ISPRS Journal of Photogrammetry and Remote Sensing 192 (2022) 189–214

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

A robust registration method for UAV thermal infrared and visible images taken by dual-cameras

Lingxuan Meng [a], Ji Zhou [a,*], Shaomin Liu [b], Ziwei Wang [a], Xiaodong Zhang [c,d], Lirong Ding [a], Li Shen [e], Shaofei Wang [c,d]

[a] School of Resources and Environment, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China
[c] Shanghai Aerospace Electronic Technology Institute, Shanghai 201109, China
[d] Shanghai Spaceflight Institute of TT&C and Telecommunication, Shanghai 201109, China
[e] Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 610031, China

ARTICLE INFO

Keywords:
Template matching
Pyramid similarity maps
Multilevel local max-pooling
Homography estimation
Image registration
UAV remote sensing

ABSTRACT

Automatic registration of unmanned aerial vehicle (UAV) thermal infrared and visible (TIR&V) images is fundamental for subsequent applications. However, few studies address this issue due to significant radiation gap, shape gap, and texture gap among TIR&V images. The area-based methods are not able to satisfy the accuracy and robustness of location at the same time, while the image pyramid-based methods are computationally expensive. To alleviate these problems, we proposed a so-called TWMM method for the registration of UAV TIR&V images taken by the camera equipped with both thermal infrared and visible sensors. TWMM is realized by combining Template matching with Weights, Multilevel local max-pooling, and Max index backtracking. TWMM consists of four steps: (1) computing similarity maps of the atomic patches using template matching with weights; (2) building pyramid similarity maps using multilevel local max-pooling; (3) deducing the corresponding points (CPs) from top to bottom using max index backtracking; and (4) eliminating outliers and estimating homography. Among the four steps, step 1 and step 2 are used to compute the similarity maps of patches with different sizes; step 3 and step 4 are used to deduce CPs and estimate homography with multiple similarity maps. TWMM was comprehensively evaluated with 600 UAV image pairs under four different scenes and also compared with current methods (i.e. SIFT, SURF, RIFT, RCB, TFeat, HardNet, RANSAC_Flow, HOPC, and CFOG). These image pairs have multiple features, i.e., different land covers, spatial resolutions, and illumination conditions, etc. Results indicate that TWMM achieves an 86.0% correct CP ratio (RCP) and a 96.0% correct matching rate (CMR) for all test images, which is a 15.1% improvement and 11.6% improvement, respectively, over the best state-of-the-art methods. TWMM also shows better robustness than other methods for weak-light images, achieving a 20.7% improvement in RCP and a 28.1% improvement in CMR. Therefore, TWMM is an effective and robust method for UAV TIR&V image registration and has good ability under different scenes.

1. Introduction

With the rapid development of UAV technology in recent years, thermal infrared and visible (TIR&V) images of large areas can be collected efficiently and economically (Meng et al., 2021; Zhao et al., 2021). TIR&V images can provide thermal and reflectance attributes of ground objects. It is useful to register and synergistically use TIR&V images in many fields (Xiang et al., 2019), including precision agriculture (Messina and Modica, 2020), wildlife protection (Chrétien et al., 2015), infrastructure inspection (Escobar-Wolf et al., 2018), emergency rescue (Ambrosia et al., 2003), building facade thermal attribute mapping (Lin et al., 2019), and defect detection (Li et al., 2017). For example, Poblete et al. (2018) and Zhang et al. (2019) extracted the thermal attributes of pure vegetation while avoiding background influences from the registered TIR&V images. The thermal attributes of vegetation can be applied to calculate the crop water stress index (Maes et al., 2016; Santesteban et al., 2017), predict the yield (Maimaitijiang et al., 2020), estimate the biomass (Khanal et al., 2017), and detect rice

* Corresponding author at: No. 2006, Xiyuan Ave, West Hi-Tech Zone, Chengdu 611731, Sichuan, China.
E-mail addresses: 201811070106@std.uestc.edu.cn (L. Meng), jzhou233@uestc.edu.cn (J. Zhou), smliu@bnu.edu.cn (S. Liu), oneziway@163.com (Z. Wang), bobtennis@sina.com (X. Zhang), dlryouxiang@163.com (L. Ding), lishen@swjtu.edu.cn (L. Shen), 201822070311@std.uestc.edu.cn (S. Wang).

https://doi.org/10.1016/j.isprsjprs.2022.08.018
Received 2 June 2021; Received in revised form 19 August 2022; Accepted 19 August 2022
Available online 28 August 2022

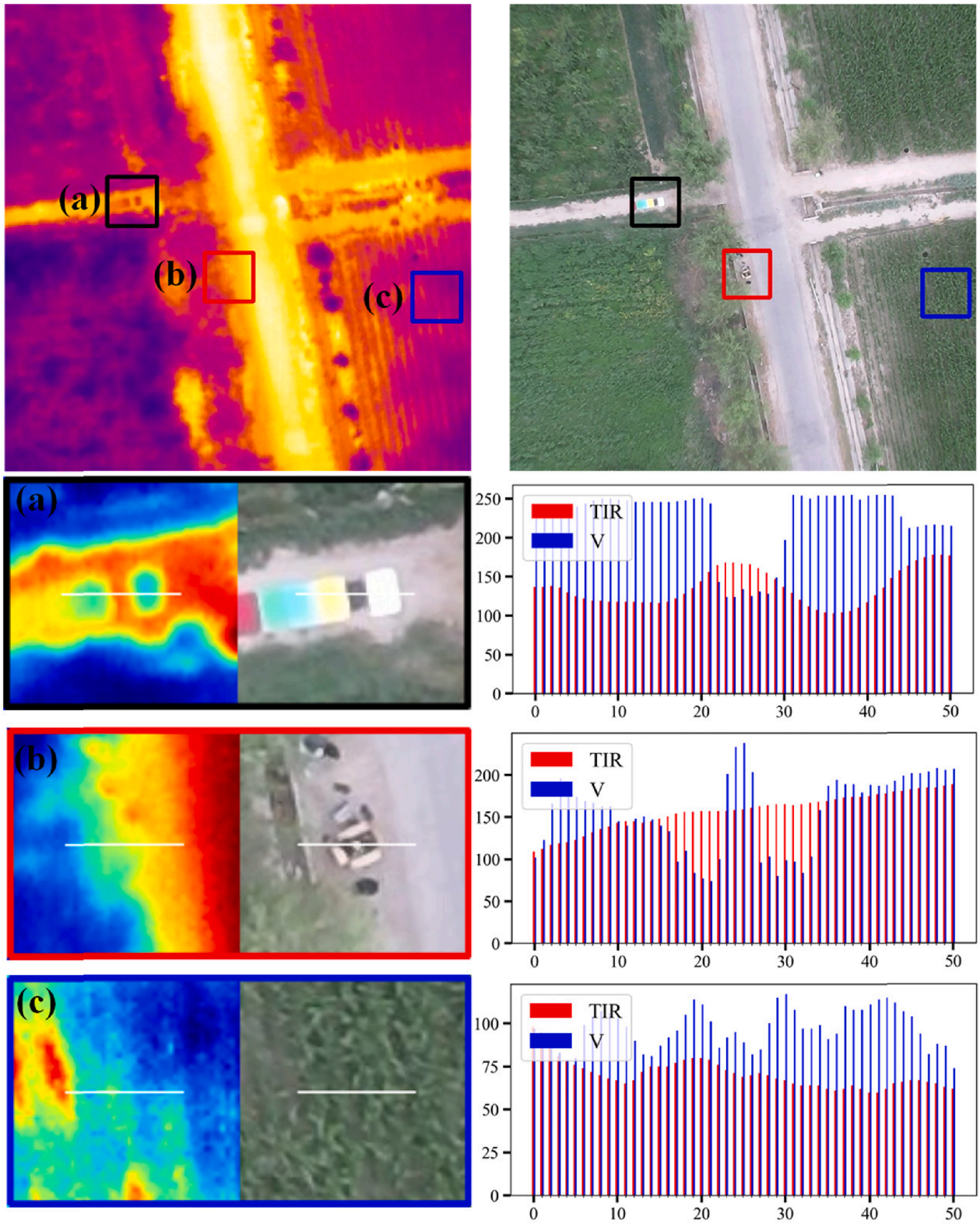


Fig. 1. Example of the UAV TIR&V images. The bar charts are the normalized pixel values on the white profiles in the enlarged TIR&V images. TIR in the bar charts means thermal infrared and V means visible. There are significant and complex radiation gaps (a), shape gaps (b), and texture gaps (c) between the TIR&V images.

lodging (Liu et al., 2018). By fusing TIR&V images, Sinha et al. (2015) and Sun and Schulz (2015) successfully improved the land cover classification accuracy by approximately 6 %. With thermal attributes, the status of the detected object can be better identified, which is useful for UAV-based real-time applications, such as pipeline leakage detection (Zhong et al., 2019) and rescue (Rudol and Doherty, 2008). Most thermal infrared (TIR) cameras, including FLIR DUO R, DJI Zenmuse XT2, and WIRIS Pro Sc, have two sensors, i.e., a TIR sensor and a visible sensor. With these cameras, one can simultaneously capture TIR&V images in one shooting. With prior knowledge, the spatial resolution

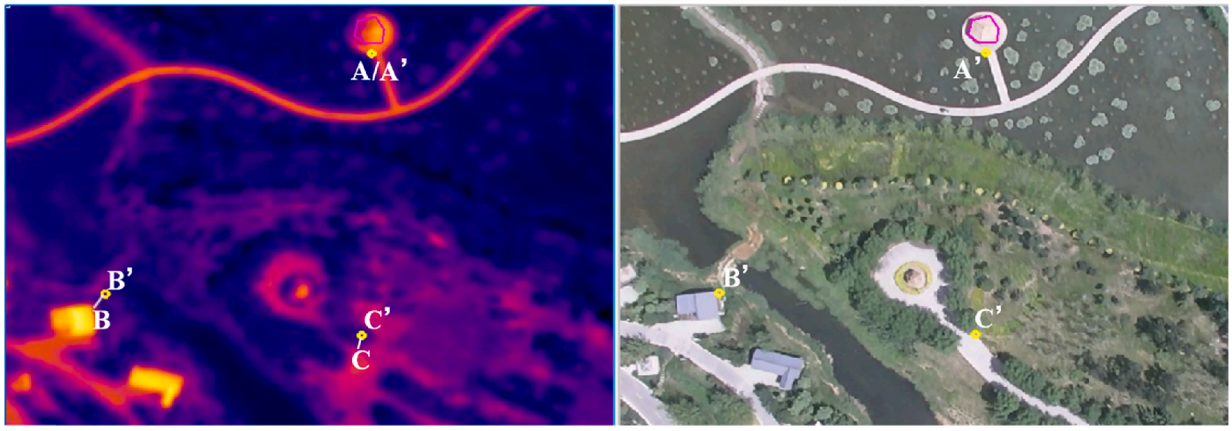


Fig. 2. An example of the deformation between the TIR&V images. Point B in the TIR image is the CP of B' in the visible image and point B' in the TIR image has the same coordinates as point B' in the visible image. This rule applies to point A and point C.

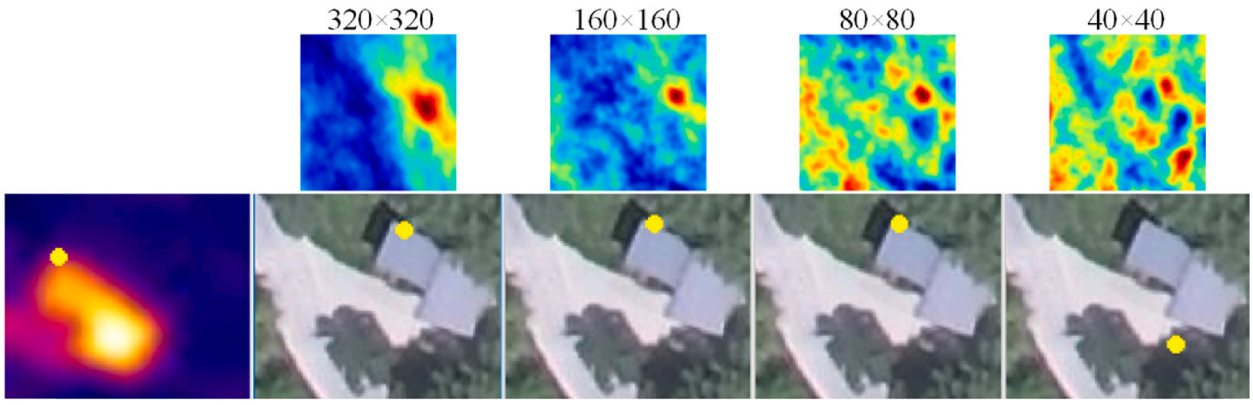


Fig. 3. An example of template matching with templates of different sizes, including 40 × 40, 80 × 80, 160 × 160, and 320 × 320. The first row shows the similarity maps of different templates. The second row is the results of different templates and the yellow points are the deduced CPs with different patches. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

difference between TIR&V images can be eliminated by resampling. However, there are still offsets of dozens of pixels between TIR&V images due to the different locations of the lenses, which further hinders the synergistic use of TIR&V images. Thus, developing a robust and effective method for registering UAV TIR&V images is an urgent issue.

Fig. 1 shows an example of UAV TIR&V images. The bar charts are the normalized pixel values on the white profiles in the enlarged TIR&V images. As seen from the bar charts, the differences in pixel intensity, gradient value, and gradient variation between the TIR&V images are significant. There are significant and complex radiation gaps (Fig. 1a), shape gaps (Fig. 1b), and texture gaps (Fig. 1c) between the TIR&V images. The radiation gap, shape gap, and texture gap mean that the same object has different radiation values, shapes, and textures in TIR&V images. In this study, the three aforementioned gaps are defined as RSTG. TIR images characterize the temperature of an object, which is completely different from the surface reflectance of the object provided by visible images (Liao et al., 2022; Ma et al., 2021; Zhang et al., 2021). It is understandable that one object has different radiation values and shapes in TIR&V images. In addition, the texture in TIR images is generally much coarser than that in visible images due to the thermal diffusion among ground objects. The RSTG decreases the correlation between the TIR&V images, making registration more challenging.

To register the TIR&V images, we employed some of the methods and had the following findings. The feature-based methods, including SIFT (Lowe, 1999) and RIFT (Li et al., 2020), can only get very few correct corresponding points (CPs) for most TIR&V image pairs, which are insufficient to evaluate the transformation model accurately. The area-based methods, including CFOG (Ye et al., 2019) and HOPC (Ye et al., 2017), require that the spatial resolution between images is approximately the same. Considering that there is a big spatial resolution gap between the TIR&V images, these area-based methods cannot be directly applied to the registration of TIR&V images. The unsupervised machine learning methods, including (Shen et al., 2020) and SCB (Cao et al., 2020), failed to achieve satisfactory performance because the difference in pixel intensity and gradient value between TIR&V images makes the inherent features unstable.

Under this context, we further analyzed the TIR&V images and had the following findings. First, the rotational difference between TIR&V images is small. This property is determined with the installation of the TIR sensor and visible sensor and works with most TIR&V dual-cameras, such as FLIR DUO R, DJI Zenmuse XT2, and WIRIS Pro Sc. Second, the spatial resolution difference between TIR&V images can be calculated from the sensor parameters. For example, according to the user manual of the WIRIS Pro Sc camera, the resolution is 0.13 m/pixel for the TIR sensor and 0.06 m/pixel for the visible sensor when the flight altitude is 100 m. Thus, up-sampling the TIR image by a factor of 2.3 can roughly eliminate the spatial resolution difference between the TIR&V images. We further center-crop the TIR&V images to eliminate the size difference between TIR&V images. Third, after up-sampling and cropping, the biases between the TIR&V images can be limited to dozens of pixels. The biases include a few rotation biases, a few scale biases, a few perspective biases, and dozens of pixels of translation biases. We term the non-translational biases as deformation in this work. Fig. 2 shows an example of the deformation. In Fig. 2, point B in the TIR image is the CP

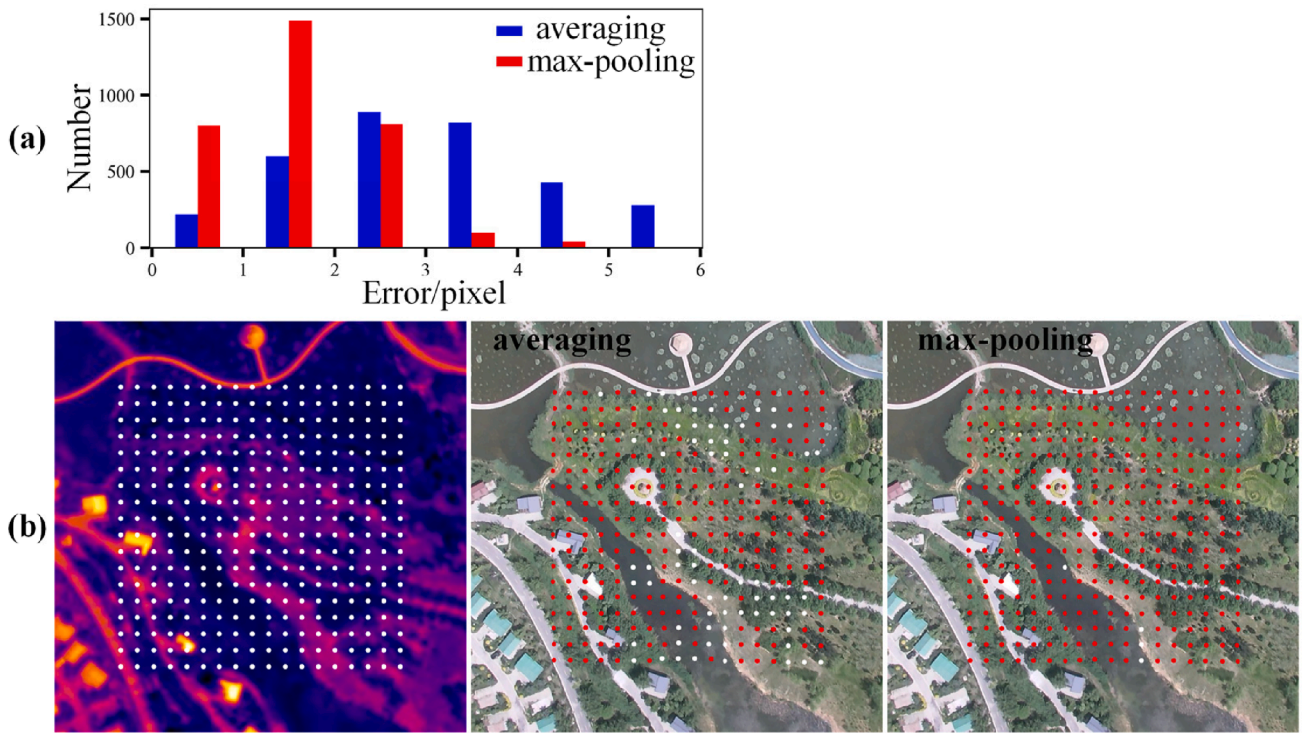


Fig. 4. The comparison of different aggregation methods (i.e., averaging and max-pooling). (a) the error statistics of different aggregation methods in ten images. The horizontal axis is the error value in pixels and the vertical axis is the number of pixels in different error intervals. (b) An example of deducing CPs with different aggregation methods. The similarity maps of the 320 × 320 patches are first calculated by different aggregation methods and the CPs are then deduced based on these similarity maps. With a threshold of 5, the correct CPs are marked in red and the incorrect CPs are marked in white. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of B' in the visible image, and point B' in the TIR image has the same coordinates as point B' in the visible image. This rule also applies to point A and point C. As seen, point A in TIR images is aligned with its CP A` in visible images by translating the TIR image, but there are still biases between point B in the TIR image and point B' in the TIR image due to the deformation. With the above findings, we decide to use the area-based method to find CPs between TIR&V images for the following two reasons: (1) with the prior knowledge that the biases between TIR&V images are dozens of pixels, the search range of template matching can be limited to dozens of pixels, which can reduce the mismatching compared with entire-image search; and (2) each point in the TIR image can meet its CP in the visible image during the template matching. Thus, the area-based method can avoid the problem of the low repetition rate of key points in the feature-based method.

