Information Fusion 73 (2021) 22–71

Contents lists available at ScienceDirect

Information Fusion

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

A review of multimodal image matching: Methods and applications

Xingyu Jiang [a], Jiayi Ma [a,*], Guobao Xiao [b], Zhenfeng Shao [c,*], Xiaojie Guo [d]

[a] Electronic Information School, Wuhan University, Wuhan, 430072, China
[b] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, College of Computer and Control Engineering, Minjiang University, Fuzhou, 350108, China
[c] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, 430079, China
[d] College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China

ARTICLE INFO

Keywords:
Multimodal
Matching
Registration
Deep learning
Medical
Remote sensing
Computer vision

ABSTRACT

Multimodal image matching, which refers to identifying and then corresponding the same or similar structure/content from two or more images that are of significant modalities or nonlinear appearance difference, is a fundamental and critical problem in a wide range of applications, including medical, remote sensing and computer vision. An increasing number and diversity of methods have been proposed over the past decades, particularly in this deep learning era, due to the challenges in eliminating modality variance and geometrical deformation that intrinsically exist in multimodal image matching. However, a comprehensive review and analysis of traditional and recent trainable methods and their applications in different research fields are lacking. To this end and in this survey, we first introduce two general frameworks, saying area- and feature-based, in terms of their core components, taxonomy, and procedure details. Second, we provide a comprehensive review of multimodal image matching methods from handcrafted to deep methods for each research field according to their imaging nature, including medical, remote sensing and computer vision. Extensive experimental comparisons of interest point detection, description and matching, and image registration are performed on various datasets containing common types of multimodal image pairs that we collected and annotated. Finally, we briefly introduce and analyze several typical applications to reveal the significance of multimodal image matching and provide insightful discussions and conclusions to these multimodal image matching approaches, and simultaneously deliver their future trends for researchers and engineers in related research areas to achieve further breakthroughs.

1. Introduction

General image matching, as defined in related surveys [1,2], aims to identify and then correspond the same or similar structure/content from two or more images. A more practical purpose is to geometrically warp a moving (sensed or target) image into the common spatial coordinate system of a fixed (reference or source) image and align their common area in pixel, *i.e.,* image registration or alignment. Multimodal image matching (MMIM) sometimes termed as Multimodal image registration (MMIR) can be seen as specific cases in the image matching community. A more universal definition is that the targets to be matched have significant nonlinear appearance differences that are typically caused by different (not limited to) imaging sensors, or by different imaging conditions (such as day–night [3–5], cross-weather [6], cross-season [7]), and input data types (such as image–paint–sketch [8,9], and image–text [10–12]).

MMIM has taken a significant role as a preprocedure requirement in many research areas and high-level tasks. Its most direct purpose is to identify and gather a wide range of physical properties from different modalities, thereby yielding richer scene representations by means of image registration and fusion [13,14]. Another goal is to recognize the differences or connections among the input images for change detection [15], target recognition/tracking [16–18], and cross-modality person re-identification [19–22]. In addition, the images captured from another modality would serve as a supplementary information supplier to achieve advanced performance in 3D reconstruction [23] and image localization (such as simultaneous localization and mapping, and place recognition) [7,24,25]. In medical domains such as radiation planning, multimodal data (*e.g.,* computed tomography (CT) and magnetic resonance imaging (MRI) scans) are often used for more accurate tumor contouring, thus reducing the risk of damaging healthy tissues during radiotherapy treatment [26,27].

* Corresponding authors.
E-mail addresses: jiangx.y@whu.edu.cn (X. Jiang), jyma2010@gmail.com (J. Ma), x-gb@163.com (G. Xiao), shaozhenfeng@whu.edu.cn (Z. Shao), xj.max.guo@gmail.com (X. Guo).

https://doi.org/10.1016/j.inffus.2021.02.012
Received 14 December 2020; Received in revised form 23 January 2021; Accepted 21 February 2021
Available online 1 March 2021

Increasing efforts have been made to propose advanced technologies over the past decades because of the high-performance requirement for MMIM or MMIR in these practical applications. As many researchers suggested, a more acceptable taxonomy for existing methods is area- and feature-based pipelines [1,2]. An area-based framework generally registers the image pairs under the guidance of a similarity metric that can measure the accuracy of image alignment to drive the optimization of the registration procedure. By contrast, feature-based framework is more steerable in the general image matching task and related applications. Such methods commonly start with distinctive feature extraction and then match the features with/without feature description, followed by a transformation model estimation and image resampling and warping, thus achieving image registration. Feature-based pipeline is used more widely due to its flexibility, robustness, and capability in a wide range of applications [2]. In recent years, deep learning has made dramatic progress on a wide range of complex tasks. Numerous researchers and engineers have also successfully addressed the image matching problem with a data-driven strategy. Learning-based methods can actually be regarded as a direct replacement of traditional frameworks in information extraction and representation, similarity measurement, and transformation parameter regression. Even though there exist many systematic and promising approaches for MMIM, it remains an open problem to develop a general pipeline with promising performance in accuracy, robustness, and efficiency due to the following challenges:

- The first limitation is insufficient or unavailable image data of different modalities. No complete and comprehensive database contains all types of multimodal image pairs together with their ground truths. To our knowledge, researchers in the medical field have provided sufficient multimodal volumes of different imaging devices and/or targets, but not in the remote sensing and computer vision communities.
- Area-based methods highly depend on the appropriate choices of similarity metric, geometrical transformation model, and optimization method. However, these components are also largely affected by overlapping areas and image contents [2]. This situation would be worse in the multimodal case due to the serious nonlinear intensity variance between two images of different modalities. In addition, these methods are extremely time consuming in the case of high-resolution image pairs or even fail if the image pairs undergo large deformations.
- The core challenges in feature-based pipeline are feature detection and description from multimodal image pairs owing to their nonlinear intensity difference. In this case, many widely used feature matchers proposed for general vision applications would be out of operation. Other limitations in this framework are consistent with those in general image matching tasks, such as the combinational nature in corresponding two feature sets, which would create a heavy computational burden, or the inevitable heavy outliers (mismatches) that appear in putative matches due to the use of local image information only.
- The deep learning framework has shown great potential in MMIM problem, but it still faces several challenges as introduced in [2, 28]. On the one hand, learning from images to directly perform image registration would be limited by the large geometrical deformations and high-resolution images. On the other hand, learning from sparse point data with convolutional strategy is still a challenging problem due to the disordered and dispersed nature of point data. In addition, this approach is limited by the insufficient real data for training to obtain a satisfying matching/registration model.
- Each pair of image modality has its own difficulties that differ from others due to the variance of imaging device and property. This would make it difficult to propose a universal paradigm to well address image registration for these common types of modalities from medical, remote sensing, and computer vision research areas simultaneously.

Few works specifically review MMIM methods and applications that simultaneously contain medical, remote sensing, and computer vision research. Existing surveys mainly focus on a general image matching or registration task, and only briefly introduce the multimodal case as a subsection [1,2,28–30]. Most works focus on medical image registration to deliver specific instruction, either aiming at 3D–2D registration [31,32], deformable registration [29], or registration of different objects such as breast [33,34], brain [35], and vascular [36]. Others typically review implementations [37,38] or deep learning frameworks [28,30]. Two survey papers related to remote sensing image matching were briefly introduced in [39,40]. In this regard, we provide an up-to-date and comprehensive review of existing MMIM methods and applications in the medical, remote sensing, and computer vision research areas, especially for the recently introduced learning-based methods.

The overall structure of this survey is presented in Fig. 1. In Section 2, we introduce two general frameworks that are commonly used in image matching-area-based and feature-based-to provide an overview of the components and flowcharts. We also review these commonly used ideas from handcrafted to deep learning techniques and analyze how they are extended to multimodal cases. In Section 3, we present a detailed review of multimodal image matching methods in regard to different research fields, including medical, remote sensing, and computer vision. In Sections 4 and 5, we respectively provide a comprehensive analysis of experimental evaluation and related applications. In Section 6, we conclude and discuss possible future developments.

2. General frameworks of image matching

As aforementioned, MMIM can be seen as a specific case in the general image matching community. In addition to the appearance difference, MMIM will face similar challenges that appear in general image matching tasks, *i.e.,* complex geometrical deformation, poor image quality, and high computation and storage burden. Therefore, we will first comprehensively introduce general matching frameworks following the taxonomy in [1,2], namely, area-based and feature-based pipelines. In each category, we will emphatically review the core idea of classical and recently published methods, particularly those that use learning techniques, then introduce the extensions of these ideas to multimodal cases and deliver their connections.

2.1. Area-based pipeline

An area-based pipeline aims to realize image registration by using the intensity information of entire images. Generally, given a predefined transform model, a similarity metric together with an optimization method is required to estimate the transform parameters and then align the common area of two images by optimizing the overall cost function. This cost function is usually created by the similarity measurement between a fixed image and a warped moving image, thereby enabling the accuracy of image alignment to be measured. A general goal of this framework can be formulated as the following form:

$$\mathcal{T}^* = \arg\min_{\mathcal{T}} \mathcal{M}(I_F, I_M \circ \mathcal{T}) + \mathcal{R}(\mathcal{T}), \tag{1}$$

where $\mathcal{T}$ indicates the transformation model from moving image I_M to fixed image I_F, and $\mathcal{M}$ qualifies the level of alignment between them. $\mathcal{R}$ regularizes the transformation with the goal of favoring any specific properties in the solution that the user requires and seeks to tackle the difficulty associated with the ill-posedness of the problem. The whole framework is illustrated in Fig. 2. In the following, we will introduce the matching methods using handcrafted and data-driven strategies, and analyze how the traditional methods develop into those trainable ones in theory and practice.

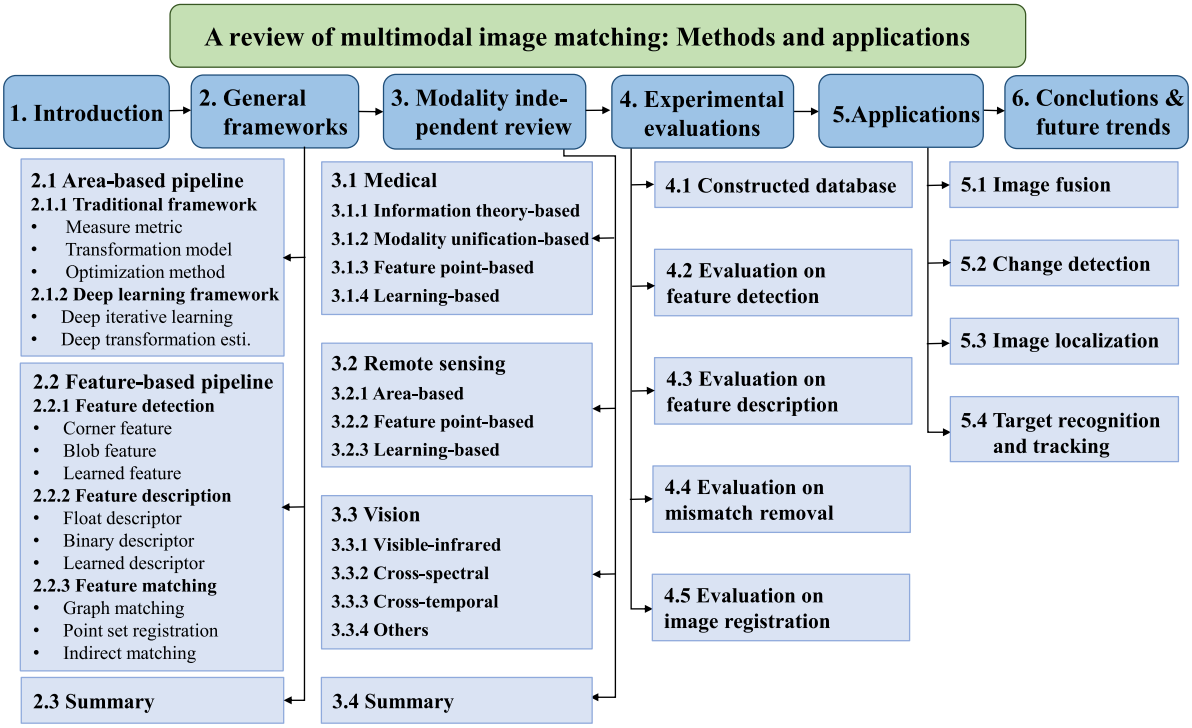

Fig. 1. Structure of this survey.

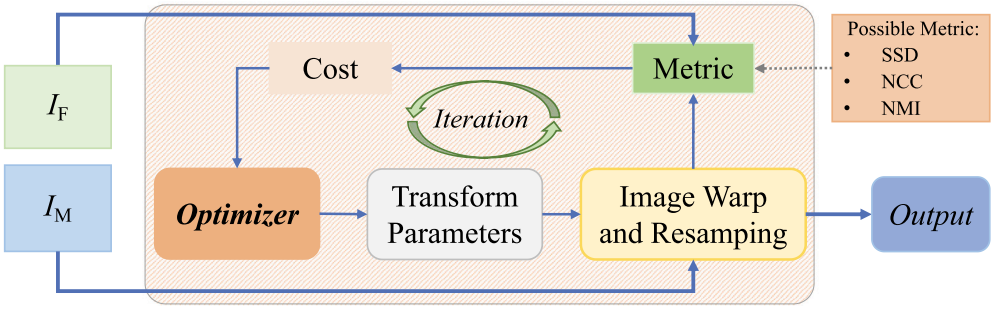

Fig. 2. General framework of traditional area-based image registration.

2.1.1. Handcrafted framework

From the above details, we can see that area-based image registration method consists of three components: (1) measure metric, (2) transformation model, and (3) optimization method. Next, we will introduce this general framework following these three key components.

(1) Measure metric

As a critical component in the area-based image registration pipeline, metrics or matching criteria for measuring the accuracy of image registration are a hot topic in guiding the entire optimization procedure. Different metrics can be devised depending on the assumptions about the intensity relationship between two images [29]. Frequently used manual metrics can be briefly classified into correlation-like and information theory-based methods.

A direct measurement in monomodal image registration is to compute the distance of corresponding pixels, such as the sum of squared or absolute differences, *i.e.,* SSD and SAD. This computation is performed under the assumption that the same anatomical structures have similar intensity values. Another idea is inspired by signal correlation, and it assumes that a linear correlation would intrinsically exist between two signal series, in which the most representative criteria are cross correlation [41–43] and normalized correlation coefficient (NCC) [44, 45].

The most representative metric in information theory-based methods is mutual information (MI) [46,47], which is typically based on a statistical comparison of the image domain. MI is particularly suitable for the registration of multimodal cases due to its statistical dependency between two images. This metric has also attracted great interest in designing advanced information-based metrics, including a normalized version of MI (NMI) [48], an upper bound on the maximum MI [49], conditional MI (cMI) [50], regional MI (RMI) [51], or some divergence-based approaches [52–54]. However, the MI experiences difficulty in determining the global maximum of the entire searching space, inevitably reducing its robustness [29].

Nevertheless, these metrics are not absolutely linear to the accuracy of image registration, which are largely affected by the size of overlapping area and image contents. Some limitations also exist when image pairs undergo serious image deformations or contain a smooth area without any prominent details.

(2) Transformation model

Another key component in image matching community is the choice of transformation model (also known as mapping function). In general, transformation models typically explain the geometrical relations between the target image pairs, whose parameters need to be accurately estimated to guide the image warping and resampling (together with an appropriate interpolation method) for ultimate registration. Aside

from area-based image matching, a good transformation model is also required for feature-based matching pipeline, such as realizing point set registration or robustly estimating the global transformation after feature matching. Although transformation models were fully studied in previous decades and formed as a series of standards, they are still worth revisiting because choosing an appropriate transformation model can not only maintain the matching accuracy but also achieve fast optimization, particularly for learning-based methods. According to the type of geometrical scene between two images, existing transformation models can be briefly classified into linear models (*e.g.,* rigid, affine, and projective) and nonlinear deformations (*e.g.,* interpolation model, elastic model, and diffusion model).

Linear models. The rigid transformation can be seen as the simplest static model that accounts for rotation and translation with 3 degrees of freedom (DOF) (or 6-DOF in the 3D case). This basic model is a common choice in the literature for matching rigid bodies, such as bones, or using similarity transform to tolerate the scale variations. A more general model (using 6-DOF in the 2D case), namely, affine transformation, can preserve the parallelism of lines but not their lengths or angles, thus additionally allowing for shearing deformations and mapping a parallelogram onto a square. Two other parametrical models, which are derived from multiview geometry and photography imaging, use more parameters to capture the camera motions, namely, fundamental matrix (epipolar geometry) and homography matrix (projective transformation). Fundamental matrix usually constrains a point in the first image to a line in the second image, while projective transformation can map a trapezoid onto a square. These two parametric models can meet the majority of requirements of natural image matching.

Nonlinear deformations. Transformation models that can explain the elastic bodies even local deformations are urgently needed in dynamic scenes. To this end, numerous nonlinear models, also known as nonrigid or deformation models, were investigated in previous decades and are widely used in the image matching community (image or point set registration, feature matching). According to the basic idea, nonlinear deformations can be classified into *physical models* and *interpolation models* [29].

Physical models are commonly derived from physical phenomena and represented by partial differential equations. An elastic body model typically regards an image grid as an elastic membrane that is deformed under the influence of two forces that compete until equilibrium is reached. An external force tries to deform the image such that matching is achieved, while an internal one enforces the elastic properties of the material. A nonlinear form of elastic model is proposed to handle large deformations by using hyperelastic material properties. Apart from the elastic body, another physical model used to recover large deformations is based on viscous fluid, in which a smoothing term is used to constrain neighboring points to deform similarly. In addition and based on Maxwell's demon, nonrigid deformation can be modeled as a diffusion procedure. This procedure typically iterates between estimating demon forces for each demon and updating the transformations guided by the calculated forces [55]. Regularization is often efficiently achieved through Gaussian kernel convolutions. Other physical models include curvature registration and flows of diffeomorphisms. In the regularization scheme of curvature methods, several boundary conditions are designed for efficient optimization, such as using second-order terms [56] or the approximation form of curvature penalty [57]. In diffeomorphism methods, the deformation is modeled by considering its velocity over time according to the Lagrange transport equation together with a regularization term that constrains the velocity field to be smooth [58,59]. A diffeomorphism framework can handle large deformation registration problems because of the gradual variation behavior of the velocity field and is thus also named large deformation diffeomorphic metric mapping [60,61].

In contrast to physical models, geometric transformations derived from interpolation or approximation theory have received considerable attention for nonrigid matching due to their low DOF, computational efficiency, and general applicability. In this family of transformations, the displacements are interpolated based on the matched control points or landmarks, thus spreading to the rest of the image domain, typically by using different spline functions or interpolation functions. The most representative function is known as radial basis function (RBF) [62], where the value at an interpolation position is estimated as its distance from the known matched landmarks. Thin-plate splines (TPS) [63, 64] first used RBF for image registration, which is still widely used in several applications that range from sparse to dense. Moreover, distance functions are usually defined variably to handle different scenarios, where the Gaussian distance function uses a kernel parameter to control the global samples to influence the local deformations. Another commonly used model is free-form deformations (FFDs) [65], which is based on gridding images into several rectangular cells that become deformed under the influence of the control points, and the dense deformation is usually conducted with cubic-B splines [66–69]. Other interpolation-based models include elastic body splines [70], basis functions from signal representation [71,72], and locally affine models [73,74]. Readers can refer to [29,75,76] for more details and evaluations.

(3) Optimization method

Once given a measure metric and transformation model, and obtaining the target or energy function like Eq. (1), it also requires an optimization method to search the optimal transformation from the solution space to best align two images. Obviously, the choice of optimization methods may largely impact the matching accuracy and efficiency. In accordance with the nature of variables that optimization methods try to infer, a brief category on them would be *continuous methods* and *discrete methods*. Continuous optimization assumes the variables as real values that require the objective function to be differentiable. Representative methods of this type are gradient descent methods, conjugate gradient methods, and quasi-Newton methods. A discrete method solves the problem by assuming its solution space as a discrete set. Representative ones are graph-based [77], message passing [78], and linear programming methods [79,80]. A probabilistic graphical model (*e.g.,* Markov random field) is often applied to formulate the matching task and solved by these discrete optimization methods. Several heuristic and metaheuristic methods, such as greedy [81,82] and evolutionary algorithms [83,84], are also investigated to explore a larger solution space, thus being able to handle a more general problem, but they cannot guarantee their optimal solutions.

In a word, the choice of optimization method is supposed to consider the nature of objective functions and the structures they can optimize. Traditional optimization methods have been sufficiently studied; we refer the readers to [29] for more details. Over recent years, an increasing number of studies have been using deep features captured by CNNs to guide the conduct of optimization. More inspiringly, these optimization methods can be replaced by deep regressors to directly estimate transformation parameters or displacement fields with data-driven strategies.

2.1.2. Learning-based framework

Traditional area-based image matching is typically performed in an iterative framework, which consists of proper designs of similarity measurement, transformation model, and optimization method. This traditional pipeline is limited by its low computational efficiency and handcrafted measure metrics. The emergence of deep learning techniques has alleviated this predicament and has been widely studied for the image matching task, particularly in the medical community. In general, existing literature has successfully embedded deep learning techniques into the traditional pipeline to drive an iterative optimization procedure or directly estimated the geometrical transformation parameters or deformative field in an end-to-end manner.

(1) Deep iterative methods

The most intuitional approach is to use deep learning networks to estimate the similarity measurement for the target image pair to drive an iterative optimization procedure. In this way, classical measure metrics, such as the correlation-like and MI methods, can be substituted with more superior deep metrics.

Many researchers attempt to train superior measure metrics with the stacked autoencoder [85–87] or some simple CNNs structures [88,89]. The combination of deep similarity metrics and handcrafted ones is also applied as an enhanced measurement for image registration [90]. These deep similarity metrics have shown promising advantages in MMIR particularly for some challenging cases where handcrafted metrics have very little success. Deep metric learning is supposed to couple with the traditional optimization method and a predefined transformation model to achieve image registration, thus requiring a long execution time and sufficient aligned image (or image patch) pairs for supervised training.

Another deep iterative method is to use the reinforcement learning (RL) paradigm to iteratively estimate the transformation parameters [91]. Given an environment and its current state, RL commonly trains artificial agents to predict best actions by maximizing the cumulative expected rewards, thus driving the iterative procedure instead of relying on traditional optimization methods. In the image registration community, the trained agent could be single [91–93] or multiple [94], and can handle both rigid [91,92,94] and non-rigid [93] image registration problem. However, these methods still suffer from time requirements due to their iterative nature, and they should conquer the limitations of optimizing from a large solution space when addressing high-resolution nonrigid image registration.

(2) Deep transformation estimation

Considering the slow registration in these iterative methods, particularly for the high-dimensional parametric space in deformable cases, an increasing number of studies are focusing on directly estimating the geometrical transformation parameters or deformative field in one step. According to the training strategies, these deep transformation estimation methods can be broadly classified into supervised and unsupervised methods.

Supervised methods. Fully supervised methods commonly require ground-truth data to define their loss functions. The biggest challenge is to obtain sufficient samples, with the ground-truth transformation parameters being known, for supervised training. To this end, in addition to using existing labeled datasets of real scenarios [95,96], many data generation strategies are proposed to enrich the diversity of training samples, which are typically synthesized by transforming aligned data through randomly or manually selected transformations [94,97] or learned plausible deformations [98]. This data synthesis strategy is more challenging for deformable registration due to added difficulties in defining their ground truth. In addition, domain adaptation modules [99] or statistical appearance models [100] are often used to ensure that these synthesized data can better simulate the real transformations.

Once enough training samples with their true transformation parameters have been obtained, another critical component is to define the loss functions. In the supervised learning procedure, the loss can be intuitively defined based on the bias between the predicted and ground-truth transformation parameters, which can be directly measured by handcrafted matching criteria or metrics. In addition, some simple CNN structures are sufficient to output few parameters to represent static transformations. In deformable cases, fully convolutional (FC) layers are usually needed to represent the high-dimensional parametric space and output deformable fields or displacement vector fields [101], such as FlowNet [100], DVFNet [102], and U-Net [103,104]. Fully supervised methods highly rely on the size and diversity of training data, which stimulate the development of more sophisticated ground truth generation [105].

