Multimodal Urban Remote Sensing Image Registration Via Roadcross Triangular Feature

Kun Yu, Xiao Zheng, Bin Fang, Pei An, Xiao Huang, Wei Luo, Junfeng Ding, Zhao Wang, and Jie Ma

***Abstract*—Automatic image registration of multimodal urban remote sensing images remains a critical challenging task in remote sensing image analysis due to significant nonlinear radiation distortions between multimodal image pairs; most of the traditional methods focus on the feature point detection and its local description and ignore the robust road information in multimodal urban remote sensing images. Motivated by this, we propose a fast and robust registration method for multimodal urban remote sensing images via road intersection triangular features. The proposed method obtains three main stages: Road lines extraction from images, intersection triangular feature construction, and triangular feature matching. The qualitative and quantitative experimental results show that the proposed method significantly outperforms other state-of-the-art methods, even when others completely fail to achieve the registration task of cross-modal images, our method still maintains good robustness and matching efficiency.**

***Index Terms*—Image matching, multimodal image registration, road intersection triangular feature, urban remote sensing.**

I. Introduction

IMAGE registration is a fundamental and challenging problem in multimodal urban remote sensing images, the primary goal of which is to align the reference image and sensed image, which are about the same target scene captured by different sensors, at different times or even from different viewpoints. Multimodal urban remote sensing image registration is a critical prerequisite in a wide range of applications, such as image mosaic, image fusion [1], [2], environmental monitoring, change detection, autonomous positioning of unmanned aerial vehicle (UAV) [3].

The registration problem of urban remote sensing image is typically addressed by two types of methods: Area-based methods [4] and feature-based methods [5]–[9]. Area-based methods mainly search the optimal geometric transform with a specified similarity metric and depend on an appropriate patch similarity measurement for creating pixel-level matches between the reference image and sensed image, in which mutual information (MI) [10] and Kullback–Leibler [11] are widely accepted. As a direct registration strategy, the area-based methods are generally sensitive to intensity change and illumination change. Another classic adopted pipeline is the feature-based method, including feature detection and descriptor, feature registration, and transform modal estimation.

Compared with area-based methods, the feature-based methods are more robust, which can overcome the above defects and establish the geometric relation more effectively via registration points [12], lines, contours, or regions [13], [14]. In particular, feature detection can extract the distinctive structure from the image, and feature description may be regarded as an image representation method that is widely used in image coding and similarity measurements. The representative method is scale-invariant feature transform (SIFT) [15], which can extract feature point as the local extrema in a DoG pyramid, filtered using the Hessian matrix of the local intensity values. Subsequently, the speed-up robust feature (SURF) [16] have been proposed that accelerates the SIFT by approximating the Hessian matrix-based detector using Haar wavelet calculation. To extract robust tie-points between multimodal remote sensing images, HOPC [17]and DLSS [18] are proposed based on phase congruency. Although HOPC performs feature detection, HOPC relies on accurate geographic information, it is essentially a template matching method, and unfortunately it is designed for a slight translation and have some limitations in case image pairs with scale and rotation issues. Besides, HOPC uses Harris [19] detector to detect the feature points; however, Harris is sensitive to nonlinear radiation distortions of multimodal images. Based on HOPC, Li proposed a feature matching method named radiation-variation insensitive feature transform (RIFT) [20]. RIFT uses phase congruency information instead of image intensity for feature point detection, and adopts maximum index map (MIM) which is constructed from the log-Gabor convolution sequence for feature description. However, RIFT does not build a scale space for feature detection and description, so that when the scale of the multimodal image is inconsistent, RIFT cannot be applied to matching at all. In order to evaluate a local patch similarity to find correspondences between multimodal images, the DoG detector [21] and a local EHD descriptor [22] are proposed. The EOH descriptor [23] can construct the feature description by using the edge distribution of four directional

Manuscript received January 8, 2021; revised February 22, 2021; accepted April 9, 2021. Date of publication April 15, 2021; date of current version May 12, 2021. This work was supported in part by National Natural Science Foundation of China under Grant 61991412 and Grant U1913602. *(Corresponding author: Jie Ma.)*

Kun Yu, Bin Fang, Pei An, Junfeng Ding, Zhao Wang, and Jie Ma are with the National Key Laboratory of Science, and Technology on Multi-spectral Information Processing, School of Artificial Intelligence, and Automation, Huazhong University of Science, and Technology, Wuhan 430074, China (e-mail: wh_ykun@hust.edu.cn; fangbin@hust.edu.cn; anpei@hust.edu.cn; djfenghust@hust.edu.cn; wangzhao@hust.edu.cn; majie@hust.edu.cn).

Xiao Zheng is with the Wuhan Institute of Marine Electric Propulsion, Wuhan 430064, China (e-mail: zhengxiaoxiao0204@163.com).

Xiao Huang and Wei Luo are with the Development and Design Center, China Ship Development, and Design Center, Wuhan 430064, China (e-mail: huangxiao_88@outlook.com; naruto_yk@163.com).

Digital Object Identifier 10.1109/JSTARS.2021.3073573

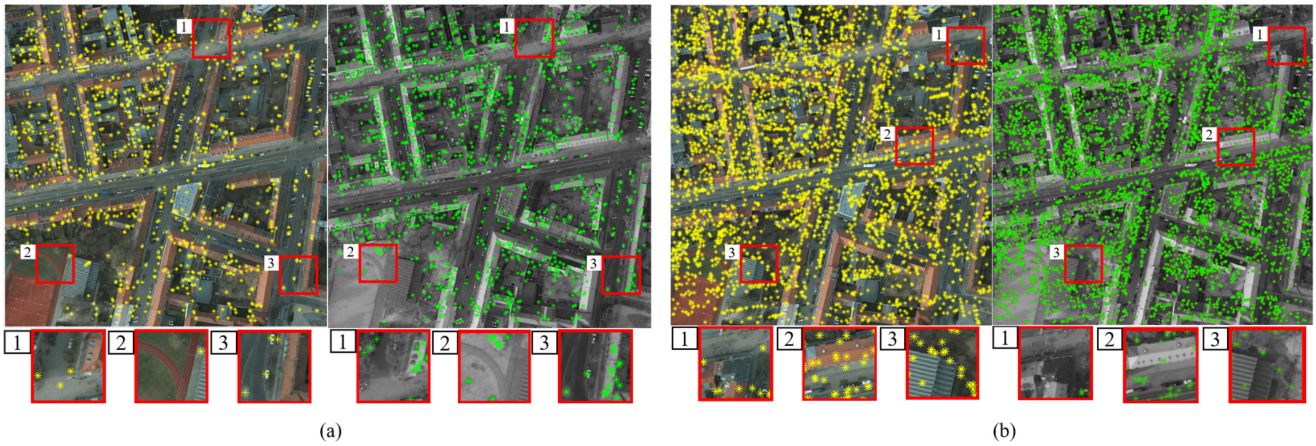


Fig. 1. (a) and (b) show the key points between optical (left) and NIR (right) urban remote sensing images, which are respectively detected by SIFT and SURF.

edges and one nondirectional edge, and it can keep structure information even when there are significant intensity variations. Different from the EOH descriptor, based on the local EHD descriptor the Log-Gabor histogram descriptor (LGHD) [24] uses multiscale and multioriented Log-Gabor filters to replace the multioriented spatial filters. Bian [25] proposed a grid-based motion statistics method (GMS), a simple means of encapsulating motion smoothness as the statistical likelihood of a certain number of matches in the region. Since GMS does not consider the size of the image when dividing the grid, the rectangular grid will be generated for images with inconsistent size, which will make the feature points distribution uniformly in the grid. In addition, GMS does not have rotation invariance in pursuit of real time. In addition, in the multimodal image matching problem, it is another important task to obtain the reliable and stable correct matching relationship while removing the mismatching relationship. Different from the traditional Ransac [26] method, the locality preserving matching method (LPM) [7] is designed, the main principle of which is to maintain the local neighborhood structures of potential true matches.

