Tsinghua Science and Technology
Doi: 10.26599/TST.2026.9010047

Review Article

Multimodal Image Registration for Low Altitude Platforms: Methods, Challenges, and Future Trends

Timing Li, Bing Cao, Pengfei Zhu (✉) and Kewen Li

Abstract Multimodal image registration is a fundamental task in computer vision and information fusion, supporting applications such as low altitude perception, medical imaging, remote sensing, and intelligent transportation. Gaps across modalities in imaging mechanisms, spectral responses, and geometric representations make cross-modal registration difficult in practice. Deployments face radiometric discrepancies, viewpoint variations, non-rigid deformations from platform motion, and mismatched resolutions. Recent progress in deep learning, cross-modal representation learning, and generative modeling has shifted conventional matching based pipelines toward end-to-end frameworks emphasizing fusion oriented modeling and joint optimization across tasks. This paper reviews the background, challenges, and methods for multimodal image registration in low altitude scenarios and synthesizes feature-level and pixel-level approaches. We summarize integration into downstream tasks such as object detection, semantic segmentation, and image fusion, and discuss limitations, and future directions toward accurate, transferable, and controllable registration in complex environments.

Keywords image alignment, image registration, low altitude perception, unmanned aerial vehicle

1 Introduction

Multimodal image registration is a fundamental task in computer vision and information fusion, supporting applications such as low altitude perception, medical imaging, remote sensing, and intelligent transportation [1–3]. Because modalities differ in imaging mechanisms, spatial resolution, spectral response, and noise statistics, the resulting data exhibit both geometric and radiometric heterogeneity. This often makes direct pixel-wise matching unreliable or misleading. Robust registration methods that can bridge modality gaps are therefore a prerequisite for effective multi-source fusion and downstream task performance [4].

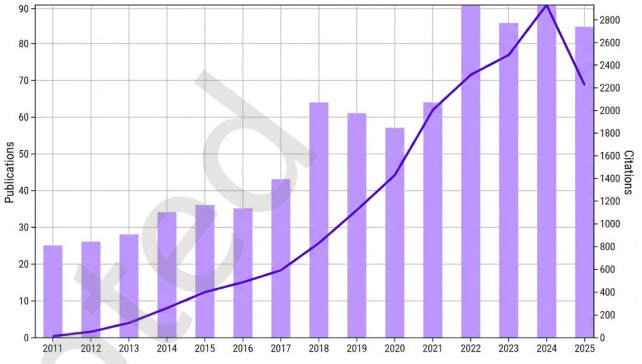

Fig. 1 Annual publications and citations in multimodal image registration from 2011 to Sep. 2025 (bars: publications; line: citations; left axis: publications; right axis: citations).

In practical applications, multimodal registration is central to medical imaging analysis, remote sensing, and urban security [5, 6]. In medical imaging, accurate registration among computed tomography, magnetic resonance imaging, and positron emission tomography helps combine complementary information for clinical decision support. In remote sensing, cooperative registration between optical imagery and synthetic aperture radar imagery, abbreviated as SAR, mitigates limitations caused by weather and surface conditions. In urban security, fusing visible and infrared imagery enhances perception in low-light environments. Recent studies on low altitude intelligent sensing also suggest the practical importance of reliable multimodal registration for UAV identification and localization in real deployment conditions [7]. Compared with settings that are more stable or have more controllable imaging conditions, low altitude platforms introduce stronger viewpoint changes, more complex local deformation, and more frequent environmental perturbations during agile motion, which makes cross-modal registration substantially more difficult [8].

Bibliometric evidence suggests sustained growth over the past fifteen years. Using Web of Science, we retrieved rele-

• Timing Li is with the School of Computer Science and Technology, Tianjin University, No.135 Yaguan Road, Haihe Education Park, Tianjin, 300372, China. E-mail: litm@tju.edu.cn
• Bing Cao is with the School of Computer Science and Technology, Tianjin University, No.135 Yaguan Road, Haihe Education Park, Tianjin, 300372, China. E-mail: caobing@tju.edu.cn
• Pengfei Zhu is with the School of Computer Science and Technology, Tianjin University, No.135 Yaguan Road, Haihe Education Park, Tianjin, 300372, China. E-mail: zhupengfei@tju.edu.cn
• Kewen Li is with the China University of Petroleum (East China), School of Computer Science and Technology, Qingdao, 266580, China. E-mail: likw@upc.edu.cn
Manuscript received: 2026-03-02; revised: 2026-04-13; accepted: 2026-05-09

vant publications with a topic search based on TS=((“Image Registration” OR “Image Alignment”) AND (“Multimodal” OR “Multi-modal”)). We then summarized annual publication volume and citation trends over the past fifteen years, as shown in Fig. 1. The overall trajectory indicates steady growth. After 2017, the rapid progress of deep learning and multimodal perception coincides with a faster increase in annual publications. Note that the 2025 statistics are counted up to September and thus represent a partial-year snapshot rather than a full-year total.

In low altitude multimodal perception, registration challenges arise from coupled geometric, scale-related, and radiometric factors. First, viewpoint changes and parallax make the geometry more ill-posed. UAV maneuvering, attitude variation, and platform vibration introduce nonlinear local deformation that undermines planar or weak-perspective assumptions, leading to local mismatches and boundary drift. Second, resolution disparity can create a semantic gap. Visible, infrared, and SAR modalities often differ greatly in imaging granularity, so textures and keypoints in high-resolution imagery may degrade or vanish in lower-resolution counterparts, which destabilizes matching under weak-feature conditions. Third, dynamic environments introduce non-stationary radiometric behavior. Sudden illumination changes, weather transitions, and asymmetric occlusions caused by moving objects can strongly perturb statistical similarity measures and reduce their reliability in real scenes [9–11]. The combined effect makes low altitude cross-modal registration a stringent test of robustness and generalization.

Despite rapid progress, low altitude multimodal registration still faces three critical bottlenecks: limited benchmark data, methodological fragility under strong deformation and radiometric heterogeneity, and the lack of tailored evaluation standards. The remainder of this paper analyzes these issues systematically and discusses their implications for reproducibility and real-world deployment.

First, benchmark data are scarce and insufficiently standardized. Existing public datasets often have limitations in modality coverage, annotation precision, and scene diversity. For example, DroneRGBT [9] is more suitable for detection studies but lacks high-precision geometric registration ground truth. Some remote sensing datasets are constrained by spatial resolution, temporal consistency, or the quality of pose-related metadata. Without a unified benchmark that covers multiple modalities and tasks and provides reliable geometric ground truth, model training, generalization assessment, and fair comparison remain difficult.

Second, strong local deformation and radiometric heterogeneity impose stringent methodological requirements. Classical affine models and conventional elastic models often fail to represent drastic local warping or region-wise independent deformation in low altitude scenarios. Deep learning-based non-rigid methods are promising, yet they can generalize poorly when encountering extreme deformation outside the training distribution, and they may lose critical structural details due to overly strong smoothness priors. Meanwhile, different imaging mechanisms lead to inconsistent texture and edge responses, making intensity-based or gradient-based consistency measures more fragile across modalities [12, 13]. Cross-scale structural discrepancies can further weaken keypoint-based approaches [14].

Third, evaluation standards tailored to cross-modal and low altitude task characteristics remain limited. Some studies still rely on metrics that are highly sensitive to radiometric consistency, such as pixel-error-based measures. These metrics do not adequately reflect semantic consistency across modalities and cannot disentangle radiometric discrepancies from geometric misalignment. The absence of comprehensive evaluation protocols spanning geometric accuracy, structural consistency, and task utility hinders standardization and broader adoption.

This paper reviews multimodal image registration with an emphasis on low altitude settings, using medical imaging and satellite remote sensing as contrastive references. We propose a taxonomy for low altitude cross-modal registration, examine how representative methods integrate with downstream tasks and where they apply, and summarize datasets and evaluation protocols with attention to reproducibility and open challenges. The remainder of the paper is organized as follows. Section 2 introduces problem definitions and background. Section 3 reviews feature-level and pixel-level methods, including end-to-end learning and generative paradigms. Section 4 summarizes datasets and downstream tasks. Section 5 discusses evaluation metrics and standardization for low altitude scenarios. Section 6 presents the future outlook, and Section 7 provides the conclusion.

2 Task Hierarchy and Core Challenges in Low Altitude Cross-Modal Registration

2.1 Registration Categories and Key Challenges

Multimodal image registration aims to establish consistent spatial or semantic correspondences across heterogeneous imaging modalities, enabling comparable, fusible, or jointly interpretable representations of the same scene under distinct sensing mechanisms. Compared with single modality registration, multimodal registration must address not only common

factors such as viewpoint and scale changes, occlusions, and noise, but also modality induced inconsistencies, including radiometric discrepancies, texture inconsistency, cross scale and cross dimensional representation gaps, and modality specific noise patterns. As a result, multimodal registration inherently couples geometric reasoning with semantic understanding, and it requires a careful balance between geometric consistency and semantic consistency.

From the perspective of alignment objectives, multimodal alignment can be discussed at the geometric level and at the semantic or representation level. Geometric alignment seeks an explicit spatial mapping across modalities, often formulated as rigid or nonrigid transformations that unify coordinate systems for comparison and fusion. Semantic or representation level alignment further targets structural association and semantic comparability. Even when spatial correspondence is established, large appearance gaps may still prevent effective downstream use, which motivates alignment strategies that emphasize structural consistency, object boundary correspondence, or shared representation spaces. Related recent progress in cross-modal semantic modeling also shows that joint semantic alignment can improve the comparability of heterogeneous representations and provide useful support for downstream tasks [8].

In terms of implementation, multimodal alignment includes both explicit and implicit paradigms. Explicit alignment aims to estimate transformation parameters or dense deformation fields and provides interpretable geometric correspondence. Implicit alignment is often embedded in downstream tasks such as fusion, detection, or tracking, where the model focuses on projecting modality specific features into a common comparable space. Consistency is induced at the representation level through mechanisms such as cross-modal attention, deformable modeling, or contrastive constraints, without necessarily producing an explicit geometric mapping.

In low altitude applications, multimodal registration becomes more challenging due to pronounced viewpoint variation and platform dynamics. Low altitude platforms operate at short ranges with agile motion, and they are susceptible to attitude jitter, heading changes, and environmental disturbances, which amplify discrepancies in temporal synchronization, spatial projection, and field of view among sensors. Without reliable registration, multimodal observations cannot support coherent spatial and semantic understanding, thereby degrading the reliability of key tasks such as object detection, scene interpretation, and situational awareness. Consequently, low altitude multimodal registration emphasizes not only accuracy, but also robustness under complex disturbances and effectiveness in supporting downstream perception.

Based on the above formulation and challenge analysis, multimodal image registration methods can be summarized into two main lines. Feature-level registration learns modality invariant representations and establishes correspondences in the feature space, which is often more robust to radiometric gaps and missing textures and is naturally compatible with implicit alignment in downstream tasks. Pixel-level registration instead targets dense correspondence in the image domain or deformation field estimation. It is preferable when explicit geometric mapping and high-precision registration are required, but its robustness to severe radiometric discrepancies and complex deformation depends more critically on model design and data support.

2.2 Task Specificity of Low Altitude Multimodal Registration

Multimodal registration across different domains shares the common objective of establishing correspondence between heterogeneous observations. In low altitude scenarios, however, this objective is realized under a distinct task configuration shaped by viewpoint-sensitive geometry, scene dynamics, unstable overlap, and platform-level deployment constraints. As a result, low altitude multimodal registration is not merely a supporting step for cross-modal fusion, but a task-critical component that directly affects the reliability of airborne perception systems.

This task specificity is first reflected in the geometric conditions of image formation. Low altitude platforms usually operate at short range and under agile motion, where attitude variation and scene depth discontinuity jointly intensify parallax, scale variation, and local projective distortion. Under such conditions, cross-modal discrepancy is often coupled with strong local geometric instability, so the misalignment cannot be adequately characterized by a single global transformation. In medical image registration, by contrast, geometric inconsistency is more commonly associated with anatomical variation, physiological motion, or inter-scan differences under scanner-based acquisition. In high-altitude remote sensing, although terrain relief and off-nadir effects remain relevant, the much longer imaging distance generally makes rapid local perspective variation less dominant at the image level. Therefore, low altitude multimodal registration places greater demands on local geometric validity and deformation-sensitive correspondence modeling.

The specificity of this task is further amplified by scene dynamics and deployment requirements. Airborne low altitude observations often contain moving objects, vegetation motion,

Table 1 Representative multimodal registration methods and typical applications.

Category	Representative methods	Strengths	Applications
	Feature-level methods		
Feature-based	SIFT, SURF, ORB, PIIFD, PSO-SIFT, LGHD, OS-SIFT, SuperGlue, KAZE, HOSS, RIFT, Matchformer, SE2-LoFTR, ASS, LoFTR, OSS, Xfeat, RIFT2, HOWP, OIM, SOFT, D2-Net, LightGlue, DKM, RoMa	Sparse robust matching	Registration; localization
Template-based	SVD-RANSAC, HOG, NCC, MI, BBS, DDIS, CoTM, CSTM-Net, HOPC, CFOG, MSTM, MOSS, QATM, GFTM, DLSTM, AESF, DCC, SIFNet, DTM	Precise local registration	Template localization; local registration
Downstream task-based	DeformConv, TDRNet, RepPoints, CMT, ProbEn	Task-supervised alignment	Detection
	AttentionFGAN, DSAN, BAPA-Net, IFA, IQSeg, ProCA		Segmentation
	DRL-Net, CMIT, CMTR, CMD-MMD, CAIL, IDKL, IMKA		Re-identification
	mmMOT, CMD, MCSR, AMNet, MRTTrack		Tracking
	Pixel-level methods		
MI-based	MM-MI, EMMA, B-spline, CMI, FNMI, Q-MI, CRE, 3D Harris, SO-MI, SMI, MINE-local, DRMIME, ACO, UDA-MIMA, MIDiffusion, GoA	Unsupervised similarity	Rigid/affine registration
Image-translation-based	UMDIR-LaGAN, RGPT, SbR, SymReg-GAN, RegGAN, Discriminator-free, AAN, TransMorph, NICE-Trans, RFNet, MURF, CAPIT, RFM-GAN, DFMIR, STABLE, OTMorph, IMF, LADDA, SSDF, AU-Net, HR4IR	Modality-gap reduction	Registration; fusion
End-to-end registration	VoxelMorph-diff, RegNet, CIRNet, VoxelMorph, JSSR, ASNet, CoCycleReg, DeepReg, KeyMorph, E2EIR, PAMRFusion, RegSeg, NIR, PFRFusion, MSFDNet, B-SR, C2RF, MulFS-CAP, MIPR, K-CMorph, Hy-CycleAlign	Direct dense registration	Dense nonrigid registration

cast-shadow variation, partial occlusion, and residual temporal mismatch across sensors, all of which weaken structural stability and reduce the reliability of valid cross-modal overlap. In contrast, medical image pairs are usually acquired over the same anatomical region under planned procedures, while high-altitude remote sensing pairs are more often dominated by relatively stable large-scale background structures. At the same time, low altitude multimodal alignment is commonly embedded in downstream tasks such as detection, scene interpretation, tracking, and situational awareness, where alignment quality directly conditions subsequent perception performance. Because onboard computation, latency, and energy budgets also constrain model complexity and inference efficiency, robustness and deployability become tightly coupled requirements in this setting.

