Measurement 258 (2026) 119340

Contents lists available at ScienceDirect

Measurement

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

Cross-scale infrared and visible image registration based on phase consistency feature for UAV scenario☆

Hao Li[1], Chenhua Liu[1], Maoyong Li, Lei Deng *, Mingli Dong *, Lianqing Zhu

Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instrument, Beijing Information Science & Technology University, Beijing 100016, China
Beijing Laboratory of Optical Fiber Sensing and System, Beijing Information Science & Technology University, Beijing 100192, China
Guangzhou Nansha Intelligent Photonic Sensing Research Institute, GuangZhou, Guangdong Province 511462, China

ARTICLE INFO

Keywords:
Infrared image
Visible image
UAV
Registration
PC-GTV
MCOG
Deformation compensation

ABSTRACT

In Unmanned Aerial Vehicle (UAV), aerial photography, hardware and imaging principle differences between multi-sensor systems cause significant variations in image resolution, field of view (FOV), and modality, leading to issues such as uneven feature point distribution, low matching rates, and local deformation in image registration. To address these, we propose a multi-stage infrared (IR) and visible (VIS) image registration pipeline. First, in the preprocessing stage, to mitigate scale inconsistencies between IR and VIS images, we introduce the Phase-Consistent Gaussian Total Variation (PC-GTV) method, which maps the original image into a Phase Consistency (PC) space. This intermediate domain aligns local phase information to unify edge structures and geometric contours. A weighting strategy combining Root Mean Square Error (RMSE) and Structural Similarity Index Measure (SSIM) is then used for scale estimation to determine the optimal scaling between images. Second, in the coarse registration stage, to overcome the uneven distribution of traditional feature points, a chunk-based detection strategy is applied in the Phase Consistent space to enhance texture response and improve feature point uniformity. Finally, in the fine registration stage, to tackle low matching rates and local distortion, a coupled constraint strategy based on the Multiscale Cyclic Oriented Gradient (MCOG) feature map is used to estimate non-rigid offsets, while an elastic compensation model completes the fine registration. Extensive experiments demonstrate that our method achieves strong visual and quantitative performance on public datasets. Our code is available at https://github.com/lh-ite/UAV_IRVIS_Registration.

1. Introduction

In Unmanned Aerial Vehicle (UAV) aerial sensing applications, precise spatial measurement and localization are critical for environmental monitoring, industrial inspection, and navigation tasks [1–3]. Integrating multimodal data from infrared (IR) and visible (VIS) cameras significantly enhances sensing capability by providing complementary information [4–7]. However, hardware limitations in IR sensors—such as lower resolution, mismatched field of view (FOV), and intrinsic modality heterogeneity introduce systematic geometric biases and feature inconsistencies, as shown in Fig. 1. These factors degrade registration accuracy, directly impacting the spatial measurement precision and reliability of subsequent analysis.

Accurate IR and VIS image registration is therefore a fundamental prerequisite for robust multi-modal measurement. In UAV remote sensing, sub-decimeter spatial registration accuracy (within 0.1 m) is generally required to ensure dependable measurement outcomes and to avoid error accumulation in downstream processing [8]. Registration errors beyond this threshold can propagate, compromising localization and quantitative analyses.

IR and VIS image registration aims to align the spatial coordinates of IR and VIS images from different viewpoints [9]. Existing methods can be categorized into deep learning methods and traditional methods. Deep learning-based [10,11] IR and VIS registration methods are mainly divided into two categories: supervised learning [12–14] and unsupervised learning [15,16]. Supervised learning relies on large

☆ **Acknowledgments**
This work is supported by the Research Project of Beijing Municipal Natural Science Foundation (No. BJXZ2021-012-00046). And they sincerely thank the reviewers and editors for their efforts in this work. Thanks to Badminton for bringing joy and giving me strength during my working days.

* Corresponding authors.
E-mail addresses: lh_010625@163.com (H. Li), dally211@163.com (L. Deng), dongml@bistu.edu.cn (M. Dong).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.measurement.2025.119340
Received 20 August 2025; Received in revised form 4 October 2025; Accepted 13 October 2025
Available online 16 October 2025

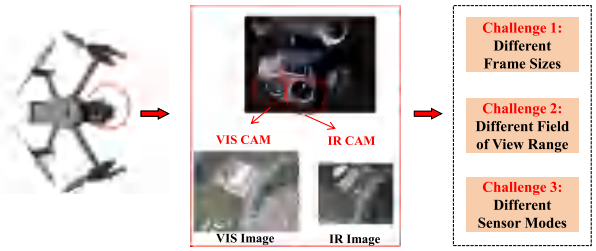


Fig. 1. UAV-based multimodal imaging system with VIS and IR cameras.

labeled datasets, such as displacement fields or segmentation labels, but data is scarce and the labeling cost is high in real scenarios, which is one of the current research bottlenecks [17]. Additionally, it is prone to noise interference, overfitting, and limited generalization. While unsupervised learning eliminates the need for labeled data, it is sensitive to data distribution, prone to local optima, and struggles to leverage cross-modal semantic information effectively.

Traditional image registration techniques can be categorized into three types: area-based, optimization-based [18] and feature point-based [19] methods. Extensive experiments indicate that the high computational complexity of the area-based method, coupled with the large difference between IR and VIS image imaging, results in poor matching accuracy and stability. Optimization-based methods are computationally intensive and sensitive to parameter tuning, and both area-based and optimization-based methods perform poorly when measuring key factors such as efficiency, stability, and practicality, especially in resource-limited UAV platforms and cross-modal scenarios. In contrast, the feature point-based method has become the current mainstream solution due to its efficient computing performance. Its core advantage lies in local feature invariance and low computational overhead, which makes it suitable for the real-time processing needs of UAV platforms. However, the resolution, FOV, and modal differences between IR and VIS images result in insufficient matching success for feature point-based registration algorithms.

To address the efficient registration challenges caused by multi-sensor heterogeneity and dynamic deformation accumulation in UAV multimodal sensing, this paper proposes a three-stage progressive registration framework. First, in the preprocessing stage, to handle resolution differences and FOV deviations between IR and VIS images, we introduce an adaptive scale search algorithm based on the joint Root Mean Square Error (RMSE) and Structural Similarity Index Measure (SSIM) metric, enabling optimal affine transformation parameter estimation. Second, in the coarse registration stage, to mitigate feature mismatches caused by differences between thermal radiation and optical reflection, we design the Phase-Consistent Gaussian Total Variation (PC-GTV) feature enhancement model. Fig. 2 illustrate how this model effectively reduces cross-modal feature energy distribution errors in the Phase Consistent space. Finally, in the fine registration stage, to address low matching rates and local deformation accumulation, we construct a Multiscale Cyclic Oriented Gradient (MCOG) descriptor combined with an elastic deformation compensation model, achieving sub-pixel-level non-rigid correction. The main contributions of this paper can be summarized as follows:

(1) We propose a robust cross-modal registration framework for UAV-based IR and VIS images, effectively mitigating spatial resolution, FOV, and modality disparities.

(2) To improve feature stability under modality and scale inconsistencies, we propose the PC-GTV method based on PC and Gaussian total variation (GTV), and introduce a weighted RMSE–SSIM dynamic scale adaptation mechanism for optimal transformation selection.

(3) We introduce the MCOG descriptor for multi-scale, multi-orientation feature representation to bridge modality gaps, and incorporate an elastic deformation compensation mechanism to correct local distortions in complex UAV IR and VIS images.

(4) Extensive experimental validation on publicly available datasets confirms the effectiveness of our method, showing both qualitative improvements in registration and quantitative superiority over state-of-the-art baselines. Moreover, we provide a user-friendly Matlab App to reduce the complexity of the registration process.

2. Related work

In the study of image registration, feature point-based methods and area-based methods are two mainstream strategies, which solve the image registration problem from different perspectives.

2.1. Area-based method

The core principle of area-based registration methods is to use a sliding window to evaluate pixel similarity between the reference image and the image to be registered, identifying the region with the highest similarity as the matching area for registration [9]. This method primarily consists of two key components: the selection of similarity metrics and the choice of transformation models.

2.1.1. Measure metric

In area-based image registration, similarity metrics are crucial for evaluating the registration between two images. Common metrics include mutual information (MI), normalized mutual information (NMI) [20], normalized cross-correlation, and mean squared error (MSE) [21]. MI and NMI measure the statistical dependence between joint and marginal probability distributions, making them effective for multimodal registration, as they are less sensitive to intensity variations. Normalized Cross Correlation (NCC) evaluates the linear correlation between image intensities and is commonly used in single-modal image registration, offering robustness to illumination changes. MSE calculates the mean squared difference between corresponding pixel intensities, which works well in scenarios with similar intensity distributions, although it is more sensitive to noise.

2.1.2. Transformation model

Another key component of area-based registration methods is the transformation model. An appropriate transformation model ensures both accurate matching and efficient optimization. Based on the geometric relationship between images, transformation models are classified into two types: rigid and non-rigid. Rigid transformations include translation and rotation, while non-rigid transformations encompass elastic deformations and local warping.

Rigid transformation. Rigid transformations are used to reposition an image without altering its intrinsic shape or the relative distances between points, involving only translations and rotations. In 1992, Besl and McKay introduced the Iterative Closest Point (ICP) algorithm, which iteratively estimates the best rotation and translation to align two 3D point sets. Their work laid the foundation for many subsequent rigid registration techniques.

Non-rigid transformation. Non-rigid transformations extend the registration process by accommodating local deformations that cannot be captured by rigid methods. These techniques are essential when registering images that exhibit complex local variations.

Based on the above information, we can know the core content of the area-based registration method. Next, we will introduce the research status of this method in recent years. In 2013, Rivaz et al. [22] proposed non-rigid registration of ultrasound and MRI using context-conditioned mutual information. In 2022, Ofvertstedt et al. [23] proposed a fast calculation of mutual information in the frequency domain for registration.

Overall, these results highlight a key trend in image registration: researchers are focusing on combining precise, complex similarity metrics with efficient optimization algorithms to address the inherent

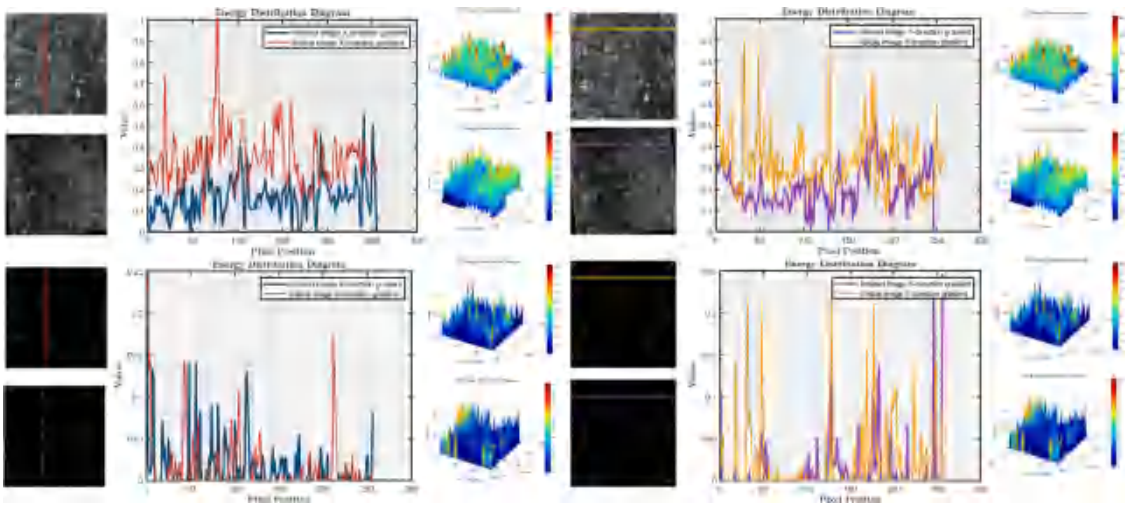


Fig. 2. Comparison of horizontal and vertical energy maps of images processed with and without PC.