The remote sensing images are generated by different imaging mechanism, there are wide differences in image pixel quality in the same urban area at different times (i.e., season), different sensors (i.e., near-infrared (NIR), SAR). Therefore, traditional methods such as SIFT use image local neighborhood information to extract feature points from the multimodal images have become unreliable. For multimodal images of the same scene, due to the difference in imaging principles and spectral ranges, the grayscale distribution between the images has obvious characteristics of the nonlinear gap. Based on this, the grayscale transformation function will be complicated, and it is difficult to express the function uniformly in a simple form. Aiming at the registration problem of urban remote sensing images, the traditional methods based on local point feature registration such as SIFT and have the following four problems.

1) A lot of unreliable feature points will inevitably be detected. As shown in Fig. 1, due to the severe variation in the statistic of gradients around the feature point, a large number of nonduplicate feature points are detected redundantly in the multimodal image, which interferes with the subsequent feature description and registration process. In Fig. 1 (a), although there are about 3819 feature points in the optical image and 4876 feature points in the NIR image detected by SIFT, only about 45% feature points are duplicate and have the same pixel location. As for Fig. 1(b), there are about 1972 feature points in the optical image and 1600 feature points in the NIR image detected by SURF, and 49.4% feature points are duplicate. A large number of unreliable duplicate points will seriously affect the efficiency of the registration process.

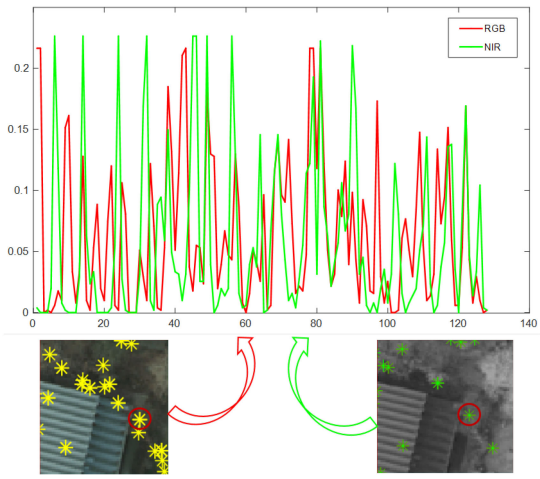


Fig. 2. 128-dimensional feature description values of corresponding key points between optical (left) and NIR (right) by SIFT. The closer the curves are, the more equivalent the description information of key points are. Due to the nonlinear radiation distortion of multimode image pairs, the description information of corresponding key points have obvious differences, which will reduce the robustness for matching.

2) The differences in local feature description between multimodal images. The feature description is a vital step for image registration; however, due to the different sensor imaging mechanisms and different imaging environmental conditions, different degrees of gradient inversion and nonlinear radiation distortions will exist between multimodal images. Hence, the traditional local feature description of the corresponding points cannot keep consistent, and the invariance of registration gets lost. It can be seen in Fig. 2 that the 128-dimensional local

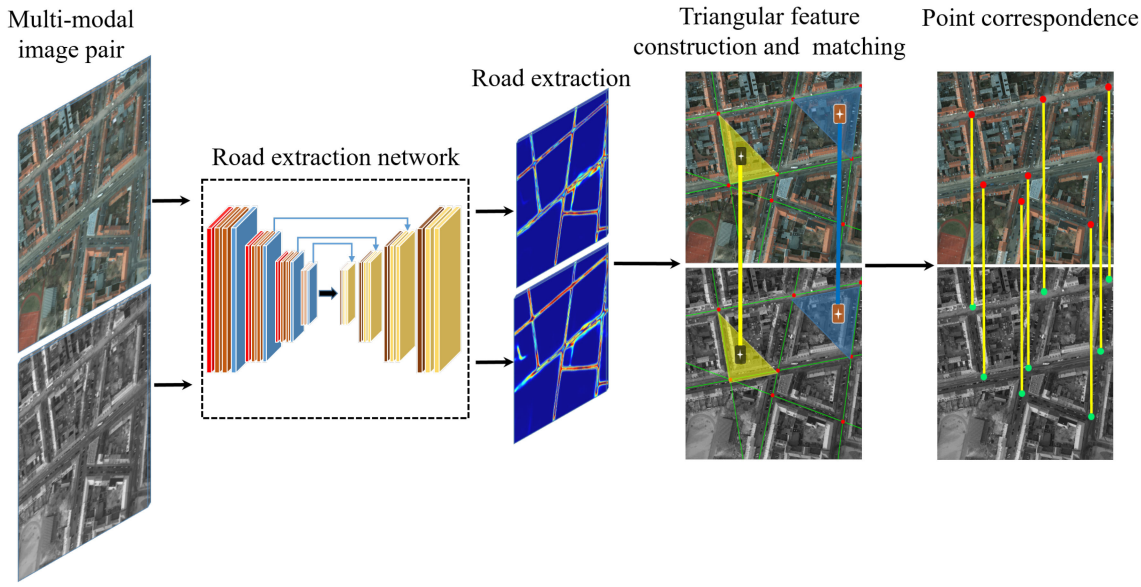


Fig. 3. Proposed method for multimodal urban remote sensing image registration.

feature description of the corresponding points marked with a red circle in the optical and NIR images have obvious differences. The lower the degree of polyline coincidence, the greater the difference in characteristics.

3) Feature point registration will be easily affected by the local repetitive structure. After using local features to describe a certain feature point, due to the lack of contextual information, it is more likely to be affected by the local repetitive structure and the symmetry structure. This will cause mismatches between feature points and ultimately affect the registration accuracy.

4) High-registration time complexity. It will cost a lot of registration time to match a mass number of feature points with high-dimensional feature description, which is the disadvantage of the local feature point registration methods. Although the emergence of fast search methods such as KD-tree alleviated this problem, the issue has not been solved fundamentally.

To address these issues, we propose a fast and robust multimodal remote sensing image registration method based on road intersection extraction and intersection triangular feature, the diagram is shown in Fig. 3. The road is a robust feature that can be seen everywhere in urban remote sensing image, although the number of the road intersections is small, these intersections are also extremely stable, which is an excellent solution to replace traditional feature points for matching. The contribution of proposed method mainly includes the following fourfold.

1) We proposed a fast and robust registration method for multimodal urban remote sensing images. The proposed method is more robust than others. Even when other methods completely fail to achieve the registration task of cross-modal images, our method still maintains good robustness and matching efficiency.

2) Road lines and intersections extraction: Based on the road segmentation model of remote sensing image and training a large number of labeled datasets, we can extract the road network from multimodal urban remote sensing image pairs and the skeleton and line fitting methods are adopted to detect more accurate road centerlines. Subsequently, the road intersections can be calculated by the road centerlines, these road intersections are used for feature points for our next description.