Another training strategy is the combinational use of ground-truth data and similarity measurements as supervision. This strategy involves using both the similarity between the predicted and ground-truth transformations, and the similarity between the warped and fixed images to train their networks [105]. To some extent, additional supervised information can achieve superior registration performance. Weakly supervised strategy, which uses segmented label similarity to construct the loss function, is also applied to learn to estimate the transformations [103,106]. Following these ideas, apart from the above-mentioned networks, generative adversarial networks (GANs) [107] are also widely used in this registration pipeline [108,109]. The generator is trained to estimate the transformations, while the discriminator identifies if the aligned image pairs are based on ground truth transformations or predicted transformations.

In a word, the principal goal of these supervised transformation estimators is to use different regularization terms to force the predicted transformations to be realistic or close to the ground truth. This type of method has greatly accelerated the registration procedure with deep techniques due to their one-step nature. This goal is more achievable in a GAN framework, but this method is still limited by the requirement of a large amount of annotated ground-truth data that typically rely on the expertise of the practitioner [100].

Unsupervised methods. To alleviate the limitations of annotated ground truth and inspired by the successful use of spatial transformer network (STN) [110], numerous unsupervised methods are proposed to predict the geometrical transformations in an end-to-end manner. Methods of this type use only traditional similarity measurements (*e.g.,* NCC, SSD, MSE, NMI), together with a regularization term that constrains the complexity or smoothness of the transformation model, to construct loss functions [111–113]. Network structures, which are similar to those used in supervised methods, are applied in an unsupervised form without using any manually annotated data. Thus, to enhance the registration performance, many researchers attempt to latently learn the similarity measurement in their networks, such as learning the relationship between image similarity metric and target registration error (TRE), applying symmetric diffeomorphic transformation-based learning [114], and using GAN framework to implicitly learn to measure the accuracy of image alignment [112,115]. Other methods also learn feature representations from the raw images then use them to train a deep transformation estimator [116].

Several efforts are also made to cope with multimodal image pairs. Typical strategies are the use of both image intensity and gradient information to feed to CNNs [113], image binarization then calculating the Dice score between warped moving images and fixed images [117], and the use of cyclic constraints that are proposed for style transfer together with crafted metrics [112].

2.2. Feature-based pipeline

Feature-based pipeline usually follows a procedure of feature detection, feature description, and feature matching [1]. This pipeline is used more widely in the image matching community because sparse features can be regarded as a simple representation for an image, thus being more flexible and robust to geometric deformation and noise [2]. In the following, we introduce classical handcrafted feature detectors, descriptors, and matchers, and those that have been proposed in recent years. The learning-based methods in each step will be emphatically reviewed. We refer the readers to a recent survey [2] for more details. A flowchart of this pipeline is illustrated in Fig. 3.

2.2.1. Feature detection

The detected features usually represent specific semantic structures in an image or the real world, and can be classified into corner [118–120], blob [121–123], line/edge [124–127], and morphological region feature [128–130]. In contrast to line and region feature, point feature is more acceptably used in the image matching community because of

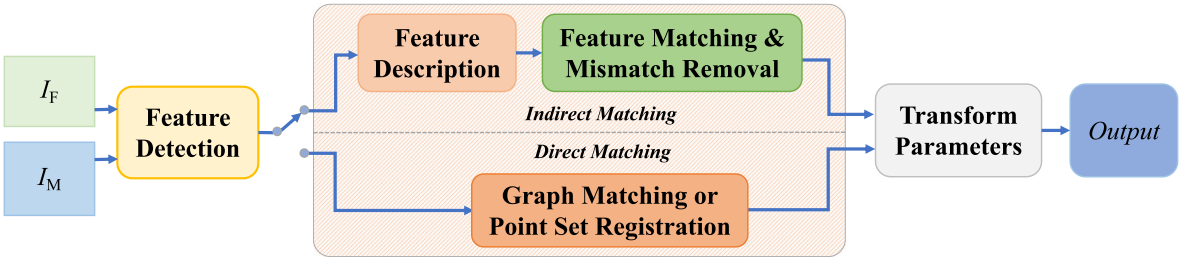

Fig. 3. General framework of traditional feature-based image registration.

its simplification and easily extracted and defined nature. The extracted line or region features are usually converted into point forms if they are used for matching [131–135]. The core idea for feature detection is to construct a response function to distinguish point, line, and region from one another, along with flat and nondistinctive image areas. This idea can be subsequently classified into *gradient-, intensity-, second-order derivative-, contour curvature-, region segmentation-, and learning-based detectors* [2]. In the following, we will briefly introduce the main procedure and typical methods for each category and mainly focus on the latest handcrafted and learning-based methods that are designed for MMIM.

(1) Corner feature

A corner feature is defined as the crossing point of two straight lines typically located in the texture area or edges. Representative responses for corner feature extraction are gradient-, intensity-, and contour curvature-based methods [1,2,136,137].

Gradient-based feature response is implemented based on the first-order information of an image, which is derived from the local intensity autocorrelation in shaking windows proposed in [138]. This strategy was improved by Harris [118] to ease the anisotropy and computational burden, thus making it invariant to orientation and illumination. The Harris detector uses a two-order moment matrix to formulate intensity changes and distinguish corner features based on the magnitude of eigenvalues. To make Harris corners' location more accurate and distributed, [139] further improved it for better tracking performance. The gradient-based strategy makes the Harris feature popularly used in MMIM because the gradient can well describe the structure information that is preserved in two images of different modalities [135,140–143].

The intensity-based corner detector, *a.k.a.* template- or intensity comparison-based detector, aims to simplify the gradient computation by comparing the intensity value with its surrounding pixels. This binary comparison strategy can largely save the execution time and is thus widely used in many storage and real-time required applications. An early approach in [119] used the brightness dissimilarity between the local radius region pixels and the nucleus to identify corner and edge features. The same concept, with a comparison of the center against the pixels along a circular pattern, was used in the famous FAST detector [144], which was further improved to enhance the repeatability [136] and robustness [145]. In particular, Rublee et al. used a grayscale centroid method to assign a main direction for each feature, making it orientation invariant, and proposed the well-known ORB feature. A more recent method called saddle-like detector was introduced in [146]. This local pixel comparison strategy computes quickly and easily extracts sufficient interest points along the texture area of the image, which is why it is also widely used in MMIM [147–149].

The curvature-based strategy aims to extract the corner point by searching a maximum curvature along the detected image curve-like edges or contours. This method usually follows a pipeline of curve (edge/contour) extraction, curve smoothing, curvature estimation, and threshold selection [150]. A curve feature detector is first needed in curvature-based corner detection, which can be conducted with several off-the-shelf methods [124–126]. Subsequently, a smoothing method is required to suppress the impact of noise. To this end, a direct approach by using Gaussian methods [151,152] can ease noise but may change curve locations, whereas an indirect approach, such as region-support-based or chord-length-based method [153,154], can better preserve the curve locations. In the next step, the design of the curvature can be regarded as a point feature response, which is further used to identify distinctive interest points through a threshold strategy. The curvature can be estimated in an algebraic or geometric form, such as cosine, local curvature, and tangential deflection [151,152,155]. A significance measurement can also be used to approximate the curvature response by counting the support structure cues, such as points [156], distance [153], and others [157,158].

(2) Blob feature

A blob feature is typically defined as a local closed region, inside which the pixels are considered similar and thus distinct from surrounding neighborhoods. Two strategies are commonly applied to extract stable blob features: second-order partial derivative (SPD)- and segmentation-based detectors [2].

SPD-based detectors are usually based on the Laplacian scale space and/or Hessian matrix calculation for scale and affine invariant. The feature extracted by this idea, *a.k.a.* blob feature, can be denoted by (x, y, θ), with (x, y) being the pixel coordinate of the feature location and θ indicating the blob shape information including scale and/or affine. This type of features is derived from the Laplacian of Gaussian (LoG) [159], which detects a local extremum point or region with normalized Gaussian response in the multiscale space. LoG is approximated by the difference of Gaussians (DoG) in the famous SIFT method [121,160] to reduce computation. SIFT extracts the potential keypoints as the local extrema in a DoG pyramid and filters them using Hessian matrix of the local intensity values. This procedure is further accelerated by the SURF method [122] using Haar wavelet calculations and integral image strategy, which can significantly simplify the construction of the SPD template. Many other improvements based on SIFT or SURF, such as enhancing affine invariant [161], efficiency [162], and repeatability [163], have been investigated.

The core idea of the segmentation-based method is to fit an optima ellipse for each segmented morphological region for blob feature detection or to use the contours or boundaries for corner feature searching. The segmented regions are usually irregular and based on constant pixel intensities or zero gradient, thus being able to remain stable against threshold changes. One of the famous region segmentation-based detectors is the maximally stable extremal region (MSER) [128]. It extracts a blob feature based on brighter or darker extremal region searching. Kimmel et al. [130] extended MSER to exploit shape structure cues, and many other improvements are introduced for higher discrimination, such as using curvature images [164,165], color information [166], or other segmentation basis [167]. More recently, by using the intersection of the boundaries of three or more regions that are built from existing superpixel segmentation methods, Mustafa et al. [168] proposed an MSFD corner detector for better wide-baseline image matching. We refer the readers to [129,169] for more comprehensive introductions.

With their robustness, discrimination, and location accuracy, SIFT-like methods are widely used in various applications. Many researchers have successfully improved SIFT or SURF to eliminate the modality

gap, thus achieving MMIM, including the matching of retina [170, 171], multispectral [172–174], optical-to-SAR images [175–177], and visible-to-infrared (VIS–IR) images [178].

(3) Learnable feature

Before deep learning, many detectors use classical learning (training a classifier) to identify more reliable and matchable features before matching, such as FAST [179], ORB [120], and others [145,180,181]. In recent years, deep learning has shown great potential in keypoint detection particularly from two images with a significant appearance difference, which often occurs in cross-modal image matching. The core idea of CNN-based detectors is to generate a response map and then search salient point locations, which is conducted as a regression problem that can be trained in a differentiable manner under geometrical transform and image appearance invariance constraints. In general, this type of method can be classified into supervised [123,182,183], self-supervised [184,185], or unsupervised methods [186–190].

Supervised methods have shown the benefits of using anchors (*e.g.,* obtained from the SIFT method) to guide their training, but the performance could be largely restricted by the method of anchor construction, because the anchor itself is intrinsically difficult to reasonably define and may prevent the network from proposing new keypoints in case no anchor exists in the proximity [190]. Self-supervised and unsupervised methods train detectors without any human annotations, and only the geometric constraints between two images are required for optimization guidance; a simple human aid is sometimes asked for pretraining [185]. In addition, many methods integrate feature detection into the entire matching pipeline by jointly training with feature description and matching [123,188,191–193], which can enhance the final matching performance and optimize the entire procedure in an end-to-end manner. We refer the readers to [2,194,195] for more details.

Several detectors, such as TILDE [182], are also trained under drastic image appearance changes of weather and lighting conditions. In TILDE, a general regressor was trained to predict a score map, whose maxima after non-maximum suppression can then be regarded as potential interest points. The learning strategy can also be combined with handcrafted methods to improve performance in MMIM [196,197]. In a word, CNNs can capture overall structure information and high-order cues such as semantic information from raw images, thus intrinsically bridging different modality images to extract matchable feature points.

2.2.2. Feature description

Feature description refers to mapping the local intensity around a feature point into a stable and discriminative vector form, enabling the fast and easy matching of the detected features. This step requires the generated descriptor of two matched features to be as close as possible and for two unmatched features to be far apart in the descriptor space, simultaneously being robust to geometrical transform, image appearance changes, and image quality. Descriptor design is the most critical part in feature-based MMIM, directly determining the final performance. Descriptors for general image matching task have difficulty building correct point correspondences between multimodal image pairs, thus requiring researchers to modify these methods for specific modality variance data. According to the used image cues (*e.g.,* gradient, intensity) and the form of descriptor generation (*e.g.,* comparison, statistic, and learning) [2], we classify existing descriptors into float, binary, and learnable descriptors, which can be subsequently classified into *gradient statistic-, local intensity comparison-, local intensity order statistic-, and learning-based methods*.

(1) Float descriptors

Float descriptors are often generated by statistic methods based on gradient or intensity cues. The core idea of gradient statistic-based descriptor is to calculate the orientation of the gradient [198] to form a float vector for feature description. The most relevant method SIFT [121] uses this strategy for each DoG feature, which is known to be scale-, rotation-, and illumination-invariant. Based on a similar concept, SURF [122] uses Haar wavelet response and integral image to simplify the gradient computation and achieve high computation efficiency. Many variants of SIFT or SURF tend to obtain better performance in computation efficiency, robustness, and discrimination, such as using color or affine information [161,199] or a square root kernel measurement [200] and gradient statistic in different domain sizes [201], just to name a few.

Another statistic strategy is based on the orders of pixel values rather than raw intensities, which has been demonstrated to be superior in monomodal visible image matching [202,203]. Pooling by intensity orders encodes ordinal information into the descriptor. This scheme may enable the descriptors to be rotation-invariant without the estimation of a reference orientation as SIFT, which appears as a major error source for most existing methods. Representative methods of this type include generating a descriptor by normalizing the captured texture information and structure information with an ordinal and spatial intensity histogram [202] or pooling local features based on their gradient and intensity orders in multiple support regions [204].

SIFT-like descriptors are widely modified for MMIM due to their distinctiveness and robustness. For example, symmetric-SIFT has been adapted to be used for multimodal registration [205], which is further improved in [206]. Other applications and modifications include performing visible and IR image registration based on morphological gradient and C-SIFT [178], improving the SIFT algorithm in its feature selection strategy [173], or adapting it to suit the characteristics of remote sensing images [175,207].

(2) Binary descriptors

Binary descriptors are typically based on the comparison strategy of local intensities. The core challenge in this method is the selection rule for comparison. A more representative method is the BRIEF descriptor [208], which is built by concatenation of the results that are created by a binary test of intensities for several random point pairs in an image patch. On the basis of this concept, the well-known ORB [120] feature is proposed by integrating a rotation-invariant strategy and using machine learning for selected robust binary tests, thus alleviating the limitations in rotation and scale change. Other binary selection and comparison rules include concentric circle sampling strategy with increasing radius [209] and comparing image intensities over a retinal sampling pattern [210].

A descriptor of this type can perform the feature description with low computation and memory requirements due to its simple comparison strategy. However, this approach may simultaneously sacrifice great discrimination and robustness, which would be worse in multimodal cases where the image pairs have remarkable nonlinear intensity variance. Therefore, binary descriptor is not a priori choice for MMIM.

(3) Learnable descriptors

The data-driven strategy in learning-based descriptors can largely enhance discrimination because high-order image cues or semantic information between two images can be extracted in CNNs, thus showing great potential to describe images of different modality. An early classical learning strategy has been adapted in many handcrafted descriptors with the aim to reduce the dimensions of descriptors [211–213], convert float descriptors into binary ones [214–216], and directly learn binary representations from raw patches [217,218].

CNN-based feature description has stimulated the emergence of an increasing number of deep descriptors in terms of training strategy, model structure, and loss design. The primary goal is to learn a representation that can enable the two matched features to be as close as possible while the unmatched ones are far apart in the measuring space [219]. In general and in accordance with the output of deep models, existing deep descriptors can be briefly classified into metric learning [220–223] and descriptor learning [5,224–229]. The former often learns a discriminative metric to predict whether a pair of raw patches or generated descriptors is matched or not. By contrast, descriptor learning tends to generate the descriptor representation from raw images or patches. This strategy is more flexible

and economical because it avoids repeated computation, unlike metric learning. Another meaningful task that integrates the feature description with the detectors into the complete matching pipeline has also been widely studied in recent years [123,185,188,191,192]. The optimization of coupled feature detection, description, and matching can create superior matching performance.

As for multimodal images, deep descriptor can be applied to predict matched labels for later registration tasks [230] or combined with handcrafted features for better registration [196,197]. Except for feature matching of the same target or scene, semantic matching [231–235] for images of similar targets/scenes (such as matching between dog and cat) has also been investigated. Data-driven strategy is a suitable choice in semantic matching and can achieve promising accuracy by using CNNs to understand the semantic similarity. In this regard, learning-based method has shown considerable potential in addressing MMIM tasks.

2.2.3. Feature matching

Suppose we have obtained a set of M interest points $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{M}$ from the moving image and N interest points $\mathcal{Y} = \{\mathbf{y}_j\}_{j=1}^{N}$ from the fixed image. Feature matching aims to establish correct feature correspondences from two extracted feature sets; this step can be conducted in an indirect or direct manner, which corresponds to the use or non-use of local image descriptors. The modality difference has been suppressed in feature detection and description parts, which is why the matching step can be performed well by general methods.

A direct method is to correspond these two sets by directly using the spatial geometrical relations and an optimization method. Two representative strategies are widely studied: *graph matching* and *point set registration*. The indirect pipeline commonly casts feature matching as a two-stage problem. In the first stage, a putative match set, *i.e.*, $\mathcal{S} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{L}$, is constructed based on the similarity of local feature descriptors, where $L \leq \min\{M, N\}$ denotes the number of established putative matches, and it depends on which matching criteria are used [120,188,236]. In the second stage, the false matches in $\mathcal{S}$ are rejected by imposing additional local and/or global geometrical constraints. Generally speaking, the indirect pipeline can be further classified into *resampling-based*, *nonparametric model-based*, and *relaxed methods*. The *learning-based methods* in each category will be reviewed in the corresponding part.

(1) Graph matching

Graph matching (GM) aims to construct a graph for each point set by defining the nodes and edges in it, then establish the point correspondences by maximizing the overall affinity score with a graph structure similarity priori. It usually formulates point set matching as a quadratic assignment problem (QAP) [237]. GM can be briefly classified into exact matching and inexact matching. The former refers to finding the bijection of two binary (sub)graphs with the requirement that all edges are strictly preserved [238–240]. This strict condition may result in poor performance in real-world applications.

By contrast, inexact matching has good flexibility and efficacy in practice, thus attracting more research interest. Most methods in this category tend to relax the constraints into an affordable form, thus giving rise to various GM solvers. In detail, spectral relaxation methods commonly convert this task as an eigenvector solving problem, which can be solved by a discretization strategy [241,242], replicator equation from evolutionary game theory [243], nonnegative matrix factorization approach [244], a probabilistic interpretation strategy [245], or by relaxing the assignment matrix to be orthogonal [246,247]. As a result of the relaxation, these methods are more efficient but less accurate. Convex relaxations are used to relax the original nonconvex QAP problem into a convex form with theoretical guarantees, which are generally solved by semi-definite programming [248–250] or linear programming [251,252]. In recent years, the dual problems of linear programming relaxation have been widely investigated to solve the GM problem [253–257]. Convex-to-concave relaxations [258,259] aim to gradually achieve convex-to-concave procedure of the original problem by using path-following approaches. Continuous relaxation methods [260–263] can approximate the QAP issue and solve it in an accurate or efficient manner, but without global optimality guarantee. Many others, such as random walk method [264] and Monte Carlo method [265,266], are also investigated for the GM problem. In addition, multigraph [267–271] and hyper-graph matching [272–274] are studied actively, which refers to jointly matching multiple graphs with consistent correspondences and formulating GM in a high-order form to mostly explore the geometrical cues, respectively.

In recent years, deep learning technique has been widely used to address the GM problem. Affinity matrix plays a key role in the GM problem, which is why most deep learning techniques tend to obtain a better representation for it in a supervised [275] or unsupervised [276] form. Many researchers even cope with the GM problem in an end-to-end manner [277,278] by simultaneously learning the node/edge affinities and solving this combinatorial optimization problem.

(2) Point set registration

Different from GM methods, point set registration (PSR) assumes that a global transformation model between point sets exists and is known beforehand, then iteratively estimates the model parameters and point-to-point correspondences. To this end, an increasing number and variety of techniques are being proposed to improve the robustness and solving efficiency. One of the most representative schemes is iterative closest point (ICP) [279] and its variants [280–283]. This method is improved by soft assignment and deterministic annealing strategy under a robust point matching (RPM) framework [284,285]. Another representative pipeline that combines RPM with Gaussian mixture models (GMMs) is also widely studied [286–289], and it is commonly optimized by expectation–maximization (EM) solvers. Many other density-based and optimization-based methods are introduced for robustness and efficiency enhancement, such as kernel density method [290], support vector parameterized strategy [291], or fuzzy clustering method [292], while optimization-based methods include stochastic optimization approaches [293,294], branch and bound [295–299], and semi-definite programming method [300].

Apart from methods that focus on exploring the model formulation and optimization methods, some PSR methods aim to construct shape descriptors from the point set, then establish sparse point correspondences by using the similarity constraint of descriptors, followed by a robust estimator for global transformation parameters estimation [301–303]. Typical points descriptors include shape context (SC) [131], spin images [304], integral volume [305], and point feature histograms [306].

(3) Indirect methods

The most classic methods for mismatch removal and parameter estimation are resampling-based methods, which are also known as random sample consensus (RANSAC) [301] and its variants [307–311]. The common idea of these approaches is to obtain the smallest consistent inlier set to fit a given transform model following a hypothesize-and-verify strategy. Many improvements based on RANSAC mainly focus on verifying the model quality, such as using a maximum likelihood procedure in MLESAC [307,308] or modifying the sampling strategy as inspired from the specific properties of inliers, such as spatially consistency-based [312], by using a grouping strategy [309] or inlier probability prediction [313]. Several studies address this issue in a local optimization procedure [314,315], or a progressive growing sampling procedure [311,316,317] to eliminate the need for user-defined thresholds such as inlier–outlier decision. A universal framework that integrates many improving strategies of RANSAC is introduced in [310]. Notably, resampling-based methods may largely rely on the resampling strategy and suffer considerably and even fail if the image pairs undergo serious nonrigid deformations. Moreover, the theoretical execution time exponentially grows with the increase in the outlier ratio.

Nonparametric model-based methods are developed to handle both rigid and nonrigid transformations, thus showing more flexibility. Such methods are representative by defining deformation functions in a high-dimensional form, such as triangulated 2D mesh [318] or a kernel representation in reproducing kernel Hilbert space with Tikhonov regularization [319–321]. These methods are typically optimized through tailored robust optimizers, such as Huber estimator [322], L2E [320], support vector regression [323], or EM solution in a Bayesian model [319,321,324–327].

Another active topic is the investigation of relaxed methods for mismatch removal. These relaxation rules are typically based on the assumption of locality or piecewise consistency [328], such as grid-based motion statistics (GMS) [329], locality preserving matching (LPM) [330] and its varieties [331,332], feature matching using spatial clustering with heavy outliers [333], and coherence-based decision boundaries [334]. Other strategies include using a filtering theory [335,336] or Markov random field formulation [337]. These methods are efficient and can handle both rigid and nonrigid image matching. Thus, they are more acceptable for real-time required applications because they achieve solutions quickly by using less-strict geometric constraints. However, these relaxed methods are commonly sensitive to parameter setting and greatly rely on a dense match set, which should well maintain the local coherency of correct matches.

In recent years, a learning technique has been widely studied and equipped to eliminate the outliers and/or estimate model parameters through training a deep regressor [338–340] or classifier [341–343]. In general, parameter regression and inlier/outlier classification are trained jointly for performance enhancement [341]. Deep regressor is inspired by the classical RANSAC and aims to estimate the transformation model, such as fundamental matrix [344] or epipolar geometry [339]. For example, a differentiable RANSAC, namely, DSAC [338], is trained in an end-to-end manner by using reinforcement learning, while other efforts are made to improve the sampling strategy [339,340]. As for classifier learning, Yi et al. [341] first introduced a pipeline called LFGC to find good feature correspondences by training a network from a set of putative match sets together with their image intrinsics. Ma et al. [342] proposed LMR, a general two-class classifier learning framework for mismatch removal, by using a few training image pairs and handcrafted geometrical representations for training and testing. Zhang et al. [345] focused more on geometrical recovery in their order-aware networks. Apart from learning with multilayer perceptron (MLP), another method conducted this task with graph convolutional networks (GCN) [346].

