Optics and Lasers in Engineering 197 (2026) 109472

Contents lists available at ScienceDirect

Optics and Lasers in Engineering

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

ECPCS: Enhanced contrast phase consistency space for visible and infrared image registration

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

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

ARTICLE INFO

Keywords:
Infrared image
Visible image
UAV
Registration
ECPCS
Flexible transformation

ABSTRACT

In Unmanned Aerial Vehicle (UAV) aerial photography scenarios, differences in lens imaging often lead to variations in image resolution, parallax, and modality, resulting in low feature point detection efficiency and reduced matching accuracy in image registration tasks. To address these challenges, we propose a progressive framework for infrared (IR) and visible (VIS) image registration. Firstly, the source image is transformed into the enhanced phase-consistent space using the Enhanced Contrast Phase Consistency Space (ECPCS) method. The optimal transformation scale of the IR image is then determined through Root Mean Square Error (RMSE) and Mutual Information (MI) optimization, thereby eliminating resolution and parallax discrepancies. Secondly, the Harris feature point detection algorithm is applied in the ECPCS to efficiently extract feature points. Finally, channel features of orientated gradients (CFOG) is employed to generate the feature map of the source image, and based on this map, Fast Fourier Transform (FFT) is used to refine feature point positions and complete the image matching. The final registration result is obtained using a flexible transformation model. Extensive experiments demonstrate that the proposed method achieves excellent performance, validated both qualitatively and quantitatively on public datasets. Our code is available at https://github.com/lh-ite/ECPC.

1. Introduction

Image registration is a key technology in the field of computer vision and plays an important role in applications [1] such as image fusion [2–5], automatic driving [6] and medical image analysis [7,8]. However, in Unmanned Aerial Vehicle (UAV) application scenarios, due to the differences in sensor types and installation locations, there are resolution differences, geometric deformations, and modality differences between infrared (IR) and visible (VIS) images, which make image registration in this scenario a significant challenge. In UAV aerial imaging scenarios, differences in sensor resolution, viewing angle, and spectral response often lead to scale inconsistency, parallax, and radiometric variation between IR and VIS images. These factors greatly increase the difficulty of achieving accurate feature correspondences, causing some registration algorithms to suffer from geometric distortions and reduced robustness. Therefore, accurate and efficient IR and VIS image registration algorithms are a prerequisite.

Mainstream methods for IR and VIS image registration can be broadly categorized into deep learning-based and traditional approaches [9]. Deep learning-based methods [10–13] have demonstrated strong nonlinear modeling capability and can learn rich feature representations from data, often yielding high accuracy in benchmark datasets. However, their performance often deteriorates in real-world applications due to domain shifts, such as varying illumination, modality inconsistencies, or scene complexity. In particular, these methods may produce unstable registration results when facing large viewpoint differences or degraded sensor conditions, as their generalization ability is highly dependent on the diversity and quality of training data. Moreover, deep models often fail to ensure geometric consistency, leading to local misalignment or distortion in the final registered images.

Traditional image registration methods mainly include two types: region-based methods and feature point-based methods. Region-based methods [9,14,15] usually match pixels or regions based on gray scale similarity to achieve registration. However, they tend to have low com-

* 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.optlaseng.2025.109472
Received 24 July 2025; Received in revised form 19 October 2025; Accepted 17 November 2025
Available online 26 November 2025

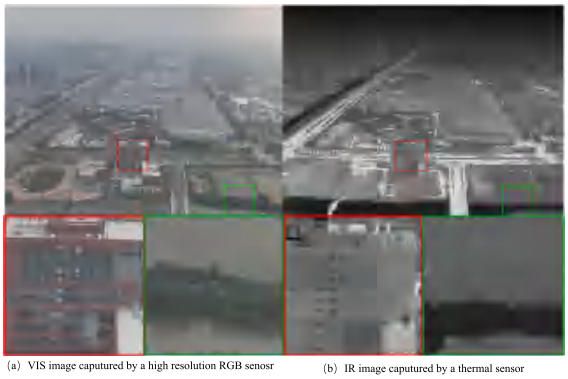


Fig. 1. IR and VIS images, parallax and modal difference display in UAV aerial photography scenes. Significant scale and parallax differences are evident due to altitude variation and sensor offset, which lead to geometric and radiometric misalignments between the two modalities.

putational efficiency and are highly sensitive to geometric variations in the image. This results in poor robustness and limits their effectiveness in complex application scenarios such as UAVs [16–18].

In contrast, feature point-based methods [19–22] have higher computational efficiency and stability by extracting key feature points and matching them to establish a spatial transformation model to accomplish image registration. However, in UAV application scenarios, the accuracy and robustness of such methods are still affected by the differences in resolution, parallax, and modality between IR and VIS images, as shown in Fig. 1.

To address the challenges of resolution discrepancies, parallax, and modality differences between IR and VIS images in UAV scenarios, this paper proposes a progressive registration framework for IR and VIS image registration. In the first stage, the images are transformed into an enhanced phase-consistent space [23] using the proposed Enhanced Contrast Phase Consistency Space (ECPCS) method, followed by optimization based on MI and RMSE to determine the optimal transformation scale. In the second stage, the ECPCS is applied again, and Harris corner detection is used to extract the feature point locations. In the final stage, the feature point positions are refined using channel features of orientated gradients (CFOG) [24], and the final registration result is obtained through a flexible transformation model. The main contributions can be summarized as follows:

(1) To address the challenges of inconsistent image resolution, parallax, and modality differences in UAV aerial imaging scenarios, we propose a progressive registration framework for IR and VIS images.

(2) We propose an optimization method that integrates MI and RMSE to mitigate resolution inconsistency and parallax. This is achieved by transforming the IR image into the coordinate system of the VIS image, thereby effectively reducing spatial misalignment.

(3) To tackle the difficulties in feature point matching caused by modality differences, we propose the ECPCS method, which transforms images into an enhanced phase-consistent space to improve the accuracy and robustness of feature point detection and registration.

2. Related work

This chapter analyzes traditional registration methods in terms of feature point detection and matching. Feature detection focuses on extracting keypoints from cross-modal images, while matching aims to overcome modality differences to establish accurate point correspondences. Together, these two aspects contribute to improving registration robustness in UAV scenarios.

2.1. Feature point detection

Feature point-based registration methods primarily work by extracting salient features from images and establishing spatial correspondences between them [25]. Therefore, the choice of feature point detection algorithm plays a particularly crucial role in ensuring accurate registration to a unified coordinate system. In 1980, Moravec et al. [26] proposed the Moravec corner point detection algorithm , which calculates the gray-scale changes of local regions in different directions through a sliding window, and obtains the final corner points. However, due to its sensitivity to noise and low directional resolution, Harris et al. [27] proposed the Harris corner detection algorithm, which is rotationally invariant by referring to the autocorrelation matrix and the corner response function. In 1999, David et al. [28] proposed the Scale-Invariant Feature Transform (SIFT) feature detection algorithm, a scale-space based feature detection algorithm, which is rotationally, scale- and illumination-invariant. Then Bay et al. [29] proposed the Speeded-Up Robust Features (SURF) algorithm based on Hessian matrix in order to solve the problem of slow computation speed of SIFT. In the same year Rosten et al. [30] proposed the Features from Accelerated Segment Test (FAST) corner point detection algorithm in order to be suitable for real-time applications. Since then, various improvements and novel feature detectors have been developed to enhance robustness, efficiency, and adaptability in different imaging conditions. For example, in 2022, Zhang et al. [31] proposed a robust registration method for SAR and optical images by combining deep learning with an improved Harris corner detector, aiming to address the challenges of cross-modality matching. In 2024, Cong et al. [32] introduced an image stitching technique tailored for police drones, which utilizes an enhanced registration pipeline based on the Oriented FAST and Rotated BRIEF (ORB) algorithm to improve efficiency and alignment accuracy in aerial surveillance tasks.

2.2. Feature matching and spatial transformation

The integration of feature matching and spatial transformation model estimation forms the core of image registration. Feature matching aims to establish correspondences between feature points in two images, typically using methods such as nearest neighbor search [28] and the ratio test. However, false matches are common, and the Random Sample Consensus (RANSAC) algorithm [33] is widely employed to robustly eliminate these outliers.

The estimation of the spatial transformation model can be broadly categorized into two types: rigid and non-rigid transformations. Rigid transformation models account only for rotation and translation, assuming that the image structure remains unchanged during the transformation. Commonly used rigid models include Euclidean and affine transformations.

Non-rigid transformation models allow for the presence of deformations within images and aim to resolve misalignments by constructing either linear or nonlinear models based on point correspondences. In 1989, Bookstein [34] introduced the Thin Plate Spline (TPS) model, which generates smooth, non-rigid deformation fields by matching corresponding points between two images, effectively mapping the source image to the target image. Later, Rueckert et al. [35] proposed the B-spline-based registration method, which constructs a grid of control points over the images and generates deformation fields by interpolating them using a smooth B-spline function. This approach enables flexible and continuous registration of image content.

Although feature point-based registration methods have demonstrated broad applicability in computer vision, existing research still faces two critical challenges. First, under complex imaging conditions, the limited discriminative power of traditional feature descriptors often leads to a significant decline in matching accuracy. Second, when addressing localized geometric deformations, global transformation based registration methods often produce noticeable resampling artifacts and

edge blurring. These issues arise not merely from interpolation or blending choices, but primarily from structural limitations of global parametric models (e.g., homography or affine), which cannot accurately accommodate local non-rigid distortions. As a result, mismatched regions are warped inconsistently, leading to visual discontinuities even when standard interpolation schemes are applied.

3. Proposed method

Fig. 2 shows our proposed framework. Our workflow consists of three stages: firstly, the IR image is transformed to the optimal scale using the ECPCS method combined with MI and RMSE optimization; secondly, ECPCS is applied to the IR image to facilitate accurate feature point extraction; and finally, the feature points are refined using the CFOG feature map and Fast Fourier Transform (FFT) [36], followed by image matching and final registration based on the transformed model.

3.1. Scale transformation

To mitigate the negative impact of scale inconsistency in IR and VIS image registration, we first apply a scale transformation to adjust both images to an appropriate scale, as shown in Fig. 3. We initialize the affine transformation matrix, as shown in Eq. (1):

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a & b & t_x \\ c & d & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \tag{1}$$

where x and y are the original coordinates, a, b, c, d control the zoom, rotate and cut operations of the image, t_x and t_y control the translation operation, and x', y' are the result of the transformation.