3) Triangular feature construction: Due to the differences in the texture pixel between the local neighborhood around the feature point, the information provided by these local neighborhood is limited, which is not conducive to distinguishing the repetitive and symmetrical structures. Therefore, we construct the intersections triangular feature to describe the relationship between feature points. These triangular features not only maintain nonneighborhood information within a certain range but also speed up the process of feature matching due to the simple feature description.

4) Triangular feature matching: We adopt the NNDR to obtain the corresponding triangular feature structures in the multimodal images. Since each triangular feature structure can estimate a set of transformation parameter $\mathbf{R}$ and $\mathbf{t}$, combined with the best matching function we proposed, we filter out the most accurate $\mathbf{R}$ and $\mathbf{t}$ as the matching result. In this step, because the dimensionality of the triangular feature structure is extremely low, and the number of feature points is much less than traditional local descriptors, the matching efficiency is higher. Besides, since the proposed repetitive rate of road centerlines is used to evaluate the $\mathbf{R}$ and $\mathbf{t}$ of each triangular feature structure, theglobal optimization method can be achieved. The remainder of this article is organized as follows. Section II introduces the proposed registration method. In Section III, the experimental results and corresponding analyses are exhibited. Finally, Section IV concludes this article.

II. Proposed Registration Method

A. Road Network Extraction From Multimodal Urban Remote Sensing Image

Road network extraction is the most important salient feature in multimodal urban satellite and aerial images, and it can also be

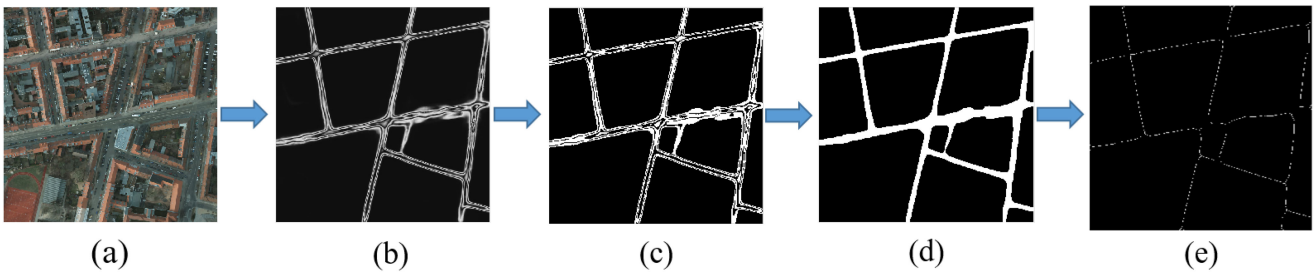


Fig. 4. Diagram of road extraction. (a) Input image. (b) Inferred road. (c) Binarization. (d) Morphological operation. (e) Road skeleton.

considered as a segment problem. Although manual extraction of roads from these images is possible and more accurate than computational work, in terms of cost and time manual, extracting road information automatically and efficiently has more practical application significance. In recent years, a variety of methods have been proposed to extract the road network. Traditional unsupervised methods mostly use different threshold segmentation schemes based on the difference between the grayscale value of the road and background [27]. However, in complex urban scenes, there is no obvious diffidence of grayscale value between the road and the backgrounds such as buildings and squares. Besides, the deep neural networks are adopted to understand the road information of remote sensing image [28], such as a semantic segmentation neural network (ResUnet), which combines the strengths of residual learning and U-Net, is proposed for road network extraction [29]. The segmentation model ResUnet has the rich skip connections, which allow designing the networks with fewer parameters and outperforms U-Net [30] and other state-of-the-art deep learning methods of road extraction [31]. ResUnet as a semantic neural network combines the strengths of both U-Net and residual neural networks.

The input training images with three channels and labels are tiled into 1300×1300 pixel chips. The topology contains several residual blocks, stacked on top each other, having different filter sizes. Residual blocks are meant to make the network to converge faster and the basic concept is to use addition to merge the initial feature map with the information from the extracted patterns resulted after one pass through the residual block. ResNet can serve as an encoder for a semantic segmentation problem. In this article, inspired by the segmentation model ResUnet and the winning SpaceNet3 [32], we use a ResNet34 encoder with a U-Net inspired decoder. The size of the convolution kernel is 7×7, and the stride is 2. In each residual block, the step size of the first convolution is 2, and the step size of the rest convolution operations is 1. We utilize SpaceNet3 satellite imagery and geocoded road centerline labels to build training datasets for our models. We include skip connections at every layer of the network, and the Adam optimizer is with the default parameter. The skip connection is to solve the problem of gradient disappearance, the skip connection can better transmit the gradient to a shallower level in the process of later transmission. Adam as an optimizer can make our model converge faster and reduce the time cost. The loss function is defined as follows [33]:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{BCE}} + (1 - \alpha)\mathcal{L}_{\text{Dice}} \tag{1}$$

where $\mathcal{L}_{\text{BCE}}$ means the loss of binary cross entropy and $\mathcal{L}_{\text{Dice}}$ is the Dice coefficient.When $\alpha = 0.75$, it can reach the best performance.

The inferred road information can be seen in Fig. 4 (b), which is predicted by our adopted model, and the grayscale value represents the coefficient level of the inferred road information. In order to refine these road vectors, we attempt to close the small gaps and remove spurious connections not already corrected via the image preprocessing methods including binarization, opening and closing procedures, and road skeleton [see in Fig. 4 (c)–(e)].

B. Road Intersection Extraction and Triangular Feature Construction

The fitting centerlines of the road, and the subsequent detection and description of intersection points are all built based on the single-pixel skeleton image. Once the skeleton images are extracted from the multimodal image pairs, the road intersection points extraction and description are needed to distinction between center lines.

In multimodal urban remote sensing images, the road intersection is still an excellent robustness feature. Even if part of the predicted road is missing, such as being blocked by clouds, as long as most of the roads can be accurately predicted, the centerline of the road can be fitted from the binary skeleton image. Subsequently, the intersection points of centerlines can be considered as point structures with distinguishability and matchability. In particular, if the intersection point is on the extension of the road centerline and not on the actual road, the point can still be regarded as a feature point and constructed triangular feature. The concept of intersection points in this article is a broad term.

To speed up the straight lines detected time, the uniform sampling operation is adopted first for the skeleton image of the road to get more sparse data. The next step is dedicated to obtaining the vector map of the centerline by using RANSAC. In RANSAC algorithm, a sample of data is randomly selected which is considered as the primary model. A set of data points within threshold distance $\mathtt{t}$ of the model has been computed by the algorithm to reestimate all of the points leading to the best-fit line selection.

In addition, we will remove the interior points from data when there is a best-fit line, to further accelerate the speed of fitting straight lines. The schematic diagram of interior point removal

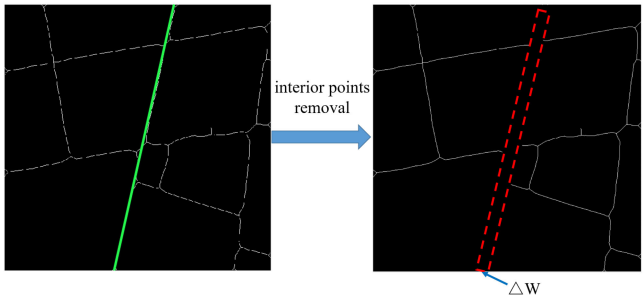


Fig. 5. Schematic diagram of interior point removal. The green line is the fitting line, the red dotted box is the range of removing the internal points, and the pixel width of the red dotted box is △W.

