Shape2Match: A Shape-to-Matching Framework for Infrared-Visible Image Matching

Maoyu Wang[1,*], Xulei Shi[1,*], Zhuolu Hou[1], Xinbo Zhao[1], Xin Huang[1], Yifan Liao[1], Yansong Duan[1,†], Pengjie Tao[1]

[1] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China - ysduan@whu.edu.cn

Keywords: Infrared-visible image matching, Shape representation, Elliptic fourier descriptor, EfficientSAM

Abstract

Traditional image matching methods rely heavily on gradient or intensity information. However, the severe nonlinear radiometric distortion (NRD) between infrared and visible images hinders the extraction of repeatable feature points, leading to poor matching performance. To address this, we propose Shape2Match, a novel framework that replaces point features with more consistent, modality-invariant shape features. Specifically, the method utilizes EfficientSAM to extract shape contours and employs elliptic fourier descriptors (EFD) to parameterize and normalize them, creating shape descriptor that is invariant to translation, rotation, and scale. Shape2Match adopts a coarse-to-fine hierarchical strategy: it first performs robust global shape matching using a weighted EFD distance, followed by precise keypoint matching—using Shape Context—within the coarsely aligned shape pairs. We validated Shape2Match on 153 image pairs from 6 datasets, comparing it against methods like SIFT, RIFT, and MS-HLMO. Experimental results demonstrate that Shape2Match achieves a 100% success rate (SR) across all datasets and significantly outperforms other methods in the number of correct matches (NCM), proving its effectiveness and robustness against NRD, rotation, and scale variations.

1. Introduction

Image matching is one of the most important upstream tasks in computer vision, playing a key role in downstream applications such as autonomous driving, industrial manufacturing, and medical imaging (Zhao et al., 2025). Over the past two decades, researchers have extensively studied the problem of image matching, proposing classical methods such as SIFT (Lowe, 2004), SURF (Bay et al., 2006), and ORB (Rublee et al., 2011), which have gradually become benchmark methods in this field. However, with the advancement of imaging technologies, image modalities have become increasingly diverse, and cross-modal image matching remains a challenging problem. Among these modalities, infrared imagery, which captures precise thermal radiation information, exhibits strong complementarity with visible imagery in both imaging mechanisms and information content (Ma et al., 2023). The fusion of the two is therefore crucial to improving the reliability of perception systems under complex conditions such as extreme weather and nighttime operations (Liu et al., 2025). Consequently, infrared–visible image matching has attracted growing research attention in recent years.

Existing image matching methods can generally be divided into two categories: region-based matching methods and feature-matching methods. Region-based matching methods directly match pixel values between images by defining a similarity measure, typically employing a template matching strategy that combines local and sliding windows to search for correspondences. The performance of this approach critically depends on the similarity measure. The two most common measures are Normalized Cross-Correlation (NCC) and Mutual Information (MI), which compute normalized correlation or mutual information between local windows based on their intensity histograms (Pratt, 1974)(Viola and Wells, 1995). In addition, Fourier transform–based methods have also been proposed. For instance, Ye et al. drew inspiration from the concept of HOG, replacing gradients with phase congruency (PC) and extending it to capture both magnitude and orientation information, thereby constructing local descriptors for window matching via NCC to achieve robust registration (Ye et al., 2017). However, template matching methods are often unable to handle complex rotations and geometric distortions between images. Moreover, their correspondence estimation is prone to local optima, leading to instability when addressing the significant NRD between infrared and visible images (Zhou et al., 2025).

Feature matching methods detect feature points in images and establish correspondences by constructing handcrafted descriptors for these points (Leng et al., 2019). The matching process is typically divided into three stages: feature detection, feature description, and feature matching (Ma et al., 2020). SIFT builds a difference of gaussian (DoG) pyramid to locate keypoints and assigns each keypoint a dominant orientation based on local gradient statistics to form a descriptor. Inspired by SIFT, Bay et al. introduced SURF, which detects feature points using the Hessian matrix and accelerates computation through the use of integral images and box type convolution filters. However, due to the distinct imaging mechanisms of infrared and visible sensors, NRD often arises between these modalities. This manifests as significant differences in intensity distributions at corresponding locations, leading to two major issues: (1) inconsistent dominant orientations between corresponding points, resulting in a loss of rotation invariance for orientation-based descriptors, and (2) inconsistent gradient distributions within local neighborhoods, causing gradient-based descriptors to exhibit low correlation. Consequently, the stability of traditional handcrafted descriptors is greatly reduced when applied to infrared–visible image registration. To address this limitation, Ma et al. proposed PSO-SIFT, which computes second-order gradients in Gaussian scale space and replaces SIFT's rectangular grid with a log-polar template to enhance robustness against radiometric variations (Ma et al., 2017). RIFT employed phase congruency (PC) (Kovesi, 2000) to generate PC maps, detected edge and corner points using FAST, and replaced image intensity with PC values, thereby enhancing its ro-

