

ML-MIFD: Multi-Level Multimodal Invariant Feature Descriptor

Zening Wang¹, Haoyu Guo¹, Yongxiang Yao^{1*}, Yongjun Zhang¹, Peihao Wu¹, and Yi Wan¹

¹ WHU, School of Remote Sensing and Information Engineering, 430079, Wuhan, Hubei, China- (wangzening, haoyu.guo, yaoyongxiang, zhangyj, wupeihao, yi.wan)@whu.edu.cn

Commission II, WG II /1

Key words: multimodal image matching; invariant feature descriptor; multi-level representation; phase congruency.

Abstract:

Existing cross-modal feature descriptors are typically limited to single-level local statistics, leading to insufficient structural representation and degraded matching performance under nonlinear radiometric differences and strong noise. To address this limitation, we propose a Multi-Level Multimodal Invariant Feature Descriptor (ML-MIFD). The proposed method introduces a hierarchical descriptor to jointly capture structural information at global, mid-level, and micro-level scales, thereby enhancing robustness in cross-modal representation. First, a modality-invariant feature space is constructed based on phase congruency, where stable keypoints are detected using the FAST operator. Then, a coarse-to-fine hierarchical encoding strategy is employed to integrate multi-scale structural information by cascading three complementary layers: a global contour layer, a mid-level spatial layout layer, and a micro-level texture layer. This multi-level design enables the descriptor to simultaneously preserve global structural consistency and local discriminative details. In addition, a dominant orientation assignment is introduced to ensure rotation invariance, and a feature value truncation strategy is adopted to suppress the influence of local extreme noise. Finally, initial correspondences are established using a bidirectional matching scheme, followed by outlier rejection with the Fast Sample Consensus (FSC) algorithm, resulting in improved matching accuracy and reliability. Extensive experiments on six cross-modal remote sensing scenarios demonstrate that the proposed method consistently outperforms state-of-the-art approaches in terms of matching success rate, number of correct matches, and matching accuracy, achieving up to 100% matching success rate.

1. INTRODUCTION

With the rapid development of multi-sensor technology, multi-source and multi-modal remote sensing data are becoming increasingly abundant, and cross-modal image matching has gradually emerged as a research hotspot. Multi-modal image matching serves as a crucial foundation for enabling the synergy of multi-source data and plays a key role in applications such as image fusion, change detection, target recognition and tracking, and image localization (Jiang et al., 2021). However, due to significant differences in the imaging mechanisms of different modalities, cross-modal image matching still faces numerous challenges. Such differences typically introduce pronounced nonlinear radiometric variations and feature heterogeneity. Moreover, complex geometric distortions further increase the difficulty of matching. Affected by the above factors, traditional feature description methods often struggle to effectively capture the common structural information across cross-modal images, leading to degraded matching performance (Li et al., 2022).

In recent years, researchers have proposed a large number of methods for cross-modal image registration. Existing methods can be broadly categorized into three types: area-based methods, deep learning-based methods, and feature-based methods. Area-based methods typically establish correspondences by evaluating the similarity between two image windows. Among these, normalized cross correlation (NCC) and mutual information (MI) are the most representative similarity measures. In cross-modal matching tasks, to mitigate the impact of nonlinear radiometric distortions, some studies compute window similarity in transformed feature spaces, such as HOPC (Ye et al., 2017), MAGD (Xiang et al., 2023), and other similar approaches. However, these methods often struggle to handle scale and rotation variations. In addition, images lack reliable

georeferencing information or exhibit significant positioning errors, their matching robustness is often difficult to guarantee. In recent years, deep learning methods have made significant progress in the field of image matching. Methods represented by SuperGlue (Sarlin et al., 2020) and LoFTR (Sun et al., 2021) explicitly model relationships between cross-image features through graph neural networks or Transformer architectures, significantly improving matching accuracy and robustness in complex scenes. Nevertheless, such methods typically rely on large-scale annotated data for training, and their generalization capability and cross-region adaptability remain limited when faced with significant nonlinear radiometric differences and geometric distortions in cross-modal remote sensing imagery. To alleviate the above issues, some studies have attempted to introduce structural priors to enhance feature representation capabilities. For example, SFA-Net (Gao, 2025) leverages edge structure information extracted by SAM as a prior to guide the fusion of CNN features with self-attention and cross-attention mechanisms, effectively improving the robustness and accuracy of multi-modal remote sensing image matching.

