimation of vector field consensus of true matches can only obtain very few correct matches in image pairs L1 and L4. By incorporating global motion consistency, BF achieves better performance than the former three methods in terms of the obtained correct matches, but the matches are still few in number. In other words, when the descriptor similarities weaken significantly in case of ground-to-aerial feature matching, all the compared methods seem unable to differentiate true and false matches. Contrary to these existing methods, the numeric results of the proposed LMPG are deemed good for all the five pairs of ground and aerial images. Taking advantage of view rectification, we can mitigate the potential deformations of ground and aerial images, thereby making images from the two views more matchable. Moreover, the obtained nearly pure correct matches reveal the strong discriminability of the transformation model that results from matching in a lattice-point mutually guided manner, which we consider as the key to successful ground-to-aerial feature matching.

The visual effects of different matching methods are shown in Figs. 16 and 17, where results from image pairs L2 and L5 are selected as a reference (with aerial images from two different large oblique images). It can be seen that the qualitative results of different comparison methods are visually consistent with the

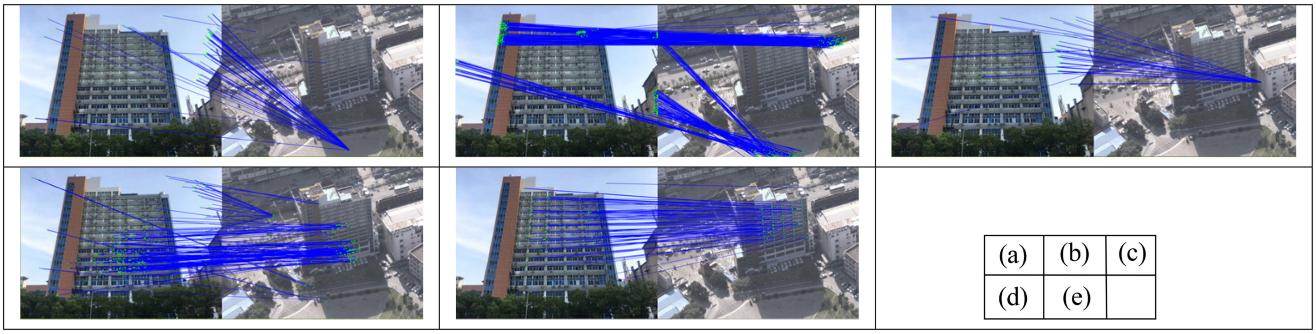


Fig. 18. Qualitative comparison results for image pair L2 under rectified views: (a) ASIFT, (b) GMS, (c) SparseVFC, (d) BF, and (e) proposed LPMG.

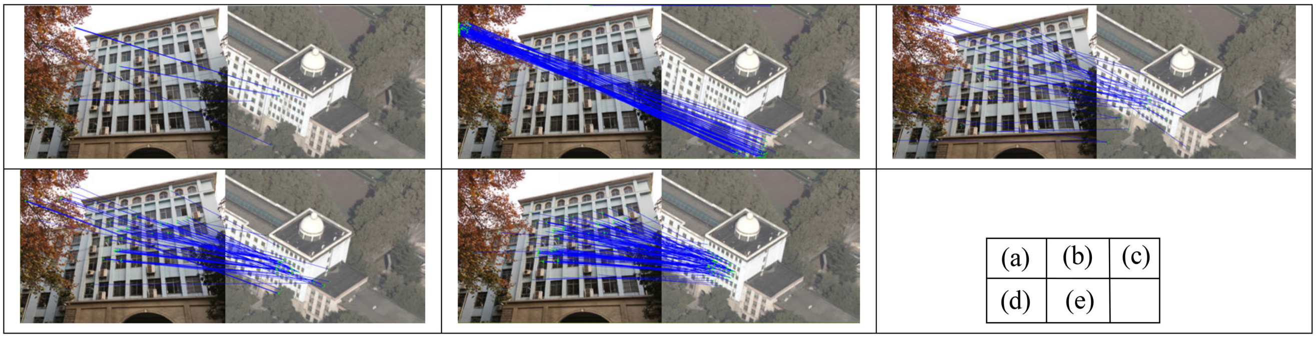


Fig. 19. Qualitative comparison results for image pair L5 under rectified views: (a) ASIFT, (b) GMS, (c) SparseVFC, (d) BF, and (e) proposed LPMG.