In the area-based method, the size of the template window affects the registration performance significantly. Considering that we do not know the optimal window size in advance, we tried four different sizes in template matching, including 40 × 40, 80 × 80, 160 × 160, and 320 × 320. According to the experimental results, each of the four sizes has its advantages and disadvantages. As seen in Fig. 3, the value of the similarity map of small patch (i.e., 40 × 40) changes quickly, and thus, has high localization accuracy. However, there is insufficient information in small patch, causing low localization robustness. The large patch (i.e., 320 × 320) has the opposite performance to the small patch, with low accuracy and good robustness in localization. We can refer to the receptive field of deep learning to understand the properties of small and large patches. In deep learning, the large receptive field contains more global information, which has good robustness and low accuracy in localization (Sun et al., 2019). To improve both the robustness and accuracy in localization, we need to combine the information of multiple sizes.

Referring to the image pyramid-based methods (Baltsavias, 1991; Gruen, 2012, 1985; Jean-Yves, 2001; Lucas and Kanade, 1981), we explored the way to combine the information of templates of multiple sizes, which can be divided into two steps: (1) at each key point, we perform template matching with templates of different sizes to get multiple similarity maps; and (2) we use the coarse-to-fine strategy to deduce the CP of the key point. The similarity map of the 320 × 320 patch gives the approximate localization, and then the similarity map of the 160 × 160 patch refine around the approximate CP to obtain a more accurate localization. The similarity map of the 80 × 80 patch and the 40 × 40 patch continues the refining process. This coarse-to-fine strategy takes advantage of both the robustness of large patches and the accuracy of small patches. However, this method is computationally expensive as we need to perform template matching multiple times on each point.

To reduce computation, we analyzed the process of template matching with templates of different sizes. We found that the large patch with the size L × L can be divided into 4 sub-patches with the size L/2 × L/2. The value of the similarity map of the large patch at (m, n) is the average of the values of the similarity maps of the 4 sub-patches at (m, n). In this way, we can only perform template matching with the 40 × 40 patch and successively deduce the similarity maps of the 80 × 80 patch,160 × 160 patch, and 320 × 320 patch by aggregating the information from their four sub-patches. However, considering the deformation between images (please see Fig. 2), it is hard for the large patch to align every pixel with its CP in the target image. Thus, just averaging the similarity maps of sub-patches may decrease the accuracy of the similarity maps of the large patch due to the deformation.

To alleviate the effects of deformation, we expected the large patch to have a certain degree of freedom, rather than an inseparable whole. With this idea, we regarded that the large patch consists of many atomic patches, and the atomic patch has the following two properties: (1) the atomic patch is small enough that the deformation in the patch is less than 1 pixel, that is, the influence of deformation can be ignored for the atomic patch in template matching; and (2) the further the atomic

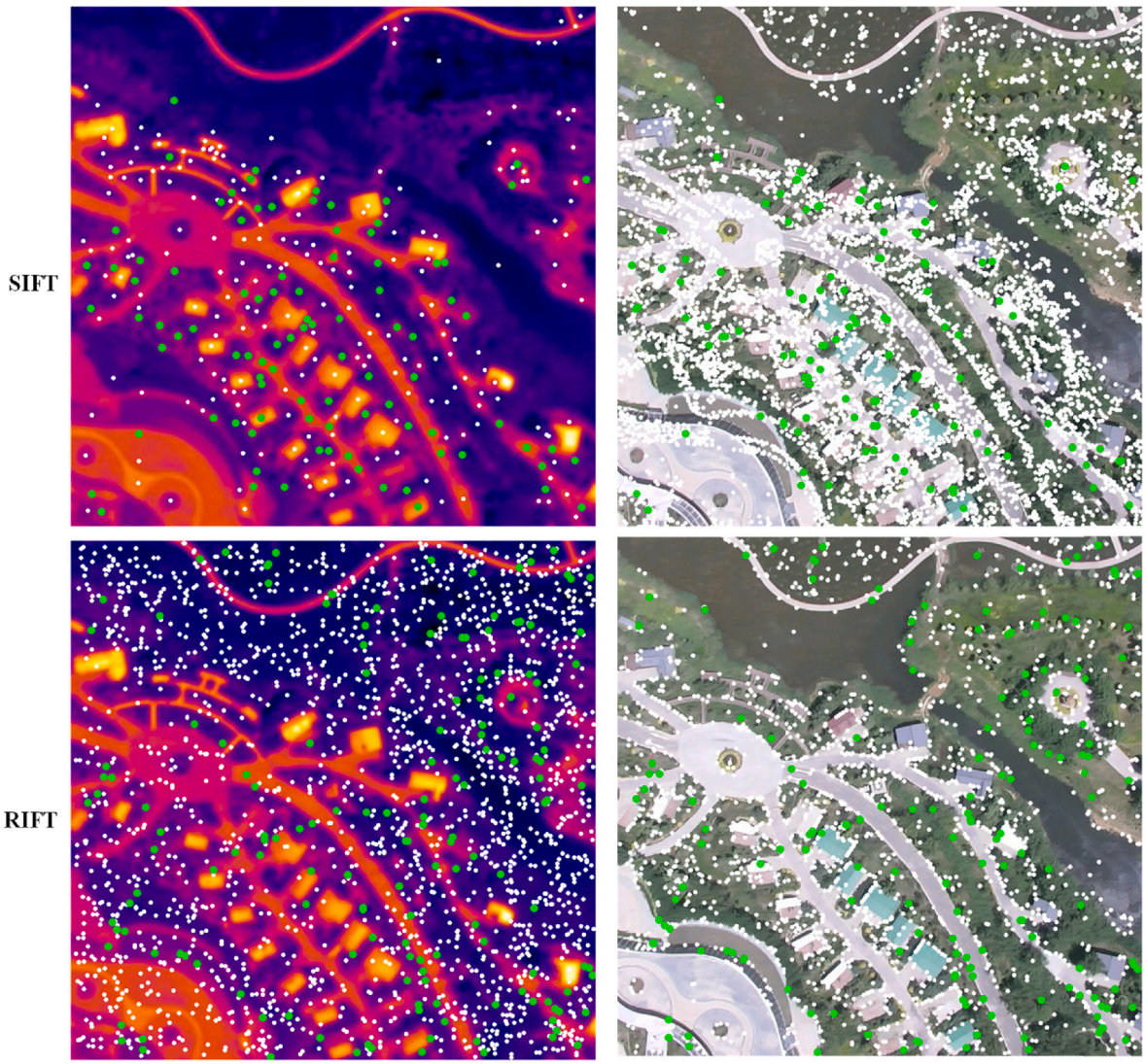


Fig. 5. The key points detected by SIFT and RIFT in the TIR&V images. The key point is considered repeatable if there is a key point detected within 5 pixels away from its CP in the other image. The repeatable key points are marked in green and the non-repeatable key points are marked in white. There are 372/7683 key points detected in the TIR/visible image by SIFT, and 110/145 of them are repeatable. For RIFT, there are 2000/2000 key points detected in the TIR/visible image by RIFT, and 198/199 of them are repeatable. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

patches are from each other, the more deformation there is between them. Different atomic patches can have different optimal corresponding locations to address the deformation between atomic patches.

Therefore, the similarity map of patches of different sizes can be calculated as follows: the 40×40 patches are regarded as the atomic patches and their similarity maps are calculated by template matching. The similarity maps of other patches are calculated by aggregation with max-pooling. The aggregation means that for the large patch, we first perform the 3×3 max-pooling on the similarity maps of its 4 sub-patches separately, and then average the pooled similarity maps to generate the similarity map of the large patch. We tested different aggregation methods (i.e., simple averaging and max-pooling) in 10 images. Fig. 4(a) shows the error statistics of different aggregation methods and Fig. 4(b) is an example of deducing CPs with different aggregation methods. Fig. 4 demonstrates that max-pooling is better than simple averaging. Therefore, we use max-pooling to aggregate information in our work. Note that we can adjust the parameters of max-pooling according to the degree of deformation, and the 3×3 max-pooling allows 2 pixels deformation for adjacent atomic patches.

After obtaining the similarity maps of different patches, we can successively deduce the correspondences of small patches from top to bottom referring to the coarse-to-fine strategy in the images pyramid-based methods. Furthermore, considering a small patch is contained by its four parent patches, we propose a scoring strategy, which improves the robustness of CPs by utilizing the high redundancy between small patch and its parent patches.

In summary, this study proposed a so-called TWMM method to register the TIR&V images taken by the camera equipped with both TIR and visible sensors. TWMM is realized by combining template matching with weights, multilevel local max-pooling (MLM), and max index backtracking (MIB). First, TWMM calculates the similarity maps of all atomic patches by template matching with weights. Second, pyramid similarity maps with different patch sizes are constructed by MLM. Last, the CPs between the TIR&V images are obtained by MIB. TWMM was comprehensively evaluated with 600 UAV image pairs under four different scenes and also compared with current methods (i.e. SIFT, SURF, RIFT, RCB, TFeat, HardNet, RANSAC_Flow, HOPC, and CFOG). Experiments demonstrate that TWMM outperforms other methods for UAV TIR&V image registration and can achieve satisfactory performance in different scenes. The datasets and source code will be released at https://github.com/mlxljz/TWMM.

2. Related work

The widely employed methods for registration of UAV TIR&V images can be divided into manual mode and automatic mode. In manual mode, the user labels the CPs one by one carefully. These CPs are used to estimate the transformation model between the TIR&V images. For example, some studies put control boards on the ground beforehand to improve the labeling accuracy (Messina and Modica, 2020; Zhang et al., 2019). However, manual labeling is costly and time-consuming for processing mass images. In addition, manual labeling may produce errors caused by human subjectivity.

Automatic mode can be approximately divided into three categories (Gruen, 2012): feature-based, area-based, and machine learning-based methods. Feature-based methods extract the features of salient objects to identify key points in two images and match points according to their feature distances. There are three steps in the feature-based methods, including key point detection, feature description, and key point matching. SIFT (Lowe, 1999), SURF (Bay et al., 2006), and ORB (Rublee et al., 2011) are three commonly employed feature-based methods in many registration studies. They employ the comparisons between pixel values to detect key points on both the reference image and the target image. The descriptor dimensions of SIFT, SURF, and ORB are 128, 64, and 256, which characterize the histogram of local gradient directions, the local gradient value, and the pixel value, respectively. In key point matching, the K-nearest neighbors algorithm is employed to obtain the top two best matches for every key point of the reference image. The ratio between the best match's distance and the second-best match's distance is applied to filter unsure CPs.

Although SIFT, SURF, and ORB have been widely used in many applications, they still have two limitations in TIR&V image registration. First, as shown in Fig. 1, the pixel intensities and the gradient values vary significantly between the TIR&V images, degrading the performance of the key point detection. In other words, the repeatability rate of the key points on TIR&V images drops due to the inconsistency of the pixel values and pixel gradient between TIR&V images. As shown in Fig. 5, the number of key points detected by SIFT in the TIR image is far smaller than that in the visible image. More importantly, the repeatability rate of the key points in the visible image is less than 2 %, which further affects the performance of key point matching. Second, SIFT, SURF, and ORB calculate the feature distance for every key point in the reference image with all key points in the target image. That is, each key point in the reference image searches for its CP in the entire target image. This entire-image search can cover large-scale transformations and rotation transformations. However, if the deviation between the TIR&V images is limited within a certain range, the entire-image search may not be suitable for TIR&V image registration due to the increasing possibility of mis-matching.

Considering the difference in pixel values and gradients between the TIR&V images, some feature-based methods have been proposed to address the nonlinear radiation distortion. For example, Ye et al. (2018) proposed a feature descriptor termed as the local histogram of oriented phase congruency, which utilizes an extended phase congruency to describe the features. Chen et al. (2019) proposed a rotation invariant feature descriptor that encodes edge information based on the multi-orientation and multi-scale Log-Gabor filter. Li et al. (2020) proposed a method referred to as the radiation-variation insensitive feature transform (RIFT), which uses phase congruency to detect key points and builds maximum index maps to describe features. Chen et al. (2022) further improved the RIFT in image preprocessing and feature description. In image processing, Chen et al. (2022) utilizes the homomorphic filtering to denoise and enhance the details of the TIR images. In feature description, Chen et al. (2022) proposed a binary pattern string based on 8 orientations Log-Gabor filter sequence to obtain more distinguishable descriptors than RIFT. These methods have been tested in different types of multimodal images, including infrared–visible, radar-visible, depth-visible, etc. Compared with SIFT, SURF, and ORB, these methods use the frequency domain information for key point detection and feature description, instead of the spatial domain information (i.e., the pixel value and gradient). We call these methods the frequency domain-based methods. The frequency domain-based methods construct the phase congruency or the maximum index map via Log-Gabor wavelet transformation, which can pay more attention to the corners, edges, textures of objects. Therefore, the frequency domain-based methods have high invariance to nonlinear radiation distortions and are more robust in key point detection and feature description. However, considering the shape gap and texture gap between the TIR&V images, the improvement of these frequency domain-based methods in key point detection is still limited. As shown in Fig. 5, less than 10 % of the key points are repeatedly detected in the TIR&V images. In addition, the frequency domain-based methods do not improve the key point matching process, and thus, still suffer from the mis-matching problem of the entire-image search.

Area-based methods identify CPs through template matching. The template window searches for CPs by evaluating the similarity of each location in the specified search area. For each search, the location with the maximum similarity is selected as the CP. Some area-based methods use pixel values to evaluate similarity, including the normalized cross-correlation (NCC) (Zhao et al., 2006), mutual information (MI) (Hirschmuller, 2008), and sum of squared difference (SSD) (Hisham et al., 2015). Considering that pixel values are vulnerable to radiation gaps (Ye et al., 2019), NCC, MI, and SSD are not suitable for TIR&V image registration. Some studies extract pixel-wise features based on the structure and shape of images and calculate the similarity map of these pixel-wise features using template matching (Revaud et al., 2016; Weinzaepfel et al., 2013). For example, Ye et al. (2017) proposed a pixel-wise feature descriptor referred to as the histogram of oriented phase congruency (HOPC). HOPC extends the phase congruency model to obtain the orientation representation of each pixel. Ye et al. (2019) proposed a descriptor referred to as channel features of orientated gradients (CFOG), which uses oriented gradients to describe each pixel. CFOG measures the similarity of features using the fast Fourier transform in the frequency domain to reduce the computational cost. Although these area-based methods show satisfactory performance in multimodal image registration, their performance is limited by the size of the template window (Ye et al., 2019). On the one hand, when the template window is too small, the information of the template window is insufficient to reflect its distinctiveness. On the other hand, the computational cost increases dramatically as the template window becomes larger. The distance between the key point and the boundary of the image also increases as the template window becomes larger, making it difficult for the estimated projection model to cover a whole image (Ye et al., 2019). In addition, due to the geometric distortion in the template window, the registration accuracy cannot keep increasing as the template window becomes larger (Ye et al., 2019).

To reduce the difficulty in determining the optimal template window size, some studies utilize the image pyramid to deduce CPs (Baltsavias, 1991; Gruen, 2012, 1985; Jean-Yves, 2001; Lucas and Kanade, 1981). The process of the image pyramid-based methods can be divided into three steps: (1) building images with multiple resolutions by successively down-sampling the original image; (2) building the pyramid similarity maps by performing template matching in each resolution with specified window size; and (3) using the coarse-to-fine strategy to deduce CPs. The correspondence of one patch is generated by refining the approximate match given by its parent patch on the lower-resolution image. However, the continuous down-sampling in the image pyramid-based methods degrades the image quality and further reduces the accuracy of the similarity maps. In addition, performing template matching in each resolution is computationally expensive. Therefore, these image pyramid-based methods still have potential to be improved.

Machine learning methods utilize gradient descent to update parameters (Haskins et al., 2020). Most machine learning methods are based on deep learning in the registration field. Deep learning methods

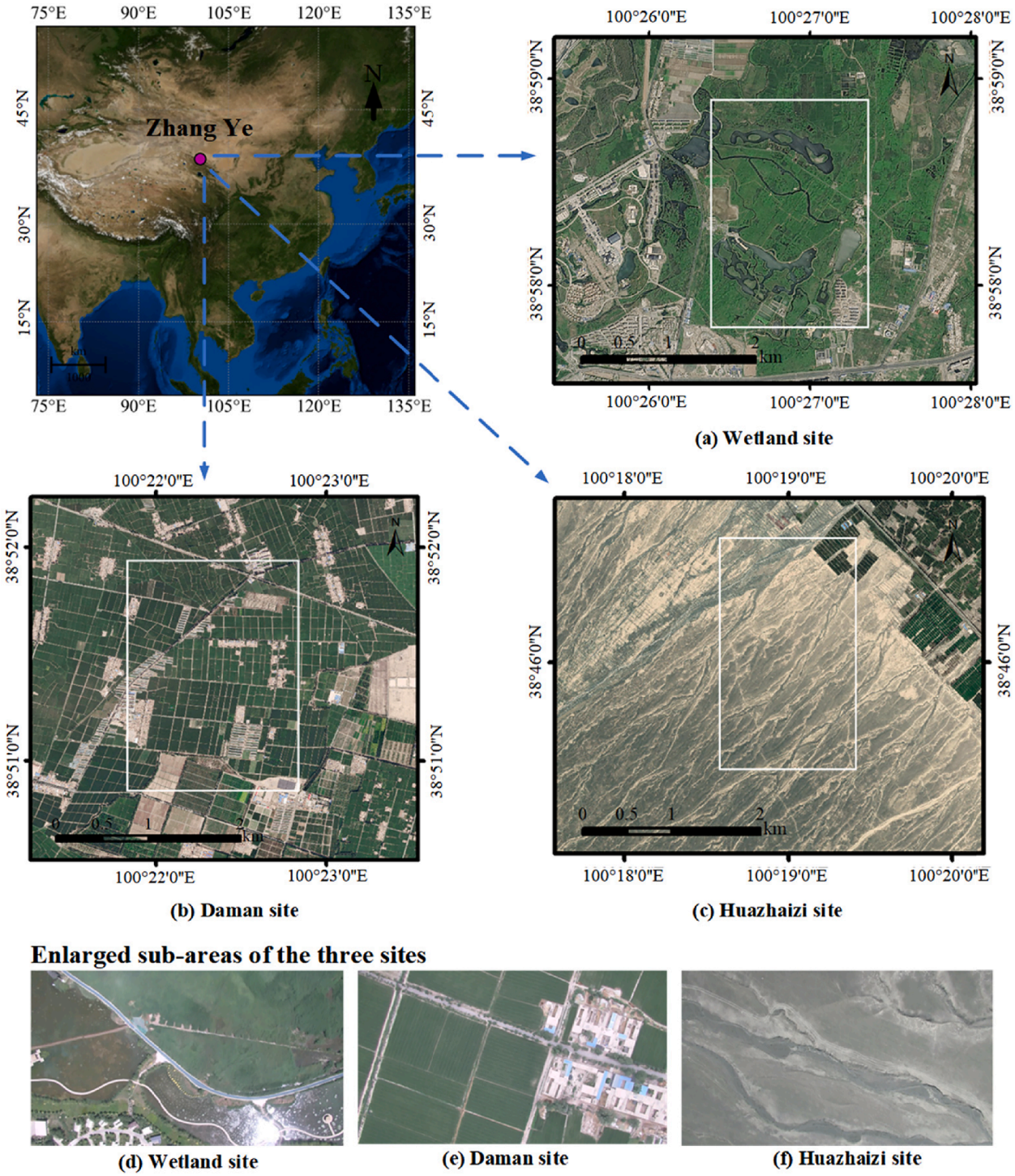


Fig. 6. Three experimental sites for collecting UAV TIR&V images: (a) Wetland, (b) Daman, and (c) Huazhaizi. The white rectangles at Wetland site, Daman site, and Huazhazi site represent the flight areas. (d), (e), and (f) are the enlarged sub-areas of the three sites.

can automatically learn feature description or projection model parameters (DeTone et al., 2016; Nguyen et al., 2018). To supervise the training, supervised methods, including TFeat (Balntas et al., 2016), LIFT (Yi et al., 2016), HardNet (Mishchuk et al., 2017), SDC (Schuster et al., 2019), and DualRC_Net (X. Li et al., 2020) use triplet margin loss based on ground truth correspondences; unsupervised methods, including unsupervised deep homography estimation (Nguyen et al., 2018), content-aware unsupervised deep homography estimation (Zhang et al., 2020), and RANSAC_flow (Shen et al., 2020), use photometric consistency loss based on pixel value or pixel gradient. Some non-deep learning methods directly optimize the parameters of the transformation model through multiple times of gradient descent. These methods do not utilize the intermediate layers, such as the fully connected layers, the convolutional layers, etc. For example, Kong et al. (2007) extracted and optimized the human face edges based on the canny detector, and then used the quasi-Newton method to directly optimize the 6 parameters of the affine model to maximize the overlap of the edges and the similarity of the shape features between visible image and warped TIR image. Similarly, Cao et al. (2020) proposed the structure consistency boosting (SCB) transform, which extracted the

(a) DJI M600 Pro (b) WIRIS Pro Sc

Fig. 7. DJI M600 Pro (a) and WIRIS Pro Sc camera (b) employed in the UAV remote sensing experiment.