Among feature-based methods, utilizing image intensity or gradient information for feature point detection and description is the most common strategy. The most representative approach is the Scale-Invariant Feature Transform (SIFT) (Lowe, 2004). A series of improved methods based on SIFT, such as SAR-SIFT (Dellinger et al., 2015) and PSO-SIFT (Ma et al., 2017), have achieved remarkable performance in same-modal image matching tasks. However, gradient-based features are sensitive to nonlinear radiometric distortions in multi-modal images, limiting their cross-modal applicability. To alleviate this issue, researchers have proposed feature representation methods based on phase consistency. Methods based on phase features, such as

* Corresponding author

LGHD (Aguilera et al., 2015) and RIFT (Li et al., 2020), have been successively proposed and have shown outstanding performance on data types such as infrared imagery, electronic maps, and depth maps. In recent years, a class of multi-modal matching methods based on handcrafted invariant features has been further developed. By modeling low-level stable information such as phase, structure, and orientation, these methods construct feature representations that are robust to nonlinear radiometric differences and geometric transformations, thereby achieving more reliable cross-modal matching (Gao et al., 2025, 2024; Wan et al., 2024; Zhang et al., 2023). However, under strong noise interference and low-texture scenarios—such as multiplicative speckle noise in SAR and optical imagery, or insufficient texture information in nighttime light imagery—the above methods still exhibit high sensitivity. In particular, the degradation of local details significantly affects the descriptor's ability to characterize the neighborhood structure around feature points. Therefore, how to effectively suppress noise interference in multi-modal images while enhancing robustness to nonlinear radiometric distortions remains a critical issue to be addressed in the field of cross-modal image matching.

It is worth noting that although existing methods have made significant progress in robust cross-modal feature extraction, most feature descriptors are still limited to single-level spatial aggregation within fixed-size local windows. Without cross-level collaborative representation, such single-level descriptions become highly vulnerable when dealing with strong nonlinear radiometric differences and complex speckle noise commonly encountered in multimodal images (e.g., SAR and optical image matching). On one hand, strong local noise can easily disrupt the statistical distribution of textures at the micro-scale, leading to

significant drift in feature vectors. On the other hand, single-level grid partitioning fails to effectively capture the macro-scale spatial layout and meso-scale structural patterns that preserve cross-modal consistency. Therefore, overcoming the representation bottleneck of traditional single-level descriptors and achieving a comprehensive and balanced characterization of multi-granularity structural information in the neighborhood of feature points, while suppressing local noise interference, remains a critical challenge for improving the robustness of cross-modal image matching.

To address the above challenges, we propose a Multi-Level Multimodal Invariant Feature Descriptor (ML-MIFD). The proposed method consists of three main stages: feature point extraction, ML-MIFD construction, and image matching. First, a modality-invariant feature space is constructed based on phase congruency, and robust keypoints are detected using the FAST operator. Second, to improve feature stability under nonlinear radiometric differences and complex textures, the ML-MIFD descriptor is constructed. Specifically, a hierarchical structure with global-level, mid-level, and micro-level layers is employed to jointly represent global structure, spatial layout, and micro-level texture information. In addition, multi-directional phase responses are integrated, and a dominant orientation assignment mechanism is introduced to ensure rotation invariance. A high-dimensional descriptor is then constructed via multi-level orientation histogram statistics, enhancing robustness to noise. Finally, initial correspondences are established using a bidirectional matching strategy, followed by outlier rejection using the Fast Sample Consensus (FSC) algorithm, resulting in stable and reliable matching performance.

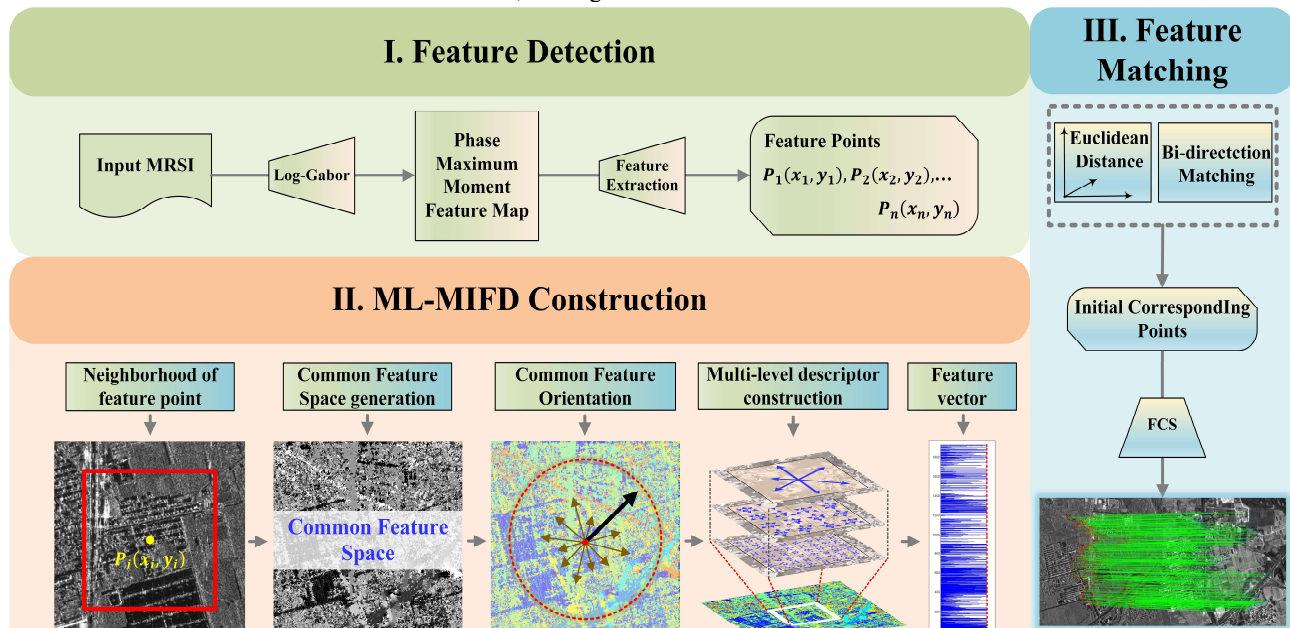


Figure 1. The pipeline of our method.