TABLE II
QUANTITATIVE COMPARISON RESULTS OF DIFFERENT METHODS ON RECTIFIED VIEWS

Method / Image	ASIFT	GMS	SparseVFC	BF	LPMG(ours)
	N_c/N_t	N_c/N_t	N_c/N_t	N_c/N_t	N_c/N_t
Pair L1	22/23	0/31	26/366	63/521	304/305
Pair L2	0/66	0/348	0/58	0/315	113/114
Pair L3	23/24	0/12	1/98	23/399	70/72
Pair L4	0/13	0/47	1/34	28/418	78/78
Pair L5	2/10	0/207	10/40	62/91	154/157

quantitative results in Table I. The erroneous matches from the results of the compared methods are clearly visible, especially for ASIFT and GMS.

Considering the uncertainty of whether existing methods can effectively function after view rectification, we further apply the four other methods on the rectified views to form a comprehensive comparison. The quantitative results are listed in Table II.

In Table II, most methods can witness a slight improvement in terms of the number of obtained true matches in some image pairs after view rectification. However, the obtained true matches are still few and the false matches increase rapidly. Thus, the analysis results imply that finding a powerful separation model for matches constructed from ground and aerial images is indispensable even when view rectification is employed. The qualitative comparison results from orthorectified views for image pairs L2 and L5 are also shown in Figs. 18 and 19. All the matching results are projected back into the original view to ensure good inspection.

In Figs. 18 and 19, all the compared methods still failed to deliver acceptable matching results, and the messy matches can be easily found in the results of the different compared methods. Although projective deformations can be mitigated by view rectification, the repetitive patterns still have a strong effect on local feature matching. This fact is often overlooked by the existing feature matching systems. For high-resolution images, the local features extracted from repeated image regions are sometimes still distinctive because of the rich detail differences between repeated elements. However, for low-resolution images, such detail differences can be smoothed, which significantly weaken the distinctiveness of the extracted local features. This condition may result in a large amount of ambiguous matches, which is exactly the case of ground-to-aerial feature matching in urban scenarios. Without a powerful separation model, it is hard to judge whether the two features of a match come from the same repeated element or different repeated elements. For example, in Figs. 18(d) and 19(d), BF seems to achieve a considerable number of true matches from ground and aerial façades, but actually few of them are correct. In contrary, by estimating the accurate homography model, LPMG can differentiate the true and false matches from those ambiguous matches.

3) Matching Results on Local Dataset II: Different from local dataset I, the aerial images in local dataset II are captured at a relatively closer range and they are not cropped. Hence, the complex backrounds of the two aerial images are clearly visible. The numeric matching results obtained by different methods are listed in Table III. The overall situation is similar to the quantitative comparison results on local dataset I. The ASIFT