To obtain the optimal transformation scale, we define a range of S values and apply an optimization algorithm to identify the best scale. The affine transformation matrix H is given by the following equation:

$$H = \begin{bmatrix} S & 0 & \text{cent}_x \\ 0 & S & \text{cent}_y \\ 0 & 0 & 1.000 \end{bmatrix}, \tag{2}$$

where S is the scaling factor. The parameters cent_x and cent_y are the translation offsets that align the center of the scaled infrared image with the center of the VIS image. In the image coordinate system, cent_x denotes the vertical translation, and cent_y denotes the horizontal translation. The definitions of $cent_x$ and $cent_y$ are shown in Eq. (3):

$$\begin{cases} \text{cent}_x = \dfrac{VI_x}{2} - \dfrac{IR_x \times S}{2} \\ \text{cent}_y = \dfrac{VI_y}{2} - \dfrac{IR_y \times S}{2} \end{cases} \tag{3}$$

where VI_x, VI_y, IR_x and IR_y represent the number of rows and columns of the VIS and IR image, respectively. After obtaining the results for different transformation scales, we first map them into a phase-consistent space using the phase congruency (PC) technique, as shown in Eq. (4):

$$PC = \frac{1}{\sum_{s=1}^{n} A_s} \times \left(\sum_{s=1}^{n} \left(E_s \cos(\phi_s) + O_s \sin(\phi_s) \right) - \sum_{s=1}^{n} \left| E_s \sin(\phi_s) - O_s \cos(\phi_s) \right| \right), \tag{4}$$

where E_s and O_s are the convolution results of the even-symmetric and odd-symmetric filters on the sth scale, respectively, ϕ_s is the phase angle at the sth scale, and A_s is the amplitude response at the sth scale, which can be described as follows:

$$A_s = \left| E_s + iO_s \right| = \sqrt{E_s^2 + O_s^2}. \tag{5}$$

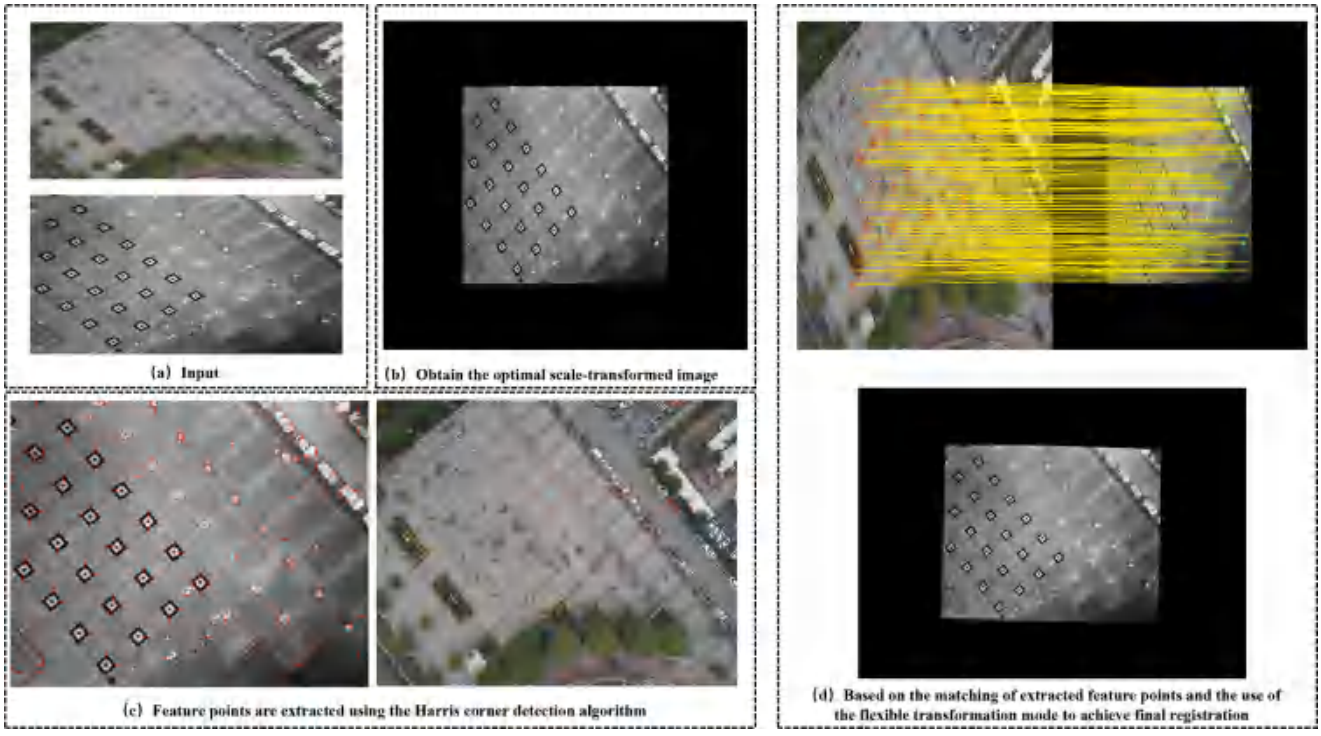


Fig. 2. Image registration framework addressing geometric misalignment. (a) Input images. (b) The optimal scale for the IR image is determined by optimizing RMSE and MI. (c) Harris feature points are detected in the ECPCS, and their positions are refined using CFOG and FFT. (d) The final registration is achieved through a flexible transformation model based on accurately matched feature points.

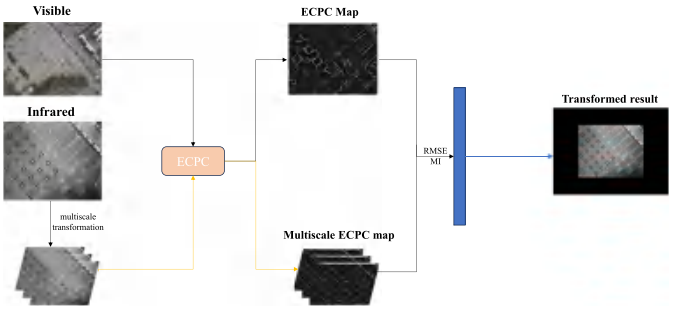


Fig. 3. Flowchart of IR image scale transformation. The IR and VIS image is first transformed into the ECPCS through the ECPCS operation, and the optimal scale-transformed infrared image is then obtained via MI and RMSE-based optimization.

In order to remove the effect of noise and ensure PC stability, we introduce a notion T of noise threshold for the energy E, as shown in Eqs. (6) and (7):

$$\begin{cases} \mathrm{EM} = \tau\sqrt{\dfrac{\pi}{2}} \\ \mathrm{ES} = \tau\sqrt{\dfrac{4-\pi}{2}} \end{cases} \tag{6}$$

$$T = \mathrm{EM} + k \cdot \mathrm{ES}, \tag{7}$$

where τ is a parameter of the Rayleigh distribution, which can be estimated from the median or plurality of the data, and EM and ES represent the estimated noise energy mean and estimated noise energy standard deviation, respectively. k is a user-defined noise compensation factor. T is then applied to the calculation of E as shown in Eq. (8):

$$E = \max(E - T, 0), \tag{8}$$

where max(·) means take the maximum operation. To suppress noise while enhancing feature significance, we propose a penalty weight coefficient ω, as shown in Eq. (9):

$$w = \frac{1}{1 + \exp\left(g\left(\mathrm{cutOff} - \frac{\sum_{s=1}^{n} A_s}{\max(A_s)}\right)\right)}, \tag{9}$$

where cutOff is the cutoff value of the frequency distribution and g is the sharpness parameter of the sigmoid function. Based on w we can get the final PC, as shown in Eq. (10):

$$PC = w \cdot \frac{1}{\sum_{s=1}^{n} A_s} \times \left(\sum_{s=1}^{n} \left(E_s \cos(\phi_s) + O_s \sin(\phi_s) \right) - \sum_{s=1}^{n} \left| E_s \sin(\phi_s) - O_s \cos(\phi_s) \right| \right), \tag{10}$$

After that we can calculate the edge strength M by PC as shown in Eq. (11):

$$M = \frac{1}{2}\left(c + a + \sqrt{b^2 + (a-c)^2}\right), \tag{11}$$

where a, b, c denote, respectively, the sum of squares of the projections of the PC in the x-direction in all directions, the sum of twice the product of the projections in the x- and y-directions, and the sum of squares of the projections in the y-direction. The terms a, b, and c defined in Eq. (12) form the components of a local structure tensor, which is built from the oriented Phase Congruency projections [23]. The largest eigenvalue of M is a classical measure for feature detection, as it quantifies the corner strength by indicating the maximum intensity of local image variations. Its calculation follows the standard eigenvalue formulation:

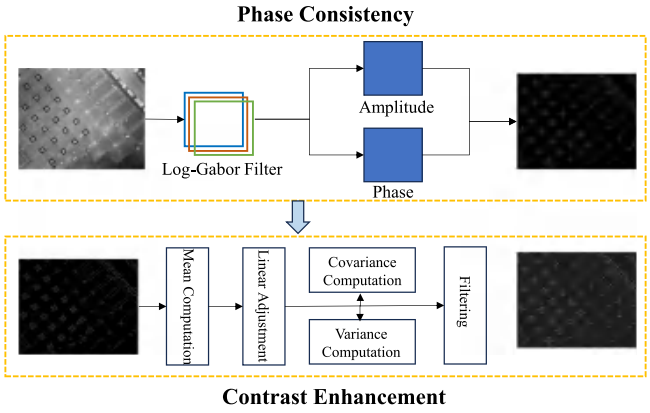


Fig. 4. The process of the ECPCS. The upper part illustrates the process of obtaining the phase-consistent map, while the lower part shows the enhancement of the phase-consistent map using contrast enhancement techniques. The ECPCS module significantly improves phase contrast and structural consistency compared with the baseline PC representation.

$$\begin{cases} a = \sum\limits_{o} \left[PC(\theta_o)\cos(\theta_o) \right]^2 \\ b = 2\sum\limits_{o} \left[PC(\theta_o)\cos(\theta_o) \cdot PC(\theta_o)\sin(\theta_o) \right] \\ c = \sum\limits_{o} \left[PC(\theta_o)\sin(\theta_o) \right]^2 \end{cases} \tag{12}$$

where o denotes the number of wavelet directions, θ_o represents the angular value corresponding to each wavelet direction, and $PC(\theta_o)$ indicates the phase-consistent response at direction θ_o.

To address the issue of blurred edges in IR images, we introduce the ECPCS method to enhance edge strength in the phase-consistent maps, as shown in Fig. 4. Given the inconsistent brightness profiles across different regions of an image, we perform adaptive contrast enhancement on a block-wise basis. Specifically, we first compute the mean and variance of each local region B, as shown in Eqs. (13) and (14):

$$\mu_B = \frac{1}{H \cdot W} \iint_B B(x,y)\,dx\,dy, \tag{13}$$

$$\sigma_B^2 = \frac{1}{H \cdot W} \iint_B B(x,y)^2\,dx\,dy - \mu_B^2, \tag{14}$$

where H and W represent the height and width of the region, respectively, and $B(x,y)$ corresponds to the image intensity at the position. We approximate the mean value of the local region as the low-frequency component, allowing the high-frequency component to be extracted by subtracting it from the original image. To enhance local contrast, we introduce an enhancement factor α, as defined in Eq. (15):

$$L_B(x,y) = \mu_B + \alpha\big(B(x,y) - \mu_B\big), \tag{15}$$

where L_B represents the enhanced pixel intensity, and $\alpha = \frac{\sigma_G}{\sigma_B}$, the ratio of the global variance σ_G to the local variance σ_B, is used to dynamically adjust the degree of enhancement.