Learning from point data is not as easy as that on raw images by using deep convolutional networks due to the unordered structure and dispersed nature of the sparse points. Even so, this approach is still worthy of attention because many recent studies have shown great potential in using the GCN and MLP network structure together with tailored normalization terms to learn to address sparse point-based tasks [2,333].

2.3. Summary

Existing matching methods for multimodal images could be systematically classified into area-based and feature-based pipelines. The area-based method would be largely affected by the choice of measure metric. Two possible strategies are commonly applied for MMIM: one is the use of a modality-independent measure metric such as MI and its varieties, and the other is the reduction of different modalities into a common domain. However, area-based methods are limited by their large computational burden and their requirement for image pairs to have large overlaps and undergo slight geometrical deformations. Deep metric learning or deep transformation estimation in this framework dramatically alleviates the predicament in matching criteria design and iterative optimization. However, this learning strategy is still restricted by the high image resolution and large or complex deformations, particularly with the requirement of sufficient training data. A feature-based pipeline can efficiently address the problem in geometrical deformation. Direct feature matching, such as graph matching and point set registration, is more suitable for image pairs containing less texture (even binary images), or heavy modality or semantic variances. In these cases, the image patch-based descriptor would be invalid. However, the graph structure among potential true point correspondences may be steadily preserved, which requires an overall corresponding matrix to be optimized to find an optimal solution. But these direct feature-based matching methods are limited by high computational burden and outlier sensitivity. As for indirect feature matching, extracting a high number and ratio of interest points then constructing distinctive descriptions and corresponding them accurately are difficult because of the significant nonlinear intensity variance between two modalities. Proposing an advanced paradigm with better registration performance in terms of both accuracy and efficiency remains an open problem for researchers.

3. Modality-independent review

As aforementioned, matching for multimodal images can be seen as one specific case. Apart from these challenges presented in the general image matching task [2], the primary one in the multimodal case also includes eliminating the domain or modality gap between two input images. However, the natures of images would vary considerably across different imaging sensors or data types in different research areas. Thus, in the following, we will review these typical and latest image matching methods that are designed for specific multimodal scenarios under different research areas, including medical, remote sensing, and computer vision. Simultaneously, the image natures of each modality will be analyzed first in the corresponding parts to reinforce the understanding of the challenges and need for image registration of different modalities.

3.1. Medical

The biggest family of MMIR may lie in the medical community. With the rapid development of visualization of computer imaging techniques, medical imaging has developed from statistic to dynamic, plane to solid, and morphological to functional imaging, which has played a significant role in modern medical diagnosis. Commonly used medical imaging techniques include X-ray, ultrasonic imaging (UI or US), computer tomography (CT), single-photon emission computed tomography (SPECT), magnetic resonance imaging (MRI), positron emission tomography (PET), and functional magnetic resonance imaging (fMRI), generating medical images of different modalities. From the perspective of medical applications, these modalities can be briefly classified into anatomical images such as CT and MRI, and functional images such as PET, fMRI, and SPECT [347]. Anatomical images have high spatial resolution that can clearly display geometrical information, such as the anatomical structures of viscera and bones, but without any functional information. In contrast, functional images can well display the functional transformation during the metabolic procedure, but the images are usually not clear enough to reveal the structure information, resulting in difficulties in anatomical structure and boundary localization. Consequently, the complementary information from the images of these two types needs to be combined, which first requires the image pairs to be spatially aligned. The registration target also includes the MRI with different weights, such as T1, T2, and proton density (PD), or the retinal images of different angiographies such as digital subtraction angiography (DSA), fundus photography and fluoroscopy angiography (FA).

In the field of medical research, MMIM is a hot topic and has been giving rise to an increasing number and diversity of registration techniques. Several main strategies have been proposed to solve this

problem, including (1) information theoretic similarity measurement; (2) reduction of the multimodal problem into a monomodal problem (modality unification); (3) interest point extraction and matching-based pipeline. All these strategies can be implemented by using deep convolutional networks, which will be the main focus in a new subsection.

3.1.1. Information theory-based

In past decades, information theoretic similarity measurements successfully alleviated the gap between multimodal image pairs in the registration task, which have been widely investigated and extended into more advanced forms. This step benefits from the successful use of MI, introduced and popularized by Viola and Wells [46,47], and Collignon and Maes [348,349]. In recent years, Maes et al. [347] recognized that the MI measure gave rise to revolutionary breakthroughs in the MMIR task. However, the widespread use and study of MI have revealed some of its shortcomings. Primarily, it is not overlap-invariant. Thus, MI may be maximized in certain cases when the images become misaligned.

Following the pipeline of maximizing the MI score for MMIR, numerous advanced information theoretic approaches have been investigated to remedy the abovementioned shortcoming. For example, Studholme et al. [48] proposed a normalized version of MI, namely, NMI, to better register slices through clinical MR and CT image volumes of the brain. An upper bound on the maximum MI [49] is studied for deformable image registration, which can provide further insight into the use of MI as a similarity metric. In addition, cMI [50] is proposed as an improved similarity metric for nonrigid registration. cMI is conducted as a 3D joint histogram based on both intensity and spatial dimensions, and incorporated in a tensor-product B-spline nonrigid registration method by using either a Parzen window or generalized partial volume kernel for histogram construction. In [350], the authors proposed a hybrid strategy that combines the spatial information with MI to achieve multimodal retinal image registration.

Many researchers utilized the divergence measures to compare the joint intensity distributions in MMIR, including Kullback–Leibler divergence (KLD) [52,53] and Jensen–Shannon divergence (JSD) [54]. The use of Renyi entropy [351,352] has also attracted great attention in the registration problem, which is conducted with minimum spanning tree or spanning graphs [353], or by integrating with KLD [354] for better generalization.

Considering that these statistic measurements are commonly based on a single pixel joint distribution model, the statistic criteria are also implemented in a global or local region. For example, building from a linear weighted sum of local evaluation of MI, Studholme et al. [51] proposed RMI to reduce the errors caused by local intensity changes. Others used octrees [355] or locally distributed functions [356].

Many researchers have been paying increasing attention to optimization methods to quickly and accurately estimate transformation models. Wachowiak et al. [83] considered that local optimization techniques frequently fail, because these metric functions with respect to transformation parameters are generally nonconvex and irregular during an area-based procedure. Hence, they modified an evolutionary approach that involves particle swarm optimization for biomedical MMIR. Arce et al. [357] used the MRF coefficient under a Bayesian formulation to model local intensity polynomial transformations, while local geometric transformations are modeled as prior information with MRF to register both rigid and nonrigid brain images of MRI T1 and T2 modalities. Moreover, Freiman et al. [358] presented a new nonuniform sampling method for the accurate estimation of MI in multimodal brain image rigid registration. This method uses 3D fast discrete curvelet transformation to reduce the sampled voxels' interdependency by sampling voxels that are less dependent on their neighborhood, thus providing a more accurate estimation of the MI. Following the NMI and FFD registration pipeline, Yang et al. [359] introduced a cooperative coevolving-based optimization method that combines the limited-memory Broyden–Fletcher–Goldfarb–Shanno with boundaries (L-BFGS-B) and cat swarm optimization for nonrigid MMIR. In this method, the block grouping strategy can capture the interdependency of all variables, thus achieving fast convergence and better registration accuracy of 3D CT, PET, and T1-, T2-, PD-weighted MR images.

In recent years and with spatial information taken into account, Legg et al. [360] proposed feature neighborhood MI in their two-stage nonrigid registration framework to align paired retinal fundus photographs and confocal scanning laser ophthalmoscope (CSLO) images. This improved MI is superior to many existing MI variants, such as original MI, gradient MI, gradient-image MI, second-order MI, regional MI, feature MI, and neighborhood incorporated MI.

The methods explored in the early part of this decade were comprehensively reviewed in [29]; readers may refer to this work for more details. The measure metrics that use deep learning are reviewed in the part of learning-based methods.

3.1.2. Modality unification-based

Another strategy aims to transform two different modalities into a common domain, making it workable for general measuring metrics that are successfully used in monomodal image matching. Two possible ways are used to reduce this problem into a monomodal one: simulating one modality from another and mapping both modalities into a third one. In this part, we review typical and related handcrafted approaches that follow this idea. Approaches that use deep networks will be introduced in the learning-based methods, such as style transfer learning and descriptor learning. Refer to the corresponding part for more details.

Following this strategy, several studies aim to map one modality to another according to the physical properties of the imaging device. To register US and MR images, Roche et al. [361] transformed MR images into the domain of US images on the basis of their intensities and gradient information. The registration is performed with a rigid model and based on the expended correlation ratio method. Another method in [362], aims to generate pseudo-US images from CT by exploiting the physical principles of US images, thus achieving CT-US rigid/affine registration optimized under a locally evaluated statistical criterion. In addition to mapping one modality to another in a global manner, the local patch-based strategy is also studied to identify the unreliable areas if directly using the MI metric, then the small patches of these areas are simulated to a common domain [363]. The mapping strategy is also conducted with a learning strategy, which will be reviewed in the part of learning-based methods.

Another way to map two different image modalities into a common one is to exploit the morphological information, such as edge or contour structures, which commonly exist in both modalities. Many approaches directly extract these morphological or structure information through filtering [364,365] or by using existing edge extractors. Gabor filtering can easily capture texture information from raw images, which is why it is widely used for modality unification [364,366], as also conducted by several local frequency representations [365].

Local descriptors can also map the target pixel or voxel into a distinctive vector form in a high-dimensional space, making similarity measurement more convenient and thus making the optimization process more effective. Inspired by this idea, many methods aim to reduce the multimodality to uniform domain on the basis of the concept of self-similarity image representation, which was first studied in local self-similarities (LSS) by Shechtman et al. [367]. LSS is a local feature descriptor that can capture the internal geometric layouts of LSS within images and indirectly represents the local image property, which is why it can be used to match two textured regions with significant appearance variance but similar layouts or geometric shapes. In addition, a descriptor called modality independent neighborhood descriptor (MIND) [368] is proposed to extract the distinct structure in a local neighborhood to generate description vectors, thus transforming the images of different modalities into a third domain, whose similarity is easily measured by arbitrary metrics such as SSD. The authors apply this descriptor within a symmetric nonparametric

Gauss–Newton registration framework, and well registered 3D CT and MRI chest scans under rigid and deformable transformations. The use of MIND makes this descriptor robust to image noise and nonlinear variance of image intensity. Therefore, the proposed technology would be applicable in the registration task of arbitrary modalities. The same authors introduced a new structural image representation called self-similarity context (SSC) [369] for efficient computation. SSC is also based on the concept of self-similarities descriptor to represent image patches in a common space, thus making the similarity metric easier to apply. This idea is conducted on the registration of 3D US and MRI brain scans by using a symmetric multiscale discrete optimization and diffusion regularization to search accurate deformation parameters.

The authors in [370] proposed entropy and Laplacian image-based structural representation together with the SSD metric, namely, eSSD, to measure patch similarity, then used it to drive the registration procedure of T2–T1–PD MRIs and PET–CT images under rigid or non-rigid transformations. This method creates an intermediate structure representation in a manifold space by means of a manifold learning scheme with a local distance-preserving constraint. The authors of [371] addressed the weakness of eSSD in terms of noise sensitivity and high computational burden, and they introduced a new similarity metric based on Weber local descriptor (WLD). Diffusion maps are used in [372] to capture the geometric and spectral properties across two images of different modalities, thereby better representing complicated image features, which can improve the registration performance. The self-similarity α-MI (SeSaMI) proposed by Rivaz et al. [373] uses local structural information in a graph-based implementation of MI for non-rigid image registration. Rivaz et al. [374] also proposed contextual conditioned MI on the basis of conditioning the estimation of MI on similar structures. This method considers both intensity values of corresponding pixels and contextual information in its neighboring pixels.

Analogous to MIND, another self-similarity-based local descriptor called Zernike moments-based local descriptor (ZMLD) [375] was proposed to enhance the robustness and discrimination of similarity metrics. The introduced self-similarity is based on the Zernike moments of image patches and also in a local neighborhood to generate descriptors, whose distance in Euclidean space is directly calculated as a measurement criterion. Extensive experiments on T1–T2–PD weighted MR brain images and real MR–CT images, together with spline-based FFD transformations and L-BFGS optimization method, demonstrated the superiority of the proposed method to NMI, ESSD, WLD, and MIND. This idea can be implemented with a binary descriptor, such as discriminative local derivative pattern (dLDP), to encode images of different modalities into similar image representations [376]. dLDP calculates a binary string for each voxel according to the pattern of intensity derivatives in its neighborhood. This descriptor similarity is evaluated using the Hamming distance, which can be efficiently computed.

3.1.3. Feature point-based

Given the reduced texture of medical images, feature-based methods are rarely used in the matching task in this field. However, the global structure and counters or edges commonly exist in both images of different modalities, thus providing matchable information. The target image pairs sometimes cover small overlapping areas or large deformation, which is significantly difficult to optimize in an area-based pipeline with a similarity measurement, thus inspiring researchers to design workable feature detection, description, and matching methods to handle these issues. Kelman et al. [377] evaluated general MMIM toward representative feature-based methods.

This idea is popularly applied to register the retina image pairs of different modalities. For instance, Chen et al. [140] proposed partial intensity invariant feature descriptor (PIIFD) for poor-quality image pairs, and it is performed with Harris features, thus also being called Harris-PIIFD. On the basis of the successful use of the PIIFD descriptor in multimodal retinal image matching, Ghassabi et al. [170] combined the use of UR-SIFT features and PIIFD descriptors, and achieved promising performance. Considering that Harris-PIIFD may fail to correctly align color retinal images with other modalities when faced with large content changes, Wang et al. [378] proposed a robust point matching framework called SURF-PIIFD-RPM for multimodal retinal image registration. In this method, the SURF detector is first exploited to extract matchable point features from two images, then the authors matched them by using the PIIFD descriptor. Subsequently, a single Gaussian robust point matching model, which is based on a kernel method conducted in reproducing kernel Hilbert space, is used to estimate the mapping function in the putative match sets with existing outliers for better matching performance. Following a similar strategy, the combined use of SIFT and PIIFD with an outlier rejection strategy is introduced in [171]. In addition, a residual-scaled-weighted least trimmed squares method is designed to enforce an affine transformation model, which significantly outperforms the Harris-PIIFD scheme. In addition to the aforementioned strategies, a symmetric-SIFT has been adapted to rapidly register CT and MR brain images through rigid transformation estimation [205], which is further improved by [206].

Recently, Li et al. [379] proposed a two-step registration procedure of color fundus (CF) and scanning laser ophthalmoscope (SLO) retina images. In the first step, the mean phase images are generated for feature descriptor matching together with the RANSAC method to estimate the global affine transformation, then MIND is exploited to refine the local accuracy, thus achieving deformable registration.

3.1.4. Learning-based

MMIR with deep networks for medical applications is an active research area and has given rise to an increasing amount and diversity of publications. Similar to traditional methods, the learning strategy is typically exploited to tackle the modality gap, thus easing the registration procedure. As mentioned in the general framework, *i.e.,* Section 2, this concept is usually performed as similarity metric learning, modality transfer learning to reduce multimodal problem as a monomodal one, and end-to-end learning to directly estimate the transformation parameters or displacement field in one step.

(1) Metric learning

Metric learning can be seen as a direct extension of handcrafted similarity metrics, such as SSD- or MI-like methods, and those learned metrics can better guide the iterative MMIR procedure. An early learning-based method in [380] proposed a learned similarity measure in a discriminative manner by means of joint kernel machine learning, such that the reference and correctly deformed images receive high similarity scores. To this end, the authors developed a method derived from max-margin structured output learning and employed the learned similarity measure within a standard rigid registration algorithm. A stacked denoising autoencoder (SAE) was introduced in [87] to learn a similarity measurement for rigid CT and MR image registration. In [89] and learning from aligned 3D T1–T2 weighted brain MR volumes, a deep similarity metric can be obtained and integrated with the gradient descent method to iteratively estimate the parameters of a predefined deformation field. To estimate the parameters of rigid model in 3D US–MR abdominal scan registration, Sedghi et al. [381] used a simple five-layer neural network to learn a similarity metric that can measure the level of registration. This learned metric is then optimized by means of Powell's method in an iterative manner. In [88], the authors proposed a novel strategy to train the networks to obtain a similarity metric by predicting registration performance evaluation such as TRE. In this strategy, the evolutionary algorithm is exploited to explore the solution space, achieving rigid registration for transrectal ultrasound (TRUS) and MR images. Unlike in the above-mentioned methods, Wright et al. [382] trained LSTM spatial co-transformer networks to guide the iteration procedure and then achieve the registration of 3D fetal MR–US brain scans. This deep method is superior to other similarity metrics and self-similarity-based descriptors.

(2) Reinforcement learning

RL is first explored in the medical image registration problem. The first team that used RL for MMIR was Liao et al. [91]. They tried to achieve the rigid registration of cardiac and spine 3D CT and cone-beam CT (CBCT) images by means of Q-learning with a single agent, in which a greedy supervised approach together with an attention strategy is exploited for end-to-end training. Ma et al. [92] performed RL for rigidly registering depth and CT image pairs, in which the Q-learning strategy is borrowed to extract feature representation to eliminate the appearance variance of two images, and the contexture information is captured to guide the registration for performance enhancement. To register X-ray and digitally reconstructed radiographs of the spine, Miao et al. [94] designed a dilated FCN in a Markov decision process and multiagent system in their RL paradigm. They used an auto-attention mechanism to capture deep information among multiple regions. The proposed method can largely improve the efficiency and accuracy of the registration.

(3) Supervised learning for transformation model estimation

To avoid the large computational burden in the iterative pipeline such as metric learning or RL-based methods, a growing number of studies estimate the geometrical transformation parameters in an end-to-end fashion. The rigid model is easier to estimate with a deep framework, as first conducted by Salehi et al. [97] and Sloan et al. [383], who used a deep regression network to generate the rigid parameters to align T1–T2 weighted brain MR. The authors in [97] constructed their loss function based on bivariant geodesic distance and initially registered the MR volume groups through a residual network before their correction network. This strategy can enlarge the capture capacity of their network. The authors validated the customized loss and their network on both multimodal 2D–3D and 3D–3D registration problems. The authors of [383] used their deep networks for both uni- and multimodal registration. The parameters of those layers for extracting deep features are shared in the monomodal case but are learned separately in MMIR.

A more challenging problem is to learn to estimate the deformable model parameters or deformation fields in one step in nonrigid cases. Yang et al. [101] solved this problem of 3D T1–T2 weighted brain MR volumes by using a low-rank Hessian approximation of the distribution of the nonrigid model parameters. In [384,385], Hu et al. trained their cooperative networks – namely, global net and local net – to estimate global and local refined deformation fields, respectively, and they applied this pipeline to register deformable MR-TRUS prostate images [384]. The whole framework is trained based on the similarity of segmented labels and is further modified as an end-to-end learning form by means of FCN networks with dense displacement field (DDF) as deformation modeling [385]. In [106], Hu et al. aimed to perform deformable MR and TRUS registration by using GAN model by maximizing segmented label similarity and minimizing the adversarial loss. Analogous with the use of label similarity to guide the network training, Hering et al. [103] exploited a U-Net-based architecture and constructed the loss function on the basis of the label and similarity metrics for deformable registration of 2D cine-MR images, which is further applied for cardiac motion tracing.

(4) Unsupervised learning for transformation model estimation

In addition to the use of annotated labels for supervised training, an increasing number of unsupervised methods have also been investigated. To perform 3D MR–US brain volume registration, an unsupervised framework [113] that uses a 3D CNN for feature extraction and deformation model regression is applied. In this method, the similarity metric based on pixel intensity and gradient information is used to achieve unsupervised training. A similar method in [111] trains the deep network guided by traditional similarity metrics such as NCC and based on prealigned image pairs to register deformable pelvic CT and MR volumes. In [386], the authors used the encoder–decoder paradigm to generate modality-independent latent representation to perform in a cycle-consistent way, and they used inverse-consistent loss to guide the STN to learn affine and nonrigid transformations under a simple similarity constraint such as MSE. The proposed method can achieve promising performance on 2D or 3D T1 and FLAIR MR brain data.

(5) GAN for transformation model estimation

A GAN model has also been studied for MMIR of medical slices or volumes. Yan et al. [108] proposed a GAN paradigm to rigidly register 3D MR and TRUS prostate biopsy volumes. The core idea is to force the trained discriminator to identify if the alignment of volume groups is performed by ground-truth transformations or those generated by a generator, and simultaneously requiring a more advanced generator whose output is more like the ground truth, thus being able to cheat the discriminator. Subsequently, Mahapatra et al. [112] constructed the loss terms by combining several similarity metrics, including NMI, SSIM, and SSD, with advanced adversarial and cyclic constraint loss in their GAN framework to cope with MMIR of retinal CF images and fluorescein angiography (FA) images.

More recently and based on the STN and GAN framework, a U-Net is trained in [104] to generate an adversarial loss, thus guiding the training procedure. To avoid the requirement of ground-truth deformation fields, annotated landmarks, or any aligned multimodal image pairs for training, Qin et al. [109] tried to learn a parameterized registration function for modality unification in a latent embedding space in an unsupervised manner. Any meaningful geometrical deformation can be directly derived in the latent space. Three components are designed in the proposed framework: an image disentangling network via unpaired image-to-image translation, a deformable registration network in the disentangled latent space, and a GAN model to implicitly learn a similarity metric in image space. The whole procedure is performed by combining self-reconstruction loss, latent reconstruction loss, cross-cycle consistency, and adversarial loss with similarity metrics defined on latent space. The GAN framework can also translate one image domain to another, thus reducing the image modalities into a common one [387–390]. In this way, an arbitrary similarity metric that is used for traditional area-based registration can be exploited to construct loss terms to guide the network training.

(6) Hybrid methods

A supervised learning approach was introduced in [391] to learn optimization updates for MMIR. This method poses the problem as a regression task to estimate the unknown transformation based on the local structure and global appearance information by means of Haar-like features, and its transformation parameters are modeled by regression forests in a large feature space. A pretrained network is used in [117] as a feature extractor, in which the interest points can be extracted and then fed to an MLP regression model to predict the geometrical transformation, and the point number is fixed and turned into a hyperparameter. This technique can perform zero-shot learning without requiring aligned image pairs or ground-truth transformations for training, thus achieving real-time registration of brain T1–T2 weighted MRs. In addition, a manifold learning approach based on the Laplacian eigenmap [392] is introduced to embed multimodal images as a monomodal problem by means of structural representation.

3.2. Remote sensing

Another group of MMIM is investigated for remote sensing application. The significant development of high-resolution sensors has resulted in an increasing number of remote sensing satellites for obtaining image data, such as Ikonos, Quickbird, TerraSAR-X, Cosmo-Skymed, and WorldView [393]. Remote sensing images are often classified into two categories according to the style of the imaging system, *i.e.,* passive manner and active manner.

Passive optical images refer to remote sensing images that are captured in the visible- and near-IR (NIR) spectral bands by using a passive sensing system, and are obtained from the reflection of electromagnetic waves to the corresponding sensor and accurately represents the color and brightness information of the target. These images can

be further classified into panchromatic and multispectral images. In an active imaging system, light detection and ranging (LiDAR) tends to construct an image in an active manner, in which electromagnetic pulses in the IR, visible, or ultraviolet ranges are emitted from a transmitter. LiDAR systems have many advantages over radio detection and ranging (RADAR) systems, which is significant in measuring the range of distant objects and surfaces. Synthetic aperture radar (SAR) is another active imaging form with microwave radar. SAR usually reflects two characteristics of targets, namely, the structure information and the electromagnetic scattering information [394].