Table 1
Instruments used in the UAV remote sensing experiment.

Instrument	Model	Specification
UAV	DJI M600 Pro	Maximum horizontal velocity: 18 m/s Maximum flight time with no load: 38 min Maximum load capacity: 6 kg
Cameras	WIRIS Pro Sc	Weight: 450 g Sensors: 2 (TIR, visible)Spectral Range (TIR) : 7.5–13.5 μmResolution (TIR) : ≈0.13 m at the flight height of 100 mResolution (visible) : ≈0.06 m at a flight height of 100 m Image Size (pixels): 640 × 512 (TIR), 1920 × 1080 (visible) Thermal Measurement Accuracy: ±2 degreesC

inherent edge structure of images by evaluating the salience of pixel gradients with respect to the local mean intensity in a neighborhood window. Then, Cao et al. (2020) employed gradient descent to optimize the parameters to minimize the structure difference between images. Kong et al. (2007) and Cao et al. (2020) are both unsupervised methods. These machine learning methods have achieved state-of-the-art performance in many registration applications (Wang et al., 2018; Zhang et al., 2020). However, it is very time-consuming for the supervised methods to collect and label a large number of TIR&V images. In addition, many kinds of photometric consistency in the unsupervised methods, including pixel intensity (Yu et al., 2016), SSIM (Meng et al., 2021; Wang et al., 2004), and SCB are hard to hold in TIR&V images due to the aforementioned RSTG. Therefore, many machine learning methods are unsuitable to register TIR&V images.

Some methods can improve the registration performance with auxiliary information. For example, Torabi et al. (2012) addressed the problem of TIR&V image registration with video information. It first performed human tracking in TIR and visible videos separately. Then, it treated the top points of the human as the trajectory points and performed the trajectory-to-trajectory matching to deduce CPs and estimate the transformation model. Maurya et al. (2020) placed the control board with square holes in advance and deduced CPs by the features of the square holes. However, considering that the auxiliary information is not easily available, these methods are not universally applicable to various registration tasks. Therefore, a robust registration method for UAV TIR&V images taken by dual-cameras is greatly needed.

3. Materials and datasets

As previously mentioned, UAV TIR&V images are useful for calculating the crop water stress index, predicting the yield, estimating the amount of biomass, etc. Many related studies have been conducted in the Heihe River Basin (HRB) in Northwest China within the framework of the Heihe Watershed Applied Telemetry Experimental Research (HiWATER) (Li et al., 2019, 2013; Liu et al., 2018; Ma et al., 2015; Xu et al., 2013). However, UAV TIR&V images in HRB are still very rare. Therefore, UAV remote sensing experiments were conducted in HRB from July to September 2020 (Meng et al., 2021). Three sites (i.e., Wetland, Daman, and Huazhaizi) with different land cover types, the core areas of HiWATER, were selected as the flight areas. As shown in Fig. 6, Wetland site is covered with large reeds, ponds, roads, etc. Daman site is covered with cropland, buildings, paths, etc. Huazhaizi site is covered with bare soil and short shrubs. In addition, the images were collected in different months; thus, even the images for the same site have different scenes due to the phenology.

The DJI Matrices 600 (M600) Pro UAV and the WIRIS Pro Sc camera

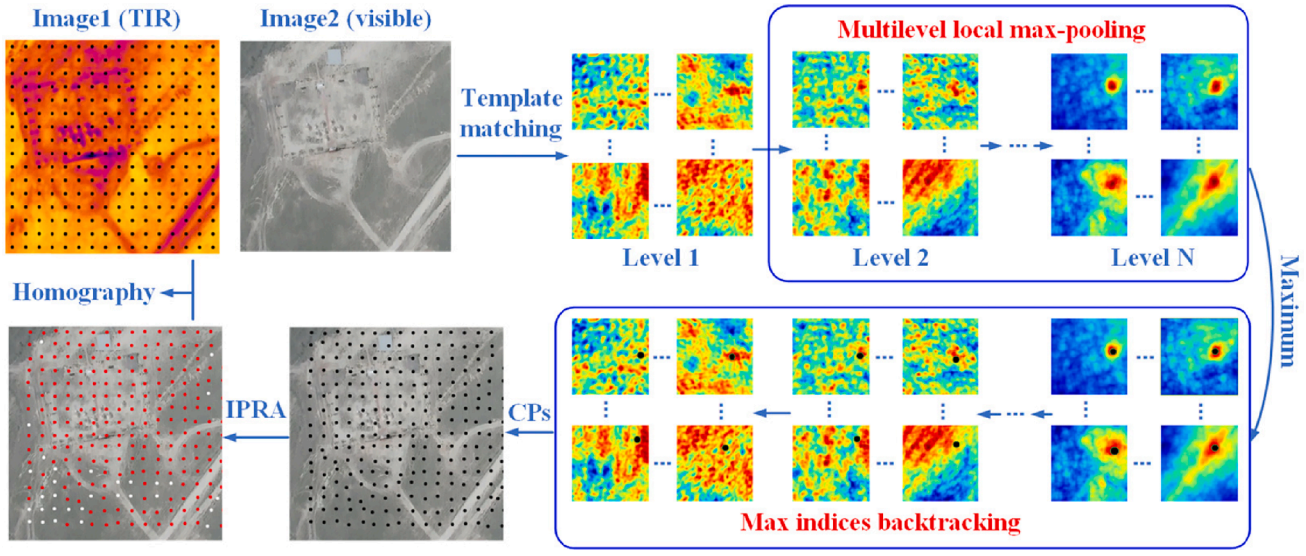


Fig. 8. The framework of TWMM. The similarity maps of Level 1 are calculated by template matching on CFOG features of Image 1 (TIR) and Image 2 (visible). The pyramid similarity maps from Level 2 to Level *N* are calculated by MLM (i.e., multilevel local max-pooling). In the pyramid similarity maps, the darker the red, the higher the similarity; the darker the blue, the lower the similarity. MIB (i.e., max index backtracking) is performed to deduce CPs from top to bottom. The black point in each similarity map in MIB represents the best correspondence of each patch deduced by MIB. The remaining CPs after outlier elimination are used to estimate the homography. On the bottom-left image, the removed CPs are marked by white points and the remaining CPs are marked by red points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

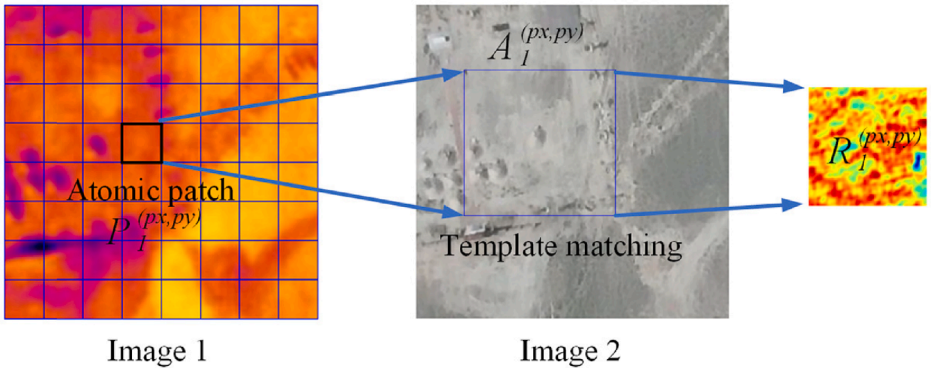


Fig. 9. Image 1 is divided into small, non-overlapping atomic patches with $S_a \times S_a$ pixels, and each atomic patch corresponds to a similarity map by performing template matching on Image 2. $P_1^{(px,py)}$ represents an atomic patch, $A_1^{(px,py)}$ represents the search area of $P_1^{(px,py)}$ in Image 2, and $R_1^{(px,py)}$ represents the corresponding similarity map of$P_1^{(px,py)}$.

were utilized to build the UAV thermal remote sensing system (Fig. 7 and Table 1). DJI M600 Pro is low-cost and highly stable, with an operation time of 35 min at a 500-g payload. The WIRIS Pro Sc camera has two sensors: a TIR sensor and a visible sensor. The camera can capture TIR&V images at one shooting. Due to the differences in lens locations, the offsets between the TIR&V images range from 20 to 60 pixels depending on the camera pose and flight altitude.

For all flights, the UAV flew at a speed range from 8 m/s to 15 m/s and at an altitude range from 80 m to 300 m above the ground. Flight missions were conducted under different weather conditions (including sunny, cloudy, and windy) and different daily time periods (including noon, afternoon, and gloaming) (Meng et al., 2021). Four sets of TIR&V images (i.e., Wetland set, Daman set, Huazhaizi set, and Dark-Daman set) were taken in the UAV remote sensing experiment, with each set containing 150 pairs of images. As previously mentioned, Wetland, Daman, and Huazhaizi sets have different land covers. The images of the Wetland set, Daman set, and Huazhaizi set were taken at noon or in the afternoon with a flight altitude of approximately 300 m, and thus, the ground spatial resolution (GSR) of the TIR images in the three sets was approximately 0.39 m. The images of the Dark-Daman set were captured at gloaming with a flight altitude of approximately 100 m, and thus, the GSR of the TIR images was approximately 0.13 m. In addition, the poor illumination conditions rendered the visible images in the Dark-Daman set dim and noisy.

4. Methodology

As mentioned previously, there is a large difference in spatial resolution and size between the TIR&V images. We use prior knowledge to eliminate the differences in resolution and size: first, the TIR images are upscaled by 2.3; second, the TIR&V images are center-cropped to 1000 × 1000 pixels to ensure that the TIR&V images have the same size. After the processing, there are still a few rotation biases, scale biases, perspective biases, and dozens of pixels of translation biases between the image pairs.

The framework of TWMM has four parts (Fig. 8): (1) computing similarity maps of atomic patches, including feature extraction and template matching with weights; (2) building pyramid similarity maps with MLM (i.e., multilevel local max-pooling), which can obtain global information while avoiding the effect of local geometric distortion; (3) deducing CPs with MIB (i.e., max index backtracking), which uses the similarity maps of all levels to increase the robustness and accuracy of the deduced CPs; and (4) eliminating outliers and calculating homography. The four parts will be illustrated in detail in the following sections.

4.1. Similarity maps of Level 1

4.1.1. Feature extraction

CFOG is selected as the feature descriptor in our experiment because of its robustness in multimodal images (Ye et al., 2019). CFOG uses oriented gradients to describe each pixel: first, the horizontal and vertical gradients are calculated by the Sobel kernel (Gao et al., 2010); second, each oriented gradient is the absolute value of the weighted sum of the horizontal and vertical gradients:

$$G_\theta = |\sin\theta \times G_h + \cos\theta \times G_v| \tag{1}$$

where G_θ is the oriented gradient; θ is the orientation; and G_h and G_v are the horizontal gradient and vertical gradient, respectively.

We divide 0 to 360 degrees into 9 orientations according to the realization of CFOG in Ye et al. (2019). Thus, θ is 0, 40…, and 320. The feature of the entire image is represented as $F^{H\times W\times 9}$, where H and W are the height of the image and width of the image, respectively, and 9 is the number of orientations. To suppress noise, $F^{H\times W\times 9}$ is filtered by a 3D Gaussian kernel in the horizontal direction, vertical direction, and gradient orientation direction. $F^{H\times W\times 9}$ is then normalized in the orientation dimension for each pixel to reduce the influence of the radiation gap between two images. For more details about the implementation of CFOG, readers can refer to Ye et al. (2019) or https://github.com/mlxljz/TWMM.

4.1.2. Template matching with weights

As shown in Fig. 8, we use Image 1 to represent the TIR image and Image 2 to represent the visible image. As shown in Fig. 9, Image 1 is divided into small and non-overlapping atomic patches with $S_a \times S_a$ pixels. Each atomic patch performs template matching in a local area of Image 2. Therefore, each atomic patch corresponds to a similarity map, and each value in the similarity map shows the similarity between the atomic patch and the corresponding area in Image 2. The similarity value at location (x, y) of the similarity map is calculated as:

$$sim(x,y) = 1 - \frac{dis(x,y) - \min(G_{dis})}{\max(G_{dis}) - \min(G_{dis})} \tag{2}$$

$$dis(x,y) = \sum_{x',y'} (F_a(x',y') - F_s(x+x',y+y'))^2 * W_a(x',y') \tag{3}$$

$$G_{dis} = \{dis(x,y) | x \in [0, H_s), y \in [0, W_s)\} \tag{4}$$

where (x', y') is a location in the atomic patch; F_a is CFOG feature of the atomic patch; F_s is CFOG feature of the search area in Image 2; W_a is the weights of the atomic patch; H_s and W_s are the height of the similarity map and width of the similarity map, respectively; $dis(x, y)$ is the

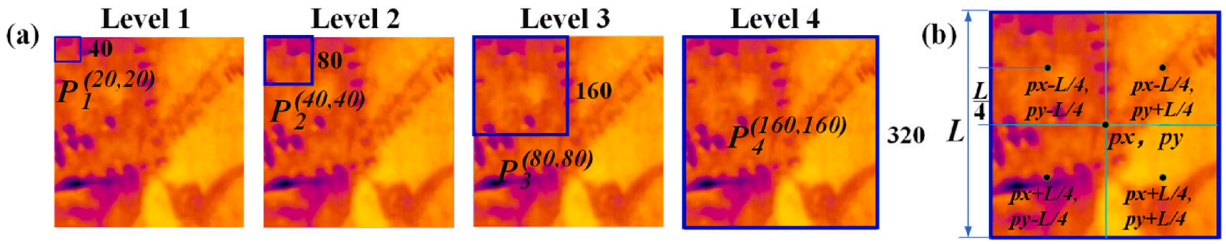


Fig. 10. The patches in different levels when S_a is set to 40 and N is set to 4. (a) The blue rectangles at Level 1, Level 2, Level 3 and Level 4 are $P_1^{(20,20)}$, $P_2^{(40,40)}$, $P_3^{(80,80)}$, and $P_4^{(160,160)}$, respectively. To obtain the similarity map of the large patch, the patch size doubles as the level increases. The patch size increases from 40 × 40 at Level 1 to 320 × 320 at Level 4. (b) A large patch with the size $L \times L$ at Level $n+1$ consists of 4 subpatches with the size $L/2 \times L/2$ at Level n. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

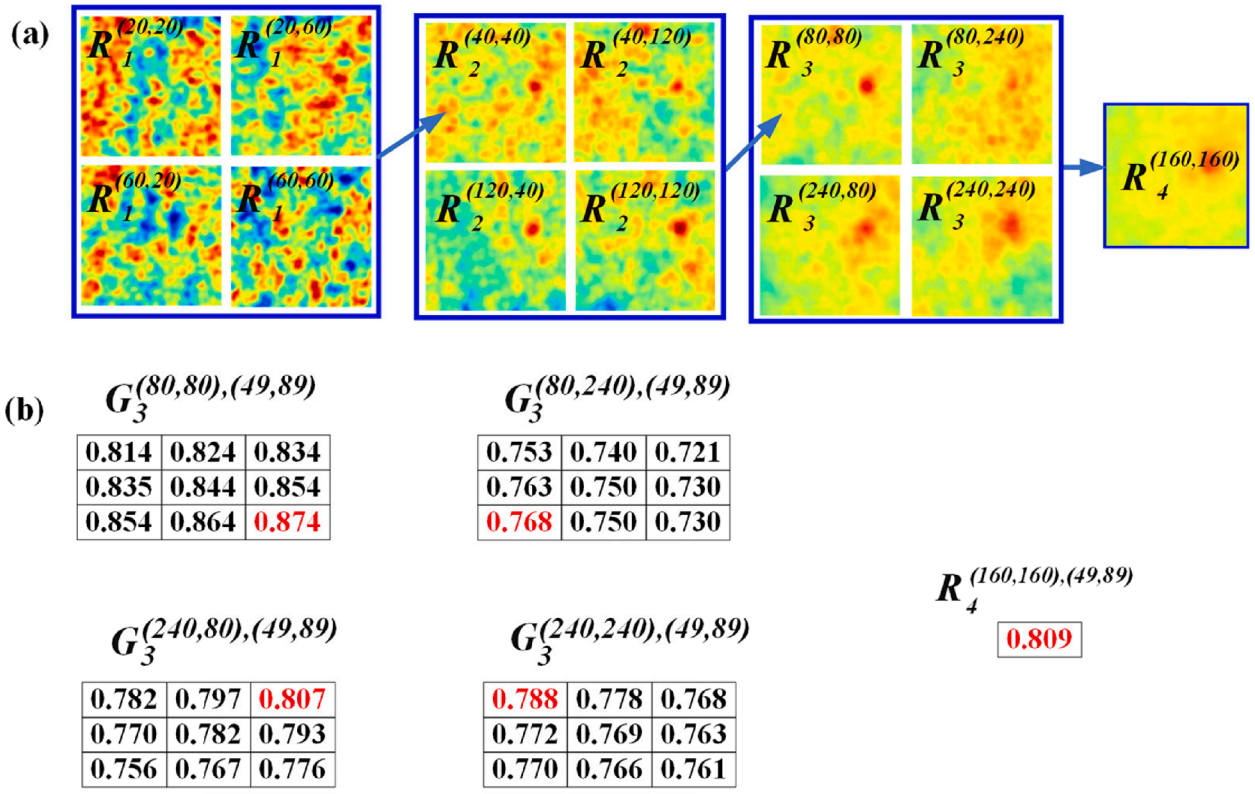


Fig. 11. MLM process when S_a is set to 40 and N is set to 4. (a) The similarity maps from Level 2 to Level 4 are calculated by MLM. The similarity maps of Level $n+1$ are generated from the similarity maps of Level n ($n \geq 1$) with local max-pooling. In the pyramid similarity maps, the darker the red, the higher the similarity; the darker the blue, the lower the similarity. (b) $R_4^{(160,160),(49,89)}$ is the mean of the maximum of $G_3^{(80,80),(49,49)}$, $G_3^{(80,240),(49,49)}$, $G_3^{(240,80),(49,49)}$, and $G_3^{(240,240),(49,49)}$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

weighted sum of the square difference of features, which represents the difference between the atomic patch and the corresponding area in Image 2; G_{dis} is a set containing all $dis(x, y)$ of the search area, which is applied to normalize $dis(x, y)$ by min–max normalization; and $sim(x, y)$ is the similarity value and equal to one minus the normalized $dis(x, y)$. Therefore, the location with the smallest feature difference has a maximum similarity value of 1, and the location with the largest feature difference has a minimum similarity value of 0.

In most existing methods, W_a in Eq. (3) is set to 1 for every location (Ye et al., 2019). However, given that it is difficult for CFOG features to be robust in every pixel between the TIR&V images, the similarity may suffer from the non-robustness of features when W_a is set to 1 for every location. Inspired by the notion that people pay attention to obvious areas when labeling CPs, we also want $dis(x, y)$ to give larger weights to obvious pixels, such as significant edges and corners. Given that the gradient is utilized to detect obvious points in many methods (Lowe, 1999), we proposed the weighting module, which uses the gradient value of each pixel in the atomic patch as the weight, namely, W_a is the gradient of the atomic patch in TWMM.

By the feature extraction and template matching that is previously described, we can obtain the similarity map set of Level 1, which is composed of similarity maps of all atomic patches. As shown in Fig. 8, the size of the similarity map set of Level 1 is $(H_p \times W_p) \times (H_s \times W_s)$, where $H_p \times W_p$ is the number of atomic patches, which is determined by the image size and the atomic patch size; $H_s \times W_s$ is the size of the similarity map, which is determined by the search radius of the template matching. The similarity maps of Level 1 provide a basis for building the similarity maps of the next levels.

4.2. Multilevel local max-pooling

We use N to represent the number of levels of the similarity maps; n

Table 2
Information of the similarity maps of all levels when the image size is 1000 × 1000 pixels and the atomic patch size, number of levels, and search radius of TWMM are set to be 40 × 40, 4, and 60, respectively.