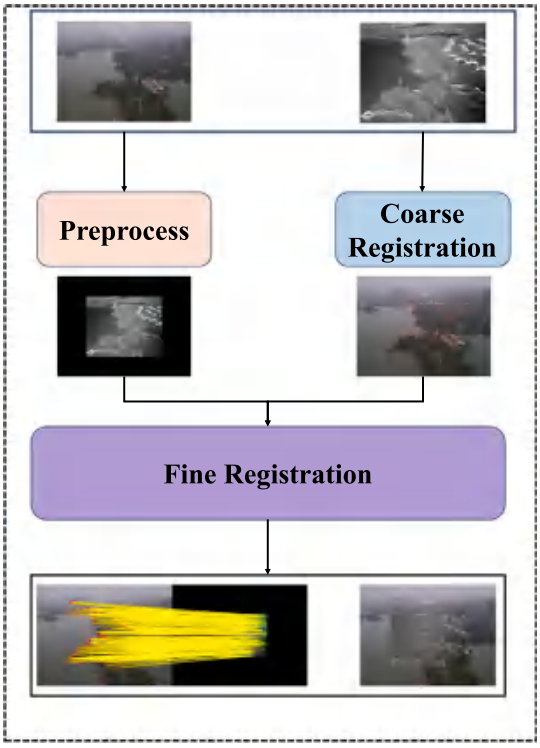


Fig. 3. Schematic diagram of the overall framework of our method. The input IR and VIS images undergo preprocessing and coarse registration to yield a resolution-transformed IR image and extracted feature points, which are then refined via fine registration to produce the final result.

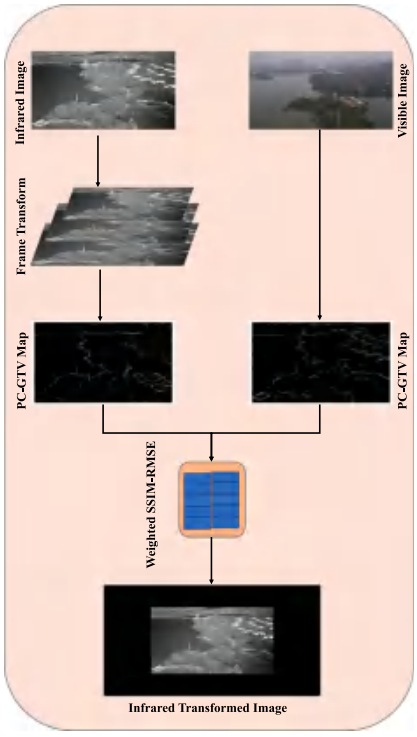


Fig. 4. Preprocessing flowchart: By applying an affine transformation to the IR image, multiple transformed scales are generated, and the optimal IR image is selected using an optimization method.

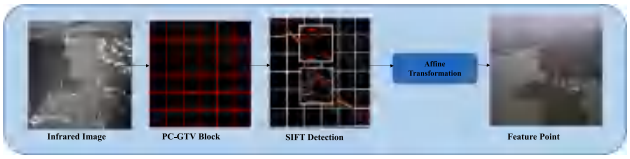


Fig. 5. Coarse registration flowchart: The IR image is processed in blocks, followed by the PC-GTV operation to enhance feature point detection efficiency, ultimately obtaining preliminary feature points.

challenges in the registration process. This approach not only enhances registration accuracy but is also expected to drive the widespread adoption of real-time registration technology across various imaging applications.

2.2. Feature point-based method

Feature point-based registration methods achieve accurate image registration by extracting and matching local salient features, then calculating the transformation matrix that describes the overall geometric relationship between the images. These methods primarily consist of feature point extraction and matching algorithms [24].

2.2.1. Feature detection

Feature points, such as corners, edges, and textured areas, are key to stable image registration under varying conditions. Effective detection algorithms identify these points reliably across different scales, rotations, and lighting changes [25].

In 1988, Harris and Stephens [26] proposed the Harris corner detection method, which identifies corners based on the image gradient's autocorrelation. In 2004, Lowe [27] proposed the Scale-Invariant Feature Transform (SIFT) algorithm that generates scale- and rotation-invariant feature descriptors. In 2006, Bay et al. [28] proposed Scale-Invariant Feature Transform (SURF) which improved SIFT's speed and robustness by using integral images and Hessian matrix approximations. Also in 2006, Rosten et al. [29] proposed Features from Accelerated Segment Test (FAST) focusing on real-time performance, detecting feature points using accelerated segmentation testing. Following these pioneering works, feature detection algorithms have continued to evolve towards higher efficiency and multimodal adaptability. In 2011, Rublee et al. [30] proposed the Oriented FAST and Rotated BRIEF (ORB) algorithm, which built upon the real-time capabilities of FAST by introducing intensity centroid calculations to achieve rotation invariance, significantly enhancing the reproducibility of feature points. In 2013, Alcantarilla et al. [31] proposed the Accelerated-KAZE (AKAZE) algorithm, which broke through the limitations of traditional linear scale spaces. Its nonlinear diffusion filtering better preserved edge and texture features. In 2018, DeTone et al. [32] proposed the SuperPoint algorithm, which for the first time achieved end-to-end joint learning of feature detection and description through a self-supervised CNN architecture, outperforming manually designed features in single-modal matching tasks. In 2025, Chen et al. [33] proposed the Joint Attention-based Multi-Modal Alignment (JAMMA) framework, enabling cross-image high-frequency interaction and global perception.

The evolution of these methods shows a balance between accuracy, robustness, and efficiency, from Harris' foundational gradient approach to the real-time applications of SuperPoint and JAMMA. These advancements are crucial for tasks like 3D reconstruction and image stitching.

2.2.2. Feature matching

Feature matching plays a critical role in image registration, with methods varying significantly between rigid and non-rigid transformations.

Rigid matching. In rigid registration, feature point matching aligns images through transformations like translation, rotation, and scaling. Traditional methods use metrics like Euclidean and Hamming distances to quantify similarity and derive transformation models. However, factors like noise and illumination changes can affect matching accuracy. To address this, Fischler et al. proposed the RANSAC algorithm, which filters out erroneous matches by evaluating geometric constraints, enhancing registration accuracy and robustness. In recent years, the rapid development of deep learning has driven significant advances in feature matching for image registration, including both cross-modal and cross-view scenarios. In 2019, Dusmanu et al. [34] introduced D2-Net, a trainable convolutional neural network that jointly performs local feature detection and description, enabling end-to-end learning of robust features. Building upon the idea of learned feature matching, Lindenberger et al. [35] proposed LightGlue in 2023, an efficient neural network architecture inspired by SuperGlue, designed to achieve high-accuracy local feature matching with reduced computational cost.

More recently, research attention has shifted towards dense correspondence estimation. Edstedt et al. [36] developed the Dense Kernelized Matching (DKM) method, targeting high-precision geometric estimation tasks. In 2024, Shen et al. [37] proposed the Generalized Image Matching (GIM) framework, which leverages large-scale internet video data for self-supervised training to improve generalization

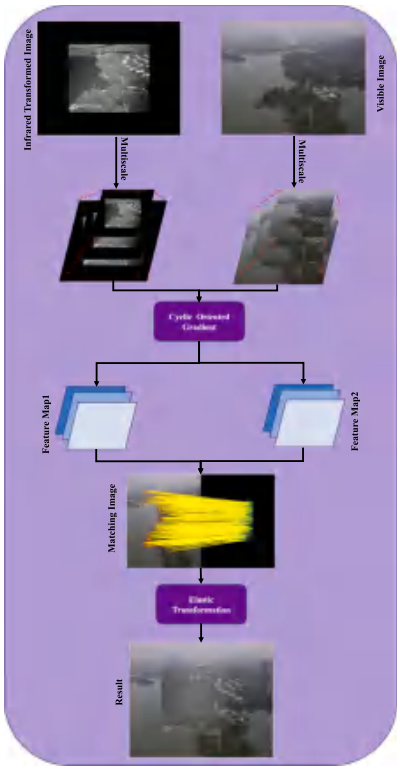


Fig. 6. Fine registration flowchart: Apply the MCOG operation to the transformed IR and VIS images to generate feature maps. Then, filter and register the feature maps to obtain matching images, followed by elastic transformation to achieve the final result.

across domains. In parallel, Edstedt et al. [38] introduced Robust dense feature matching (RoMa), which integrates DINOv2 and ConvNet features, coupled with a customized Transformer decoder and refined loss function, to achieve state-of-the-art dense matching performance. Most recently, Rendered et al. [39] presented Modality invariant image matching (MINIMA), which pushes the boundaries of modality transfer by enabling high-quality cross-modal image matching.

These methods not only represent the state of the art in the CV community but are also highly relevant to UAV-based IR and VIS images registration. Their robustness to appearance changes, illumination variations, and viewpoint differences offers valuable insights for enhancing cross-modal feature correspondence in challenging UAV scenarios.

Non-rigid matching. Non-rigid registration handles local deformations by considering both global transformations and local image variations. Bookstein et al. introduced Thin Plate Spline (TPS), which minimizes deformation energy to align images with local changes. Rueckert et al. [40] proposed Elastic Transformation, using elastic mechanics to introduce local deformation constraints for non-rigid registration.

Based on the above, we will now introduce the feature-based registration methods in recent years. Wang et al. [41] proposed a method to combine optimized SAR-SIFT features with RD model for multi-source SAR image registration. Yu et al. [42] proposed a method a novel SIFT framework based on nonlinear diffusion and polarimetric spatial frequency descriptors for universal SAR and optical image registration.

3. Proposed method

Fig. 3 illustrates the main process framework of our method, which is primarily divided into three stages: preprocessing, coarse registration, and fine registration.

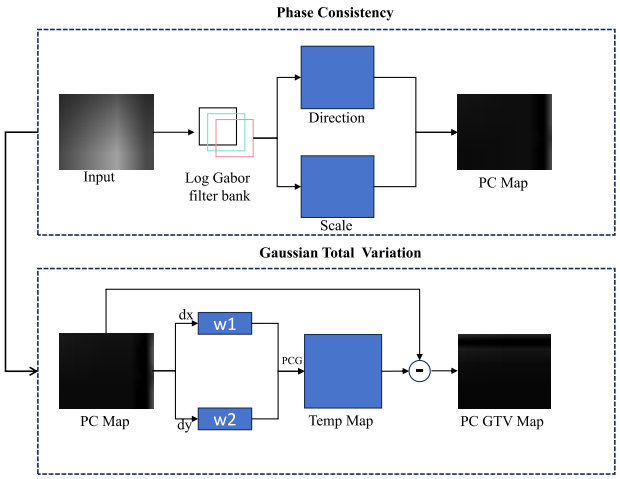


Fig. 7. Processing flow of PC-GTV method: the input image is decomposed by Log Gabor filter bank for scale and direction information to generate PC map; Subsequently, the PC-GTV results are output through gradient weighting (w_1/w_2), PCG solving, and Gaussian full-variance constraints.

3.1. Preprocess

In Fig. 4, to account for varying resolutions and FOV of UAV images, we preprocess inconsistent sizes for uniformity, improving registration efficiency. If sizes are consistent, no preprocessing is applied. This is achieved through similarity transformation.