* These authors contributed equally to this work.
† Corresponding author

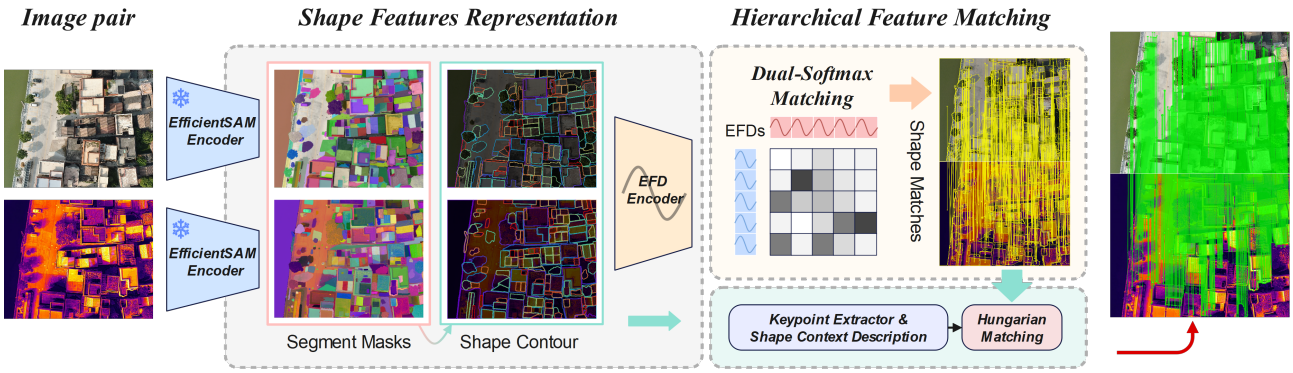

Figure 1. The Shape2Match Framework. Our pipeline consists of two main stages: Shape Feature Representation and Hierarchical Feature Matching. We employ a coarse-to-fine strategy, this two-stage process yields a final set of precise sparse correspondences, as visualized in the accurate registration overlay (far right).

bustness to NRD (Li et al., 2020). To mitigate RIFT's instability under specific rotation angles, Li et al. proposed Shift Rotation Invariant Feature Descriptor (sRIFD) which enhanced rotation continuity by increasing the number of orientations in the Log-Gabor filter and introducing a bidirectional fusion strategy (Li et al., 2023). Gao et al. proposed the Multiscale Histogram of Local Main Orientation (MS-HLMO), which calculated ASG across scales, fusing weighted gradient orientations into a partial main orientation map (PMOM) to assign dominant directions to Harris corners (Gao et al., 2022). Furthermore, Li et al. proposed LNIFT, which applies a local normalization filter to map multimodal images into a common intermediate modality, enabling gradient-based descriptor matching while reducing the computational cost associated with frequency-domain operations (Li et al., 2022).

However, the key to infrared–visible image matching lies in identifying sketch features between the two modalities (Liao et al., 2024). In existing feature matching methods, this role is predominantly fulfilled by point features, whose effectiveness has been extensively validated within single-modal settings. Nevertheless, due to the severe NRD between infrared and visible images, the same feature point does not always appear simultaneously in both modalities (Cui et al., 2020). As a result, relying solely on point features is insufficient to ensure stable matching performance.

Based on the above analysis, this paper proposes a shape-based infrared and visible image matching method named Shape2Match. First, instead of relying on feature points, Shape2Match adopts shape features, which raises higher requirements for precise shape extraction. To this end, both infrared and visible images are segmented using EfficientSAM (Xiong et al., 2024), maximizing the accuracy of shape boundary detection while maintaining efficiency. Next, edge extraction is performed on the shape masks output by EfficientSAM, and EFD (Kuhl and Giardina, 1982) is employed to construct rotation- and scale-invariant feature descriptors for the shape boundaries. On this basis, high-curvature points along the shape boundaries are identified as feature points, and the Shape Context descriptor is applied to these points to achieve shape- and scale-invariant characterization. This strategy significantly broadens the spatial distribution of feature points available for matching in Shape2Match, thereby enhancing overall matching performance. Finally, Shape2Match is evaluated on 6 datasets containing 153 image pairs, and the experimental results demonstrate that Shape2Match achieves excellent performance in terms of NCM, SR, RMSE, and ME, substantially improving the stability of matching under variations in rotation, scale, and NRD.

2. Methodology

In this section, we propose Shape2Match, a novel image matching framework that fundamentally rethinks the paradigm of feature correspondence in cross-modal imagery. As illustrated in Fig. 1, Shape2Match pioneers a shape-centric strategy, fundamentally diverging from conventional point-based approaches. The core insight of Shape2Match lies in leveraging semantic segmentation to extract modality-invariant shape primitives, which serve as stable anchors for subsequent matching processes.

The Shape2Match pipeline comprises three cohesive stages: First, shape primitive extraction through advanced segmentation to obtain structural contour. Subsequently, in section 3.1, we parameterize contour using EFD to derive an scale and rotation invariant shape representation. Finally, we implement a hierarchical matching strategy, this coarse-to-fine process ensures both high recall and high precision, yielding geometrically consistent sparse matches.

2.1 Shape Contour Extraction via Semantic Segmentation

The Shape2Match framework is predicated on the hypothesis that, although infrared-visible image pairs (I_A, I_B) may exhibit significant radiometric differences, their underlying geometric structures remain largely invariant. The robustness of our approach therefore critically depends on the extraction of these structural elements as high-fidelity shape contours, denoted as $\mathcal{C}$.

To achieve this, we employ EfficientSAM, a foundational zero-shot segmentation model selected for its favorable trade-off between segmentation performance and computational efficiency. By leveraging its prompt-free, category-agnostic segmentation capability, we effectively decouple the shape extraction process from modality-specific appearance features. For each image $I \in I_A, I_B$, the model produces a comprehensive and multi-granular set of binary masks M_i, resulting in an over-complete collection of candidate shape primitives.