Furthermore, we introduce an enhancement cutoff parameter to prevent over-enhancement when α is excessively large. The enhancement process is described as follows:

$$L_B(x,y) = \begin{cases} \mu_B + \frac{\sigma_G}{\sigma_B}\big(I(x,y) - \mu_B\big), & \text{if } \frac{\sigma_G}{\sigma_B} < \beta \\ \mu_B + \beta\big(I(x,y) - \mu_B\big), & \text{if } \frac{\sigma_G}{\sigma_B} \geq \beta \end{cases} \tag{16}$$

To further refine edge details in localized areas, we introduce a guided filter to enhance the refinement of the enhancement results, as

shown in Eq. (17):

$$L_{gfB}(x,y) = \kappa(x,y) \cdot L_B(x,y) + \nu(x,y), \tag{17}$$

where $\kappa(x,y)$ and $\nu(x,y)$ are linear coefficients obtained by minimizing the error function.

Finally, we normalize the enhanced result to ensure that pixel values remain within the range [0, 1], as shown in Eq. (18):

$$\mathrm{E}(x,y) = \frac{\mathrm{E}\,(x,y) - \min(\mathrm{E}\,)}{\max(\mathrm{E}\,) - \min(\mathrm{E}\,)}, \tag{18}$$

As shown in Fig. 4, ECPCS enhances the phase consistency features by improving edge sharpness and structural contrast compared with the baseline PC representation. The quantitative benefit of ECPCS is further validated in Tables 7 and 8, where the ablation results demonstrate consistent improvements across all evaluation metrics.

The proposed ECPCS algorithm structurally mitigates the modality discrepancy between IR and VIS sensors. Conventional PC-based methods are sensitive to contrast and gradient variations across modalities, which often lead to inconsistent feature responses under different radiation or illumination conditions. In contrast, ECPCS introduces a local adaptive contrast enhancement on the phase consistent maps, which balances local intensity variations and amplifies salient structural cues while suppressing modality-specific artifacts. This process effectively aligns the structural representations of both modalities within a unified contrast-invariant feature space, thereby overcoming sensor-type differences at the feature level.

To determine the transformation scale that best matches the IR image, we calculate the weighted mutual information (MI) and root mean square error (RMSE) between the transformed IR and VIS image, and select the scale with the optimal combined metric as the final solution. The computational description of RMSE and MI is shown in Eqs. (19) and (20):

$$MI(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right), \tag{19}$$

where X and Y represent two image inputs, $p(x,y)$ is the joint probability distribution of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distributions of X and Y, respectively.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{20}$$

where y_i is the true value, $\hat{y}_i$ is the predicted value and N represents the overall number of pixel points.

Since RMSE tends to be smaller and MI tends to be larger, it is necessary to perform an optimization transformation on both metrics. The optimization of RMSE is carried out as shown in the following equation:

$$RMSE_n = 1 - \left(\frac{R_L}{max(R_L)}\right), \tag{21}$$

where R_L represents the list of RMSE for different transformation scales, and $max(\cdot)$ represents the maximize operation. Subsequently, we introduce an adaptive parameter α to balance the influence of the two metrics, thereby determining the final transformation scale, as shown in Eq. (22):

$$Scale = max(\alpha MI_l + (1-\alpha)\cdot RMSE_n), \tag{22}$$

where α is the adaptive parameter, MI_l denotes the list of MI values corresponding to different scale cases, and $\max(\cdot)$ indicates the scale associated with the maximum values of the weighted mutual information and RMSE.

3.2. Feature point extraction

In the previous section, we obtained the IR image after the scale transformation, and then we processed it to obtain the feature points, the overall process is shown in Fig. 5. First we convert it to ECPCS and do local adaptive contrast enhancement on it as shown in Eq. (23):

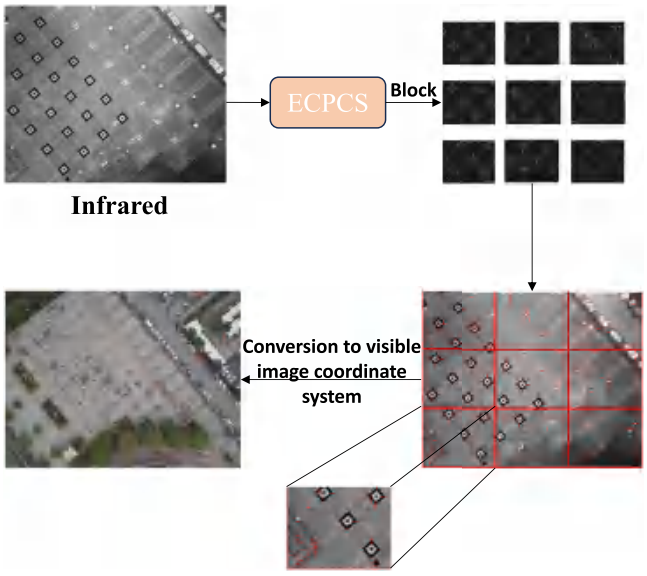


Fig. 5. Feature Point Extraction Process. The image is first converted into the ECPCS, and then aligned for feature point detection using a blocking strategy.

$$E_m = ECPCS(I), \tag{23}$$

where E_m denotes the enhanced phase-consistent map, ECPCS represents the Enhanced Contrast Phase Consistency Space operation, and I is the input image.

Since traditional feature point detection methods operate on the entire source image, the low resolution of IR images often leads to the extraction of many redundant points, negatively impacting subsequent matching. To address this, we adopt a block-based strategy for feature point detection, as shown in Eq. (24):

$$B_{i,j}(x,y) = I((i-1)\cdot h + x, (j-1)\cdot w + y), \tag{24}$$

where i, j represent the coordinates of the blocks, which are all in the range of $1-n$, I is the original input matrix, and $h = \frac{H}{N}$, $w = \frac{W}{n}$ represent the height and width of the small blocks, respectively.

Next, we apply the Harris feature point detection algorithm to each block individually. However, since the detected feature point coordinates are relative to each block, a coordinate system transformation is required to map them back to the global image space, as shown in the following equation:

$$\begin{cases} x = (i-1)\cdot h + x' \\ y = (j-1)\cdot w + y' \end{cases} \tag{25}$$

where x', y' denotes the feature points of the coordinates inside the block, and x, y represents the transformed coordinates.

3.3. Feature point matching

After extracting the feature points in the previous step, we proceed to the feature point matching process, as shown in Fig. 6. First, we perform CFOG operations on the input IR and VIS images to obtain CFOG feature maps.

The CFOG descriptor encodes the local gradient distribution by constructing orientation histograms over multiple Gaussian-smoothed gradient channels. This process effectively suppresses modality-specific intensity variations and highlights structural edges that are invariant to illumination and radiation differences. As a result, the CFOG representation enhances the mutual consistency between IR and VIS modalities,

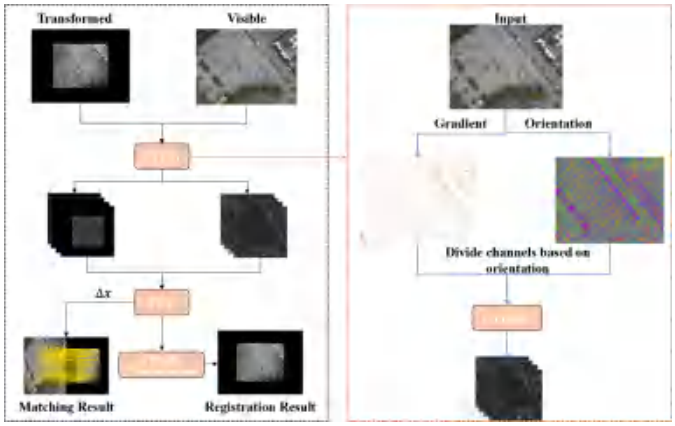


Fig. 6. Image matching and registration process. The feature map is generated using CFOG, and the feature point deviations are subsequently corrected with FFT to complete the matching and registration process.

making the feature responses more stable and reliable for cross-modal correspondence. To further illustrate this effect, Fig. 6 includes intermediate visualizations of the CFOG maps, where edges and contours are more distinctly aligned between modalities compared to raw gradient responses.

Then, for each detected feature point, we perform Fast Fourier Transform (FFT)-based matching within a region of size batchsize centered at the point, enabling accurate position correction of the feature points. The following equation is used to calculate the corresponding area:

$$B(x,y)=\begin{cases} C\left(\max\left(x-\frac{b}{2},1\right):\min\left(x+\frac{b}{2},H\right),\right. \\ \left.\max\left(y-\frac{b}{2},1\right):\min\left(y+\frac{b}{2},W\right)\right)\end{cases} \tag{26}$$

where b denotes the batch size, C represents the input feature map, and $B(x,y)$ refers to the local region centered at the corresponding point (x,y). The fast Fourier matching is shown in Eqs. (27) and (28):

$$\text{correlation}=\left|\mathcal{F}^{-1}\left(F_{\text{correlation}}\right)\right|, \tag{27}$$

where $F_{\text{correlation}}=F_{\text{vis}}\cdot\overline{F_{\text{ir}}}$ is the dot product of two Fourier transforms used to compute the correlation, F stands for the Fourier transform, $\overline{F}$ stands for the complex conjugate of Fourier, and $\mathcal{F}^{-1}$ is the inverse Fourier transform that transforms it back to the spatial domain.

$$(m,n)=\text{argmax}(\text{correlation}), \tag{28}$$

We obtain the offset between the two images by locating the maximum value in the correlation map, and then use this offset to correct the coordinates of the feature points.

After global scale correction, residual geometric differences between local patches are predominantly translational. While our current approach uses FFT-based phase correlation for local refinement, we acknowledge that in certain UAV scenarios, residual rotation or shear may occur due to parallax. Addressing such cases with rotation-aware methods (e.g., log-polar phase correlation) is left for future work.

Next we introduce a flexible transformation model to complete the transformation process. The model maps the position of each point to a new position as shown in the following equation:

$$\begin{cases} x'=a_1x+a_2y+a_3+\sum_{k=1}^{n}w_k\cdot U(r_k) \\ y'=b_1x+b_2y+b_3+\sum_{k=1}^{n}v_k\cdot U(r_k)\end{cases} \tag{29}$$

where a_1, a_2, a_3, b_1, b_2, b_3 represent the affine transformation parameters, ω_k and μ_k are the weighting coefficients of the transformation, and $U(r_k)$ denotes the radial basis function used to model the nonlinear deformation between control points, where r is the Euclidean distance from the current point (x,y) to the control point (x_k,y_k).