The combined use of these data is required in several remote sensing applications, such as land cover mapping [395], change detection [15, 17], and data fusion [396]. Image registration of multiple modalities can provide the ability to jointly use multiple information, leading to a larger data volume, shorter revisit time, and the utilization of complementary characteristics [397]. Take the widely used optical and SAR image matching as an example. With the distinct variance of these two imaging styles, a great appearance difference exists among optical and SAR images, which can explain different properties of the imaging area. In addition, the SAR system can work in both day and night, and see through fogs and clouds, which passive optical sensors are incapable of doing. In this case, combining the information of the historical optical images and the currently captured SAR images is important for analyzing the imaged area, making the registration of optical and SAR images a core and inevitable problem. However, the strong speckle noise in SAR images would make it difficult to extract effective features for the matching task [175].

In addition to the matching between optical and SAR or optical and LiDAR images, other matching targets include map-to-optical, UAV cross-season, optical day–night, and cross-temporal image pairs. In the following, we will provide a detail and comprehensive review of the matching methods on these multimodal data, with the taxonomy of area-, feature point-, and learning-based pipelines.

3.2.1. Area-based

As introduced in Section 2 and similar to medical image registration, an area-based framework is used to handle multimodal remote sensing image pairs. However, the core challenge is how to design and use a suitable similarity metric to drive the iteration procedure, thus accurately estimating the geometrical transformation. One direct solution is to utilize or modify popularly used metrics that are designed for multimodal images under information theory, such as MI and NMI. Another way is to indirectly use similarity metrics by reducing multimodal images into a uniform domain, which is commonly performed with domain transferring techniques, such as fast Fourier transform (FFT), structural information extraction, and mapping image intensity into a high-dimensional space by using descriptors. Advanced formulation of the registration problem and optimization methods to search for optimal solutions have also been investigated.

(1) Information theory-based methods

The successful use of information theory-based similarity metrics in MMIR have prompted many researchers to follow this strategy to handle registration problem of remote sensing image pairs with significant nonlinear radiation difference. The similarity metric that uses MI is useful for optical and SAR image registration [398–401] and is further improved by using it in combination with feature-based methods [402, 403]. An MI-peak metric was introduced in [393] to achieve automatic registration of high-resolution TerraSAR-X and Ikonos images acquired specifically over urban areas. The proposed metric involves estimating the joint histogram directly from image intensity values, which might have been generated from different sensor geometries and/or modalities. Xu et al. [404] proposed a symmetric form of MI, namely, Jeffrey's divergence (JD) as the similarity measure, and conducted mathematical analysis and experiments on the registration of SPOT image, Landsat TM image, ALOS PalSAR image, and digital elevation model (DEM) data with the affine model. By providing a larger feasible search space, the registration model based on JD is more capable in registering multimodal image pairs of a small overlap region.

(2) Frequency-based modality unification

To avoid the limitation in using the MI-like method, a more effective strategy is to construct image representation, thus reducing multimodal images into a uniform one. This approach is successfully conducted in the frequency domain by using FFT, in which the phase correlation or phase congruency (PC) algorithm developed by De [405] and Reddy [406] is widely used to register multimodal images that require extracting commonly existed structures for modality gap elimination. Inspired by the successful use of phase-based methods and to address significant modality and geometric difference, Xie et al. [407] extended early methods [405,406,408] based on multiscale log-Gabor [409] filtering and called their proposed method MLPC. This method is further improved in LGEPC [410] and used to register optical-IR, LiDAR depth-optical, and panchromatic to multispectral image pairs of urban areas or buildings.

(3) Descriptor-based modality unification

Another strategy to achieve modality unification is to use a feature descriptor to embed the image intensity into a high-dimensional feature space. To register different channels in MODIS data captured from island and coastline scenes, the authors in [411] proposed a novel MI scheme based on an adapted weighted strategy, which is conducted on the feature map constructed by efficient LoG and guided filter methods for salient feature extraction. In [412], the authors aimed to densely register optical and SAR images by means of optical flow-like algorithm. In this method, a dense descriptor constructed based on consistent gradient computation (*i.e.*, GLOH) is exploited for optical and SAR image representation and matching. Two cooperative frameworks conducted on global and local image content are used for global objective function optimization and local flow vector calculation, respectively. With the used coarse-to-fine strategy in this method, the method is applicable in large displacement scenarios. More recently, inspired by the feature-based method and the PC strategy, Xiang et al. [413] combined robust feature representations of optical and SAR images and 3D PC. To accurately estimate the 3D PC, two solutions are proposed in the spatial and Fourier domains. The constrained energy minimization method is used to seek the Dirac delta function in the spatial space after inverse FFT, and a fast sample consensus fitting scheme is applied to estimate linear phase coefficients in the frequency domain.

(4) Methods focusing on formulation and solution

In addition to focusing on better using similarity metrics following a traditional area-based framework, an advanced problem formulation or solution is critical to enhance the accuracy and efficiency of registration, thus prompting great attention among the remote sensing community due to the high resolution, noise, and large-scale nature. In [414], the authors mainly focused on the optimization method adapted to the MI cost function and provided a practical solution. The proposed inverse compositional optimization method has shown that using a specific optimization approach based on Hessian matrix can make the registration more robust and less computationally intensive. Hasan et al. [415] borrowed cross-cumulative residual entropy (CCRE) for remote sensing SAR and Google satellite image registration. In this method, a novel extension to the Parzen-window optimization approach based on partial volume interpolation is studied for solution space searching in the calculation of the gradients of the similarity measure, in which the geometric transformation is modeled as a second-degree polynomial or affine matrix.

Karantzalos et al. [416] proposed an automated registration framework for optical-radar satellite data, which is based on MRF formulation and linear programming solution, and similarity metrics such as NCC and NMI are exploited as a spectral preservation constraint. The experimental data cover urban, agricultural, coastal, and forest areas. Uss et al. [417] proposed a new area-based method based on registration with accuracy estimation (RAE), and defined the Cramer–Rao

lower bound (CRLB) of registration error for each local correspondence between coarsely registered pair of images. In this method, CRLB is estimated based on local image texture and noise properties, which can help tolerate the outliers and enhance the accuracy of transformation estimation. The whole pipeline is applied to register optical-radar-DEM images under the affine and second-order polynomial model.

3.2.2. Feature-based

Several limitations may exist in the area-based pipeline when it comes to performing the registration of multimodal remote sensing images. Remote sensing images are imaged with high resolution and typically contain heavy noise caused by imaging sensors and atmosphere, leading to a large computational burden in using similarity metrics to guide the optimization. Moreover, image pairs are often captured under significant geometrical variance, such as large rotation, scaling, deformations, and small overlapping area, making the solution space complex and difficult to optimize. Feature-based pipeline is more popularly accepted for handling remote sensing images, in which the main challenge is to extract repeatable feature points across multimodal images and then match them correctly.

(1) Methods based on existing feature operator

The most direct strategy achieves feature detection and description by directly modifying off-the-shelf methods such as SIFT and Harris. For example, the authors in [172] first coarsely registered multisource image pairs with the SIFT feature operator together with affine model estimation. Then, in the process of fine-scale registration, they extracted Harris corners followed by piecewise linear transformation. Finally, the triangular irregular network (TIN) and affine estimation are exploited for local deformation rectification. The proposed method is validated on registration for Quickbird panchromatic image, SPOT5 panchromatic image, SPOT4, and TM multispectral images, captured in city, lake, or river areas. In addition, Yi et al. [418] proposed SR-SIFT, a gradient orientation modification description and scale restriction strategy, to adapt SIFT for multispectral image registration.

To modify the classical SIFT method so that it can efficiently register cross-bands of multispectral or panchromatic images, Sedaghat et al. [173] improved SIFT algorithm in its feature selection strategy, called uniform robust SIFT (UR-SIFT), under the full distribution of feature location and scale. In this method, features are qualified and selected based on stability and distinctiveness constraints. Subsequently, an initial cross-matching process together with a consistency check strategy during projective transformation estimation is utilized for correct feature correspondence construction and image alignment. The tested data cover a variety of spatial resolution from 1 to 30 m, from both urban and rural areas. Another SIFT improvement was introduced in [419] by enlarging the pool range of the descriptor to adapt it to suit the characteristics of multisensor remote sensing images. By applying a similar process as that used to improve SIFT to make it workable in the optical and SAR image matching problem, Fan et al. [175] described the extracted features under multiple support regions and introduced a spatial consistent matching strategy to obtain reliable feature correspondences, followed by RANSAC to estimate a homography model for image registration.

(2) Methods based on advanced feature descriptor

A growing number of studies are focusing on designing valid descriptors to construct more reliable feature matches so that transformation parameters can be estimated correctly. Inspired by the self-similarity strategy [367], Sun et al. [174] introduced a multiscale self-similarity (MSS descriptor to initially construct a feature point set. Then, they conducted coherent point set analysis based on GMM model to correspond point sets under affine parameter estimation. This is considered as a probability density estimation problem together with EM solution. The superiority of the proposed method is demonstrated on multispectral and visible-spectrum images of a city.

Following a similar pipeline, Sedaghat et al. [207] used UR-SIFT [173] to uniformly detect local features and introduced an advanced self-similarity descriptor called distinctive order-based self-similarity (DOBSS descriptor to match the detected feature points. Then the reliable matches are identified by descriptor cross matching and consistency checking strategy constrained by projective transform. A rank-based local self-similarity (RLSS) descriptor is introduced in [420] to address the severe nonlinear radiometric differences between optical and SAR images for registration. The proposed RLSS is inspired from Spearman's rank correlation coefficient and further utilized for template matching of several subregions centered on the point locations extracted by block-based Harris detector from the image pairs.

Another strategy was proposed to describe detected features by means of PC in the frequency domain, which is commonly used to design advanced descriptor by using it in combination with PC and classical feature embedding methods. Ye et al. [421] integrated PC with an orientation histogram strategy to describe extracted features based on the structural properties of images; this process is called HOPC. The authors first detected control points by means of block-based Harris operator and top k selection, then applied a fast template matching scheme around these control points. A similarity metric named $HOPC_{ncc}$ is defined from the NCC of the HOPC descriptors to guide this matching procedure. Unreliable point matches are eliminated through a global constraint in the projective transformation model, and the final nonrigid registration is implemented with a piecewise linear model estimation based on the fine matches of control points, which is validated on several visible, IR, map, and LiDAR data covering urban and suburban areas.

Fan et al. [135] proposed a feature point detection, description, and matching pipeline by combining improved Harris and phase congruency structural descriptor (PCSD) to register SAR-optical, optical-LiDAR, and IR-optical images. Uniform nonlinear diffusion-based Harris feature extraction was designed to reduce the influence of speckle noise. On the basis of the observation that the structure features are less sensitive to modality variation, they proposed PCSD by using PC structural images in a grouping manner. With the advantage of the PC method in alleviating the modality difference, multiscale PC (MS-PC) descriptors [422], which are more robust to the radiation differences between images, are used as a similarity metric to achieve correspondences construction. Cui et al. [423] constructed a scale space based on nonlinear diffusion function to make the proposed method workable on different resolutions. Then a pixel-wise local PC method was used to extract distinctive feature points. These point features are matched through the proposed rotation invariance descriptor.

(3) Methods based on conjugated structural map

Apart from the methods that focus on designing better feature detectors and descriptors, several methods attempt to extract the conjugated structures first, such as gradient, edge, and counter. On the basis of a structural map, a feature-based matching framework can reliably perform feature matching and image registration.

Following this idea and to register airborne optical and C-band SAR images, Huang et al. [142] first extracted edge features to suppress noise. Subsequently, the preprocessed features are considered as point sets and are matched using the improved SC, which are later used to estimate an affine transformation. In [177], the classical SIFT is modified to avoid failed registration caused by poor feature extraction by using a segmentation method based on an iterative level set to extract conjugated features; which is called ILS-SIFT. This method is validated to register optical and SAR images under polynomial transformation estimation with the proposed improved RANSAC. Another strategy in [424] involves performing optical and SAR image registration based on iterative line extraction inspired by [425], followed by line intersection matching under a coarse-to-fine registration scheme. In this method, the Voronoi approach [426] together with spectral point matching strategy [427] is utilized to enhance the matching accuracy.

Another novel registration method for optical and SAR images is proposed in [428], which is based on straight line feature extraction and MI. This method first uses different edge detectors to perform the

line segments in both optical and SAR images. Then, through Hough transform and straight-line fitting, the main straight lines of each image are extracted, and their intersections are obtained and taken as the candidate matching points. Finally, RANSAC is employed to coarsely register the image pairs to generate patch pairs, which are subsequently optimized for fine registration under MI measurement.

Given the superiority of PC in conjugated structure extraction over gradient [429], Zhang et al. [147] proposed a novel method to register multitemporal UAV images. The authors first utilized PC to describe the images with respect to their structural characteristics, on the basis of which the FAST corners are detected and then matched via the similarity of feature descriptors constructed by a maximum index map, followed by the RANSAC method for affine model estimation and mismatch elimination.

Considering that structural similarity between images could be well preserved and utilized for image registration across different modalities, Ye et al. [143,430] first generated a densely described image based on existing local descriptors such as HOG and LSS. Then, they defined a similarity metric in the frequency domain by using 3D-FFT together with oriented gradients to determine and match control points from extracted Harris corners through a template matching scheme. An iterative mismatch removal procedure is performed in cubic polynomial model estimation and consistency judging, and the final performance of the entire large-size image pair is verified under a piecewise linear transformation model based on TINs and local affine estimation.

Inspired from MSER and by merging phase congruency theory, Liu et al. [431] proposed a novel affine and contrast invariant descriptor called maximally stable phase congruency (MSPC), which integrates the affine invariant region extraction with structure features to achieve image registration. In particular, The interest points are first detected via moment ranking analysis of the PC images. The structure features are constructed from the multiorientation PC via the proposed SFE method. Subsequently, the extracted points are matched based on the similarity of the introduced MSPC descriptor, followed by RANSAC algorithm for refined matching, affine model estimation, and image registration. On the basis of a similar strategy, the authors in [432] extracted the common structure features from scale-adapted IR and visible images based on PC and significance ranking space. These features were subsequently matched with the proposed kernelized correlation filter.

A more general method based on PC and maximum index map (MIM) was proposed by Li et al. [148], namely, radiation-invariant feature transform (RIFT). In RIFT, the authors first detected better repeated corner and edge feature points based on the generated PC map. Then, they performed MIM based on log-Gabor convolution sequence for feature description, thus achieving rotation invariance by constructing multiple MIMs. The proposed method has shown promising performance in matching general multimodal images, such as optical-to-SAR, IR-to-optical, depth- or map-to-optical, and day-to-night image pairs.

(4) Others

Some studies mainly focused on the formulation and framework design of image registration. In [176], the authors proposed a two-stage (coarse-to-fine) framework for multispectral remote sensing image registration called pre- and fine registration. In the coarse stage, the SR-SIFT [418] is utilized to initially estimate and rectify geometrical differences such as scale and rotation. In the second stage, the Harris corners are detected from the prealigned image pairs, and point correspondences are constructed based on the local self-similarity descriptor [367] followed by a global consistency check strategy to remove false matches. Finally, the authors achieved registration through a piecewise linear transform model and tested it on multispectral image pairs.

The structure variances between optical and LiDAR images pose difficulty in detecting repeated points across these two modalities. Wong et al. [394] proposed an effective technique to align optical and LIDAR images captured from highlands ranch or city area. In this method, the control point (CP) candidates in LiDAR image are first detected by selecting top responses of Harris extractor. Then, a set of region correspondences around CPs between LIDAR and optical image pairs are determined by means of local feature mapping transform optimization and FFT acceleration, which are conducted based on similarity matching with the SSD cost metric of local regions. They also used the RANSAC algorithm to seek reliable feature correspondences, which are further used to estimate the nonrigid transformation model and then align the image pair via the normalized direct linear transformation (DLT) algorithm. Moreover, Marcos et al. [397] aimed to perform domain adaption via extracting a domain-invariant feature representation for each superpixel in multisensor remote sensing images. They proposed a midlevel representation based on spatial distribution of spectral neighbors and formulated this in an MRF model that is optimized by the iterated conditional mode algorithm.

More recently and considering that traditional matching methods are not able to construct a high number and ratio of correct point matches for multisource images due to the large radiometric and geometric distortion among them, Deng et al. [433] were inspired by graph theory and proposed a two-stage mismatch elimination method. They first used cluster strategy to represent the local geometric similarity without considering any global geometric model in advance. The cluster is constructed from matched triangles in a complete graph formulation. Then, they used TIN to approximate a complete graph, which can greatly simplify the computational complexity.

3.2.3. Learning-based

An increasing number of studies are focusing on the MMIR of remote sensing images with deep techniques. The common strategies in this type of method are to (1) generate deep and high-level image representation to obtain more repeated feature points and/or advanced descriptors to obtain a high number and ratio of accurate feature matches; (2) learn to transfer one modality to another, thus enabling traditional methods to successfully perform MMIR; and (3) directly predict the underlying transformation model in an end-to-end paradigm.

A more direct method is to integrate CNN into traditional image matching pipeline, such as generating trainable feature detector, descriptor, or similarity measurement, to enhance the registration performance. Yang et al. [434] proposed a CNN feature-based multitemporal remote sensing image registration method by learning for multiscale feature descriptors and gradually increasing the selection of inliers to improve the registration performance. The multiscale feature descriptor is generated from a pretrained VGG network, and the TPS model is integrated to explain the nonrigid transformation and estimated under a GMM and EM framework. To compensate for the weakness of the classical SIFT in terms of its use of only local low-level image information, [196] aimed to exploit middle- or high-level information by adopting the advantages of emerging CNN and fusing SIFT and CNN features for multispectral and multisensor image registration under a simple similarity transformation. Similarly, Ma et al. [197] cast the registration task as a coarse-to-fine problem in a two-stage framework. They initially approximated the spatial relationship with a deep architecture based on the image feature layer, such as using VGG-16. This deep model can serve as a feature extractor to achieve pyramid feature detection. Later, the authors improved the accuracy of registration under the combined use of deep and handcrafted local features together with the RANSAC method for refined feature matching and transformation estimation. In addition, a densely connected CNN was introduced in [435] for visible and IR remote sensing image registration. A channel-stacked network with densely connected convolutional building blocks was designed to capture low-level features and drive the template matching. Moreover, an augmented cross-entropy loss was proposed to guide the training procedure with better learning ability and stability.

Another learning strategy introduced in [436] involves addressing the requirement of immense data for training. The authors proposed a generative matching network (GMN) to generate the coupled simulated optical image for the real SAR image or generated a pseudo SAR for a single optical image. These generated patch pairs are then fed into a deep matching network to exploit the latent and coherent features between multimodal patch pairs to infer their matching labels. This method is perfected and improved in [230] by proposing an end-to-end architecture to directly learn from generated patch pairs and their matching labels for later registration. A similar idea is adopted in [437], in which the authors combined conditional GANs (cGANs) and several handcrafted methods such as NCC metric, SIFT, or BRISK to improve the registration performance for optical and SAR images. The core idea is to train to generate SAR-like image patches from optical images by using cGANs, thus achieving modality unification to make it workable for many methods designed for monomodal image matching. The final refined feature matches are obtained through RANSAC with an underlying affine model constraint.

Considering the difficulty in feature design and slow optimization by gradient descent of classical methods, Zampieri et al. [438] designed easy-to-train, fully convolutional neural networks to learn scale-specific features, thus achieving nonrigid MMIR in linear time. The proposed method could directly predict the optimal parameters of deformation modeled as diffeomorphisms, thus avoiding the iterative process. This idea is validated on registered visible images and binary maps of buildings or houses.

We can see that CNN-based methods designed for MMIR of remote sensing application are not as rich and active as those in the medical field. The main reason is the difficulty in capturing sufficient multimodal remote sensing images for training and testing, as well as the complexity of remote sensing images in terms of their high resolution, hybrid noise, and large geometrical variance. These challenges require researchers to pay more attention to collecting available datasets and introducing more effective deep registration frameworks, loss functions, and training strategies to obtain advanced matching performance for multimodal remote sensing images.

3.3. Vision

In the field of computer vision, the most active research on image matching involves unimodality image pairs. The core challenges are how to handle large geometrical deformations and the low image quality caused by viewpoint changes and negative imaging conditions. Typical barriers include image rotation, scaling, affine, local distortion, background noise, occlusion, abnormal illumination, and low texture [2].

In MMIM, geometrical deformation can be addressed easily due to the great efforts and achievements of researchers in general image matching. In this regard, more attention to serious appearance differences is required. In the computer vision research area, a representative type of MMIM may be IR and visible (IR–VIS) image pairs, aside from cross-spectral, cross-temporal (such as cross-weather/season, day–night), and other modalities.

3.3.1. Visible-to-infrared

The most popular topic in the visual area is VIS–IR image matching, which is widely used in various visual applications (*e.g.,* image fusion) [13] due to the complementary information provided by the two image types. Visible images capture reflected light, while IR images can capture thermal radiation, thus providing information or properties of the same target/scene from different aspects. However, due to the differences in imaging sensors, visible images typically have high spatial resolution and considerable details and chiaroscuro, but they are highly affected by light and weather conditions. In contrast, IR images are independent of these disturbances due to the nature of thermal radiation-based imaging, but they typically have low resolution and poor texture. Therefore, IR and visible image matching remains an open problem that requires further attention.

(1) Area-based methods

Infrared images have few details. Thus, many techniques extract common structures first from both images and then use the aforementioned area- or feature-based matching methods or directly design deep learning frameworks to handle the IR–VIS image registration problem.

To extract consistent features and better apply NMI for area-based registration, Yu et al. [439] detected the edge features through a grayscale weighted window strategy from the image pair to reduce the joint entropy and local extrema of NMI, thus enhancing the registration performance. Another method in [440] rapidly and accurately identified canthi regions for fever screening. The authors proposed an area-based registration approach for VIS–IR registration by converting the original image pairs into edge maps first. Then, they used affine and FFD to explain the geometrical relations for coarse and fine registration. This approach is optimized by maximizing the overall similarity of the MI metric between the edge maps.

(2) Feature-based methods

To perform the registration of IR and visible image pairs that are taken from slight viewpoint changes of the same buildings, Hrkać et al. [441] conducted an experiment in which the scenario was that the corners are more stable in both images. They detected Harris corners from these two images, then used a simple similarity transformation to register IR and visible building images. They chose partial Hausdorff distance as the similarity measurement. Ma et al. [133] proposed a nonrigid pipeline for registering visible and thermal IR face images. In this paper, an edge map of each image is extracted then converted into a point set as the inherent feature to represent an image. Subsequently, the Gaussian field criterion is analyzed and validated for the point set registration problem, and a regularized form of this criterion is generated under a reproducing kernel Hilbert space for both rigid and nonrigid registration. In [442], the corners were extracted based on extremal moments of phase congruency images and described with log-Gabor filters. Then, the matching step was performed with the similarity of descriptors, then the RANSAC method was used to determine reliable point correspondences. In addition, a fast visual salient feature detector together with a descriptor-rearranging strategy is proposed in [443] to register the visible and IR image pairs.

In building diagnostics, IR and visible images need to be combined to obtain more comprehensive information. The authors [444] registered the image pairs of these two modalities by segmenting the edge lines to extract quadrilateral features first and introduced a forward selection algorithm to identify reliable feature correspondences for transformation model estimation. Considering the significant variance in terms of resolution and appearance caused by different imaging sensors, Du et al. [141] proposed a scale-invariant PIIFD for corner feature description and matching. In addition, a locality preservation constraint was coupled to remove false matches to better estimate the affine matrix under a Bayesian framework. In [178], the authors first extracted edges by using the morphological gradient method, and then they applied C_SIFT detector on edge maps for distinct point searching and BRIEF for description, thus achieving scale- and orientation-invariant matching.