However, these raw masks, frequently contain topological noise and boundary artifacts that compromise contour quality. We therefore introduce a dedicated contour refinement pipeline.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-37-2026

This process begins by filtering out geometrically insignificant regions based on area and perimeter thresholds(the minimum area is set to 0.1% of the image area and the minimum perimeter is set to 50 pixels in our experiments). Subsequently, it applies a sequence of morphological operations (e.g., closing and opening) to suppress boundary noise and ensure topological closure. Finally, we extract the shape representation of the external boundary from each refined mask. The output of this stage consists of two sets of clean, closed contours, $\mathcal{C}_A = \{C_i^A\}$ and $\mathcal{C}_B = \{C_j^B\}$, which serve as the geometric foundation for the subsequent matching stages.

2.2 Shape Representation with EFD

The contours $\mathcal{C} = \{C_i\}$ derived from the segmentation stage, represented as ordered point sequences $C_i = \{p_k\}_{k=1}^{L}$, are ill-suited for direct matching. They are of variable length L and, more critically, lack invariance to rotation and scale. To overcome this, we parameterize these contours into a fixed-dimensional, transformation-invariant feature space.

The Elliptic Fourier Descriptor. We address this by adopting EFD, which perform a harmonic decomposition of a 2D closed contour, modeling the shape as the superposition of rotating phasors that trace a series of ellipses. Given a contour C with L points and total perimeter T, we decompose the contour path $C(t)$ into two separate Fourier series for its x and y projections. By truncating the series at the N-th harmonic, the contour can be compactly reconstructed and represented by $4N$ coefficients $\{a_n, b_n, c_n, d_n\}_{n=1}^{N}$, plus the DC components (a_0, c_0):

$$\begin{aligned} x(t) &= \frac{a_0}{2} + \sum_{n=1}^{N}\left(a_n \cos\left(\frac{2n\pi t}{T}\right) + b_n \sin\left(\frac{2n\pi t}{T}\right)\right) \\ y(t) &= \frac{c_0}{2} + \sum_{n=1}^{N}\left(c_n \cos\left(\frac{2n\pi t}{T}\right) + d_n \sin\left(\frac{2n\pi t}{T}\right)\right) \end{aligned} \tag{1}$$

where t is the cumulative path length, this provides a fixed-dimension representation $\mathbf{C} \in \mathbf{R}^{N\times 4}$ with DC components for any arbitrary closed contour.

Normalization of EFD. The raw coefficients, however, are still variant to rotation, scale and contour starting point. We achieve invariance through a standard normalization process. Translation invariance is obtained by discarding the DC components (a_0, c_0), which simply represent the contour's centroid (we store the centroid c_i separately for later use). Scale invariance is achieved by dividing all remaining $4N$ coefficients by the semi-major axis length of the first harmonic ellipse, a_1, which encodes the contour's overall size. Rotation and starting-point invariance are jointly achieved by analytically rotating and shifting the coefficients of all harmonics based on the orientation and phase of this fundamental ellipse. This normalization yields a final, invariant feature vector $\mathbf{F}_i \in \mathbf{R}^{4N-3}$ for each contour C_i.

2.3 EFD-based Shape Matching Algorithm

Given the two sets of normalized EFD feature vectors, $\mathcal{F}_A = \{\mathbf{F}_i^A\}$ and $\mathcal{F}_B = \{\mathbf{F}_j^B\}$, we establish coarse shape correspondences through a specialized matching strategy. While the standard L_2 distance $\|\mathbf{F}_i - \mathbf{F}_j\|_2$ treats all harmonics equally, we hypothesize that lower-order harmonics—encoding shape structure—exhibit greater robustness to noise and modality variations than high-order harmonics capturing fine details.

We therefore introduce a Weighted EFD Distance D_W. For two EFD feature $\mathbf{F}_i, \mathbf{F}_j$, we compute the per-harmonic distance $d_n = \|\mathbf{C}_i^{(n)} - \mathbf{C}_j^{(n)}\|_2$, where $\mathbf{C}^{(n)} = [a_n, b_n, c_n, d_n]^\top$. The weighted metric is defined as:

$$D_W(\mathbf{F}_i, \mathbf{F}_j) = \sum_{n=1}^{N} w_n \cdot d_n \tag{2}$$

with weights w_n following a normalized exponential decay $w_n \propto \exp(-\lambda n)$ to emphasize fundamental harmonics. In our experiments, λ is set to 0.1 to emphasize low-order harmonics that capture global shape structures while suppressing high-frequency noise. The pairwise distance matrix $\mathbf{D}$, where $\mathbf{D}_{ij} = D_W(\mathbf{F}_i^A, \mathbf{F}_j^B)$, is converted to similarity scores via $\mathbf{S} = \exp(-\mathbf{D}/\tau)$. The temperature parameter τ is set to 0.05 in our experiments to ensure stable similarity scaling. We then apply a dual-softmax operation with temperature scaling and numerical stabilization:

$$\mathbf{P}_{\text{mutual}} = \text{Softmax}_{row}(\mathbf{S}) \odot \text{Softmax}_{col}(\mathbf{S}) \tag{3}$$

where $\mathbf{P}_{\text{mutual}}$ represents the final mutual matching probability. This yields high-confidence shape correspondences defined as $\mathcal{M}_{\text{shape}} = \{(i,j) \mid \mathbf{P}_{\text{mutual}}[i,j] > \tau_{\text{conf}}\}$, which are subsequently passed to the geometric verification stage.

2.4 Hierarchical Geometric Refinement Strategy

The global shape correspondences $\mathcal{M}_{\text{shape}}$ provide robust but coarse alignment. To obtain precise sparse matches for registration, we introduce a hierarchical refinement stage operating within each matched shape pair $(C_i^A, C_j^B) \in \mathcal{M}_{\text{shape}}$.