2. Image Registration Based on ML-MIFD

The ML-MIFD image registration method proposed in this paper mainly consists of three stages: (i) feature point extraction; (ii) ML-MIFD feature descriptor construction; and (iii) feature matching and outlier removal. The overall workflow of the method is illustrated in Figure 1. Among these stages, the construction of the ML-MIFD feature descriptor is the core part of this paper and will be elaborated in detail below.

2.1 Feature point extraction

To obtain robust structural features under nonlinear radiometric distortion conditions, the phase congruency (PC) of the image is first computed using multi-scale and multi-orientation Log-Gabor filters. PC features provide a local structural representation that is insensitive to contrast variations, and thus have been widely adopted in multi-modal image analysis. The detailed calculation process of PC can be found in the literature (Kovesi, 2000) and will not be reiterated in this paper.

To enhance feature saliency, this paper follows the work (Li et al., 2020) by performing moment analysis on the phase congruency responses to construct the maximum moment response map.

Finally, feature points are detected on the maximum moment response map using the FAST operator (Rosten and Drummond, 2006). Since the maximum moment map highlights structurally salient regions, this strategy effectively improves the stability of feature point detection and mitigates the impact of noise and nonlinear radiometric distortions on feature detection.

2.2 Construction of ML-MIFD descriptor

To enhance the robustness of feature descriptors in multi-modal remote sensing imagery, this paper proposes a Multi-Level Multimodal Invariant Feature Descriptor. The descriptor mainly consists of three steps: (i) common feature space generation; (ii) dominant orientation assignment; and (iii) hierarchical descriptor construction and truncation. It is worth noting that RIFT adopts a single-layer histogram structure to construct descriptors in the common feature space, whereas the proposed ML-MIFD models structural information at different granularities through a hierarchical descriptor structure.

2.2.1 Common Feature Space Generation: To enhance the adaptability of the descriptor to nonlinear radiometric distortions, a Common Feature Space (CFS) is adopted to represent the input images. Following the method of Li et al. (2020), multi-scale and multi-orientation Log-Gabor filters are first applied to compute the amplitude responses.

Let $A_{s,\omega}(x,y)$ denote the amplitude response at scale s and orientation ω . The orientation response aggregated across scales can be expressed as:

$$A_{\omega}(x,y) = \sum_{s=1}^{N_s} A_{s,\omega}(x,y), \quad (1)$$

where N_s and N_o denote the numbers of scales and orientations, respectively.

Furthermore, the value of pixel (x,y) in the Common Feature Space is defined as the index of the orientation with the maximum amplitude response:

$$\omega_{\max}(x,y) = \arg \max_{\omega \in \{1,2,\dots,N_o\}} A_{\omega}(x,y). \quad (2)$$

Therefore, each pixel in the Common Feature Space encodes the dominant structural orientation at that location. Compared with gradient-based representations, the Common Feature Space integrates multi-scale and multi-orientation Log-Gabor responses, providing stronger robustness to nonlinear radiometric distortions. As a result, it is more suitable for multimodal remote sensing image registration.

2.2.2 Dominant Orientation Assignment : To achieve rotation invariance, a dominant orientation assignment strategy similar to RIFT is adopted. In the CFS, a local neighborhood centered at each feature point is extracted, and the dominant orientation is determined as the mode of the orientation values within this neighborhood.

Based on the assigned dominant orientation, an initial rotation angle is computed, and the local orientation map is aligned to a reference direction $\theta = 0^\circ$. However, due to the inherent 180° ambiguity of edge orientations, relying solely on the dominant orientation for rotation alignment may result in orientation

flipping. To address this issue, a consistency verification strategy inspired by RIFT is further applied to the rotated local orientation map.

Unlike the original RIFT method, a circular neighborhood is adopted for dominant orientation estimation in this work, which better preserves rotational symmetry. The rotated local orientation map is then divided into upper and lower subregions, and the consistency of their orientation distributions is evaluated. If the lower subregion exhibits higher orientation consistency, an additional 180° rotation is applied to the initial alignment to resolve the directional ambiguity.

Finally, the local orientation map is rotated according to the determined angle, and the central region is cropped as the orientation-normalized local patch.