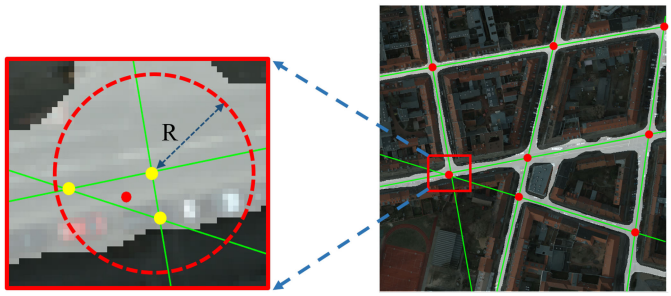


Fig. 6. Result of intersection point extraction. The white layer overlaid on the optical layer means inferred road. The green lines mean fitting straight lines. The red dots are road intersections as feature points. The yellow dots are the intersection of the straight lines.

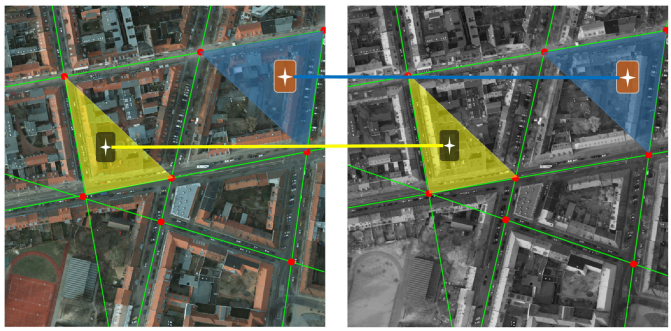


Fig. 7. Diagram of triangular feature matching.

is shown in Fig. 5. When the RANSAC is used to successfully fit the road center line, we take the centerline of the road as the benchmark and remove the redundant internal points within a certain pixel width range △W, so that we can fit all the center lines more quickly and accurately. The fitted straight lines are shown as the green line in Fig. 6 , and the yellow dots mean the intersections of straight lines within the image size. Obviously, there are some yellow dots in the neighborhood which are shown in the left diagram of Fig. 6. Hence, we construct a search tree for all intersection points of lines and replace these points in the Euclidean distance threshold with the center point in (2). The result of road intersections is shown in Fig. 6

$$P(x_c,\ y_c) = \frac{1}{n}\sum_{i=1}^{n} P(x_i\ ,\ y_i), \quad n = 1,\ 2,\ 3,\ \cdots \tag{2}$$

C. Triangular Feature Matching

1) Definition of Triangular Feature Matching: Given two multimodal urban remote sensing image pairs $\mathbf{S}$ and $\mathbf{T}$, the fitting road line sets $\mathbf{S}_L$ and $\mathbf{T}_L$ can be extracted, respectively. We use the intersection of randomly nonparallel road lines to build triple elements denoted as $\mathbf{S}_{L_i}$ and $\mathbf{T}_{L_i}$. By comparing the similarity of the orderly interior angles $\{\theta_1, \theta_2\}$ of the triples, the initial matching set $C = \{(\mathbf{S}_{Ti}, \mathbf{T}_{Ti})\}_{i=1}^{M}$ can be obtained. Given set C, the task is to estimate the 4DOF similarity transformation parameters

$$f(\mathbf{T}_{Ti}|s, \theta, \mathbf{t}) = s[\mathbf{R}(\theta)\mathbf{S}_{Ti} + \mathbf{t}] \tag{3}$$

where θ is the rotation parameter and its range is $\theta \in [0, 2\pi]$. $\mathbf{t}$ is the translation vector and $\mathbf{t} \in \mathbf{R}^2$. s is the scale parameter. After the similarity transformation, our goal is to align as many road lines as possible.

2) Nearest Neighbor Line and Overlap Rate of Lines: The line on the 2D plane can be represented by the two parameters r and θ_r. The parameter r means the distance from the line to the original point, and the parameter θ_r is the angle between the positive direction of the X axis and the vertical line which is from the original point to the road line. Given two lines in the multimodal image pair $\mathbf{L}_i = (r, \theta_r)_i$ and $\mathbf{L}_j = (r, \theta_r)_j$, the similarity degree between these two lines can be defined as

$$\mathbf{LD}_{Li,Lj} = 0.5 * \left(|\tfrac{r_i}{r_j} - 1| + |\tfrac{\theta_i}{\theta_j} - 1|\right) \tag{4}$$

when $\mathbf{LD}$ is less than a given threshold ε, the two lines $\mathbf{L}_i$ and $\mathbf{L}_j$ are considered as corresponding lines.

When the transformation parameters s, θ, and $\mathbf{t}$ are solved, the road line set $\mathbf{S}_L$ can be transformed into the target road image space, and the transformed road lines can be denoted as $\mathbf{S}'_L$. For each transformed road line, the best similar line can be found in $\mathbf{T}_L$ by (4), the measure distance of the line pair can be denoted as

$$\mathrm{Dist}(\mathbf{S}'_{Li}, \mathbf{T}_L) = \min(\mathbf{LD}_{Li,Lj},\ L_j \in \mathbf{T}_L). \tag{5}$$

When $\mathrm{Dist}(\mathbf{S}'_L, \mathbf{T}_L)$ is less than the threshold ε, there will exist a pair of overlapping road lines in the multimodal image pair. After executing all the lines, the number of road lines which satisfy the transformation can be presented as $\mathbf{NLO}_{s,\theta,\mathbf{t}}$. Finally, the normalized overlap rate is used for measuring the parameters s, θ, and $\mathbf{t}$, and the normalized overlap rate is defined as follows:

$$\mathbf{NLO}_{s,\theta,\mathbf{t}} = \frac{\mathbf{NLO}_{s,\theta,\mathbf{t}}}{\max(|\mathbf{S}_L|, |\mathbf{T}_L|)}. \tag{6}$$

For each triplet, it can get a normalized overlap rate of lines, and we take the transformation parameter at the maximum overlap rate as the final initial transformation parameter

$$s^*, \theta^*, \mathbf{t}^* = \arg\max_{s,\theta,\mathbf{t}}(\mathbf{NLO}_{s,\theta,\mathbf{t}}|\varepsilon). \tag{7}$$

III. Experiments

A. Datasets and Evaluation Metrics

1) Train Dataset and Test Image Pairs: **Train dataset**: We use the SpaceNet 3 datasets and road labels [34] including Las

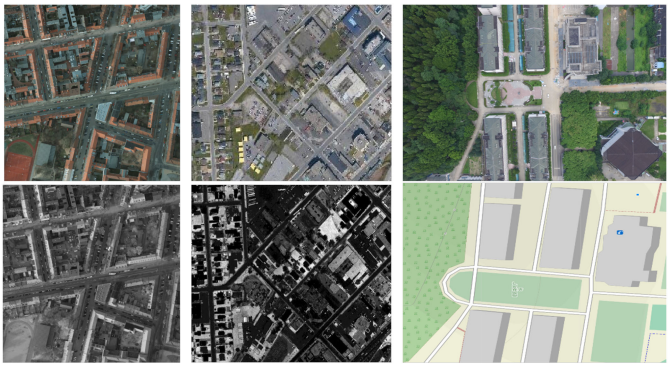


Fig. 8. Examples of multimodal image pairs from different datasets.