Min et al. [445] proposed enhanced affine transformation (EAT) for nonrigid IR and visible image registration. They first extracted interest points from the edge maps of input images, thus casting image registration into a point set registration problem, in which they used SC to describe the local structure of the point set to simplify the registration procedure. In this method, the transform model, objective function, and optimization method are needed to achieve promising performance. The optimal EAT model is estimated from the local feature to explain the global deformations, and a Gaussian field-based objective function is established and simplified by using the potential true correspondences between image pairs to guide the matching procedure. A coarse-to-fine strategy based on quasi-Newton method

is also designed and applied to determine the optimal transformation coefficients from point sets of IR and visible images. To cope with nonrigid registration, a feature point-based method was proposed by Min et al. [446], in which a Gaussian weighted SC is introduced to quickly extract matching point pairs from edge maps of image pairs.

(3) Video sequences-based methods

Another major category of IR and visible image matching methods is on the basis of video sequence. The idea is to use target tracking information, thus integrating the temporal domain with the image domain to improve the overall registration performance. Bilodeau et al. [447] proposed a registration method based on trajectory points and a novel error function to align two multisensor images before image fusion. The authors detected feature points based on the trajectories of the moving objects, which are obtained using the background and a simple tracking strategy. Then, they matched the trajectory points by using a RANSAC-based method and a novel registration criterion. Han et al. [448] registered IR and visible image pairs by aligning hybrid visual features, including straight lines and interest points, which are used to estimate the global perspective transformation and local transform adaptation, respectively. In [449], a strategy that uses the local shape of noisy polygon vertices for frame pair matching and then estimates the homography transformation is introduced for VIS–IR video registration. To register the sequences of IR and visible videos, an integrated global-to-local framework [450] is proposed to address this dynamic scene matching problem. An overall MI measurement of two sequences is optimized for global homography estimation, and frame-to-frame registration is performed to refine the local transformation for each frame pairs. In addition, a smooth strategy is used to enforce the temporal consistency in the temporal domain, thus smoothing the local homographies.

Similarly combining the motion and feature information, a coarse-to-fine framework [451] is proposed for VIS–IR video registration. The motion information, which is captured by curvature scale space keypoint extraction, is applied to estimate the global scale and rotation for coarse matching. The interest points are relocated and matched with normalized descriptors, together with a mismatch removal strategy based on coherence checking for fine registration. These descriptors are created from the histogram of edge orientation. To register planar VIS–IR image sequences through spatiotemporal association, Zhao et al. [149] bypassed the use of feature extraction while first coarsely registering the frame pairs by using the motion vector distribution as descriptors to present the temporal motion information of foreground contours. The fine matching stage was performed by FAST corner matching and similar transformation estimation. The method proposed in [452] registers nonplanar VIS–IR frame pairs by segmenting the salient targets and matching the blob features.

(4) Learning-based methods

In complex scenes, detecting matchable features or directly training deep networks from IR and visible images is not easy. Wang et al. [453] proposed a two-stage adversarial network, which includes a domain transfer network and a geometrical transformer module, to map images across different modalities and obtain refined warped images. Baruch et al. [454] aimed to jointly detect and match interest feature points in one step. In this method, the authors introduced a hybrid CNN architecture that consists of a Siamese CNN and a dual non-weight-sharing CNN, which can well capture and leverage the joint and disjoint cues from multimodal image patches.

Considering the shortcomings of handcrafted similarity measurements in a traditional nonlinear optimization pipeline, an unsupervised procedure is proposed in [455] to simultaneously train an image-to-image translation network and a registration network on two given modalities. This learned translation allows the transfer of one image domain to another image then training the registration network by using simple and reliable mono-modality metrics. A GAN framework that encourages the generator to preserve the geometry information is applied for the translation task and then modified to generate smooth and accurate spatial transformation, thus completing the registration task.

(5) Others

Considering that image fusion quality is bound not only by the quality of the algorithm, but also by the outcome of the required image registration algorithm, [456] presented a combined method called MIRF, which can register and fuse multimodal images for area-based image registration and object-based image fusion. It is implemented by dual-tree complex wavelet transform. Another integrated framework that simultaneously addresses registration, fusion. and people tracking of thermal and visible videos is introduced in [457]. The registration involves maximizing the overlapped area of image pairs by using trajectory feature matching with RANSAC to estimate the affine matrix. A nonrigid registration approach for IR and visible images with mixed features is proposed in [23] and integrated in a 3D reconstruction scheme to acquire a more informative and complete 3D model. In [458], the authors introduced a long-wave IR and visible-light spectrum image matching method to detect the object from the images of these two modalities. This method is based on edge detection and binary temple matching strategy, followed by the use of a local fuzzy threshold to identify high-similarity objects.

3.3.2. Cross-spectral

Cross-spectral image matching is defined as the target images are taken from different spectral bands, such as in multispectral and visible-to-NIR image matching.

In multispectral matching and considering that the traditionally used measuring metrics, such as SSD or SAD, are computationally efficient but perform poorly in multispectral image alignment, Cao et al. [459] proposed a structure consistency boosting (SCB) transform to enhance the structural similarity in the multispectral/modality image registration problem. This approach can help avoid spectral information distortion caused by misalignment across spectral bands due to imaging device movement or alternation. In [460], the authors proposed RegiNet, which uses a gradient map of the reference image to guide the target image for registration. This method was optimized by the structure loss that forces the networks to efficiently capture gradient information from the reference image. PSNR and structural similarity (SSIM) were used as evaluation metrics in the experimental part.

Another spectral image called NIR can be used to recognize the chemical compositions of the visual target in a nondestructive and efficient way. Many researchers study the matching problem between NIR and visible images, and propose meaningful techniques. To deal with the occlusion and heterogeneous problem in NIR and visible face image matching, Yi et al. [461] introduced an edge-enhancing filtering to capture the common structures for image pairs of different modalities, and template matching is conducted on the segmented image patches. [462] aims to develop a new solution to meet the accuracy requirement of face-based biometric recognition. On the basis of an analysis of properties of NIR and visible face images, a mechanism of correlation between NIR and visible faces is learned from NIR–VISible face pairs, then the learned correlation is used to evaluate the similarity to guide face registration under different illumination conditions.

3.3.3. Cross-temporal

Another type of multimodality involves appearance variance due to temporal changes, such as day to night, cross-weather, or cross-season. In such a scenario, the appearance variance is more significant than that caused by different sensors due to the disappearance of common details and structures in the image pairs. An image pair captured from day and night may have fewer details in night images but more details in the daytime images. Fortunately, contours of buildings or trees obviously coexist in both day and night images, or the textural details in night images are revealed by artificial light. For cross-weather or season image matching, the appearance of the same scene would be largely changed, such as in rainy and sunny days, and summer and snowy

winter days. Such methods is significance for the application of change detection, image localization, and place recognition.

To investigate good cross-temporal matching for vision application, Shrivastava et al. [8] studied the visual similarity between images across different domains, such as photos taken in different seasons, paintings, and sketches. The authors hypothesized that the important parts of the image are those that are more unique or rare within the visual world, following the work in [463]. Inspired by the success of classic intensity-constancy-based image alignment methods and the modern GAN technology, Zhou et al. [6] proposed a latent generative model for cross-weather image alignment. In this method, the image registration task is formulated as a constrained deformation flow estimation problem with a latent encoding procedure based on the use of intensity constancy and image manifold property.

As for matching with different lighting conditions such as day and night, Luo et al. [5] presented a trainable point descriptor by combining the local patch similarity constraint with spatial geometrical constraint of the detected points, achieving promising matching performance for multimodal image pairs. To detect repeatable point features with drastic appearance variances caused by different weather and lighting conditions, Verdie et al. [182] trained multiple piecewise linear regression models from aligned image pairs by using handcrafted DoG for training set collection and then training general regressors to generate a feature score map. The keypoints are identified as the local maxima of this map with non-maximum suppression.

3.3.4. Others

In addition to the above-mentioned modality pairs used for matching or registration, the nonlinear intensity variance caused by data type or domain is also studied in the computer vision community. This approach is typically represented by image–painting–sketch matching, image-point cloud matching, semantic matching, and cross-domain matching from image to text.

Matching for real image and paint or sketch is a challenging task not only because of the great appearance difference but also because of nonrigid deformations. An early relevant work involved registering photos taken in different seasons, paintings, and sketches. Shrivastava et al. [8] proposed an interesting work by defining visual similarity between images across different domains. A method proposed by Aubry et al. [9], developed a new compact representation of complex 3D scenes. The 3D model of the scene is represented by a small set of discriminative visual elements that are automatically learned from rendered views.

Semantic matching denotes the input image pairs that have different targets but similar properties, such as a dog and a cat. This problem has attracted increasing attention, with researchers using deep techniques to understand the semantic similarity. Choy et al. [231] proposed UCN by using deep metric learning to directly learn a feature space that preserves either geometric or semantic similarity for semantic matching. Another method called NCN was proposed by Rocco et al. [464]. In this method, the authors trained an end-to-end deep architecture with the use of semilocal constraints to create reliable feature correspondences. Other similar methods for semantic matching include [232–235].

Another interesting research topic is the matching for different domains between real images and texts, which has been a hot topic in recent years. We refer interested readers to [10–12] for detailed reading.

3.4. Summary

MMIM has broad taxonomies regarding different imaging devices or conditions in the medical, remote sensing, and computer vision research areas. Methods in different modalities are often partial to a specific type of registration pipeline. For instance, most medical images can be better registered under an area-based framework because these images usually have large overlaps, slight image deformation, and low resolution. However, due to the distinct vessel structures present in both retinal images, the feature-based strategy can achieve more accurate registration results. As for remote sensing and computer vision research, the images for matching typically have high-resolution and large transformations; thus, an area-based framework would cause low accuracy and high computational burden. Therefore, a universal method that can simultaneously handle the image matching problem of all types of modality to satisfy different application requirements needs to be designed. To this end, Zimmer et al. [466] tried to develop a general metric based on the commutativity of image graph Laplacians, to guide the optimization for MMIR of synthetic data, visible to infrared data, and medical data. The typical methods reviewed above are summarized in Tables 1 2 3, which correspond to the fields of medical, remote sensing, and computer vision. The modality pair, method taxonomy, transformation model, scene or target of test images, and core idea of each method are listed in the tables. In each table, we use '–' to indicate the information that cannot be queried in its literature. And we use many abbreviations for better view, which can be easily inferred from body text.

4. Experimental evaluations

Over the past decades, MMIM has attracted increasing attention in the fields of medical, remote sensing, and computer vision, serving as a prerequisite procedure for many high-level applications such as image fusion, image localization, and object recognition and tracking. As a result, many diverse matching methods have been developed to address nonlinear intensity variances and geometrical deformations of two or more multimodal images. Hence, these typical, recently developed methods need to undergo qualitative and quantitative comparison to determine their strengths and weaknesses.

However, due to the distinctiveness among multimodal image pairs in different research areas, to the best of our knowledge, no such literature comprehensively reviews MMIM methods in details and across the medical, remote sensing, and computer vision fields. Existing related surveys typically focus on general methods in a single research field, either for general medical image registration, visual image matching, or application-based review. Many survey papers provided comprehensive analysis on method introduction and taxonomy but ignore practical experimental evaluations. Researchers also have difficulty in directly evaluating their techniques on public MMIM with unified standards.

In this survey, we try our best to collect multimodal image pairs from public websites to ensure that all typical modalities used by researchers are covered. In the following, we will introduce experimental details about the constructed MMIM datasets, evaluation metrics, and evaluation performance of typical methods in feature detection, feature description and matching, mismatch removal, and image registration.

4.1. Constructed evaluation database

To satisfy the requirement of our experimental evaluation and construct a uniform standard for future researches, we collect a complete database that covers all typical multimodal image pairs across the medical, remote sensing, and computer vision fields. The collected database contains 18 modality pairs: (1) The medical research database includes cross-matching of MR T1-, T2- and Pd-weighted images, MRI–PET, SPECT–CT, and Retina images with different imaging methods; (2) The remote sensing research database contains UAV cross-season image pairs, optical day–night, LiDAR depth-optical, IR-optical, map-optical, optical cross-temporal, and SAR-optical image pairs; (3) The computer vision-related research database consists of VIS–IR image pairs, visible–NIR, visible cross-season, day–night, and image–paint image pairs. We have created in total 164 image pairs for our experimental evaluation. Our collected database and its source are introduced in detail below.

Table 1
Multimodal image matching methods in medical research area.

Refer.	Modality	Method type	Transform.	Target/Scene	Core idea
[350]	Fundus-FA	Area-based	Similarity	Retina	EM-PCA-MI
[360]	Fundus-CSLO	Area-based	Deformation	Retina	Feature neighborhood MI;
[379]	Fundus-SLO	Feature-based	Deformation	Retina	Mean phase image generation + RANSAC + MIND
[140]	Fundus-FA	Feature-based	Affine	Retina	Harris + PIIFD
[170]	Fundus-FA	Feature-based	–	Retina	UR-SIFT + PIIFD
[171]	Fundus-FA	Feature-based	Affine	Retina	SIFT; PIIFD; RSW-LTS
[358]	T1–T2-PD; CT–MRI	Area-based	Rigid	Brain	Sampling strategy: 3D Fast discrete curvelet transform + MI
[376]	T1–T2-PD; MRI–US	Area-based	Deformation	Brain	dLDP + MRF
[369]	CT–MRI; MRI–US	Area-based	Deformation	Brain	Patch-based SSC + Discrete optimization
[359]	CT/PET–MR	Area-based	Deformation	–	L-BFGS-B + CSO + FFD + Block grouping strategy
[357]	T1–T2	Area-based	Polynormal	Brain	Bayesian formulation + MRF
[465]	T1–T2	Area-based	FFD	Brain	Geodesic active fields + Polyakov energy
[362]	CT-US	Area-based	Rigid/Affine	Liver/Kidney	Transfer CT to US; Locally statistical metric
[368]	CT–MRI	Area-based	Rgid/Deformation	Lung	Self similarity Descriptor: MIND + SSD + Diffusion + Gauss–Newton
[375]	T1–T2-PD	Area-based	Deformation	Brain	Self similarity Descriptor: ZMLD + FFD + L-BFGS
[370]	T2-T1-PD; MRI–CT	Area-based	Rigid/Deformation	Brain	Entropy and Laplacian Image + SSD
[380]	CT–MR; PET–MR	Learning	Rigid	Brain	Similarity function learning
[87]	CT–MR	Learning	Deformation	Head	Metric Learning: SAE
[88]	TRUS–MR	Learning	Rigid	Transrectal	Metric learning
[89]	T1–T2	Learning	Deformation	Brain	Metric learning+Gradient descent optimization
[381]	US–MR	Learning	Rigid	Abdominal	Metric learning with 5-layer CNN + Powells optimization
[382]	MR–US	Learning	Rigid	Fetal Brain	Metric Learning with LSTM + STN
[91]	CT–CBCT	Learning	Rigid	Spine/Cardiac	RL: Q-Learning with single RL agent + Greedy supervised + Attention
[92]	Depth-CT	Learning	Rigid	Spine	RL: Q-Learning for Modality reduction + Contexture information
[94]	X-ray-DRRs	Learning	Rigid	Spine	RL: Multi agents + FCN in Markov decision process + Auto-attention
[97]	T1–T2	Learning	Rigid	Brain	Deep Transformation estimation: CNN-11+ResNet-18+Bivariant geodesic distance loss
[383]	T1–T2	Learning	Rigid	Brain	Deep Transformation estimation: CNN-6 + FCN-10
[101]	T1–T2	Learning	Deformation	Brain	Deep Transformation estimation: FCN + LDDMM
[384]	MR-TRUS	Learning	Deformation	Prostate	Deep Transformation estimation: FCN-30 + DDF+Label similarity-based Loss
[106]	MR-TRUS	Learning	Deformation	Prostate	Deep Transformation estimation: GAN+ DDF +Label Similarity-based Loss
[103]	cine-MR	Learning	Deformation	Cardiac	Deep transformation estimation: Unet + Label/Similarity metrics-based Loss
[113]	MR–US	Learning	Deformation	Brain	Deep Transformation estimation: 3D CNN+Intensity and Gradient-based similarity loss
[111]	CT–MR	Learning	Deformation	Pelvic	Deep Transformation estimation: Unet + NCC-based Loss
[386]	T1-FLAIR	Learning	Affine/Non-rigid	Brain	Deep transformation estimation: Encoder–Decoder + STN + Cycle-consistent and Inverse-consistent loss
[108]	MR-TRUS	Learning	Rigid	Prostate Biopsy	GAN: Synthetic Transforms + Adversarial loss
[112]	Fundus-FA	Learning	Deformation	Retina	GAN: Simulated deformations + Content loss using NMI, SSIM and VGG + Adversarial and Cycle-constraint Loss
[109]	T1–T2;T2-FLAIR	Learning	Deformation	Brain/Lung	patch GAN: Self-reconstruction Loss + Cross-cycle consistency + Adversarial loss + similarity metrics
[117]	T1–T2	Learning	Affine	Brain	CNN for points extraction+ MLP for transformation estimation

– BrainWeb [467].[1] BrainWeb is also called simulated brain database (SBD), which contains a set of realistic MRI data volumes produced by an MRI simulator. The full 3D volumes were simulated using three sequences (T1-, T2-, and PD-weighted) and a variety of slice thickness, noise levels, and levels of nonuniform

[1] BrainWeb: https://brainweb.bic.mni.mcgill.ca/brainweb/.

Table 2
Multimodal image matching methods in the remote sensing research area.

Ref.	Modality	Method type	Transformation	Target/Scene	Core idea
[414]	Map–Opti.	Area-based	Homography	Urban	MI cost function + Inverse compositional optimization
[416]	Optical–Radar	Area-based	Deformation	Urban/Farmland/Coastal	MRF + Linear programming + Uniform deformation grid + NCC & NMI
[415]	Optical-SAR	Area-based	Affine/Polynormal	Sydney	CCRE Metric + Parzen-window Optimization
[393]	Opti.–SAR	Area-based	Deformation	Urban	MI-peak metric
[404]	SPOT–TM–SAR-DEM	Area-based	Affine	Terrestrial region	Symmetric Jeffrey's divergence as similarity metric
[410]	Opti.–IR/LiDAR; MS	Area-based	Similarity	Urban/Building	Multi-scale Log-Gabor + PC
[417]	Opti.–Radar–DEM	Area-based	Affine/Polynormal	–	Cramer–Rao Lower Bound (CRLB) + Registration with Accuracy Estimation (RAE)
[411]	Cross-channel in MODIS	Area-based	Deformation	Island/Coastline	Salient feature map construction via LoG & GF + Adapted weighted MI
[412]	Opti.–SAR	Area-based	Deformation	Urban/Suburban/Fishery	Image representation: GLOH dense description + Optical Flow-like framework
[413]	Opti.–SAR	Area-based	Translation	Urban/Suburban/Fishery	Feature representation + 3D PC + Constrained energy minimization + Fitting
[142]	Opti.–SAR	Feature-based	Deformation	Farmland	Canny + Harris + SC + Affine estimation + TPS
[433]	GF1-NIR;Landsat-Pan	Feature-based	Affine	Mountain/Urban/Hill	Graph-based local geometrical similarity: Cluttering + TIN
[172]	Ms/Pan	Feature-based	Nonrigid	City/Lake/River	Coarse Step: SIFT + Affine Esti.; Fine step: Harris + Piecewise linear transfor. + TIN & Affine Esti.
[394]	Opti.–LiDAR	Feature-based	Polynomial/Projec.	Highlands Ranch/City	Region matching: Harris-based region + SSD + FFT + RANSAC + DLT
[173]	MS/Pan	Feature-based	Projec.	Urban/Rural	Improved SIFT via Uniform Robust Feature Selection (UR-SIFT)
[175]	Opti.–SAR	Feature-based	Homography	City	Improved SIFT via Multiple support region description + Spatial consistent matching
[174]	MS–VIS	Feature-based	Affine	City	Points selection via Multi-scale Self-similarity + GMM + EM
[176]	Cross-band of MS	Feature-based	Piecewise Linear	Plain/Mountain/Urban	SR-SIFT+Harris+LSS
[207]	MS/Pan	Feature-based	Projec.	Plain/Mountain/Urban	UR-SIFT + Distinctive Order-based Self Similarity Descriptor (DOBSS)
[177]	Opti.–SAR	Feature-based	Polynomial	City/District/Airport	Iterative Level Set-SIFT + Improved RANSAC
[424]	Opti.–SAR	Feature-based	Affine	Airport/Suburb/Urban	Iterative line extraction + Line intersection matching
[421]	Opti.–IR–Map–LiDAR	Feature-based	Piecewise linear	Urban/Suburban	Block-based Harris + Histogram of Orientated PC (HOPC) + NCC + Projective Transform Constraints
[147]	Multitemporal UAV	Feature-based	Affine	City	PC + FAST + Maximum Index Map + RANSAC
[135]	Opti.–SAR/IR/LiDAR	Feature-based	–	–	UND-Harris + PCSD Description + Mismatch removal
[143]	Opti.-SAR-Map-LiDAR	Feature-based	Piecewise linear	Urban/Suburban/Plain	Harris + HOG/LSS + Template Matching via 3D-FFT + Mismatch removal + TINs &Local Affine Esti.
[420]	Opti.–SAR	Feature-based	Projec.	Urban/Suburban	Block-based Harris + Rank-based LSS (RLSS) + Template matching via SSD Metric
[423]	Opti.–SAR–LiDAR–IR	Feature-based	Rigid	Urban/Vegetation/Farmland	Local PC + Nonlinear diffusion scale space + Rotation invariance descriptor
[431]	VIS–IR	Feature-based	Affine	Urban	PC + Moment Ranking + Struct. Extrac. + Maximally Stable PC (MSPC) Descrip. +RANSAC
[432]	VIS–IR	Feature-based	Affine	Urban	PC + Moment ranking + Struct. Extrac. + Kernelized correlation filter matching + RANSAC
[434]	Cross-Temporal	Learning	Nonrigid (TPS)	Mountain/Plain	Learnable descriptor: VGG-16 + GMM-EM
[438]	VIS–Binary Map	Learning	Deformation	Building	Deformation Estim. : U-Net + FCN + Content consistency & Estim. Error & Smooth-based loss
[196]	Cross-band of MS	Learning	Similarity	–	Hybrid feature: VGG16+SIFT
[436]	Opti.–SAR	Learning	Rigid	–	Modality transfer: GAN + FC Layers + Adversarial & Mapping & Reconstruction & Cross-Entropy loss
[197]	Opti.–SAR	Learning	Affine	Suburban	VGG-16+SIFT + RANSAC
[435]	VIS–IR	Learning	Similarity	Suburban	CNN-7+FC-2 +Augmented entropy loss + Template matching
[437]	Opti.–SAR	Learning	Affine	Urban	Modality transfer learning: cGAN + SIFT/BRISK/NCC + RANSAC

Table 3
Multimodal image matching methods in the computer vision research area.