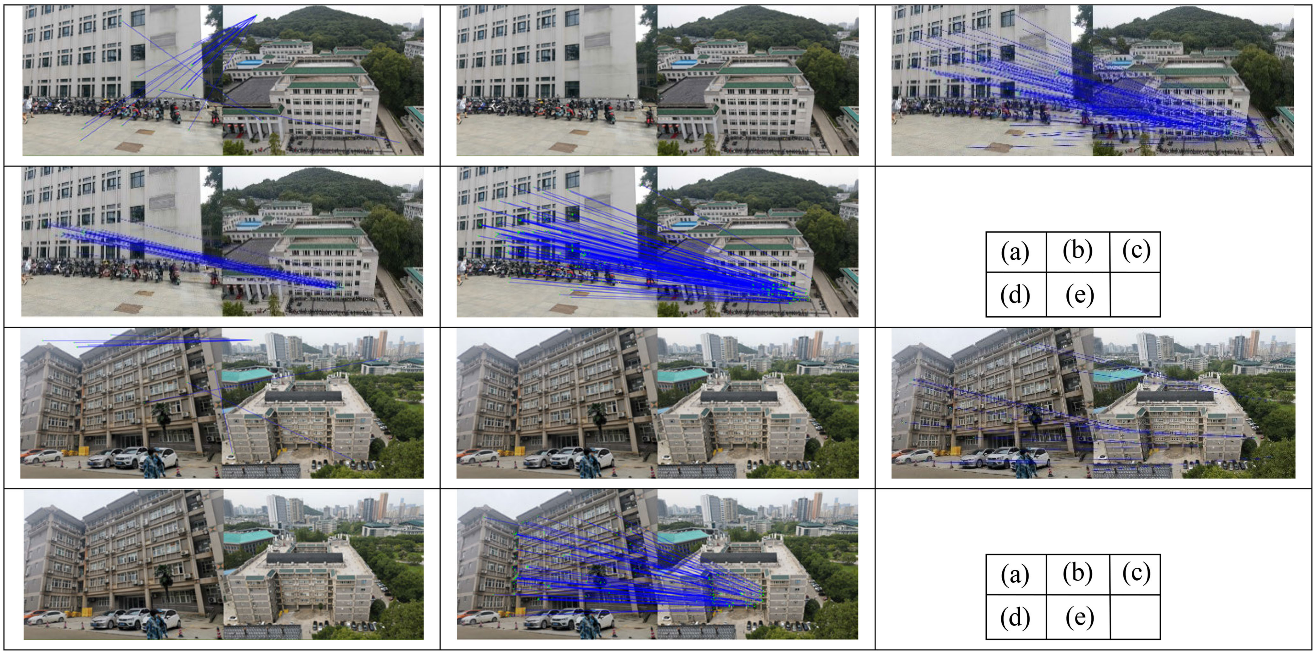


Fig. 20. Matching results obtained by different methods on local dataset II: (a) ASIFT, (b) GMS, (c) SparseVFC, (d) BF, and (e) proposed LPMG.

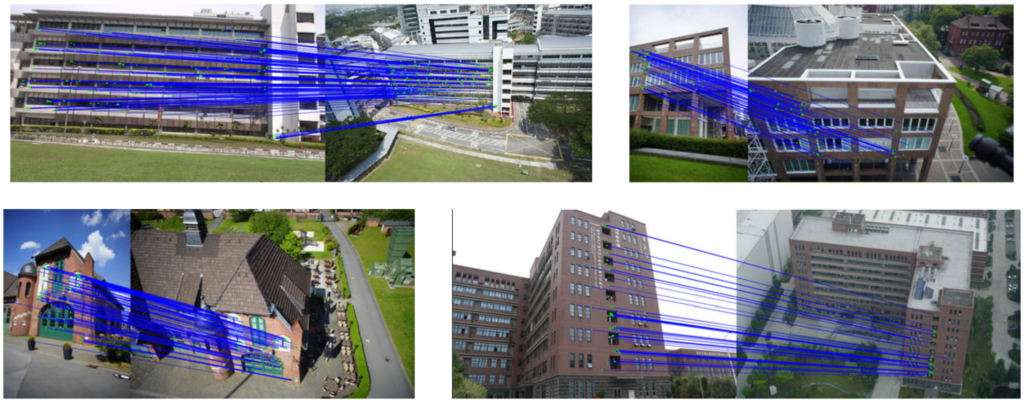


Fig. 21. Matching results of LMPG on public datasets.

TABLE III
MATCHING RESULTS OBTAINED BY DIFFERENT METHODS ON LOCAL DATASET II

Method Image	ASIFT N_c/N_t	GMS N_c/N_t	SparseVFC N_c/N_t	BF N_c/N_t	LPMG(ours) N_c/N_t
Pair L6	0/18	0/0	12/105	29/37	170/181
Pair L7	0/10	0/0	0/16	0/0	114/118

N_c denotes the number of correct matches, and N_t represents the total number of obtained feature matches.

and GMS are completely failed in two scenes. The SparseVFC and BF can obtain a few correct matches on pair L6, but obtain no correct matches on pair L7. The proposed LPMG relies on information from both local features and geometric structure of building facades, is able to deliver stable matching results. The visual effects are shown in Fig. 20. All the results confirm the potential of the proposed LPMG in ground-to-aerial feature matching for urban building images.

C. Results on Public Dataset

To validate the effectiveness and generalizability of the proposed LMPG, we also conducted experiments on the four image pairs from those public datasets, as shown in Fig. 11. The final matching results obtained by the proposed LMPG for the four image pairs from public datasets are presented in Fig. 21. The results show a similar performance to that on the local dataset, revealing the robustness of the proposed LMPG for images from different architectural scenes and with different viewing angles. LMPG is capable of processing façades that are partially or completely covered by near-regular and repetitive textures.

To evaluate the proposed LMPG, we also conducted the ASIFT, GMS, SparseVFC, and BF on the public datasets for comparison. The quantitative results of different methods performed on the four image pairs from the public datasets are listed in Table IV.

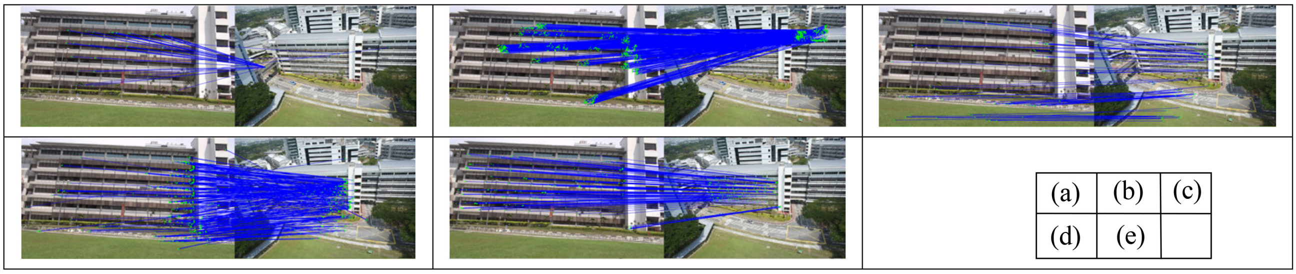


Fig. 22. Qualitative comparison results for image pair P1: (a) ASIFT, (b) GMS, (c) SparseVFC, (d) BF, and (e) proposed LPMG.