Taken together, the specificity of low altitude multimodal registration does not lie in any single difficulty taken in isolation, but in the concurrent presence of viewpoint-sensitive geometry, unstable overlap, dynamic scene interference, and deployment-oriented resource constraints. This specificity is of both methodological and practical significance, as it affects the construction of representative datasets, the validity of modeling assumptions, the choice of robustness-oriented alignment strategies, and the design of evaluation protocols that reflect downstream utility. For this reason, low altitude multimodal registration constitutes an application-critical problem setting that warrants dedicated investigation.

3 Multimodal Image Registration Methods

Image registration methods can be divided into two levels based on how cross-modal correspondence is established, including feature-level registration and pixel-level registration. Feature-level methods construct modality-consistent correspondence in the representation space and are more closely related to implicit alignment in downstream perception tasks. Pixel-level methods focus on explicit spatial correspondence or deformation field estimation and are therefore more suitable for scenarios requiring accurate geometric mapping.

Accordingly, the methods discussed in this section include feature-based, template-based, and downstream task-based approaches at the feature-level, and mutual-information-

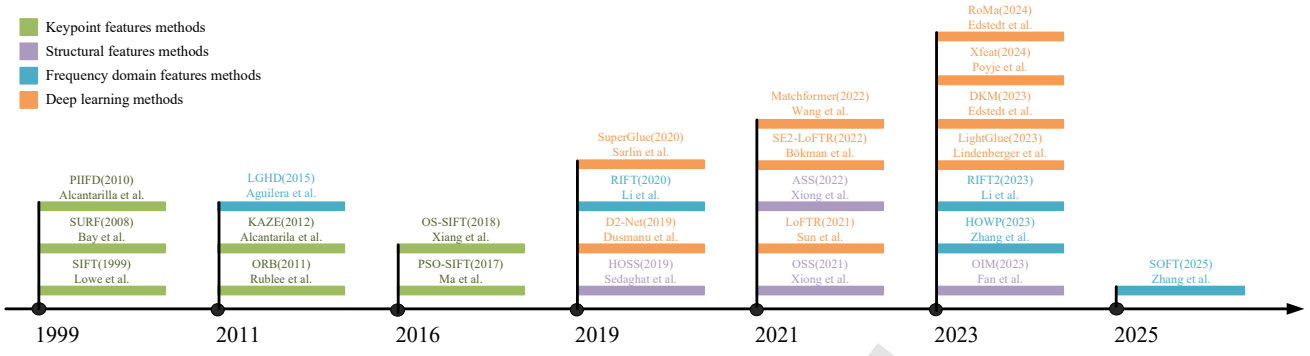

Fig. 2 Feature-level image registration strategies.

based, image-translation-based, and end-to-end registration approaches at the pixel level. These categories differ in correspondence modeling, alignment form, and typical applications in low altitude multimodal settings. Table 1 summarizes the representative studies and their typical applications, while the following subsections discuss these categories in detail.

3.1 Feature-Level Image Registration Strategies

3.1.1 Feature-Based Registration Methods

Feature-based matching is a long-standing backbone of image registration, as shown in Fig. 2. A typical pipeline includes keypoint detection, feature description, cross-image matching, and robust model estimation with outlier rejection. Early methods mainly relied on sparse point features, where salient structures such as corners and edge points are detected and encoded by local descriptors. The SIFT algorithm proposed by Lowe in 1999 established a representative baseline due to its strong scale and rotation invariance [15]. To improve efficiency, SURF introduced integral-image-based acceleration [16], while ORB combined fast detectors with binary descriptors to support real-time applications [17].

In multimodal settings, the modality gap often manifests as radiometric heterogeneity, inconsistent edge and texture responses, and modality-specific noise patterns. These factors reduce descriptor comparability and increase matching ambiguity. To improve robustness, PIIFD models local intensity distributions to mitigate radiometric differences [18], and PSO-SIFT constructs more robust gradient cues to tolerate nonlinear intensity variations [19]. KAZE introduces nonlinear diffusion filtering to build a scale space that preserves salient structures, which can be beneficial under radiometric changes [20]. Beyond generic designs, modality-pair-specific strategies have also been explored. For optical–SAR registration, Xiang et al. constructed a bimodal Harris scale space and used exponentially weighted gradients to seek cross-modal stable features [21]. Despite these improvements, point-feature-based methods still rely heavily on local low-level cues such as intensity and gradients. Under large modality gaps, texture-poor regions, or strong artifacts, reliable sparse correspondences become difficult to obtain, which can degrade matching accuracy and destabilize geometric estimation. This limitation motivates methods that emphasize more stable structural or frequency-domain information.

Structural feature methods exploit geometric shapes, contours, and self-similarity patterns, focusing on the stability of topological relations rather than photometric consistency. They are particularly useful when grayscale or intensity statistics differ substantially across modalities while coarse structural layouts remain relatively stable. Representative approaches include HOSS [22] and OSS [23]. ASS further improves OSS by generating feature maps with local self-similarity descriptors and enhancing rotation tolerance via orientation voting and descriptor rotation [24]. However, structural self-similarity can still be challenged by severe nonlinear radiometric discrepancies and complex clutter. OFM applies multi-directional filtering with additional filter optimizations, which improves robustness to nonlinear intensity changes in complex multimodal imagery [25].

Frequency-domain feature methods perform matching by transforming images into the frequency domain and extracting phase-related information. Phase-based cues tend to be less sensitive to radiometric inconsistencies than raw intensities, and they can be implemented using transforms such as the Fourier transform or Log-Gabor filtering. LGHD uses multi-scale, multi-orientation Log-Gabor filters to extract features [26]. For visible and long-wave infrared matching, its performance can be affected by limitations of the FAST detector, including sensitivity to noise and redundant clustered detections [27]. RIFT introduces a radiation-variation-insensitive feature transform and uses a maximum index map to improve robustness under radiometric discrepancy [28]. Its

rotation handling based on ring-feature computation may increase computational overhead. To address more complex nonrigid deformation, HOWP and SOFT enhance matching via weighted phase-orientation modeling or tensor-based orientation feature maps [29, 30]. These methods are often competitive under severe radiometric discrepancy and can provide accurate displacement estimates in applications such as medical imaging and remote sensing.

With the rise of deep learning, learning modality-robust representations through end-to-end training has become increasingly common. Early attempts replaced handcrafted features with deep features extracted by convolutional neural networks or graph neural networks, but robustness to large modality gaps and matching efficiency were still limiting factors in many settings. Recent learning-based methods can be organized by how they generate correspondences. One line focuses on learning keypoint detection and description. D2-Net unifies detection and description in a single network and improves matching performance across sources [31], while XFeat emphasizes efficient architectures for keypoint extraction [32]. A second line performs matching more directly on feature maps. LoFTR uses attention mechanisms to establish correspondences without an explicit keypoint detection stage and is effective in weak-texture scenes [33]. SE2-LoFTR improves rotation tolerance [34], and MatchFormer integrates feature extraction with similarity learning to streamline the pipeline [35]. A complementary line refines correspondences using matching backbones inspired by graph reasoning. LightGlue, building on SuperGlue, improves runtime while maintaining strong accuracy through self-attention and cross-attention mechanisms [36, 37]. In addition, dense matching methods aim to produce richer correspondences. DKM [38] and RoMa [39] can generate a large number of matches and provide stronger constraints for nonrigid alignment, but they often require higher computation.

In summary, feature-based multimodal registration has evolved from handcrafted descriptors to learning-based matching. Point-feature methods are efficient but can fail under large modality gaps and strong local deformation. Structural and frequency-domain methods are typically more tolerant to radiometric discrepancy, yet they can be sensitive to complex deformation and may struggle under real-time constraints. Learning-based methods improve robustness by learning cross-modal representations, but they depend on training data coverage and can face deployment challenges on resource-limited airborne platforms. These issues are especially pronounced in low altitude scenarios, where large viewpoint changes, platform-induced motion, partial field-of-view overlap, and time synchronization errors jointly increase alignment difficulty. Therefore, improving robustness under the low altitude modality gap while maintaining efficiency and reliability remains an important direction, and Fig. 2 summarizes the methodological evolution and key characteristics of representative approaches.

3.1.2 Template-Based Image Registration Methods

Template based registration aligns images by searching for the most similar region in a target image with respect to a reference template. A typical workflow includes template construction and preprocessing, similarity measure design, hierarchical search to narrow candidates, and refinement around the similarity peak for sub pixel localization. Due to its simplicity and suitability for large area coarse search, template matching is widely used for initial positioning and coarse registration in remote sensing.

Early studies mainly adopted pixel based similarity measures, such as the sum of squared differences, abbreviated as SSD, and the normalized correlation coefficient, abbreviated as NCC, to evaluate match quality [40]. Under cross-modal conditions, these measures can be sensitive to nonlinear radiometric discrepancy, which may lead to ambiguous matches and false peaks. Mutual information, abbreviated as MI, reduces the impact of radiometric mapping changes to some extent, but it is sensitive to the choice of window size and has relatively high computational cost, which limits its practicality in large area search [41]. To improve robustness, researchers proposed more reliable metrics and matching strategies. BBS reduces background interference through a nearest neighbor mechanism [42]. DDIS improves robustness by accounting for template deformation and by using the diversity of deep features [43]. CoTM measures matching error using co occurrence matrix statistics, which reduces direct dependence on color differences [44]. Xiong et al. proposed a method named CSTM-Net, which quantifies similarity between SAR and optical images through spatial search and cosine similarity [45]. Despite these advances, many methods still rely on pixel values or pixel statistics, so performance may degrade under complex radiometric distortion, occlusion, or cluttered backgrounds.

To reduce dependence on pixel-level metrics, research gradually shifted toward region level templates and similarity measures based on structural cues. Structural cues in feature matching are often used to stabilize local keypoint descriptors, whereas structural cues in template matching aim to build region level structured similarity or structural templates for improved cross-modal robustness. The main motivation is that boundaries, orientations, and local shapes are often less

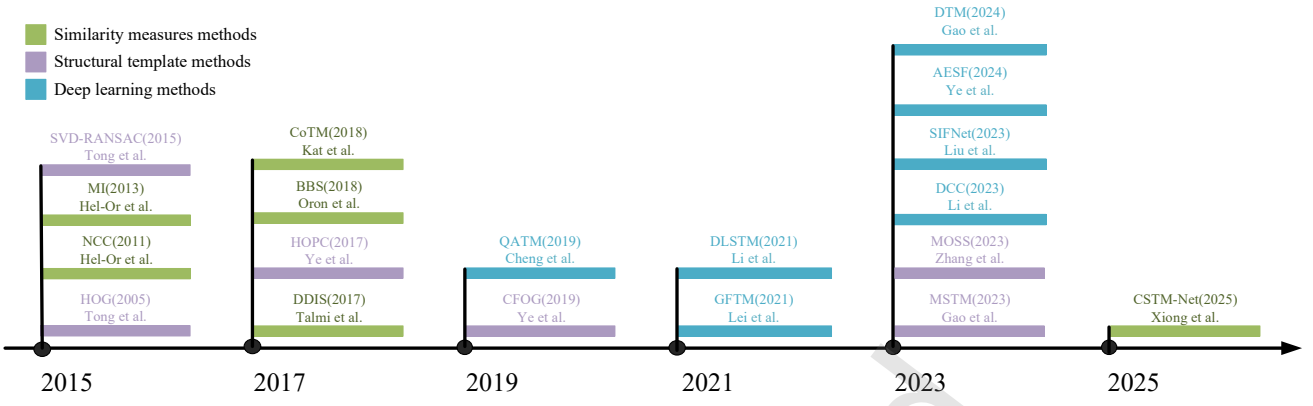

Fig. 3 Template-based image registration methods.

affected by radiometric discrepancy than raw intensity. Tong et al. extracted displacement information via singular value decomposition and improved registration under noise [46]. The HOPC descriptor proposed by Ye et al. combines phase congruency with HOG and constructs template descriptors by statistically characterizing local structural patterns, enabling high precision registration [13, 47]. Its pixel-wise description, however, increases computational cost. CFOG improves speed through frequency domain computation, but it remains sensitive to initial position deviation [48]. MSTM suppresses nonlinear radiometric distortion using frequency domain convolution maps and combines rotation robust omnidirectional aggregation with multi-scale matching for efficient and robust registration [49]. MOSS further uses the self similarity map from coarse matching as a structural template and refines alignment via orientation aggregation [50]. Overall, structural template methods effectively exploit structural cues to mitigate modality gaps, but their performance can still be limited by the expressiveness and cross scene generalization of handcrafted features.