Level	Patch number	Patch center	Patch size	Map size
R_4	16 × 16	x, y∈{160, 200 …, 800, 840}	320	121 × 121
R_3	21 × 21	x, y∈{80, 120, …, 840, 920}	160	121 × 121
R_2	24 × 24	x, y∈{40, 80, …, 920, 960}	80	121 × 121
R_1	25 × 25	x, y∈{20, 60, …, 940, 980}	40	121 × 121

Note: The patch number column represents the number of patches at each level. The patch center column represents the centers of the patches at each level. The patch size and map size represent the size of the patch and size of the similarity map, respectively, at each level.

to represent a particular level; $P_n^{(px,py)}$ to represent the patch centered at (px, py) at Level n in Image 1; $A_n^{(px,py)}$ to represent the search area of $P_n^{(px,py)}$ in Image 2; $A_n^{(px,py),(sx,sy)}$ to represent the area at the location (sx, sy) in $A_n^{(px,py)}$; R_n to represent all similarity maps of Level n; $R_n^{(px,py)}$ to represent the similarity map of $P_n^{(px,py)}$; and $R_n^{(px,py),(sx,sy)}$ to represent the similarity value located at (sx, sy) of $R_n^{(px,py)}$, which is the similarity between $P_n^{(px,py)}$ and $A_n^{(px,py),(sx,sy)}$.

The size of atomic patches is usually small to avoid local geometric distortions, causing the features of atomic patches to lack global information and the similarity maps of atomic patches to lack discrimination. Thus, it is very hard to obtain the correct CPs from similarity maps of atomic patches, which are shown in Fig. 8. To overcome this shortcoming, we use MLM to increase the size of the patch layer by layer and obtain more global information. The similarity maps of Level $n+1$ are generated from the similarity maps of Level n ($n \geq 1$) with local max-pooling.

Fig. 10 and Fig. 11 show the MLM process. As shown in Fig. 10(a), we double the patch size as the level increases to obtain the similarity map of the large patch, which makes the patch of Level $n+1$ contain 4 subpatches of Level n. Thus, the patch size increases from 40 × 40 in Level 1 to 320 × 320 in Level 4 when S_a is set to 40. As shown in Fig. 10 (b), the large patch with the size $L \times L$ at Level $n+1$ consists of 4 subpatches with the size $L/2 \times L/2$ at Level n. The relationship between the coordinates of the center points of the large patch and its 4 subpatches is:

$$cpx_i, cpy_i = (px, py) + o_i, \text{with} \begin{cases} o_1 = (\frac{L}{4}, \frac{L}{4}) \\ o_2 = (\frac{L}{4}, -\frac{L}{4}) \\ o_3 = (-\frac{L}{4}, \frac{L}{4}) \\ o_4 = (-\frac{L}{4}, -\frac{L}{4}) \end{cases} \quad (5)$$

where L is the height and width of the large patch; (px, py) is the center of the large patch; and (cpx_i, cpy_i) is the center of the i-th subpatch.

As shown in Fig. 11(a), we generate the patch's similarity map at Level $n+1$ by aggregating its 4 subpatches' similarity maps at Level n, namely, $R_{n+1}^{(px,py)}$ can be calculated by $R_n^{(cpx_1,cpy_1)}$, $R_n^{(cpx_2,cpy_2)}$, $R_n^{(cpx_3,cpy_3)}$, and $R_n^{(cpx_4,cpy_4)}$ as follows:

$$R_{n+1}^{(px,py),(sx,sy)} = \frac{1}{4}\sum_{i=1}^{4} \max(G_n^{(cpx_i,cpy_i),(sx,sy)}) \quad (6)$$

$$G_n^{(cpx_i,cpy_i),(sx,sy)} = \left\{ R_n^{(cpx_i,cpy_i),(sx+si,sy+sj)} \,\middle|\, si \in (-1,0,1),\ sj \in (-1,0,1) \right\} \quad (7)$$

where $G_n^{(cpx_i,cpy_i),(sx,sy)}$ represents the 3 × 3 neighborhood centered on (sx, sy) in $R_n^{(cpx_i,cpy_i)}$.

According to Eq. (6), $R_{n+1}^{(px,py),(sx,sy)}$ is the average of the maximum of $G_n^{(cpx_1,cpy_1),(sx,sy)}$, $G_n^{(cpx_2,cpy_2),(sx,sy)}$, $G_n^{(cpx_3,cpy_3),(sx,sy)}$, and $G_n^{(cpx_4,cpy_4),(sx,sy)}$, namely, $R_{n+1}^{(px,py)}$ is the average of the 3 × 3 max-pooling of $R_n^{(cpx_1,cpy_1)}$, $R_n^{(cpx_2,cpy_2)}$, $R_n^{(cpx_3,cpy_3)}$, and $R_n^{(cpx_4,cpy_4)}$. With local max-pooling, similarity maps of large patches can aggregate global information while avoiding the effect of local geometric distortion. After obtaining R_1 by template matching, as mentioned in Section 4.1, we can successively deduce R_2 to the R_N level by level according to Eq. (6) and Eq. (7).

We further describe the MLM process in detail. $P_2^{(40,40)}$ has 4 subpatches, $P_1^{(20,20)}$, $P_1^{(20,60)}$, $P_1^{(60,20)}$, and $P_1^{(60,60)}$, as shown in Fig. 10(a). Therefore, $R_2^{(40,40)}$ is generated by aggregating $R_1^{(20,20)}$, $R_1^{(20,60)}$, $R_1^{(60,20)}$, and $R_1^{(60,60)}$ according to Eq. (6) and Eq. (7), as shown in Fig. 11(a). We can generate $R_4^{(160,160)}$ in the same way as $R_2^{(40,40)}$. Fig. 11(b) shows the local max-pooling process described by Eq. (6) and Eq. (7). $R_4^{(160,160),(49,89)}$ is the mean of the maximum of $G_3^{(80,80),(49,49)}$, $G_3^{(80,240),(49,49)}$, $G_3^{(240,80),(49,49)}$, and $G_3^{(240,240),(49,49)}$, where 0.874 is the maximum of $G_3^{(80,80),(49,49)}$; 0.768 is the maximum of $G_3^{(80,240),(49,49)}$; 0.807 is the maximum of $G_3^{(240,80),(49,49)}$; 0.788 is the maximum of $G_3^{(240,240),(49,49)}$; and 0.809 is the mean of 0.768, 0.874, 0.807, and 0.788.

Table 2 shows the similarity maps of all levels with specified parameters. When the image size is 1000 × 1000 pixels and the atomic patch size, number of levels (N), and search radius of TWMM are set to 40 × 40, 4, and 60, respectively, the patch size in each level is 40 × 40, 80 × 80, 160 × 160, and 320 × 320, respectively. The number of patches at Level 1 is 25 × 25, which is equal to the image size divided by the atomic patch size. The step between two patches at each level is equal to the atomic patch size. The size of the similarity map is equal to 121 × 121, which is determined by the search radius. The atomic patch size and number of levels can be adjusted, and the influence of the two parameters is discussed in Section 5.2.

4.3. Max index backtracking

After obtaining the similarity maps of patches with different sizes, we need to obtain the CPs between two images. On the one hand, similarity maps of small patches are not suitable for obtaining CPs due to a lack of discrimination, as previously mentioned. On the other hand, similarity maps of very large patches are also unsuitable for obtaining CPs for two reasons. First, as the patch size increases, the global information of the similarity map increases while the locating accuracy decreases, which is a common problem for max-pooling (Sun et al., 2019). As shown in Fig. 11(a), the difference between the maximum and its neighbor value in the similarity map $R_4^{(160,160)}$ is very small. Therefore, the location of the maximum in $R_4^{(160,160)}$ may not be an accurate correspondence. Second, as the level increases, the patch size increases and the number of patches decreases (Table 2). Thus, we can only obtain a few CPs using large patches, which may fail to cover the complicated geometric transformation between two images. To overcome this drawback, we use MIB (i.e., max index backtracking) to obtain a sufficient number of accurate CPs between two images. We use $C_n^{(px,py),(cx,cy),score}$ to represent the correspondence between $P_n^{(px,py)}$ and $A_n^{(px,py),(sx,sy)}$, and the correspondence has an attribute (i.e., *score*) to represent the matching degree of $P_n^{(px,py)}$ and $A_n^{(px,py),(sx,sy)}$.

Fig. 12 shows the MIB process. As shown in Fig. 8 and Fig. 12(a), MIB starts from the highest level, continues level by level, and ends at Level 1. For the highest level (i.e., Level N), we build $C_N^{(px,py),(cx,cy),score}$ using the maximum of $R_N^{(px,py)}$, where (px, py) is the center of $P_N^{(px,py)}$, (sx, sy) is the location of the maximum of $R_N^{(px,py)}$, and the *score* is the maximum of

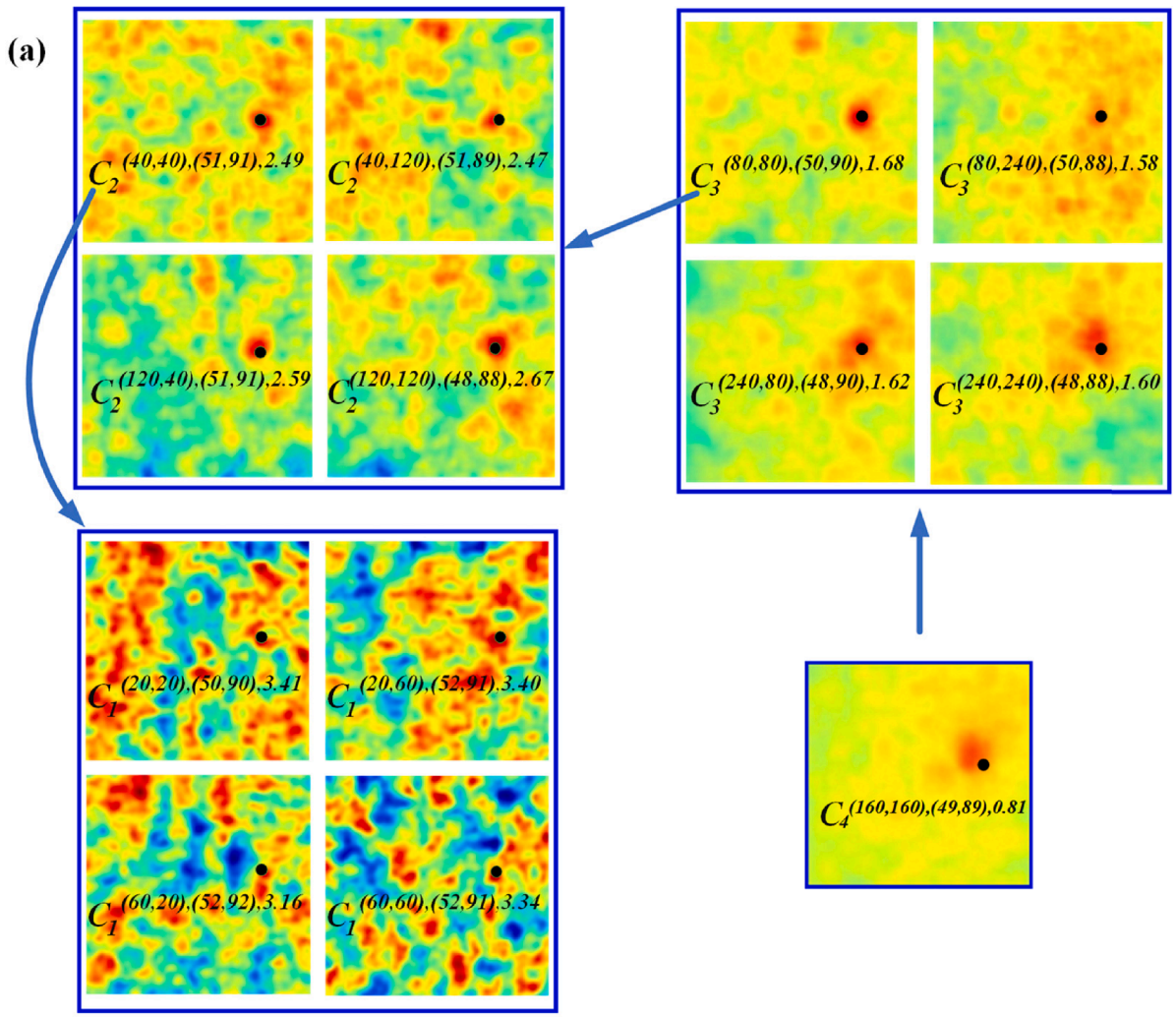


Fig. 12. MIB process when S_a is set to 40 and N is set to 4. (a) MIB can successively deduce the correspondences of atomic patches from top to bottom with the pyramid similarity maps. In the pyramid similarity maps, the darker the red, the higher the similarity; the darker the blue, the lower the similarity. (b) CPs can be deduced using the correspondences of atomic patches. (c) $C_4^{(160,160),(49,89),0.809}$ deduces 4 subcorrespondences $C_3^{(80,80),(50,90),1.683}$, $C_3^{(80,240),(50,88),1.577}$, $C_3^{(240,80),(48,90),1.617}$, and $C_3^{(240,240),(48,88),1.597}$ according to Eq. (8). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$R_N^{(px,py)}$. The correspondence at Level $n+1$ can backtrack 4 sub-correspondences at Level n, namely, $C_{n+1}^{(px,py),(cx,cy),score}$ can deduce 4 sub-correspondences (i.e., $\{ C_n^{(cpx,cpy),(csx,csy),cscore} \}_{i=1}^{4}$) by MIB as follows:

$$C_{n+1}^{(px,py),(sx,sy),score} \xrightarrow{\max \text{ indices}} \{ C_n^{(cpx,cpy),(csx,csy),cscore} \}_{i=1}^{4} \tag{8}$$

$$csx, csy = \underset{sx+i,sy+j}{\operatorname{argmax}} \left\{ R_n^{(cpx,cpy),(sx+i,sy+j)} \,\middle|\, i=(-1,0,1),\ j=(-1,0,1) \right\} \tag{9}$$

$$cscore = score + R_n^{(cpx,cpy),(csx,csy)} \tag{10}$$

where (csx, csy) is the max index of the 3×3 neighborhood at the location (sx, sy) of $R_n^{(cpx,cpy)}$; and $cscore$ is the sum of $score$ and $R_n^{(cpx,cpy),(csx,csy)}$.

After obtaining $C_N^{(px,py),(cx,cy),score}$, we can successively backtrack $C_{N-1}^{(px,py),(cx,cy),score}$ to $C_1^{(px,py),(cx,cy),score}$ level by level according to Eq. (8). In

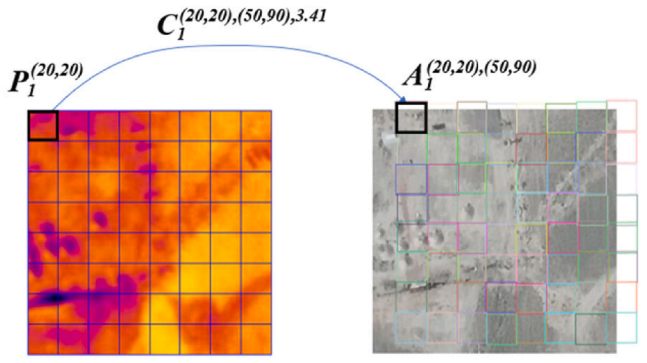


Fig. 13. CPs can be deduced using the correspondences of atomic patches.

Table 3
Results of TWMM with different S_a values when N is fixed to 4.

Index	S_a value			
	20	40	60	80
RMSE	2.82	**0.72**	1.43	1.76
CMR	81.3 %	**96.0 %**	92.5 %	88.4 %

Table 4
Results of TWMM with different N values when S_a is fixed to 40.

Index	N value			
	2	3	4	5
RMSE	4.83	2.80	**0.72**	1.52
CMR	68.5 %	85.3 %	**96.0 %**	91.3 %

Table 5
Results of TWMM with different features (i.e., CFOG, SIFT, ResNet).

Index	Features		
	CFOG	SIFT	ResNet
RMSE	0.72	**0.58**	2.8
CMR	96.0 %	**96.8 %**	92.7 %

the backtracking process, if multiple $C_n^{(px,py),(sx,sy),score}$ are equal in n and (px, py), we only retain the correspondence with the highest *score* and remove the other correspondences to accelerate the backtracking. As previously described, MIB uses the similarity maps of all levels, which can solve the problem of a lack of discrimination of small patches and low locating accuracies of large patches. After obtaining the correspondence for every atomic patch $P_1^{(px,py)}$, the CP can be deduced with $C_1^{(px,py),(sx,sy),score}$ as follows:

$$px_2 = px + sx - sr \tag{11}$$

$$py_2 = py + sy - sr \tag{12}$$

where (px, py) and (px_2, py_2) are the CPs between Image 1 and Image 2; and sr is the search radius during template matching. The number of CPs obtained by MIB is equal to the number of atomic patches, as shown in Fig. 13.

We further describe the MLM process in detail. As shown in Fig. 12 (a), $C_4^{(160,160),(49,89),0.809}$, which is marked by a black point in $R_4^{(160,160)}$, is the start of the MIB process. $C_4^{(160,160),(49,89),0.809}$ means that the similarity map $R_4^{(160,160)}$ has the maximum value of 0.809 at location (49,89). As shown in Fig. 12(b), $C_4^{(160,160),(49,89),0.809}$ deduces 4 subcorrespondences $C_3^{(80,80),(50,90),1.683}$, $C_3^{(80,240),(50,88),1.577}$, $C_3^{(240,80),(48,90),1.617}$, and $C_3^{(240,240),(48,88),1.597}$ according to Eq. (8), Eq. (9), and Eq. (10). Taking $C_3^{(80,80),(50,90),1.683}$ as an example, where $P_3^{(80,80)}$ is one of the subpatches of $P_4^{(160,160)}$ (please refer to Eq. (5)), (50,90) is the max index of $G_3^{(80,80),(49,49)}$ (please refer to Eq. (9)), 1.683 is the sum of 0.809 (i.e., the *score* of $C_4^{(160,160),(49,89),0.809}$) and 0.874 (i.e., $R_3^{(80,80),(50,90)}$) (please refer to Eq. (10)). With MIB, we can successively backtrack the correspondences of all atomic patches, such as $C_1^{(20,20),(50,90),3.41}$, $C_1^{(20,60),(52,91),3.40}$, $C_1^{(60,20),(52,92),3.16}$, and $C_1^{(20,20),(52,91),3.34}$ in Fig. 12(a). The CPs can be deduced using the correspondences of all atomic patches according to Eq. (11) and Eq. (12), as shown in Fig. 13.

4.4. Homography calculation

After obtaining the CPs between two images, we use the imprecise points removing algorithm (IPRA) to remove the outliers (Wu et al., 2015). In each iteration, IPRA calculates the transformation model by all CPs and then removes the CP with the largest transformation error. The iteration process will not terminate until all remaining CPs' transformation errors are smaller than the specified threshold. The specified threshold is set to 5 pixels in our experiment. As shown in Fig. 8, the removed CPs are marked by white points, and the remaining CPs are marked by red points.

We use homography to cover the geometric transformation between two images, which has been employed in many registration studies (Meng et al., 2021; Zhang et al., 2020). The homography has 8 degrees of freedom, which is parameterized by a 3 × 3 matrix (Baker et al., 2006):

$$\mathrm{H} = \begin{pmatrix} h1 & h2 & h3 \\ h4 & h5 & h6 \\ h7 & h8 & 1 \end{pmatrix} \tag{13}$$

where h_1, h_2, h_4, and h_5 are rotation terms and scale terms; h_3 and h_6 are translation terms; and h_7 and h_8 are perspective terms. The remaining CPs are used to calculate the homography by the least squares method (Marquardt, 1963).

4.5. Similarities and differences with other methods

Area-based and image pyramid-based methods are well studied in image registration. In the following, we will discuss the similarities and differences between TWMM and these methods.

Both TWMM and the area-based methods generate the similarity maps by template matching (Gruen, 2012; Ye et al., 2017). However, it is difficult for the single-size template to have both robustness and accuracy in localization for the area-based methods. To alleviate this problem, TWMM uses the small patches to perform template matching, the MLM to construct the pyramid similarity maps, and the MIB to deduce CPs. In general, TWMM alleviates the geometric distortion problem of the area-based method to a certain extent, and thus, improves the area-based methods. At the same time, TWMM alleviates the difficulty in determining the appropriate template window size.