2.2.3 Hierarchical Descriptor Construction and Truncation : Traditional feature descriptors, such as SIFT, typically construct feature vectors by partitioning the neighborhood around a keypoint into several grids and computing orientation histograms within each subregion. Such methods can encode local spatial structural information to a certain extent, thereby achieving good discriminative capability. However, most existing descriptors rely primarily on single-level local statistics. As a result, they are vulnerable to noise interference, texture variations, and nonlinear radiometric distortions in cross-modal image matching, which degrades descriptor stability. To achieve collaborative representation of local structural information at multiple granularities, hierarchical orientation histograms are constructed within the orientation-normalized image patch. By cascading spatial statistical features at three levels—global contour, mid-level layout, and micro-level texture—a multi-level descriptor with strong discriminative capability is ultimately formed, as illustrated in Fig. 2.

To further improve robustness, a Multi-Level Multimodal Invariant Feature Descriptor (ML-MIFD) is proposed. Within the orientation-aligned local neighborhood in the CFS, the proposed method jointly characterizes the overall orientation distribution, spatial structural information, and fine-grained texture features through multi-level orientation histogram statistics. Unlike the RIFT descriptor, which adopts a single-level statistical structure, ML-MIFD employs a hierarchical modeling strategy to more comprehensively represent local structures in complex cross-modal scenarios.

Step 1: Hierarchical Descriptor Construction. For each feature point, a $p \times p$ orientation-aligned local patch is extracted in the CFS. A hierarchical statistical descriptor is then constructed over this patch. Let N_o denote the number of orientation bins; each histogram contains N_o bins to represent the distribution of dominant orientations. Specifically, the local structure is described at three levels:

Global level: A single orientation histogram is computed over the entire $p \times p$ patch to capture the overall directional characteristics of the neighborhood.

Mid-level: The patch is evenly divided into $m \times m$ subregions, and an orientation histogram is computed within each subregion to characterize mid-scale spatial layout.

Micro-level: The patch is further divided into $2m \times 2m$ subregions, where orientation histograms are computed to capture finer-grained local texture information.

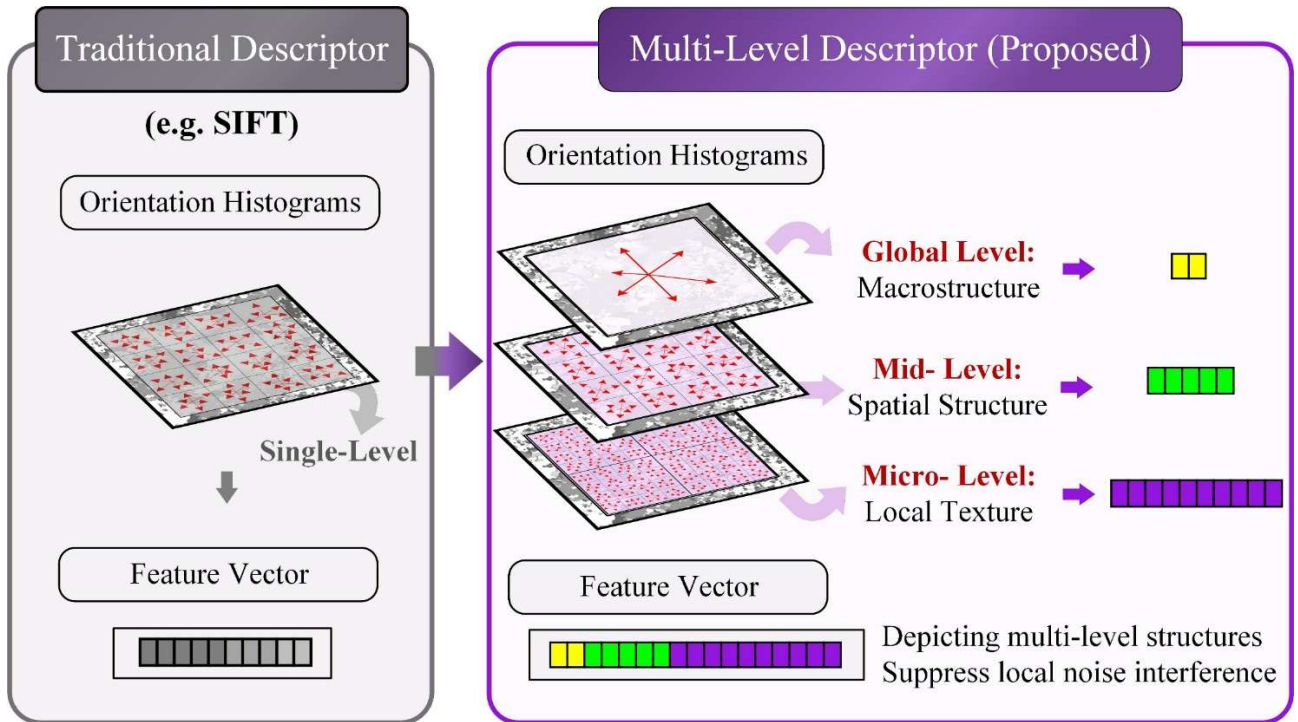


Figure 2. The multi-level orientation histogram statistical strategy constructed by ML-MIFD.