Given the difficulty of covering complex cross-modal variation with handcrafted designs, deep learning provides a more adaptive solution for template matching. QATM models matching quality as a soft ranking score and can be used as a standalone module or embedded into networks [51]. GFTM and DLSTM used deep networks to extract cross-modal features and improved robustness and sub pixel localization through multi loss optimization [52, 53]. Recent studies further introduced attention mechanisms and contrastive constraints. AESF adopts multi branch global attention to enhance structural representations and uses a multi crop matching loss to exploit both global and local information [54]. DCC introduced a dense consistent InfoNCE contrastive loss to improve fine grained feature discriminability and suppress overfitting, achieving strong results in cross-modal matching [55]. More recent work integrates template matching into end-to-end architectures. SIFNet uses self attention for multi-scale feature fusion and formulates registration as a regression task [56]. Gao et al. proposed a method named DTM, which combines edge aware modules with Transformer based structural cues and achieves sub pixel registration via differentiable hierarchical optimization [57]. Overall, learning-based template matching improves cross-modal adaptability, but it depends on training data coverage and computational resources, and interpretability remains an important concern for applications that require high reliability.

In summary, template matching has evolved from single pixel-level metrics to structurally enhanced templates and further to learning-based end-to-end adaptation. In low altitude scenarios, these methods are often more suitable for coarse registration or initialization. Low altitude imagery commonly exhibits strong parallax and platform induced motion that lead to local deformation. It also suffers from inconsistent sensor coverage and dynamic occlusion, which can violate the fixed template assumption. Structural templates may face structural confusion in cluttered backgrounds, while learning-based methods can be constrained by limited data and onboard computation budgets. Fig. 3 summarizes the major development stages, representative works, and their key characteristics and limitations.

3.1.3 Downstream Task-Based Image Registration Methods

Unlike explicit registration, implicit alignment is usually not implemented as an independent registration pipeline. Instead, it is embedded as an intrinsic mechanism within downstream tasks. The key idea is to drive cross modality consistency in space or in representation by optimizing task objectives or self supervised constraints. For this reason, implicit alignment is best summarized from the perspective

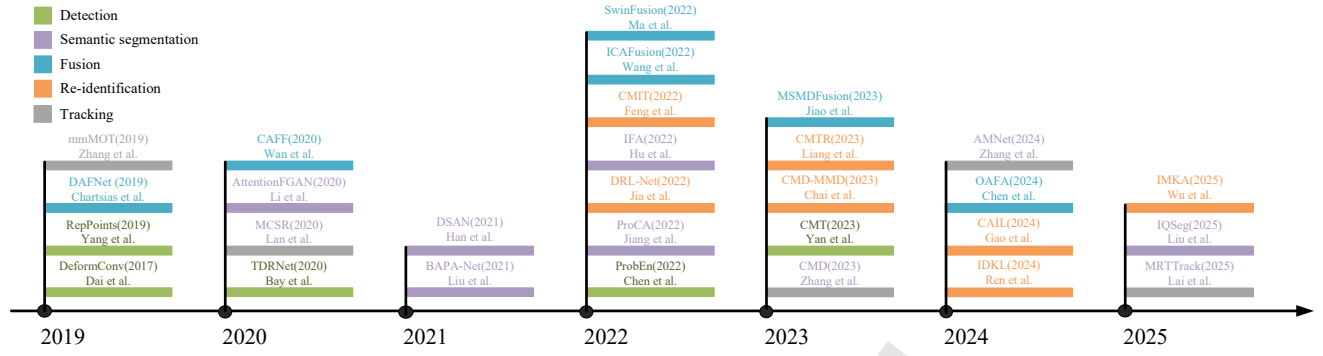

Fig. 4 Downstream task-based image registration methods.

of application tasks, including detection, segmentation, re-identification, and tracking.

In object detection, cross modality alignment is critical for improving object saliency and boundary clarity in low altitude scenarios, which directly affects localization and classification performance [58]. Early solutions often employed deformable convolutional networks, abbreviated as DeformConv, to compensate local geometric offsets by learning sampling displacements [59]. Representative works such as TDRNet and RepPoints learn local offsets so that sampling points adapt to object shapes, reducing pixel-level misalignment [60, 61]. These methods are computationally efficient and easy to integrate, but they mainly perform local correction and can be limited under large global displacement or rotation. More recently, attention based models provide stronger global interaction for implicit alignment. CMT fuses multimodal features into unified tokens and aligns them in three dimensional space, enabling global interaction and improving robustness to missing modalities, while requiring higher computation [62]. ProbEn further shows that the focus of implicit alignment is expanding from spatial alignment to temporal consistency in dynamic scenes [63].

In semantic segmentation, implicit alignment primarily aims to reduce pixel-level spatial offsets and semantic drift between multimodal features. Early work focused on unsupervised domain adaptation, abbreviated as UDA, to align feature distributions between source and target domains. DSAN achieved cross-modal medical image segmentation through symmetric or bidirectional feature alignment, showing that implicit alignment can be effective under limited annotation [64]. Later studies moved toward finer grained alignment. BAPA-Net introduced boundary adaptation and prototype alignment to address missing boundary cues and class level distribution drift [65]. Hu et al. proposed an implicit feature alignment function, abbreviated as IFA, to aggregate multi level feature maps efficiently for arbitrary resolution segmentation [66]. For visible and thermal segmentation, Liu et al. proposed IQSeg, a deformable alignment module to implicitly align spatial features across modalities [67]. With the rise of unsupervised and self supervised learning, implicit alignment has improved in generality. Jiang et al. combined contrastive learning with domain adaptation and reduced prototype level discrepancies to align cross domain features implicitly [68].

In cross-modal data fusion, implicit alignment is usually achieved during fusion through architectural design or feature learning. CAFF and AttentionFGAN use cross-modal attention or generative adversarial networks to align features and fuse complementary information, improving diagnosis or recognition performance [69, 70]. Fusion is also commonly used as an upstream component for detection. OAFA and MSMDFusion aggregate LiDAR and camera features into a unified representation space through spatial alignment, which improves three-dimensional detection accuracy with the aid of gated convolutions [58, 71]. DAFNet learned modality invariant representations by decoupling anatomical structure from modality related factors, enabling alignment and fusion across modalities [72]. To handle style discrepancy, ICAFusion and SwinFusion are used to align fusion features and maintain style consistency [73, 74].

In multimodal object re-identification, abbreviated as Re ID, the key challenge is to obtain feature consistency and discriminability across modalities without strict geometric registration. Existing methods often combine feature decoupling, distribution alignment, and cross modality interaction for implicit alignment. DRL-Net adopts Transformer architectures with object queries and contrastive learning, and improves robustness under occlusion through local semantic reasoning [75]. IMKA generates implicit modality data and aligns their distributions, and incorporates uncertainty modeling to stabilize retrieval [76]. Additionally, IDKL exploits modality specific cues by purifying and injecting them into shared representations to enhance cross-modal embeddings

[77]. CAIL reduces modality discrepancy and improves intra class compactness via multi level channel fusion and modality center alignment losses [78]. Distribution level alignment is another important line. TransVI introduces CMD-MMD constraints to improve cross-modal consistency at the distance distribution level [79]. CMTR encodes modality features using modality embeddings and modality aware enhancement losses [80]. CMIT uses cross modality attention and modality discrimination losses to strengthen modality invariant representations [81]. Across these designs, task objectives encourage feature space consistency and thus achieve implicit alignment.

In multimodal object tracking, naive concatenation or fusion often yields inconsistent representations, so recent work emphasizes feature-level implicit alignment and cross modality association. The mmMOT framework proposed by Zhang et al. adopts a sensor agnostic design that represents each modality independently and achieves alignment in a multimodal adjacency estimator, improving association robustness [82]. For visible and thermal tracking, CMD transfers modality specific and shared knowledge from a dual stream network to a lightweight single stream network, thereby achieving implicit feature alignment [83]. AMNet introduces mutual interaction spatial alignment modules and information matching fusion modules to maintain spatial consistency [84]. Lan et al. employed MCSR, sparse representation and modality correlation modeling so that features from different modalities converge in a unified space [85]. Lai et al. proposed MRTTrack and combined separation collaboration design with cross modality discrepancy constraints within a Transformer framework, mitigating offsets caused by modality gaps and improving synergistic tracking [86]. Although technical choices differ, these methods share the goal of consistent cross modality representations.

In summary, implicit alignment is a cross modality consistency modeling strategy that does not require explicit registration labels, and it has been widely adopted across detection, segmentation, fusion, re-identification, and tracking. Common mechanisms include structural modules such as deformable sampling and attention interaction, as well as optimization objectives such as contrastive constraints and distribution alignment losses. These mechanisms capture geometric and distributional deviations across modalities at the feature-level and realize spatial and semantic consistency within downstream tasks. Different tasks emphasize different aspects. Detection and segmentation highlight boundary quality and semantic consistency. Fusion prioritizes complementarity and style consistency. Re-identification and tracking emphasize discriminability and temporal consistency. Overall, current research indicates a trend of expanding cross-modal alignment from spatial modeling toward semantic and temporal modeling.

3.1.4 Low Altitude Adaptability Analysis of Feature-Level Registration Methods

Feature-level image registration methods establish cross-modal correspondences by constructing local or global features with modal robustness and thus occupy a central position in multimodal image registration research. However, when these methods are directly applied to low altitude scenarios, significant discrepancies arise between their theoretical assumptions and the actual imaging environment. These discrepancies reveal a series of severe adaptability issues which are primarily manifested in four dimensions including geometric viewpoint, environmental interference, computational resources, and model generalization.

The first challenge arises from geometric instability and viewpoint disparity. Low altitude platforms typically rely on unmanned aerial vehicles namely unmanned aerial vehicles (UAVs) or light aircraft for data acquisition characterized by low imaging altitudes and frequent viewpoint changes. Attitude jitter and heading adjustments during flight often introduce significant scale variations, perspective distortions, and local non-rigid deformations. This characteristic directly challenges point-based feature methods that rely on local geometric stability. Although SIFT, ORB, and their cross-modal variants perform stably under medium-to-high altitude or near-nadir imaging conditions, the projection differences of 3D terrain structures are significantly amplified in low altitude environments. Such differences lead to a sharp decline in the repeatability and geometric consistency of local features and consequently render feature matching highly susceptible to viewpoint variations and background interference.

The second challenge is the interference introduced by complex environments and dynamic objects. Low altitude scenes are often characterized by complex backgrounds and drastic scale variations accompanied by numerous dynamic targets such as vehicles, pedestrians, and rotating blades. This poses new difficulties for feature methods based on structure or frequency domain. Structural features and self-similarity features possess strong cross-modal robustness under the premise of relatively stable scene structures. However, under low altitude rapid imaging conditions, frequent changes in object occlusion and the destruction of local structures by dynamic targets or incompleteness tend to weaken the discriminative power of structural consistency. While frequency

domain feature methods are insensitive to radiometric differences, they are highly dependent on noise patterns and imaging stability. Consequently, their matching stability is severely limited in low altitude data plagued by strong motion blur, sensor jitter, or complex scattering effects.

The third challenge is the contradiction between computational efficiency and resource constraints. Low altitude multimodal perception often requires real-time operation under limited computing power and energy consumption constraints which places extremely high demands on the computational efficiency of feature-level registration methods. Traditional frequency domain methods and certain structural feature methods typically involve complex filtering, transformations, or multi-scale calculations. Meanwhile, dense feature matching and deep learning methods offer richer correspondences but are often accompanied by prohibitive computational and storage overheads. This dichotomy between precision and efficiency is particularly pronounced on resource-constrained low altitude embedded platforms and directly impedes the deployment of many high-precision feature-level methods in practical systems.

The final challenge involves data distribution and model trustworthiness bottlenecks. For deep learning-based feature alignment methods, adaptability in low altitude scenarios is severely constrained by training data distribution and model generalization capabilities. Most existing deep matching models are trained on standard datasets or relatively standardized remote sensing and natural images. In contrast, low altitude scenario data are characterized by variable viewpoints, diverse modal combinations, and high annotation costs. This results in training samples that fail to cover the complex long-tail cases found in real-world applications. Furthermore, deep models generally suffer from a lack of interpretability. This makes it difficult to effectively evaluate and verify the reliability of registration results in high-stakes applications such as low altitude emergency monitoring and security patrols.

In summary, feature-level image registration methods in low altitude scenarios face not a single technical bottleneck but a systemic obstacle composed of multiple factors including imaging geometric instability, dynamic environmental interference, real-time and computing power constraints, and insufficient data and model generalization capabilities. This indicates that existing feature-level methods cannot simply be directly migrated to effectively adapt to low altitude multimodal perception tasks. Instead, they require deep integration with motion modeling, temporal constraints, task-driven implicit alignment mechanisms, or lightweight network architectures.

3.2 Pixel-Level Image Registration Strategies

According to the level at which cross-modal correspondence is established, multimodal alignment methods for low altitude platforms can be broadly organized into two methodological lines, namely feature-level registration and pixel-level registration. Feature-level registration emphasizes the construction of modality-robust correspondences in representation space and is mainly reflected in feature-based methods, template-based methods, and downstream task-coupled implicit alignment. Pixel-level registration focuses on explicit spatial correspondence in the image domain and mainly includes mutual-information-based methods, image-translation-based methods, and end-to-end registration methods.

These method families differ not only in implementation form, but also in the assumptions they make about radiometric consistency, geometric deformation, overlap stability, and deployment conditions. Feature-level methods are generally more tolerant to modality discrepancy and are more naturally compatible with downstream perception pipelines, whereas pixel-level methods are more suitable when explicit geometric mapping and dense alignment are required. In low altitude scenarios, where viewpoint variation, dynamic interference, partial overlap, and resource constraints are often coupled, the practical effectiveness of each method family depends on the specific balance among robustness, geometric precision, and deployability. The following subsections review representative methods under this taxonomy, and Table 1 summarizes their main characteristics and typical applicability.

3.2.1 Mutual-Information-Based Registration Methods

Mutual information, abbreviated as MI, is an information theoretic measure that quantifies statistical dependence between two random variables. Unlike similarity measures that assume photometric consistency, MI builds an unsupervised alignment criterion by modeling joint statistical dependency between images. As a result, MI has become a widely used similarity measure for multimodal registration and has been applied in medical imaging, remote sensing, and industrial vision [87, 88]. MI based registration typically estimates transformation parameters by maximizing the mutual information between a reference image and a moving image, which yields geometric registration.