To alleviate the effect of geometric distortion, both TWMM and the image pyramid-based methods construct the pyramid similarity maps with different patch sizes and deduce CPs from top to bottom on the pyramid similarity maps (Baltsavias, 1991; Gruen, 2012, 1985; Jean-Yves, 2001; Lucas and Kanade, 1981). The main differences between TWMM and the image pyramid-based methods include two aspects. First, to construct the pyramid similarity maps, TWMM builds the similarity maps of Level 1 using template matching and builds the similarity maps of other levels with max-pooling; the image pyramid-based methods first build the image pyramid with down-sampling and then build the pyramid similarity maps by performing template matching in

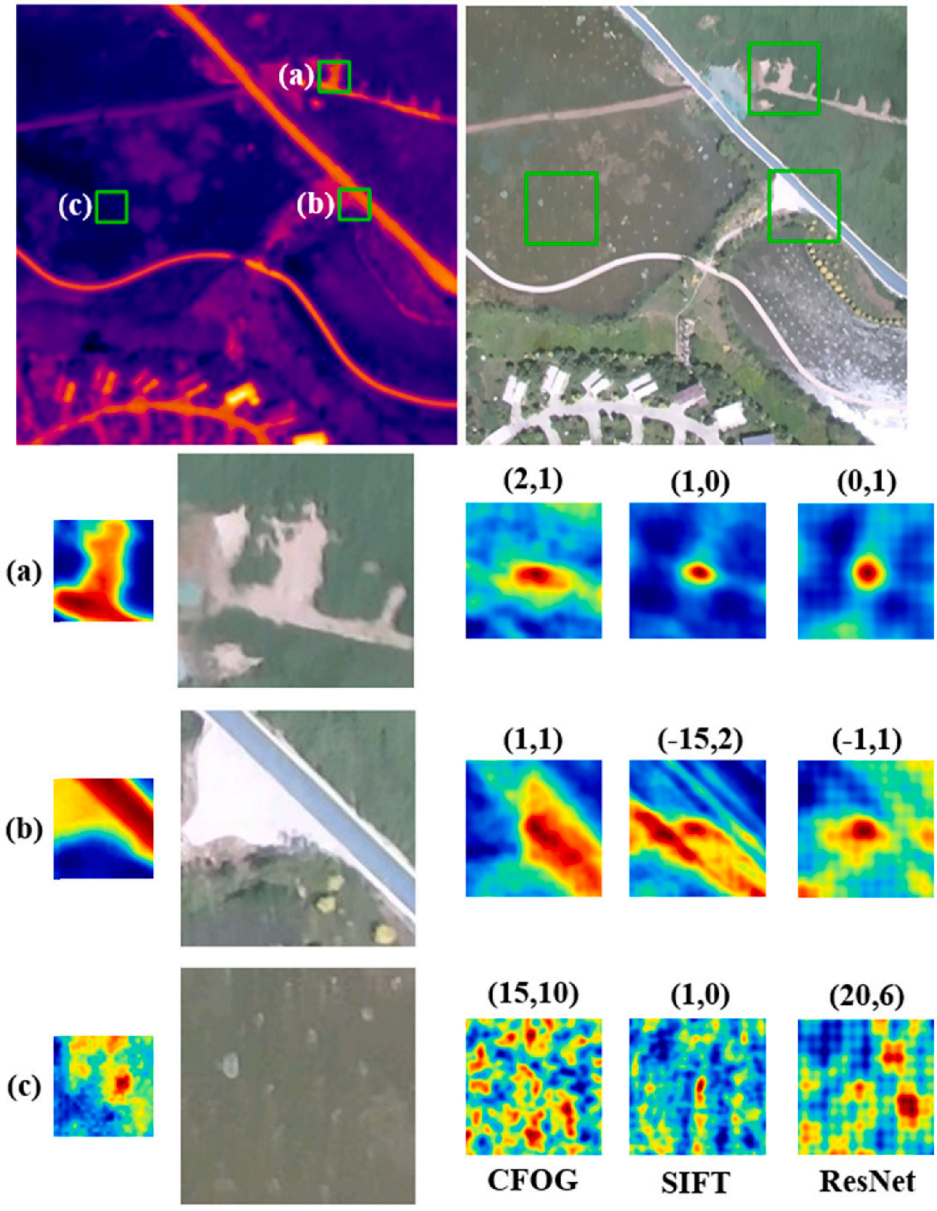


Fig. 14. Similarity maps obtained by template matching with different features. The patch size used for template matching is 40 × 40 and the search radius is 30. The number above the similarity map is the error in the horizontal direction and vertical direction between the correct corresponding location and the calculated corresponding location.

each level. Compared with max-pooling, the continuous down-sampling degrades the image quality seriously, causing the similarity maps of the image pyramid-based methods less accurate than TWMM. In addition, the multiple times of template matching in the image pyramid-based methods requires more computational cost compared with TWMM. Second, to deduce CPs with the pyramid similarity maps, TWMM uses the scoring strategy, namely, the correspondence of one patch is decided by comparing the score of the paths from its four parent patches; the image pyramid-based methods use the coarse-to-fine strategy, namely, the correspondence of one patch is generated by refining the approximate match given by its parent patch. The difference in deducing CPs is mainly due to the high matching redundancy of TWMM, i.e., the patch of Level n is contained by four parent patches of Level $n + 1$ in TWMM while only one parent patch in image pyramid-based methods. In other words, TWMM can derive the correspondence of the atomic patch from multiple paths and select the one with the highest score. Overall, the high matching redundancy makes TWMM more robust than the image pyramid-based methods.

5. Results

5.1. Evaluation metrics

As mentioned in Section 3, we took four sets of TIR&V images in HRB under different scenes to comprehensively evaluate TWMM and other methods. Each set contains 150 pairs of TIR&V images. To obtain the quantitative results of all methods, a ground truth homography is needed for every image pair. In general, we need to manually label the CPs. However, it is very difficult to manually label CPs accurately due to the significant RSTG. Therefore, we use a semiautomatic method to obtain the ground truth homography. First, we use multiple methods (TWMM, CFOG, RIFT, and SIFT) to estimate CPs and remove the outliers according to the IPRA method. The remaining CPs are applied to calculate the homography by the least square method. Second, we

Table 6
Quantitative results of all methods. The higher the CCP, RCP, and CMR and the lower the RMSE are, the better the method. The bold numbers show the best performance in all methods. The symbol - means larger than 100 in column RMSE and smaller than 1% in column CMR. The symbol * means that the method does not have the TCP, CCP, and RCP metrics.

Method	Wetland					Daman					Huazhaizi					Dark-Daman				
	TCP	CCP	RCP	RMSE	CMR	TCP	CCP	RCP	RMSE	CMR	TCP	CCP	RCP	RMSE	CMR	TCP	CCP	RCP	RMSE	CMR
SIFT	8	5	67.5 %	–	2.7 %	9	5	58.8 %	–	3.3 %	10	1	10 %	–	1.3 %	3	0	0	–	–
SURF	17	5	30.0 %	–	1.3 %	20	4	22.0 %	–	1.3 %	14	1	5 %	–	–	7	0	0	–	–
RIFT	27	16	61.4 %	–	3.3 %	12	5	40.0 %	–	4.6 %	10	4	40.5 %	–	4.0 %	5	0	0	–	–
SCB	*	*	*	–	2 %	*	*	*	–	2.0 %	*	*	*	–	1.3 %	*	*	*	–	–
TFeat	12	7	61.0 %	–	4.0 %	16	9	56.2 %	–	4.7 %	9	2	18.2 %	–	1.3 %	5	1	9 %	–	–
HardNet	16	10	63.8 %	–	6.7 %	22	13	58.6 %	–	8.0 %	9	2	22.0 %	–	4.0 %	6	1	15.0 %	–	3.3 %
RANSAC_Flow	*	*	*	–	7.3 %	*	*	*	–	5.3 %	*	*	*	–	6.6 %	*	*	*	–	4.0 %
HOPC	11	3	32.0 %	–	3.3 %	11	5	39.6 %	–	3.3 %	10	2	23.4 %	–	4.7 %	11	2	16.8 %	–	2.0 %
CFOG	289	213	73.5 %	1.15	91.3 %	325	244	75.1 %	1.36	93.7 %	289	205	71.0 %	1.35	90.0 %	163	104	64.2 %	4.50	62.6 %
TWMM	333	**285**	**85.7 %**	**0.40**	**98.7 %**	361	**310**	**86.0 %**	**0.57**	**98.7 %**	351	**307**	**87.5 %**	**0.82**	**96.0 %**	211	**179**	**84.9 %**	**1.08**	**90.7 %**

carefully check the registration performance achieved by the calculated homography. If the registration is inaccurate, we will remove some incorrect CPs or manually label some CPs to ensure the accuracy of the ground truth homography.

For methods that estimate the homography by determining the CPs between two images (i.e., TWMM, RIFT, SIFT, TFeat, HardNet, HOPC, and CFOG), we can obtain the determined CPs. We compute the residuals of these CPs under the ground truth homography and take the CPs with residuals less than 5 pixels as the correct CPs. The average number of total CPs (TCP), average number of correct CPs (CCP), and correct CP ratio (RCP) are applied to quantitatively evaluate the CP detection results. The RCP is equal to the CCP divided by the TCP. In general, the greater the CCP and the higher the RCP, the better the method.

We use the average root mean square error (RMSE) over all images and correct matching ratio (CMR) to evaluate the registration result. The CMR is equal to the number of correct matching images divided by the number of total images. For each image, we evenly selected 36 points in the image and calculated the RMSE according to:

$$RMSE = \sqrt{\frac{1}{36}\sum_{i=1}^{36}(rxi)^2 + (ryi)^2} \tag{14}$$

where rx and ry are the residuals (in pixels) in the x-direction and y-direction, respectively, between the ground truth correspondence and the estimated correspondence.

Considering that the original TIR images are upscaled by 2.3 for registration, as mentioned in Section 4, the image pair is considered to be matched correctly if the RMSE is less than 2.3 pixels, which is the subpixel accuracy for the original TIR images. The subpixels of the original TIR images are approximately 0.39 m in the ground space for the Wetland set, Daman set, and Huazhaizi set and approximately 0.13 m in the ground space for the Dark-Daman set, as mentioned in Section 3.

5.2. Parameter and feature analysis

5.2.1. Parameter analysis

As shown in Table 2, TWMM contains two main parameters: the size of the atomic patch (S_a) and the number of levels (N). We tested the performance of different combinations of S_a and N using all images described in Section 4.1. If S_a is too small, the atomic patch cannot contain sufficient information to reflect its distinctiveness, which yields low discrimination of the similarity map of Level 1. If S_a is too high, the atomic patch may contain local geometric distortion, rendering the similarity map inaccurate. Furthermore, a larger S_a means fewer CPs. Table 3 shows the RMSE and CMR versus different S_a when N is set to 4. TWMM achieves the best performance when S_a is 40. Thus, we set S_a to 40. Determination of the optimal value of Sa is affected by several factors, such as the TIR&V camera, the flight altitude, the deformation between images, etc. In practice, one can first randomly select some images to test the performance of different Sa, and then choose the optimal value according to the results.

After S_a is determined, N can control the patch size of the highest level of TWMM. For each additional level, the patch size of the highest level will be doubled. As shown in Table 4, the CMR of TWMM substantially increases when N increases from 1 to 4. The CMR of TWMM with N set to 5 is slightly worse than that with N set to 4 because the locations of the maximum of the similarity maps at Level 5 cannot cover the correct correspondence in some images. Therefore, N is set to 4 in our experiment. Based on Table 3 and Table 4, the two parameters are fixed to $S_a = 40$ and $N = 4$ in our experiment.

We can determine the other parameters in TWMM based on the available information. The up-sampling factor (i.e., 2.3) for eliminating spatial resolution differences between TIR&V images is determined

Table 7
Inference time of all methods. The methods are implemented on a computer with an Intel i7 8700 CPU and an NVIDIA GeForce GTX 2080Ti graphics card.

Method	SIFT	SURF	RIFT	SCB	TFeat	HardNet	RANSAC_Flow	HOPC	CFOG	TWMM
Time(s)	1.3	0.9	5.3	9.5	0.9	1.5	1.3	126.5	10.5	16.3

according to the user manual of the WIRIS Pro Sc camera. As shown in Table 1, the resolution is 0.13 m/pixel for the TIR sensor and 0.06 m/pixel for the visible sensor when the flight altitude is 100 m. Thus, the up-sampling factor is the ratio of the two resolutions. The feature dimension of CFOG (i.e., 9) is determined according to the corresponding paper (Ye et al., 2019). In general, the increasing of dimensionality of features can improve the robustness of features to some extent, but it increases the computational cost at the same time. Thus, to balance the robustness and the efficiency, the dimension of CFOG is set to 9.

5.2.2. Feature analysis

As mentioned in Section 4.1, TWMM employs the template matching with pixel-wise features of TIR&V images to calculate the similarity maps of Level 1. To analyze the effect of different features on TWMM, we tested three different features, including CFOG, SIFT, and ResNet. The implementation of CFOG refers to Section 4.1.1. The CFOG feature has 9 dimensions for each pixel, which represents the distribution of local gradients in 9 orientations. The local gradients are calculated in the 3 × 3 neighborhood around each pixel by the Sobel kernel. The implementation of SIFT refers to Lowe, (1999) and Liu et al. (2011). The SIFT feature has 128 dimensions for each pixel, which represents the histogram of local gradient directions in the 16 × 16 neighborhood around the pixel. The implementation of ResNet can refer to He et al. (2016) and Shen et al. (2020). The ResNet feature represents the deep feature in the conv4 layer of ResNet-50 (He et al., 2016), which has 256 dimensions for each pixel. The ResNet-50 network is trained in ImageNet datasets and extracts the image features using multiple convolutional layers.

Table 5 is the RMSE and CMR of TWMM with different pixel-wise features. Fig. 14 is an example of the similarity maps obtained by template matching with different features. As shown in Table 5, CFOG feature performs better than ResNet feature. We think the reason is that the depth-based feature (i.e., ResNet) is weaker than local feature in localization accuracy (i.e., CFOG, SIFT) due to continuous convolution and down sampling in the deep network. Compared with CFOG feature, SIFT feature has more dimensions and represents a larger neighborhood. Thus, SIFT feature is more robust and performs better than CFOG feature. However, SIFT feature requires more computation during template matching because its dimensions (i.e., 128) are much more than CFOG (i.e., 9). Thus, the inference time of TWMM with SIFT features is about 13.8 s longer than CFOG feature. Considering both the performance and the inference time, we use CFOG features in our following experiments.

5.3. Comparison with other methods

5.3.1. Realization of other methods

We compare TWMM with two widely employed methods (i.e., SIFT and SURF) and seven state-of-the-art methods (RIFT, SCB, TFeat, HardNet, RANSAC_flow, HOPC, and CFOG). One should note that most of these selected methods are designed for visible images or multispectral images and achieve good performance in various scenes. They are not dedicated to TIR&V image registration.

SIFT and SURF, as feature-based methods, are realized by Python and OpenCV libraries. For SIFT and SURF, first, the features of key points are extracted by corresponding algorithms. Second, the features are matched by the Euclidean distance and the first-to-second-nearest neighbor ratio with the threshold set to 0.75 according to the performance. In addition, a similar threshold value is adopted by Wang et al. (2018), Zhang et al. (2020), and Meng et al. (2021). The random sample consensus (RANSAC) algorithm is applied to filter outliers and estimate the homography (Fischler and Bolles, 1981).

RIFT, SCB, CFOG, and HOPC are implemented according to the officially released codes[1],[2] and the corresponding papers (Cao et al., 2020; Li et al., 2020; Ye et al., 2019, 2017). The parameters of these methods are fine-tuned for the best performance. For RIFT, the number of scales of Log-Gabor filter, number of orientations of Log-Gabor filter, and size of the local image patch are set to 6, 4, and 72, respectively. SCB is an unsupervised machine learning-based method, which uses the inherent edge structure of images to optimize the parameters of the affine model directly. The window size of SCB transforms is set to 3 × 3. HOPC and CFOG are area-based methods. For HOPC, the number of CPs, orientation bins for feature description, search radius of template matching, and threshold for outlier elimination of IPRA are set to 100, 8, 60, and 5, respectively. For CFOG, the search radius of template matching and threshold for outlier elimination of IPRA are the same as those for HOPC. The template window size of CFOG is set to 100 × 100. The number of CPs of CFOG is set to 625, which is the same as that of TWMM.

The deep learning methods (i.e., TFeat, HardNet, and RANSAC_Flow) are implemented based on their officially trained models. For TFeat and HardNet, features with strong semantic information are extracted by convolutional neural networks, and other parts, including key point detection, feature matching, and homography estimation, are carried out in the same way as SIFT. For RANSAC_flow, first, it uses the deep features in the conv4 layer of ResNet-50 (He et al., 2016) to register TIR&V images coarsely; then, it performs the fine alignment by estimating the flow between two coarsely registered TIR&V images.

5.3.2. Quantitative comparison

The quantitative comparison results for four datasets among TWMM and other methods are shown in Table 6 and Table 7. TWMM outperforms other methods in TIR&V image registration. For the three feature-based methods (i.e., SIFT, SURF, and RIFT), RIFT performs best and achieves approximately 1 % CMR on all image pairs. The unsatisfactory performance of the three feature-based methods may be attributed to the following reasons. First, there are significant texture and shape gaps between the TIR&V images, degrading the robustness of features. Second, many images are covered with reeds, farmlands, or bare soil, making it difficult to obtain discriminating key points. The performance of SCB and HOPC demonstrates that the features of SCB and HOPC are not robust on TIR&V images. Although TFeat, HardNet, and RANSAC_flow achieve satisfactory performance in most visible-to-visible registration tasks, they are unsuitable for TIR&V image registration, which may be attributed to the notion that the features learned in visible images are hard to apply to TIR images.

CFOG and TWMM perform well on TIR&V image registration for the following reasons. First, with prior knowledge that the offset between the TIR&V images is dozens of pixels, these methods search for CPs in a local area instead of the entire images, which eliminates the disturbances of irrelevant areas. Second, CFOG features are robust to the RSTG between the TIR&V images to a certain extent.

TWMM achieves the best performance among all methods. The CMR of TWMM in all test images is approximately 96.0 %, which is a 11.6 % improvement compared with CFOG. The CMR of TWMM in the Dark-Daman set is 90.7 %, which is a 28.1 % improvement compared with

[1] RIFT: https://www.ivlab.org/publications.html.
[2] SCB: https://ljy-rs.github.io/web/.