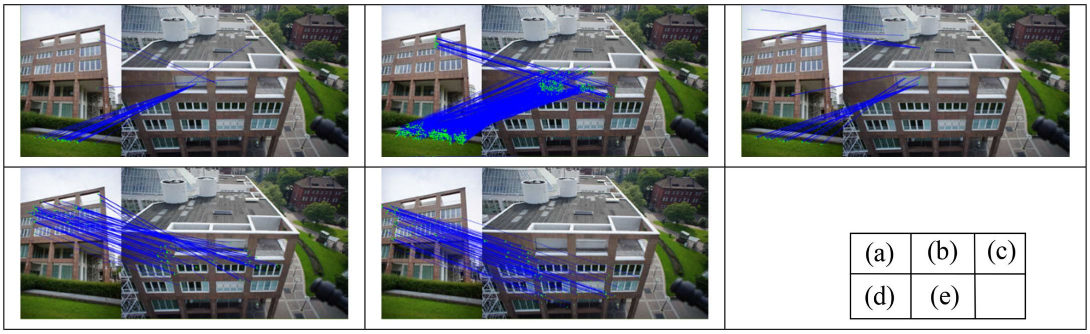


Fig. 23. Qualitative comparison results for image pair P2: (a) ASIFT, (b) GMS, (c) SparseVFC, (d) BF, and (e) proposed LPMG.

TABLE IV
QUANTITATIVE COMPARISON RESULTS OF DIFFERENT METHODS

Method	ASIFT	GMS	SparseVFC	BF	LPMG(ours)
Image	N_c/N_t	N_c/N_t	N_c/N_t	N_c/N_t	N_c/N_t
Pair P1	0/52	0/689	12/94	37/431	250/251
Pair P2	0/60	14/715	0/76	26/160	113/114
Pair P3	0/158	62/788	0/199	60/385	134/136
Pair P4	9/10	0/252	1/101	0/76	95/95

N_c denotes the number of correct matches, and N_t represents the total number of obtained feature matches.

The results in Table IV show that the performance of ASIFT and VFC are similar to that of the local dataset, revealing that those two methods are unable to match images with repetitive patterns and abrupt changes in viewpoint and illumination. The results of GMS show that compared with the complete failure on the local dataset, GMS is able to find correct matches from image pairs P2 and P3 on the public datasets. The reason may be that image pairs P2 and P3 has a high resolution and the façades are small and have a relatively weak repetitive pattern, which may allow GMS to derive correct motion statistics in some local regions. The numeric results of BF show that BF are able to identify a number of true matches from each pair of images, achieving a relatively more stable performance than on the local dataset. Evidently, the enhanced resolution mitigates the scale variation between ground and aerial images, thus improving the description of local features and resulting in a less noisy scattered putative match set. Compared with the other methods, BF models the global motion consistency in a bilateral domain, which is more capable of finding true matches from the improved putative match set. However, all these methods still fail to deliver a satisfactory result, revealing that ground-to-aerial feature matching is beyond the processing ability of those existing methods. Meanwhile, the consistent performance of the proposed LMPG on both local and public datasets again demonstrates that this dedicatedly designed method is suitable to urban ground-to-aerial feature matching task. Some comparison results from image pairs P1 and P2 are provided in Figs. 22 and 23.

IV. DISCUSSION

As mentioned in the abstract, this work specifically addresses two key technical issues: 1) reducing descriptor variance, and 2) enhancing true–false match discriminability. In this section, an in-depth analysis of the challenging local dataset is also provided to further reveal whether and how the proposed method functions on the two issues.

A. *Analysis of Influence of View Rectification*

The drastic viewpoint variation is one of the major problems that impede the ground-to-aerial feature matching. In the following, we analyze how viewpoint variation degrades the matching performance and verify if view rectification is effective based on observing changes in descriptor similarities. Taking image pair L1 as an example, we first analyze the descriptor distance changes on local features before and after view rectification, as shown in Fig. 24.