Vegas, Paris, Shanghai, and Khartoum datasets to train our model of road network extraction. In order to improve the ability of generalization, the selected datasets and road labels contain a large number of road remote sensing and aerial imagery at different times of the world, and cover a huge geographic area from 400 to 3600 square kilometer. Training images with three channels and labels are tiled into 1300 × 1300 pixel chips and the image resolution is about 30 cm/pixel.

Test image pairs: The experiment test image pairs include three kinds of multimodal urban remote sensing images as shown in Fig. 8. $optical - NIR$ image pair is from the Potsdam dataset [35] which contains 38 the same size patches, each patch consisting of a true orthophoto extracted from a larger mosaic. The dataset mainly represents distant view remote sensing images and has a large number of ground repeatable structures. Each patch image has the same size (6000×6000) with optical and NIR channels, and the resolution is 5 cm/pixel. The resolution of the test image must be consistent with that of the training image. To achieve this, we reduced the resolution of the high-resolution test image through the down-sampling method. $optical - Intensity$ image pair as aerial image contains optical spectral band and LiDAR intensity information on the Niagara city in Canada. The image pair has certain difficulty in image registration, due to the grayscale values of the corresponding areas between the visible and intensity images present an obvious nonlinear difference. $optical - OSM$ image pair reflects the square scene of the Huazhong University of Science and Technology which contains the drone aerial photography and open street map (OSM) area. As an abstract representation of remote sensing image features, OSM map information has certain common characteristics; high-resolution remote sensing images and OSM map registration can be better applied to mapping and geographic information update, and can also be used for autonomous positioning of UAV. However, these image types are completely different, and the OSM image is mainly composed of simple geometric shapes that cannot reflect the real remote sensing scene, this kind of cross-modal image is a huge challenge for registration. To establish the ground truth of $optical - OSM$ image pair, we recorded the altitude, resolution, and direction angle of the UAV.

2) Evaluation Metrics: To compare the robustness of the feature point detection and matching by our proposed method and five other comparison methods (SIFT, SURF, EHD, LGHD, and LPM). We use the repeatability Rep and the number of repetitive correspondences N^c that can be established for detected features as the evaluation metrics [36]. According to the homography matrix between the two images, it can be calculated to determine whether the feature points are repetitive. The repeatability Rep is a ratio between N^c and the average number of features detected in two image pairs $\mathbf{S}$ and $\mathbf{T}$, repeatability Rep can be defined as follows:

$$\text{Rep} = \frac{N^c}{(n_s + n_t)/2} = \frac{|\{||\mathbf{x}_i^s - \mathbf{H}\mathbf{x}_i^t|| < 3\}_{i=1}^{n_s}|}{(n_s + n_t)/2} \tag{8}$$

where $\mathbf{H}$ is the ground truth transformation between $\mathbf{S}$ and $\mathbf{T}$; $\mathbf{x}_i^s$ and $\mathbf{x}_i^t$ are the homogeneous coordinates of a feature in $\mathbf{S}$ and $\mathbf{T}$, respectively; n_s and n_t are the number of feature points in $\mathbf{S}$ and $\mathbf{T}$, respectively; $|\{||\mathbf{x}_i^s - \mathbf{H}\mathbf{x}_i^t|| <\}_{i=1}^{n_s}|$ represents the number of matches that satisfy $||\mathbf{x}_i^s - \mathbf{H}\mathbf{x}_i^t|| < 3$.

The correct matches rate (CMR) is chosen as the evaluation criterion [17]. The CMR is defined as follows:

$$\text{CMR} = \frac{\#CM}{\#C} \tag{9}$$

where #CM represents the number of correct matching points; #C represents the total number of matching point pairs. The point pair with localization error less than threshold is regarded as the correct match. The value of threshold is set to 3.0 pixel.

In order to directly evaluate the accuracy of the alignment of the transformed image, we use the root mean square error (RMSE) and mean error (ME) as the evaluation metrics.

B. Performance With Respect to Feature Point Detection

To demonstrate the advantages of our proposed road intersections for multimodal images, experiments were performed on the above test image pairs. Some state-of the-art methods are compared which are SIFT, SURF, EHD, LGHD, and LPM, among other five comparison methods, SIFT and SURF find extreme points as feature points on the constructed scale space, EHD and LGHD both adopt FAST [37] feature points. LPM is used to remove the mismatching relationship when SIFT is usually adopted to establish putative feature correspondences. In order to evaluate the algorithm fairly and effectively, the parameters of each comparison method are fine-tuned to obtain the best performance and are consistent in all test image pairs. SIFT is implemented by the open-source VLFEAT toolbox [38], other comparison methods are obtained from the authors' website.

The Table I has shown evaluation results of the feature point detection achieved by these six methods, the highest repeatability values are highlighted with boldface font. As for all test image pairs, although our method has a lower number of the repetitive correspondences which depends on the road intersection points, it can still maintain a high-feature point repetition rate. SIFT, SURF, EHD, and LGHD almost have the same level of repetition rate.

C. Performance With Respect to Matching Accuracy

To demonstrate the matching accuracy of our proposal, we compare it with the above five state-of-the art methods. The

TABLE I
REPEATABILITY RATE (REP) AND THE NUMBER OF REPETITIVE CORRESPONDENCES(N^c) ACHIEVED BY COMPARISON METHODS AND OUR PROPOSED METHOD

	optical-NIR		optical-Intensity		optical-OSM	
Method	*Rep*	N^c	*Rep*	N^c	*Rep*	N^c
SIFT	45.0%	540	43.0%	553	4.8%	62
SURF	42.0%	511	42.9%	528	4.1%	87
EHD	44.7%	524	43.2%	549	11.7%	165
LGHD	44.7%	524	43.2%	549	11.7%	165
LPM	45.0%	540	43.0%	553	4.8%	62
Ours	**100%**	8	**90%**	9	**100%**	12

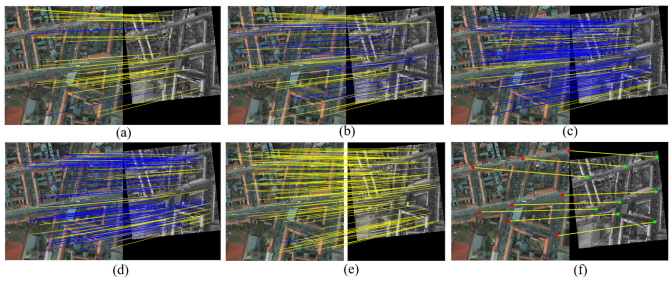


Fig. 9. Registration results of proposed method and other five methods on optical-NIR image pair. The NIR image has affine transformation with rotate transform ($\theta = 6°$) and scale transform ($\sigma = 0.75$). The feature matching points in two images have been marked as red dots and green dots, and yellow and blue matching lines mean true positive and true negative.