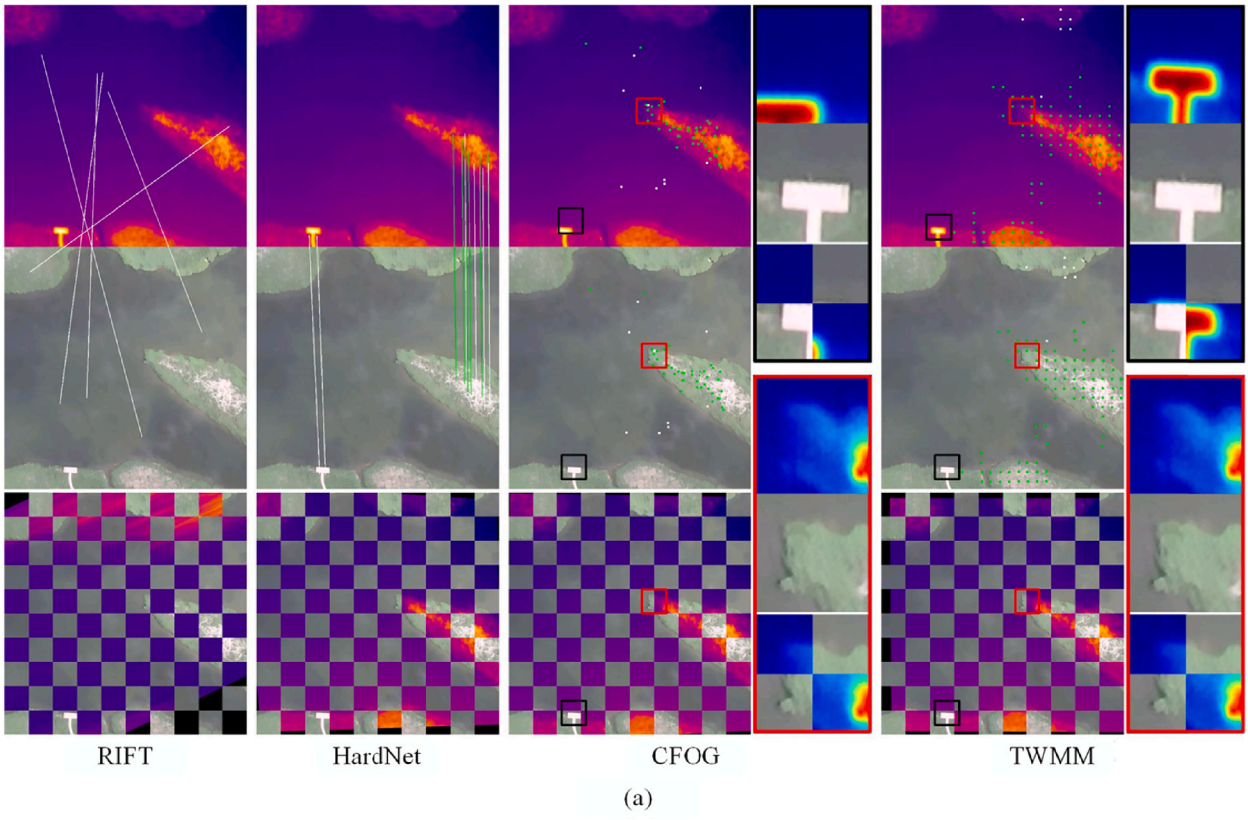
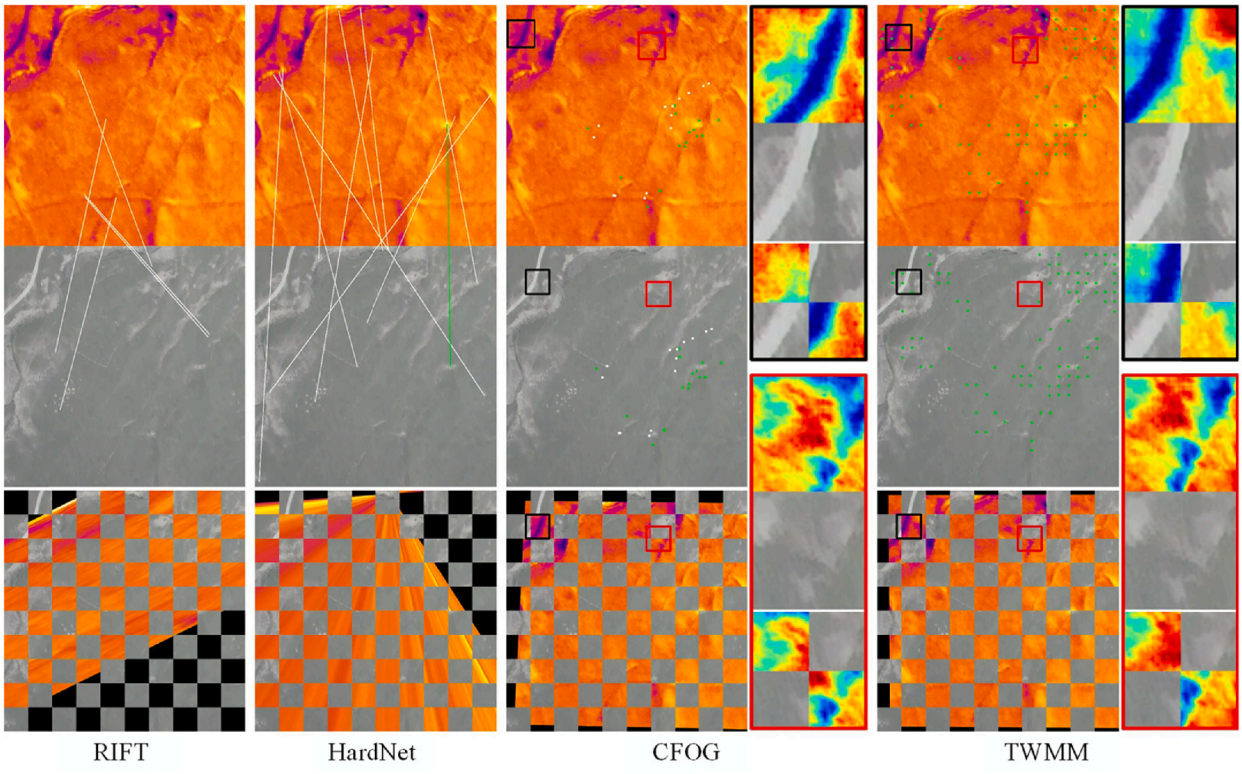


Fig. 15. Qualitative comparison of different methods. For each case, the third row is the mosaic image of the visible image and warped TIR image. For RIFT and HardNet, we use green lines to represent correct CPs and white lines to represent incorrect CPs. For CFOG and TWMM, we use green dots to represent correct CPs and white dots to represent incorrect CPs for better observation. The black rectangle and red rectangle on the right side of CFOG or TWMM are the enlarged views of the corresponding areas in the images. For each rectangle, from top to bottom are the enlarged TIR image, enlarged visible image, and mosaic of the enlarged TIR image and visible image. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



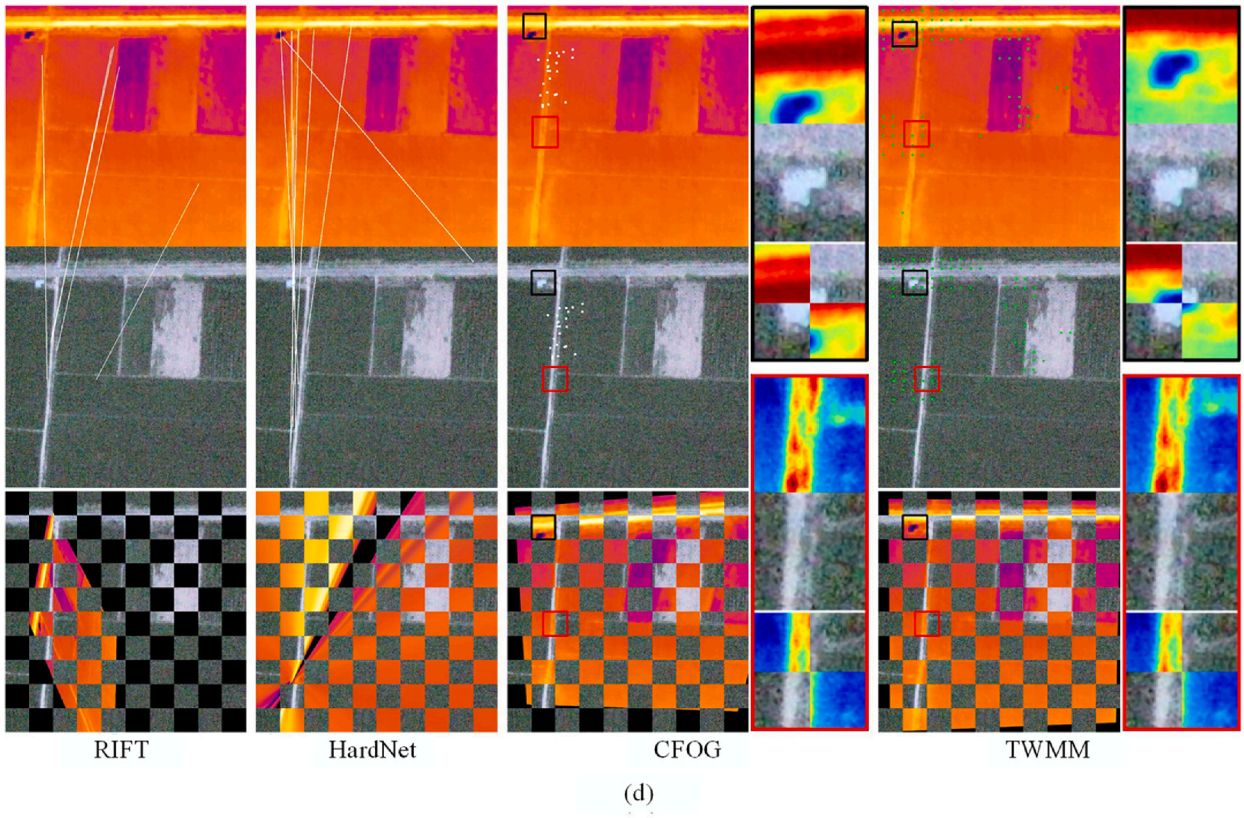


Fig. 15. (*continued*).

CFOG. The improvement in TWMM is mainly attributed to three reasons: (1) TWMM can aggregate global information by MLM, which makes TWMM robust to various conditions, such as unobvious texture and heavy noise caused by weak light; (2) MIB of TWMM uses both global information and local information when searching for CPs, which can improve the robustness and accuracy of detected CPs; and (3) CPs

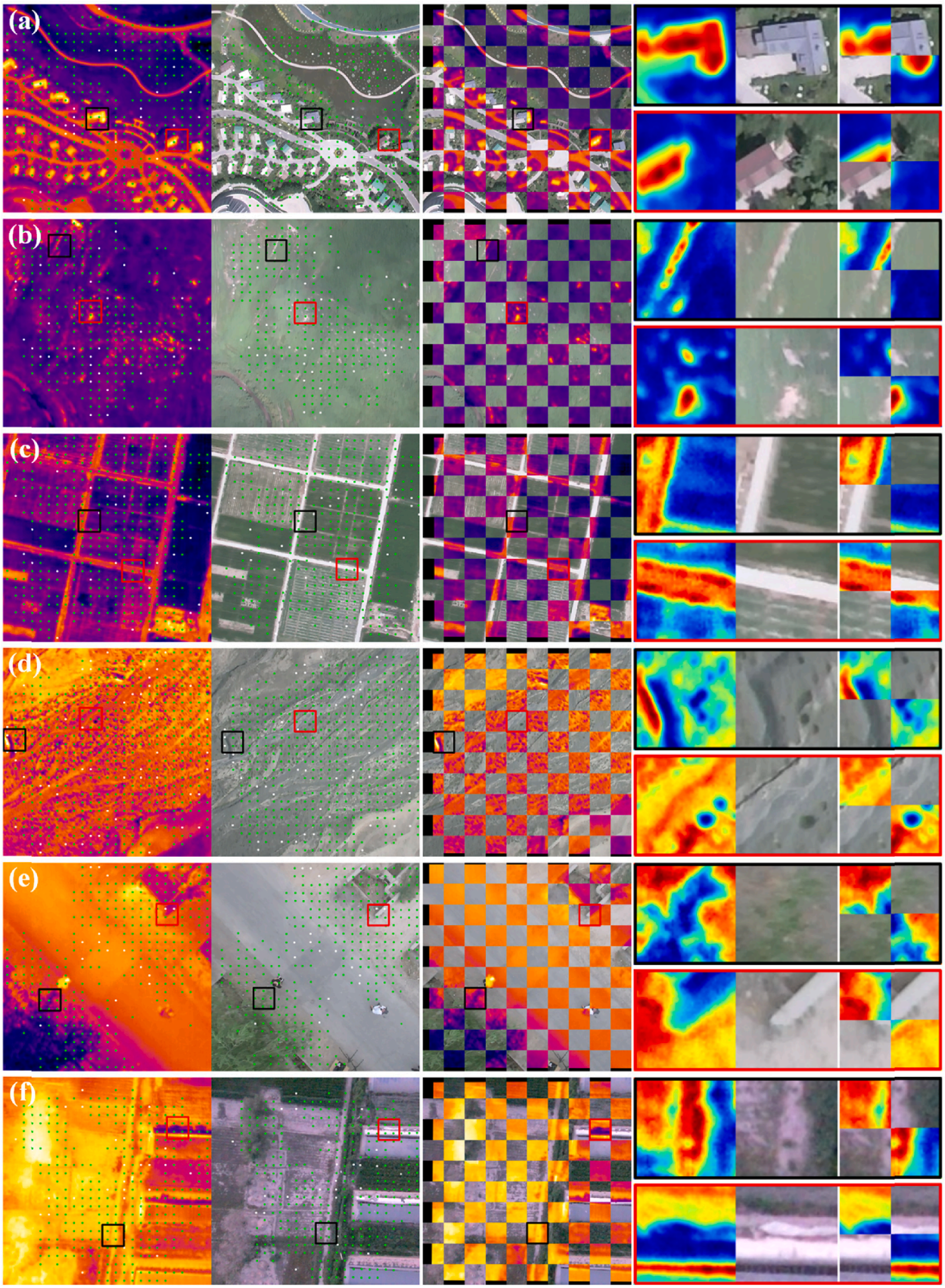


Fig. 16. Qualitative results of TWMM. For each case, the third column is the mosaic image of the visible image and warped TIR image. We use green dots to represent the correct CPs and white dots to represent incorrect CPs. The black rectangle and red rectangle in the fourth column are the enlarged views of the corresponding areas in the image. For each rectangle, from left to right are the enlarged TIR image, enlarged visible image, and mosaic of the enlarged TIR image and visible image. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 8
The effect of the weighting module, the MIM module, and the MIB module.

Method	+ weighting	+MIM and MIB	RCP	RMSE	CMR	Time (s)
CFOG/Baseline			70.9	2.09	84.4	10.5
CFOG^{+W}	✔		72.7	1.51	87.5	12.8
CFOG^{+M}		✔	81.1	1.01	91.7	15.6
CFOG^{+WM}/ TWMM	✔	✔	86.0	0.72	96.0	16.3

detected by TWMM are evenly distributed in the entire image, which can cover the complicated geometric transformation between two images. In addition, the CCP and RCP of TWMM are approximately 260 and 85 %, respectively, in all images, which are higher than those of other methods. The inference time of TWMM is slightly longer than that of CFOG, as shown in Table 7. In summary, TWMM is superior to other methods in CP detection and image registration. TWMM is a robust and effective method in TIR&V image registration.

5.3.3. Qualitative comparison

Fig. 15 shows the qualitative results of RIFT, HardNet, CFOG, and TWMM on four different images captured under different scenes: Fig. 15 (a) for Wetland site covered with ponds and large reeds; Fig. 15(b) for Daman site covered with cropland and paths; Fig. 15(c) for Huazhaizi site covered with bare soil and short shrubs; and Fig. 15(d) for Daman site at gloaming, thus, the low illumination makes the visible image dim and noisy. Therefore, the four images in Fig. 15 have different land covers, different texture conditions, and different illumination conditions. In addition, due to the different imaging mechanisms of TIR&V sensors and the thermal diffusion among ground objects, the RSTG is an intrinsic difference between TIR&V images and exists more or less in the images in Fig. 15. For example, the green duckweed in the pond can be seen in the visible image in Fig. 15(a), but it is difficult to be seen in the TIR image. The texture of the cropland in the visible image in Fig. 15(b) is hard to identify in the TIR image. The road is brighter in the visible image in Fig. 15(c), but its temperature is lower in the TIR image. There is heavy noise in the visible image in Fig. 15(d) due to low illumination, but the TIR image is not affected by the illumination condition. There

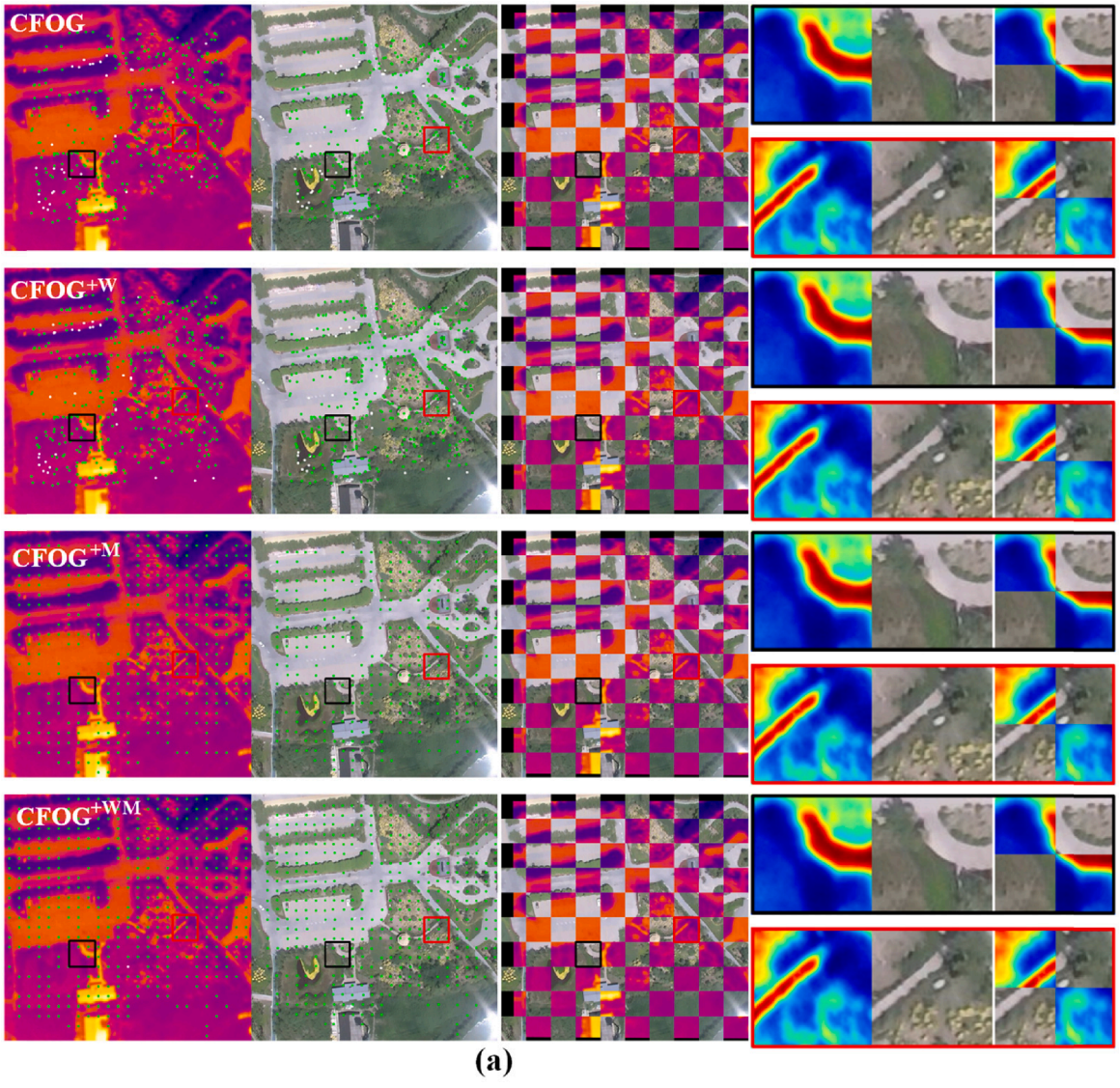


Fig. 17. Qualitative result of four methods (i.e., CFOG, CFOG^{+W}, CFOG^{+M}, and CFOG^{+WM}). For each case, the third column is the mosaic image of the visible image and warped TIR image. We use green dots to represent the correct CPs and white dots to represent incorrect CPs. The black rectangle and red rectangle in the fourth column are the enlarged views of the corresponding areas in the image. For each rectangle, from left to right are the enlarged TIR image, enlarged visible image, and mosaic of the enlarged TIR image and visible image. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

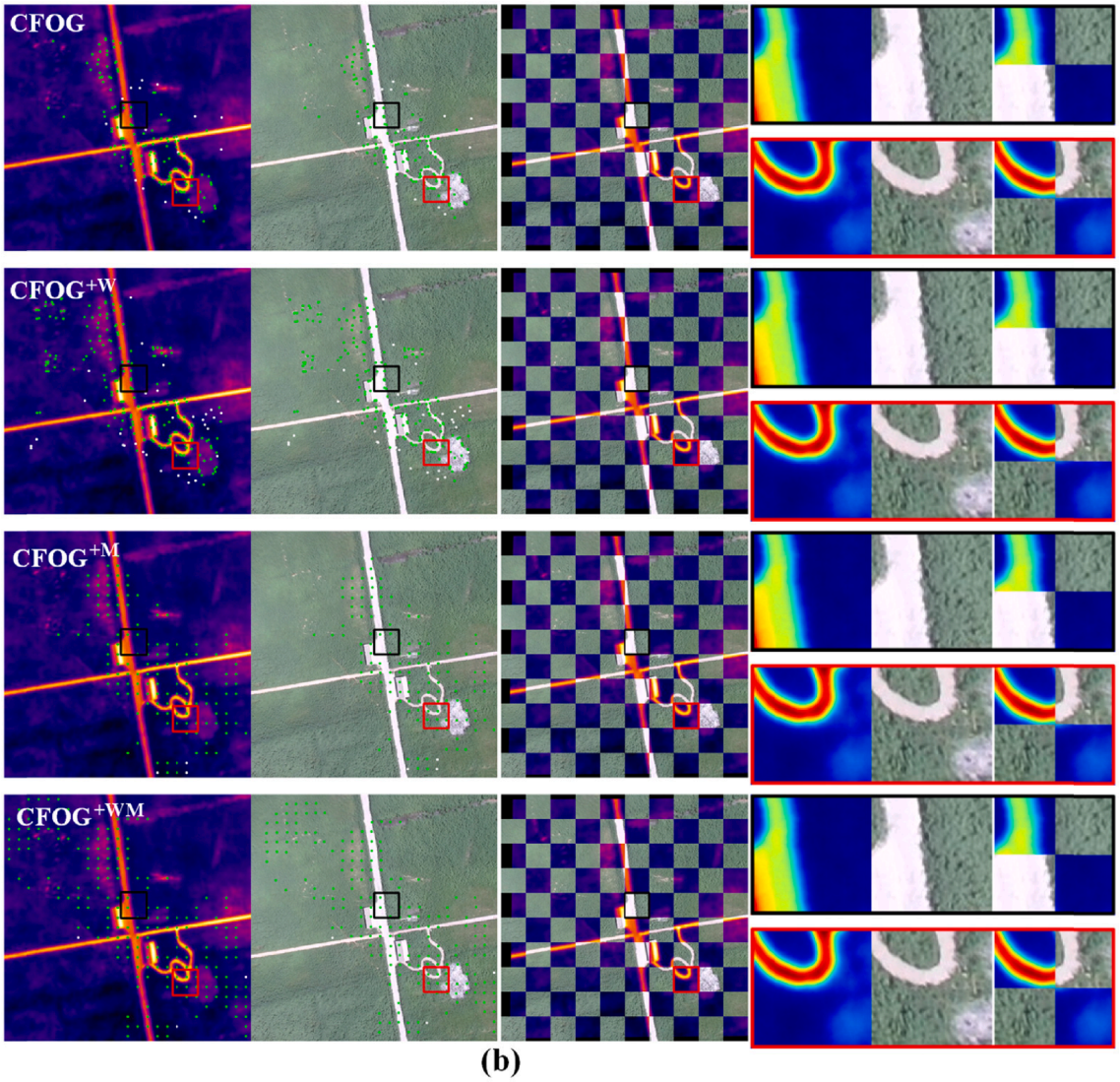


Fig. 17. (*continued*).

are a few rotation biases, a few scale biases, a few perspective biases, and dozens of pixels of translation biases between the TIR&V images in Fig. 15.