Accordingly, the dimensionality of the final descriptor is given by:

$$n = (1 + m^2 + (2m)^2)N_o = (1 + 5m^2)N_o, \quad (3)$$

where n denotes the descriptor dimension. This coarse-to-fine hierarchical structure enables the descriptor to simultaneously encode global structural information and local detailed features, thereby enhancing its representation capability in complex multimodal scenarios.

Step 2: Feature Value Outlier Truncation. Multimodal images are often affected by complex noise, such as multiplicative speckle noise in SAR images and measurement noise in depth images. These disturbances may cause certain orientation components to exhibit abnormally large responses, allowing a few dominant components to overwhelm the feature vector and reduce descriptor stability.

To mitigate this issue, each histogram within the subregions is first normalized individually. The concatenated feature vector is then processed using an outlier truncation strategy to suppress excessively large components, defined as:

$$c_i = \min(b_i, \theta), \quad i \in 1, 2, \dots, n, \quad (4)$$

where b_i and c_i denote the i -th component of the descriptor before and after truncation, respectively, and θ is the truncation threshold. In this work, $\theta = 0.6$.

This truncation effectively limits the influence of abnormally strong responses on the overall distribution, thereby improving robustness to noise and local perturbations.

Step 3: Descriptor Aggregation. The histograms from the global, mid-level, and micro-level layers are concatenated sequentially to form the combined feature vector $\mathbf{f}$. The final ML-MIFD descriptor is then obtained by applying L_2 normalization:

$$\text{ML-MIFD} = \frac{\mathbf{f}}{\|\mathbf{f}\|_2}. \quad (5)$$

The resulting ML-MIFD descriptor jointly encodes global structure, mid-level layout, and fine-grained texture information

within the feature neighborhood. Combined with the truncation normalization strategy, it achieves improved robustness against cross-modal noise interference and nonlinear radiometric variations.

2.3 Matching of Images

After constructing the ML-MIFD descriptor, feature matching is performed. First, the Euclidean distance is used as the similarity measure, and initial correspondences are established using the mutual nearest neighbor strategy.

To further eliminate mismatches, the Fast Sample Consensus (FSC) algorithm is employed for outlier rejection. FSC is a robust estimation method based on the random sample consensus principle, which effectively removes incorrect matches while preserving correspondences that satisfy geometric constraints, thereby yielding stable and reliable matching point pairs.

3. Experimental result

In this study, six state-of-the-art methods are selected for comparison, including RIFT (Li et al., 2020), HOWP (Zhang et al., 2023), WSSF (Wan et al., 2024), MS-HLMO (Gao et al., 2024), HOMO (Gao et al., 2025), and SFA-NET (Gao, 2025). In the experiments, the parameters of the proposed ML-MIFD method are set as follows: the neighborhood size $p = 72$, and the number of orientation bins $N_o = 6$. All competing methods are tuned to their optimal parameter settings.

For fair evaluation of matching performance, the number of detected feature points is uniformly limited to fewer than 4000. Since SFA-NET does not allow strict control over the number of extracted feature points within its framework, the maximum number of matched feature pairs before outlier removal is set to 1500.

The proposed ML-MIFD method and the traditional comparison methods (RIFT, HOWP, WSSF, MS-HLMO, and HOMO) are implemented using publicly available code and executed in

MATLAB R2023a. All experiments are conducted on a platform equipped with a 13th Gen Intel(R) Core(TM) i9-13900H processor (2.60 GHz), 16 GB RAM, and a Windows 11 x64 operating system.

For each image pair, a homography matrix H is provided as ground truth to evaluate the localization accuracy of the matched points. The performance of different methods is quantitatively assessed using four metrics: success rate (SR), number of correct matches (NCM), root mean square error (RMSE) of correct matches, and matching time (MT).

3.1 Image Datasets

A total of 80 pairs of multimodal remote sensing images (MRSI) are used for validation, covering six scenarios: optical–optical, infrared–optical, synthetic aperture radar (SAR)–optical, map–optical, depth–optical, and day–night images. The image sizes range from 400×400 to 817×728 pixels. The dataset includes images with nonlinear radiometric differences (NRD) and contrast variations, as well as geometric transformations such as translation, scale, and rotation changes.

3.2 Quantitative Results

Figure 3 presents the quantitative evaluation results of NCM and RMSE across six cross-modal scenarios for seven different methods. Representative matching results are randomly selected and illustrated in Fig. 4.