The introduction of MI to image registration dates back to the mid 1990s. Collignon et al. and Viola et al. proposed MI based approaches for automatic rigid registration and demonstrated effective registration for multimodal medical imagery such as CT with MRI and PET with CT [89, 90]. Wells et al. further developed an automated framework that iteratively

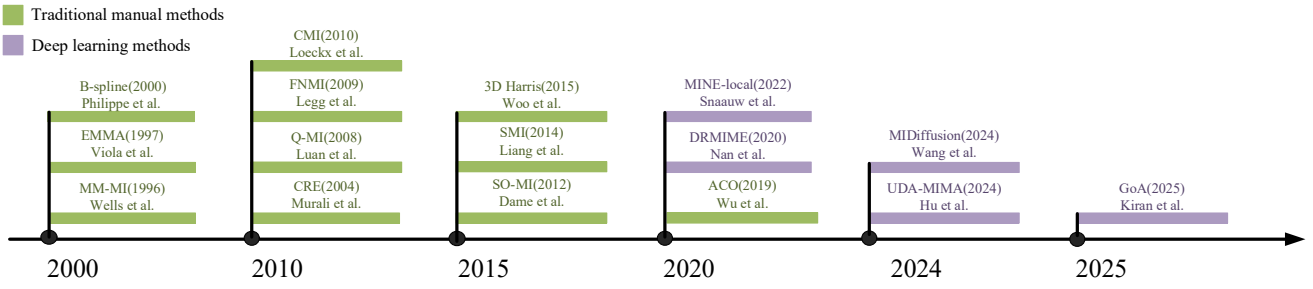

Fig. 5 Mutual-information-based image registration methods.

updates pose parameters until MI is maximized [91]. Because this information theoretic formulation makes fewer assumptions about the imaging process, it is often more robust than correlation based measures across devices and modalities. Tsao et al. analyzed the impact of interpolation artifacts on MI registration and motivated subsequent improvements in numerical stability and implementation practice [92].

Standard MI still has limitations, including sensitivity to local intensity variations, limited spatial priors, and strong dependence on estimation and optimization details. For metric enhancement and prior modeling, Luan et al. proposed a quantitative qualitative MI measure that combines region and structure cues with grayscale statistics to address the limitations of purely intensity based MI [93]. Loeckx et al. introduced conditional mutual information, abbreviated as CMI, and treated spatial position as a conditional variable to model local statistical dependency, which is useful for non-rigid registration [94]. Woo et al. and Legg et al. embedded geometric constraints into the MI objective by incorporating three dimensional Harris cues or feature neighborhood information, which improves local consistency in complex scenes [95, 96]. For optimization and numerical computation, Dame and Marchand accelerated gradient and Hessian evaluation using second order Taylor approximation [97]. Liang et al. and Wu et al. employed heuristic strategies such as ant colony optimization to mitigate nonconvexity and improve global search behavior [88, 98]. In addition, joint histogram estimation, binning strategies, interpolation schemes, and multiresolution search can substantially affect the stability and accuracy of MI in practice.

In the deep learning era, MI has gradually been used not only as a direct objective, but also as a differentiable loss term or an alignment module within trainable frameworks. Hu et al. introduced an MI maximization module in an unsupervised domain adaptation framework to learn domain invariant features and enable cross domain knowledge transfer [99]. Wang et al. proposed a differentiable local MI layer to constrain an iterative denoising process and provide guidance through statistical cross-modal consistency without relying on an explicit modality mapping [100]. These studies indicate that MI can serve as a regularization term that complements representation learning and generative models by enforcing cross-modal feature consistency.

Despite its long standing success in medical imaging and remote sensing, purely MI driven registration has become less common in recent years. One reason is that low altitude multimodal data often exhibit limited overlap and field of view mismatch, dynamic occlusion and increased noise, and resolution imbalance with local deformation. These factors degrade the reliability of joint statistics and make MI optimization more prone to local optima and unstable convergence. Meanwhile, learning-based methods provide end-to-end feature modeling and multiscale context fusion. Therefore, MI is more frequently used as an auxiliary loss or an evaluation metric rather than the sole driving objective. Fig. 5 summarizes the development trajectory and representative directions of MI based multimodal registration.

3.2.2 Image-translation-based Registration Methods

Image translation aims to learn a correspondence between the source domain and the target domain. In image registration, its main role is to reduce nonlinear appearance variations induced by modality differences while preserving structural cues in medical images that are critical for accurate geometric alignment [101]. Translation-based methods for multimodal registration typically follow a modality normalization paradigm. A generative model first synthesizes the moving image so that its appearance is consistent with the fixed image. The subsequent registration step then estimates the transformation using similarity measures defined within a single-modality setting. This two-stage design effectively reformulates multimodal registration as unimodal registration, thereby alleviating the optimization difficulty caused by cross modality intensity inconsistencies and improving the robustness of similarity-driven alignment.

Early research predominantly adopted generative adversarial networks (GANs) as translation modules, employing

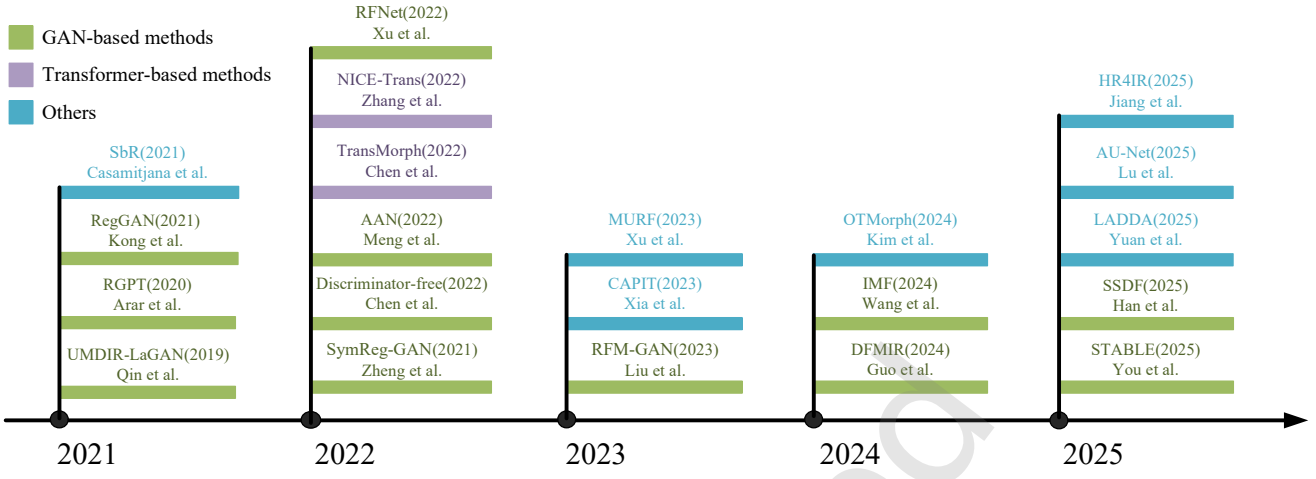

Fig. 6 Image translation registration methods.

adversarial training to learn cross-modal appearance transfer [102]. However, optimizing primarily for visual fidelity can lead to spatial and structural inconsistencies, producing geometry-inconsistent artifacts that deviate from the original spatial structure. To address this, subsequent studies have incorporated geometric consistency constraints into the translation process. SymReg-GAN [103] utilizes cycle-consistency to suppress structural drift, while cGAN [104] and RegGAN [105] treat correspondence errors under weakly paired conditions as observation noise, introducing auxiliary registration modules to constrain generator training and reduce misalignment between synthesized and target structures. These methods regularize the translation network through explicit geometric constraints, aiming to ensure the physical plausibility of the generated pseudo-images. Addressing the risk of structural misalignment in repetitive texture regions, AAN [106] further introduces edge-gradient constraints and local patch contrastive mechanisms to mitigate propagation errors in fine-grained structures.

The evolution of network architectures also reflects a shift from generation quality to structural fidelity. INNReg [107] introduces Invertible Neural Networks, leveraging their bijective properties to minimize information loss during cross-domain transformation and ensure strict topological correspondence between input and output. This pursuit of structural fidelity is particularly critical in medical image registration, which is highly sensitive to density statistics and local anatomical consistency. The semi-supervised framework proposed by Han et al. [108] significantly reduces geometric distortion caused by translation through structure-preserving mapping, indicating that high-quality pseudo-image generation is crucial for enhancing the reliability of multimodal registration.

Recent research has further leveraged Diffusion Models and Optimal Transport theory to reinforce structural priors. Yuan et al. [109] achieved explicit disentanglement of anatomical information and modal style based on Latent Diffusion Models, ensuring the adaptability of translation results for downstream registration via frozen attention mechanisms. More recently, AU-Net incorporates DDPM-based bidirectional cross-modal translation as supervision in a unified registration–fusion framework, supporting joint learning of alignment and fusion [110]. Similarly, HR4IR builds a harmonized domain through invertible cross-modal translation and combines coarse-to-fine deformation correction with alternate search to improve infrared-visible alignment [111]. Meanwhile, Kim et al. [112] introduced Neural Optimal Transport to achieve precise statistical distribution alignment at the domain level. These methods are no longer limited to pixel-level mapping but strive to achieve modality unification at the feature distribution level, thereby providing more robust matching references than traditional GANs in scenarios characterized by drastic illumination changes, such as low altitude remote sensing.

In addition, translation has been extended to collaborative learning frameworks and challenging natural scenes. Xu et al. [113, 114] employed a translation network as a coarse registration stage combined with an affine estimation network, which accelerated convergence speed while enhancing overall registration precision. Xia et al. [115] explored the use of image translation to normalize sensor inputs under adverse weather conditions by converting them into standard visibility representations to improve downstream tasks including semantic segmentation, depth estimation, and localization. Their training scheme based on coarsely aligned image pairs suggests that effective image translation can enhance the performance of subsequent perception and alignment tasks

even in the absence of strictly paired data. This observation further supports the transferability of domain normalization and structural-prior ideas to low altitude imagery, where radiometric distortions and imperfect pairing are common.

In summary, translation-based multimodal registration has evolved from GAN-driven appearance transfer aimed at enabling unimodal registration to structure-centric paradigms that suppress translation-induced drift via explicit geometric constraints and consistency losses. More recently, diffusion-based generation and neural optimal transport have further enhanced the physical plausibility and statistical alignment of cross-modal mappings. Although these innovations originated largely outside low altitude settings, their core concepts regarding domain normalization and structural priors are particularly transferable to mitigating radiometric distortions in low altitude imagery, such as drastic illumination and appearance variations. At the same time, translation by itself does not address core geometric difficulties, including large viewpoint changes and occlusion. When such factors dominate, sequential translation and registration pipelines can suffer from error accumulation that degrades alignment quality. A promising direction is to couple translation and registration within a unified differentiable framework that supports joint optimization, thereby improving geometric fidelity and overall robustness.

3.2.3 End-to-End Registration Methods

End-to-end image registration uses deep neural networks to learn a direct mapping from an image pair to a spatial transformation. This paradigm integrates feature extraction, correspondence estimation and transformation optimisation into a single feed-forward inference procedure. It improves inference efficiency and facilitates practical deployment, and has become an important research direction in recent years. Figure 7 summarises the temporal evolution of the end-to-end registration methods reviewed in this section, and shows the progression from early convolutional neural network based unsupervised frameworks to methods that incorporate distribution level constraints, semantic guidance and explicit refinement schemes.

Early work such as VoxelMorph [116, 117] established an unsupervised framework that learns dense deformation fields by combining image similarity objectives with deformation regularisation, without requiring explicit correspondences. However, predicting a dense deformation field with many degrees of freedom from intensity information alone is often ill posed. Ambiguity is severe in regions characterised by weak texture, repeated structures and occlusions, because the available image evidence is insufficient to determine correspondences in a unique way. These conditions can lead to foldings of the deformation field, reflected by regions with non-positive Jacobian determinant values, as well as unrealistic distortions near structural boundaries and reduced generalisation when there is a shift between training and test domains, for example due to changes in acquisition protocol or modality. Subsequent studies have mainly followed two complementary routes. The first route introduces explicit priors and consistency constraints to restrict the solution space and to promote deformations that are physically meaningful. The second route integrates semantic information and improves network architectures to better model cross-scale and cross modality correspondences.

To reduce ambiguity in direct prediction, one major line of work strengthens regularisation and introduces distribution level constraints. Bidirectional consistency has been widely enforced through cycle consistency losses. CIRNet [118] promoted approximate inverse consistency between forward and backward deformations. Lian et al. [119] proposed CoCycleReg, which couples registration with image translation in a cyclic framework and improves robustness through complementary constraints across tasks. In multimodal scenarios where intensity distributions differ substantially, ASNet [120] and PAMRFuse [121] used adversarial learning to align registered outputs with the target modality at the distribution level. These approaches may introduce training instability and discriminator bias, and may also encourage alignment in appearance space that is not accompanied by correct geometric correspondence. For this reason, adversarial objectives are commonly combined with geometric consistency constraints, regularisation that favours invertible mappings, or uncertainty estimation mechanisms in order to improve overall reliability.