To determine the above transformation parameters and weights, we establish the following system of linear equations, as shown in Eq. (30):

$$\begin{bmatrix}\Phi & P+\lambda I\\ P^T & 0\end{bmatrix}\begin{bmatrix}w\\ a\end{bmatrix}=\begin{bmatrix}\Delta x\\ 0\end{bmatrix}$$
$$\begin{bmatrix}\Phi & P+\lambda I\\ P^T & 0\end{bmatrix}\begin{bmatrix}v\\ b\end{bmatrix}=\begin{bmatrix}\Delta y\\ 0\end{bmatrix} \tag{30}$$

where Φ is an $n*n$ matrix with elements $\Phi_{ij}=U\left(\sqrt{\left(x_i-x_j\right)^2+\left(y_i-y_j\right)^2}\right)$, $U(r)=r^2\log r$, P is an $n*3$ matrix with elements $P_i=\left[x_i,y_i,1\right]^T$, I is the identity matrix, λ is introduced to avoid ill-conditioned problems and to balance between fitting accuracy and transformation smoothness, and Δx and Δy are the coordinate differences of the control point pairs x and y.

After calculating these parameters, we can then calculate the position of each pixel after the transformation, as shown in Eq. (29).

Finally, to smooth the transition, we introduce a parameter η, which is calculated as follows:

$$\eta=\frac{\eta_{d1}-dist}{\eta_{\text{d1}}-\eta_{\text{d0}}}, \tag{31}$$

where $dist$ is the minimum Euclidean distance from the current pixel to the boundary of the overlapping region. This metric quantitatively defines how far a pixel is from the transition zone. η_{d1} denotes the inner boundary threshold. Pixels with $dist>\eta_{d1}$ are considered to be well inside the overlapping region. For these pixels, $\eta=0$, ensuring that the full local deformation from the flexible transformation model is applied for optimal alignment. η_{d0} represents the outer boundary threshold. Pixels with $dist<\eta_{d0}$ are considered to be outside the primary overlap. For these pixels, $\eta=1$, prioritizing global transformation consistency over local deformation. The region where $\eta_{d0}\le dist\le\eta_{d1}$ defines a smooth transition zone. Within this zone, η varies linearly from 1 to 0, ensuring a gradual and visually seamless blend between the two transformation models. The final transformation result is described in the following equation:

$$\begin{cases} u_{\text{final}}=u-\text{gx}\cdot\eta \\ v_{\text{final}}=v-\text{hy}\cdot\eta\end{cases} \tag{32}$$

where (u,v) are the preliminary coordinates warped by the flexible transformation model. The terms g_x and h_y represent the global translation offsets in the x and y directions, respectively. Thus, the terms $g_x\cdot\eta$ and $h_y\cdot\eta$ gradually pull the local non-linear warp back towards the global translation near the boundary, ensuring geometric consistency in non-overlapping areas and preventing noticeable distortions.

4. Experiments

This section details the dataset, comparison methods, evaluation metrics, parameter settings, experimental results, and ablation study analysis. To ensure experimental fairness, all traditional methods were conducted in MATLAB R2023b on a personal workstation equipped with an AMD R7-6800H processor, while deep learning-based methods were implemented and executed on an NVIDIA RTX 3090 GPU server.

4.1. Datasets

To comprehensively validate the performance advantages of the proposed method in UAV scenarios, this study utilizes the IR and VIS cross-modal, cross-resolution dataset introduced by [37]. This dataset was collected using a DJI M600Pro UAV equipped with a Zenmuse XT2 gimbal

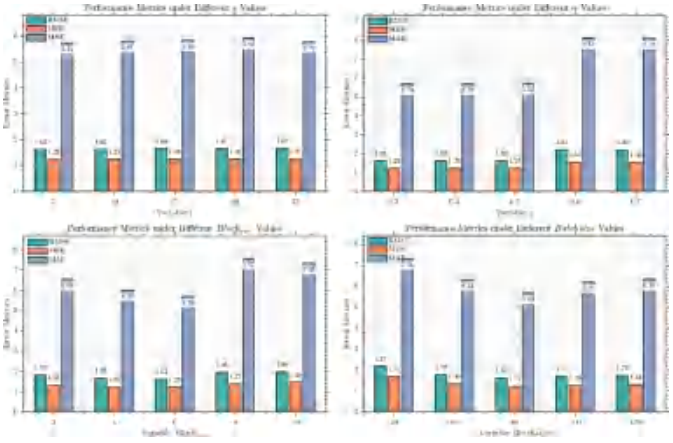


Fig. 7. Parameter sensitivity analysis. The performance (RMSE, MEE, MAE) of the proposed method with respect to different parameter settings: (a) the control radius r, (b) the weight α, (c) the block number $Block_{num}$, and (d) the $Batchsize$.

system and includes three data groups acquired at different flight altitudes. The VIS images in the first and third groups have an original resolution of 4000×3000 and are uniformly downsampled to 1600×1200 to reduce computational complexity. The second group's VIS images have a native resolution of 1920×1080. All corresponding IR images maintain a fixed resolution of 640×512.

To further assess the generalization ability of the proposed method, we also conduct comparative experiments on a multimodal dataset constructed by [38]. This dataset comprises IR and VIS images captured in 2020 and 2021, respectively. All images are normalized to a resolution of 256×256, and the IR images are augmented with random rotations within a 0-90° range. Experiments on these two datasets with distinct characteristics not only demonstrate the method's effectiveness in UAV-based multimodal scenarios but also highlight its robustness and adaptability under varying imaging conditions.

It is worth noting that due to the resolution discrepancy between visible and infrared images, the infrared images were temporarily resampled to match the visible image size for feature detection and matching. After feature correspondence estimation, all infrared feature coordinates were inversely transformed to their original scale using the provided homography matrices. Therefore, the evaluation metrics were computed on the original infrared image coordinate system, using the ground-truth points provided in the dataset. This ensures that no pseudo ground truth was introduced and that geometric consistency was preserved throughout the process.

4.2. Comparison methods and evaluation metrics

To verify the effectiveness of the proposed method, we selected nine representative approaches for comparison, namely CoFSM [39], MS-HLMO [40], CAO-C2F [41], 3MRS [42], LNIFT [43], RIFT [44], PIIFD [45],SRIF [38], D2-Net [46], LightGlue [47] and LOFTR [48]. CoFSM extracts features in a novel scale space using a Co-Occurrence Filter and accomplishes image registration through multidimensional logarithmic-polarimetric descriptors. MS-HLMO achieves matching by extracting HLMO features at Harris feature points with high intensity, rotation, and scale invariance. CAO-C2F proposes a coarse-to-fine automatic registration framework for IR and VIS images based on contour angle orientation. 3MRS is a multimodal remote sensing image matching method that first performs coarse matching through phase-consistent feature detection and Log-Gabor descriptors [49], followed by fine matching optimization via 3D correlation matching. LNIFT introduces a simple yet highly effective spatial-domain multimodal feature matching algorithm, termed Local Normalized Image Feature Transform, which applies a local normalization filter to transform the original image into a normalized form for feature detection and description, thereby significantly reducing the Nonlinear Radiometric Differences (NRD) between multimodal images. RIFT presents a robust multimodal image matching method based on PC feature detection and Maximum Index Map (MIM) description; stable feature points are detected on PC maps and rotationally invariant MIM descriptors are constructed using Log-Gabor convolutional sequences, markedly improving feature matching robustness. Finally, MS-PIIFD achieves overall registration by combining the Harris detection algorithm with the Partial Intensity Invariant Feature Descriptor (PIIFD). SRIF detects FAST keypoints and estimates their scales in a simplified pyramid space to achieve scale invariance. It then computes rotation-invariant descriptors using a Local Intensity Binary Transform (LIBT), which enhances structural consistency across modalities for robust matching. D2-Net proposes a single convolutional neural network that simultaneously detects and describes features, thus improving keypoint stability under extreme appearance changes. LightGlue introduces a deep neural matcher that learns sparse feature correspondence with adaptive network depth and width, achieving high accuracy and runtime efficiency. Finally, LOFTR develops a transformer-based framework to produce semi-dense matches without explicit keypoint detection, enabling robust matching in low-texture or large-viewpoint scenarios.

To ensure the accuracy of the registration results, this paper adopts three evaluation metrics: RMSE, mean absolute error (MAE), and maximum estimated error (MEE). Specifically, RMSE reflects the overall deviation, MAE measures the mean displacement, and MEE evaluates the maxmimum local misalignment. These three complementary indicators collectively verify the robustness of the registration method from different perspectives, and the formula is described in the following equation:

$$\begin{cases} \mathrm{RMSE} = \sqrt{\dfrac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \\ \mathrm{MEE} = \max\limits_{i=1}^{N} |y_i - \hat{y}_i| \\ \mathrm{MAE} = \dfrac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \end{cases} \tag{33}$$

where N represents the number of detected points, y_i and $\hat{y}_i$ represent the coordinates of the reference points and the corrected points, respectively, and $\max(\cdot)$ denotes the maximization operation. For the computation of RMSE, MAE, and MEE, the ground-truth correspondences are obtained differently for the two datasets. Specifically, in the dataset [37], the ground-truth matching points are provided as manually annotated tie points in the dataset. In contrast, for the dataset [38], the ground truth is derived from the provided homography matrices, and all transformed coordinates are calculated accordingly. These ground-truth correspondences are used as the reference to compute point-wise errors between estimated and true matching points.

It is worth noting that due to the resolution discrepancy between IR and VIS images, the IR images were temporarily resampled to match the VIS image size for feature detection and matching. After feature correspondence estimation, all IR feature coordinates were inversely transformed to their original scale using the provided homography matrices. Therefore, the evaluation metrics were computed on the original IR image coordinate system, using the ground-truth points provided in the dataset. This ensures that no pseudo ground truth was introduced and that geometric consistency was preserved throughout the process.

4.3. Parameter setting

In the experimental process, in order to ensure that the configuration of each parameter reaches the optimal effect, this study uses the univariate control method for parameter optimization. Next we sat control variable experiments on each of the four parameters, the control radius r in ECPCS, the weighting coefficient α between RMSE and MI,

Table 1
Quantitative results of various methods on [37].