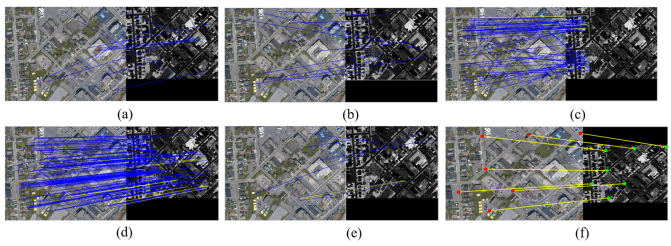


Fig. 10. Registration results of proposed method and other five methods on optical-Intensity image pair. The Intensity image has affine transformation with rotate transform ($\theta = 0°$) and scale transform ($\sigma = 0.75$). The feature matching points in two images have been marked as red dots and green dots, and yellow and blue matching lines mean true positive and true negative.

qualitative evaluation on the matching performance of the proposed method and others are shown as Figs. 9, 10, and 11. The test image pairs have different affine transformation, the NIR image has affine transformation with rotate transform ($\theta = 6°$) and scale transform ($\sigma = 0.75$) and the intensity image has affine transformation with rotate transform ($\theta = 0°$) and scale transform ($\sigma = 0.75$). For SIFT and SURF, they were matched by the Euclidean distance ratio between the nearest neighbor and the second nearest neighbor of corresponding features, and the ratios is set as 0.7. The RANSAC algorithm was adopted for SIFT, SURF, EHD, and LGHD these four comparison methods to remove the mismatching points, LPM can remove mismatches by preserving the local structure consistency of correct correspondences matching. As for our method, we adopted the proposed overlap rate of road lines as a global optimization method to evaluate the rotate and scale result of each triangular feature structure.

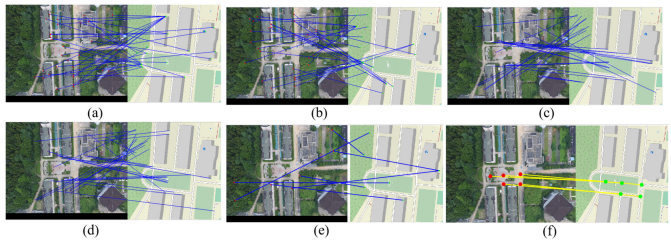


Fig. 11. Registration results of proposed method and other five methods on optical-OSM image pair. The feature matching points in two images have been marked as red dots and green dots, and yellow lines mean correct matching, blue lines mean the fault matching and too many outliers to satisfy the RANSAC conditions.

1) Qualitative Comparisons: In Figs. 9, 10, and 11, these image pairs have different registration transformation including rotation and scale changes. Therefore, matching on these multimodal urban remote sensing image pairs is very challenging, Figs. 9, 10, and 11 plot the matching results of SIFT, SURF, EHD, LGHD, LPM, and proposed method.

As seen, SIFT, SURF, EHD, LGHD, and LPM have a certain matching effect on $optical - NIR$ image pair, the reason is that the spectral range of the NIR image is closer to the optical image's, the feature description of image pair has a certain degree of similarity. The number of correct matches (NCM) of EHD is smaller among five comparison methods and NCM of LPM is larger. The difference of NCM also proves that LPM is better than RANSAC at eliminating mismatches, and RANSAC cannot work well when the image transformation does not satisfy a parametric model. LPM can obtain more robust correct matches and perform the best among the five comparison methods. Our proposed method can accurately extract the road intersection points to complete the matching. In Fig. 10, due to the nonlinear radiation distortions of optical and intensity image pair, the traditional feature descriptions of correspondences such as SIFT have a great difference, and the robustness of five comparison methods is greatly reduced. Our method can still achieve a good matching effect, accurately extract the main road intersection points, and the robustness performance is far superior to other comparison methods. In addition, the Fig. 11 represents a completely different cross-modal image pair. The OSM map to be matched is a semantic label map, the map is structured by the simple geometrical shape of buildings and roads which are not real scenes, on this occasion other local description operators are powerless. This image type is completely different from UAV optical aerial image, this kind of image pair puts forward higher and more difficult requirements for multimodal urban remote sensing matching task. It can be seen that SIFT, SURF, EHD, LGHD, and LPM completely fail to match on $optical - OSM$ image pair; however, only our method can extract feature points effectively and match them accurately. The proposed method has a unique advantage over the cross-modal matching problem.

The registration results of different multimodal urban remote sensing image pairs are shown in Fig. 12, and affine transformation is used as the geometric model for checkerboard mosaic or overlap images. As can be seen from the magnification displays of the local areas, the alignment precision of our proposed

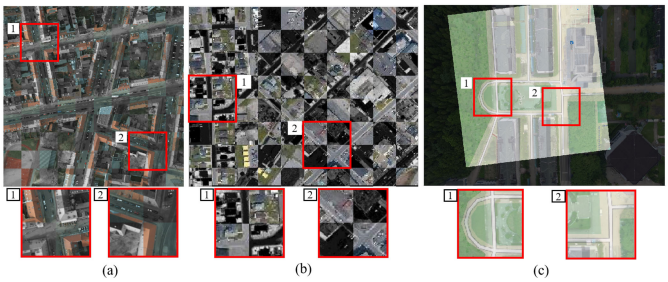


Fig. 12. Registration results of the proposed method. (The top row is the registration results with checkerboard or overlap, and the bottom row is the magnification displays of the local area.).

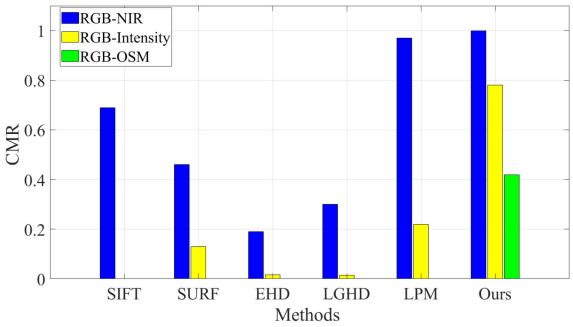


Fig. 13. CMR scores of the proposed and comparison methods on three multimodal image datasets. Some bars not shown mean that there are no correct matches and CMR value is 0 by default.

method is very high and the local geometric distortion is very small, which will meet the practical application.

The matching performance of the proposed method on the multimodal urban remote sensing image pairs with nonlinear radiation distortions is far superior to other state-of-the-art feature matching methods. The reasons may be as follows. a) We choose the deep learning model to segment road information and extract road intersection points from multimodal image pairs. Although the number of road intersection points is not large, these intersection points are robust enough to be considered as feature points. b) The local description of the feature points is abandoned, we propose to construct the intersection points triangular features and describe the global information between triangular features to prevent the description differences caused by the local grayscale level information. In this way, the nonlinear radiation distortions of multimodal images can be avoided to disturb feature description. Therefore, the proposed method is more robust than others on different multimodal datasets, even when comparison methods completely fail to achieve the registration task of cross-modal images, the proposed method still maintains good robustness and matching efficiency.

2) Quantitative Comparisons: Fig. 13 is the quantitative results of the CMR metric, where shows the results of all methods on three kinds of multimodal image pairs. As seen, the matching performance of the proposed method is very stable and robust, it is hardly affected by the type of radiation distortions, our method is far superior to other methods. In $optical - NIR$ image pair, the difference in imaging mechanism between images is smaller than that of the other datasets and image matching is relatively

TABLE II
QUANTITATIVE EVALUATION RESULTS OF THE PROPOSED METHOD

metrics	optical-NIR	optical-Intensity	optical-OSM
ME /pixels	1.32	1.60	2.52
RMSE /pixels	1.47	1.85	2.78

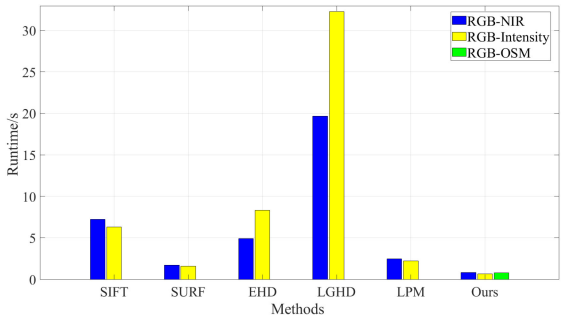


Fig. 14. Running time results of the proposed and comparison methods.

easy. In all comparison methods, LPM achieves better results than SIFT, SURF, EHD, and LGHD, five comparison methods can obtain higher matching scores on $optical - NIR$, EHD performs the worst. In $optical - Intensity$ image pair, the difference is that the nonlinear radiation distortions are more complex, CMR of most comparison methods declined significantly. However, the proposed method still has the highest scores because of its resistance to nonlinear radiation distortion changes. In $optical - OSM$ image pair, five comparison methods do not have the correct matches at all. Accordingly, the green bar of each comparison method in the figure is not marked, indicating that CMR value of the method is 0. CMR result demonstrates that the comparison methods are completely ineffective for this kind of cross-modal image pair.