Keypoint Extraction. We extract geometrically salient keypoints from contours by identifying high-curvature extremes. We compute the discrete curvature $\kappa(s)$ at each vertex p_s along the contour C_i. Keypoints are then identified as the local extrema of $\kappa(s)$ that satisfy a minimum prominence τ_p and spatial separation τ_{dist} to suppress noise:

$$\mathcal{K}_i = \{p_k \in C_i \mid \mathcal{E}(\kappa(p_k)) \wedge \mathcal{P}(\kappa(p_k)) > \tau_p\} \tag{4}$$

where $\mathcal{E}$ detects the local extreme and $\mathcal{P}$ denotes the curvature prominence calculation function.

Local Description. For each keypoint $p_k \in \mathcal{K}$, we employ the Shape Context descriptor (Belongie et al., 2002) to encode its local geometric context. This descriptor constructs a normalized log-polar histogram $\hat{h}_k$ that captures the relative distribution of all other contour points $\{p_i\}_{i\neq k}$:

$$\hat{h}_k(r,\theta) = \frac{1}{L-1}\sum_{i\neq k} \mathbb{I}(\text{bin}(p_i - p_k) = (r,\theta)) \tag{5}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, $\text{bin}(\cdot)$ maps relative vectors to their corresponding log-polar coordinates, r is the radial bin index, and θ is the angular bin index. Radial distances are normalized by a global scale factor, while angular bins are aligned with the local tangent direction at p_k to ensure rotation invariance.

https://doi.org/10.5194/isprs-annals-XI-2-2026-37-2026

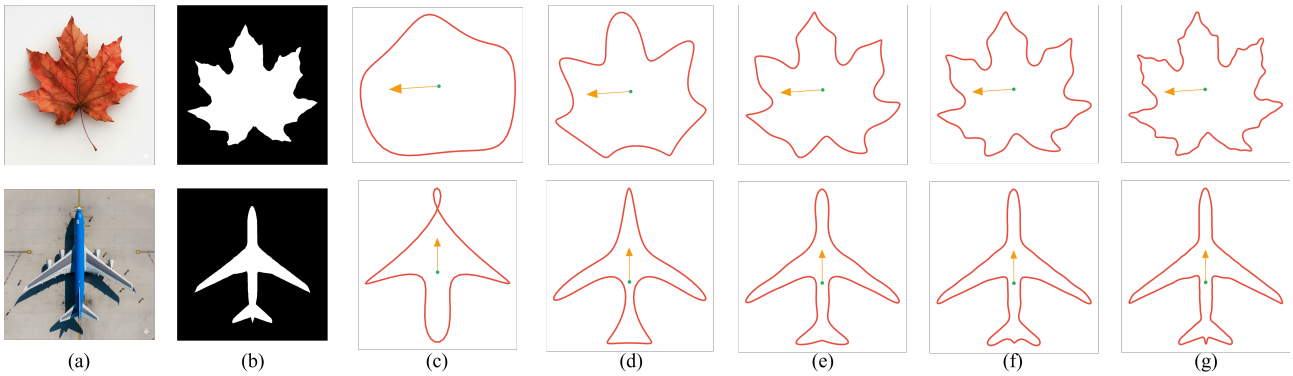

Figure 2. The reconstruction results of different shapes by different Fourier orders. (a) Original images created by Google Gemini and (b) their shape masks. (c)-(g) show the reconstructed contours using EFD with orders 4, 8, 16, 32, and 64, respectively.

Figure 3. Sample infrared and visible images from datasets.

Keypoint Matching. For each shape pair, we compute pairwise costs using the χ^2 distance between descriptors:

$$C_{mn} = \mathcal{D}_{\chi^2}(\hat{h}_m^A, \hat{h}_n^B) = \frac{1}{2}\sum_k \frac{(\hat{h}_m^A(k) - \hat{h}_n^B(k))^2}{\hat{h}_m^A(k) + \hat{h}_n^B(k) + \epsilon} \quad (6)$$

We solve this optimal assignment problem using the Hungarian algorithm to find the minimum-cost, one-to-one matching π^*:

$$\pi^* = \arg\min_\pi \sum_m C_{m,\pi(m)} \quad (7)$$

This process is repeated for all pairs in $\mathcal{M}_{\text{shape}}$, and all resulting keypoint matches are aggregated into $\mathcal{M}_{\text{keypoint}}$, with a final RANSAC stage computing the global transformation M_{final} and identifying inliers $\mathcal{M}_{\text{final}}$ as the Shape2Match output.

3. Experiment

We first introduce the experimental setup, detailing the datasets and the evaluation environment. Then, to validate the matching performance and demonstrate the superiority of Shape2Match, we conducted comprehensive qualitative and quantitative comparisons against several conventional methods: SIFT, sRIFD, RIFT, and MS-HLMO. All baseline methods were implemented using the official implementations released by their respective authors. The default parameter settings provided in the original implementations were adopted to ensure fair comparison across all datasets.

3.1 Experiment Settings

The infrared and visible image pairs used in our experiments were collected from several representative scenes and open-source datasets, including the IVDZ dataset, MSRS dataset (Tang et al., 2025a)(Tang et al., 2025b)(Tang et al., 2022b)(Tang et al., 2022a)(Ma et al., 2022), FLIR dataset, METU Multimodal dataset (Yaman and Kalkan, 2013)(Yaman and Kalkan, 2015), and TNO Multiband dataset (Toet, 2017). The image resolutions range from 600×800 to 2460×1936 pixels. Most of these image pairs exhibit varying degrees of scale differences, some show slight perspective variations, and the majority contain minimal rotational differences. Sample data are shown in Fig. 3.