As shown in Fig. 3 and Table 1, the proposed ML-MIFD method successfully matches all image pairs, achieving an average NCM of 529.78 and an RMSE of 1.83 pixels, both outperforming the competing methods. The matching performance of ML-MIFD is highly robust, consistently delivering high accuracy across different cross-modal scenarios. This is mainly attributed to its multi-level feature description strategy, which integrates macro-level, mid-level, and micro-level representations to effectively capture modality-invariant features. Furthermore, it is observed that ML-MIFD demonstrates particularly superior performance in day–night image matching compared to other scenarios, achieving both the lowest RMSE and significantly higher NCM than other methods. This can be attributed to the fact that, under reduced texture details in nighttime images, the proposed multi-level descriptor is able to jointly exploit information across different scales, enabling more reliable capture of stable

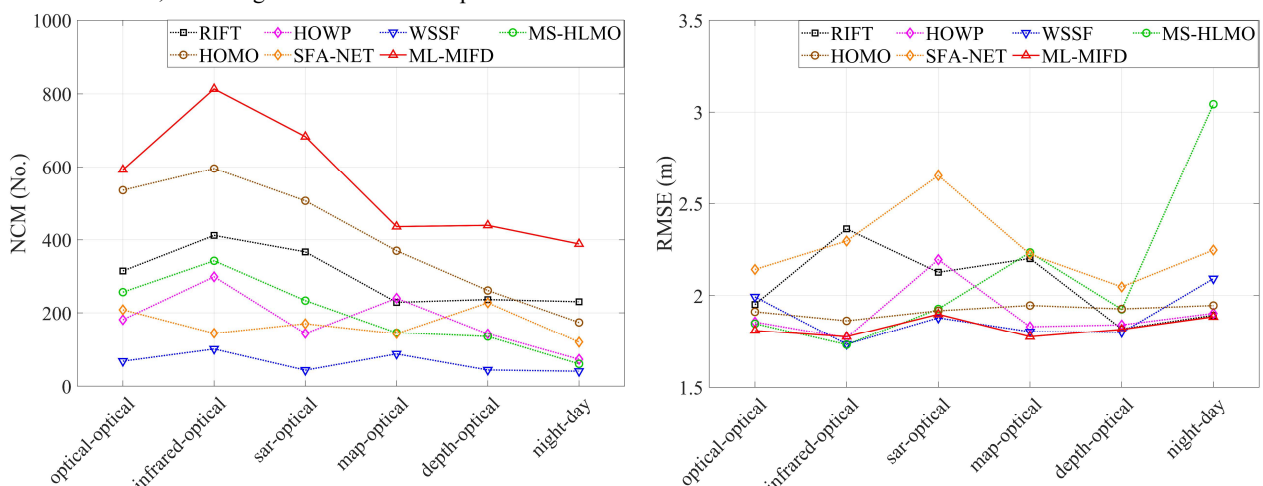
geometric structures, edge contours, and spatial layouts than conventional single-level descriptors.

In contrast, the other methods exhibit certain limitations across different metrics and scenarios. The deep learning-based method SFA-NET enhances cross-modal feature representation through a SAM-guided attention mechanism, achieving relatively high SR and stable NCM. However, its feature representation, which primarily focuses on regional semantics and structural information, is insufficient for fine-grained geometric modeling, leading to reduced localization accuracy of matching points and consequently higher RMSE.

In the synthetic aperture radar (SAR)–optical image scenario, the significant differences in imaging mechanisms between the two sensors result in pervasive speckle noise and distinct scattering characteristics in SAR images. Such strong nonlinear radiometric differences (NRDs) further exacerbate inconsistencies in edge structure representation, thereby reducing matching accuracy. Traditional feature matching methods such as RIFT demonstrate relatively stable performance in cross-modal scenarios (SR = 90.00%, RMSE = 2.00). However, their average NCM is only 286.35, indicating that their feature descriptors are insufficient to accurately characterize the local neighborhood of feature points.

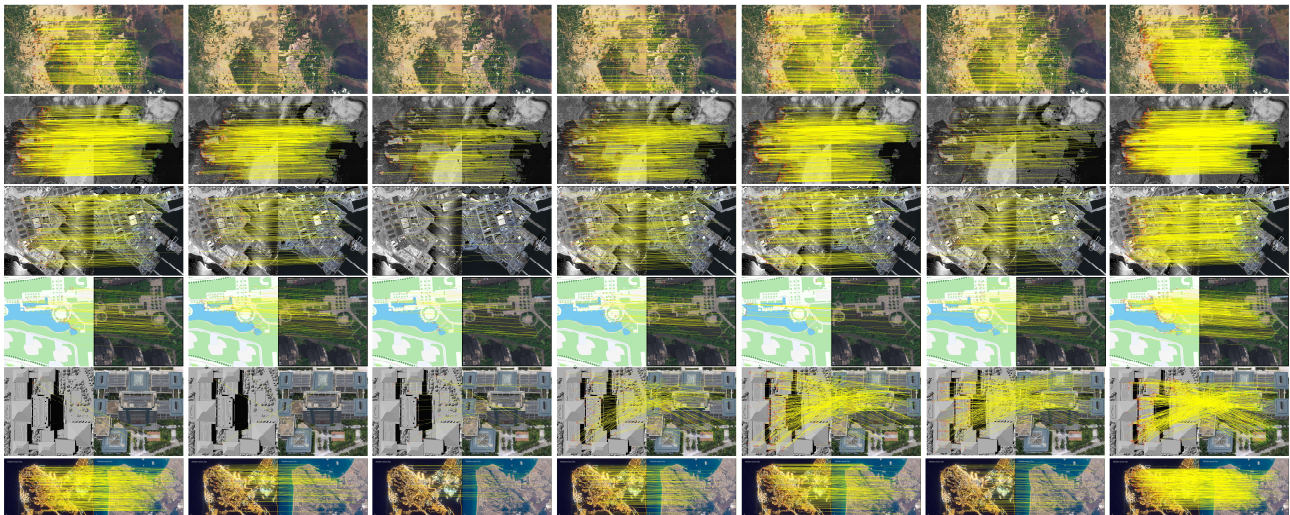
Method	SR(%)	NCM(No.)	RMSE(px)	MT(s)
RIFT	90.00	286.35	2.00	3.86
HOWP	87.50	161.37	1.91	5.55
WSSF	91.25	58.08	1.90	13.64
MS-HLMO	98.75	178.81	2.13	10.42
HOMO	95.00	388.17	1.92	15.66
SFA-NET	96.25	179.88	2.24	8.36
ML-RIFD	100.00	529.78	1.83	4.89