In order to evaluate the alignment accuracy of the proposed method, Table II reports the ME and RMSE of the proposed method on all datasets. From the table, the precision of registration alignment is high, where the ME are approximately 1.32, 1.60, and 2.52 pixels on three datasets, and RMSE are approximately 1.47, 1.85, and 2.78 pixels. Due to $optical - OSM$ dataset is acquired by UAV, the ground truth of affine transformation in each image pair is mainly obtained by manually marking a large number of control points, so ME and RMSE values of $optical - OSM$ dataset are higher than $optical - NIR$ and $optical - Intensity$, respectively. In spite of this, the ME and RMSE values of $optical - OSM$ dataset still remain within the range of 3 pixels, which meets the matching requirements.

Furthermore, all methods are evaluated by the scale error ε_S and rotate error ε_R metrics on three datasets in Table III, the lowest values are highlighted with boldface font. As seen, our proposed method compared with the other five methods performs best and has the lowest error on each image pair.

D. Performance With Respect to Running Time

As well as the matching accuracy, the computational efficiency is another important metric for evaluating the matching performances. Fig. 14 reports the running time of each compared

TABLE III
SCALE ERROR ε_S AND THE ROTATE ERROR ε_R ACHIEVED BY OUR AND COMPARISON METHODS

	Scale error $\varepsilon_S(10^{-3})$			Rotate error ε_R		
Method	optical-NIR	optical-Intensity	optical-OSM	optical-NIR	optical-Intensity	optical-OSM
SIFT	0.15	-	-	0.03	-	-
SURF	0.30	15.45	-	0.08	118.73	-
EHD	1.20	31.50	-	0.04	9.51	-
LGHD	0.60	19.10	-	0.06	0.06	-
LPM	0.10	6.42	-	0.02	4.40	-
Ours	**0.02**	**0.10**	**0.13**	**0.02**	**0.01**	**0.01**

method on three datasets. The running time experiment has been implemented in Matlab using a PC equipped with a 3.4 GHz CPU and 4 GB of RAM. The running time metric of each method contains all the matching processes including feature detection, feature description, matching measurement, and mismatching elimination.

As can be seen in Fig. 14, the proposed method performs best and costs the smallest running time than other compared methods on three multimodal image datasets, the running time of the proposed method is about half of SURF, which is the fastest of all comparison methods. The green bars of comparison methods on $optical - OSM$ in the figure are not marked, indicating that the methods cannot achieve the matching task, the running times are infinite by default. Therefore it is meaningless to discuss the running time results of comparison methods on $optical - OSM$ dataset. On the other two datasets, LGHD takes a longer running time than other comparison methods.

It can be seen that our proposed method spent less time than compared methods on the same multimodal image pair, the reason is that our segmentation network model can quickly extract the road information from the multimodal image pairs, and several robust intersection points are used to construct triangular feature. The constructed triangular features not only maintain nonneighborhood information within a certain range but also speed up the process of feature matching due to the global feature description. In contrast, the comparison methods have a lot of redundant feature points and local descriptions, the redundant information will spend a lot of unproductive time.

In general, the proposed method can have a smaller running time while maintaining a higher CMR score, and the matching efficiency is the best than other compared methods.

E. Failure Cases

Although the proposed method has many advantages over other comparison methods in multimodal urban remote sensing image registration, there are some limitations and main failure cases shown in Fig. 15. The limitations mainly include the following aspects. 1) The resolution of the multimodal image pairs must be consistent with that of training dataset images. Fig. 15(a) shows that when the resolution of experimental multimodal image pairs is higher than that of training images, the width of segmentation road is smaller than actual road, the red dotted lines indicate the width of actual road. 2) Since the proposed method uses the straight lines to fit the segmentation road information and obtains the intersection points, there are errors and uncertainties in the coordinates of intersection points

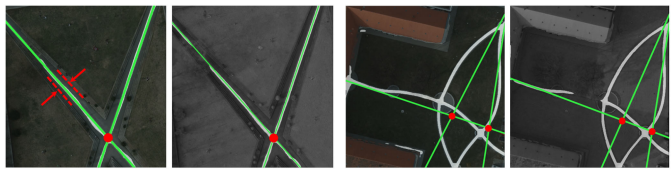


Fig. 15. Failure cases that illustrate the limitations of the proposed method. The red dotted lines indicate the width of actual road.

fitted by straight lines as shown in Fig. 15(b). 3) The proposed triangular feature construction and matching need more than three intersection points, so the method is not applicable to the situation where the number of intersection points is less than three as seen in Fig. 15(a) and (b).

IV. CONCLUSION

In this article, we proposed a fast and robust multimodal remote sensing image registration method via road intersection triangular feature. Our proposed registration method has mainly three stages: Automatic road lines extraction from multimodal urban remote sensing images, road intersection triangular feature construction, and triangular feature matching. The experimental results show that the proposed method is more efficient and robust than other state-of-the-art methods. Even when other comparison methods completely fail to achieve the registration task of cross-modal image pairs, the proposed method still maintains good robustness and matching accuracy. The qualitative and quantitative comparisons on different multimodal urban remote sensing image datasets demonstrate that our method has superiority over the comparison methods.

Although the proposed registration method has many advantages, there are some limitations and failure cases that have been analyzed and discussed. In future, we will further study how to improve the adaptation to the condition of fewer road intersection points, we can assign a method that only one or two road intersections are used to construct a matching structure. Multiscale feature integration can be adopted to combine features of different scales to improve the robustness of feature extraction under the conditions of complex backgrounds.

ACKNOWLEDGMENT

The authors would like to thank Y. Zhu and H. Liu for providing SpaceNet data as training datasets. The authors declare no conflicts of interest.

REFERENCES

[1] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, 2019.

[2] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2020.3012548.

[3] T.-C. Su, "A study of a matching pixel by pixel (MPP) algorithm to establish an empirical model of water quality mapping, as based on unmanned aerial vehicle (UAV) images," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 58, pp. 213–224, 2017.

[4] K. Yu, J. Ma, F. Hu, T. Ma, S. Quan, and B. Fang, "A grayscale weight with window algorithm for infrared and visible image registration," *Infrared Phys. Technol.*, vol. 99, pp. 178–186, 2019.

[5] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 23–79, 2020.

[6] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, and J. Tian, "Robust feature matching for remote sensing image registration via locally linear transforming," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, pp. 6469–6481, Dec. 2015.

[7] J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo, "Locality preserving matching," *Int. J. Comput. Vis.*, vol. 127, pp. 512–531, 2018.