Fig. 24(a) reports the descriptor distances of two sets of correct matches from image pairs before and after rectification, and matches with the same number on horizontal axis share same feature locations (location distance is less than 1.5 pixels) on

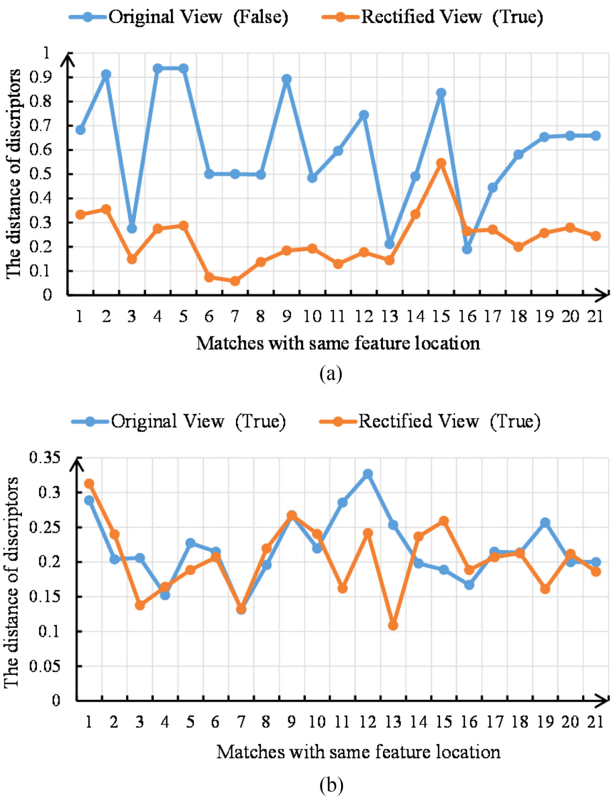


Fig. 24. Analysis of changes of descriptor distances (a) before and (b) after view rectification.

the two image pairs. To guarantee visual quality, the figure shows only the results of 21 matches. We can observe that the descriptor distances do not show an obvious change for most of the matches that remain correct in the original and rectified views. However, for matches that are false in the original views but are true in the rectified views, the descriptor distances show a clear decrease after rectification, as illustrated in Fig. 24(b). The reason may be that when transforming ground and aerial view images into an orthorectified view, both the geometry and appearance of the underlying planar region in 3-D are recovered, thereby leading to more similar descriptors of the features of a match. This is helpful to increase the number of potential true matches in the putative match set. To demonstrate this condition, Fig. 25 shows a comparison of the number of true matches involved in the putative set constructed from original and rectified views. For each pair of images, the transformation model estimated from all the obtained correct matches is used as the geometric constraint to identify the true matches from the original putative set. Clearly, each pair of rectified image pairs obtains an increasing amount of latent true matches in varying degrees.

B. Analysis of Homography Model Estimation With LMPG

Aside from improving the descriptor similarity of local features, finding a homography model with sufficient separability for true and false matches is another vital step. In this section, we discuss the separability of homography models obtained at different stages of our matching pipeline. The models are, respectively, computed from point matching seeds, coarsely aligned lattices, and finely aligned lattices termed as H_{seed}, H_{cl}, and H_{fl}. Then, we apply the models to the putative set to observe how each stage influences the performance of the final feature matching. The statistical results are reported in Fig. 26.

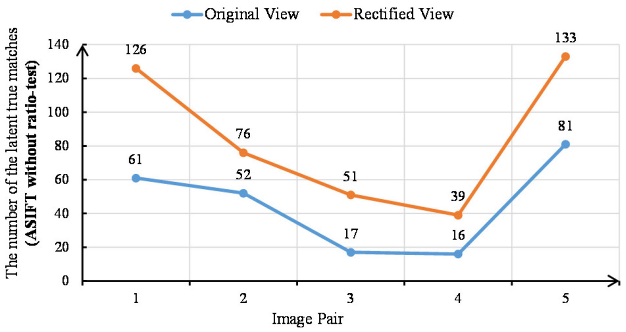


Fig. 25. Number of latent true matches in original and rectified views for image pairs from local dataset.