To analyze the effect of EFD on shape contour reconstruction at different orders, we compared the reconstruction results for various Fourier orders, $N = 4, 8, 16, 32, 64$, as shown in Fig. 2. As the order increases, the reconstructed contour progressively approaches the original shape, and fine-grained features (such as leaf margins and aircraft tailfins) are clearly recovered. At lower orders ($N = 4$ or $N = 8$), the reconstruction primarily retains global contour information while smoothing out details. Conversely, at higher orders ($N = 32$ or $N = 64$), the local convex/concave characteristics and high-frequency details of the contour are well-described. However, an excessively high order can lead to a sharp increase in feature dimensionality and potential overfitting to shape noise. Balancing reconstruction accuracy and computational complexity, this study selected $N = 32$ as the default order. This choice ensures precise shape description while maintaining a manageable feature dimensionality.

To quantitatively evaluate the performance of the proposed method, we required ground-truth geometric transformations between each image pair. Before conducting experiments, each image pair was subjected to rotation and scaling operations, after which five uniformly distributed correspondences were manually selected on both images to compute an affine transformation as ground truth. For quantitative analysis, we adopted four primary evaluation metrics: NCM, SR, Root Mean Square Error (RMSE), and Mean Error (ME). A match is considered correct if the Euclidean distance between corresponding points after geometric transformation is less than 5 pixels; a matching attempt is regarded as a failure if the NCM between two images is fewer than 3.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-37-2026 | © Author(s) 2026. CC BY 4.0 License.

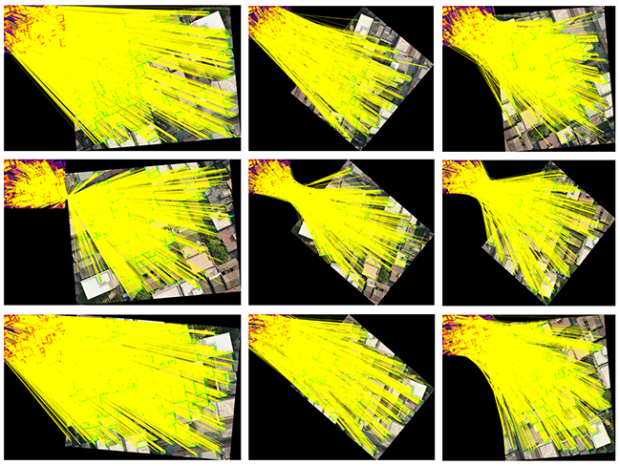

Figure 4. Rotation invariance test: Several representative registration results.

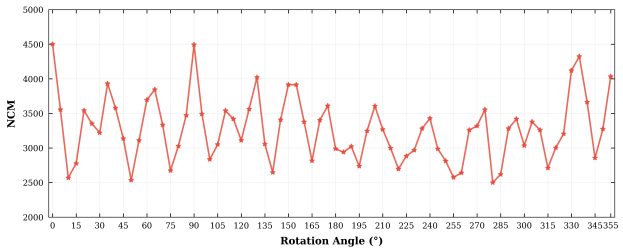

Figure 5. Rotation invariance test: NCM values for 72 image pairs.

3.2 Invariance Test

3.2.1 Rotation Invariance To evaluate the robustness of Shape2Match against rotation variations, we selected an image pair, holding one image fixed while progressively rotating the other from 0° to 355° in 5° increments. This procedure generated 72 distinct test pairs. We then applied Shape2Match to all 72 pairs and recorded the NCM for each rotation angle. Fig. 4 presents part of the registration results.

Fig. 5 illustrates the performance of Shape2Match under varying rotation conditions. As the rotation angle spans from 0° to 355°, the NCM values fluctuate within a high range (2500 to 4500), indicating that the method maintains relatively stable matching performance. The oscillations observed in the curve are primarily attributed to the geometric distortion introduced by the rotation. This distortion can slightly alter local feature structures and, more critically, degrade segmentation repeatability. Specifically, the consistency of shape feature detection—which is inherently constrained by EfficientSAM’s automatic mask generation—is compromised, leading to the observed NCM variations.

3.2.2 Scale Invariance We employed the same experimental strategy to evaluate Shape2Match’s scale invariance: an image pair was selected from the dataset, keeping one image fixed while scaling the other from 1.0× to 6.0× in 0.2× increments, generating 21 image pairs. Shape2Match was performed on these 20 pairs, with the NCM recorded for each pair. Fig. 6 presents part of the registration results.

Fig. 7 illustrates the performance variation of Shape2Match under different scale conditions. Overall, as the scale factor increases from 1.0× to 6.0×, the NCM values remain within the

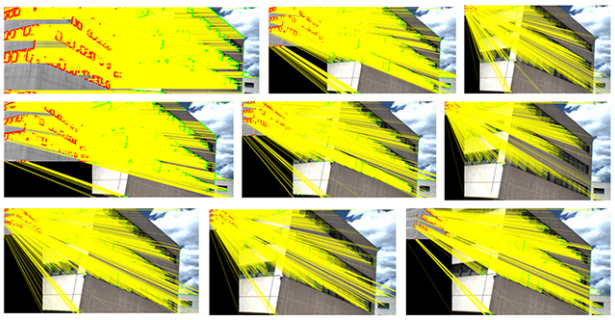

Figure 6. Scale invariance test: Several representative registration results.