Table 1. Average quantitation matching scores of seven methods with respect to four evaluation indicators. The value of NCM is positively associated with the performance, the higher the number, the better the performance. As of RMSE, the RMSE results higher than 5 are set to 5 pixels.



(a) NCM results of several methods; (b) RMSE results of several methods

Figure 3. Quantitative results of several methods



(a) RIFT; (b) HOWP; (c) WSSF; (d) MS-HLMO; (e) HOMO; (f) SFA-NET; (g) ML-MIFD

Figure 4. Matching results from the seven compared methods.

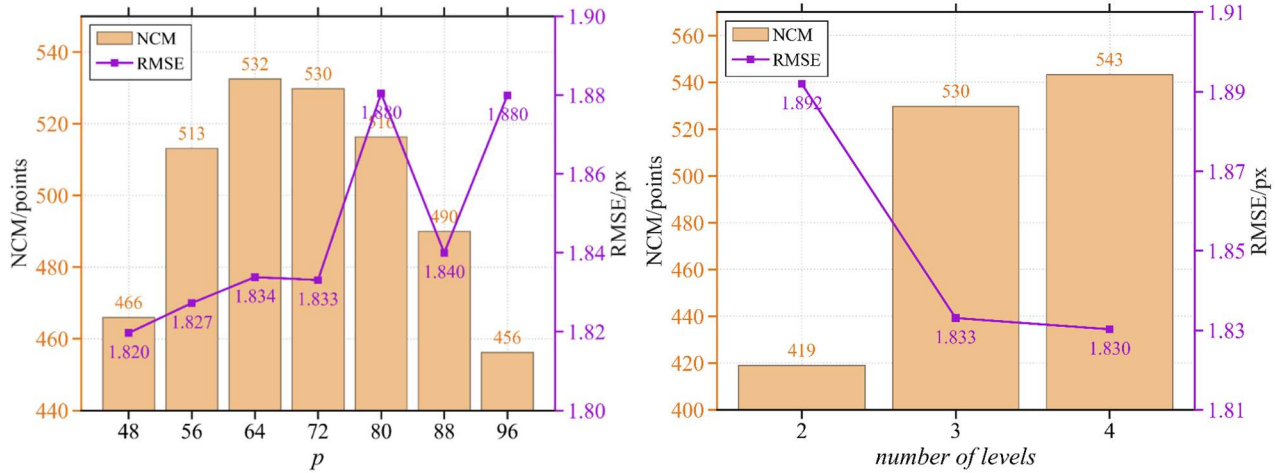
4. Parametric Analysis

The neighborhood size of the feature descriptor and the number of levels in the multi-level feature representation are key parameters of the ML-MIFD method, and different parameter settings can significantly affect its matching performance. To this end, a quantitative analysis is conducted to evaluate the performance of ML-MIFD under different parameter configurations.

A total of 80 pairs of cross-modal images are used for testing, and the influence of these parameters is assessed by examining the average number of correct matches (NCM) and RMSE. Figure 5

illustrates the impact of different parameter settings on ML-MIFD.

As shown in Fig. 5(a), when p is smaller than 64 or larger than 72, the NCM decreases rapidly. When $p = 72$, a good balance between NCM and RMSE is achieved. Regarding the number of levels, when it is set to 2 or 3, both NCM and RMSE improve significantly as the number of levels increases. Although a slight improvement is observed when the number of levels is increased to 4, the performance gain does not justify the doubled computational cost introduced by the additional level. Therefore, the optimal number of levels is determined to be 3.



(a) Assessment result of p ; (b) Assessment result of number of levels

Figure 5. Evaluation results of two indicators.

5. CONCLUSION

This paper investigates the bottleneck of insufficient representation capability in single-level feature descriptors for multimodal remote sensing image matching and proposes a Multi-Level Multimodal Invariant Feature Descriptor (ML-MIFD). The proposed method overcomes the limitations of traditional approaches that rely on single-scale local statistics by constructing a hierarchical “global–mid-level–micro-level” representation framework, achieving an effective integration of macro-structural stability and micro-level texture discriminability.