Method	RMSE	MEE	MAE
RIFT [44]	598.00 ± 59.55	597.13 ± 57.12	654.52 ± 109.63
CoFSM [39]	14.50 ± 4.07	13.70 ± 3.84	23.25 ± 3.42
LNIFT [43]	7818.01 ± 687.54	7356.63 ± 668.17	12617.60 ± 1037.28
3MRS [42]	2.29 ± 5.50	2.18 ± 5.53	3.46 ± 5.14
MS-PIIFD [45]	599.57 ± 659.65	552.02 ± 616.35	990.25 ± 1119.00
MS-HLMO [40]	443.78 ± 414.26	404.15 ± 379.59	778.67 ± 776.86
CAO-C2F [41]	490.60 ± 356.88	452.56 ± 338.64	847.36 ± 591.14
SRIF [38]	8.15 ± 2.34	7.23 ± 2.01	15.89 ± 4.67
D2-Net [46]	1003.12 ± 27.74	992.54 ± 28.32	1250.70 ± 39.07
LightGlue [47]	7.20 ± 1.05	5.74 ± 0.68	18.30 ± 3.70
LOFTR [48]	6.85 ± 0.92	5.41 ± 0.75	16.73 ± 3.25
Ours	**1.63± 0.27**	**1.23 ± 0.18**	**5.70 ± 1.85**

Table 2
Quantitative results of various methods on [38].

Method	RMSE	MEE	MAE
RIFT [44]	3.71 ± 0.11	3.60 ± 0.13	4.49 ± 0.14
CoFSM [39]	40.57 ± 11.19	40.30 ± 11.37	45.80 ± 11.48
LNIFT [43]	16.74 ± 7.08	16.72 ± 7.12	16.88 ± 6.78
3MRS [42]	2.29 ± 5.50	2.18 ± 5.53	3.46 ± 5.14
MS-PIIFD [45]	50.59 ± 60.69	44.27 ± 54.30	86.39 ± 94.09
MS-HLMO [40]	7.83 ± 6.26	7.09 ± 5.98	14.84 ± 9.68
CAO-C2F [41]	59.33 ± 65.84	49.62 ± 59.18	92.05 ± 93.31
SRIF [38]	2.08 ± 0.06	1.95 ± 0.08	2.99 ± 0.01
D2-Net [46]	105.00 ± 9.97	80.09 ± 12.81	267.90 ± 26.62
LightGlue [47]	54.06 ± 18.84	42.88 ± 18.19	137.55 ± 36.79
LOFTR [48]	2.00 ± 0.23	1.88 ± 0.37	**2.76 ± 0.28**
Ours	**1.67 ± 0.51**	**1.53± 0.50**	2.76 ± 0.42

the number of blocks in block detection $Block_{num}$, and the block size $batchsize$ in fine registration. In Fig. 7, the experimental results indicate that each parameter influences the algorithm's performance by regulating characteristics across different dimensions. Specifically, the control radius r, as a key parameter for contrast enhancement, achieves the optimal image enhancement effect when $r = 5$. The weight coefficient α, responsible for balancing the trade-off between RMSE and MI, reaches the optimal equilibrium at $\alpha = 0.4$. In terms of feature processing, the 6×6 block strategy significantly improves the efficiency of feature point detection, while setting the batch size to 80 optimizes the computational efficiency of regional matching. The synergistic effect of these parameters ensures the optimal overall performance of the algorithm across all aspects of image processing. In the scale selection process, the scale factor $Scale$ is searched within the range [1.0, 1.5] using a step size of 0.01 through a discrete grid search strategy rather than continuous optimization. At each iteration, the moving image is geometrically warped with bilinear interpolation to ensure alignment consistency. Although no explicit anti-aliasing filter is applied, the interpolation introduces moderate smoothing that effectively suppresses aliasing artifacts. A smaller step size improves the precision of scale estimation but increases computational cost.

4.4. Experiment results and analysis

This section focuses on four perspectives: visualization results, quantitative results, ablation experiments, and time.

4.4.1. Visualization results and analysis

As shown in Fig. 8, the visual analysis of the matching and registration results reveals that although the three methods, RIFT, LNIFT, SRIF and D2-Net are able to detect a large number of feature points, there is a significant error in their matching process. Similarly, although the LoFTR method detects a comparable number of keypoints, its correspondence accuracy is limited, leading to noticeable geometric misalignments in the final registration results. This lack of matching accuracy directly leads to obvious geometrical variations in the final registration results. In contrast to the first two methods, MS-PIIFD, MS-HLMO, and CAO-C2F detect too few feature points, resulting in poor registration accuracy. Although the number of feature points detected by the CoFSM method is relatively small, it achieves higher registration accuracy compared to RIFT and LNIFT, albeit with some residual errors. In contrast, 3MRS, LightGlue and the proposed method detect a greater number of feature points and achieve superior registration performance with more accurate visual results. However, the 3MRS and LightGlue methods exhibit slight blurring artifacts around the car edges in the example image, as shown in Fig. 9, whereas our method effectively mitigates such artifacts while maintaining high registration accuracy.

In Fig. 10, the RIFT and LNIFT algorithms exhibit high sensitivity during the feature point detection stage, enabling the extraction of a large number of feature points. However, they suffer from notable stability issues during the feature matching and registration processes. In contrast, CoFSM, MS-PIIFD, and CAO-C2F demonstrate higher computational efficiency but detect relatively fewer feature points, leading to considerable geometric distortions in the final registration results, as shown in Fig. 9. The performance of MS-HLMO shows noticeable improvement on the current dataset compared to the previous one, with an increased number of detected feature points and enhanced registration accuracy; nonetheless, some degree of matching uncertainty still persists. Similarly, the SRIF method detects a dense distribution of feature points and achieves reasonably good alignment results. However, minor geometric deviations remain near object boundaries, indicating limited robustness under local illumination variations. D2-Net and LightGlue also detect a relatively large number of feature points. However, they suffer from poor matching precision, resulting in significant distortions in the final registration outputs. In particular, although LightGlue performed well on the previous dataset, its effectiveness degrades noticeably on the current dataset, indicating a lack of robustness across different imaging conditions.

Table 3
Feature point statistics and matching accuracy on [37].

Method	Point Number	Accuracy(%)
RIFT [44]	**10582**	8.99
CoFSM [39]	42	83.3
LNIFT [43]	5000	7.38
3MRS [42]	156	87.8
MS-PIIFD [45]	8	25
MS-HLMO [40]	9	55.6
CAO-C2F [41]	6	66.7
SRIF [38]	2758	10.9
D2-Net [46]	3046	10.9
LightGLue [47]	363	**100**
LOFTR [48]	121	60.4
Ours	620	**100**

LoFTR, 3MRS, and the proposed method represent the top-performing approaches in our comparison, all producing visually accurate and geometrically consistent registration results under complex UAV imaging conditions. LoFTR demonstrates strong learning-based correspondence capabilities and yields precise alignment across multimodal scenes. Similarly, 3MRS achieves robust geometric consistency but introduces slight blurring artifacts along object boundaries, as shown in Fig. 9. In contrast, the proposed method not only maintains high geometric accuracy comparable to LoFTR and 3MRS, but also effectively preserves edge sharpness and structural integrity in the final registration outputs.

In UAV scenarios, the large viewpoint changes and sensor discrepancies cause conventional algorithms to rely heavily on intensity or gradient information, resulting in severe misalignments. In contrast, the proposed ECPCS-based method enhances modality-invariant structural features derived from phase congruency, which remain stable despite spectral and radiometric differences. Consequently, ECPCS provides more

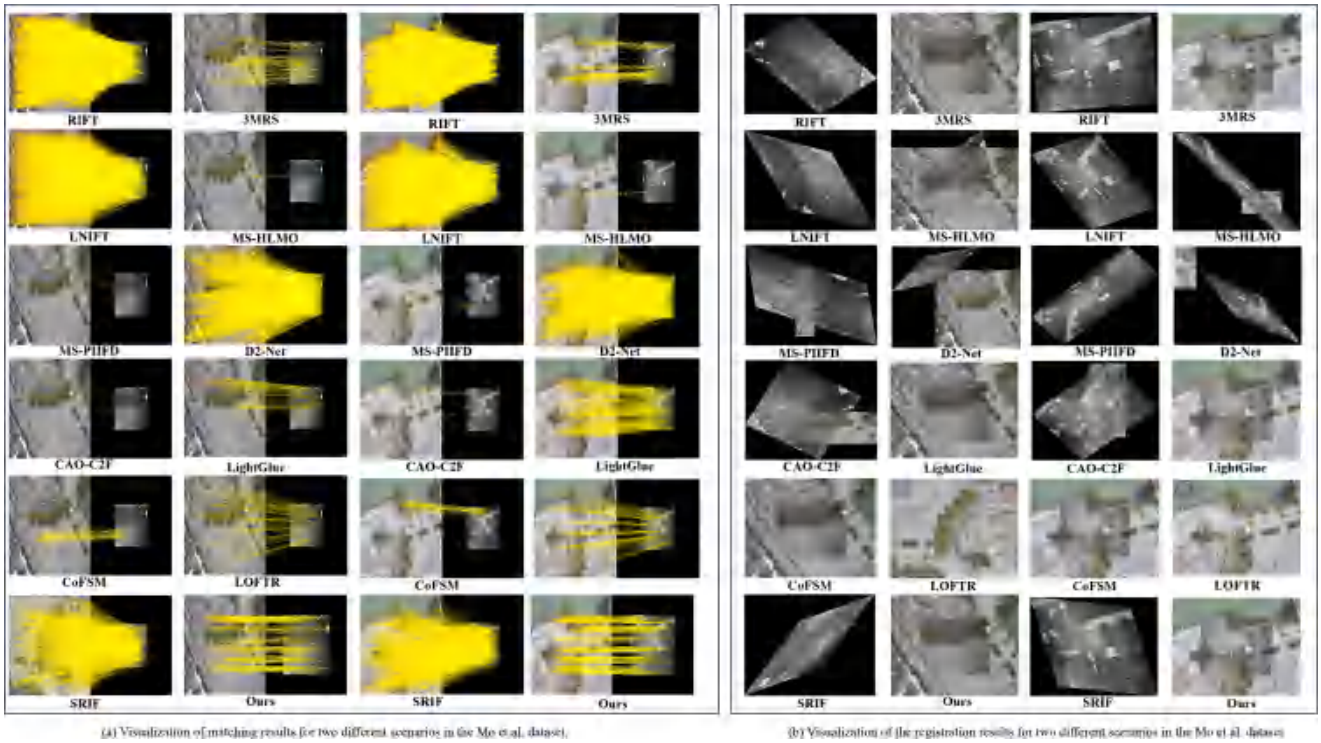


Fig. 8. Visualization results of different methods for matching as well as registration in the dataset of [37].