We apply similarity transformation to align the moving image I_m with the reference image I_f, obtaining the transformed image I_t as shown in Eq. (1):

$$I' = \begin{bmatrix} s & 0 & t_x \\ 0 & s & t_y \\ 0 & 0 & 1 \end{bmatrix} I \tag{1}$$

where I represents the input image, s represents the scaling size, Δx and Δy represent the translation coordinates in the x and y directions and I' represents the transformed image.

For the translation transformation matrix, since the two sensors share the same optical axis during the dataset acquisition process, the image pairs requiring registration have a common image center. The translation parameters Δx and Δy are shown in Eqs. (2) and (3):

$$\Delta x = \frac{1}{2}(C_r - scale_x C_m) \tag{2}$$

$$\Delta y = \frac{1}{2}(R_r - scale_x R_m) \tag{3}$$

where C and R express the width and height of the corresponding image, respectively, while r and c represent row and column indices. For the scale factor $scale_x$, an affine transformation is applied to the image being registered. During this process, the scale factor $scale_x$ is incrementally adjusted from 1 to 0.01, up to a maximum value.

To improve scale transformation and subsequent feature point detection for images from different resolutions and FOV, we propose the PC-GTV. PC-GTV extracts the PC of the source image and then applies variational enhancement, facilitating the acquisition of scale transformation amplitudes. PC-GTV combines PC and GTV. In the following, we first introduce the formulation of the PC module, then describe the GTV module, and finally present the combination strategy.

To enhance feature robustness and structural consistency across modalities, we introduce a PC model that measures the coherence of phase information in local neighborhoods. Specifically, the PC at each pixel location is defined as:

$$\mathbf{PC}(x) = \frac{\int_\Omega A(x)\cos(\phi(x) - \bar{\phi}(x))dx}{\int_\Omega A(x)dx} \tag{4}$$

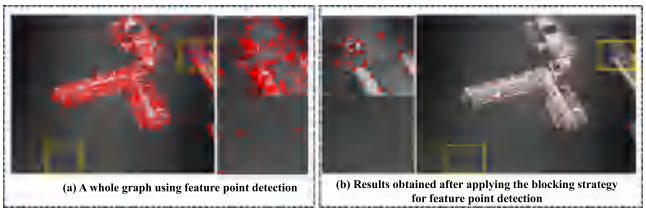


Fig. 8. Comparison between global and block-based feature point detection.

where $A(x)$ denotes the amplitude of the local Fourier component, $\phi(x)$ is the local phase, $\bar{\phi}(x)$ is the mean phase over the neighborhood Ω, and x indicates the spatial location. PC value close to 1 indicates strong phase coherence, implying stable and salient local features.

To improve robustness against noise and outliers, we further introduce a noise compensated version of PC, denoted as $PC'(x)$, as shown in Eq. (5):

$$\mathbf{PC}'(x) = \frac{\int_\Omega W(x)\lfloor A(x)(\cos(\Delta\phi(x)) - |\sin(\Delta\phi(x))|) - T\rfloor dx}{\int_\Omega A(x)dx + \varepsilon} \tag{5}$$

where $W(x)$ represents the spreading weight, and $\lfloor\cdot\rfloor$ denotes a conditional operator where negative values are set to 0, and non-negative values remain unchanged. $\Delta\phi(x)$ indicates the difference between $\phi(x)$ and $\bar{\phi}(x)$. To prevent division by zero, a small constant ε is added in the denominator. The threshold T is designed to mitigate the impact of noise. With the above Equations, we can calculate the intensity of the image in different directions, as shown in Eq. (6):

$$\begin{cases} C_1 = \int_0^{2\pi}((PC'(\theta)\cos(\theta)))^2 d\theta \\ C_2 = 2\int_0^{2\pi}((PC'(\theta)\cos(\theta)) \cdot (PC'(\theta)\sin(\theta)))d\theta \\ C_3 = \int_0^{2\pi}((PC'(\theta)\sin(\theta)))^2 d\theta \end{cases} \tag{6}$$

where C_1, C_2, and C_3 represent the contributions of the cosine component, the sine–cosine cross component, and the sine component, respectively. The maximum covariance matrix of PC M_{max} is utilized to characterize the edge strength of the image, while the minimum covariance matrix M_{min} represents the corner strength, as shown in Eq. (7):

$$\begin{aligned} M_{\max}(I) &= \frac{1}{2}\left(C_3 + C_1 + \sqrt{C_2 + (C_1 - C_3)^2}\right) \\ M_{\min}(I) &= \frac{1}{2}\left(C_3 + C_1 - \sqrt{C_2 + (C_1 - C_3)^2}\right). \end{aligned} \tag{7}$$

As shown in Fig. 7, we can observe the working mechanism of PC. Having understood the principles of PC, we now turn to the principles of GTV. GTV is based on total variation, which as shown in Eq. (8):

$$\left|\nabla_{x,y}F \cdot \exp\left(\frac{(\nabla_{xy}g)^2}{2\sigma_1^2}\right)\right| \tag{8}$$

where F denotes the filtered image, ∇_{xy} represents the differential in the x and y directions, $g = G_{\sigma_2}(F)$, $G_{\sigma_2}(F)$ is the result of Gaussian filtering with a specified standard deviation, and σ_1 is a parameter used to adjust the influence of the gradient in the exponential function.

Direct optimization of Eq. (8) is difficult due to the presence of the non-smooth L_1 norm in the regularization term. To address this, we reformulate the total variation term using an equivalent integral form:

$$\int_\Omega \left|\frac{\nabla_{x,y}F}{\exp\left(-\frac{(\nabla_{x,y}g)^2}{2\sigma_1^2}\right)}\right| dxdy = \int_\Omega \left|(\nabla_{x,y}F)\exp\left(\frac{(\nabla_{x,y}g)^2}{2\sigma_1^2}\right)\right| dxdy \tag{9}$$

To facilitate efficient computation, we introduce a surrogate formulation that enables quadratic optimization. Specifically, we multiply

both the numerator and the denominator by $\nabla_{x,y}F$, and approximate $\nabla_{x,y}F \approx \nabla_{x,y}T$, as shown in Eq. (10):

$$\begin{aligned}&\int_{\Omega}\left|\frac{\left(\nabla_{x,y}F\right)^{2}}{\left(\nabla_{x,y}F\right)\exp\left(-\frac{\left(\nabla_{x,y}g\right)^{2}}{2\sigma_{1}^{2}}\right)}\right|dxdy\\&\approx\int_{\Omega}\frac{\left(\nabla_{x,y}T\right)_{2}^{2}}{\max\left(\left(\nabla_{x,y}F\right),\epsilon\right)\exp\left(\frac{\left(\nabla_{x,y}g\right)^{2}}{2\sigma_{1}^{2}}\right)}dxdy\end{aligned}\tag{10}$$

where ϵ is a constant, to prevent the denominator from being 0, we set it to 0.0001. Then we decomposes the GTV regularization approximation into two quadratic terms, $\|\nabla_{x,y}T\|^2$ and $\frac{1}{\max\left(\left\|\nabla_{x,y}F\right\|,\epsilon\right)\exp\left(\frac{\left(\nabla_{x,y}g\right)^{2}}{2\sigma_{1}^{2}}\right)}$, where $\omega_{x,y}=\frac{1}{\max\left(\left\|\nabla_{x,y}F\right\|,\epsilon\right)\exp\left(\frac{\left(\nabla_{x,y}g\right)^{2}}{2\sigma_{1}^{2}}\right)}$ represents the nonlinear weight definition. Substituting this into the regularized objective function leads to a weighted quadratic optimization problem, as shown in Eq. (11):

$$\arg\min_{F}\|F-S\|_{F}^{2}+\lambda\left(\omega_{x}\left\|\nabla_{x}F\right\|_{F}^{2}+\omega_{y}\left\|\nabla_{y}F\right\|_{F}^{2}\right)\tag{11}$$

where $\|*\|_F$ represents the Frobenius norm. This can be reformulated using discrete matrix operators. Let D_x and D_y denote the forward-difference gradient operators in matrix form, and let the weights ω_x, ω_y be embedded as diagonal matrices. Then, the objective function becomes:

$$\arg\min_{F}\|F-S\|_{F}^{2}+\lambda\left(\left\|D_{x}W_{D_{x}}F\right\|_{F}^{2}+\left\|D_{y}W_{D_{y}}F\right\|_{F}^{2}\right)\tag{12}$$

Solving this least-squares problem leads to the closed-form solution, as shown in Eq. (13):

$$F=(1+\lambda L)^{-1}S\tag{13}$$

where 1 denotes the identity matrix with dimensions matching the original image S, and S represents the sparse five-point fixed Laplacian matrix.

Utilizing the analytical solution of Eq. (13), we can design an iterative filter to generate piecewise smoothing results F_k in Eq. (14):

$$F_{k}=(1+\lambda L_{k-1})^{-1}S\tag{14}$$

The overall GTV process is shown in Fig. 7. By combining the above, we obtain the optimized image through PC-GTV, as shown in Eq. (15):

$$GTVPC(I)=I-GTV(M_{max}(I))\tag{15}$$

where I represents the input image. To determine the optimal scale for transformation, we employ a weighted combination of RMSE and SSIM metrics, as this strategy balances both pixel-level accuracy and structural similarity. RMSE quantifies the overall pixel-wise difference, ensuring low absolute errors, while SSIM evaluates perceptual and structural consistency, which is crucial for preserving important image features. By weighting and combining these two complementary metrics, the scale conversion process achieves a more reliable and robust registration that reflects both precise intensity matching and structural integrity between the transformed image and the VIS image. The definitions of RMSE and SSIM are as shown in Eqs. (16) and (17), respectively:

$$RMSE(A,B)=\sqrt{\frac{1}{M*N}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(a_{ij}-b_{ij}\right)^{2}}\tag{16}$$

where A and B denote the two input images for computation, M and N represent the number of rows and columns of the images, respectively,

Fig. 9. PC-GTV blocks feature points are shown on IR and VIS images.

and a_{ij} and b_{ij} refer to the pixel values at position (i, j) in the two images.

$$SSIM(A,B)=\frac{(2\mu_{A}\mu_{B}+n_{1})*(\sigma_{AB}+n_{2})}{(\mu_{A}^{2}+\mu_{B}^{2}+n_{1})*(\sigma_{A}^{2}+\sigma_{B}^{2}+n_{2})}\tag{17}$$

where A and B represent the two input pictures, μ_A and μ_B represents the average pixel value of the corresponding picture, $n1, n2$ prevent the denominator from being zero, σ_{AB} represents the covariance between the A and B and σ_A, σ_B represent the standard deviation of the corresponding image.

Since the larger the SSIM, the better, and the smaller the RMSE, the better, we take the inverse and normalize it, as shown in Eq. (18):

$$RMSE_{norm}=1-\frac{RMSE_{l}}{max(RMSE_{l})}\tag{18}$$

$RMSE_l$ represents the RMSE list of all amplitudes, $max(\cdot)$ represents the maximum value operation, and $RMSE_{norm}$ represents the RMSE list after normalization. We introduce a parameter α as the weight between RMSE and SSIM. The final weighting coefficient is shown in Eq. (19):

$$SR=\alpha*SSIM_{l}+(1-\alpha)*RMSE_{norm}\tag{19}$$

where α is the weight balancing SSIM and RMSE, $SSIM_l$ represents the SSIM sequence of all scale, and SR denotes the final weighted sequence.