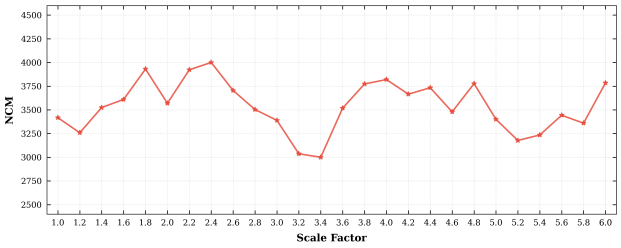

Figure 7. Scale invariance test: NCM values for 26 image pairs.

range of 3000 to 4000, indicating that the method maintains good stability and exhibits strong scale invariance even under large-scale variations. The curve exhibits noticeable fluctuations at scale factors near 3.2× and 3.4×, which are probably attributed to the interpolation process during image enlargement. Interpolation introduces smoothing effects that reduce edge sharpness, thereby decreasing the precision of shape-based feature extraction. Consequently, some feature points exhibit spatial displacement and are subsequently identified as outliers and eliminated during geometric consistency filtering.

3.3 Registration Performance Test

3.3.1 Qualitative Analysis For this evaluation, we selected six experimental image pairs. Fig. 8 presents comparative results for SIFT, RIFT, sRIFD, MS-HLMO, and the proposed Shape2Match method. The experimental results reveal significant performance differences among the evaluated methods. Among them, SIFT performed the worst, successfully matching only two out of six image pairs, with a very limited number of correspondences (57 and 95, respectively). As discussed earlier, SIFT is more suitable for optical image matching. The third and fourth image pairs originate from the METU dataset, where the infrared band (830 nm) is spectrally close to the visible band; thus, SIFT could still produce marginally acceptable results. However, as the modality gap widens, SIFT fails completely. RIFT and sRIFD exhibited nearly identical yet limited performance, each achieving an 83% SR due to their failure on the first image pair. This specific pair exhibits a large scale difference, which poses a significant challenge to the algorithms’ scale-invariant capabilities. While MS-HLMO and the proposed Shape2Match both achieved 100% SR across all datasets, Shape2Match consistently produced a higher NCM in every test case, with uniformly distributed correspondences across the image domain. Moreover, Shape2Match demonstrated superior rotation and scale invariance, confirming its robustness under multimodal matching conditions.

3.3.2 Quantitative Analysis In this section, we evaluate five methods, including Shape2Match, across all 6 datasets. To

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-37-2026 | © Author(s) 2026. CC BY 4.0 License.

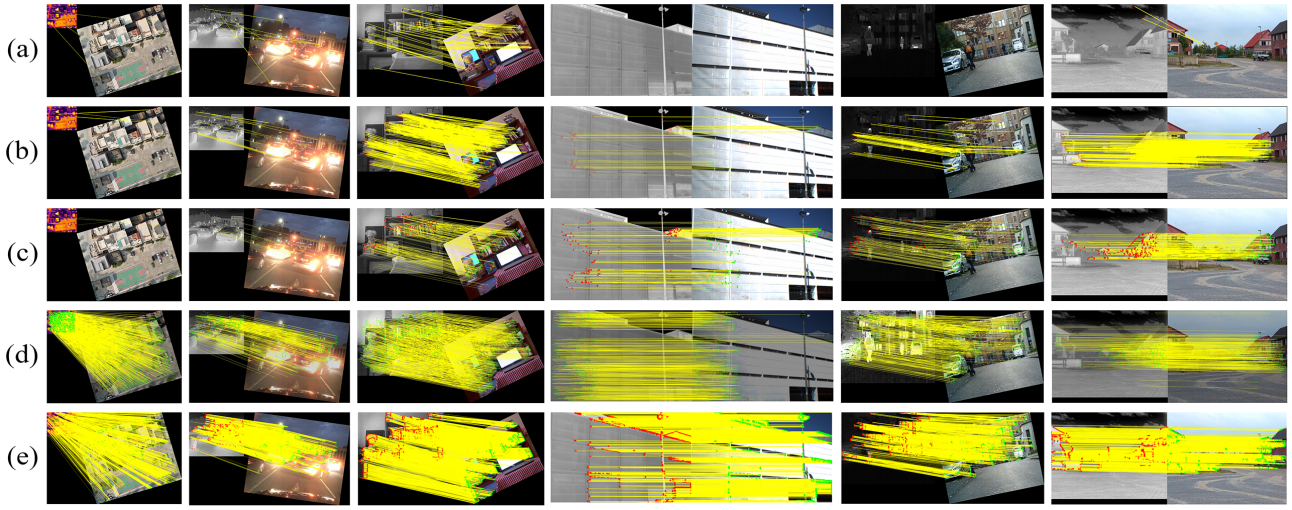

Figure 8. The qualitative results of different feature matching method across IVDZ, FLIR, METU, LGHD, MSRS and TNO datasets:(a) SIFT. (b) RIFT. (c) sRIFD. (d) MS-HLMO and our (e) Shape2Match.