Refer.	Modality	Method type	Transform.	Scene/Target	Core idea
[439]	VIS–IR	Area-based	Nonrigid	–	Edge + NMI
[440]	VIS–IR	Area-based	Affine + FFD	Person	Edge + MI
[441]	VIS–IR	Feature-based	Similarity	Buildings	Harris + Robust estimation
[133]	VIS–IR	Feature-based	Deformation	Face	Points from Edge Extraction + GFC + RKHS
[445]	VIS–IR	Feature-based	Affine	Multiple	Points from edge extraction + GFC
[442]	VIS–IR	Feature-based	Rigid	–	PC + RANSAC
[444]	VIS–IR	Feature-based	Affine + FFD	Building	Edge/Quadrilateral features
[141]	VIS–IR	Feature-based	Affine	Multiple	Harris + PIIFD; Relaxed matching
[178]	VIS–IR	Feature-based	Rigid	Multiple	Edge/C_SIFT+BRIEF+RANSAC
[431]	VIS–IR	Feature-based	Affine	Building	MSPC + RANSAC
[432]	VIS–IR	Feature-based	Affine	Building	Moment ranking + KCF + RANSAC
[452]	VIS–IR	Feature-based	–	–	Video sequence + Segmentation-blob feature
[447]	VIS–IR	Feature-based	Rigid	–	Video sequence + Points from Trajectories+RANSAC
[448]	VIS–IR	Feature-based	Perspective	–	Video sequence + Lines + Points
[451]	VIS–IR	Feature-based	Rigid	–	Video sequence + Points from motion
[450]	VIS–IR	Area-based	Homography	–	Video sequence + MI; Global to local
[149]	VIS–IR	Feature-based	Similarity	–	Video sequence + Motion + FAST
[449]	VIS–IR	Feature-based	Homography	–	Video sequence + Local shape + Polygon vertices
[455]	VIS–IR	Learning	Deformation	Plants	Transfer GAN + STN
[453]	VIS–IR	Learning	Deformation	pedestrian	GAN+FlowNet
[454]	VIS–IR	Learning	Rigid	Buildings	Hybrid CNN + Cross entropy & Hinge Loss
[461]	VIS–NIR	Area-based	–	Face	Edge map/Patch + Template matching
[462]	VIS–NIR	Area-based	–	Face	Learned correlation metric
[182]	Day–night	Learning	–	Multiple	Piecewise linear regression models + DoG for training + NMS
[5]	Day–night	Learning	–	Multiple	Descriptor Learning + Local patch similarity + Spatial constraint
[459]	MultiSpectral	Area-based	–	–	Structure Consistency Boosting (SCB) Transform
[460]	MultiSpectral	Learning	–	–	Gradient map + Structural loss
[8]	CS/image–paint	Learning	–	–	SVM + Defined visual similarity
[6]	Cross-Weather	Learning	Deformation	Road	GAN + Deformation flow estimation
[9]	3D–Paint	Learning	–	Building	3D Scece representation + Visual element matching

intensity. On the basis of this database, we construct our T1, T2, and Pd-weighted MR images as a part of our test data.

- Atlas.[2] This database consists of real CT, MRI, PET and SPECT brain volumes, from which we select 10 MRI–PET and SPECT–CT slice pairs. For each image pair, we warp the moving image with a random affine matrix, thus stimulating a geometric deformation.
- Retina [330]. This database consists of 65 retina image pairs undergoing nonrigid transformations, which are imaged under different angiography techniques. In this database and for some image pairs with slight deformations, we can use an affine model to explain their geometrical transformations, thus preserving it as our test data for feature detection and description.
- CoFSM.[3] This is a newly published multimodal remote sensing image database that consists of 6 types of modality pairs, namely, optical-optical (cross-temporal), IR-optical, depth-optical, map-optical, SAR-optical, and night–day. These raw image pairs help us construct the test data of remote sensing community. Several UAV cross-season image pairs from the 720Yun platform[4] are also collected for experimental evaluation.
- VIS–IR. We collected this VIS–IR subdataset ourselves. It contains rich road scenes such as roads, vehicles, and pedestrians. These images are highly representative scenes from FLIR video.[5] From these image pairs, we also construct an image fusion dataset available at,[6] which are registered by using our manually labeled landmarks and computed affine matrixes.
- VIS–NIR [468].[7] This dataset consists of 477 images in 9 categories captured using separate exposures from modified SLR cameras with visible and NIR filters. The scene categories include country, field, forest, indoor, mountain, old building, street, urban, and water. From each category, we select several typical raw image pairs as our test data. For each image pair, we similarly warp the moving image with a random affine matrix, thus stimulating a geometric deformation.
- WxBS [469].[8] The wide baseline dataset consists of 31 image pairs, simultaneously combining several nuisance factors such as geometry, illumination, and IR–VISible. Each image pair already provides true matched landmarks (points) for evaluation, from which we select the ones that undergo day–night and cross-season changes as a part of our evaluation dataset.
- Vision Cross-Weather/Season.[9] This dataset, which contains a large number of paired images undergoing day–night, weather, or season changes, was originally used for place recognition or visual localization. Following the description of Zhou et al. in [6], we select some representative image pairs from the Philly-Commuting Road Scene (PRS) Dataset, the Nordland Railroad Scene (NRS) Dataset [470], and the RobotCar Seasons (RCS) Dataset [471] as the test data for cross-weather or season image matching.
- image–paint [8].[10] In this dataset, the authors originally provided rich types of multimodal image pairs for cross-domain image matching. We use the painting queries together with some collected from the Internet as our test data. The used images are of 8 places or targets, including Tower Bridge in London, the Sydney Opera, and the Arc de Triomphe. Each target contains a real image, sketch, painting, or drawing.

From these collected raw image pairs, we manually label 15 to 20 matched landmarks (*i.e.,* point locations) for each one (except the WxBS dataset), which could be used to evaluate the registration accuracy

[2] Atlas: http://www.med.harvard.edu/aanlib/home.html.
[3] CoFSM: https://skyearth.org/publication/project/CoFSM/.
[4] 720Yun: https://720yun.com/.
[5] VIS–IR: https://www.flir.com/oem/adas/adas-dataset-form/.
[6] VIS–IR Fusion: https://github.com/jiayi-ma/RoadScene.
[7] VIS–NIR: https://ivrlwww.epfl.ch/supplementary_material/cvpr11/index.html.
[8] WxBS: https://pgram.com/dataset/cmp-wxbs-dataset/home/.
[9] Vision Cross-Weather/Season: https://www.visuallocalization.net/datasets/.
[10] image–paint: http://graphics.cs.cmu.edu/projects/crossDomainMatching/.

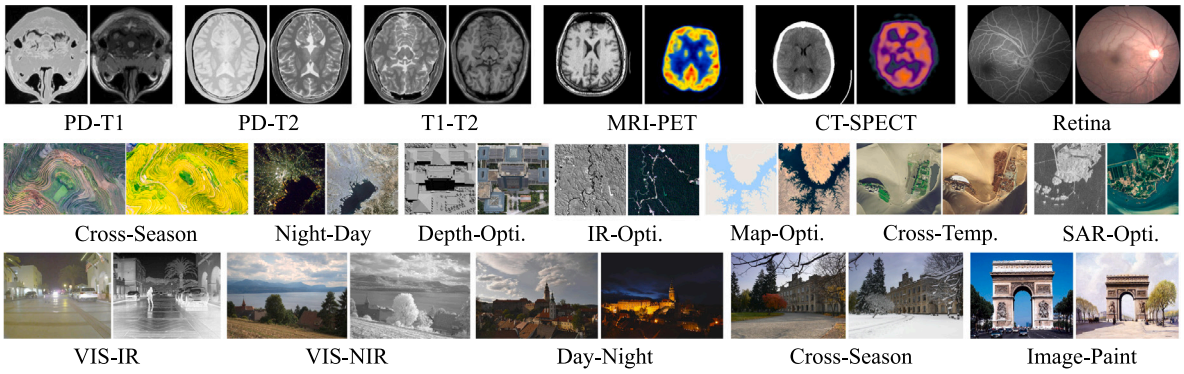

Fig. 4. Selected raw image pairs from our collected database, which covers 18 types of modality pairs regarding the research areas of medical (top line), remote sensing (middle line), and computer vision (bottom line).

based on the distance of these matched landmarks. In our experiments, we first try our best to manually label these matched landmarks that are dispersive and distinct, to estimate the affine matrix to best register each image pair without any visible misalignment, then preserve these landmarks and affine parameters as our Ground Truths. In particular, we estimate the affine parameters based on direct linear transformation (DLT) by using Matlab Toolbox "cp2tform". On the basis of the transformation matrix for each image pair, we can know which two points are matched in the fixed and moving images, and judge whether the matched points created by the feature descriptor are correct or not. While for non-rigid cases or that an affine model cannot explain the geometrical transformation anymore, we just preserve these matched landmarks as ground truths and only use them for image registration test. In fact, the correctness of putative matches for non-rigid cases could be labeled manually one by one, as performed in Refs. [2,333]. And the registration is conducted under a non-rigid model such as TPS [336]. All the collected data and their ground-truth landmarks, as well as transformation matrices are integrated and available at.[11] Some selected raw image pairs in this database are shown in Fig. 4.

4.2. Evaluation on feature detection

On the basis of our collected multimodal image data and with the help of these given transformation matrices, we first conduct the performance comparisons of feature detection. With the existing literature in image matching taken into consideration [129,236], three metrics are commonly used for performance evaluation in keypoint detection task, namely, *repeatability* (*Rep.*), *entropy* (*En.*), and *efficiency* or *runtime* (*RT*).

Suppose the feature detector extracts M and N keypoints from the fixed image and moving image, respectively. Hence, the detected point number (*DPN*) is defined as $DPN = M + N$. The repeatable point number (*RPN*) is identified as the number of matchable points that have the same location in the real world (simultaneously extracted in two images), which are measured through our given ground-truth transformation matrix by transforming these points in the moving image then searching their nearest points in the fixed image within a pixel distance threshold (in our experiment, the threshold equals 5). On the basis of these definitions, repeatability can be defined as

$$Repeatability = \frac{RPN}{DPN}. \tag{2}$$

The entropy can evaluate the influence of the detector on a descriptor in terms of spatial distribution, which measures if the detected keypoints are sufficiently distributed in the image. A high value of this metric indicates that the extracted points are easier to be distinguished by local descriptors due to their nonclustered properties [472]. Following the instruction in [472,473], we first create a $2D$ evenly spaced binning of the feature points and denote the center of each bin as $p = (x, y)$. Each point's contribution to a given bin is weighted by a Gaussian relative to its distance to the bin's center. A bin $b(p)$ at position p can be calculated with $b(p) = \frac{1}{Z}\sum_{m\in M} G(\|p - m\|)$, where m is a keypoint in the full set M of detected interest points, and G is a Gaussian function. A constant of $\frac{1}{Z}$ is added to enable the sum of all bins to evaluate to 1. From these bins, we can obtain

$$Entropy = \sum_{p} -b(p)\log b(p). \tag{3}$$

In this experiment, we choose 12 typical detectors to represent the detection of corner feature, blob feature, and learnable feature. The compared methods include Harris [118], FAST [144], and BRISK [209]. We use three corner detectors with the input of the phase congruency map, as performed by many other researchers [147,148,432] (*i.e.*, PC-Harris, PC-FAST, and PC-BRISK). The blob feature detectors for comparison are DoG (SIFT) [121], SURF [122], and MSER [128]. For learning-based detectors, we choose TILDE [182], LFnet [188], SuperPoint [185] for evaluation. Harris, FAST, BRISK, and SURF are directly implemented with the MATLAB toolbox, in which we restrict the number of detected features in each image within $2k$ by selecting the top responses. DoG (SIFT) and MSER are implemented with VLFeat ToolBox[12] [474]. We directly use the authors' source codes to apply these learning-based detectors in our evaluation. All the handcrafted detectors are performed on a desktop with 4.0 GHz Intel Core i7-6700K CPU, 16 GB memory. The deep methods are performed on a server with 2.0 GHz Intel Xeon CPU, 128 GB memory.

The qualitative results of representative feature detectors on typical multimodal image pairs are presented in Figs. 5 6 7. For each image pair, the repeated points under a threshold of 3 and 5 pixels are denoted by blue and green stars, respectively, while the unmatchable points whose distances are beyond 5 pixels are shown with red dots. All the quantitative results on 18 datasets are shown in Tables 4–9, which are classified into three parts in regard to medical, remote sensing, and computer vision communities. For each dataset, the average DPN, RPN, $Rep.$ (%), $En.$, and RT (ms) of each method are depicted in these tables. The overall detection performance for these three research communities is also shown at the end of the tables, and the first, second, and third best results are indicated in bold red, green, and blue, respectively.

From the results, we can see that FAST and SURF detectors can still achieve promising detection performance in MMIM in terms of repeatability and execution efficiency. The combination of PC map can greatly enhance the performance of corner detectors but requires additional

[11] Available at: https://github.com/StaRainJ/Multi-modality-image-matching-database-metrics-methods.

[12] VLFeat ToolBox: https://www.vlfeat.org/.

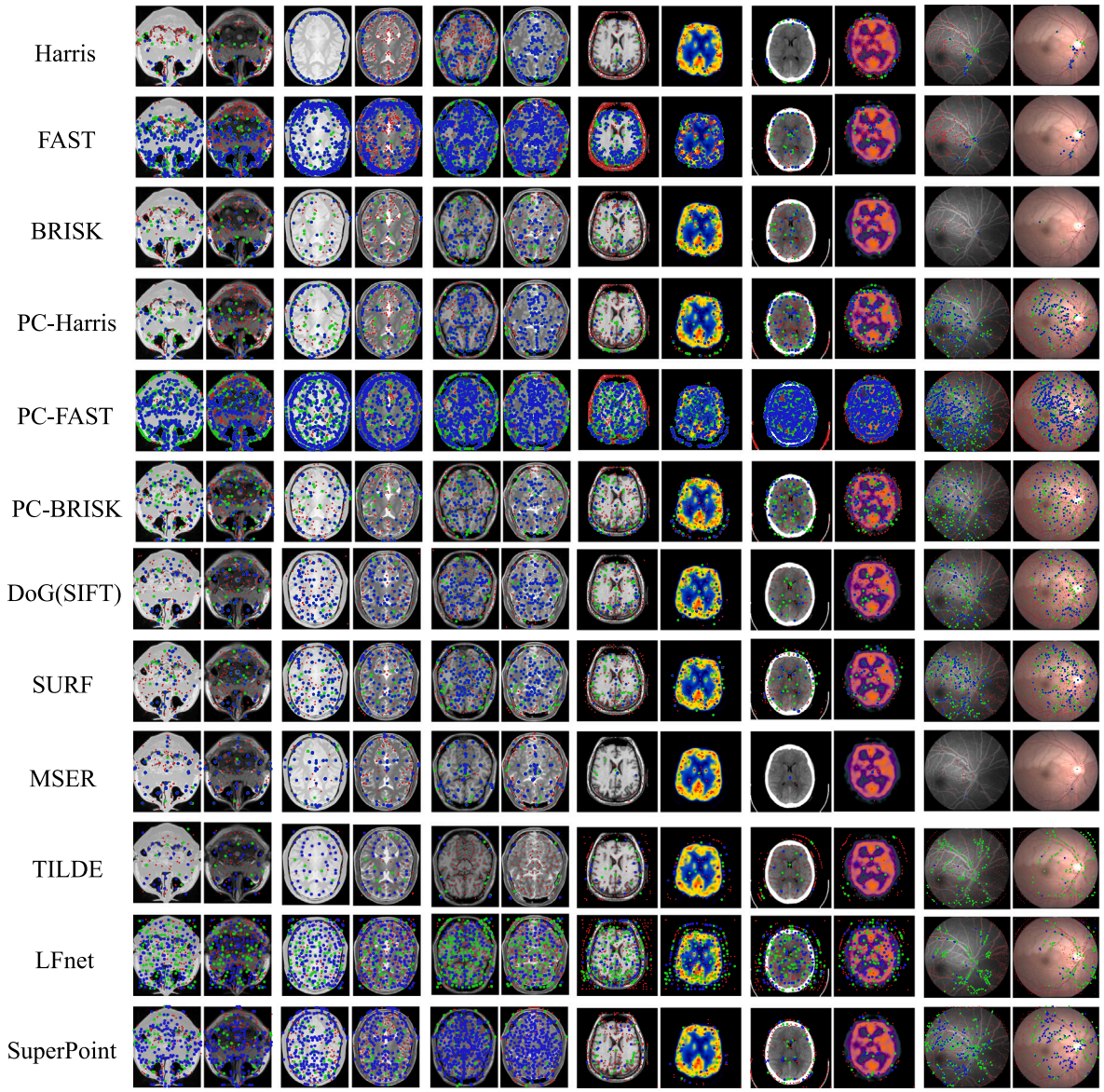

Fig. 5. Qualitative results of 12 feature detectors on typical multimodal image pairs in the medical research area. (blue = repeated points with threshold of 3 pixels, green = repeated points with threshold of 5 pixels, red = unmatchable points). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

computational burden. As for learning-based methods, SuperPoint can maintain good generalization ability. Readers can refer to the tables for more detailed results or test other detectors with our public datasets and ground truth.

4.3. Evaluation on feature description

In this part, we will test several representative feature descriptors on these multimodal image datasets. As introduced in [472], three metrics are mainly used for evaluation, namely, *matching score*, *precision*, and *recall*. Before evaluation, suppose we have obtained a set of N_0 putative matches $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N_0}$ created by a combination of feature detector and descriptor, where $\mathbf{x}_i$ and $\mathbf{y}_i$ are pixel coordinates of interest points in fixed and moving images, respectively, from which the correct match set C is identified by using our ground-truth transformation within a pixel distance threshold ϵ, *i.e.*, $C = \{(\mathbf{x}_j, \mathbf{y}_j)\}_{j=1}^{N_c} | \ \|\mathbf{x}_j - \mathcal{T}(\mathbf{y}_j)\| \leq \epsilon$, where $N_c \leq N_0$ and $\mathcal{T}$ indicates the ground-truth geometrical transformation, and in our experiment, ϵ equals 5. Therefore, the putative match number (PMN) is defined as the cardinality of the putative match set, *i.e.*, $PMN = |S| = N_0$. The correct match number (CMN) is defined as the cardinality of the correct match set, *i.e.*, $CMN = |C| = N_c$. On the basis of these definitions, the *matching score* (MS) can be calculated as the ratio between the number of features that belong to correct matches and the number of all detected features. The *Precision*, also known as inlier ratio, defines the number of correct matches out of the set of putative matches, which is calculated with

$$Precision = \frac{CMN}{PMN}. \tag{4}$$

Recall quantifies how many of the ground-truth correct matches are actually found by descriptor matching. Inspired by [472], we define *Recall* as the ratio between the number of features belonging to correct matches and the number of all repeatable features. The *RunTime* (RT) during the entire procedure of feature detection and description is used to evaluate the execution efficiency.

In this experiment, we select three well-known classical feature matchers in the computer vision area for comparison, namely, SIFT [121], SURF [122], and ORB [120], as well as two deep matchers, *i.e.,* SuperPoint [185] and LFnet [188]. Two widely used descriptors

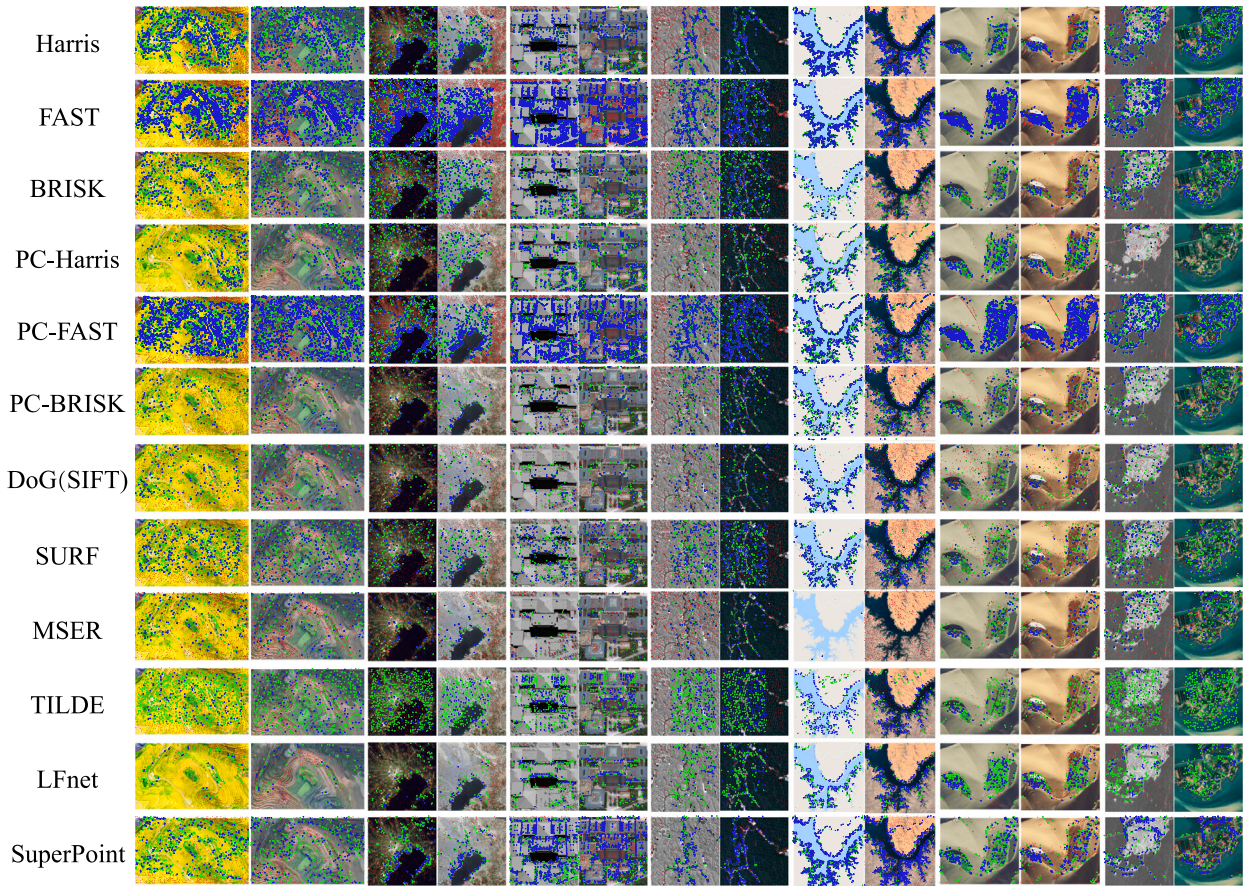

Fig. 6. Qualitative results of 12 feature detectors on typical multimodal image pairs in the remote sensing research area. (blue = repeated points with threshold of 3 pixels, green = repeated points with threshold of 5 pixels, red = unmatchable points). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

tailored for multimodal images called PIIFD [140] (only test on medical data) and RIFT [148] are also used for comparison. Five top-performing detectors that are measured in the detection experiment are used in combination with PIIFD and RIFT for evaluation. The chosen detectors are SURF, PC-FAST, PC-BRISK, TILDE, and SuperPoint. Similar to the feature detection experiment, SIFT is implemented with VLFeat Tool-Box, and the deep methods are implemented with their source codes provided by the authors themselves. ORB is implemented with OpenCV library [475], and the others are implemented by MATLAB Toolbox. All the handcrafted methods are performed on a desktop with 4.0 GHz Intel Core i7-6700K CPU, 16 GB memory, while the deep methods are performed on a server with 2.0 GHz Intel Xeon CPU, 128 GB memory.

The qualitative results of representative feature descriptor matching methods on typical multimodal image pairs are presented in Figs. 8 9 10. For each image pair, the correct matches under a threshold of 3 and 5 pixels are denoted by blue and green lines, respectively, while incorrect matches whose distances are beyond 5 pixels are shown with red lines. All the quantitative results for feature descriptor matching on 18 datasets are listed in Tables 10–15, which are classified into three parts in regard to medical, remote sensing, and computer vision community. For each dataset, the average correct match number, putative match number, matching score (%), precision (%), recall (%), and runtime (ms) of each method are listed in these tables. The overall descriptor matching performance for these three research communities is shown at the end of the tables. The first, second, and third best results are indicated in bold red, green, and blue, respectively.

From the results, we can see that classical descriptors such as SIFT, SURF, and ORB are not workable anymore in most multimodal cases, while RIFT and SuperPoint can handle most types of MMIM tasks, but RIFT is typically time consuming. In particular and for the medical data, feature matching on different weighted MR images can be successfully achieved due to the good preservation of structure information between two modalities. In contrast, the matching for MRI–PET or SPECT–CT modality pair would not be easy to realize, because few cues can be used in these two cases even by using manual matching. As for retina images, the RIFT can achieve satisfying performance, while PIIFD can also perform well but consumes too much time. In these datasets, we use PC-FAST detector and RIFT descriptor to construct a putative match set for subsequent mismatch removal and image registration. The comparison experiments on remote sensing datasets show that only the RIFT descriptor can achieve promising performance with a high number and ratio of correct matches. We also choose PC-FAST and RIFT for putative match set construction. This superiority is quite different in the visual domain. Several matchers proposed for visual applications maintain their feasibility in visible–NIR, cross-season, and day–night image matching due to the unremarkable difference of image intensity. In our evaluation and VIS–IR image matching, only RIFT can achieve promising performance. More efforts are still required to handle the matching of image and painting. Among these vision datasets, we use SuperPoint's detector and RIFT to create putative matches.