Finally, the most suitable $scale_x$ is obtained by optimizing the problem of finding the max SR, as shown in Eq. (20):

$$scale_{x}=Max(SR)\tag{20}$$

where $scale_x$ represents the final transformation ratio, $Max(\cdot)$ represents the maximum operation. We can obtain the final transformation matrix M, as shown in Eq. (21):

$$M=\begin{bmatrix}\text{scale} & 0 & \frac{1}{2}(C_{r}-scaleC_{m})\\0 & \text{scale} & \frac{1}{2}(R_{r}-scaleR_{m})\\0 & 0 & 1\end{bmatrix}\tag{21}$$

By M we execute an affine transformation on IR to get the result.

$$IR_{T}=Affine(IR,M)\tag{22}$$

3.2. Coarse registration

Fig. 5 presents the flowchart of our coarse registration process. Then we can use this matrix to perform an affine transformation on the image to be registered.

Traditional registration methods detect feature points in the image, often resulting in a large number of mismatched features. To solve this problem, we first extract feature points from the image to be registered. Considering that various images may have problems such as blur, low

resolution, or inconsistent FOV, we use the previously introduced PC-GTV to process the image to obtain a Phase Consistent map G, as shown in Eq. (23):

$$G = GTVPC(I) \tag{23}$$

To enhance the sufficiency of feature point detection, we propose a block-based strategy.

$$B_{k,l} = I\left[(k-1)\frac{W}{N} : k\frac{W}{N}, (l-1)\frac{H}{N} : l\frac{H}{N}\right] \tag{24}$$

where i and j represent the block at row k and column l respectively, and N represents the number of blocks per row and column. In Fig. 8, we use global feature point detection and chunked detection on the source image respectively, we can find that the result of global detection compared to our chunked result has more detection points, but a large number of invalid or redundant feature points are present, which negatively impact the matching efficiency and accuracy. By setting the image to be divided into $N \times N$ blocks, we enable more comprehensive and uniform feature detection across the entire image. Then the SIFT feature point detector is applied to the map to effectively extract feature points, especially the edges of G. We can obtain all feature points P_{IR} after using the block strategy:

$$P_{IR} = \bigcup_{k=1}^{N}\bigcup_{l=1}^{N} SIFT(B_{k,l}) \tag{25}$$

Finally, we transform P_{IR} to the VIS image coordinate system to obtain the coarse point P_T.

$$P_T = \left(M \cdot P_{IR}^T\right)^T \tag{26}$$

In Fig. 9, the feature points detected using the block-based strategy exhibit a higher correspondence with the VIS image, demonstrating the effectiveness of localized feature detection in handling modality differences.

3.3. Fine registration

Previously, we obtained the feature points extracted by coarse registration. In order to make the extracted feature points more accurate, we plan to adopt a fine registration solution, as shown in Fig. 6.

3.3.1. Feature representation

In order to filter and match the previously acquired feature points, we propose a method based on MCOG, as shown in Fig. 10. We apply gradient processing to the source image, as shown in Eq. (27):

$$\begin{cases} G_x(x,y) = \dfrac{\partial I}{\partial x} \approx \dfrac{I(x+1,y) - I(x-1,y)}{2} \\ G_y(x,y) = \dfrac{\partial I}{\partial y} \approx \dfrac{I(x,y+1) - I(x,y-1)}{2} \end{cases} \tag{27}$$

where G_x and G_y denote the gradients in the x and y directions, respectively, and $I(x,y)$ represents the grayscale value of the image at position (x,y). Using the gradient magnitudes in the x and y directions, we can compute the overall gradient and its direction, as shown in Eq. (28):

$$\begin{cases} G(x,y) &= \sqrt{G_x^2 + G_y^2} \\ \theta(x,y) &= \arctan\left(-\frac{G_y}{G_x}\right) \end{cases} \tag{28}$$

where $G(x,y)$ denotes the gradient magnitude of the source image, and $\theta(x,y)$ represents the gradient direction.

Then, we partition the obtained $\theta(x,y)$, as shown in Eq. (29):

$$\begin{cases} k_1 = \left\lfloor \dfrac{\theta}{\Delta\theta} \right\rfloor + 1 \\ k_2 = \left(k_1 \bmod N\right) + 1 \\ w_2 = \dfrac{\theta - \left(k_1 - 1\right)\Delta\theta}{\Delta\theta} \\ w_1 = 1 - w_2 \end{cases} \tag{29}$$

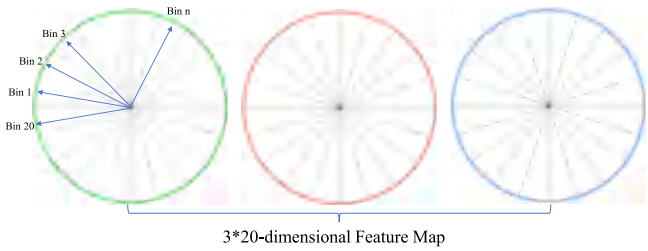


Fig. 10. The MCOG schematic diagram.

where k_1, k_2 represent the indices of adjacent bins, w_1, w_2 represent the weight distribution of the two bins, N denotes the total number of bins, and $\Delta\theta$ represents the sharing width, given by $\Delta\theta = \pi/N$. To more accurately recognize the gradient strength as well as the orientation of each pixel point, we set N to 20.

By binning the image, we obtain the histograms of different bins. To suppress noise and ensure spatial continuity, we apply Gaussian smoothing, as shown in Eq. (30):

$$H'_k(x,y) = H_k(x,y) * G_\sigma(x,y) \tag{30}$$

where $H_k(x,y)$ represents the histogram of the kth bin, $G_\sigma(x,y)$ denotes the Gaussian filter kernel with a standard deviation of σ, $*$ represents the convolution operation, and $H'_k(x,y)$ denotes the smoothed histogram.

To better extract the feature information of the image, we introduce multi-scale technology and perform multi-scale transformation operations on it based on the computed histogram, as shown in Eq. (31):

$$\begin{aligned} D(x,y) = \bigcup_{s\in S} \text{Resize}\left(H'_{s,k}(x,y)\Big|_{k=1}^{N}, \right. \\ \left. \text{mode} = \begin{cases} \text{upsample}, & s<1 \\ \text{downsample}, & s>1 \end{cases}\right) \end{aligned} \tag{31}$$

where S represents the scale set, Resize$(\cdot)$ represents the selection of upsampling or downsampling according to the scale direction, and $D(x,y)$ represents the feature tensor obtained by multi-scale concatenation. After obtaining the feature tensor through multi-scale adjustment, to ensure the stability and effectiveness of the feature descriptor, the tensor is subsequently cropped and normalized, as shown in Eq. (32):

$$\begin{cases} D_{\text{clip}}\ (x,y,c) &= \min(D(x,y,c),\ \text{clip}\ ) \\ D_{\text{norm}}\ (x,y,c) &= \frac{D_{\text{clip}}\ (x,y,c)}{\sum_{c=1}^{3N} D_{\text{clip}}\ (x,y,c)+\epsilon} \end{cases} \tag{32}$$

where clip represents the feature value cutoff threshold, θ represents the minimum value to avoid the denominator being zero, 3N represents the total feature dimension, c represents the dimension order, D_{clip} represents the feature after truncation, and D_{norm} represents the final feature descriptor after normalization.

So we perform the MCOG operation on the image to be registered and the reference image to obtain their feature maps F_1 and F_2, as shown in Eq. (33):

$$FeatureMap = MCOG(I) \tag{33}$$

Subsequently, we perform a traversal of the feature points obtained from the coarse registration. For each feature point, a block of size batchsize×batchsize centered around the point is extracted. The similarity between these blocks is then calculated using the fast Fourier transform (FFT), allowing us to identify the optimal matching position. This process ultimately yields the final location of the matched point.

3.3.2. Elastic compensation deformation

Due to the inherent differences in imaging principles and sensor characteristics between IR and VIS images, traditional rigid transformation methods often fail to achieve accurate registration. To address this issue, we introduce Elastic Compensation Deformation for non-rigid transformation, which ensures global smoothness while allowing for precise local deformation adjustments. The goal of Elastic Compensation Deformation is to find a mapping function $(u', v') = f(u, v)$ so that the transformed point $(u', v') = f(u, v)$ is closest to the target point (u', v'), which as shown in Eq. (34):

$$\begin{cases} u' = f_u(u,v) = \sum_{i=1}^{n} w_{x,i}\phi\left(r_i\right) + a_1 u + a_2 v + a_3 \\ v' = f_v(u,v) = \sum_{i=1}^{n} w_{y,i}\phi\left(r_i\right) + b_1 u + b_2 v + b_3 \end{cases} \tag{34}$$

where n is the number of matching points, (u_i, v_i) denotes the coordinates of the matching points, $\phi(r) = r^2 \log(r)$ is the radial basis function, and $r_i = \sqrt{(u-u_i)^2 + (v-v_i)^2}$ represents the Euclidean distance from the current point to the matching point. The terms $w_{x,i}$ and $w_{y,i}$ are weight coefficients used to control local deformation, while a_1, a_2, a_3 and b_1, b_2, b_3 are the parameters for the affine transformation that ensures global registration (see Fig. 11). To solve the above coefficients, we use the least squares method to solve the linear system and obtain the weight coefficients and transformation parameters, as shown in Eq. (35):

$$\begin{cases} \begin{bmatrix} \Phi & P \\ P^T & 0 \end{bmatrix} \begin{bmatrix} w \\ a \end{bmatrix} = \begin{bmatrix} g_x \\ 0 \end{bmatrix} \\ \begin{bmatrix} \Phi & P \\ P^T & 0 \end{bmatrix} \begin{bmatrix} w \\ b \end{bmatrix} = \begin{bmatrix} g_y \\ 0 \end{bmatrix} \end{cases} \tag{35}$$

where $\Phi_{ij} = \phi\left(r_{ij}\right)$, $P_i = \left[u_i, v_i, 1\right]$, $g_{x,i} = u_i^* - u_i, g_{y,i} = v_i^* - v_i$. We then convert this into a pixel-level displacement field:

$$\begin{bmatrix} \Delta x(u,v) \\ \Delta y(u,v) \end{bmatrix} = \sum_{i=1}^{n} w_i \phi\left(\left\|(u,v) - \left(u_i, v_i\right)\right\|\right) + \begin{bmatrix} a_0 + a_1 u + a_2 v \\ b_0 + b_1 u + b_2 v \end{bmatrix} \tag{36}$$

where (u, v) is the source image pixel coordinate, and $\Delta x, \Delta y$ are the displacement of each pixel. The image after local deformation can then be obtained using the following equation:

$$I_{\text{warped}}(u,v) = \sum_{i=-1}^{1}\sum_{j=-1}^{1} I_{\text{sen}}\left(u' + \Delta x + i, v' + \Delta y + j\right) \cdot B(i,j) \tag{37}$$

where $u' = u + \Delta x(u, v)$ and $v' = v + \Delta y(u, v)$. It is then weighted by the double cubic basis function $B(i, j)$ to obtain I_{warped}. In order to consider the susceptibility to mutation at the junction of overlapping and non-overlapping regions, we achieve smooth fusion in the transition region by exponential decay weighting.

$$\eta = \frac{\eta_d 1 - dist_{sub}}{\eta_d 1 - \eta_d 0} \tag{38}$$

where $\eta_d 0$ and $\eta_d 1$ are the upper and lower bounds of the smooth transition, and $dist_{sub}$ is the distance from the current point to the boundary of the overlapping region. The final registered image is obtained by applying this smoothing coefficient, as shown in Eq. (39):

$$I_{\text{final}}(x,y) = \eta \cdot I_{\text{warped}}(x,y) + (1-\eta) \cdot I_{\text{global}}(x,y) \tag{39}$$

where $I_{final}(x, y)$ denotes the final registration result and $I_{global}(x, y)$ denotes the global change result obtained by affine transformation

4. Experiments

This section details the datasets, evaluation metrics, comparison methods, parameter settings, experiment result analysis, and complexity analysis. All experiments were conducted under the same computing environment, with all methods executed on a personal computer featuring an AMD R7-6800H processor (3.2 GHz), 16 GB of RAM, and MATLAB 2023b.