[8] R. Li, H. Zhao, X. Zhang, X. Ge, Z. Yuan, and Q. Zou, "Automatic matching of multispectral images based on nonlinear diffusion of image structures," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 762–774, 2021.

[9] 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.

[10] M. Gong, S. Zhao, L. Jiao, D. Tian, and S. Wang, "A novel coarse-to-fine scheme for automatic image registration based on SIFT and mutual information," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 4328–4338, Jul. 2014.

[11] S. Dawn, V. Saxena, and B. D. Sharma, "Advanced free-form deformation and Kullback–Lieblier divergence measure for digital elevation model registration," *Signal, Image Video Process.*, vol. 9, pp. 1625–1635, 2015.

[12] J. Ma, X. Jiang, J. Jiang, and Y. Gao, "Feature-guided Gaussian mixture model for image matching," *Pattern Recognit.*, vol. 92, pp. 231–245, 2019.

[13] J. Ma, J. Jiang, H. Zhou, J. Zhao, and X. Guo, "Guided locality preserving feature matching for remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4435–4447, Aug. 2018.

[14] B. Fang, K. Yu, J. Ma, and P. An, "EMCM: A novel binary edge-feature-based maximum clique framework for multispectral image matching," *Remote. Sens.*, vol. 11, 2019, Art no. 3026.

[15] S. A. K. Tareen and Z. Saleem, "A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, AND BRISK," *in Proc. Int. Conf. Comput., Math. Eng. Technol.*, 2018.

[16] H. Bay, T. Tuytelaars, and L. Gool, "Surf: Speeded up robust features," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.

[17] Y. Ye and L. Shen, "HOPC: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. III-1, pp. 9–16, 2016.

[18] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, "Robust registration of multimodal remote sensing images based on structural similarity," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2941–2958, May 2017.

[19] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, 1988, pp. 147–152.

[20] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Trans. Image Process.*, vol. 29, pp. 3296–3310, 2020.

[21] G. LoweDavid, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[22] B. S. Manjunath, J. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 703–715, Jun. 2001.

[23] C. A. Aguilera-Carrasco, F. Barrera, F. Lumbreras, A. Sappa, and R. Toledo, "Multispectral image feature points," *Sensors*, Basel, Switzerland, vol. 12, pp. 12 661–12 672, 2012.

[24] C. A. Aguilera-Carrasco, A. Sappa, and R. Toledo, "LGHD: A feature descriptor for matching across non-linear intensity variations," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 178–181.

[25] J. Bian, W.-Y. Lin, Y. Matsushita, S. Yeung, T. 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. 2828–2837.

[26] H. Kamangir, M. Momeni, and M. Satari, "Automatic centerline extraction of coverd roads by surrounding objects from high resolution satellite images," vol. XLII-4/W4, pp. 111–116, 2017.

[27] J. D. D. Jayaseeli and D. Malathi, "An efficient automated road region extraction from high resolution satellite images using improved Cuckoo search with multi-level thresholding schema," *Procedia Comput. Sci.*, vol. 167, pp. 1161–1170, 2020.

[28] F. Bastani *et al.*, "Roadtracer: Automatic extraction of road networks from aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4720–4728.

[29] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[30] X. Yang, X. Li, Y. Ye, R. Y. K. Lau, X. Zhang, and X. Huang, "Road detection and centerline extraction via deep recurrent convolutional neural network U-Net," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7209–7220, Sep. 2019.

[31] S. Saito, T. Yamashita, and Y. Aoki, "Multiple object extraction from aerial imagery with convolutional neural networks," *J. Imag. Sci. Technol.*, vol. 60, 2016, Art no. 0 10402.

[32] Albu, "SpaceNet round 3 winner: Albu's implementation," 2018. [Online]. Available: https://github.com/spacenetchallenge/roaddetector/tree/master/albu-solution

[33] A. V. Etten, "City-scale road extraction from satellite imagery V2: Road speeds and travel times," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1775–1784.

[34] "Spacenet on Amazon web services (AWS). "datasets" the spacenet catalog," 2018. [Online]. Available: https://spacenetchallenge.github.io/datasets/datasethomepage.html

[35] "Potsdam dataset of remote sensing images, distributed by the international society for photogrammetry and remote sensing," [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/2d-semlabel-potsdam.html

[36] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[37] D. Viswanathan, "Features from accelerated segment test (fast)," 2011.

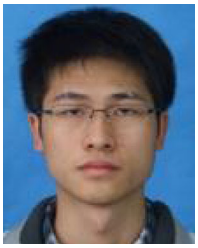
[38] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1469–1472.

Kun Yu received the B.Sc. and M.Sc. degrees in measurement control technology and instruments from Hubei University of Technology, Wuhan, China, in 2015 and 2018, respectively. He is currently working toward the Ph.D. degree in control science and engineering with the School of Artificial Intelligence and Automation of Huazhong University of Science and Technology, Wuhan, China.

His research interests include computer vision, multimodal/multispectral image matching, image fusion, and 3D point cloud registration.

Xiao Zheng received the B.Sc. and M.Sc. degrees in mechanical design from Hubei University of Technology, Wuhan, China, in 2014 and 2017, respectively.

Since 2021, she has been a Engineer with Wuhan Institute of Marine Electric Propulsion, Wuhan, China. Her research interests include image processing, image matching, and image quality evaluation.



Bin Fang received the bachelor's degree from Huazhong University of Science and Technology, Wuhan, China, where he is currently working toward the doctor's degree with the School of Artificial Intelligence and Automation, since 2014 .

His research interest includes point cloud description, recognition, and registration.

Pei An received the bachelor's degree from Wuhan Institute of Technology, Wuhan, China. He is currently working toward the doctor's degree with the School of Artificial Intelligence and Automation of Huazhong University of Science and Technology, Wuhan, China.

His research interest includes 3D object detection.

Xiao Huang received the B.E. degree in automation and the Ph.D. degree in control science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2009 and 2018, respectively.

He is currently an Engineer with China Ship Development and Design Center, Wuhan, China. His research interests include image processing, computer vision, and computer graphics.

Wei Luo received the B.Sc. degree in computer science from Wuhan University of Technology, Wuhan, China, in 2002, and the M.Sc. and Ph.D. degrees in computer science from Huazhong University of Science and Technology, Wuhan, China, in 2005 and 2008, respectively.

He is currently an Associate Researcher with the Department of Innovation Center, China Ship Development and Design Center, Wuhan. His research interests are reliability-aware scheduling, cluster and fault tolerant computing, parallel and distributed systems, real-time/embedded systems, and artificial intelligence.

Junfeng Ding received the bachelor's degree from Huazhong University of Science and Technology, Wuhan, China, where he is currently working toward the doctor's degree with the School of Artificial Intelligence and Automation, since 2018.

His research interests include 2D/3D matching and point cloud registration.

Zhao Wang received the bachelor's degree from Huazhong University of Science and Technology, Wuhan, China, where he is currently working toward the master's degree with the School of Artificial Intelligence and Automation, since 2018 .

His research interest includes 3D target detection, recognition, and tracking.

Jie Ma received the Ph.D. degree in control science and engineering with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China, in 2004.

He is currently a Professor with the School of Artificial Intelligence and Automation of Huazhong University of Science and Technology, Wuhan, China. He has more than 30 years of experience in image processing. His research interests include computer vision, pattern recognition and machine learning. His research field is also involved in 3D LiDAR target detection and recognition, 3D environment perception, image registration, image fusion, image enhancement, and video stabilization.