As seen in Fig. 15, the performance of RIFT is unsatisfactory in most cases. RIFT can get a small number of correct CPs between TIR&V images, which are not sufficient to evaluate the transformation model. RIFT is the feature-based method, the performance of which is mainly determined by the repetition rate of key point and the robustness of the feature. As mentioned in Section 2, RSTG decreases both the repetition rate of key point of RIFT and the robustness of feature. Furthermore, the entire-image search degrades the performance of key point matching. HardNet is the feature-based method that employs the deep learning network as the feature extractor. HardNet can correctly match some key points with significant features, such as the vegetation on the bare ground (Fig. 15a) or trees near the road (Fig. 15b). The performance of HardNet decreases in the areas with heavy noise (Fig. 15d) or low texture, such as the large ponds (Fig. 15a), crops (Fig. 15b), bare soil (Fig. 15c). In addition, considering that the transformation model between TIR&V images is complex, the homography calculated by a few CPs may be inaccurate, resulting in poor alignment in areas without CPs (Fig. 15a, c, d).

CFOG is the area-based method, which uses template matching to search CPs within a certain range. CFOG outperforms RIFT and HardNet in the three cases (Fig. 15a, b, c). The superior performance of CFOG demonstrates that the area-based method can alleviate the problem of the low key point repetition rate and the high mis-matching rate of the feature-based method. As shown in Fig. 15, the distribution of CP in CFOG has the following characteristics: the CPs mainly exist in areas with evident features, such as the edges of the island (Fig. 15a) or paths (Fig. 15b). Few CPs exist in areas with low texture (Fig. 15c) or heavy noise (Fig. 15d). The reason for this phenomenon is that the information contained in a patch with a size of 100×100 is sufficient for searching the CP in areas with significant features, while insufficient in areas with poor features. Because of these characteristics, CFOG performs well in areas with significant features while unsatisfactorily in areas with poor features. In addition, heavy noise degrades the robustness of CFOG, causing CFOG to fail in searching for the correct CP.

As seen from Fig. 15, TWMM has two advantages in CCP compared with CFOG: (1) in areas with evident features, TWMM can find more CCPs; and (2) in areas with poor features or in images with heavy noise, TWMM can also find some CCPs. Therefore, TWMM can find more CCPs than CFOG and these CCPs can a cover larger area, resulting in better registration performance. The reason for the advantages of TWMM is that TWMM can successfully employ the information of patches with different sizes and can improve both the robustness and accuracy of CP.

Fig. 16 shows more results of TWMM under various conditions.

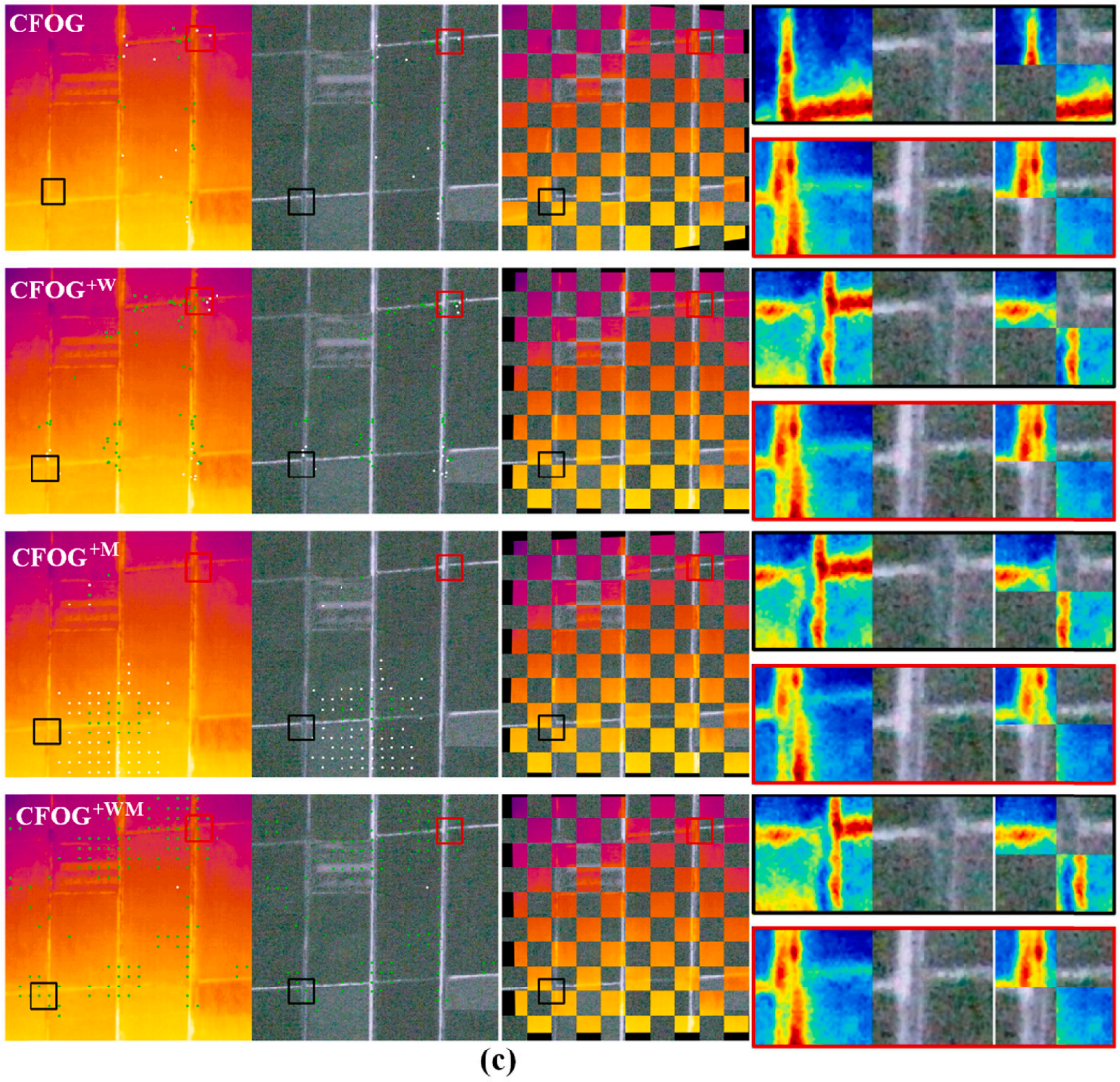


Fig. 17. (*continued*).

Combined with Fig. 15, we can get the performance of TWMM on various land cover types, textures, illumination conditions, and spatial resolutions. The land cover types include buildings (Fig. 16a), large reeds (Fig. 16b), large ponds (Fig. 15a), large crop lands (Fig. 15b), paths (Fig. 16c), roads (Fig. 16e), bare soil without short shrubs (Fig. 15c), bare soil with short shrubs (Fig. 16d), etc. The buildings and bare soil with short shrubs have good textures; the paths, roads, and reeds have middle textures; the large ponds, large crop lands, and bare soil without short shrubs have poor textures. Fig. 15(d) and Fig. 16(f) are taken at dusk and their visible images have lower illumination and heavier noise than images captured during the day. The resolutions include 0.03 m/pixel with a flight altitude of 50 m (Fig. 16e), the 0.06 m/pixel with a flight altitude of 100 m (Fig. 15d, Fig. 16f), and the 0.19 m/pixel with a flight altitude of 300 m (Fig. 15a–c, Fig. 16a–d).

As seen from the mosaic images and the enlarged views of TWMM in Fig. 15 and Fig. 16, TWMM shows high registration accuracy between the TIR&V images in these cases, demonstrating that TWMM is robust under various conditions. In addition, TWMM obtains many correct CPs in images with good or middle textures, and these CPs are approximately evenly distributed on the whole image. It is worth noting that the poor textures and heavy noise does reduce the number of correct CP of TWMM due to the robustness of the CFOG feature being degraded under these conditions.

5.4. Ablation study

In this section, we conduct several experiments to analyze the effect of the two modules in TWMM, including (1) the weighting module, and (2) the MIM module and the MIB module. The weighting module is employed to improve the accuracy of the similarity maps in Level 1. The MIM and the MIB module are employed to improve the accuracy of CPs by combining the similarity maps of patches with different sizes. Table 8 shows the results on all test images by successively applying the two modules. The baseline is the CFOG method, which deduces CPs by template matching with 100 × 100 window. The second column (+weighting) means that we employ the weighting module during the template matching. The third column (+MIM and MIB) means that we first replace the size of the template window from 100 × 100 to 40 × 40, and then use the MIM and MIB to combine the similarity maps of different patches.

As shown in Table 8, by comparing CFOG with $CFOG^{+W}$, and $CFOG^{+M}$ with $CFOG^{+WM}$, we find that the proposed weighting module improves the CMR of CFOG by 3.1 % and improves the CMR of $CFOG^{+M}$ by 4.3 %, which illustrates the effectiveness of the weighting module. Thus, it is useful to put more weights on areas with large gradients in the template window when calculating the similarity maps. Compared with CFOG, $CFOG^{+M}$ improved the CMR by 7.3 % with the MIM and MIB

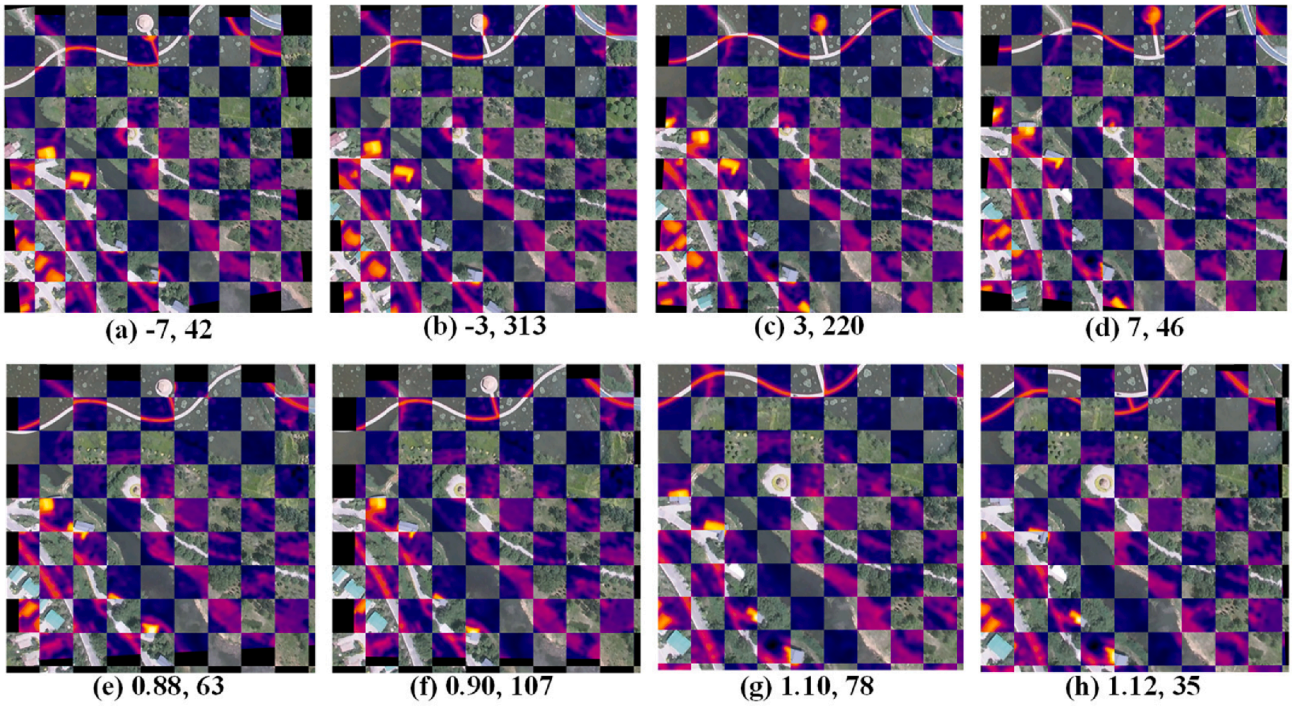


Fig. 18. Qualitative result of TWMM in TIR&V images with rotation differences (the first row) or scale differences (the second row). The two numbers under every image represent the rotation angle/resampling factor of the visible image and the number of CCP of TWMM, respectively.

module, demonstrating the effectiveness of combining the information from different patches. By using the above two modules, CFOG^{+WM} improves the CFOG by 15.1 % in RCP and 11.6 % in CMR.

Fig. 17 shows the qualitative result of all the four methods (i.e., CFOG, CFOG^{+W}, CFOG$^{+M,}$ and CFOG^{+WM}) in images with different conditions. In images with rich and clear texture (Fig. 17a), all four methods perform well with many correct CPs and high registration accuracy. In images with unobvious texture (Fig. 17b) or heavy noise (Fig. 17c), CFOG^{+W} and CFOG^{+WM} perform better than CFOG and CFOG^{+M}. The weight module can achieve better performance in areas with a few obvious objects by utilizing the salient features. For example, in the top-left region of Fig. 17(b), CFOG^{+W} and CFOG^{+WM} outperform CFOG and CFOG^{+M} because the features of the bare soil in the reeds are utilized by the weighting module. Similarly, the features of the path in Fig. 17(c) are also utilized by the weighting module to improve the performance. With MIM and MIB, the correct CP can cover a larger area in the image, demonstrating that the combination of information from multiple similarity maps can improve the robustness and accuracy of CPs. These qualitative results are consistent with the quantitative results in Table 8. In addition, considering the inference time of CFOG^{+WM} is 6.3 s longer than CFOG, we can choose the appropriate method according to the land cover and light conditions to balance the efficiency and performance. If the land cover is obvious objects and the images are taken in good light condition, the baseline (i.e., CFOG) is recommended for the TIR&V image registration; otherwise, CFOG^{+WM} (i.e., TWMM) is recommended for better performance.

6. Discussion

Although TWMM achieves good performance in UAV TIR&V image registration, there are still some limitations in TWMM. First, TWMM uses template matching to obtain correlations between the TIR&V images. Template matching calculates the similarity between the pixel in the template and the pixel in the corresponding location in the search area, as mentioned in Section 4.1. When there are large rotation or scale differences between images, it is difficult to align every pixel with its CP in the template matching, decreasing the reliability of the similarity maps. Fig. 18 shows the qualitative result of TWMM in TIR&V images with rotation differences or scale differences. According to our experience, TWMM is not suitable for the registration of these TIR&V images if the angle difference exceeds 6 degrees (Fig. 18a, Fig. 18d) or the scale difference exceeds 10 % (Fig. 18e, Fig. 18h). In many UAV flight missions, TIR&V images are usually captured by dual-cameras. The relative location of the two sensors in dual-cameras is fixed, which can avoid rotations between the two corresponding images. Besides, considering that the resolution of the TIR&V images can be calculated respectively with prior knowledge, the scale differences between TIR&V images can be largely eliminated by image sampling. In summary, TWMM can be used for most UAV TIR&V image registration tasks. Besides, our future study will focus on improving TWMM to be more robust to geometric variations between images. For example, we can utilize the gradient information (Lowe, 1999) or phase information (Ye et al., 2018) of the TIR&V images to calculate their dominant orientations, which may avoid the influence of the rotation differences to some degree. In addition, the image pyramids (Lowe, 1999) may be helpful to address the scale difference issue.

Second, TWMM is time-consuming for each atomic patch to perform template matching in a certain area, which is a common problem in many studies (Schweitzer et al., 2002; Ye et al., 2017). The inference time of TWMM is slightly long, as mentioned in Table 7. Therefore, TWMM is only suitable for image post-processing applications, such as land cover classification (Sun and Schulz, 2015), precision agriculture (Maimaitijiang et al., 2020), power line inspection (Wang et al., 2010), and quantitative remote sensing (Maes et al., 2016), etc. TWMM is not recommended for image real-time-processing applications, such as fire fighting (Shamsoshoara et al., 2020) and rescue of missing people (Dong et al., 2021). TWMM could be accelerated by graphic computation (Ye et al., 2018), which our upcoming study focuses on.

7. Conclusions

Automatic registration of UAV TIR&V images is fundamental for

subsequent applications. However, few methods have addressed this issue due to the significant RSTG between the TIR&V images. Therefore, we proposed the TWMM method to register UAV TIR&V images taken by the camera equipped with both TIR and visible sensors. TWMM uses template matching with weights to CFOG features of TIR&V images to obtain the similarity maps of the atomic patches. The MLM builds pyramid similarity maps with different patch sizes. The similarity maps of Level $n + 1$ are generated by aggregating its 4 subpatches' similarity maps at Level n ($n \geq 1$) with local max-pooling. MLM can obtain global information while avoiding the effect of local geometric distortion. To deduce CPs, MIB starts from the highest level of the pyramid similarity maps, continues level by level, and ends at Level 1. MIB uses similarity maps of all levels, which can solve the problem of a lack of discrimination of small patches and low locating accuracy of large patches. Thus, MIB can obtain a sufficient number of accurate CPs between two images. The remaining CPs filtered by IPRA are used to estimate the homography between the TIR&V images.

TWMM was comprehensively evaluated with 600 UAV image pairs under four different scenes and also compared with current methods (i.e. SIFT, SURF, RIFT, RCB, TFeat, HardNet, RANSAC_Flow, HOPC, and CFOG). One should keep in mind that since current methods dedicated to UAV TIR&V image registration are very rare, most of the selected methods are designed for the registration of visible images or multispectral images with good ability. Results indicate that TWMM outperforms these methods. It achieves 86.0 % in RCP and 96.0 % in CMR in all test images with 15.1 % and 11.6 % improvement over CFOG, respectively. TWMM is more robust than existing methods in weak-light images, achieving 20.7 % improvement in RCP and 28.1 % improvement in CMR compared to CFOG. Thus, TWMM is an effective and robust method for TIR&V image registration, which can perform well under different scenarios.

For clearer to the readers, the main contributions of this study are summarized as follows. First, we design a weighting module to improve the performance of the template matching. During template matching, the designed weighting module puts more weights on the important areas in the template window. Second, we propose a general image registration framework named TWMM. With MLM and MLB, TWMM can combine the information from patches of different sizes to improve both the robustness and accuracy of CPs. Third, we evaluate TWMM and other methods extensively on a variety of TIR&V images. Among all methods, TWMM achieves the best registration performance. Therefore, TWMM has great potential in massive UAV TIR &V image registration missions. Considering that TWMM is not invariant to scale and rotation changes, our future study will focus on improving TWMM to be more robust to geometric variations between images.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the Strategic Priority Research Program of the Chinese Academy of Sciences (grant number: XDA20100101), the Key Research and Development Projects of Sichuan Science and Technology Department (grant number: 2022YFG0209), the Sichuan Science and Technology Program (2022YFS0593), and the Fundamental Research Funds for the Central Universities of China, University of Electronic Science and Technology of China (grant number: ZYGX2019J069).