Fig. 11. Partial scenario demonstration of two datasets.

4.1. Datasets, comparison methods and evaluation metrics

To demonstrate the superiority of our method across different resolutions and FOV, we conducted experiments on the VIR-UAV dataset proposed by [43], as shown in Fig. 11. This dataset was captured using the DJI M600Pro UAV equipped with the Zenmuse XT2 dual-sensor imaging platform. It contains 19 image pairs, mainly divided into three groups, each differing in altitude, FOV, rotation angle, and scenario complexity. The resolution of the VIS images is 1600 × 1200, while the IR images have a resolution of 640 × 512. The imaging system operates at a unified frame rate of 30 Hz for both modalities. Specifically, the IR sensor has a focal length of 13 mm and a pixel pitch of 17 μm, while the VIS sensor employs a fixed 8 mm focal length and a pixel pitch of 3.92 μm. These variations in sensor specifications further increase the challenge of accurate image registration, making this dataset well-suited for evaluating multimodal UAV-based registration methods.

To verify the stability of our method, we conducted experiments on the Li et al.'s dataset [44], which consists of 100 co-registered IR and VIS image pairs with a fixed resolution of 256 × 256, as shown in Fig. 11. The VIS images were captured by Landsat −8's Band 2 (Blue, 0.45–0.51 μm) in 2020, providing 30 m ground resolution data resampled to 256 × 256 pixels, while the IR images were acquired from Band 5 (NIR, 0.85–0.88 μm) in 2021, with the same spatial resolution and resampling process. The dataset covers diverse landscapes and includes rotation angles ranging from 0° to 90° to test robustness. The information for these two datasets is shown in Table 1, where we summarize information on their modality, resolution, quantity, nonlinear radiation differences (NRDs), rotation, and scale.

The evaluation was conducted on the two datasets introduced earlier. The VIR-UAV dataset contains 19 IR and VIS image pairs, which were collected from a UAV platform in various outdoor scenes, including urban areas, rivers, and forests. This dataset exhibits significant

Table 1
Summary of dataset information.

Dataset	Description	Modality	Dataset size	Image size	NRDs	Rotation change	Scale change
VIR-UAV [43]	UAV aerial image	Optical-Infrared	19 pairs	640*512-1600*1200	Yes	No	Yes
Li [44]	Satellite image	Optical-Infrared	200 pairs	256*256	Yes	Yes	NO

Algorithm 1 Cross-Scale Image Registration

Require: Infrared (I_{IR}), Visible (I_{VIS}) images
Ensure: Registered result R
for $s = 1$ to $S_{\max}$ **do**
  Generate $I_{IR}^{(s)}$ via scale-space framing
  Compute PC-GTV maps Eq(15)
  Evaluate $M(s)$ via weighted RMSE/SSIM
end for
Determine optimal scale $s^* \leftarrow \arg\max M(s)$ Eq(20)
Calibrate $I'_{IR} \leftarrow T_{affine}(I_{IR}, s^*)$
Partition I'_{IR} into blocks $\{B_i\}$
for each B_i **do**
  Extract SIFT keypoints $\{P_i\}$ via PC-GTV
end for
Aggregate $P \leftarrow \bigcup P_i$, align via T_{affine}
Estimate shift $\Delta \leftarrow FFT(MCOG(I'_{IR}, I_{VIS}))$ Eq((33))
Correct $P^{corr} \leftarrow P + \Delta$
Output $R \leftarrow ElasticDeform(P^{corr})$ Eq((39))
return R

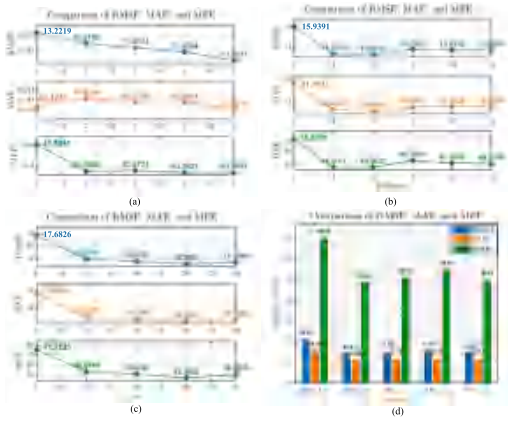


Fig. 12. The parameter influence on the metrics is shown in the plot, where the horizontal axis represents the varied parameter, and the vertical axis indicates the corresponding change in its value. Specifically: (a) shows the influence of the λ parameter; (b) depicts the effect of the *blocknum* parameter; (c) presents the impact of the *Bin* parameter; and (d) demonstrates the influence of the *Scale* parameter.

differences in spatial resolution, FOV, and scene complexity, making cross-modal registration extremely challenging. The Li et al.'s dataset includes 200 IR and VIS image pairs collected from a satellite perspective, covering scenes such as plains, hills, and vegetation. These two datasets exhibit different characteristics in terms of resolution and texture complexity, enabling a comprehensive evaluation of registration accuracy and robustness.

In order to validate the performance differences between the different algorithms, we intend to compare the CoFSM [45], MS-HLMO [46], CAO-C2F [47], 3MRS [48], LNIFT [9], RIFT [49], MS-PIIFD [50], D2-Net [34,51], and LightGlue [35,52] methods.

To evaluate the registration algorithm's performance in terms of spatial registration accuracy, local deviation control, and extreme error suppression, we use three metrics: RMSE, mean absolute error (MAE),

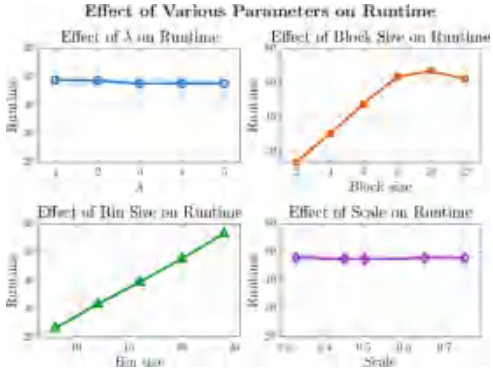


Fig. 13. Parameter impact on runtime. The four subplots show the effect of varying λ, *Block*, *Bin*, and *Scale*, in that order, on runtime.

and maximum estimated error (MEE). RMSE measures overall registration accuracy by calculating the root mean square of the Euclidean distances between matched points. MAE quantifies systematic deviation with the mean absolute error, and MEE captures extreme errors by identifying the maximum residual point, as shown in Eq. (40):

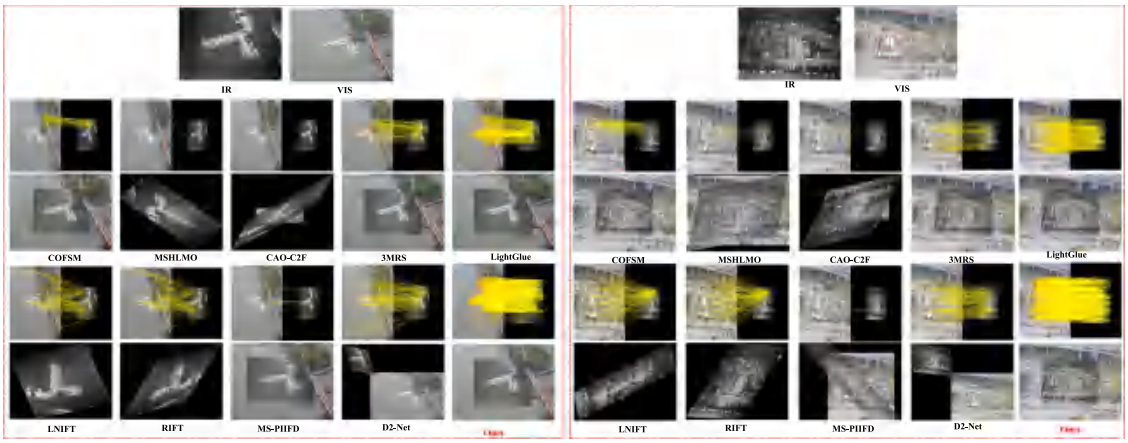
$$\begin{cases} \text{RMSE} = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n} \left\| \boldsymbol{X}_i - \boldsymbol{X}'_i \right\|_2^2}, \\ \text{MAE} = \dfrac{1}{n}\sum_{i=1}^{n} \left\| \boldsymbol{X}_i - \boldsymbol{X}'_i \right\|_2, \\ \text{MEE} = \max\limits_{1 \le i \le n} \left\| \boldsymbol{X}_i - \boldsymbol{X}'_i \right\|_2. \end{cases} \tag{40}$$

where n represents the number of corresponding points involved in the evaluation, $\left\| \boldsymbol{X}_i - \boldsymbol{X}'_i \right\|_2^2$ represents the Euclidean distance between two points.

In practice, we observed that when any of these error values exceed 20 pixels, the registration results become significantly degraded and visually unacceptable. Therefore, to clearly denote such failure cases and maintain the readability and consistency of our result tables, all values above this threshold are uniformly capped at 20.00. This threshold was empirically determined based on extensive experiments and corresponds to a practical upper bound beyond which registration results lose application value.

Building upon these pixel-level error metrics, we further introduce the number of correct matches and the matching accuracy to form a more comprehensive, multi-dimensional evaluation system. The number of correct matches reflects the density of reliable correspondences established between images, which is crucial for providing sufficient geometric constraints, especially in large-scale scenarios. The matching accuracy, defined as the ratio of correct matches to the total number of initial matches, directly quantifies the reliability of the feature matching stage. A high number of correct matches coupled with high matching accuracy ensures that the subsequent transformation estimation is based on both abundant and trustworthy geometric information.

These metrics collectively form a comprehensive registration quality assessment system. Inspired by [53], we randomly select 50 IR and VIS image pairs from the datasets and manually annotate five corresponding landmark points in each pair (250 point pairs in total). After registration, each VIS domain landmark should coincide with its IR



(a) Results of different algorithms on scenario 1 of the VIR-UAV dataset.

(b) Results of different algorithms on scenario 2 of the VIR-UAV dataset.

Fig. 14. Matching results and overlap comparisons of different algorithms across various scenarios on the VIR-UAV Dataset [43].

Table 2
Performance comparison of different methods on the VIR-UAV dataset [43].

Method	RMSE	MAE	MEE
LightGlue [35]	7.44	5.95	18.77
D2-Net [34]	20.00	20.00	20.00
MS-PIIFD [50]	20.00	20.00	20.00
RIFT [49]	20.00	20.00	20.00
LNIFT [9]	20.00	20.00	20.00
3MRS [48]	6.91	5.93	16.90
MS-HLMO [46]	20.00	20.00	20.00
CAO-C2F [47]	20.00	20.00	20.00
CoFSM [45]	10.94	10.02	17.27
Ours	**1.44**	**1.15**	**4.71**

Table 3
Performance comparison of different methods on the Li *et al.*'s dataset [44].

Method	RMSE	MAE	MEE
LightGlue [35]	14.05	13.98	14.46
D2-Net [34]	20.00	20.00	20.00
MS-PIIFD [50]	19.28	19.28	19.32
RIFT [49]	3.40	3.28	4.21
LNIFT [9]	17.35	17.34	17.46
3MRS [48]	2.24	2.12	3.39
MS-HLMO [46]	7.97	7.87	8.63
CAO-C2F [47]	20.00	20.00	20.00
CoFSM [45]	20.00	20.00	20.00
Ours	**1.77**	**1.64**	**2.84**

counterpart. We therefore compute the average Euclidean Error (AEE) between the matched points as our registration-error metric, as shown in Eq. (41). Because any residual spatial offset propagates directly into size, distance, or temperature measurements extracted from the aligned imagery, a lower AEE indicates higher spatial localization precision and thus better measurement capability for downstream metrology-oriented applications.

$$e_{AEE} = \sum_{p=1}^{M} \left\| p_{gt} - p'(p, \hat{a}) \right\|_2 \tag{41}$$

4.2. Experiment results and analysis

This subsection presents the experimental results and analysis, which are divided into five parts: parameter setting, quantitative evaluation, visual result analysis, ablation study, and runtime analysis.