4.4. Evaluation on mismatch removal

As shown in the feature matching results based on the similarity of local descriptors, the use of only image local information to search matched interest features would inevitably create a large number and high ratio of false matches. This situation may largely damage the accuracy of image transformation or deformation estimation and affect

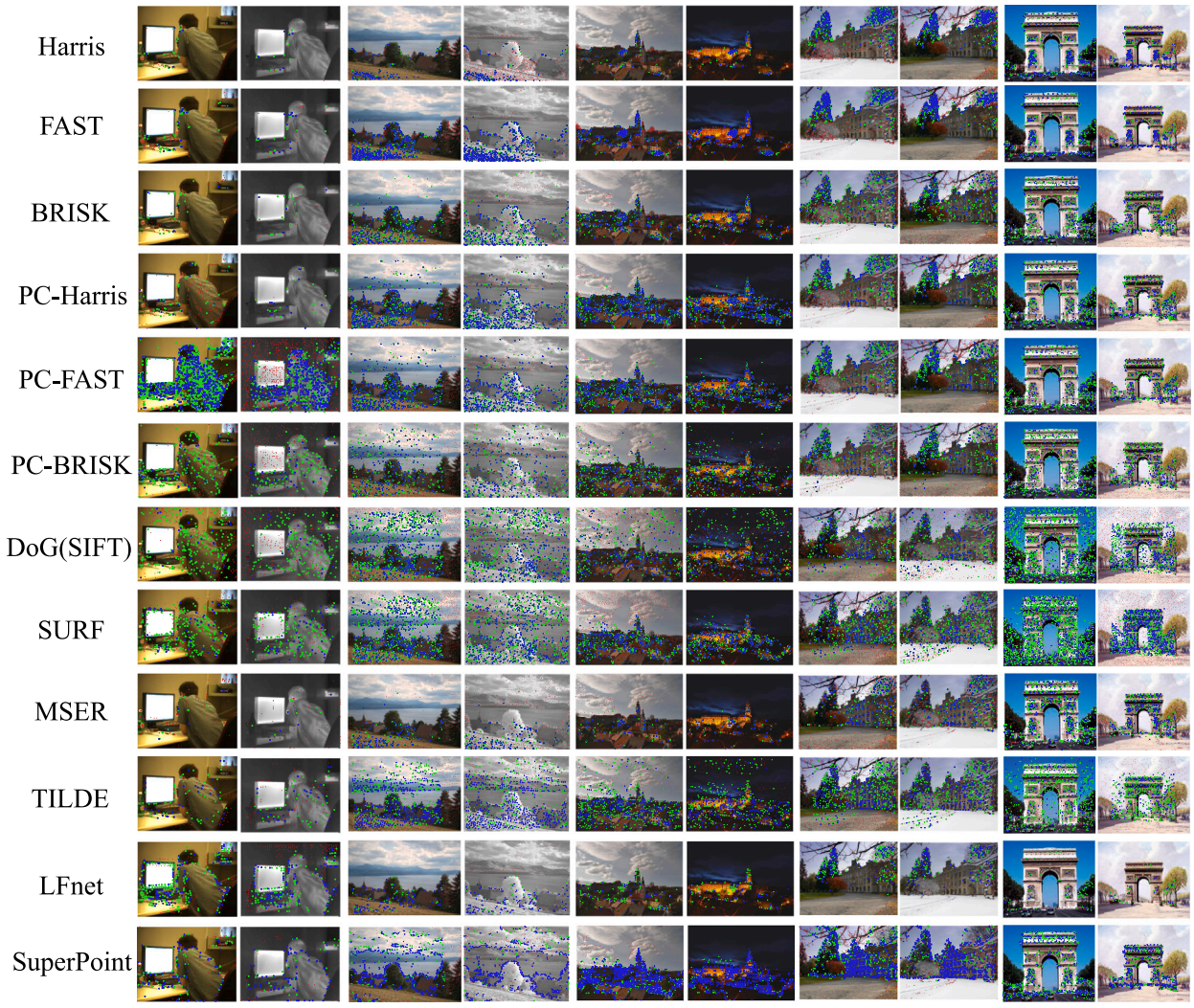

Fig. 7. Qualitative results of 12 feature detectors on typical multimodal image pairs in the computer vision research area. (blue = repeated points with threshold of 3 pixels, green = repeated points with threshold of 5 pixels, red = unmatchable points). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the registration considerably. A mismatch removal procedure needs to be integrated to find as many correct matches as possible and keep the mismatches to a minimum. We first create a set of putative matches for each image pair as introduced in the last experiment, from which we can know which match is correct in advance, referring to the definition of a correct match set C. This correct match set would serve as our ground truth in the mismatch removal experiment. On the basis of the putative match set and the correct match labels, three evaluation metrics are employed, namely, *Precision*, *Recall*, and *F-score*. By verifying the consistency between the matches identified by mismatch removal methods and ground-truth correct match sets, we can obtain the number of true positives (TP), true negatives (TN), false positives (TP), and false negatives (FN). Thus, the *Precision* and *Recall* can be obtained by

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}. \tag{5}$$

While *F-score*, as a summary statistic of precision and recall, is calculated as follows:

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{6}$$

In this experiment, 12 mismatch removal methods are used for comparison, namely, RANSAC [301], MAGSAC++ [317], ICF [323], VFC [319], LLT [325], GS [243], SM [241], LPM [330], GMS [329], mTopKRP [332], LFGC [341], and LMR [342]. These methods are representative of resampling-based, nonparametric model-based, graph-based, relaxed, and learning-based methods. All the methods are performed on a desktop with 4.0 GHz Intel Core i7-6700K CPU, 16 GB memory by using the source codes provided by their authors.

The qualitative results of representative mismatch removal methods on typical multimodal image pairs are presented in Figs. 11 12 13. For visibility, in each image pair, at most 100 randomly selected matches are presented, and the true negatives are not shown. The blue, green, and red lines in these three plots represent TP, FN, and FP matches, respectively, preserved by the evaluated mismatch removal methods. All the quantitative results for *Precision*, *Recall*, *F-score*, and *RunTime* on each dataset are shown in Fig. 14. Methods that use resampling or nonparametric model can achieve satisfying precision due to their global geometrical constraints on these test data that are of linear transformations. Graph-based methods are also acceptable, but they are limited by the large computational burden. Relaxed methods such as GMS and LPM are easy to implement and surprisingly efficient due to their relaxed geometrical constraints. Learning-based methods show promising ability in mismatch removal task by learning from sparse point sets. LMR can also achieve satisfying results, because the handcrafted high-dimensional match representations can easily learn to identify outliers. However, directly learning geometrical properties

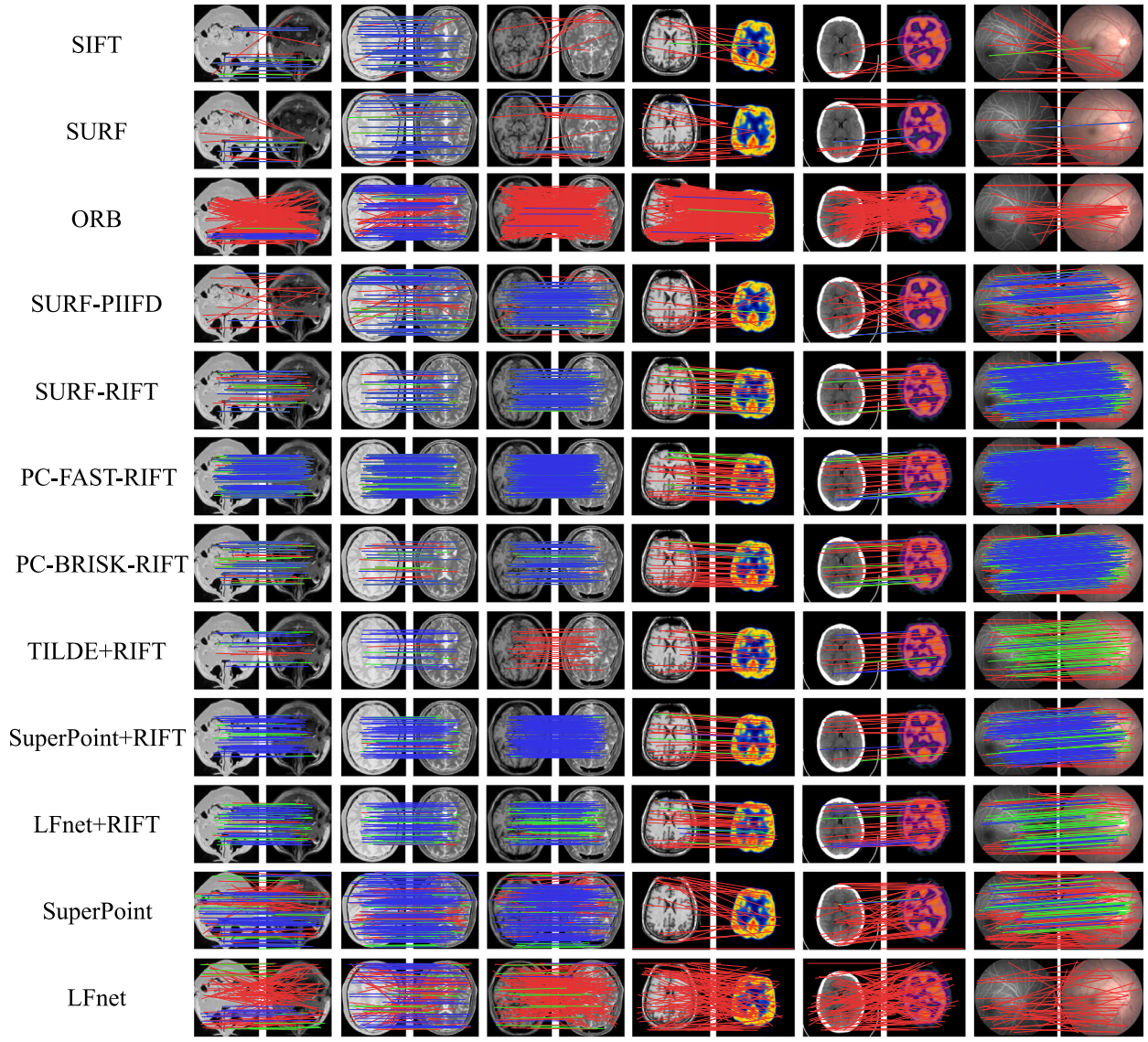

Fig. 8. Qualitative descriptor matching results of 12 methods on typical multimodal image pairs in the medical research area. (blue = correct matches with threshold of 3 pixels, green = correct matches with threshold of 5 pixels, red = incorrect matches). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to guide the network to filter outliers by using a deep convolutional approach, such as LFGC, remains a challenging problem.

as performed in Refs. [2,333]. And the registration is conducted under a non-rigid model such as TPS [336].

4.5. Evaluation on image registration

In this part, we will test the image registration performance by using the final matched features created by the above feature detectors, descriptors, and mismatch removal methods. As aforementioned, for image pairs with ground-truth affine matrices, we directly use direct linear transformation to estimate their optimal transformation parameters. While for non-rigid cases, we model their transformations as TPS, and directly estimate parameters for non-rigid registration, as performed in [336]. To evaluate the registration accuracy, two types of metrics are chosen. The first one is obtained from the distance of our ground-truth matched landmarks, which is inspired from [476] and denoted as TRE (*i.e.,* target registration error). Metrics of this type are more objective for measuring registration accuracy [336]. In our experiment, this is indicated by the *root mean square error* (*RMSE*), *maximum error* (*MAE*), and *median error* (*MEE*) between the landmark pairs with the following definitions:

$$RMSE = \sqrt{1/L \sum_{i=1}^{L} (\mathbf{f}_i - \mathcal{T}(\mathbf{m}_i))^2}, \tag{7}$$

$$MAE = \max\left\{ \sqrt{(\mathbf{f}_i - \mathcal{T}(\mathbf{m}_i))^2} \right\}_{i=1}^{L}, \tag{8}$$

$$MEE = median\left\{ \sqrt{(\mathbf{f}_i - \mathcal{T}(\mathbf{m}_i))^2} \right\}_{i=1}^{L}, \tag{9}$$

where $\{\mathbf{f}_i, \mathbf{m}_i\}$ is the $i-$th ground-truth matched landmark, $\mathbf{f}_i$ and $\mathbf{m}_i$ are respectively annotated from fixed image and moving image. $\mathcal{T}$ is the transformation function from the moving image to the fixed image, L represents the number of used landmarks, and $\max(\cdot)$ and $median(\cdot)$ return the maximal and median value of a set, respectively.

Another type of metric is typically designed consistent with the perception of human beings, which are widely used in case no landmark or other golden standard is available to evaluate the accuracy of image registration. These evaluation metrics commonly include *peak signal-to-noise ratio* (*PSNR*), *structural similarity* (*SSIM*) [411], and *mutual information* (*MI*) between the warped moving image and fixed image.

Table 4
Feature detection results about Detected Point Number, Repeated Point Number, Repeatability, Entropy metrics and RunTime of 12 feature detectors on multimodal datasets of medical field (Part I).

Datasets	PD–T1					PD-T2					T1–T2				
Detectors\Metrics	*DPN*	*RPN*	*Rep.* (%)	*En.*	*RT* (ms)	*DPN*	*RPN*	*Rep.* (%)	*En.*	*RT* (ms)	*DPN*	*RPN*	*Rep.* (%)	*En.*	*RT* (ms)
Harris	149.55	36.40	23.94	5.18	14.36	157.60	81.20	50.13	5.18	14.28	218.86	132.18	59.78	5.40	14.73
FAST(ORB)	466.70	317.80	67.69	5.51	2.81	645.90	387.40	59.67	5.54	2.99	719.05	508.64	69.96	5.60	2.83
BRISK	123.50	50.60	39.36	5.34	6.55	135.40	68.30	48.32	5.38	6.71	148.59	83.09	55.02	5.39	6.93
PC-Harris	197.45	90.50	45.36	5.54	130.36	193.50	118.00	60.56	5.51	131.30	226.05	146.09	64.50	5.57	130.77
PC-FAST	717.45	576.80	80.35	5.67	117.44	730.25	636.00	86.89	5.69	118.30	771.45	685.45	88.90	5.68	119.61
PC-BRISK	169.95	82.60	47.62	5.54	121.27	157.95	85.80	52.85	5.52	121.52	180.23	103.36	57.10	5.55	127.35
DoG(SIFT)	157.20	85.60	54.43	5.41	65.04	158.05	89.90	56.61	5.42	64.56	165.27	93.27	55.86	5.45	64.31
SURF	308.55	221.30	71.08	5.60	10.44	352.50	256.10	72.45	5.63	10.96	354.73	258.36	72.41	5.66	11.25
MSER	101.90	59.80	58.38	5.16	44.96	147.40	85.60	58.06	5.30	59.45	147.59	75.82	51.60	5.26	61.07
TILDE	66.95	25.90	39.00	5.46	831.76	71.85	51.70	71.59	5.48	806.44	74.05	23.64	32.24	5.50	798.43
LFnet	323.45	230.70	71.17	5.79	500.84	345.95	273.50	78.98	5.79	495.96	351.41	244.73	69.61	5.78	564.56
SuperPoint	270.25	212.60	78.04	5.58	150.64	289.65	237.70	81.86	5.61	151.09	282.73	235.27	83.08	5.58	154.40

Table 5
Feature detection results about Detected Point Number, Repeated Point Number, Repeatability, Entropy metrics and RunTime of 12 feature detectors on multimodal datasets of medical field (Part II).

Datasets	MRI–PET					SPECT–CT					Retina					All medical data				
Detectors\Metrics	*DPN*	*RPN*	*Rep.* (%)	*En.*	*RT* (ms)	*DPN*	*RPN*	*Rep.* (%)	*En.*	*RT* (ms)	*DPN*	*RPN*	*Rep.* (%)	*En.*	*RT* (ms)	*DPN*	*RPN*	*Rep.* (%)	*En.*	*RT* (ms)
Harris	200.75	28.40	14.02	5.07	19.88	214.45	56.50	25.55	5.17	20.22	377.66	132.64	32.39	4.95	77.10	246.83	87.88	34.23	5.12	35.14
FAST(ORB)	966.20	391.80	40.44	5.22	4.59	243.00	74.10	30.02	5.08	2.80	755.41	359.05	33.22	4.75	3.73	653.03	343.33	47.39	5.19	**3.36**
BRISK	221.85	65.20	29.69	5.15	11.74	126.00	32.00	25.32	5.05	6.34	246.86	71.73	23.89	4.64	12.07	180.77	64.08	35.06	5.07	**9.01**
PC-Harris	236.70	75.10	30.36	5.40	192.45	308.35	158.10	50.79	5.43	179.35	912.61	555.91	57.09	5.72	1342.81	440.53	251.79	52.44	5.56	516.39
PC-FAST	902.90	672.70	73.11	5.46	159.89	1091.85	938.30	85.53	5.44	155.08	1990.80	1565.55	78.63	5.81	1244.39	1193.19	965.22	**81.61**	**5.65**	473.35
PC-BRISK	254.45	96.90	37.85	5.45	158.68	264.05	108.20	40.96	5.42	159.73	1214.52	825.41	66.21	5.82	1254.32	513.94	318.81	53.20	**5.60**	479.00
DoG(SIFT)	191.40	71.30	36.81	5.23	91.39	116.05	28.10	24.07	5.15	73.22	1238.23	834.18	66.47	5.80	521.63	487.67	305.88	51.89	5.48	209.17
SURF	380.30	166.20	43.11	5.30	14.42	333.60	156.40	46.91	5.41	13.22	1153.02	832.18	69.85	5.54	54.94	592.01	400.64	**63.77**	5.52	**25.16**
MSER	77.45	6.10	8.08	4.66	45.22	27.15	2.00	7.48	3.83	21.48	187.73	53.36	30.28	4.36	114.39	127.18	48.17	34.70	4.69	67.29
TILDE	112.75	26.90	23.71	5.51	1066.23	127.40	19.90	15.63	5.50	1138.04	533.41	207.59	37.61	5.77	4424.79	225.98	83.60	36.19	5.58	1996.94
LFnet	407.00	228.60	56.12	5.72	563.92	419.85	235.50	56.10	5.70	544.66	500.00	224.50	44.90	5.52	574.69	409.66	237.08	59.78	**5.68**	546.56
SuperPoint	189.65	51.10	27.17	5.37	188.88	188.40	60.90	31.52	5.34	183.14	494.14	304.41	60.32	5.68	455.30	320.13	203.72	**60.41**	5.55	254.01

Table 6
Feature detection results about Detected Point Number, Repeated Point Number, Repeatability, Entropy metrics and RunTime of 12 feature detectors on multimodal datasets of remote sensing field (Part I).

Datasets	UAV Cross-season					Day–night					Depth-optical					Infrared-optical				
Detec-tors\Metrics	*DPN*	*RPN*	*Rep. (%)*	*En.*	*RT* (ms)	*DPN*	*RPN*	*Rep. (%)*	*En.*	*RT* (ms)	*DPN*	*RPN*	*Rep. (%)*	*En.*	*RT* (ms)	*DPN*	*RPN*	*Rep. (%)*	*En.*	*RT* (ms)
Harris	1585.50	1121.75	68.74	5.67	61.43	1032.57	391.00	31.91	5.33	77.74	1313.25	1021.17	76.29	5.51	61.78	1287.30	571.80	41.10	5.58	109.18
FAST(ORB)	2000.00	1340.00	67.00	5.73	10.52	1462.29	681.57	44.66	5.57	6.01	1997.33	1614.83	80.87	5.60	6.98	1994.30	1245.20	62.44	5.71	12.45
BRISK	1844.25	1262.75	68.47	5.93	34.05	808.07	199.86	24.65	5.60	19.33	1488.83	1183.00	76.30	5.76	26.64	1156.20	674.60	55.09	5.83	41.97
PC-Harris	699.88	361.75	51.85	5.56	979.52	1359.79	866.43	62.00	5.82	1159.92	1121.67	877.67	77.63	5.66	941.46	1005.90	529.20	51.74	5.79	1459.39
PC-FAST	1982.13	1378.50	69.52	5.76	904.56	2000.00	1466.14	73.31	5.85	1082.35	2000.00	1625.00	81.25	5.75	892.10	2000.00	1413.40	70.67	5.87	1339.21
PC-BRISK	776.88	399.50	51.16	5.74	907.03	1395.36	749.43	53.22	5.86	1099.85	1074.50	801.17	73.53	5.79	901.23	1061.90	593.00	55.25	5.89	1369.46
DoG(SIFT)	835.00	437.00	49.83	5.86	391.41	1450.93	704.86	49.59	5.82	585.07	1171.25	836.33	68.92	5.91	483.21	1508.20	969.60	64.63	5.92	609.47
SURF	2000.00	1378.25	68.91	5.83	77.40	1577.07	910.71	56.16	5.74	59.12	1828.33	1471.17	80.23	5.76	61.76	1892.50	1396.80	74.36	5.84	90.14
MSER	1307.25	808.75	52.39	5.82	507.71	632.43	87.43	13.70	5.47	269.39	1360.17	985.33	67.53	5.68	529.16	1934.30	946.80	48.93	5.85	785.95
TILDE	813.50	467.25	57.39	5.91	3759.70	643.86	331.57	48.99	5.83	4256.35	604.33	430.33	69.33	5.82	3408.24	844.30	514.80	58.89	5.87	4574.00
LFnet	500.00	241.75	48.35	5.67	551.71	500.00	226.14	45.23	5.63	624.99	500.00	318.17	63.63	5.53	475.40	500.00	222.60	44.52	5.59	629.17
SuperPoint	785.38	431.25	58.01	5.55	402.83	1380.79	926.43	64.48	5.75	480.68	1531.83	1219.17	78.67	5.79	414.55	1075.50	638.20	57.71	5.85	500.01

Table 7
Feature detection results about Detected Point Number, Repeated Point Number, Repeatability, Entropy metrics and RunTime of 12 feature detectors on multimodal datasets of remote sensing field (Part II).

Datasets	Map-optical					Optical cross-temporal					SAR-optical					All remote sensing data				
Detec-tors\Metrics	*DPN*	*RPN*	*Rep. (%)*	*En.*	*RT* (ms)	*DPN*	*RPN*	*Rep. (%)*	*En.*	*RT* (ms)	*DPN*	*RPN*	*Rep. (%)*	*En.*	*RT* (ms)	*DPN*	*RPN*	*Rep. (%)*	*En.*	*RT* (ms)
Harris	1493.25	902.13	61.56	5.73	78.42	1387.08	828.33	59.55	5.58	114.05	1266.50	771.17	53.04	5.41	69.24	1327.46	786.29	55.47	5.54	81.75
FAST(ORB)	2000.00	1210.13	60.51	5.74	7.90	2000.00	1391.00	69.55	5.63	9.42	1804.50	1347.33	74.28	5.64	8.53	1881.39	1241.83	**64.88**	5.66	**8.55**
BRISK	1687.06	1129.38	67.21	5.84	30.69	1646.33	1089.50	68.36	5.80	36.50	1293.58	849.67	60.27	5.72	25.24	1401.99	895.02	59.27	5.77	**29.93**
PC-Harris	1344.06	870.38	63.98	5.81	1194.83	1193.17	670.83	56.99	5.70	1948.15	762.33	458.50	56.15	5.53	1101.15	1108.64	694.36	60.87	5.71	1258.04
PC-FAST	2000.00	1367.13	68.36	5.86	1104.13	2000.00	1442.50	72.13	5.77	1798.92	1837.83	1410.00	75.94	5.68	1035.48	1975.13	1443.95	**73.03**	5.79	1168.64
PC-BRISK	1253.31	842.88	65.40	5.88	1122.60	1166.50	602.67	56.85	5.82	1836.75	866.33	565.67	60.06	5.71	1049.66	1115.60	675.45	59.98	**5.82**	1187.64
DoG(SIFT)	1112.63	578.00	50.90	5.86	501.25	2123.67	1116.00	55.69	5.90	879.75	956.50	552.00	54.42	5.87	445.31	1320.17	742.38	55.98	**5.88**	561.15
SURF	2000.00	1451.50	72.57	5.80	86.89	1992.33	1471.17	73.90	5.76	109.71	1815.92	1528.67	84.50	5.73	68.72	1864.80	1364.52	**72.69**	5.78	**78.82**
MSER	1238.63	514.38	39.62	5.75	496.54	1569.08	800.17	58.90	5.74	637.66	1117.50	607.83	46.19	5.61	447.29	1274.21	644.19	45.30	5.69	511.99
TILDE	946.44	672.88	68.70	5.85	4189.59	1206.42	670.67	66.09	5.90	6184.59	792.33	513.83	61.16	5.88	4011.00	837.44	519.90	61.81	**5.86**	4353.40
LFnet	500.00	223.00	44.60	5.65	545.30	500.00	223.00	44.60	5.60	519.15	500.00	254.67	50.93	5.58	663.33	500.00	243.38	48.68	5.61	572.32
SuperPoint	1893.00	1233.88	65.13	5.85	547.57	1152.08	658.17	55.91	5.72	664.04	1027.25	601.00	57.03	5.66	439.72	1323.70	860.52	62.92	5.75	499.20

Table 8
Feature detection results about Detected Point Number, Repeated Point Number, Repeatability, Entropy metrics and RunTime of 12 feature detectors on multimodal datasets of computer vision field (Part I).