References

Ambrosia, V.G., Wegener, S.S., Sullivan, D.V., Buechel, S.W., Dunagan, S.E., Brass, J.A., Stoneburner, J., Schoenung, S.M., 2003. Demonstrating UAV-Acquired Real-Time Thermal Data over Fires. Photogramm. Eng. Remote Sens. 69, 391–402. https://doi.org/10.14358/PERS.69.4.391.

Baker, S., Datta, A., Kanade, T., 2006. Parameterizing Homographies. Tech. Rep. CMU-RI-TR-06-11 23.

Balntas, V., Riba, E., Ponsa, D., Mikolajczyk, K., 2016. Learning local feature descriptors with triplets and shallow convolutional neural networks. In: Procedings of the British Machine Vision Conference 2016. Presented at the British Machine Vision Conference 2016, British Machine Vision Association, York, UK, p. 119.1-119.11. https://doi.org/10.5244/C.30.119.

Baltsavias, E.P., 1991. Multiphoto geometrically constrained matching. ETH Zurich. https://doi.org/10.3929/ETHZ-A-000617558.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. SURF: Speeded Up Robust Features. In: Leonardis, A., Bischof, H., Pinz, A. (Eds.), Computer Vision – ECCV 2006. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 404–417. https://doi.org/10.1007/11744023_32.

Cao, S.-Y., Shen, H.-L., Chen, S.-J., Li, C., 2020. Boosting Structure Consistency for Multispectral and Multimodal Image Registration. IEEE Trans. Image Process. 29, 5147–5162. https://doi.org/10.1109/TIP.2020.2980972.

Chen, J., Cheng, B., Zhang, X., Long, T., Chen, B., Wang, G., Zhang, D., 2022. A TIR-Visible Automatic Registration and Geometric Correction Method for SDGSAT-1 Thermal Infrared Image Based on Modified RIFT. Remote Sens. 14, 1393. https://doi.org/10.3390/rs14061393.

Chen, H., Nan, X., Yipeng, Z., Qikai, L., GuiSong, X., 2019. Robust visible-infrared image matching by exploiting dominant edge orientations. Pattern Recognit. Lett. 8.

Chrétien, L.-P., Théau, J., Ménard, P., 2015. WILDLIFE MULTISPECIES REMOTE SENSING USING VISIBLE AND THERMAL INFRARED IMAGERY ACQUIRED FROM AN UNMANNED AERIAL VEHICLE (UAV). ISPRS – Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XL-1/W4, 241–248. https://doi.org/10.5194/isprsarchives-XL-1-W4-241-2015.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2016. Deep Image Homography Estimation, in: RSS Workshop on Limits and Potentials 216 of Deep Learning in Robotics. Presented at the RSS Workshop on Limits and Potentials of Deep Learning in Robotics.

Dong, J., Ota, K., Dong, M., 2021. UAV-Based Real-Time Survivor Detection System in Post-Disaster Search and Rescue Operations. IEEE J. Miniaturization Air Space Syst. 2, 209–219. https://doi.org/10.1109/JMASS.2021.3083659.

Escobar-Wolf, R., Oommen, T., Brooks, C.N., Dobson, R.J., Ahlborn, T.M., 2018. Unmanned Aerial Vehicle (UAV)-Based Assessment of Concrete Bridge Deck Delamination Using Thermal and Visible Camera Sensors: A Preliminary Analysis. Res. Nondestruct. Eval. 29, 183–198. https://doi.org/10.1080/09349847.2017.1304597.

Fischler, M.A., Bolles, R.C., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Comm ACM 24, 42.

Gao, W., Zhang, X., Lei, Y., Liu, H., 2010. An improved Sobel edge detection. In: Presented at the 2010 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT 2010). IEEE, Chengdu, China, pp. 67–71. https://doi.org/10.1109/ICCSIT.2010.5563693.

Gruen, A.W., 1985. ADAPTIVE LEAST SQUARES CORRELATION: A POWERFUL IMAGE MATCHING TECHNIQUE. South Afr. J. Photogramm. Remote Sens. Cartogr. 14, 13.

Gruen, A., 2012. Development and Status of Image Matching in Photogrammetry: Development and status of image matching in photogrammetry. Photogramm. Rec. 27, 36–57. https://doi.org/10.1111/j.1477-9730.2011.00671.x.

Haskins, G., Kruger, U., Yan, P., 2020. Deep learning in medical image registration: a survey. Mach. Vis. Appl. 31, 8. https://doi.org/10.1007/s00138-020-01060-x.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

Hirschmuller, H., 2008. Stereo Processing by Semiglobal Matching and Mutual Information. IEEE Trans. Pattern Anal. Mach. Intell. 30, 328–341. https://doi.org/10.1109/TPAMI.2007.1166.

Hisham, M.B., Yaakob, S.N., Raof, R.A.A., Nazren, A.B.A., Wafi, N.M., 2015. Template Matching using Sum of Squared Difference and Normalized Cross Correlation. In: 2015 IEEE Student Conference on Research and Development (SCOReD). Presented at the 2015 IEEE Student Conference on Research and Development (SCOReD), IEEE, Kuala Lumpur, pp. 100–104. https://doi.org/10.1109/SCORED.2015.7449303.

Jean-Yves, B., 2001. Pyramidal implementation of the affine Lucas Kanade feature tracker description of the algorithm. Intel Corp. 5, 1–10.

Khanal, S., Fulton, J., Shearer, S., 2017. An overview of current and potential applications of thermal remote sensing in precision agriculture. Comput. Electron. Agric. 139, 22–32. https://doi.org/10.1016/j.compag.2017.05.001.

Kong, S.G., Heo, J., Boughorbel, F., Zheng, Y., Abidi, B.R., Koschan, A., Yi, M., Abidi, M. A., 2007. Multiscale Fusion of Visible and Thermal IR Images for Illumination-Invariant Face Recognition. Int. J. Comput. Vis. 71, 215–233. https://doi.org/10.1007/s11263-006-6655-0.

Li, X., Cheng, G., Liu, S., Xiao, Q., Ma, M., Jin, R., Che, T., Liu, Q., Wang, W., Qi, Y., Wen, J., Li, H., Zhu, G., Guo, J., Ran, Y., Wang, S., Zhu, Z., Zhou, J., Hu, X., Xu, Z., 2013. HEIHE WATERSHED ALLIED TELEMETRY EXPERIMENTAL RESEARCH (HiWATER). Bull. Am. Meteorol. Soc. 94, 16.

Li, J., Hu, Q., Ai, M., 2020a. RIFT: Multi-Modal Image Matching Based on Radiation-Variation Insensitive Feature Transform. IEEE Trans. Image Process. 29, 3296–3310. https://doi.org/10.1109/TIP.2019.2959244.
Li, X., Yang, Q., Chen, Z., Luo, X., Yan, W., 2017. Visible defects detection based on UAV-based inspection in large-scale photovoltaic systems. IET Renew. Power Gener. 11, 1234–1244. https://doi.org/10.1049/iet-rpg.2017.0001.
Li, X., Han, K., Li, S., Prisacariu, V., 2020. Dual-Resolution Correspondence Networks. Conf. Neural Inf. Process. Syst. NeurIPS 12.
Li, M., Zhou, J., Peng, Z., Liu, S., Göttsche, F.-M., Zhang, X., Song, L., 2019. Component radiative temperatures over sparsely vegetated surfaces and their potential for upscaling land surface temperature. Agric. For. Meteorol. 276–277, 107600. https://doi.org/10.1016/j.agrformet.2019.05.031.
Liao, Y., Shen, X., Zhou, J., Ma, J., Zhang, X., Tang, W., Chen, Y., Ding, L., Wang, Z., 2022. Surface urban heat island detected by all-weather satellite land surface temperature. Sci. Total Environ. 811, 151405. https://doi.org/10.1016/j.scitotenv.2021.151405.
Lin, D., Jarzabek-Rychard, M., Tong, X., Maas, H.-G., 2019. Fusion of thermal imagery with point clouds for building façade thermal attribute mapping. ISPRS J. Photogramm. Remote Sens. 151, 162–175. https://doi.org/10.1016/j.isprsjprs.2019.03.010.
Liu, S., Li, X., Xu, Z., Che, T., Xiao, Q., Ma, M., Liu, Q., Jin, R., Guo, J., Wang, L., Wang, W., Qi, Y., Li, H., Xu, T., Ran, Y., Hu, X., Shi, S., Zhu, Z., Tan, J., Zhang, Y., Ren, Z., 2018a. The Heihe Integrated Observatory Network: A Basin-Scale Land Surface Processes Observatory in China. Vadose Zone J. 17, 180072. https://doi.org/10.2136/vzj2018.04.0072.
Liu, T., Li, R., Zhong, X., Jiang, M., Jin, X., Zhou, P., Liu, S., Sun, C., Guo, W., 2018b. Estimates of rice lodging using indices derived from UAV visible and thermal infrared images. Agric. For. Meteorol. 252, 144–154. https://doi.org/10.1016/j.agrformet.2018.01.021.
Liu, C., Torralba, A., Yuen, J., 2011. SIFT Flow: Dense Correspondence across Scenes and its Applications. IEEE Trans. Pattern Anal. Mach. Intell. 33, 17.
Lowe, D.G., 1999. Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision. Presented at the Proceedings of the Seventh IEEE International Conference on Computer Vision, IEEE, Kerkyra, Greece, pp. 1150–1157 vol.2. https://doi.org/10.1109/ICCV.1999.790410.
Lucas, B.D., Kanade, T., 1981. An Iterative Image Registration Technique with an Application to Stereo Vision. Int. Jt. Conf. Artif. Intell. 674–679.
Ma, M., Che, T., Li, X., Xiao, Q., Zhao, K., Xin, X., 2015. A Prototype Network for Remote Sensing Validation in China. Remote Sens. 7, 5187–5202. https://doi.org/10.3390/rs70505187.
Ma, J., Zhou, J., Liu, S., Göttsche, F.-M., Zhang, X., Wang, S., Li, M., 2021. Continuous evaluation of the spatial representativeness of land surface temperature validation sites. Remote Sens. Environ. 265, 112669. https://doi.org/10.1016/j.rse.2021.112669.
Maes, W.H., Baert, A., Huete, A.R., Minchin, P.E.H., Snelgar, W.P., Steppe, K., 2016. A new wet reference target method for continuous infrared thermography of vegetations. Agric. For. Meteorol. 226–227, 119–131. https://doi.org/10.1016/j.agrformet.2016.05.021.
Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., Fritschi, F.B., 2020. Soybean yield prediction from UAV using multimodal data fusion and deep learning. Remote Sens. Environ. 237, 111599. https://doi.org/10.1016/j.rse.2019.111599.
Marquardt, D.W., 1963. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. J. Soc. Ind. Appl. Math. 11, 431–441. https://doi.org/10.1137/0111030.
Maurya, L., Mahapatra, P., Chawla, D., Verma, S., 2020. An Automatic Thermal and Visible Image Registration Using a Calibration Rig. In: Jain, S., Paul, S. (Eds.), Recent Trends in Image and Signal Processing in Computer Vision, Advances in Intelligent Systems and Computing. Springer Singapore, Singapore, pp. 67–76. https://doi.org/10.1007/978-981-15-2740-1_5.
Meng, L., Zhou, J., Liu, S., Ding, L., Zhang, J., Wang, S., Lei, T., 2021. Investigation and evaluation of algorithms for unmanned aerial vehicle multispectral image registration. Int. J. Appl. Earth Obs. Geoinformation 102, 102403. https://doi.org/10.1016/j.jag.2021.102403.
Messina, G., Modica, G., 2020. Applications of UAV Thermal Imagery in Precision Agriculture: State of the Art and Future Research Outlook. Remote Sens. 12, 1491. https://doi.org/10.3390/rs12091491.
Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J., 2017. Working hard to know your neighbor's margins: Local descriptor learning loss. Adv. Neural Inf. Process. Syst, NIPS.
Nguyen, T., Chen, S.W., Shivakumar, S.S., Taylor, C.J., Kumar, V., 2018. Unsupervised Deep Homography: A Fast and Robust Homography Estimation Model. IEEE Robot. Autom. Lett. 3, 2346–2353. https://doi.org/10.1109/LRA.2018.2809549.
Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C., 2016. DeepMatching: Hierarchical Deformable Dense Matching. Int. J. Comput. Vis. 120, 300–323. https://doi.org/10.1007/s11263-016-0908-3.
Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. ORB: An efficient alternative to SIFT or SURF. In: 2011 International Conference on Computer Vision. Presented at the 2011 IEEE International Conference on Computer Vision (ICCV), IEEE, Barcelona, Spain, pp. 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544.
Rudol, P., Doherty, P., 2008. Human Body Detection and Geolocalization for UAV Search and Rescue Missions Using Color and Thermal Imagery. In: 2008 IEEE Aerospace Conference. Presented at the 2008 IEEE Aerospace Conference, IEEE, Big Sky, MT, USA, pp. 1–8. https://doi.org/10.1109/AERO.2008.4526559.
Santesteban, L.G., Di Gennaro, S.F., Herrero-Langreo, A., Miranda, C., Royo, J.B., Matese, A., 2017. High-resolution UAV-based thermal imaging to estimate the instantaneous and seasonal variability of plant water status within a vineyard. Agric. Water Manag. 183, 49–59. https://doi.org/10.1016/j.agwat.2016.08.026.
Schuster, R., Wasenmuller, O., Unger, C., Stricker, D., 2019. SDC – Stacked Dilated Convolution: A Unified Descriptor Network for Dense Matching Tasks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, pp. 2551–2560. https://doi.org/10.1109/CVPR.2019.00266.
Schweitzer, H., Bell, J.W., Wu, F., 2002. Very Fast Template Matching. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (Eds.), Computer Vision — ECCV 2002, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 358–372. https://doi.org/10.1007/3-540-47979-1_24.
Shamsoshoara, A., Afghah, F., Razi, A., Zheng, L., Fulé, P.Z., Blasch, E., 2020. Aerial Imagery Pile burn detection using Deep Learning: the FLAME dataset.
Shen, X., Darmon, F., Efros, A.A., Aubry, M., 2020. RANSAC-Flow: generic two-stage image alignment, in: RANSAC-Flow: Generic Two-Stage Image Alignment. Presented at the 2020 European Conference on Computer Vision, Glasgow English.
Sinha, S., Sharma, L.K., Nathawat, M.S., 2015. Improved Land-use/Land-cover classification of semi-arid deciduous forest landscape using thermal remote sensing. Egypt. J. Remote Sens. Space Sci. 18, 217–233. https://doi.org/10.1016/j.ejrs.2015.09.005.
Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, pp. 5686–5696. https://doi.org/10.1109/CVPR.2019.00584.
Sun, L., Schulz, K., 2015. The Improvement of Land Cover Classification by Thermal Remote Sensing. Remote Sens. 7, 8368–8390. https://doi.org/10.3390/rs70708368.
Torabi, A., Massé, G., Bilodeau, G.-A., 2012. An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications. Comput. Vis. Image Underst. 116, 210–221. https://doi.org/10.1016/j.cviu.2011.10.006.
Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Trans. Image Process. 13, 600–612. https://doi.org/10.1109/TIP.2003.819861.
Wang, B., Chen, X., Wang, Q., Liu, L., Zhang, H., Li, B., 2010. Power line inspection with a flying robot. In: 2010 1st International Conference on Applied Robotics for the Power Industry (CARPI 2010). Presented at the 2010 1st International Conference on Applied Robotics for the Power Industry (CARPI 2010), IEEE, Montreal, QC, Canada, pp. 1–6. https://doi.org/10.1109/CARPI.2010.5624430.
Wang, S., Quan, D., Liang, X., Ning, M., Guo, Y., Jiao, L., 2018. A deep learning framework for remote sensing image registration. ISPRS J. Photogramm. Remote Sens. 145, 148–164. https://doi.org/10.1016/j.isprsjprs.2017.12.012.
Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C., 2013. DeepFlow: Large Displacement Optical Flow with Deep Matching. IEEE Int. Conf. Comput. Vis. 8.
Wu, Y., Ma, W., Gong, M., Su, L., Jiao, L., 2015. A Novel Point-Matching Algorithm Based on Fast Sample Consensus for Image Registration. IEEE Geosci. Remote Sens. Lett. 12, 43–47. https://doi.org/10.1109/LGRS.2014.2325970.
Xiang, T., Xia, G., Zhang, L., 2019. Mini-Unmanned Aerial Vehicle-Based Remote Sensing: Techniques, applications, and prospects. IEEE Geosci. REMOTE Sens. Mag. 36.
Xu, Z., Liu, S., Li, X., Shi, S., Wang, J., Zhu, Z., Xu, T., Wang, W., Ma, M., 2013. Intercomparison of surface energy flux measurement systems used during the HiWATER-MUSOEXE: INTERCOMPARISON OF FLUX INSTRUMENTS. J. Geophys. Res. Atmospheres 118, 13140–13157. https://doi.org/10.1002/2013JD020260.
Ye, Y., Shan, J., Bruzzone, L., Shen, L., 2017. Robust Registration of Multimodal Remote Sensing Images Based on Structural Similarity. IEEE Trans. Geosci. Remote Sens. 55, 2941–2958. https://doi.org/10.1109/TGRS.2017.2656380.
Ye, Y., Shan, J., Hao, S., Bruzzone, L., Qin, Y., 2018. A local phase based invariant feature for remote sensing image matching. ISPRS J. Photogramm. Remote Sens. 142, 205–221. https://doi.org/10.1016/j.isprsjprs.2018.06.010.
Ye, Y., Bruzzone, L., Shan, J., Bovolo, F., Zhu, Q., 2019. Fast and Robust Matching for Multimodal Remote Sensing Image Registration. IEEE Trans. Geosci. Remote Sens. 57, 9059–9070. https://doi.org/10.1109/TGRS.2019.2924684.
Yi, K.M., Trulls, E., Lepetit, V., Fua, P., 2016. LIFT: Learned Invariant Feature Transform. Eur. Conf. Comput. Vis. ECCV.
Yu, J.J., Harley, A.W., Derpanis, K.G., 2016. Back to Basics: Unsupervised Learning of Optical Flow via Brightness Constancy and Motion Smoothness. In: Hua, G., Jégou, H. (Eds.), Computer Vision – ECCV 2016 Workshops, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 3–10. https://doi.org/10.1007/978-3-319-49409-8_1.
Zhang, J., Wang, C., Liu, S., Jia, L., Ye, N., Wang, J., Zhou, J., Sun, J., 2020. Content-Aware Unsupervised Deep Homography Estimation, in: Content-Aware Unsupervised Deep Homography Estimation. Presented at the 2020 European Conference on Computer Vision, Glasgow English.
Zhang, L., Niu, Y., Zhang, H., Han, W., Li, G., Tang, J., Peng, X., 2019. Maize Canopy Temperature Extracted From UAV Thermal and RGB Imagery and Its Application in Water Stress Monitoring. Front. Plant Sci. 10, 1270. https://doi.org/10.3389/fpls.2019.01270.
Zhang, X., Zhou, J., Liang, S., Wang, D., 2021. A practical reanalysis data and thermal infrared remote sensing data merging (RTM) method for reconstruction of a 1-km all-weather land surface temperature. Remote Sens. Environ. 260, 112437 https://doi.org/10.1016/j.rse.2021.112437.
Zhao, F., Huang, Q., Gao, W., 2006. Image Matching by Normalized Cross-Correlation. In: 2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings. Presented at the 2006 IEEE International Conference on Acoustics

Speed and Signal Processing, IEEE, Toulouse, France, p. II-729-II–732. https://doi.org/10.1109/ICASSP.2006.1660446.

Zhao, Y., Sun, B., Liu, S., Zhang, C., He, X., Xu, D., Tang, W., 2021. Identification of mining induced ground fissures using UAV and infrared thermal imager: Temperature variation and fissure evolution. ISPRS J. Photogramm. Remote Sens. 180, 45–64. https://doi.org/10.1016/j.isprsjprs.2021.08.005.

Zhong, Y., Xu, Y., Wang, X., Jia, T., Xia, G., Ma, A., Zhang, L., 2019. Pipeline leakage detection for district heating systems using multisource data in mid- and high-latitude regions. ISPRS J. Photogramm. Remote Sens. 151, 207–222. https://doi.org/10.1016/j.isprsjprs.2019.02.021.