4.2.1. Parameters setting

In this paper, we evaluate the PC-GTV balance coefficient λ, the number of feature point detection chunks $N \times N$, and the number of histogram groupings and multi-scale transformation parameters of MCOG. On the VIR-UAV dataset, we analyze the parameter sensitivities through four sets of independent experiments: only one parameter is adjusted at a time, and the rest are fixed to default values.

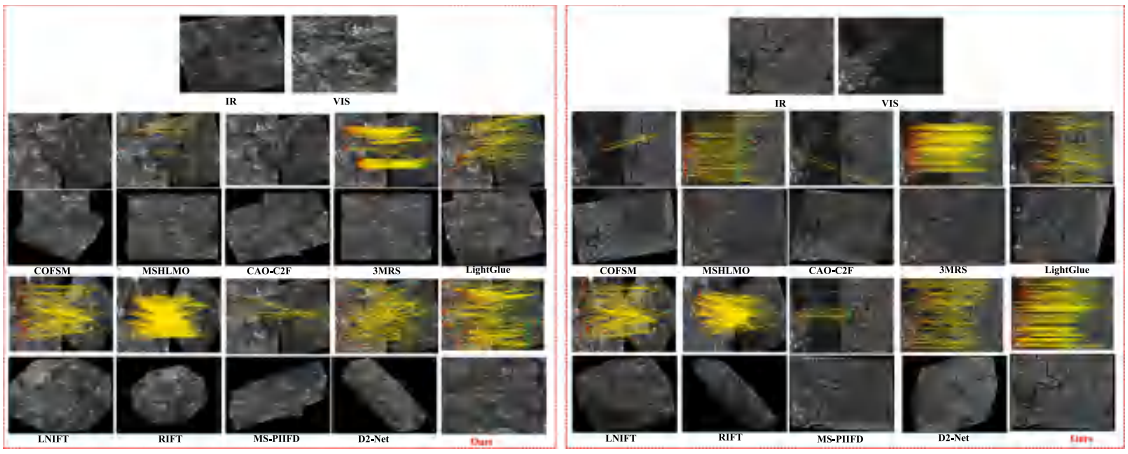
In Fig. 12, the experimental results show that when λ increases, each index shows a decreasing trend, and $\lambda = 5$ is chosen after weighing edge retention and noise suppression; the three indexes are synchronized and optimal when the number of coarse registration chunks is 6×6. In the stage of fine registration, the MCOG parameter is set to the number of histogram subgroups 20, and the multi-scale transformation coefficients [0.45,1,2]. When the MCOG parameters are set to the number of histogram groups 20 and the multi-scale transformation coefficient [0.45,1,2], the feature matching accuracy reaches the peak. The weighting coefficients for RMSE and SSIM were determined empirically through a systematic search over a representative subset of images, selecting the ratio that yielded the optimal overall performance in terms of RMSE, MAE, and MEE. Ultimately, the coefficients were set to 0.3.

To comprehensively evaluate the computational efficiency of our method, we conducted a systematic runtime analysis across all critical parameters. In Fig. 13, the value transformations of λ and Scale have basically no effect on the running time, and *Block* and *Bin* basically conform to the rule that the larger the value, the longer the running time.

Based on these analyses of accuracy, we select the following parameter settings for our experiments: $\lambda = 5$, coarse registration chunk number 6×6, and MCOG parameters with histogram groups set to 20 and multi-scale transformation coefficients of [0.45, 1, 2].

4.2.2. Visualize results and analysis

In Fig. 14, in the qualitative comparison of heterogeneous scenarios between the VIR-UAV datasets, the existing methods generally face feature matching deficiencies. MSHLMO and CAO-C2F suffer from insufficient number of matched feature points resulting in serious deformation in overlapping regions after the registration. LNIFT, RIFT and D2-Net although generating more feature points, still trigger significant registration bias with a cross-modality. MS-PIIFD produces local distortion in some scenarios, such as scenario b, due to low feature density, while COFSM, 3MRS and LightGlue, although their coverage is close to that of our method, cause edge artifacts due to the parallax of the sensor, and the offset error is large. In contrast, our method outperforms the comparison algorithms in terms of the number of features, matching accuracy and deformation suppression through the progressive registration architecture.



(a) Results of different algorithms on scenario (1) of the Li et al.'s dataset.

(b) Results of different algorithms on scenario (2) of the Li et al.'s dataset.

Fig. 15. Matching results and overlap comparisons of different algorithms across various scenarios on the Li et al.'s Dataset [44].

Table 4
The number of matching points and matching accuracy results of each method on the VIR-UAV dataset [43].

Method	Success rate (%)	Correct match number
LightGlue [35]	69.9	265
D2-Net [34]	42.5	602
MS-PIIFD [50]	85.7	18
RIFT [49]	6.8	933
LNIFT [9]	3.7	185
3MRS [48]	88.5	123
CAO-C2F [47]	14.0	7
MS-HLMO [46]	66.6	8
CoFSM [45]	75.0	45
Ours	**93.6**	**1559**

Table 5
The number of matching points and matching accuracy results of each method on the Li *et al.*'s dataset [44].

Method	Success rate (%)	Correct match number
LightGlue [35]	31.8	21
D2-Net [34]	18.4	153
MS-PIIFD [50]	82.1	23
RIFT [49]	11.7	51
LNIFT [9]	18.0	54
3MRS [48]	93.9	154
CAO-C2F [47]	62.5	5
MS-HLMO [46]	75.4	52
CoFSM [45]	33.3	2
Ours	**99.5**	**205**

Failure Scenario (a) Failure Scenario (b)

Fig. 16. Representative failure cases of the proposed method.

In Fig. 15, in the second dataset with the same resolution of IR and VIS images, comparing the samples with different rotation angles, it can be seen that: the overlapping region is distorted and deformed after the registration of CoFSM and CAO-C2F, which is mainly due to the sparse and uneven distribution of the feature points. The MS-HLMO has localized stretching deformations despite the increase in the number of matching points. The LNIFT, D2-Net and the RIFT generate dense matching points, but the large number of mismatches leads to the misalignment of the features. The MS-PIIFD fails cross-modal matching in Scenario a, and Scenario b can be partially aligned. It is worth mentioning that the LightGLue method, which performed excellently in the previous dataset, has a large deformation problem in this dataset due to the inappropriateness of this data scenario problem. While 3MRS and this paper's method achieve a natural transition in both scenarios without obvious visual defects.

In summary, our method demonstrates consistent superiority in both same-resolution and different-resolution scenarios. This advantage can be attributed to three key factors: First, the incorporation of PC-GTV-based preprocessing enhances the quality of feature point detection, ensuring reliable and robust keypoints. Second, for images with varying resolution sizes, we introduce an optimization strategy that normalizes resolution dimensions, thereby improving adaptability to inter-modal image registration challenges. Lastly, the proposed MCOG descriptor effectively captures the characteristics of feature points, significantly enhancing matching accuracy. Combined with the final elastic transformation, this leads to more precise registration results.

While our method achieves competitive performance across diverse scenarios, there remain challenging cases where it struggles, as shown in Fig. 16. Specifically, in low-texture regions the lack of discriminative features limits reliable correspondence extraction, and in repetitive patterns the descriptors may yield ambiguous matches. These conditions occasionally result in local misalignments. We explicitly highlight these limitations to provide a more comprehensive evaluation.

4.2.3. Quantitative results and analysis

The quantitative results for the two datasets are shown in Tables 2 and 3, where lower values of RMSE, MAE, and MEE indicate better performance, and the bolded numbers represent the best results among all methods.

In Table 2, existing methods exhibit performance limitations in different resolution image registration. The RMSE, MAE, and MEE metrics of CoFSM, MS-HLMO, CAO-C2F, LNIFT, RIFT, MS-PIIIFD and D2-Net fall significantly below benchmark levels. CAO-C2F underperforms as it is tailored for IR and VIS registration in power maintenance, while MS-HLMO and MS-PIIIFD struggle due to their remote sensing-optimized feature descriptors. Although LNIFT, RIFT and D2-Net are theoretically

Table 6
Conversion of acquired AEE results to real-world physical misalignment.

Method	LightGlue [35]	D2-Net [34]	MS-PIIFD [50]	RIFT [49]	LNIFT [9]	3MRS [48]	CAO-C2F [47]	MS-HLMO [46]	CoFSM [45]	Ours
Error (m)	0.13	14.87	116.25	178.16	250.73	0.14	33.99	8.98	0.24	**0.06**

Table 7
Performance comparison of different metric selection strategies in determining optimal transformation scale. Note that values larger than 20 are uniformly capped at 20 for readability, since such high errors correspond to severely distorted results.

Metric strategy	RMSE	MAE	MEE
RMSE	1.45	1.15	4.73
SSIM	1.48	1.16	5.06
RMSE&SSIM	**1.44**	**1.15**	**4.71**

Table 8
Quantitative evaluation results of various filter on the VIR-UAV dataset [43], obtained from ablation experiments.

Filter	RMSE	MAE	MEE
Gauss	8.47	6.49	21.22
RTVD	20.00	20.00	20.00
RGF	3.38	2.39	11.36
Ours	**1.44**	**1.15**	**4.71**

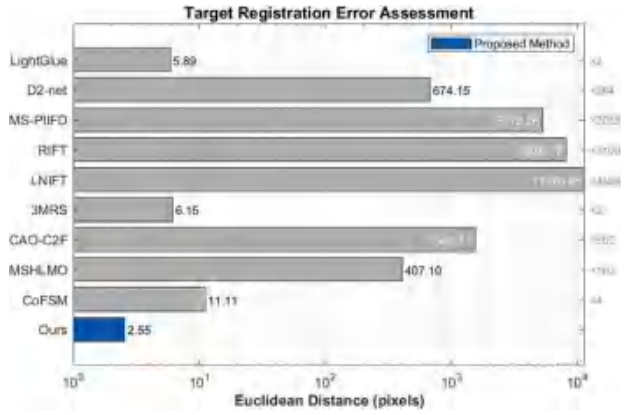


Fig. 17. Evaluation of registration error based on selected points correspondence. The rightmost annotations indicate how many times higher the error is for each method compared to ours.

Table 9
Quantitative evaluation results of various filter on the Li *et al.*'s dataset [44], obtained from ablation experiments. Note that values larger than 20 are uniformly capped at 20 for readability, since such high errors correspond to severely distorted results.

Filter	RMSE	MAE	MEE
Gauss	20.00	20.00	20.00
RTVD	13.99	13.95	14.27
RGF	4.59	4.48	5.51
Ours	**1.76**	**1.63**	**2.78**

compatible, they fail to address resolution size differences, leading to systematic spatial deviations.

Among the compared methods, 3MRS and LightGlue shows relative advantages but lacks elastic deformation modeling, causing local error accumulation. Our method, with dynamic deformation compensation, stabilizes RMSE and MAE within 2 and 3 pixels, respectively. Experimental results confirm its effectiveness in nonlinear deformation correction for cross-resolution registration.

Fig. 18. Demonstration of results of filter ablation experiments.