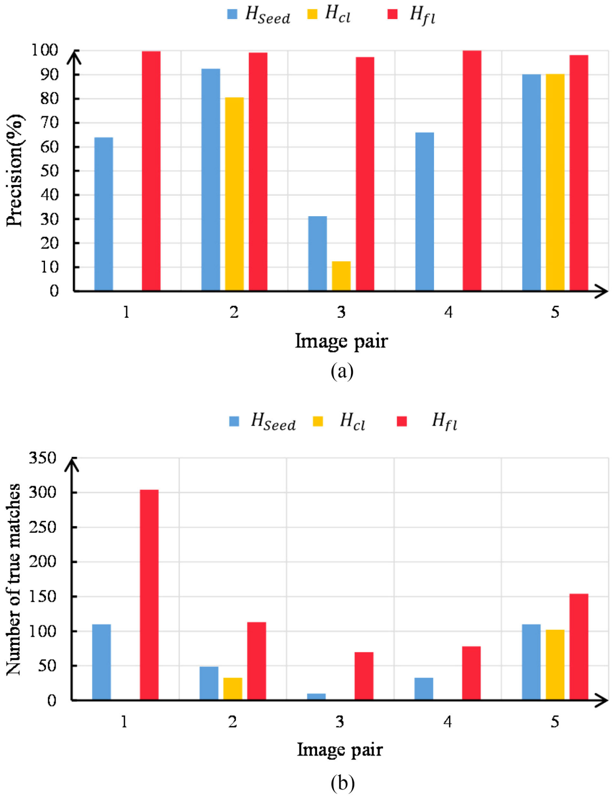


Fig. 26. Statistical results for transformation models estimated at different stages (for local dataset). (a) Matching precision. (b) Number of true matches.

As shown in Fig. 26(a), the application of H_{seed} on different image pairs produces a large discrepancy in the derived matching precision. For example, it can achieve nearly 90% matching precision for image pairs L2 and L5 while only obtaining a nearly 30% precision in image pair 3. Furthermore, the number of true matches returned by H_{seed} is also unsatisfactory, particularly for image pairs L2, L3, and L4, which are too few for a successful matching. The reason may be that point matching seeds are usually very few in number and distributed randomly and/or locally, thereby leading to unstable model estimation. For H_{cl}, the results imply that a large deviation from the real underlying image transformation models occurs. The reason is the extremely large difference between ground and aerial view images, and the lattice tiles extracted from the two views usually cannot accurately localize the same physically repeated element. Thus, a coarse alignment is not enough to derive accurate tile correspondences. From the statistical results of H_{fld}, we can observe that both the matching precision and number of obtained true matches are at the highest level. The results demonstrate that the fine alignment of ground and aerial lattices based on minimizing the transformation error of point matching seeds is feasible. To summarize, LMPG demonstrates superior standing on the ground-aerial image matching task. The multistage transformation model estimation and optimization toward LMPG proves to be not only effective but also necessary.

V. Conclusion

In this article, we presented a LMPG feature matching method for urban ground and aerial images that need no 3-D information as prior. The experimental results show that the proposed LMPG can be successfully applied to a variety of ground and aerial image pairs that have drastic variations in viewpoint, scale, and appearance, as well as contain urban buildings with different heights, architectural styles, and degrees of occlusion. The qualitative and quantitative comparison results also demonstrate that by incorporating the local features with the pattern information of urban buildings, our LMPG outperforms those existing feature matching methods with a large margin in terms of quantity and accuracy. Specifically, most of the existing matching methods failed to deliver available matching results for urban ground and aerial images, whereas our LMPG can obtain a considerable number of purely correct matches. The achievements of LPMG are mainly attributed to view rectification for descriptor similarity, and LPMG matching strategy for reliable transformation model estimation and true–false match differentiation. The in-depth analytical results in the discussion also verified these findings, thereby emphasizing that the most important contribution we have to consider is the mutually guided feature matching, which solves the ambiguous matching problem of repeated elements and local features for urban ground and aerial images.

The direct ground-to-aerial feature matching is however, a notoriously hard problem, which needs substantial effort to promote its robustness and efficiency in different scenarios. In the current stage, we still miss a more advanced method of extracting any regular or nonregular repetitive pattern for various types of urban objects, and LMPG can match only one pair of façades at a time. In future studies, we intend to match multiple façades at once and also extend the idea of mutually guided feature matching to highly complex urban objects.

References

[1] L. Liu and H. Li, “Lending orientation to neural networks for cross-view geo-localization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5624–5633.

[2] S. Hu and G. H. Lee, “Image-based geo-localization using satellite imagery,” *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1205–1219, 2020.

[3] Q. Shan, C. Wu, B. Curless, Y. Furukawa, C. Hernandez, and S. M. Seitz, “Accurate geo-registration by ground-to-aerial image matching,” in *Proc. 2nd Int. Conf. 3D Vis.*, vol. 1, 2014, pp. 525–532.

[4] Q. Zhu, Z. Wang, H. Hu, L. Xie, X. Ge, and Y. Zhang, “Leveraging photogrammetric mesh models for aerial-ground feature point matching toward integrated 3 D reconstruction,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 26–40, 2020.

[5] X. Gao, S. Shen, Z. Hu, and Z. Wang, “Ground and aerial meta-data integration for localization and reconstruction: A review,” *Pattern Recognit. Lett.*, vol. 127, pp. 202–214, 2019.