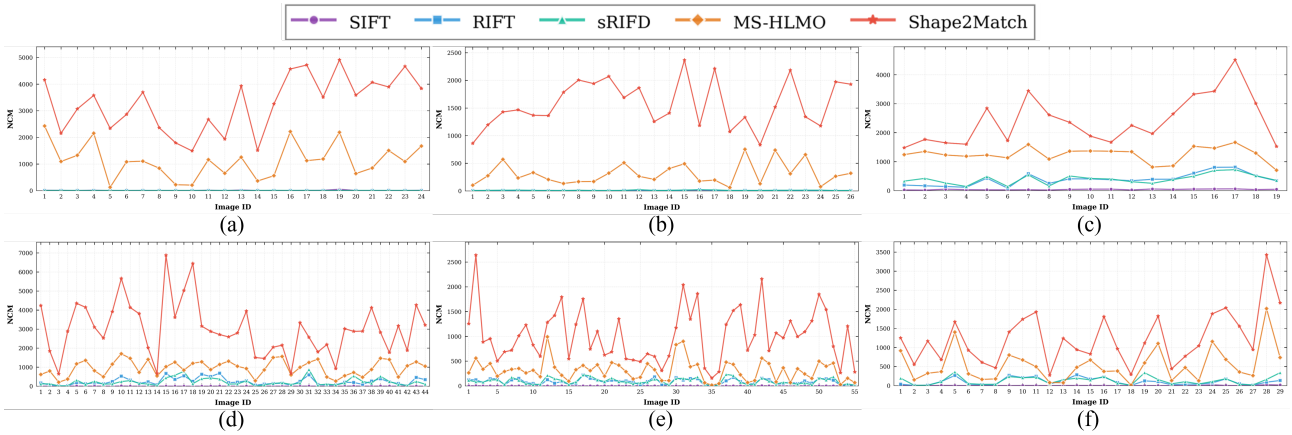

Figure 9. Quantitative results of NCM values for different feature matching method across six experiment datasets:(a) IVDZ. (b) FLIR. (c) METU. (d) LGHD (e) MSRS and (f) TNO dataset

facilitate a quantitative comparison of their matching performance, we summarize the NCM, SR, and RMSE achieved by each method on each dataset in Fig. 9, Table 1, and Table 2, respectively. It is important to note that RMSE represents the average values of each method across every single datasets, and image pairs that were not successfully registered are excluded from the statistics.

As shown in the figure, SIFT continues to perform well on the METU dataset, consistent with the qualitative analysis and in line with expectations. RIFT and sRIFD again demonstrate similar behavior, achieving relatively high and comparable SR values across most datasets except for FLIR and IVDZ, where neither method successfully matched any image pairs. Beyond the evident impact of large scale differences, this result also raises the possibility that both methods are sensitive to variations in pseudocolor mappings. It appears that these intensity discrepancies, which originate outside the sensor, impose additional challenges to their NRD robustness. MS-HLMO successfully completed matching on all datasets, achieving an SR of 100%, which indicates good overall performance. However, compared to other datasets, the NCM values of MS-HLMO dropped significantly on most image pairs from the MSRS and FLIR datasets, often falling below 500. In contrast, the proposed Shape2Match also achieved 100% SR across all datasets and consistently outperformed other methods in terms of NCM in nearly every case. These results demonstrate that replacing point features with shape features enables Shape2Match to achieve superior matching performance.

Table 1. Quantitative Comparison: SR for 5 methods across IVDZ, FLIR, METU, LGHD, MSRS, and TNO.

method	SR/%					
	IVDZ	FLIR	METU	LGHD	MSRS	TNO
SIFT	62.5	7.7	100.0	4.5	0.0	17.2
RIFT	8.3	50.0	100.0	10.0	100.0	93.1
sRIFD	4.2	46.2	100.0	95.5	98.2	89.7
MS-HLMO	100.0	100.0	100.0	100.0	100.0	100.0
Shape2Match	100.0	100.0	100.0	100.0	100.0	100.0

Regarding registration accuracy, SIFT performed the worst overall. However, SIFT surprisingly achieved significantly better results than RIFT and sRIFT on the IVDZ dataset. In fact, its SR on this dataset also surpassed those of the other two methods—a counterintuitive outcome that warrants further investigation in future work. RIFT and sRIFD exhibited nearly identical performance across almost all datasets, consistent with the earlier analysis. Shape2Match and MS-HLMO achieved

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-37-2026 | © Author(s) 2026. CC BY 4.0 License.

Table 2. Quantitative Comparison: RMSE for 5 methods across IVDZ, FLIR, METU, LGHD, MSRS, and TNO.

method	RMSE/pixel					
	IVDZ	FLIR	METU	LGHD	MSRS	TNO
SIFT	3.40	4.05	2.72	3.97	-	3.62
RIFT	3.54	3.54	2.34	3.38	2.28	2.51
sRIFD	3.91	3.59	2.56	3.22	2.34	2.60
MS-HLMO	2.86	2.78	1.93	2.17	1.97	2.36
Shape2Match	2.63	2.96	2.29	2.06	1.88	2.52

comparable accuracy, with each outperforming the other on roughly half of the datasets. The shape-based matching strategy of Shape2Match enhances its robustness to NRD and ensures stable registration, as strongly evidenced by its high NCM and SR metrics. On the other hand, its reliance on segmentation for feature point localization makes the resulting points sensitive to segmentation quality, which in turn slightly compromises the registration accuracy. This trade-off is, in fact, consistent with our expectations.

3.4 Limitations

The matching performance of Shape2Match depends heavily on the quality of shape feature extraction. In this study, EfficientSAM was employed to segment images and obtain target masks. Although this approach improves computational efficiency to some extent, the segmentation results produced by EfficientSAM are less accurate than those of the original SAM, which in turn adversely affects Shape2Match’s overall matching performance. Moreover, replacing point features with shape features renders Shape2Match particularly effective for registering images containing abundant stable shapes and well-defined structural patterns. However, its performance is limited when applied to images dominated by fine textures or indistinct edges. The noticeably higher RMSE of Shape2Match on the TNO dataset corroborates this limitation.

4. Conclusion

In this paper, we propose a novel shape-based infrared and visible image matching method, termed Shape2Match. We first analyze why conventional feature matching methods often perform poorly on infrared–visible image pairs: point features are unstable across modalities and difficult to detect reliably. To address this limitation, we present in detail how Shape2Match overcomes the issue. During the feature detection stage, Shape2Match introduces stable shape features to replace point features, fundamentally improving robustness to NRD. In the feature description stage, we employ EFD to characterize shape features, granting the method rotation and scale invariance. Finally, we conduct both qualitative and quantitative experiments on 153 image pairs from five datasets, validating the effectiveness and superiority of Shape2Match, followed by an in-depth analysis of its applicability and limitations.