In Table 3, results show that same-resolution scenarios yield higher registration accuracy than cross-resolution tasks, while relative performance rankings remain stable. CoFSM, CAO-C2F, MS-PIIFD and D2-Net consistently perform the worst, while MS-HLMO and LNIFT, despite reducing errors, still suffer from localization biases in regions lacking structural features. It is worth mentioning that although the LightGlue method performed well in the previous dataset, it performed poorly this time due to the type of dataset. RIFT and 3MRS retain their advantages but lag behind our method in keypoint matching accuracy.

Experimental results demonstrate that our method outperforms the compared deep learning-based methods and shows competitive advantages over traditional-based methods. This superiority can be attributed to several factors: unlike deep models that depend heavily on large-scale annotated datasets and often face generalization challenges in unseen UAV IR and VIS scenarios, our approach does not require extensive training data. Moreover, it enhances cross-modal robustness by explicitly boosting PC features through PC-GTV preprocessing and employs the MCOG descriptor for stable keypoint matching. Additionally, the elastic deformation compensation effectively corrects local distortions, enabling accurate registration across different resolutions, viewpoints, and scene complexities without extra retraining or parameter tuning. By addressing resolution inconsistencies, scenario specificity, and nonlinear deformation via multi-scale feature representation and elastic deformation modeling, our method achieves leading performance in RMSE, MAE, and MEE metrics across both cross-resolution and same-resolution scenarios.

To further validate the robustness of the proposed method, we evaluate the matching performance in terms of both the success rate and the number of correct correspondences, as summarized in Tables 4 and 5. In Table 4, our method achieves the highest success rate of 93.6%. More importantly, the number of correct matches obtained by our method is substantially the largest among all competing methods. It significantly exceeds that of the second-best method in terms of correct matches. This result demonstrates that our method is capable

Table 10
Quantitative evaluation results of various feature detection methods on the VIR-UAV dataset [43], obtained from ablation experiments.

Detecotr	RMSE	MAE	MEE
FAST	1.77	1.31	6.41
SURF	4.52	3.97	9.89
Harris	5.32	4.28	13.95
Ours	**1.44**	**1.15**	**4.71**

Table 11
Quantitative evaluation results of various feature detection methods on the Li *et al.*'s dataset [44], obtained from ablation experiments.

Detecotr	RMSE	MAE	MEE
FAST	8.92	8.84	9.55
SURF	20.00	20.00	20.00
Harris	2.42	2.29	3.45
Ours	**1.76**	**1.63**	**2.78**

of generating a large number of highly reliable correspondences, which provides stronger geometric constraints for subsequent transformation estimation.

In Table 5, further confirm the effectiveness of our approach. Our method attains the highest success rate of 99.5%, while also maintaining the largest number of correct matches. It is noteworthy that while 3MRS achieves a relatively high success rate and a considerable number of correct matches, and D2-Net produces a similar quantity of matches, their success rates are significantly lower than that of our method. This indicates that our method excels not only in generating a sufficient number of correspondences but, more critically, in ensuring their high geometric correctness.

In summary, across both datasets, our method consistently achieves the highest or nearly the highest success rate while simultaneously yielding the largest absolute number of correct correspondences. This combination ensures dense and reliable coverage of the overlap regions. Such a property is particularly advantageous in large-scale UAV image registration, where an abundance of correct matches contributes to more robust transformation estimation and ultimately leads to lower global registration errors.

In Fig. 17, our method significantly reduces the AEE compared to other approaches. Specifically, the AEE achieved by our method is controlled within 3 pixels, demonstrating a substantial improvement in registration accuracy. Even when compared with relatively effective methods such as LightGlue and 3MRS, our approach achieves over a 50% improvement in precision, highlighting its robustness and superiority under complex UAV imaging conditions.

To further interpret the AEE in terms of real-world displacement, we convert it into physical error using the Ground Sample Distance (GSD). The resulting values are summarized in Table 6. Under the real-world physical error metric, our method achieves an registration error of only 0.0562 m, significantly outperforming the second-best method, LightGlue (0.1299 m), by 56.7%. In stark contrast, traditional methods such as RIFT (178.16 m) and LNIFT (250.73 m) exhibit catastrophic failures, with errors exceeding those of the top-performing methods by more than three orders of magnitude. The proposed method is capable of maintaining multimodal registration error within the sub-decimeter range (<0.1 m), which is sufficient to meet the requirements of high-precision remote sensing applications.

These results demonstrate that, although validated on UAV platforms, our method exhibits strong generalizability across a wide range of IR and VIS scenes, underscoring its robustness to varying imaging conditions and sensor configurations.

4.2.4. Ablation experiment

This study systematically validates the proposed algorithmic enhancement module through comprehensive ablation experiments on the two datasets. The ablation process follows the algorithmic workflow, starting from the optimal transformation scale selection, followed by the frequency domain filtering module, the feature point detection strategy, and finally the elastic deformation compensation model.

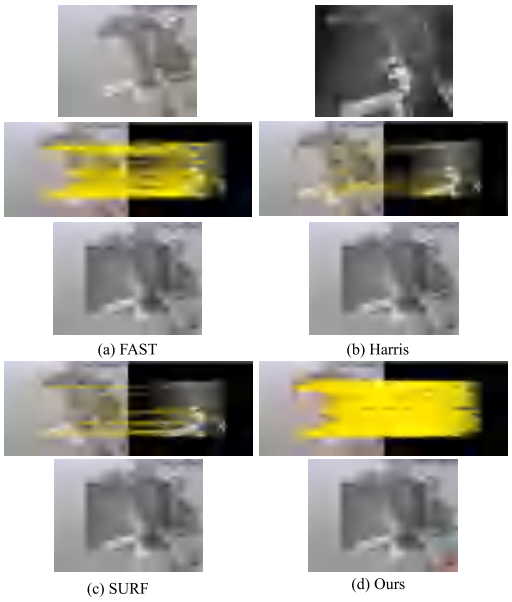


Fig. 19. Feature point detection method ablation experiment showing results.

To evaluate the effectiveness of our weighted RMSE–SSIM criterion for optimal transformation scale selection, we compare three strategies: (1) using RMSE only, (2) using SSIM only, and (3) using the proposed RMSE+SSIM weighted combination. RMSE is widely used in image registration to measure geometric alignment accuracy, while SSIM captures structural similarity and is sensitive to perceptual image quality. By combining them, we aim to simultaneously ensure precise spatial alignment and preserve structural consistency across modalities.

Table 7 summarizes the performance of these strategies on the VIR-UAV datasets. The RMSE+SSIM strategy consistently outperforms the single-metric approaches on dataset, with improvements observed in all three error metrics. This demonstrates that combining geometric and structural measures produces more robust scale estimation, especially under cross-modal UAV imaging conditions.

The algorithmic components include Gaussian Filter (Gauss), Relative Total Variation Denoising (RTVD), and Rolling Guidance Filter (RGF) for frequency domain processing. Feature detection employs FAST, SURF, Harris Corner Detection, and SIFT.

In Tables 8 and 9, our filter outperforms others in all metrics. Fig. 18 further illustrates that our method achieves a significantly higher feature point matching rate than Gauss and RTVD.

In Tables 10 and 11, while different feature point detection methods show some fluctuations, our method remains stable and outperforms others. Fig. 19 visually confirms this superiority.

To highlight the advantages of our elastic compensation model, we compare the registration results with and without it in Fig. 20. The images processed with elastic transformation exhibit clearer building edges, finer structural details, and smoother boundaries. In contrast, those without elastic compensation appear blurrier and show noticeable misregistration, especially in complex regions.

4.2.5. Run time analysis

We analyzed runtime on both datasets, as shown in Tables 12 and 13. On the VIR-UAV dataset, our method ranks mid-tier among eight compared methods, whereas on the Li et al. dataset, the runtime is generally lower, with our method ranking in the third tier. This

Table 12
Time results of various methods on the VIR-UAV dataset [43] (seconds).

Method	LightGlue [35]	D2-Net [34]	MS-PIIFD [46]	RIFT [49]	LNIFT [9]	3MRS [48]	CAO-C2F [47]	MS-HLMO [50]	CoFSM [45]	Ours
Time (s)	9.86	4.62	116.00	80.40	**1.09**	6.68	62.92	137.31	141.56	48.07

Table 13
Time results of various methods on the Li *et al.*'s dataset [44] (seconds).

Method	LightGlue [35]	D2-Net [34]	MS-PIIFD [50]	RIFT [49]	LNIFT [9]	3MRS [48]	CAO-C2F [47]	MS-HLMO [46]	CoFSM [45]	Ours
Time (s)	6.69	**0.40**	23.17	16.34	0.90	1.24	8.20	26.49	3.10	5.48

Fig. 20. Difference between the elastic transformation model and the normal transformation model.

discrepancy is primarily attributed to differences in image resolution and scene complexity between the two datasets. The VIR-UAV dataset contains high-resolution images with large FOV, leading to increased computational cost during both the resolution selection stage where multiple transformations are evaluated to determine the optimal scale and the MCOG feature extraction stage. In contrast, the Li et al. dataset comprises lower-resolution images with simpler backgrounds, which reduces the number of features to be processed and shortens the runtime. Overall, our method's computational cost is sensitive to image size and structural complexity.

5. Conclusions

In this paper, we propose an IR and VIS image registration algorithm specifically designed to address resolution and FOV inconsistency in UAV scenarios. The proposed method leverages a PC-GTV-based technique to enhance the robustness of feature points between cross-modal images and introduces a weighted scale transformation strategy to automatically determine the optimal transformation scale. Furthermore, by integrating our MCOG feature map with an elastic deformation compensation model, we achieve precise pixel-level registration. Experiments on benchmark datasets show that our approach outperforms recent methods, with RMSE, MAE, and MEE improvements exceeding 30% and real-world ground error reduced by over 50%.

However, the proposed method may encounter challenges in scenes with extremely low texture, repetitive patterns, significant illumination changes, or severe sensor noise, which can limit reliable feature extraction and matching. Abrupt or non-rigid object movements may also exceed the elastic deformation model's capability. In future work, we plan to design more robust feature descriptors, develop adaptive strategies for noisy or low-texture regions, and extend the deformation model to handle non-rigid motion. We also aim to explore parallel processing techniques to further accelerate computational efficiency.

In addition, it is worth noting that UAV-based imaging systems may also be affected by intentional or unintentional electromagnetic interference, which can degrade image quality and lead to distortion or other errors [54,55]. Although this work primarily focuses on registration issues caused by resolution, field-of-view, and modality differences, evaluating the robustness of the proposed method under noisy environments represents an important direction for future research.

In summary, our method effectively reduces cross-modal registration error, improving the spatial accuracy of localized structures and enabling more reliable measurements in tasks such as thermal distribution analysis, structural sizing, and multi-modal inspection.

CRediT authorship contribution statement

Hao Li: Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Data curation, Conceptualization. **Chenhua Liu:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Data curation, Conceptualization. **Maoyong Li:** Formal analysis, Data curation, Conceptualization. **Lei Deng:** Visualization, Software, Project administration, Methodology. **Mingli Dong:** Resources, Project administration, Funding acquisition. **Lianqing Zhu:** Resources, Project administration, Funding acquisition.

Consent for publication

Written informed consent for publication was obtained from all participants.

Declaration of competing interest

No potential conflict of interest was reported by the authors.

Data availability

Within the scope of academic research, the datasets and related research source codes used in this article have been made public and can be used freely without any conflict of interest.

References

[1] Xingfang Zhou, Zujun Yu, Tao Ruan, Baoqing Guo, Dingyuan Bai, Tao Sun, Robust IR–VIS image registration with different FOVs in railway intrusion detection, Measurement 225 (2024) 113928.

[2] Long Liu, Bing Yi, Jia Liu, A robust scaling registration method for rail profile inspection, Measurement 249 (2025) 116972.