A second line of research focuses on introducing higher level semantic information and refining network architectures. The aim is to reduce the dependence on low level intensity cues, which can be unreliable under large deformations and in weak texture regions. DeepReg [122] and RegSeg [123] integrated anatomical labels or segmentation networks into the registration pipeline and provided structural supervision through label overlap constraints or semantic feature consistency. Liu et al. [124] further developed this idea in the JSSR system through joint optimisation of image synthesis, registration and segmentation, which strengthens the interaction between representation learning and geometric alignment. Architectural innovations also address limitations of convolutional networks in modelling long range dependencies. E2EIR [125] introduced self-attention to improve global con-

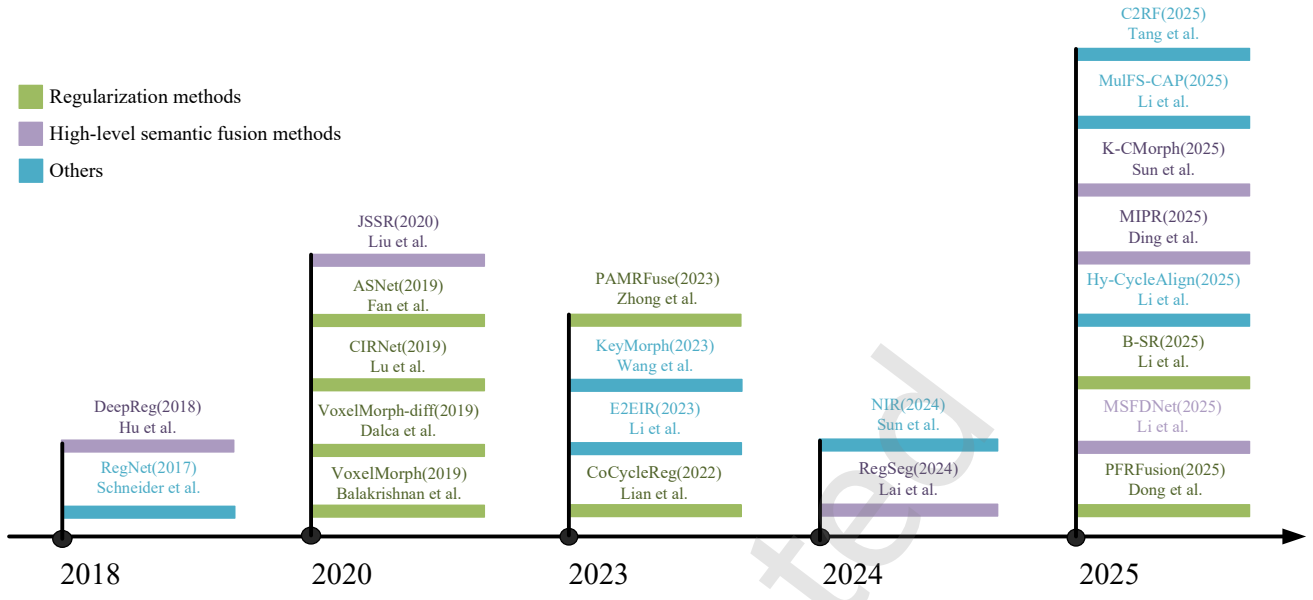

Fig. 7 End-to-end registration methods.

text modelling for large-scale multimodal alignment in remote sensing imagery. NIR [126] employed neural fields to model deformation with implicit continuous functions. KeyMorph [127] incorporated sparse and interpretable control through keypoint prediction, providing geometric anchors in addition to dense deformation estimates. MulFS-CAP performs single-stage fusion for unregistered inputs by modeling cross-modality alignment perception during fusion, alleviating the dependence on explicit pre-registration [128]. Meanwhile, C2RF connects registration and fusion through commonality mining and fusion-guided contrastive learning, enabling their joint optimization in a mutually reinforcing manner [129].

Consistent with the temporal pattern in Figure 7, recent end-to-end multimodal registration methods have evolved from single-stage prediction pipelines towards frameworks that explicitly incorporate constraints and multi-step refinement at the architectural level. Representation disentanglement methods seek to separate modality-invariant structural information from modality-specific appearance, as in MSFDNet [130] and PFRFusion [131]. Multiscale progressive strategies improve robustness by first correcting global geometric discrepancies and then refining local correspondences, as in MIPR [132] and K-CMorph [133]. Feedback based designs provide additional refinement steps during inference, including bidirectional constraints [134] and residual feedback attention [135]. In parallel, Hy-CycleAlign [136] introduces the first hyperbolic space based multimodal image registration framework, extending multimodal alignment from Euclidean space to hyperbolic space to better capture hierarchical and cross-modal structural relationships. Together, these studies introduce explicit consistency conditions and repeatable refinement mechanisms that improve stability under various perturbations.

These developments provide useful guidance for registration in low altitude scenarios. Terrain-induced parallax, rapid changes in viewpoint and scale, and disturbances such as platform jitter, motion blur and occlusions increase correspondence ambiguity and reduce the reliability of single-scale alignment. Multiscale progressive registration is therefore well suited to first correct global geometric discrepancies and then refine local structure. Consistency constraints and feedback mechanisms help suppress unreliable local matches and adjust deformation estimates within multi-step refinement procedures during training or inference. Robust registration under these conditions supports applications that require high spatio-temporal consistency, including disaster emergency monitoring and fine-grained land cover change detection.

3.2.4 *Low Altitude Applicability of Pixel-Level Registration Strategies*

Pixel-level registration establishes cross modality correspondences in the intensity domain or through intensity induced statistical distributions. It uses raw intensities and their statistical relationships as primary constraints and avoids reliance on hand crafted geometric descriptors or explicitly defined semantic entities. While widely employed in multimodal medical imaging and satellite remote sensing, where imaging conditions remain comparatively controlled, their efficacy diminishes in low altitude scenarios. When transferred to such dynamic platforms, the inherent photometric constancy

assumption often fails due to complex radiometric variations and large-scale geometric distortions, thereby severely limiting robustness and generalization capabilities.

Low altitude imagery is characterized by significant viewpoint and scale variations, accompanied by perspective distortion and pronounced terrain relief. Although mutual information and related statistical similarity measures are theoretically robust to cross modality radiometric differences, their practical effectiveness relies on the estimation of a joint distribution that remains sufficiently stationary over adequate spatial support. In low altitude scenarios, limited spatial support combined with strong spatial heterogeneity and a high fraction of dynamic objects compromises the stability of local statistics. Moreover, the compounding effects of occlusion, dynamic motion, and sensor noise severely compromise the integrity of distribution estimation, thereby inducing a highly multimodal objective landscape. This increases the susceptibility to convergence towards incorrect local optima, particularly under conditions of drastic viewpoint changes or non-rigid deformations.

Translation based pixel-level registration faces a persistent tension between structural preservation and geometric consistency. Even structure aware translation models rely on priors learned from training data. Distribution shifts may arise from changes in flight altitude, trajectory, platform attitude, or sensor configuration. Under such shifts, translated target modality surrogates can exhibit local structural distortions or semantic misplacement. These artifacts increase geometric uncertainty and may propagate to the subsequent registration stage. This issue is accentuated in unmanned aerial vehicle remote sensing, where distribution gaps across missions and acquisition campaigns are often substantial.

End-to-end pixel-level registration offers efficient inference, but its low altitude applicability depends strongly on training coverage and regularisation design. Low altitude data combine global geometric variation with local nonrigid effects and dynamic perturbations, which increases the ill posedness of direct deformation prediction. Without sufficiently strong geometric constraints or semantic priors, learned deformation patterns may be unstable under real conditions and may lack geometric validity. Practical deployment is further constrained by onboard computation and energy budgets.

Overall, the dominant limitations in low altitude environments reflect the concurrent degradation of statistical stability, structural consistency, and geometric validity. Mutual-information-based objectives rely on stationarity that is difficult to satisfy in dynamic heterogeneous scenes. Translation-based methods can reduce modality gaps but may introduce geometric artifacts. End-to-end models offer high expressiveness yet remain sensitive to data distributions and regularization design. Consequently, robust low altitude multimodal registration necessitates unified formulations that couple pixel-level constraints with multi-scale modeling, structural or semantic priors, and task-driven constraints. Such integration is essential to enforce geometric consistency and semantic stability in a manner that aligns with the challenging operating conditions of low altitude platforms.

4 Datasets

Multimodal datasets for low-altitude vision can be divided according to cross-modal alignment quality, including registered datasets and coarsely aligned datasets. Registered datasets provide relatively accurate spatial correspondence and are therefore more suitable for evaluating geometric alignment and registration-oriented fusion methods, while coarsely aligned datasets more closely reflect practical UAV sensing conditions, where residual parallax, local deformation, and unstable correspondence are often unavoidable. These two categories differ in acquisition difficulty, annotation cost, alignment precision, and typical usage, and together define the current data foundation for low-altitude multimodal registration research. Table 2 summarizes the representative datasets, their acquisition scenarios, and primary applications, while the following subsections discuss these two categories in detail.

4.1 Low Altitude Registered Multimodal Datasets

Low altitude registered multimodal datasets provide pixel-level geometric ground truth, which is essential for evaluating alignment errors and fusion algorithms. Building such datasets at scale in low altitude scenarios is constrained by three factors, namely platform dynamics, sensor heterogeneity, and annotation cost. On the UAV side, rapid attitude changes and platform motion, together with residual inter-sensor timing offsets, introduce nonlinear misalignment between modalities. On the sensing side, visible and thermal infrared cameras often differ in field of view, spatial resolution, and lens distortion characteristics, making accurate registration difficult and typically requiring hardware-level synchronization as well as careful geometric calibration and post-processing. In addition, producing or validating pixel-accurate correspondences often relies on manual annotation or manual quality control, which incurs substantial time and financial cost and further limits dataset scale and category diversity.

Although high-quality registered datasets such as LLVIP [137] and MSRS [138] have been developed in general vision,

Table 2 Representative datasets discussed in this review, their acquisition scenarios, and primary applications. RGB denotes red-green-blue images, T denotes thermal images, HSI denotes hyperspectral images, D denotes depth maps, and 3D denotes 3D point clouds.

Dataset	Modality type	Acquisition scenario	Low-altitude	Alignment	Primary applications
LLVIP	RGB-T	Surveillance	No	Yes	Fusion and object detection
MSRS	RGB-T	Road scenes	No	Yes	Fusion and semantic segmentation
M3FD	RGB-T	Mixed scenes	No	Yes	Fusion and object detection
VIFB	RGB-T	Mixed scenes	No	Yes	Fusion
HyKo	3D-HSI	Autonomous driving	No	-	Tracking
SensatUrban	3D-RGB	Aerial scenes	Yes	-	Semantic segmentation
UAVid-3D-Scenes	RGB-D	Aerial scenes	Yes	-	Depth estimation
DroneVehicle	RGB-T	Aerial scenes	Yes	No	Vehicle detection
UAVmatch	RGB-T	Aerial scenes	Yes	Yes	Registration and fusion

their acquisition settings differ substantially from low altitude UAV scenarios. These datasets are commonly collected from fixed or near-static ground-based viewpoints with limited camera motion and modest viewpoint change, which reduces parallax and facilitates pixel-level registration, and in this setting resolutions up to 1280×1024. In contrast, low altitude aerial imagery is characterized by pronounced depth variation and viewpoint changes, where terrain relief and scene depth discontinuities induce large parallax, and camera motion introduces strong perspective effects. As a result, models trained or evaluated on ground-view registered datasets often exhibit limited robustness when transferred to low altitude tasks, since they have not been exposed to the geometric deformation patterns that dominate UAV-based multimodal alignment.

Native registered datasets for low altitude UAV scenarios remain scarce. Existing attempts often improve local registration accuracy by reducing the effective field of view, typically through center cropping, to avoid regions with large viewpoint-induced misalignment. For example, the UAVmatch benchmark constructed by Gao et al.[139] derives pixel-level ground truth from originally unregistered sources such as DroneVehicle [140], but the resulting resolution is limited to 480×480, which is substantially lower than the resolution ground-view benchmarks. This reduction is closely tied to the difficulty of low altitude registration. To obtain pixel-level alignment under strong geometric variation, it is often necessary to discard peripheral areas where alignment errors become dominant. The resulting tradeoff between field of view and alignment accuracy reduces the amount of large-scale structural context preserved in the data, and highlights the absence of systematic benchmarks that jointly provide high resolution and high-precision registration. Consequently, the availability of high-quality registered data remains a key bottleneck for progress in this area.

4.2 Low Altitude Coarsely Aligned Multimodal Datasets

Low altitude multimodal datasets are often only coarsely aligned after applying approximate geometric constraints and preliminary extrinsic calibration. They do not provide pixel-level correspondences and typically retain noticeable residual misalignment and image deformation, mainly due to imperfect inter-sensor time synchronization, high-frequency platform vibration, and rapid attitude changes during flight. In many multimodal methods, these residual errors are implicitly ignored and the inputs are treated as if they were well aligned. This mismatch between the data assumption and the actual sensing condition can be harmful, because residual parallax and local distortion reduce cross-modal consistency and introduce implicit noise when transferring supervision across modalities, which in turn degrades the reliability and accuracy of downstream tasks. Across existing datasets, the remaining misalignment and deformation can be summarized into three recurring factors. Spatio-temporal bias causes large parallax and multi-scale offsets. Platform motion coupled with imaging characteristics produces local warping and nonrigid deformation. Cross-modal association is further complicated by weak texture and occlusion, which leads to missing or unstable correspondences between modalities.

For two-dimensional image modalities, UAV-view RGB-T pairs and hyperspectral imagery mainly reflect the effects of spatiotemporal bias and local deformation. In RGB-T datasets such as DroneVehicle [140], approximate extrinsic calibration and pose information can support coarse alignment between modalities, yet noticeable residual misalignment remains. In particular, near-range objects often exhibit depth-dependent offsets and parallax that cannot be compensated by a single global affine transform or homography, leading to spatially varying cross-modal inconsistency. Related RGB-T datasets such as M3FD [141] and VIFB [142] further exhibit practical degradations including motion blur, rapid illumination

changes, and different levels of initial misalignment, which compound the geometric inconsistency introduced by platform motion and imperfect synchronization.

In hyperspectral datasets such as HyKo [143], pushbroom acquisition introduces modality-specific geometric artifacts. Even after mapping to a common reference frame, scanline-related distortion and local nonuniform scaling can persist, and band-to-band misregistration is frequently observed due to sensor motion and timing differences across spectral channels. These factors make it difficult to maintain consistent correspondences across bands and require subsequent processing to address spatially varying misalignment together with spectral variation.

For cross-dimensional data, joint datasets of low altitude optical images and LiDAR point clouds highlight the third challenge. Although extrinsic calibration and trajectory estimation enable approximate alignment in a common coordinate frame, residual calibration error and imperfect time synchronization often produce noticeable reprojection error on the image plane, especially in regions with large depth variation. SensatUrban [144] and UAVid-3D-Scenes [145] further reflect these issues in large urban scenes with complex structures and frequent occlusions, where reliable correspondences are difficult to maintain and cross-modal fusion becomes sensitive to residual misalignment.