[6] A. Li, H. Hu, P. Mirowski, and M. Farajtabar, “Cross-view policy learning for street navigation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8100–8109.

[7] X. Gao, S. Shen, Y. Zhou, H. Cui, L. Zhu, and Z. Hu, “Ancient chinese architecture 3 D preservation by merging ground and aerial point clouds,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 143, pp. 72–84, 2018.

[8] N. Xue, G.-S. Xia, X. Bai, L. Zhang, and W. Shen, “Anisotropic-scale junction detection and matching for indoor images,” *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 78–91, Jan. 2018.

[9] M. Chen, A. Habib, H. He, Q. Zhu, and W. Zhang, “Robust feature matching method for SAR and optical images by using Gaussian-gamma-shaped bi-windows-based descriptor and geometric constraint,” *Remote Sens.*, vol. 9, no. 9, 2017, Art. no. 882.

[10] X. Gao, L. Hu, H. Cui, S. Shen, and Z. Hu, “Accurate and efficient ground-to-aerial model alignment,” *Pattern Recognit.*, vol. 76, pp. 288–302, 2018.

[11] S. Jiang and W. Jiang, “Reliable image matching via photometric and geometric constraints structured by delaunay triangulation,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 153, pp. 1–20, 2019.

[12] T.-Z. Xiang, G.-S. Xia, X. Bai, and L. Zhang, “Image stitching by line-guided local warping with global similarity constraint,” *Pattern Recognit.*, vol. 83, pp. 481–497, 2018.

[13] B. Fan, H. Liu, H. Zeng, J. Zhang, X. Liu, and J. Han, “Deep unsupervised binary descriptor learning through locality consistency and self distinctiveness,” *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2020.3016122.

[14] Y. Tian, X. Yu, B. Fan, F. Wu, H. Heijnen, and V. Balntas, “Sosnet: Second order similarity regularization for local descriptor learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11016–11025.

[15] M. Donoser and H. Bischof, “Efficient maximally stable extremal region (MSER) tracking,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2006, pp. 553–560.

[16] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: An efficient alternative to sift or surf,” in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2564–2571.

[17] G. Lowe, “Sift-the scale invariant feature transform,” *Int. J*, vol. 2, pp. 91–110, 2004.

[18] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.

[19] J.-M. Morel and G. Yu, “Asift: A new framework for fully affine invariant image comparison,” *SIAM J. Imag. Sci.*, vol. 2, no. 2, pp. 438–469, 2009.

[20] G.-R. Cai, P.-M. Jodoin, S.-Z. Li, Y.-D. Wu, S.-Z. Su, and Z.-K. Huang, “Perspective-sift: An efficient tool for low-altitude remote sensing image registration,” *Signal Process.*, vol. 93, no. 11, pp. 3088–3110, 2013.

[21] D. Mishkin, J. Matas, and M. Perdoch, “MODS: Fast and robust method for two-view matching,” *Comput. Vis. Image Understanding*, vol. 141, pp. 81–93, 2015.

[22] Y. Liu, L. De Dominicis, B. Wei, L. Chen, and R. R. Martin, “Regularization based iterative point match weighting for accurate rigid transformation estimation,” *IEEE Trans. Visualization Comput. Graph.*, vol. 21, no. 9, pp. 1058–1071, Sep. 2015.

[23] P. H. Torr and A. Zisserman, “Mlesac: A new robust estimator with application to estimating image geometry,” *Comput. Vis. Image Understanding*, vol. 78, no. 1, pp. 138–156, 2000.

[24] J. Ma, J. Zhao, J. Tian, A. L. Yuille, and Z. Tu, “Robust point matching via vector field consensus,” *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1706–1721, Apr. 2014.

[25] K. Adamczewski, Y. Suh, and K. Mu Lee, “Discrete tabu search for graph matching,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 109–117.

[26] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J.-M. Frahm, “Usac: A universal framework for random sample consensus,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 2022–2038, Aug. 2013.

[27] W.-Y. D. Lin, M.-M. Cheng, J. Lu, H. Yang, M. N. Do, and P. Torr, "Bilateral functions for global motion modeling," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 341–356.

[28] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T.-D. Nguyen, and M.-M. Cheng, "GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4181–4190.

[29] Z. Zhang, A. Ganesh, X. Liang, and Y. Ma, "Tilt: Transform invariant low-rank textures," *Int. J. Comput. Vis.*, vol. 99, no. 1, pp. 1–24, 2012.

[30] M. Kushnir and I. Shimshoni, "Epipolar geometry estimation for urban scenes with repetitive structures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2381–2395, Dec. 2014.

[31] Q. Zhang, Y. Li, R. S. Blum, and P. Xiang, "Matching of images with projective distortion using transform invariant low-rank textures," *J. Vis. Commun. Image Representation*, vol. 38, pp. 602–613, 2016.