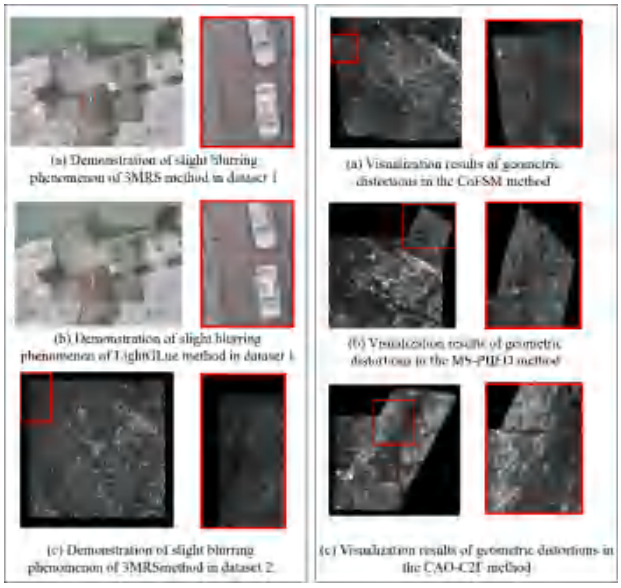


Fig. 9. Visualization results of slight blurring and geometric distortions.

consistent feature correspondences and significantly reduces misalignment errors in challenging UAV conditions, leading to clearer structural alignment and improved robustness across datasets.

4.4.2. Quantitative results and analysis

To ensure clarity, in all quantitative tables, the best performance is highlighted in **bold**, and the second-best result is underlined.

Tables 1 and 2 present the quantitative results on the datasets of Mo et al. and Li et al., respectively. In this study, all RMSE, MAE, and MEE values are fully reported to ensure transparent and comprehensive comparison across different methods. To further enhance the statistical interpretability of the results, the standard deviation of each metric is also provided, reflecting the variability and reliability of registration performance under different UAV imaging conditions.

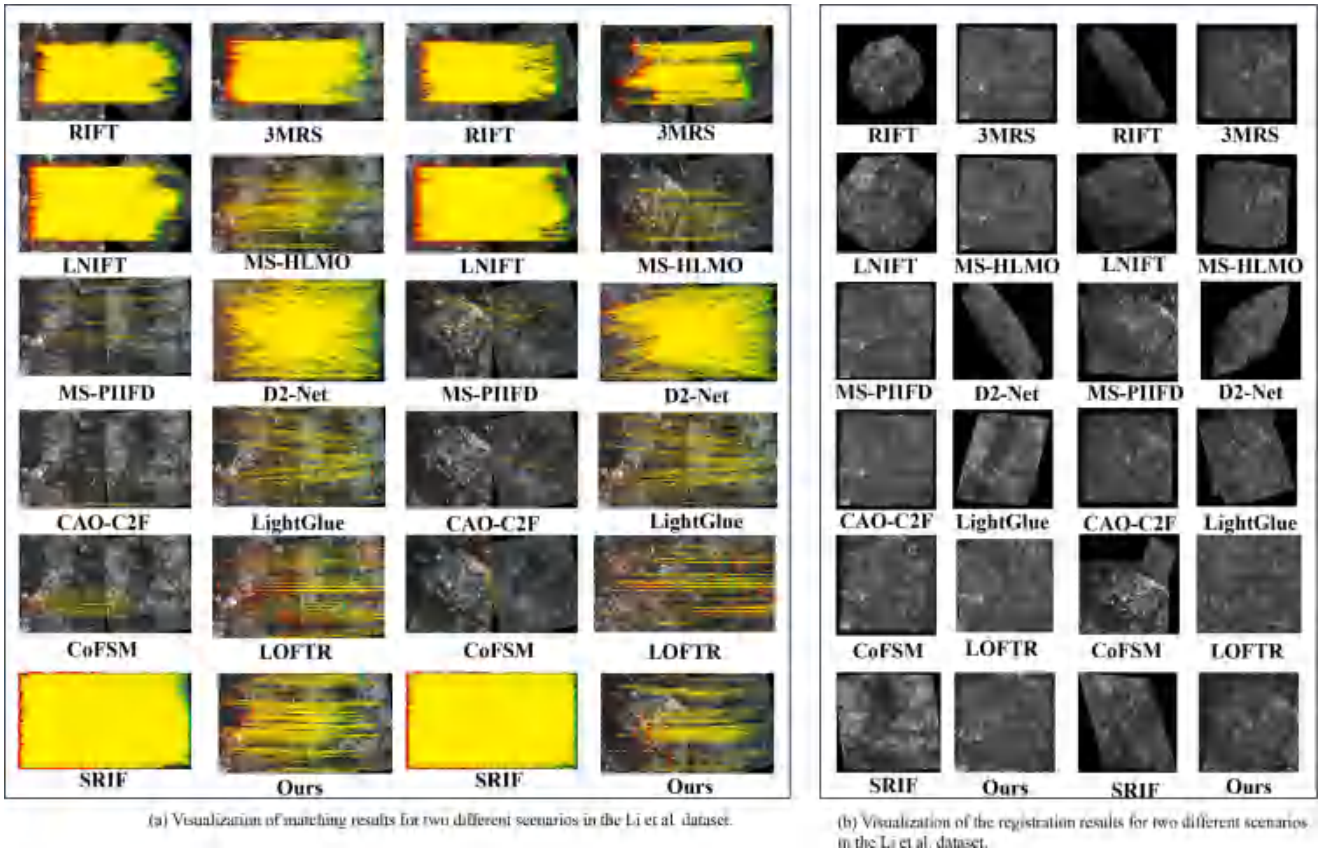
As shown in Table 1, our method effectively addresses the challenges of varying resolutions, parallax differences, and modality discrepancies in UAV scenarios using the proposed dataset. Compared to other methods, our approach achieves significantly superior results across all evaluated aspects. Among them, although the CoFSM, 3MRS, LightGlue and LOFTR methods demonstrate certain effectiveness, they still exhibit relatively large registration errors. The remaining algorithms suffer from

Table 4
Feature point statistics and matching accuracy on [38].

Method	Point Number	Accuracy(%)
RIFT [44]	3406	22.5
CoFSM [39]	3	**100**
LNIFT [43]	**5000**	30.8
3MRS [42]	156	92.5
MS-PIIFD [45]	26	80.9
MS-HLMO [40]	107	98.1
CAO-C2F [41]	6	83.3
SRIF [38]	2231	27.9
D2-Net [46]	741	9.5
LightGLue [47]	73	71.1
LOFTR [48]	341	99.7
Ours	217	**100**

Table 5
Ablation experiment results with different feature point detection algorithms on [37].

Detector	RMSE	MEE	MAE
FAST	1.89	1.43	7.08
SURF	1.82	1.34	6.73
SIFT	1.94	1.45	7.05
Our	**1.63**	**1.23**	**5.70**



(a) Visualization of matching results for two different scenarios in the Li et al. dataset.

(b) Visualization of the registration results for two different scenarios in the Li et al. dataset.

Fig. 10. Visualization results of different methods for matching as well as registration in the dataset of [38].

Table 6
Ablation experiment results with different feature point detection algorithms on [38].

Detector	RMSE	MEE	MAE
FAST	1.71	1.58	2.84
SURF	10.66	10.59	11.17
SIFT	1.68	1.55	2.86
Our	**1.67**	**1.53**	**2.76**

Table 7
Ablation experiment results with different enhancement algorithms on [37].

Enhancement	RMSE	MEE	MAE
PC	3.35	2.32	12.29
Our	**1.63**	**1.23**	**5.70**

Table 8
Ablation experiment results with different enhancement algorithms on [38].

Enhancement	RMSE	MEE	MAE
PC	1.67	1.54	2.77
Our	**1.67**	**1.53**	**2.76**

Table 9
Quantitative evaluation of fusion quality before and after registration on [37].

Fusion Method	Metric	Before Registration	After Registration	Improvement(%)
ADF [50]	MI	2.49	**2.51**	0.8
	PSNR	61.32	**61.35**	0.1
	Qcv	278.33	**267.91**	3.9
CBF [51]	MI	2.99	**3.03**	1.3
	PSNR	60.50	**60.56**	0.1
	Qcv	420.89	**399.47**	5.4
GAN [52]	MI	1.07	**1.09**	1.9
	PSNR	60.83	**60.85**	0.0
	Qcv	763.92	**714.70**	6.9
GTF [53]	MI	1.75	**1.85**	5.7
	PSNR	61.12	**61.16**	0.1
	Qcv	681.71	**650.43**	4.7

significant limitations when applied to UAV scenarios, resulting in substantial misalignments in their outputs.

As shown in Table 2, some algorithms demonstrate notable improvements compared to their performance on the first dataset. Among them, RIFT and 3MRS achieve relatively good results, yet still fall short when compared to our proposed method. The SRIF method also shows competitive performance, achieving slightly lower average accuracy than ours but maintaining remarkably low standard deviations across all metrics, indicating strong result stability. While LNIFT and MS-HLMO exhibit a certain degree of effectiveness, their lack of robustness leads to considerable registration errors. In contrast, the performance of CoFSM, MS-PIIFD, and CAO-C2F is notably limited due to challenges associated with their applicability in the given scenario, resulting in suboptimal outcomes. Additionally, although D2-Net and LightGlue extract a large number of keypoints, their matching precision is poor on this dataset, resulting in significant registration errors across all three metrics. Notably, LightGlue, which previously performed well on another dataset, demonstrates a marked performance drop here, suggesting limited adaptability to diverse data conditions. In contrast, the LoFTR algorithm achieves results on par with our method in terms of MAE, while exhibiting even

Table 10
Quantitative comparison of fusion results obtained from different registration methods on the UAV IR-VIS dataset.

Method	Qcv	PSNR	MI
RIFT [44]	941.85	60.18	1.71
CoFSM [39]	651.40	61.11	1.76
LNIFT [43]	974.73	60.70	1.43
3MRS [42]	665.5	61.12	1.78
MS-PIIFD [45]	979.83	61.23	1.37
MS-HLMO [40]	1035.53	60.29	1.85
CAO-C2F [41]	919.76	61.13	1.61
SRIF [38]	965.39	60.44	1.12
D2-Net [46]	827.97	60.44	1.12
LightGlue [47]	653.56	61.14	1.80
LOFTR [48]	859.46	60.65	0.35
Ours	**650.43**	**61.16**	**1.85**

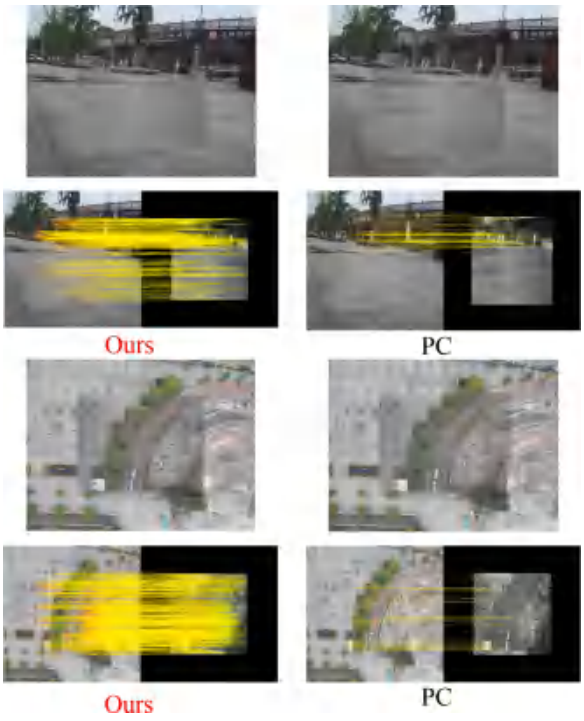


Fig. 11. ECPCS ablation experiment visualization results.

lower standard deviations, reflecting its high stability and reliable geometric consistency under complex UAV imaging conditions.