Overall, the prevalence of coarsely aligned low altitude multimodal datasets is primarily due to the difficulty of obtaining high-precision registered data at scale. Because these datasets retain residual parallax and local deformation, treating them as well aligned creates a systematic mismatch between modeling assumptions and the actual sensing condition, weakening cross-modal feature consistency and degrading downstream tasks performance. Consequently, methods deployed on such data should explicitly address residual misalignment, either by incorporating an alignment module or by designing fusion and training strategies that are tolerant to spatial offsets and local distortion, so that the adverse impact of imperfect registration can be mitigated in practical settings.

5 Evaluation Metrics

Accurate evaluation is paramount in low altitude multimodal registration, as residual misalignment is often subtle, spatially variant, and highly sensitive to scene depth and platform motion. In the context of pixel-level registration, registration quality is quantified directly within the image plane by assessing discrepancies between corresponding pixels following the warping process. This approach necessitates either dense pixel-level correspondences or reliable geometric references. Accordingly, this section synthesizes similarity metrics and pixel-wise discrepancy measures computed on warped image pairs. Furthermore, it addresses geometric and boundary consistency measures derived from sparse correspondences or segmentation contours, provided such reference data are available. Conversely, feature-level registration is typically integrated within downstream perception pipelines and is, therefore, evaluated using task-specific performance criteria.

5.1 Pixel-level Registration Metrics

Pixel-level registration evaluates alignment quality directly in the image plane. The corresponding metrics therefore focus on the consistency between the reference image and the warped image, including intensity similarity, pixel-wise discrepancy, and geometric or boundary agreement when suitable annotations are available.

Mutual Information (MI). Mutual information is widely used for cross modality registration because it measures statistical dependence between intensity distributions. Let X and Y denote discretized intensity values of the reference image I_r and the warped image I_m, with marginals $p_X(x)$ and $p_Y(y)$ and joint distribution $p_{XY}(x,y)$. The mutual information is

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}. \quad (1)$$

In practice, p_{XY} is estimated from the joint histogram of (I_r, I_m), and the summation is taken over bins with $p_{XY}(x,y) > 0$.

Normalized Cross-Correlation (NCC). Normalized cross-correlation quantifies linear correlation. Let $\bar{I}_r$ and $\bar{I}_m$ be the mean intensities of I_r and I_m over Ω. Then

$$\mathrm{NCC}(I_r, I_m) = \frac{\sum_{x\in\Omega} \big(I_r(x) - \bar{I}_r\big)\big(I_m(x) - \bar{I}_m\big)}{\sqrt{\sum_{x\in\Omega} \big(I_r(x) - \bar{I}_r\big)^2} \sqrt{\sum_{x\in\Omega} \big(I_m(x) - \bar{I}_m\big)^2}}. \quad (2)$$

When the denominator is close to zero, numerical stabilization is required. NCC is most suitable for same-modality data or cases with approximately linear intensity relationships.

Mean Absolute Error (MAE). Mean absolute error measures the average absolute intensity discrepancy between the reference image I_r and the warped image I_m over the pixel domain Ω

$$\mathrm{MAE}(I_r, I_m) = \frac{1}{N} \sum_{x\in\Omega} |I_r(x) - I_m(x)|, \quad (3)$$

where $N = |\Omega|$. MAE provides a linear penalty on deviations and is relatively less sensitive to isolated large errors.

Mean Squared Error (MSE). Mean squared error measures the average squared intensity discrepancy between I_r and the warped image I_m

$$\mathrm{MSE}(I_r, I_m) = \frac{1}{N} \sum_{x \in \Omega} \left(I_r(x) - I_m(x)\right)^2 . \tag{4}$$

By applying a quadratic weighting to deviations, MSE assigns greater importance to large residual errors and is therefore well suited for scenarios where larger misalignments or severe local discrepancies need to be emphasized.

Dice Similarity Coefficient (DSC). For region-based metrics, let Ω denote the image domain. Let $\mathcal{R}_p \subset \Omega$ and $\mathcal{R}_g \subset \Omega$ denote the predicted and ground-truth foreground regions, respectively, where each region is represented as the set of pixels classified as foreground in the corresponding binary mask. The associated boundary point sets are denoted by $\partial\mathcal{R}_p$ and $\partial\mathcal{R}_g$, and $|\cdot|$ denotes set cardinality.

$$\mathrm{DSC}(\mathcal{R}_p, \mathcal{R}_g) = \frac{2\,|\mathcal{R}_p \cap \mathcal{R}_g|}{|\mathcal{R}_p| + |\mathcal{R}_g|}. \tag{5}$$

Hausdorff Distance (HD) and 95th-percentile Hausdorff Distance (HD95). The Hausdorff distance characterizes the worst-case boundary deviation by taking the maximum of the two directed distances between $\partial\mathcal{R}_p$ and $\partial\mathcal{R}_g$. Let $d(x, S) = \min_{y \in S} \|x - y\|_2$ denote the point-to-set distance, and define the symmetric distance set

$$\mathcal{D} = \{d(x, \partial\mathcal{R}_g)\}_{x \in \partial\mathcal{R}_p} \cup \{d(y, \partial\mathcal{R}_p)\}_{y \in \partial\mathcal{R}_g}. \tag{6}$$

The Hausdorff distance is then given by

$$\mathrm{HD}(\partial\mathcal{R}_p, \partial\mathcal{R}_g) = \max \mathcal{D}. \tag{7}$$

To reduce sensitivity to isolated outliers, the 95th-percentile Hausdorff distance replaces the maximum with a high-order quantile. Let $Q_{0.95}(\cdot)$ denote the 95th percentile operator. Then

$$\mathrm{HD95}(\partial\mathcal{R}_p, \partial\mathcal{R}_g) = Q_{0.95}(\mathcal{D}) . \tag{8}$$

Average Symmetric Surface Distance (ASSD). The average symmetric surface distance summarizes the mean boundary deviation by averaging distances in both directions, where $\partial\mathcal{R}_p$ and $\partial\mathcal{R}_g$ denote the boundary point sets, $|\cdot|$ denotes set cardinality, and $d(x, \partial\mathcal{R}) = \min_{z \in \partial\mathcal{R}} \|x - z\|_2$ denotes the Euclidean distance from a boundary point x to the closest point on the boundary $\partial\mathcal{R}$

$$\mathrm{ASSD}(\partial\mathcal{R}_p, \partial\mathcal{R}_g) = \frac{1}{2}\left(\frac{1}{|\partial\mathcal{R}_p|} \sum_{x \in \partial\mathcal{R}_p} d(x, \partial\mathcal{R}_g) + \frac{1}{|\partial\mathcal{R}_g|} \sum_{y \in \partial\mathcal{R}_g} d(y, \partial\mathcal{R}_p) \right). \tag{9}$$

5.2 Feature-level Registration Evaluation Based on Downstream Tasks

Feature-level registration focuses on alignment in the feature space. The corresponding evaluation is therefore commonly conducted through downstream task performance, including detection, segmentation, tracking, and re-identification, which provides a more practical measure of the quality of learned cross-modal correspondences.

Object Detection Metrics. For detection tasks, precision and recall are commonly used to quantify the correctness and completeness of predicted objects. Let TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively. Then

$$\mathrm{Precision} = \frac{TP}{TP + FP}, \tag{10}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN}. \tag{11}$$

Based on the precision–recall curve, the average precision for one category can be written as

$$AP = \int_0^1 P(R)\, dR, \tag{12}$$

where $P(R)$ denotes precision as a function of recall. Over C categories, the mean average precision is

$$mAP = \frac{1}{C} \sum_{c=1}^{C} AP_c. \tag{13}$$

Semantic Segmentation Metrics. For segmentation tasks, feature-level registration can be evaluated using mean intersection over union. Let TP_c, FP_c, and FN_c denote the numbers of true positive, false positive, and false negative pixels for class c, respectively, where $c \in \{1, \ldots, C\}$ and C is the total number of semantic classes. Then

$$\mathrm{IoU}_c = \frac{TP_c}{TP_c + FP_c + FN_c}, \tag{14}$$

and the mean intersection over union is

$$mIoU = \frac{1}{C} \sum_{c=1}^{C} \mathrm{IoU}_c. \tag{15}$$

Multi-Object Tracking Metrics. For tracking tasks, feature-level registration affects both object detection quality and cross-frame identity association. A commonly used metric is multi-object tracking accuracy, defined as

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t GT_t}, \tag{16}$$

where FN_t, FP_t, and $IDSW_t$ denote the numbers of false negatives, false positives, and identity switches at frame t, respectively, and GT_t is the number of ground-truth objects. Another important metric is IDF1, which emphasizes identity

preservation:

$$IDF1 = \frac{2\,IDTP}{2\,IDTP + IDFP + IDFN}, \tag{17}$$

where $IDTP$, $IDFP$, and $IDFN$ denote identity true positives, false positives, and false negatives, respectively.

Re-identification Metrics. For re-identification tasks, feature-level registration is expected to improve the consistency of shared embeddings under modality discrepancy. Let $\mathrm{rank}(i)$ denote the rank position of the first correct match for query i among N_q queries. Then Rank-1 accuracy can be written as

$$\text{Rank-1} = \frac{1}{N_q}\sum_{i=1}^{N_q}\mathbf{1}\big(\mathrm{rank}(i)=1\big), \tag{18}$$

where $\mathbf{1}(\cdot)$ is the indicator function. The mean average precision over all queries is

$$mAP = \frac{1}{N_q}\sum_{i=1}^{N_q} AP_i, \tag{19}$$

where AP_i is the average precision of query i.

6 Future Outlook

We review the main research directions and representative methods for multimodal image registration and provide a basis for low altitude applications. However, low altitude environments involve complex imaging conditions, changing observation scales, and strong platform constraints, so methods developed in more stable settings often show limited robustness when transferred directly. To achieve further progress in multimodal image registration for low altitude scenes, future research should continue to advance in data resources, method design, and evaluation frameworks, with greater emphasis on practical generalization and deployment.

In terms of data resources, low altitude datasets typically have a wide field of view and strong scale imbalance, so many targets occupy only a few pixels and fine structures are difficult to preserve. This weak evidence reduces the reliability of feature extraction and correspondence estimation, and it also increases labeling uncertainty for both boundaries and cross-modal matches. Existing datasets further remain limited in modality coverage, annotation quality, and scene diversity, while illumination, weather, land cover, and flight states can introduce substantial distribution shifts across regions and time. Synthetic data can still serve as a useful supplement for rare conditions and controlled analysis, but future progress will depend more on large-scale, high-quality real-world datasets with broader sensor coverage and more consistent collection and annotation protocols. In particular, future datasets should better support cross-region, cross-season, and cross-platform evaluation so that model transferability and reproducibility can be assessed more reliably.

In terms of methods, future work should place greater emphasis on robust and efficient alignment under the specific constraints of low altitude platforms. Building on the progress of existing methods, further advances may also come from improving cross-modal correspondence modeling under nonrigid deformation, dynamic interference, partial overlap, and limited onboard computation. For pixel-level registration, promising directions include physics-consistent modeling, uncertainty-aware similarity estimation, dynamic-region suppression, and coarse-to-fine optimization that remains reliable under noise, motion blur, compression artifacts, and resolution loss. For feature-level registration, an important direction is to develop stronger cross-modal representations that make better use of global context while preserving sensitivity to local structures. Large pretrained models, cross-modal pretraining across optical, infrared, and SAR imagery, and lightweight adaptation strategies may therefore provide an effective path toward better transfer across regions and seasons. In addition, multimodal alignment should gradually extend from static spatial registration to spatiotemporal consistency and semantic association, so that future systems can better support dynamic low altitude scenes.

In terms of evaluation frameworks, existing geometric and statistical metrics remain essential for assessing multimodal registration performance. However, when used alone, they are often not sufficient to fully reflect the practical effectiveness of low altitude multimodal alignment. Future evaluation should therefore further integrate conventional registration metrics with downstream task indicators, including detection, segmentation, and tracking performance, so as to establish a more comprehensive evaluation framework. In addition, robustness and uncertainty should be reported more explicitly. Practical evaluation protocols may include progressively increasing noise, motion blur, compression, or resolution degradation, and then measuring how registration accuracy and downstream task performance change with disturbance strength. Such an evaluation framework would better reflect model reliability and practical value in real low altitude environments.

A conclusion is not restatement of the abstract, but to stress the importance of the work, to give the paper a sense of completeness, and leave a final impression on the readers. The conclusion section is the last section of the paper to be numbered.

7 Conclusion

This survey summarizes multimodal image registration techniques for low altitude scene requirements and analyzes the applicability of existing methods under data conditions, scene complexity, and deployment constraints. Overall, the main challenges in low altitude multimodal registration include insufficient data resources with complex distributions, prominent effects from moving objects and nonrigid factors, and a mismatch between common evaluation metrics and downstream task benefit. Future research should advance jointly across data, methods, and evaluation. It is necessary to build datasets with broader coverage and more consistent labeling, to develop more robust and interpretable models, and to extend alignment objectives from the spatial level to the spatiotemporal level and the semantic level. It is also necessary to establish standardized and task oriented evaluation, and to provide measurable descriptions of robustness and reliability in complex environments. Through cross discipline collaboration and open sharing, multimodal image registration can support reliable deployment and wider adoption in key domains such as urban governance, agricultural monitoring, and disaster response.

[1] J. Chen, H. Wang, J. Tang, and J. Wang, Adaptfly: Prompt-guided adaptation of foundation models for low-altitude uav networks, *IEEE Transactions on Cognitive Communications and Networking*, 2026.

[2] Y. Shi, A. Abulizi, H. Wang, K. Feng, N. Abudukelimu, Y. Su, and H. Abudukelimu, Diffusion models for medical image computing: A survey, *Tsinghua Science and Technology*, vol. 30, no. 1, pp. 357–383, 2024.

[3] J. Zhu, H. Li, and T. Zhang, Camera, lidar, and imu based multi-sensor fusion slam: A survey, *Tsinghua Science and Technology*, vol. 29, no. 2, pp. 415–429, 2023.