[32] B. Wu, L. Xie, H. Hu, Q. Zhu, and E. Yau, "Integration of aerial oblique imagery and terrestrial imagery for optimized 3 d modeling in urban areas," *ISPRS J. Photogrammetry Remote Sens.*, vol. 139, pp. 119–132, 2018.

[33] L. Yue, H. Li, and X. Zheng, "Distorted building image matching with automatic viewpoint rectification and fusion," *Sensors*, vol. 19, no. 23, 2019, Art. no. 5205.

[34] C. Wu, J.-M. Frahm, and M. Pollefeys, "Detecting large repetitive structures with salient boundaries," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 142–155.

[35] M. Bansal, H. S. Sawhney, H. Cheng, and K. Daniilidis, "Geo-localization of street views with aerial image databases," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 1125–1128.

[36] M. Bansal, K. Daniilidis, and H. Sawhney, "Ultrawide baseline facade matching for geo-localization," in *Large-Scale Visual Geo-Localization*. Springer, 2016, pp. 77–98.

[37] M. Wolff, R. T. Collins, and Y. Liu, "Regularity-driven facade matching between aerial and street views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1591–1600.

[38] M. Park, Y. Liu, and R. T. Collins, "Efficient mean shift belief propagation for vision tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[39] D. Ceylan, N. J. Mitra, Y. Zheng, and M. Pauly, "Coupled structure-from-motion and 3 D symmetry detection for urban facades," *ACM Trans. Graph.*, vol. 33, no. 1, pp. 1–15, 2014.

[40] D.-G. Kim, W.-J. Nam, and S.-W. Lee, "A robust matching network for gradually estimating geometric transformation on remote sensing imagery," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, 2019, pp. 3889–3894.

[41] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5007–5015.

[42] J. Ma, X. Jiang, J. Jiang, J. Zhao, and X. Guo, "LMR: Learning a two-class classifier for mismatch removal," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4045–4059, Aug. 2019.

[43] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 883–890.

[44] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[45] B. Grünbaum and G. C. Shephard, *Tilings and Patterns*. USA: Courier Dover Publications, 1987.

[46] J. Hays, M. Leordeanu, A. A. Efros, and Y. Liu, "Discovering texture regularity as a higher-order correspondence problem," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 522–535.

[47] M. Park, K. Brocklehurst, R. T. Collins, and Y. Liu, "Deformed lattice detection in real-world images using mean-shift belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1804–1816, Oct. 2009.

[48] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 128–142.

[49] D. Mishkin, F. Radenovic, and J. Matas, "Repeatability is not enough: Learning affine regions via discriminability," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 284–300.

[50] A. L. Yuille and N. M. Grzywacz, "The motion coherence theory," in *Proc. IEEE Int. Conf. Comput. Vis.*, 1988, pp. 344–353.

[51] F. Nex, F. Remondino, M. Gerke, H.-J. Przybilla, M. Bäumker, and A. Zurhorst, "ISPRS benchmark for multi-platform photogrammetry," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. II-3/W4, pp. 135–142, 2015.

Xianwei Zheng received the M.S. degree in geographic information system and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 2010 and 2015, respectively.

He is currently working as an Associate Professor in computer vision and 3D geographic information system at Wuhan University. His current research interests include indoor and outdoor scene parsing, 3-D computer vision and reconstruction, and geovisualization.

Hongjie Li received the M.S. degree in geographic information system from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 2020. He is currently working toward the Ph.D degree in geographic information system with LIESMARS, Wuhan University.

His research interests include image matching, point cloud registration, structure from motion, and 3-D surface modeling.

Hanjiang Xiong received the B.S. degree from the School of Remote Sensing and Engineering from Wuhan University of Surveying and Mapping, Wuhan, China, in 1995, the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 2002.

He has been working as a Visiting Scholar at the Queensland University of Technology for three months in 2011. He is currently working as a Full Professor in 3-D geographic information system at Wuhan University. His current research interests include geospatial data management, 3-D visualization, augmented reality, and indoor and outdoor geographic information system.

Xiao Xie received the Ph.D degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2016.

She has been a Research Fellow with the Department of Cartography of Technical University of Munich, Munich, Germany from 2014 to 2016 and currently a Senior Engineer with the Key Lab of Environmental Computing and Sustainability, Liaoning province as well as an Assistant Professor in urban and environmental computation with the Institute of Applied Ecology, Chinese Academy of Sciences, China. She is also a Post-Doctoral Researcher with the School of Geodesy and Geomatics, Wuhan University. Her research interests include 3-D geographic information system and smart cities.