To further evaluate the matching reliability of different registration methods, we statistically analyzed the number of detected keypoints and the corresponding matching accuracy on both datasets. Tables 3 and 4 summarize the results. It can be observed that our method achieves the highest proportion of correct correspondences while maintaining a moderate number of detected features, indicating a better balance between keypoint density and matching precision. This confirms that the proposed feature extraction and matching strategy yields more robust and geometrically consistent correspondences.

Overall, the proposed method achieves the best performance across all metrics, demonstrating both high accuracy and robustness in cross-modal image registration. Comprehensive evaluations on two datasets with distinct scenarios demonstrate that our method consistently outperforms existing approaches in terms of both accuracy and stability.

4.4.3. Ablation experiment

We conduct ablation experiments on two distinct datasets to validate the effectiveness and necessity of our module design choices. Specifically, we evaluate the impact of the feature point detection algorithm and the proposed ECPCS module. As shown in Tables 5 and 6, among several widely used feature point detection algorithms, the Harris detector achieves the best performance, confirming its suitability for our method.

Additionally, the results in Tables 7 and 8 demonstrate that the inclusion of the ECPCS module significantly enhances the registration accuracy, thereby substantiating its critical role in our framework. Compared with the conventional PC, ECPCS introduces a local adaptive contrast enhancement on the maps, which effectively amplifies weak structural responses and preserves edge saliency under varying illumination. This enhancement improves the distinctiveness of feature representations, leading to more reliable feature matching and higher registration precision. As illustrated in Fig. 11, ECPCS notably enhances both the accuracy and robustness of feature point detection and matching, especially in low-texture or low-contrast regions.

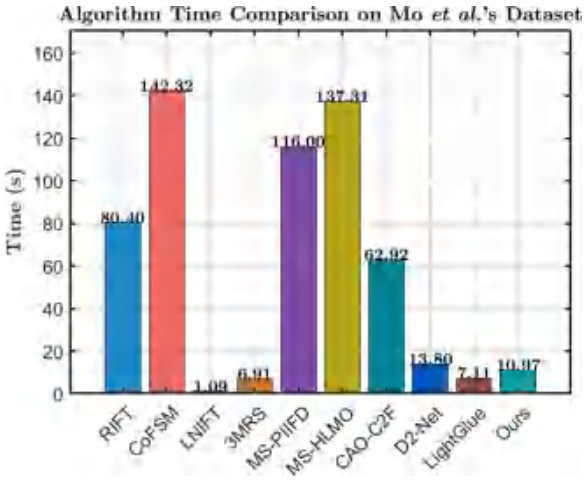


Fig. 12. Comparison of run times of different methods on the dataset of [37].

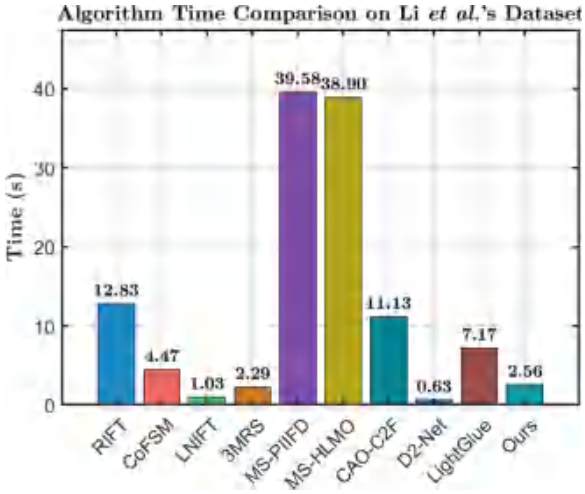


Fig. 13. Comparison of run times of different methods on the dataset of [38].

4.4.4. Time analysis

To evaluate the runtime of various algorithms, we record the average execution time of each method on two datasets separately. Figs. 12 and 13 show the average running times of our proposed method on the two datasets, respectively. As described previously, the two datasets have significantly different image resolutions.

As shown in Fig. 12, the average running time of the proposed method lies in the mid-range among all compared methods and is sig-

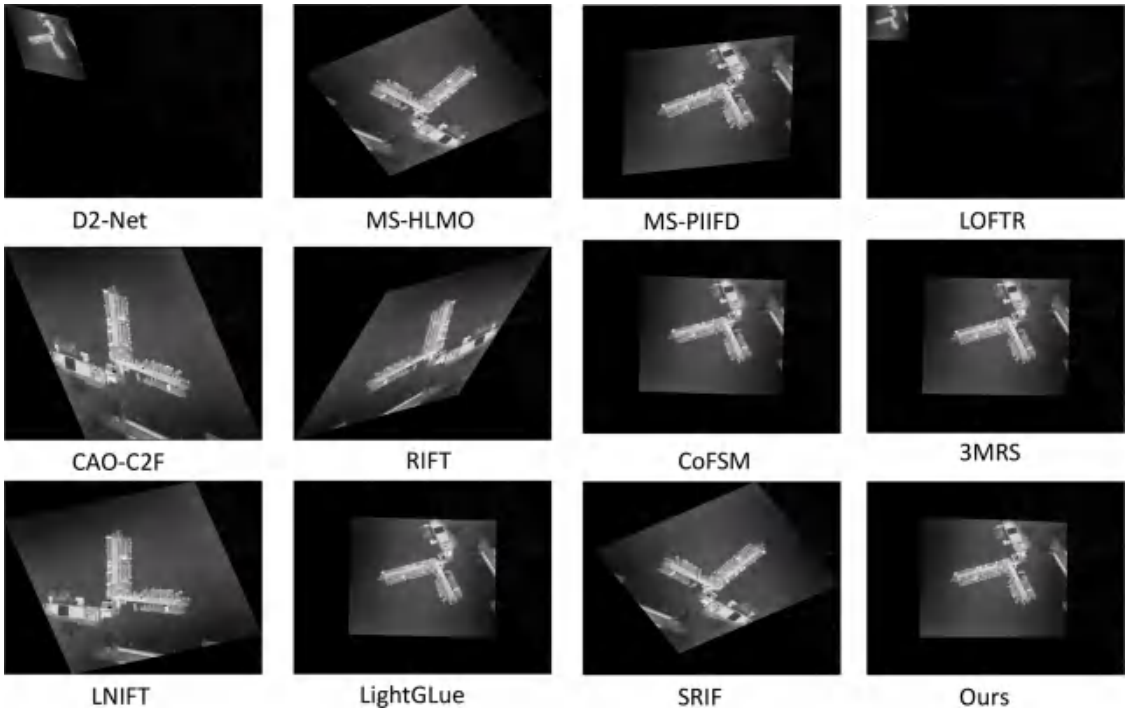


Fig. 14. Visual comparison of IR and VIS fusion results under different registration methods.

nificantly faster than those of RIFT, CoFSM, MS-PIIFD, and MS-HLMO. The experimental results in Fig. 13 further validate this trend. Compared to the previous dataset, the average running time of the proposed method is very close to that of 3MRS but still lags behind LNIFT and D2-Net. This difference mainly stems from the fact that LNIFT employs a locally normalized filter with a time complexity of O(1), and D2-Net benefits from a fully convolutional architecture that enables efficient, parallelized computation on GPUs, making its runtime largely insensitive to image resolution.

4.5. Downstream validation

To further establish an application-level closed loop between registration and downstream tasks, a fusion-based evaluation was conducted on the UAV infrared-visible dataset. The goal of this experiment is to quantitatively assess how accurate geometric alignment improves the quality of subsequent image fusion results.

Four representative IR and VIS fusion methods were selected for evaluation, covering both traditional and learning-based approaches: Anisotropic Diffusion and Karhunen-Loeve Transform Filter (ADF) [50], Cross Bilateral Filter (CBF) [51], a Generative Adversarial Network-based method (GAN) [52], and the Gradient Transfer and Total Variation minimization method (GTF) [53]. Each method was applied to image pairs before and after registration, while all other parameters remained identical to ensure fairness.

The fusion quality was evaluated using Mutual Information (MI), Peak Signal-to-Noise Ratio (PSNR), and the human visual system-based metric Qcv [54]. The quantitative results are summarized in Table 9.

As shown in Table 9, all fusion algorithms achieve higher MI and PSNR values after registration, indicating enhanced structural consistency and reduced misalignment artifacts between modalities. The improvements in Qcv also demonstrate that the registered images yield better global contrast and detail preservation in the fused results. Among the tested methods, GTF and GAN-based fusion show the most significant gains, suggesting that accurate alignment particularly benefits feature-level and learning-based fusion schemes that rely on fine spatial correspondence.

In addition, to further validate the robustness of the proposed registration algorithm, a complementary experiment was conducted using the same fusion method (GTF) while varying the registration approaches. The visual comparison shown in Fig. 14 illustrates that the fused images based on our registration pipeline preserve sharper edges and clearer structural boundaries than those derived from other registration methods. The quantitative comparison, as summarized in Table 10, shows that the fused results derived from our registration pipeline consistently outperform those based on other registration methods across all three fusion quality metrics. This demonstrates that precise geometric alignment achieved by the proposed algorithm contributes directly to the perceptual quality and structural fidelity of the fused image, establishing a more complete registration-fusion closed loop.

Overall, the experimental results verify that precise registration not only improves geometric accuracy but also enhances the perceptual and structural quality of fused infrared-visible images, thereby validating the effectiveness of the proposed registration method in real UAV application scenarios.

5. Conclusions

In this paper, we propose an image registration algorithm for IR and VIS light images. The method first performs a uniform transformation of the source image into an enhanced phase-consistent space using the ECPCS technique. Based on this space, the optimal transformation angle of the IR image is determined, and accurate feature points are extracted. Finally, a flexible transformation model is employed to achieve the final registration results. Extensive experiments conducted on two datasets across multiple scenes demonstrate the effectiveness and robustness of the proposed method.

While this work is primarily focused on aerial remote sensing scenarios, the proposed ECPCS transformation, which excels in enhancing structural edges and suppressing modality-specific noise, may offer a promising algorithmic insight for other multimodal imaging challenges. For instance, the fundamental principle of extracting modality-invariant features could be potentially explored in other fields, such as medical

image analysis [55], for tasks like registering functional photoacoustic images with anatomical scans.