[4] G. Lei, T. Liang, Y. Ping, X. Chen, L. Zhou, J. Wu, X. Zhang, H. Ding, X. Zhang, W. Yuan *et al.*, Towards secure low-altitude airspace: Mllm-enabled uav intent recognition, *IEEE Internet of Things Magazine*, 2026.

[5] W. Du and S. Tian, Transformer and gan-based super-resolution reconstruction network for medical images, *Tsinghua Science and Technology*, vol. 29, no. 1, pp. 197–206, 2023.

[6] H. Siddiqui, C. Banerjee, E. Blasch, E. Pasiliao Jr, and T. Mukherjee, Deep feature learning with concatenated rectified pooling units, *Big Data Mining and Analytics*, 2026.

[7] Y. Tao, Z. Gao, F. Ye, J. Xu, T. Song, W. Li, Y. Su, L. Peng, X. Wu, T. Qin *et al.*, Intelligent multimodal multi-sensor fusion-based uav identification, localization, and countermeasures for safeguarding low-altitude economy, *arXiv preprint arXiv:2510.22947*, 2025.

[8] W. Zhang, J. Ding, H. Liu, T. Han, Y. Liu, and L. T. Yang, Improving cross-modal semantic alignment with cross-modal joint semantic transformer for multimodal sentiment analysis, *Big Data Mining and Analytics*, 2026.

[9] T. Peng, Q. Li, and P. Zhu, Rgb-t crowd counting from drone: A benchmark and mmccn network, in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[10] A. Blali, S. Dargaoui, M. Azrour, A. Guezzaz, F. Amounas, and A. Alabdulatif, Hfxl-model: A hybrid deep learning and ensemble boosting framework for binary intrusion detection in iot networks, *Big Data Mining and Analytics*, 2025.

[11] F. Abbasi, M. Muzammal, Q. Qu, F. Riaz, and J. Ashraf, Snca: Semi-supervised node classification for evolving large attributed graphs, *Big Data Mining and Analytics*, vol. 7, no. 3, pp. 794–808, 2024.

[12] A. A. Cole-Rhodes, K. L. Johnson, J. LeMoigne, and I. Zavorin, Multiresolution registration of remote sensing imagery by optimization of mutual information using a stochastic gradient, *IEEE Transactions on Image Processing*, vol. 12, no. 12, pp. 1495–1511, 2003.

[13] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, Robust registration of multimodal remote sensing images based on structural similarity, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2941–2958, 2017.

[14] Y. Zhu, Y. Huang, M. Yang, D. Mao, Y. Zhang, L. Jiao, Y. Zhang, and J. Yang, Sar image super-resolution based on multi-scale edge texture-oriented gan approach, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[15] D. G. Lowe, Object recognition from local scale-invariant features, in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2. IEEE, 1999, pp. 1150–1157.

[16] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, Speeded-up robust features (surf), *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

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

[18] J. Chen, J. Tian, N. Lee, J. Zheng, R. T. Smith, and A. F. Laine, A partial intensity invariant feature descriptor for multimodal retinal image registration, *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1707–1718, 2010.

[19] W. Ma, Z. Wen, Y. Wu, L. Jiao, M. Gong, Y. Zheng, and L. Liu, Remote sensing image registration with modified sift and enhanced feature matching, *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 1, pp. 3–7, 2016.

[20] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, Kaze features, in *European Conference on Computer Vision*. Springer, 2012, pp. 214–227.

[21] Y. Xiang, F. Wang, and H. You, Os-sift: A robust sift-like algorithm for high-resolution optical-to-sar image registration

in suburban areas, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 6, pp. 3078–3090, 2018.

[22] A. Sedaghat and N. Mohammadi, Illumination-robust remote sensing image matching based on oriented self-similarity, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 153, pp. 21–35, 2019.

[23] X. Xiong, G. Jin, Q. Xu, and H. Zhang, Self-similarity features for multimodal remote sensing image matching, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 12 440–12 454, 2021.

[24] X. Xiong, G. Jin, Q. Xu, H. Zhang, L. Wang, and K. Wu, Robust registration algorithm for optical and sar images based on adjacent self-similarity feature, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[25] Z. Fan, M. Wang, Y. Pi, Y. Liu, and H. Jiang, A robust oriented filter-based matching method for multisource, multitemporal remote sensing images, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[26] C. A. Aguilera, A. D. Sappa, and R. Toledo, Lghd: A feature descriptor for matching across non-linear intensity variations, in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 178–181.

[27] E. Rosten, R. Porter, and T. Drummond, Faster and better: A machine learning approach to corner detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 105–119, 2008.

[28] J. Li, Q. Hu, and M. Ai, Rift: Multi-modal image matching based on radiation-variation insensitive feature transform, *IEEE Transactions on Image Processing*, vol. 29, pp. 3296–3310, 2019.

[29] Y. Zhang, Y. Yao, Y. Wan, W. Liu, W. Yang, Z. Zheng, and R. Xiao, Histogram of the orientation of the weighted phase descriptor for multi-modal remote sensing image matching, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 1–15, 2023.

[30] Y. Zhang, P. Wu, Y. Yao, Y. Wan, W. Zhang, Y. Li, and X. Yan, Multi-modal remote sensing image robust matching based on second-order tensor orientation feature transformation, *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[31] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, D2-net: A trainable cnn for joint description and detection of local features, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8092–8101.

[32] G. Potje, F. Cadar, A. Araujo, R. Martins, and E. R. Nascimento, Xfeat: Accelerated features for lightweight image matching, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2682–2691.

[33] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, Loftr: Detector-free local feature matching with transformers, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8922–8931.

[34] G. Bökman and F. Kahl, A case for using rotation invariant features in state of the art feature matchers, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5110–5119.

[35] Q. Wang, J. Zhang, K. Yang, K. Peng, and R. Stiefelhagen, Matchformer: Interleaving attention in transformers for feature matching, in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 2746–2762.

[36] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, Lightglue: Local feature matching at light speed, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 627–17 638.

[37] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, Superglue: Learning feature matching with graph neural networks, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4938–4947.

[38] J. Edstedt, I. Athanasiadis, M. Wadenbäck, and M. Felsberg, Dkm: Dense kernelized feature matching for geometry estimation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 765–17 775.

[39] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg, Roma: Robust dense feature matching, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 790–19 800.

[40] Y. Hel-Or, H. Hel-Or, and E. David, Fast template matching in non-linear tone-mapped images, in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1355–1362.

[41] ——, Matching by tone mapping: Photometric invariant template matching, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 317–330, 2013.

[42] S. Oron, T. Dekel, T. Xue, W. T. Freeman, and S. Avidan, Best-buddies similarity—robust template matching using mutual nearest neighbors, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1799–1813, 2017.

[43] I. Talmi, R. Mechrez, and L. Zelnik-Manor, Template matching with deformable diversity similarity, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 175–183.

[44] R. Kat, R. Jevnisek, and S. Avidan, Matching pixels using co-occurrence statistics, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1751–1759.

[45] W. Xiong, M. Sun, H. Du, B. Xiong, C. Zhang, Q. Ou, and Z. Rao, Cosine similarity template matching networks for optical and sar image registration, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[46] X. Tong, Z. Ye, Y. Xu, S. Liu, L. Li, H. Xie, and T. Li, A novel subpixel phase correlation method using singular value

decomposition and unified random sample consensus, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 8, pp. 4143–4156, 2015.

[47] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[48] Y. Ye, L. Bruzzone, J. Shan, F. Bovolo, and Q. Zhu, Fast and robust matching for multimodal remote sensing image registration, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 9059–9070, 2019.

[49] T. Gao, C. Lan, W. Huang, L. Wang, Z. Wei, and F. Yao, Multiscale template matching for multimodal remote sensing image, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 10 132–10 147, 2023.

[50] Y. Zhang, W. Zhang, Y. Yao, Z. Zheng, Y. Wan, and M. Xiong, Robust registration of multi-modal remote sensing images based on multi-dimensional oriented self-similarity features, *International Journal of Applied Earth Observation and Geoinformation*, vol. 127, p. 103639, 2024.

[51] J. Cheng, Y. Wu, W. AbdAlmageed, and P. Natarajan, Qatm: Quality-aware template matching for deep learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 553–11 562.

[52] R. Lei, B. Yang, D. Quan, Y. Li, B. Duan, S. Wang, H. Jia, B. Hou, and L. Jiao, Deep global feature-based template matching for fast multi-modal image registration, in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2021, pp. 5457–5460.

[53] L. Li, L. Han, M. Ding, H. Cao, and H. Hu, A deep learning semantic template matching framework for remote sensing image registration, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 181, pp. 205–217, 2021.

[54] Y. Ye, C. Yang, G. Gong, P. Yang, D. Quan, and J. Li, Robust optical and sar image matching using attention-enhanced structural features, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[55] B. Li, L. Y. Wu, D. Liu, H. Chen, Y. Ye, and X. Xie, Image template matching via dense and consistent contrastive learning, in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 1319–1324.

[56] M. Liu, G. Zhou, L. Ma, L. Li, and Q. Mei, Sifnet: A self-attention interaction fusion network for multisource satellite imagery template matching, *International Journal of Applied Earth Observation and Geoinformation*, vol. 118, p. 103247, 2023.

[57] Z. Gao, R. Yi, Z. Qin, Y. Ye, C. Zhu, and K. Xu, Learning accurate template matching with differentiable coarse-to-fine correspondence refinement, *Computational Visual Media*, vol. 10, no. 2, pp. 309–330, 2024.

[58] C. Chen, J. Qi, X. Liu, K. Bin, R. Fu, X. Hu, and P. Zhong, Weakly misalignment-free adaptive feature alignment for uavs-based multimodal object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 836–26 845.

[59] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, Deformable convolutional networks, in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 764–773.

[60] X. Chen, J. Yu, S. Kong, Z. Wu, and L. Wen, Joint anchor-feature refinement for real-time accurate object detection in images and videos, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 594–607, 2020.

[61] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, Reppoints: Point set representation for object detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9657–9666.

[62] J. Yan, Y. Liu, J. Sun, F. Jia, S. Li, T. Wang, and X. Zhang, Cross modal transformer: Towards fast and robust 3d object detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 268–18 278.

[63] Y.-T. Chen, J. Shi, Z. Ye, C. Mertz, D. Ramanan, and S. Kong, Multimodal object detection via probabilistic ensembling, in *European Conference on Computer Vision*. Springer, 2022, pp. 139–158.

[64] X. Han, L. Qi, Q. Yu, Z. Zhou, Y. Zheng, Y. Shi, and Y. Gao, Deep symmetric adaptation network for cross-modality medical image segmentation, *IEEE Transactions on Medical Imaging*, vol. 41, no. 1, pp. 121–132, 2021.

[65] Y. Liu, J. Deng, X. Gao, W. Li, and L. Duan, Bapa-net: Boundary adaptation and prototype alignment for cross-domain semantic segmentation, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8801–8811.

[66] H. Hu, Y. Chen, J. Xu, S. Borse, H. Cai, F. Porikli, and X. Wang, Learning implicit feature alignment function for semantic segmentation, in *European Conference on Computer Vision*. Springer, 2022, pp. 487–505.

[67] C. Liu, H. Liu, J. Zhuo, B. Zou, J. Chen, Q. Zhao, and H. Ma, Implicit alignment and query refinement for rgb-t semantic segmentation, *Pattern Recognition*, vol. 169, p. 111951, 2026.

[68] Z. Jiang, Y. Li, C. Yang, P. Gao, Y. Wang, Y. Tai, and C. Wang, Prototypical contrast adaptation for domain adaptive semantic segmentation, in *European Conference on Computer Vision*. Springer, 2022, pp. 36–54.

[69] Y. Wan, W. Wang, G. Zou, and B. Zhang, Cross-modal feature alignment and fusion for composed image retrieval, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8384–8388.

[70] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng, Attentionfgan: Infrared and visible image fusion using attention-based generative adversarial networks, *IEEE Transactions on Multimedia*, vol. 23, pp. 1383–1396, 2020.

[71] Y. Jiao, Z. Jie, S. Chen, J. Chen, L. Ma, and Y.-G. Jiang,

Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 643–21 652.

[72] A. Chartsias, G. Papanastasiou, C. Wang, S. Semple, D. E. Newby, R. Dharmakumar, and S. A. Tsaftaris, Disentangle, align and fuse for multimodal and semi-supervised image segmentation, *IEEE Transactions on Medical Imaging*, vol. 40, no. 3, pp. 781–792, 2020.

[73] Z. Wang, W. Shao, Y. Chen, J. Xu, and X. Zhang, Infrared and visible image fusion via interactive compensatory attention adversarial learning, *IEEE Transactions on Multimedia*, vol. 25, pp. 7800–7813, 2022.

[74] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer, *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 7, pp. 1200–1217, 2022.

[75] M. Jia, X. Cheng, S. Lu, and J. Zhang, Learning disentangled representation implicitly via transformer for occluded person re-identification, *IEEE Transactions on Multimedia*, vol. 25, pp. 1294–1305, 2022.

[76] S. Wu, S. Shan, G. Xiao, M. S. Lew, and X. Gao, Implicit modality knowledge alignment and uncertainty estimation for visible-infrared person re-identification, *Expert Systems with Applications*, vol. 259, p. 125291, 2025.

[77] K. Ren and L. Zhang, Implicit discriminative knowledge learning for visible-infrared person re-identification, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 393–402.

[78] H. Gao, Y. Yan, Y. He, J. Zhou, Z. Zhang, and Y. Yang, Cail: Cross-modal vehicle reidentification in aerial images using the centroid-aligned implicit learning network, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 18, pp. 2577–2588, 2024.

[79] Z. Chai, Y. Ling, Z. Luo, D. Lin, M. Jiang, and S. Li, Dual-stream transformer with distribution alignment for visible-infrared person re-identification, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 11, pp. 6764–6776, 2023.

[80] T. Liang, Y. Jin, W. Liu, and Y. Li, Cross-modality transformer with modality mining for visible-infrared person re-identification, *IEEE Transactions on Multimedia*, vol. 25, pp. 8432–8444, 2023.