[3] Zekun Sun, Li Li, Ning Chu, Huajiang Ren, Keke Tu, Caifang Cai, Ali Mohammad-Djafari, An infrared-optical image registration method for industrial blower monitoring based on contour-shape descriptors, Measurement 240 (2025) 115634.

[4] Chenhua Liu, Hanrui Chen, Lei Deng, Chentong Guo, Xitian Lu, Heng Yu, Lianqing Zhu, Mingli Dong, Modality specific infrared and visible image fusion based on multi-scale rich feature representation under low-light environment, Infrared Phys. Technol. 140 (2024) 105351.

[5] Chentong Guo, Chenhua Liu, Lei Deng, Zhixiang Chen, Mingli Dong, Lianqing Zhu, Hanrui Chen, Xitian Lu, Multi-scale infrared and visible image fusion framework based on dual partial differential equations, Infrared Phys. Technol. 135 (2023) 104956.

[6] Hanrui Chen, Lei Deng, Zhixiang Chen, Chenhua Liu, Lianqing Zhu, Mingli Dong, Xitian Lu, Chentong Guo, SFCFusion: Spatial-frequency collaborative infrared and visible image fusion, IEEE Trans. Instrum. Meas. (2024).

[7] Hao Li, Shengkun Wu, Lei Deng, Chenhua Liu, Yifan Chen, Hanrui Chen, Heng Yu, Mingli Dong, Lianqing Zhu, Enhancing infrared and visible image fusion through multiscale Gaussian total variation and adaptive local entropy, Vis. Comput. (2025).

[8] Sayed Ishaq Deliry, Uğur Avdan, Accuracy assessment of UAS photogrammetry and structure from motion in surveying and mapping, Int. J. Eng. Geosci. 9 (2) (2024) 165–190.

[9] Jiayuan Li, Wangyi Xu, Pengcheng Shi, Yongjun Zhang, Qingwu Hu, LNIFT: Locally normalized image for rotation invariant multimodal feature matching, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–14.

[10] Wujie Zhou, Yusen Wang, Xiaohong Qian, Knowledge distillation and contrastive learning for detecting visible-infrared transmission lines using separated stagger registration network, IEEE Trans. Circuits Syst. I. Regul. Pap. (2025).

[11] Jiangang Ding, Yuanlin Zhao, Lili Pei, Yihui Shan, Yiquan Du, Wei Li, Modal-invariant progressive representation for multimodal image registration, Inf. Fusion 117 (2025) 102903.

[12] Xu Zhang, Felix X. Yu, Svebor Karaman, Shih-Fu Chang, Learning discriminative and transformation covariant local feature detectors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6818–6826.

[13] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, Krystian Mikolajczyk, Key. net: Keypoint detection by handcrafted and learned cnn filters, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5836–5844.

[14] Alessa Hering, Sven Kuckertz, Stefan Heldmann, Mattias P Heinrich, Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking, in: Bildverarbeitung FÜr Die Medizin 2019: Algorithmen–Systeme–Anwendungen. Proceedings Des Workshops Vom 17. Bis 19. MÄrz 2019 in LÜbeck, Springer, 2019, pp. 309–314.

[15] Inwan Yoo, David GC Hildebrand, Willie F Tobin, Wei-Chung Allen Lee, Won-Ki Jeong, ssEMnet: Serial-section electron microscopy image registration using a spatial transformer network with learned features, in: International Workshop on Deep Learning in Medical Image Analysis, Springer, 2017, pp. 249–257.

[16] Yuanxin Ye, Tengfeng Tang, Bai Zhu, Chao Yang, Bo Li, Siyuan Hao, A multiscale framework with unsupervised learning for remote sensing image registration, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15.

[17] Masato Ito, Fumihiko Ino, An automated method for generating training sets for deep learning based image registration, in: Bioimaging, 2018, pp. 140–147.

[18] Junchi Bin, Heqing Zhang, Zhila Bahrami, Ran Zhang, Huan Liu, Erik Blasch, Zheng Liu, The registration of visible and thermal images through multi-objective optimization, Inf. Fusion 95 (2023) 186–198.

[19] Yuandong Ma, Meng Yu, Hezheng Lin, Chun Liu, Mengjie Hu, Qing Song, Efficient networks for textureless feature registration via free receptive field, Inf. Fusion 108 (2024) 102371.

[20] Colin Studholme, Derek L.G. Hill, David J. Hawkes, An overlap invariant entropy measure of 3D medical image alignment, Pattern Recognit. 32 (1) (1999) 71–86.

[21] Alexis Roche, Grégoire Malandain, Xavier Pennec, Nicholas Ayache, The correlation ratio as a new similarity measure for multimodal image registration, in: Medical Image Computing and Computer-Assisted Intervention—MICCAI'98: First International Conference Cambridge, MA, USA, October 11–13, 1998 Proceedings 1, Springer, 1998, pp. 1115–1124.

[22] Hassan Rivaz, Zahra Karimaghaloo, Vladimir S Fonov, D Louis Collins, Non-rigid registration of ultrasound and MRI using contextual conditioned mutual information, IEEE Trans. Med. Imaging 33 (3) (2013) 708–725.

[23] Johan Öfverstedt, Joakim Lindblad, Nataša Sladoje, Fast computation of mutual information in the frequency domain with applications to global multimodal image alignment, Pattern Recognit. Lett. 159 (2022) 196–203.

[24] Xingyu Jiang, Jiayi Ma, Guobao Xiao, Zhenfeng Shao, Xiaojie Guo, A review of multimodal image matching: Methods and applications, Inf. Fusion 73 (2021) 22–71.

[25] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, Junchi Yan, Image matching from handcrafted to deep features: A survey, Int. J. Comput. Vis. 129 (1) (2021) 23–79.

[26] Chris Harris, Mike Stephens, et al., A combined corner and edge detector, in: Alvey Vision Conference, vol. 15, (50) Citeseer, 1988, pp. 10–5244.

[27] David G. Low, Distinctive image features from scale-invariant keypoints, J. Comput. Vis. 60 (2) (2004) 91–110.

[28] Herbert Bay, Tinne Tuytelaars, Luc Van Gool, Surf: Speeded up robust features, in: Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9, Springer, 2006, pp. 404–417.

[29] Edward Rosten, Tom Drummond, Machine learning for high-speed corner detection, in: Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9, Springer, 2006, pp. 430–443.

[30] Ethan Rublee, Vincent Rabaud, Kurt Konolige, Gary Bradski, ORB: An efficient alternative to SIFT or SURF, in: 2011 International Conference on Computer Vision, 2011, pp. 2564–2571.

[31] Pablo Fernández Alcantarilla, Adrien Bartoli, Andrew J. Davison, KAZE features, in: Computer Vision – ECCV 2012, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 214–227.

[32] Daniel DeTone, Tomasz Malisiewicz, Andrew Rabinovich, SuperPoint: Self-supervised interest point detection and description, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2018, pp. 337–33712.

[33] Xiaoyong Lu, Songlin Du, Jamma: Ultra-lightweight local feature matching with joint mamba, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 14934–14943.

[34] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, Torsten Sattler, D2-net: A trainable cnn for joint description and detection of local features, in: Proceedings of the Ieee/Cvf Conference on Computer Vision and Pattern Recognition, 2019, pp. 8092–8101.

[35] Philipp Lindenberger, Paul-Edouard Sarlin, Marc Pollefeys, LightGlue: Local feature matching at light speed, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2023, pp. 17627–17638.

[36] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, Michael Felsberg, DKM: Dense kernelized feature matching for geometry estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 17765–17775.

[37] Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, Cheng Wang, Gim: Learning generalizable image matcher from internet videos, 2024, arXiv preprint arXiv:2402.11095.

[38] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, Michael Felsberg, Roma: Robust dense feature matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 19790–19800.

[39] Jiangwei Ren, Xingyu Jiang, Zizhuo Li, Dingkang Liang, Xin Zhou, Xiang Bai, Minima: Modality invariant image matching, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 23059–23068.

[40] Daniel Rueckert, Paul Aljabar, Non-rigid registration using free-form deformations, in: Handbook of Biomedical Imaging: Methodologies and Clinical Research, Springer, 2015, pp. 277–294.

[41] Mengmeng Wang, Jixian Zhang, Kazhong Deng, Fenfen Hua, Combining optimized SAR-SIFT features and RD model for multisource SAR image registration, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–16.

[42] Qiuze Yu, Dawen Ni, Yuxuan Jiang, Yuxuan Yan, Jiachun An, Tao Sun, Universal SAR and optical image registration via a novel SIFT framework based on nonlinear diffusion and a polar spatial-frequency descriptor, ISPRS J. Photogramm. Remote Sens. 171 (2021) 1–17.

[43] Yan Mo, Xudong Kang, Shuo Zhang, Puhong Duan, Shutao Li, A robust infrared and visible image registration method for dual-sensor UAV system, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–13.

[44] Jiayuan Li, Qingwu Hu, Yongjun Zhang, Multimodal image matching: A scale-invariant algorithm and an open dataset, ISPRS J. Photogramm. Remote Sens. 204 (2023) 77–88.

[45] Yongxiang Yao, Yongjun Zhang, Yi Wan, Xinyi Liu, Xiaohu Yan, Jiayuan Li, Multi-modal remote sensing image matching considering co-occurrence filter, IEEE Trans. Image Process. 31 (2022) 2584–2597.

[46] Chenzhong Gao, Wei Li, Ran Tao, Qian Du, MS-HLMO: Multiscale histogram of local main orientation for remote sensing image registration, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–14.

[47] Qian Jiang, Yadong Liu, Yingjie Yan, Jun Deng, Jian Fang, Zhe Li, Xiuchen Jiang, A contour angle orientation for power equipment infrared and visible image registration, IEEE Trans. Power Deliv. 36 (4) (2020) 2559–2569.

[48] Zhongli Fan, Yuxian Liu, Yuxuan Liu, Li Zhang, Junjun Zhang, Yushan Sun, Haibin Ai, 3MRS: An effective coarse-to-fine matching method for multimodal remote sensing imagery, Remote. Sens. 14 (3) (2022) 478.

[49] Jiayuan Li, Qingwu Hu, Mingyao Ai, RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform, IEEE Trans. Image Process. 29 (2019) 3296–3310.

[50] Chenzhong Gao, Wei Li, Multi-scale PIIFD for registration of multi-source remote sensing images, 2021, arXiv preprint arXiv:2104.12572.

[51] Song Cui, Ailong Ma, Liangpei Zhang, Miaozhong Xu, Yanfei Zhong, MAP-Net: SAR and optical image matching via image-based convolutional network with attention mechanism and spatial pyramid aggregated pooling, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–13.

[52] Yuan Li, Chuanfeng Wei, Dapeng Wu, Yaping Cui, Peng He, Yuan Zhang, Ruyan Wang, A robust multisource remote sensing image matching method utilizing attention and feature enhancement against noise interference, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–21.

[53] Feiyan Cheng, Yiteng Zhou, Xiaoqiao Huang, Ruimin Huang, Yonghang Tai, Junsheng Shi, CycleRegNet: A scale-aware and geometry-consistent cycle adversarial model for infrared and visible image registration, Measurement 242 (2025) 116063.

[54] Huamin Jie, Zhenyu Zhao, Yu Zeng, Yongqi Chang, Fei Fan, Changdong Wang, Kye Yak See, A review of intentional electromagnetic interference in power electronics: Conducted and radiated susceptibility, IET Power Electron. 17 (12) (2024) 1487–1506.

[55] Huamin Jie, Zhenyu Zhao, Hong Li, Theng Huat Gan, Kye Yak See, A systematic three-stage safety enhancement approach for motor drive and gimbal systems in unmanned aerial vehicles, IEEE Trans. Power Electron. (2025).