Experimental results demonstrate that ML-MIFD exhibits strong robustness and generalization across six typical cross-modal scenarios, including optical, SAR, infrared, map, and depth images. Compared with six representative methods—RIFT, HOWP, WSSF, MS-HLMO, HOMO, and the deep learning-based SFA-NET—the proposed method achieves a 100% matching success rate while significantly improving both the number of correct matches and registration accuracy. These results validate the superiority of multi-level joint modeling in handling nonlinear radiometric distortions and strong noise interference.

In summary, ML-MIFD provides a robust feature representation solution for multimodal remote sensing data processing in complex environments. Future work will explore integrating multi-level structural priors into deep learning frameworks to balance physical interpretability and computational efficiency for large-scale data.

ACKNOWLEDGEMENTS

This work was supported by the Key Program of the National Key Research and Development Program of China Grant 2024YFB3909001, the Key Program of the National Natural Science Foundation of China under project number 42530106, the Program of the National Natural Science Foundation of China under Project 42401534 and Project 42471470, the LIESMARS Special Research Funding, and the Key Laboratory of the Ministry of Education on Application of Artificial Intelligence in Equipment(2024-AAIEKF02-02).

References

- Aguilera, C.A., Sappa, A.D., Toledo, R., 2015. LGHD: A feature descriptor for matching across non-linear intensity variations, in: 2015 IEEE International Conference on Image Processing (ICIP). Presented at the 2015 IEEE International Conference on Image Processing (ICIP), IEEE, Quebec City, QC, Canada, pp. 178–181.
- Dellinger, F., Delon, J., Gousseau, Y., Michel, J., Tupin, F., 2015. SAR-SIFT: A SIFT-Like Algorithm for SAR Images. *IEEE Trans. Geosci. Remote Sens.* 53, 453–466.
- Gao, C., Li, W., Li, Y., 2024. A New Invariant Feature for Multi-Modal Images Matching, in: IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium. Presented at the IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium, IEEE, Athens, Greece, pp. 8374–8378.
- Gao, C., Li, W., Weng, D., 2025. HOMO-Feature: Cross-Arbitrary-Modal Image Matching with Homomorphism of Organized Major Orientation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10538–10548.
- Gao, T., 2025. SFA-Net: A SAM-guided focused attention network for multimodal remote sensing image matching. *ISPRS J. Photogramm. Remote Sens.*
- Jiang, X., Ma, J., Xiao, G., Shao, Z., Guo, X., 2021. A review of multimodal image matching: Methods and applications. *Inf. Fusion* 73, 22–71. <https://doi.org/10.1016/j.inffus.2021.02.012>
- Kovesi, P., 2000. Phase congruency: A low-level image invariant. *Psychol. Res.* 64, 136–148.
- Li, J., Hu, Q., Ai, M., 2020. RIFT: Multi-Modal Image Matching Based on Radiation-Variation Insensitive Feature Transform. *IEEE Trans. Image Process.* 29, 3296–3310.
- Li, L., Liu, M., Ma, L., Han, L., 2022. Cross-Modal feature description for remote sensing image matching. *Int. J. Appl. Earth Obs. Geoinformation* 112, 102964.
- Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* 60, 91–110.
- 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 Geosci. Remote Sens. Lett.* 14, 3–7.
- Rosten, E., Drummond, T., 2006. Machine Learning for High-Speed Corner Detection, in: Leonardis, A., Bischof, H., Pinz, A. (Eds.), *Computer Vision – ECCV 2006*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 430–443.
- Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. SuperGlue: Learning Feature Matching With Graph Neural Networks, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, pp. 4937–4946.
- Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X., 2021. LoFTR: Detector-Free Local Feature Matching with Transformers, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, TN, USA, pp. 8918–8927.
- Wan, G., Ye, Z., Xu, Y., Huang, R., Zhou, Y., Xie, H., Tong, X., 2024. Multimodal Remote Sensing Image Matching Based on Weighted Structure Saliency Feature. *IEEE Trans. Geosci. Remote Sens.* 62, 1–16.
- Xiang, Y., Wang, X., Wang, F., You, H., Qiu, X., Fu, K., 2023. A Global-to-Local Algorithm for High-Resolution Optical and SAR Image Registration. *IEEE Trans. Geosci. Remote Sens.* 61, 1–20.
- Ye, Y., Shan, J., Bruzzone, L., Shen, L., 2017. Robust Registration of Multimodal Remote Sensing Images Based on Structural Similarity. *IEEE Trans. Geosci. Remote Sens.* 55, 2941–2958.
- Zhang, Y., Yao, Y., Wan, Y., Liu, W., Yang, W., Zheng, Z., Xiao, R., 2023. Histogram of the orientation of the weighted phase descriptor for multi-modal remote sensing image matching. *ISPRS J. Photogramm. Remote Sens.* 196, 1–15.