[81] Y. Feng, J. Yu, F. Chen, Y. Ji, F. Wu, S. Liu, and X.-Y. Jing, Visible-infrared person re-identification via cross-modality interaction transformer, *IEEE Transactions on Multimedia*, vol. 25, pp. 7647–7659, 2022.

[82] W. Zhang, H. Zhou, S. Sun, Z. Wang, J. Shi, and C. C. Loy, Robust multi-modality multi-object tracking, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2365–2374.

[83] T. Zhang, H. Guo, Q. Jiao, Q. Zhang, and J. Han, Efficient rgb-t tracking via cross-modality distillation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5404–5413.

[84] T. Zhang, X. He, Q. Jiao, Q. Zhang, and J. Han, Amnet: Learning to align multi-modality for rgb-t tracking, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 7386–7400, 2024.

[85] X. Lan, M. Ye, S. Zhang, H. Zhou, and P. C. Yuen, Modality-correlation-aware sparse representation for rgb-infrared object tracking, *Pattern Recognition Letters*, vol. 130, pp. 12–20, 2020.

[86] P. Lai, D. Gao, S. Wang, and G. Cheng, Mining representative tokens via transformer-based multi-modal interaction for rgb-t tracking, *Pattern Recognition*, p. 112162, 2025.

[87] F. Maes, D. Vandermeulen, and P. Suetens, Medical image registration using mutual information, *Proceedings of the IEEE*, vol. 91, no. 10, pp. 1699–1722, 2003.

[88] J. Liang, X. Liu, K. Huang, X. Li, D. Wang, and X. Wang, Automatic registration of multisensor images using an integrated spatial and mutual information (smi) metric, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 603–615, 2013.

[89] A. Collignon, Automated multi-modality image registration based on information theory, in *Information processing in medical imaging*, 1995, pp. 263–274.

[90] P. Viola and W. M. Wells III, Alignment by maximization of mutual information, *International Journal of Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997.

[91] W. M. Wells III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, Multi-modal volume registration by maximization of mutual information, *Medical image analysis*, vol. 1, no. 1, pp. 35–51, 1996.

[92] J. Tsao, Interpolation artifacts in multimodality image registration based on maximization of mutual information, *IEEE Transactions on Medical Imaging*, vol. 22, no. 7, pp. 854–864, 2003.

[93] H. Luan, F. Qi, Z. Xue, L. Chen, and D. Shen, Multimodality image registration by maximization of quantitative–qualitative measure of mutual information, *Pattern Recognition*, vol. 41, no. 1, pp. 285–298, 2008.

[94] D. Loeckx, P. Slagmolen, F. Maes, D. Vandermeulen, and P. Suetens, Nonrigid image registration using conditional mutual information, *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 19–29, 2009.

[95] J. Woo, M. Stone, and J. L. Prince, Multimodal registration via mutual information incorporating geometric and spatial context, *IEEE Transactions on Image Processing*, vol. 24, no. 2, pp. 757–769, 2014.

[96] P. A. Legg, P. L. Rosin, D. Marshall, and J. E. Morgan, A robust solution to multi-modal image registration by combining mutual information with multi-scale derivatives, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2009, pp. 616–623.

[97] A. Dame and E. Marchand, Second-order optimization of mutual information for real-time image registration, *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4190–4203, 2012.

[98] Y. Wu, W. Ma, Q. Miao, and S. Wang, Multimodal continuous ant colony optimization for multisensor remote sensing image registration with local search, *Swarm and Evolutionary Computation*, vol. 47, pp. 89–95, 2019.

[99] Q. Hu, Y. Wei, J. Pang, and M. Liang, Unsupervised domain adaptation for brain structure segmentation via mutual information maximization alignment, *Biomedical Signal Processing and Control*, vol. 90, p. 105784, 2024.

[100] Z. Wang, Y. Yang, Y. Chen, T. Yuan, M. Sermesant, H. Delingette, and O. Wu, Mutual information guided diffusion for zero-shot cross-modality medical image translation, *IEEE Transactions on Medical Imaging*, vol. 43, no. 8, pp. 2825–2838, 2024.

[101] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, Image-to-image translation with conditional adversarial networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[102] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, Improved techniques for training gans, *Advances in neural information processing systems*, vol. 29, 2016.

[103] Y. Zheng, X. Sui, Y. Jiang, T. Che, S. Zhang, J. Yang, and H. Li, Symreg-gan: symmetric image registration with generative adversarial networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5631–5646, 2021.

[104] M. Arar, Y. Ginger, D. Danon, A. H. Bermano, and D. Cohen-Or, Unsupervised multi-modal image registration via geometry preserving image-to-image translation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 410–13 419.

[105] L. Kong, C. Lian, D. Huang, Y. Hu, Q. Zhou *et al.*, Breaking the dilemma of medical image-to-image translation, *Advances in Neural Information Processing Systems*, vol. 34, pp. 1964–1978, 2021.

[106] M. Meng, L. Bi, M. Fulham, D. D. Feng, and J. Kim, Enhancing medical image registration via appearance adjustment networks, *NeuroImage*, vol. 259, p. 119444, 2022.

[107] M. Guo, Unsupervised multi-modal medical image registration via invertible translation, in *European Conference on Computer Vision*. Springer, 2024, pp. 22–38.

[108] J. Y. Han, S. Yang, S. Kim, S. Kim, S.-H. Lim, H. Yun, D. Kim, and W.-J. Yi, Semi-supervised deformation-free image-to-image translation for realistic ct synthesis from cbct, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2025, pp. 577–586.

[109] P. Yuan, J. Dong, W. Zhao, F. Lyu, C. Xue, Y. Zhang, C. Yang, Z. Wu, Z. Gao, T. Lyu *et al.*, Ladda: Latent diffusion-based domain-adaptive feature disentangling for unsupervised multi-modal medical image registration, *IEEE Journal of Biomedical and Health Informatics*, 2025.

[110] M. Lu, M. Jiang, X. Tao, and J. Kong, Au-net: Adaptive unified network for joint multi-modal image registration and fusion, *IEEE Transactions on Image Processing*, 2025.

[111] Z. Jiang, Z. Zhang, and J. Liu, Harmonized domain enabled alternate search for infrared and visible image alignment, *IEEE Transactions on Image Processing*, 2025.

[112] B. Kim, Y. Zhuang, T. S. Mathai, and R. M. Summers, Otmorph: unsupervised multi-domain abdominal medical image registration using neural optimal transport, *IEEE Transactions on Medical Imaging*, 2024.

[113] H. Xu, J. Ma, J. Yuan, Z. Le, and W. Liu, Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 679–19 688.

[114] H. Xu, J. Yuan, and J. Ma, Murf: Mutually reinforcing multi-modal image registration and fusion, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 148–12 166, 2023.

[115] Y. Xia, J. Monica, W.-L. Chao, B. Hariharan, K. Q. Weinberger, and M. Campbell, Image-to-image translation for autonomous driving from coarsely-aligned image pairs, in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 7756–7762.

[116] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, Voxelmorph: a learning framework for deformable medical image registration, *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.

[117] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces, *Medical image analysis*, vol. 57, pp. 226–236, 2019.

[118] Z. Lu, G. Yang, T. Hua, L. Hu, Y. Kong, L. Tang, X. Zhu, J.-L. Dillenseger, H. Shu, and J.-L. Coatrieux, Unsupervised three-dimensional image registration using a cycle convolutional neural network, in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2174–2178.

[119] C. Lian, X. Li, L. Kong, J. Wang, W. Zhang, X. Huang, and L. Wang, Cocyclereg: Collaborative cycle-consistency method for multi-modal medical image registration, *Neurocomputing*, vol. 500, pp. 799–808, 2022.

[120] J. Fan, X. Cao, Q. Wang, P.-T. Yap, and D. Shen, Adversarial learning for mono-or multi-modal registration, *Medical image analysis*, vol. 58, p. 101545, 2019.

[121] Y. Zhong, S. Zhang, Z. Liu, X. Zhang, Z. Mo, Y. Zhang, H. Hu, W. Chen, and L. Qi, Unsupervised fusion of misaligned pat and mri images via mutually reinforcing cross-modality image generation and registration, *IEEE Transactions on Medical Imaging*, vol. 43, no. 5, pp. 1702–1714, 2023.

[122] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C. M. Moore, M. Emberton *et al.*,

Weakly-supervised convolutional neural networks for multimodal image registration, *Medical image analysis*, vol. 49, pp. 1–13, 2018.

[123] W. Lai, F. Zeng, X. Hu, S. He, Z. Liu, and Y. Jiang, Regseg: An end-to-end network for multimodal rgb-thermal registration and semantic segmentation, *IEEE Transactions on Image Processing*, 2024.

[124] F. Liu, J. Cai, Y. Huo, C.-T. Cheng, A. Raju, D. Jin, J. Xiao, A. Yuille, L. Lu, C. Liao *et al.*, Jssr: A joint synthesis, segmentation, and registration system for 3d multi-modal image alignment of large-scale pathological ct scans, in *European Conference on Computer Vision*. Springer, 2020, pp. 257–274.

[125] L. Li, L. Han, M. Ding, and H. Cao, Multimodal image fusion framework for end-to-end remote sensing image registration, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[126] S. Sun, K. Han, C. You, H. Tang, D. Kong, J. Naushad, X. Yan, H. Ma, P. Khosravi, J. S. Duncan *et al.*, Medical image registration via neural fields, *Medical Image Analysis*, vol. 97, p. 103249, 2024.

[127] A. Q. Wang, M. Y. Evan, A. V. Dalca, and M. R. Sabuncu, A robust and interpretable deep learning framework for multimodal registration via keypoints, *Medical Image Analysis*, vol. 90, p. 102962, 2023.

[128] H. Li, Z. Yang, Y. Zhang, W. Jia, Z. Yu, and Y. Liu, Mulfs-cap: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 5, pp. 3673–3690, 2025.

[129] L. Tang, Q. Yan, X. Xiang, L. Fang, and J. Ma, C2rf: Bridging multi-modal image registration and fusion via commonality mining and contrastive learning, *International Journal of Computer Vision*, vol. 133, no. 8, pp. 5262–5280, 2025.

[130] T. Li, Z. Zhou, Q. Zhu, J. Luo, and Y. Wang, Multiscale spherical feature decoupling network for multimodal image registration, *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[131] A. Dong, J. Xu, and L. Wang, A fusion network for multimodality medical image registration with progressive feature alignment, *Knowledge-Based Systems*, vol. 317, p. 113427, 2025.

[132] J. Ding, Y. Zhao, L. Pei, Y. Shan, Y. Du, and W. Li, Modal-invariant progressive representation for multimodal image registration, *Information Fusion*, vol. 117, p. 102903, 2025.

[133] X. Sun, H. Ding, C. Liu, J. Lu, F. Li, Z. Shao, and J. Yang, K-cmorph: Integrating k-space consistency and complex-valued processing for improved mri deformable registration, in *International Conference on Intelligent Computing*. Springer, 2025, pp. 413–424.

[134] T. Li, B. Cao, P. Zhu, B. Xiao, and Q. Hu, Bi-directional self-registration for misaligned infrared-visible image fusion, *arXiv preprint arXiv:2505.06920*, 2025.

[135] M. K. Hasan, Y. Luo, G. Yang, and C. H. Yap, Feedback attention to enhance unsupervised deep learning image registration in 3d echocardiography, *IEEE Transactions on Medical Imaging*, 2025.

[136] T. Li, B. Cao, J. Feng, H. Cao, Q. Hu, and P. Zhu, Hyperbolic cycle alignment for infrared-visible image fusion, *arXiv preprint arXiv:2507.23508*, 2025.

[137] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, Llvip: A visible-infrared paired dataset for low-light vision, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3496–3504.

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

[139] Z. Gao, D. Li, Y. Kuai, R. Chen, and G. Wen, Visible-infrared image alignment for unmanned aerial vehicles: Benchmark and new baseline, *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[140] Y. Sun, B. Cao, P. Zhu, and Q. Hu, Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6700–6713, 2022.

[141] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5802–5811.

[142] X. Zhang, P. Ye, and G. Xiao, Vifb: A visible and infrared image fusion benchmark, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 104–105.

[143] C. Winkens, F. Sattler, V. Adams, and D. Paulus, Hyko: A spectral dataset for scene understanding, in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 254–261.

[144] Q. Hu, B. Yang, S. Khalid, W. Xiao, N. Trigoni, and A. Markham, Sensaturban: Learning semantics from urban-scale photogrammetric point clouds, *International Journal of Computer Vision*, vol. 130, no. 2, pp. 316–343, 2022.

[145] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, Uavid: A semantic segmentation dataset for uav imagery, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 165, pp. 108–119, 2020.

Author biography

Timing Li received the B.S. degree in microelectronics science and engineering from Lanzhou University, Lanzhou, China, in 2017, and the M.S. degree in integrated circuit engineering from Tianjin University, Tianjin, China, in 2021. He is currently pursuing the Ph.D. in computer science and technology at Tianjin University, Tianjin, China. His research interests are focused on computer vision, pattern recognition, and multi-modality learning.

Bing Cao received the B.S. degree in electrical engineering and automation from Hebei University, Baoding, China, in 2015, and the Ph.D. degree in information and telecommunications engineering from Xidian University, Xi'an, Shaanxi, in 2020. From April 2019 to July 2020, he was a Visiting Ph.D. Student with the University of North Caolina at Chapel Hill, Chapel Hill, NC, USA. He is currently Professor with the School of Computer Science and Technology, Tianjin University, Tianjin, China. His research interests include computer vision, pattern recognition, and machine learning.

Pengfei Zhu received the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong, China, in 2015. He received his B. S. and M. S. from Harbin Institute of Technology, Harbin, China in 2009 and 2011, respectively. Now he is Professor with the School of Computer Science and Technology, Tianjin University. His research interests are focused on machine learning and computer vision.

Kewen Li received the Ph.D. degree from Tianjing University, Tian Jing, China, in 2011. He received his M. S. from China University of Petroleum, Qingdao, China in 2001. Now he is Professor with the School of Computer Science and Technology, China University of Petroleum. His research interests are focused on machine learning and computer vision.