In future work, we plan to extend the registration framework toward closed-loop perception tasks, such as IR and VIS fusion-based object detection and semantic segmentation, to further validate its benefits for UAV vision systems. Additionally, while residual local transformations in the current work are predominantly translational, scenarios with larger parallax may introduce local rotation or shear. Investigating rotation-aware phase correlation or other techniques to handle such residuals represents a promising direction for future improvement.

Nevertheless, the primary application and immediate value of our work remain in advancing the capabilities of UAV vision systems for applications like precision agriculture, infrastructure inspection, and surveillance.

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.

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.

Declaration of competing interest

The authors declare that they have no potential or explicit financial interests in this paper and have no objections to the order of authorship.

Acknowledgement

This work is supported by the Research Project of Beijing Municipal Natural Science Foundation (No. BJXZ2021-012-00046).

References

[1] Liu C, Li H, Li M, Deng L, Dong M, Zhu L. Multi-stage non-uniformity correction pipeline for single-frame infrared images based on hybrid high-order directional and low-rank prior information. IEEE Sens J 2025.
[2] Liu C, Chen H, Deng L, Guo C, Lu X, Yu H, et al. Modality specific infrared and visible image fusion based on multi-scale rich feature representation under low-light environment. Infrared Phys Technol 2024;140:105351.
[3] Li H, Wu S, Deng L, Liu C, Chen Y, Chen H, et al. Enhancing infrared and visible image fusion through multiscale gaussian total variation and adaptive local entropy. Vis Comput 2025;41:7817–38.
[4] Chen H, Deng L, Chen Z, Liu C, Zhu L, Dong M, et al. Sfcfusion: spatial-frequency collaborative infrared and visible image fusion. IEEE Trans Instrum Meas 2024;73:1–15. https://doi.org/10.1109/TIM.2024.3370752
[5] Wu S, Li H, Deng L, Yu H, Chen H, Chen Z, et al. Foggyfuse: infrared and visible image fusion method based on saturation line prior in foggy conditions. Opt Laser Technol 2025;190:113075.
[6] Zheng Y, Li Y, Yang S, Lu H. Global-PBNet: a novel point cloud registration for autonomous driving. IEEE Trans Intell Transp Syst 2022;23(11):22312–19.
[7] Maes F, Vandermeulen D, Suetens P. Medical image registration using mutual information. Proc IEEE 2003;91(10):1699–722.
[8] Chen J, Frey EC, He Y, Segars WP, Li Y, Du Y. Transmorph: transformer for unsupervised medical image registration. Med Image Anal 2022;82:102615.
[9] Jiang X, Ma J, Xiao G, Shao Z, Guo X. A review of multimodal image matching: methods and applications. Inf Fusion 2021;73:22–71.
[10] Santarossa M, Koch R, Tavakoli M. Robust multimodal retinal image registration in diabetic retinopathy using a light-weight neural network and improved RANSAC algorithm. IEEE Sens J 2025;25:13469–79.
[11] Zhou H, Ma J, Tan CC, Zhang Y, Ling H. Cross-weather image alignment via latent generative model with intensity consistency. IEEE Trans Image Process 2020;29:5216–28.
[12] Ding J, Zhao Y, Pei L, Shan Y, Du Y, Li W. Modal-invariant progressive representation for multimodal image registration. Inf Fusion 2025;117:102903.
[13] Lian Z, Gu Y, You K, Xie X, Guo M, Gu Y, et al. An adaptive point cloud registration method with self-cross attention and hierarchical correspondence filtering. Opt Laser Technol 2025;184:112384.
[14] Zeng L, Du Y, Lin H, Wang J, Yin J, Yang J. A novel region-based image registration method for multisource remote sensing images via CNN. IEEE J Sel Top Appl Earth Obs Remote Sens 2020;14:1821–31.
[15] Okorie A, Makrogiannis S. Region-based image registration for remote sensing imagery. Comput Vision Image Understanding 2019;189:102825.
[16] Meng L, Zhou J, Liu S, Wang Z, Zhang X, Ding L, et al. A robust registration method for UAV thermal infrared and visible images taken by dual-cameras. ISPRS J Photogramm Remote Sens 2022;192:189–214.
[17] Khan MA, Menouar H, Eldeeb A, Abu-Dayya A, Salim FD. On the detection of unauthorized drones-techniques and future perspectives: a review. IEEE Sens J 2022;22(12):11439–55.
[18] Cui J, Liu M, Zhang Z, Yang S, Ning J. Robust UAV thermal infrared remote sensing images stitching via overlap-prior-based global similarity prior model. IEEE J Sel Top Appl Earth Obs Remote Sens 2020;14:270–82.
[19] Ye Y, Zhu B, Tang T, Yang C, Xu Q, Zhang G. A robust multimodal remote sensing image registration method and system using steerable filters with first-and second-order gradients. ISPRS J Photogramm Remote Sens 2022;188:331–50.
[20] Wu K. Creating panoramic images using ORB feature detection and RANSAC-based image alignment. Adv Comput Commun 2023;4(4):220–4.
[21] Xu M, Ma H, Zhong X, Zhao Q, Chen S, Zhong R. Fast and accurate registration of large scene vehicle-borne laser point clouds based on road marking information. Opt Laser Technol 2023;159:108950.
[22] Li H, Liu C, Li M, Deng L, Dong M, Zhu L. Cross-scale infrared and visible image registration based on phase consistency feature for UAV scenario. Measurement 2026;258:119340. https://doi.org/10.1016/j.measurement.2025.119340
[23] Kovesi P, et al. Image features from phase congruency. Videre J Comput Vis Res 1999;1(3):1–26.
[24] Ye Y, Bruzzone L, Shan J, Bovolo F, Zhu Q. Fast and robust matching for multimodal remote sensing image registration. IEEE Trans Geosci Remote Sens 2019;57(11):9059–70.
[25] Brown LG. A survey of image registration techniques. ACM Comput Surv 1992;24(4):325–76.
[26] Moravec HP. Obstacle avoidance and navigation in the real world by a seeing robot rover. Stanford University; 1980.
[27] Harris C, Stephens M. A combined corner and edge detector. In: Proceedings of the Alvey vision conference. 1988, pp. 147–51.
[28] Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2004;60:91–110.
[29] Bay H, Tuytelaars T, Van Gool L. 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. 404–17.
[30] Rosten E, Drummond T. 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. 430–43.
[31] Zhang W. Robust registration of SAR and optical images based on deep learning and improved harris algorithm. Sci Rep 2022;12(1):5901.
[32] Cong Y. Image stitching technology for police drones using an improved image registration method incorporating ORB algorithm. Informatica 2024;48(2):269–82.
[33] Fischler MA, Bolles RC. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 1981;24(6):381–95.
[34] Bookstein FL. Principal warps: thin-plate splines and the decomposition of deformations. IEEE Trans Pattern Anal Mach Intell 1989;11(6):567–85.
[35] Rueckert D, Sonoda LI, Hayes C, Hill D LG, Leach MO, Hawkes DJ. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans Med Imaging 1999;18(8):712–21.
[36] Reddy BS, Chatterji BN. An FFT-based technique for translation, rotation, and scale-invariant image registration. IEEE Trans Image Process 1996;5(8):1266–71.
[37] Mo Y, Kang X, Zhang S, Duan P, Li S. A robust infrared and visible image registration method for dual-sensor UAV system 61:1–13. https://doi.org/10.1109/TGRS.2023.3306558
[38] Li J, Hu Q, Zhang Y. Multimodal image matching: a scale-invariant algorithm and an open dataset. ISPRS J Photogramm Remote Sens 2023;204:77–88.
[39] Yao Y, Zhang Y, Wan Y, Liu X, Yan X, Li J. Multi-modal remote sensing image matching considering co-occurrence filter 31:2584–97. https://doi.org/10.1109/TIP.2022.3157450
[40] Gao C, Li W, Tao R, Du Q. MS-HLMO: multiscale histogram of local main orientation for remote sensing image registration 60:1–14. . https://doi.org/10.1109/TGRS.2022.3193109
[41] Jiang Q, Liu Y, Yan Y, Deng J, Fang J, Li Z, et al. A contour angle orientation for power equipment infrared and visible image registration. IEEE Trans Power Delivery 2020;36(4):2559–69.
[42] Fan Z, Liu Y, Liu Y, Zhang L, Zhang J, Sun Y, et al. 3MRS: an effective coarse-to-fine matching method for multimodal remote sensing imagery. Remote Sens 2022;14(3):478.
[43] Li J, Xu W, Shi P, Zhang Y, Hu Q. Lnift: locally normalized image for rotation invariant multimodal feature matching. IEEE Trans Geosci Remote Sens 2022;60:1–14.

[44] Li J, Hu Q, Ai M. Rift: multi-modal image matching based on radiation-variation insensitive feature transform. IEEE Trans Image Process 2019;29:3296–310.
[45] Gao C, Li W. Multi-scale PIIFD for registration of multi-source remote sensing images.2021 arXiv preprint arXiv:210412572.
[46] Dusmanu M, Rocco I, Pajdla T, Pollefeys M, Sivic J, Torii A, et al. 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. 8092–101.
[47] Lindenberger P, Sarlin P-E, Pollefeys M. Lightglue: local feature matching at light speed. In: Proceedings of the IEEE/CVF international conference on computer vision. 2023, pp. 17627–38.
[48] Sun J, Shen Z, Wang Y, Bao H, Zhou X. LoFTR: detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, pp. 8922–31.
[49] Field DJ. Relations between the statistics of natural images and the response properties of cortical cells. J Opt Soc Am A 1987;4(12):2379–94.
[50] Bavirisetti DP, Dhuli R. Fusion of infrared and visible sensor images based on anisotropic diffusion and Karhunen-Loeve transform. IEEE Sens J 2015;16(1):203–9.
[51] Shreyamsha Kumar BK. Image fusion based on pixel significance using cross bilateral filter. Signal Image Video Process 2015;9(5):1193–204.
[52] Ma J, Yu W, Liang P, Li C, Jiang J. FusionGAN: a generative adversarial network for infrared and visible image fusion. Inf fusion 2019;48:11–26.
[53] Ma J, Chen C, Li C, Huang J. Infrared and visible image fusion via gradient transfer and total variation minimization. Inf Fusion 2016;31:100–9.
[54] Chen H, Varshney PK. A human perception inspired quality metric for image fusion based on regional information. Inf fusion 2007;8(2):193–207.
[55] Kratkiewicz K, Manwar R, Zafar M, Mohsen Ranjbaran S, Mozaffarzadeh M, de Jong N, et al. Development of a stationary 3D photoacoustic imaging system using sparse single-element transducers: Phantom study. Appl Sci 2019;9(21). https://doi.org/10.3390/app9214505