References

Bay, H., Tuytelaars, T., Van Gool, L., 2006. *SURF: Speeded Up Robust Features*. Springer Berlin Heidelberg, 404–417.

Belongie, S., Malik, J., Puzicha, J., 2002. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4), 509-522.

Cui, S., Xu, M., Ma, A., Zhong, Y., 2020. Modality-Free Feature Detector and Descriptor for Multimodal Remote Sensing Image Registration. *Remote Sensing*, 12(18). https://www.mdpi.com/2072-4292/12/18/2937.

Gao, C., Li, W., Tao, R., Du, Q., 2022. MS-HLMO: Multiscale Histogram of Local Main Orientation for Remote Sensing Image Registration. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-14.

Kovesi, P., 2000. Phase congruency: A low-level image invariant. *Psychological Research*, 64(2), 136-148.

Kuhl, F. P., Giardina, C. R., 1982. Elliptic Fourier Features of a Closed Contour. *Computer Graphics and Image Processing*, 18(3), 236–258.

Leng, C., Zhang, H., Li, B., Cai, G., Pei, Z., He, L., 2019. Local Feature Descriptor for Image Matching: A Survey. *IEEE Access*, 7, 6424-6434.

Li, J., Hu, Q., Ai, M., 2020. RIFT: Multi-Modal Image Matching Based on Radiation-Variation Insensitive Feature Transform. *IEEE Transactions on Image Processing*, 29, 3296-3310.

Li, J., Xu, W., Shi, P., Zhang, Y., Hu, Q., 2022. LNIFT: Locally Normalized Image for Rotation Invariant Multimodal Feature Matching. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–14.

Li, Y., Li, B., Zhang, G., Chen, Z., Lu, Z., 2023. sRIFD: A Shift Rotation Invariant Feature Descriptor for multi-sensor image matching. *Infrared Physics & Technology*, 135, 104970.

Liao, Y., Tao, P., Chen, Q., Wang, L., Ke, T., 2024. Highly adaptive multi-modal image matching based on tuning-free filtering and enhanced sketch features. *Information Fusion*, 112, 102599.

Liu, J., Wu, G., Liu, Z., Wang, D., Jiang, Z., Ma, L., Zhong, W., Fan, X., Liu, R., 2025. Infrared and Visible Image Fusion: From Data Compatibility to Task Adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lowe, D. G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 91-110.

Ma, J., Jiang, X., Fan, A., Jiang, J., Yan, J., 2020. Image Matching from Handcrafted to Deep Features: A Survey. *International Journal of Computer Vision*, 129(1), 23-79.

Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., Ma, Y., 2022. SwinFusion: Cross-domain Long-range Learning for General Image Fusion via Swin Transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7), 1200–1217.

Ma, W., Wang, K., Li, J., Yang, S. X., Li, J., Song, L., Li, Q., 2023. Infrared and visible image fusion technology and application: A review. *Sensors*, 23(2), 599.

Ma, W., Wen, Z., Wu, Y., Jiao, L., Gong, M., Zheng, Y., Liu, L., 2017. Remote Sensing Image Registration With Modified SIFT and Enhanced Feature Matching. *IEEE Geoscience and Remote Sensing Letters*, 14(1), 3-7.

Pratt, W. K., 1974. Correlation Techniques of Image Registration. *IEEE Transactions on Aerospace and Electronic Systems*, AES-10(3), 353-358.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. Orb: An efficient alternative to sift or surf. *2011 International Conference on Computer Vision*, 2564–2571.

Tang, L., Li, C., Ma, J., 2025a. Mask-DiFuser: A Masked Diffusion Model for Unified Unsupervised Image Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-18.

Tang, L., Yan, Q., Xiang, X., Fang, L., Ma, J., 2025b. C2RF: Bridging Multi-modal Image Registration and Fusion via Commonality Mining and Contrastive Learning. *International Journal of Computer Vision*, 133, 5262–5280.

Tang, L., Yuan, J., Ma, J., 2022a. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82, 28-42.

Tang, L., Yuan, J., Zhang, H., Jiang, X., Ma, J., 2022b. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*.

Toet, A., 2017. The tno multiband image collection.

Viola, P., Wells, W., 1995. Alignment by maximization of mutual information. *Proceedings of IEEE International Conference on Computer Vision*, 16–23.

Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F., Krishnamoorthi, R., Chandra, V., 2024. Efficientsam: Leveraged masked image pretraining for efficient segment anything. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16111–16121.

Yaman, M., Kalkan, S., 2013. Multimodal Stereo Vision Using Mutual Information with Adaptive Windowing | Request PDF. *ResearchGate*.

Yaman, M., Kalkan, S., 2015. An iterative adaptive multimodal stereo-vision method using mutual information. *Journal of Visual Communication and Image Representation*, 26, 115-131.

Ye, Y., Shan, J., Bruzzone, L., Shen, L., 2017. Robust Registration of Multimodal Remote Sensing Images Based on Structural Similarity. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5), 2941-2958.

Zhao, Y., Chen, T., Dai, J., Gao, X., Chen, X., 2025. A survey of deep-learning-based image matching algorithms. *Other Conferences*.

Zhou, X., Pan, Q., Jiang, D., Zhang, J., 2025. Research on Flexible Material Identification and Positioning Based on Machine Vision. *IEEE Access*, 13, 38584-38592.