Datasets	Visible–infrared					Visible-near infrared					Cross-season				
Detectors\Metrics	*DPN*	*RPN*	*Rep. (%)*	*En.*	*RT* (ms)	*DPN*	*RPN*	*Rep. (%)*	*En.*	*RT* (ms)	*DPN*	*RPN*	*Rep. (%)*	*En.*	*RT* (ms)
Harris	84.85	10.30	10.90	4.17	43.20	1131.43	766.27	65.62	5.03	154.57	736.25	349.17	51.00	5.07	132.45
FAST(ORB)	359.85	121.30	27.64	4.86	2.60	1819.87	1328.27	72.74	5.25	7.88	1764.17	1200.33	69.86	5.20	11.04
BRISK	135.00	28.90	17.70	4.97	7.30	1355.05	907.17	65.53	5.50	31.01	1061.00	523.17	44.88	5.43	85.02
PC-Harris	403.90	183.50	42.02	5.56	664.90	1634.27	1185.87	71.72	5.59	2869.17	1071.00	613.67	53.36	5.33	2208.21
PC-FAST	1553.05	1300.80	80.11	5.74	613.72	2000.00	1473.57	73.68	5.64	2816.17	1477.50	954.50	67.81	5.44	2081.55
PC-BRISK	668.45	403.20	51.87	5.77	659.41	1756.70	1138.73	64.79	5.73	2827.00	1118.17	501.67	39.33	5.65	2206.35
DoG(SIFT)	810.65	478.90	48.82	5.85	314.07	2416.33	1602.37	65.01	5.90	1052.03	2342.75	1411.33	42.19	5.81	967.41
SURF	570.30	356.60	52.27	5.40	26.66	1963.47	1447.93	73.74	5.55	137.05	1261.92	694.50	57.81	5.35	122.41
MSER	179.60	53.60	18.76	4.94	84.75	1481.98	1040.53	65.67	5.45	589.16	1218.33	536.50	35.66	5.29	509.21
TILDE	268.85	96.30	33.13	5.65	2536.29	1457.37	1018.37	68.81	5.83	9449.45	1181.00	580.83	40.74	5.65	7774.93
LFnet	497.70	246.90	49.60	5.64	609.28	500.00	290.33	58.07	5.31	539.01	500.00	237.17	47.43	5.38	573.25
SuperPoint	493.60	322.00	57.40	5.59	327.15	1587.42	1181.40	73.63	5.60	861.56	1077.58	658.83	57.16	5.40	770.94

Table 9
Feature detection results about Detected Point Number, Repeated Point Number, Repeatability, Entropy metrics and RunTime of 12 feature detectors on multimodal datasets of computer vision field (Part II).

Datasets	Day–night					image–paint					All vision data				
Detectors\Metrics	*DPN*	*RPN*	*Rep. (%)*	*En.*	*RT* (ms)	*DPN*	*RPN*	*Rep. (%)*	*En.*	*RT* (ms)	*DPN*	*RPN*	*Rep. (%)*	*En.*	*RT* (ms)
Harris	1232.50	242.50	22.31	5.14	271.96	2000.00	1079.00	53.95	5.36	219.53	891.31	545.92	50.65	4.87	135.25
FAST(ORB)	2000.00	624.00	31.20	5.21	11.32	2000.00	1027.00	51.35	5.38	14.14	1526.11	1031.39	61.05	5.16	**7.46**
BRISK	1667.50	499.00	33.36	5.47	50.15	2000.00	852.00	42.60	5.58	58.45	1095.97	663.12	51.46	5.38	**34.12**
PC-Harris	2000.00	956.00	47.80	5.55	4482.43	2000.00	918.00	45.90	5.55	4150.50	1336.59	896.39	61.91	5.55	2430.38
PC-FAST	2000.00	913.50	45.67	5.59	4406.55	2000.00	865.00	43.25	5.55	3919.05	1844.81	1339.47	**72.51**	5.63	2364.16
PC-BRISK	2000.00	823.00	41.15	5.81	4477.91	2000.00	869.00	43.45	5.67	4212.88	1471.32	892.22	57.63	**5.73**	2404.30
DoG(SIFT)	5160.75	2729.00	56.21	5.94	2012.37	3888.00	1951.00	50.18	5.95	1706.27	2221.68	1402.80	58.25	**5.88**	943.61
SURF	2000.00	743.00	37.15	5.40	248.86	2000.00	1200.00	60.00	5.42	222.35	1595.48	1099.12	**65.63**	5.49	**119.04**
MSER	2323.75	691.50	32.17	5.48	947.33	3058.50	1654.00	54.08	5.49	1159.11	1250.44	775.67	50.81	5.33	502.68
TILDE	2457.00	1109.50	49.72	5.86	14473.13	2333.50	987.00	42.30	5.89	13359.46	1239.65	779.69	56.77	**5.77**	8118.40
LFnet	500.00	116.00	23.20	5.20	552.80	500.00	110.00	22.00	5.28	495.38	499.53	264.16	52.88	5.38	557.21
SuperPoint	2000.00	1176.50	58.83	5.50	1435.65	2000.00	1225.00	61.25	5.51	1284.13	1327.02	942.71	**67.44**	5.57	773.46

Table 10
Descriptor matching results about Correct Match Number, Putative Match Number, Match Score, Precision, Recall metrics of 12 feature matchers on multimodal datasets of medical field (Part I).

Datasets	PD–T1					PD-T2					T1–T2				
Matchers\Metrics	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*
SIFT	3.40	10.00	2.33	24.50	4.53	38.90	44.90	24.64	86.62	43.62	4.09	10.45	2.38	17.37	4.11
SURF	55.60	80.20	17.57	72.03	24.54	81.90	103.20	23.21	80.90	32.02	77.45	96.91	21.42	79.79	29.11
ORB	11.80	211.30	2.56	5.27	3.81	148.50	310.40	23.90	49.57	39.57	15.64	347.00	2.19	6.94	3.29
SURF-PIIFD	37.00	61.20	11.43	55.87	15.60	77.40	93.50	22.02	82.71	30.58	75.64	96.18	20.89	76.83	28.41
SURF-RIFT	54.00	59.40	17.92	89.61	25.00	62.00	67.10	18.04	92.39	24.97	68.09	71.82	19.45	93.98	26.89
PC-FAST-RIFT	114.80	118.40	16.61	96.58	20.62	118.60	122.70	16.94	96.05	19.74	158.36	159.18	21.09	99.36	23.68
PC-BRISK-RIFT	26.20	34.10	15.91	76.67	34.15	25.90	33.00	17.26	76.91	33.17	36.00	41.09	20.66	86.60	36.12
TILDE-RIFT	4.80	15.40	7.68	27.02	14.17	21.10	22.70	29.24	92.26	40.69	5.27	18.00	7.53	25.44	16.47
SuperPoint-RIFT	73.50	76.80	28.97	95.45	36.70	77.50	80.30	28.87	96.66	34.67	95.91	96.45	36.41	99.30	42.88
LFnet-RIFT	55.2	69.1	17.05	79.52	23.96	78.3	83.4	22.62	93.88	28.66	68.64	80.73	19.47	84.75	27.98
SuperPoint	68.80	131.90	25.24	51.76	32.28	161.90	197.50	54.95	80.51	67.04	114.73	162.73	40.33	69.79	48.53
LFnet	14.1	116.7	4.35	11.88	6.13	84.6	164.6	24.52	50.88	31.01	17.18	148.45	4.94	11.8	7.01

Table 11
Descriptor matching results about Correct Match Number, Putative Match Number, Match Score, Precision, Recall metrics of 12 feature matchers on multimodal datasets of medical field (Part II).

Datasets	MRI–PET					SPECT–CT					Retina					All Medical Data					
Match-ers\Metrics	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*RT (ms)*
SIFT	0.90	12.60	0.45	5.16	1.08	0.00	6.10	0.00	0.00	0.00	35.55	70.82	3.94	21.54	5.71	16.92	32.83	5.05	24.12	8.66	**244.12**
SURF	0.00	20.60	0.00	0.00	0.00	0.60	29.30	0.18	2.06	0.38	16.55	112.91	0.99	9.49	1.66	34.86	80.10	8.96	35.57	12.47	**65.69**
ORB	3.40	438.30	0.42	0.78	1.00	0.90	103.20	0.42	0.86	1.50	34.41	195.95	4.85	15.16	9.88	34.29	258.04	5.41	12.72	9.59	**4.66**
SURF-PIIFD	0.40	22.40	0.12	2.46	0.38	0.80	16.90	0.24	4.74	0.50	106.18	175.05	9.46	53.66	13.04	59.07	93.85	10.53	47.31	14.54	36665.26
SURF-RIFT	7.20	42.50	1.85	16.45	4.21	7.20	34.60	2.18	21.19	4.75	227.59	256.09	19.80	86.79	28.95	**97.35**	116.68	14.40	**70.18**	20.90	1496.17
PC-FAST-RIFT	9.30	55.20	1.00	16.25	1.40	12.30	54.00	1.15	23.43	1.33	492.82	551.68	24.75	88.44	31.13	**208.96**	240.25	15.59	**73.14**	18.94	2805.57
PC-BRISK-RIFT	5.90	38.00	2.23	14.82	6.14	7.50	39.40	2.93	19.28	7.28	315.45	383.41	25.44	79.80	38.25	**110.76**	143.15	**16.12**	62.79	**28.13**	1968.58
TILDE-RIFT	3.30	22.00	2.90	14.14	11.82	3.10	20.80	2.43	14.80	16.35	88.09	136.86	16.04	59.17	39.80	31.92	55.49	11.49	41.28	25.65	2908.89
SuperPoint-RIFT	4.80	29.60	2.55	16.65	9.43	3.60	25.30	1.91	14.71	5.95	140.14	159.23	28.92	86.25	45.58	78.63	91.82	**22.73**	**71.21**	**32.12**	1226.33
LFnet-RIFT	7.7	50.8	1.91	15.24	3.43	11.8	57.7	2.81	20.62	5.01	84.32	115.82	16.86	69.77	35.54	56.58	83	14.02	62.02	23.24	1608.8
SuperPoint	0.70	42.80	0.36	1.54	1.30	1.40	58.20	0.80	2.65	2.76	72.18	173.86	15.76	36.71	25.09	69.75	134.83	**21.58**	39.81	**28.54**	689.07
LFnet	0.1	93	0.02	0.1	0.04	0.6	81.6	0.14	0.76	0.25	9.68	87.18	1.94	6.99	3.56	18.6	110.78	5.14	12.19	7.03	532.98

Table 12
Descriptor matching results about Correct Match Number, Putative Match Number, Match Score, Precision, Recall metrics of 10 feature matchers on multimodal datasets of remote sensing field (Part I).

Datasets	UAV Cross-season					Day–night					Depth-optical					Infrared-optical				
Match-ers\Metrics	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*
SIFT	1.25	16.75	0.17	16.07	0.44	31.29	60.29	2.33	33.33	4.70	12.83	36.33	1.32	33.65	2.08	26.00	55.00	2.96	16.51	4.32
SURF	84.75	395.75	2.37	20.93	4.82	0.86	49.29	0.04	9.52	0.69	49.00	152.50	2.09	33.32	3.19	8.00	532.20	0.32	2.67	0.58
ORB	2.00	808.50	0.10	0.25	0.16	71.57	660.86	4.63	9.80	8.67	76.33	955.83	3.83	8.31	4.73	15.40	894.80	0.78	1.75	1.25
SURF-RIFT	269.00	311.50	13.45	72.54	18.87	94.00	124.29	6.34	68.88	21.26	283.17	321.00	15.47	87.70	19.31	147.60	302.20	7.92	50.69	10.99
PC-FAST-RIFT	257.50	293.00	13.02	76.53	19.22	194.71	256.71	9.74	74.04	13.24	299.00	334.17	14.95	88.99	18.80	157.00	313.60	7.85	53.63	11.57
PC-BRISK-RIFT	95.75	123.75	12.33	66.68	23.75	136.29	187.43	9.61	70.58	17.47	169.83	212.00	16.09	79.17	22.03	92.00	232.60	8.11	44.49	14.74
TILDE-RIFT	116.75	166.00	14.66	63.38	24.93	94.00	134.57	14.62	65.99	30.38	159.33	201.00	26.07	78.16	38.22	121.60	273.20	12.54	49.64	21.34
SuperPoint-RIFT	118.75	134.00	17.89	76.90	25.76	170.00	212.14	12.05	74.73	17.30	286.33	319.67	19.77	88.80	24.12	98.80	218.00	10.44	48.90	17.84
SuperPoint	15.50	188.00	2.39	9.83	3.93	124.86	504.00	7.79	19.53	10.54	191.17	660.83	15.46	32.77	19.77	57.60	370.00	7.82	15.71	13.29
LFnet	1.00	155.50	0.20	0.62	0.46	3.43	80.00	0.69	3.39	1.25	31.17	243.33	6.23	12.88	9.80	10.40	212.40	2.08	5.12	3.85

Table 13
Descriptor matching results about Correct Match Number, Putative Match Number, Match Score, Precision, Recall metrics of 10 feature matchers on multimodal datasets of remote sensing field (part II).

Datasets	Map-optical					Optical cross-temporal					SAR-optical					All remote sensing data					
Match-ers\Metrics	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*RT* (ms)
SIFT	3.38	21.75	0.33	13.67	0.64	25.33	65.17	1.26	31.67	2.18	1.00	16.67	0.12	5.90	0.28	14.67	39.21	1.21	21.83	2.11	846.55
SURF	3.25	309.13	0.09	1.26	0.19	12.83	216.50	0.53	10.41	0.75	36.50	287.00	0.94	9.98	2.06	23.83	261.86	0.80	11.81	1.54	**349.72**
ORB	25.00	891.50	1.25	2.81	1.91	30.67	924.50	1.53	3.35	2.03	5.50	707.83	0.28	0.69	0.40	34.79	833.21	1.92	4.16	3.00	**9.53**
SURF-RIFT	210.25	258.63	10.51	80.22	14.31	126.33	188.00	6.36	48.25	7.45	266.33	304.50	14.66	84.17	17.05	**195.45**	251.83	10.50	**71.15**	15.63	4756.21
PC-FAST-RIFT	194.50	235.00	9.73	81.71	14.08	131.17	194.17	6.56	49.41	7.80	253.00	285.00	13.60	85.70	17.11	**210.31**	268.98	10.67	**73.59**	14.34	6168.90
PC-BRISK-RIFT	143.75	193.00	11.32	73.11	17.45	69.33	121.33	7.38	44.60	11.23	132.83	163.00	14.41	75.44	23.09	123.31	178.38	**11.31**	65.79	**18.30**	4826.76
TILDE-RIFT	135.63	192.25	14.96	70.79	22.56	78.67	133.00	11.11	46.18	14.68	127.33	169.67	16.03	69.32	26.20	119.29	179.33	**15.78**	64.10	**25.57**	7240.91
SuperPoint-RIFT	231.13	275.50	12.60	83.04	18.36	42.83	74.67	5.66	40.53	10.95	153.67	177.67	15.10	83.27	24.84	**164.40**	208.26	**13.14**	**71.79**	**19.52**	3860.86
SuperPoint	83.25	663.38	4.52	12.57	6.97	51.50	369.50	4.70	13.94	9.00	45.00	274.00	4.35	15.57	8.00	86.10	458.64	6.82	17.35	10.29	1398.05
LFnet	5.00	190.88	1.00	2.69	2.39	7.17	187.33	1.43	4.75	3.51	1.33	120.00	0.27	0.63	0.50	8.52	168.45	1.70	4.36	3.14	**548.35**

Table 14
Descriptor matching results about Correct Match Number, Putative Match Number, Match Score, Precision, Recall metrics of 11 feature matchers on multimodal datasets of computer vision field (Part I).

Datasets	Visible–infrared					Visible-near infrared					Cross-season				
Matchers\Metrics	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*
SIFT	1.50	36.70	0.42	12.35	1.29	388.33	470.33	16.24	73.08	23.69	32.33	123.00	2.10	27.40	7.47
SURF	0.10	28.50	0.02	0.17	0.02	17.97	402.23	0.42	4.68	1.22	7.83	359.00	1.26	10.61	2.02
ORB	1.60	164.40	0.36	0.77	1.38	379.60	1004.03	21.40	35.26	27.70	47.33	731.67	3.14	8.60	3.82
SURF-RIFT	48.90	79.50	8.64	60.82	16.91	459.17	503.17	23.28	84.18	30.22	86.67	166.00	6.58	48.68	10.94
PC-FAST-RIFT	116.30	168.70	6.67	66.08	8.03	501.53	552.03	25.08	83.44	32.63	91.83	176.67	5.67	52.12	8.73
PC-BRISK-RIFT	46.80	88.90	6.53	47.95	12.63	360.27	432.77	20.82	76.58	31.14	43.67	131.17	4.65	35.19	12.49
TILDE-RIFT	20.30	49.40	7.70	41.79	23.69	493.50	551.20	33.56	81.07	46.70	89.17	178.67	10.32	53.79	28.86
SuperPoint-RIFT	47.90	70.70	10.39	67.19	17.02	541.77	576.00	33.54	86.84	41.91	140.33	217.50	11.52	62.61	18.82
LFnet-RIFT	46.90	76.60	9.39	58.40	18.79	156.77	178.93	31.35	80.50	51.78	44.67	94.00	8.93	44.61	18.06
SuperPoint	16.00	154.30	4.25	11.81	8.04	734.37	957.10	44.19	71.44	57.82	114.33	523.00	11.54	23.72	20.32
LFnet	0.90	56.20	0.18	1.42	0.34	118.23	246.17	23.65	42.37	37.89	24.83	224.17	4.97	10.72	9.33

Table 15
Descriptor matching results about Correct Match Number, Putative Match Number, Match Score, Precision, Recall metrics of 11 feature matchers on multimodal datasets of computer vision field (Part II).

Datasets	Day–night					image–paint					All Vision Data					
Matchers\Metrics	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*CMN*	*PMN*	*MS (%)*	*Pre. (%)*	*Rec. (%)*	*RT* (ms)
SIFT	57.00	182.00	0.95	26.54	1.78	4.00	135.00	0.10	2.96	0.21	244.43	320.69	10.32	51.76	15.76	1407.17
SURF	120.50	503.50	3.24	41.45	11.88	4.00	770.00	0.05	0.52	0.33	16.98	332.31	0.55	5.90	1.49	**615.4**
ORB	3.00	919.00	0.15	0.31	0.36	3.00	991.00	0.15	0.30	0.29	238.71	795.59	13.57	22.82	17.73	**9.17**
SURF-RIFT	114.50	156.00	5.73	41.49	11.24	16.00	212.00	0.80	7.55	1.33	306.71	355.31	17.07	**71.76**	23.78	6646.36
PC-FAST-RIFT	193.00	242.50	9.65	44.06	14.33	23.00	191.00	1.15	12.04	2.66	**350.39**	407.84	17.83	**73.00**	23.33	9944.77
PC-BRISK-RIFT	145.50	203.50	7.28	41.57	11.66	15.00	188.00	0.75	7.98	1.73	241.71	311.31	14.96	62.84	23.68	9334.33
TILDE-RIFT	135.50	224.00	10.08	40.45	16.85	13.00	244.00	0.56	5.33	1.32	323	383.55	**23.81**	66.51	**37.67**	13539.31
SuperPoint-RIFT	240.00	265.50	12.04	46.78	17.74	24.00	203.00	1.21	11.82	1.96	**368.94**	408.69	**24.58**	**76.70**	32.2	6474.79
LFnet-RIFT	38.00	75.00	7.60	35.51	19.79	2.00	77.00	0.40	2.60	1.82	112.61	141.33	22.52	68.17	**38.59**	4360.62
SuperPoint	120.00	778.50	6.00	15.47	10.51	38.00	786.00	1.90	4.83	3.10	**472.55**	729.33	**29.62**	49.79	**40.02**	2298.28
LFnet	1.50	211.00	0.30	0.70	1.77	1.00	268.00	0.20	0.37	0.91	75.69	203.71	15.14	27.58	24.5	**556.83**

PSNR is the ratio between the maximum power of a signal and the power of noise that affects the fidelity of its representations. It is most commonly used to measure the quality of reconstruction errors in image evaluation. SSIM is often used to model image distortion and loss, measuring the structural similarity between the warped source image and the original fixed image. MI determines the degree of similarity between the image intensity distribution in two images. Higher values of these three metrics represent better registration results. We refer the readers to [6,13,477] for more details. In our evaluation, we first convert all images from RGB space (for single-channel images, we repeat this channel into three to simulate RGB) into YCbCr, then use the channel Y to calculate these three metrics.

Figs. 15 16 show the average results for the aforementioned six metrics for 12 state-of-the-art methods. As shown in Fig. 15, the advanced resampling-based method MAGSAC++ obtains the best registration accuracy in terms of distance-based metrics, followed by RANSAC, nonparametric, and graph-based methods, because of their global geometrical or strict local graph-consistent constraints. For the relaxed methods, LPM and GMS are limited in remote sensing, VIS–IR, and image–paint image pairs due to their loose constraints and the dominant outliers in putative match sets. mTopKRP is more robust because of its enhanced outlier filtering rules. The learning-based method LMR can also achieve satisfying registration results owing to its good mismatch removal performance. However, the poor result for correct match finding obviously leads to a large registration error when LFGC is performed.

Fig. 16 shows that the difference among these methods is not so remarkable because PSNR, SSIM, and MI are highly related to the overlapping area between two images in the image registration task, resulting in a large range among different datasets, particularly if the values changed slightly along with the registration accuracy. MAGSAC++ achieves the best results in SSIM and MI measurement. Overall, these three metrics are less discriminative and objective than the landmark-based metrics when measuring the registration accuracy of general multimodal image pairs. Note that, in our early test, area-based methods can only achieve promising registration performance in medical data (except retina image pairs), and they require significant time for an image pair with large size. Thus we ignore the area-based methods for our general test of MMIM in medical, remote sensing and computer vision researches. In addition, for some types of modalities, such as MRI–PET, SPECT–CT, Cross Season in remote sensing, even the feature matching results are not so satisfactory, we can still find that the false matches can keep weak consistence with inliers, allowing the transformation models to be accurately estimated thus obtaining good registration results as the RMSE MAE and MEE shown in Fig. 15.

5. Applications

In this part, we will briefly introduce several typical applications based on MMIM, including image fusion, change detection, image localization, target recognition, and tracking. This part would deliver more insightful understanding for the significance of MMIM or registration.

5.1. Image fusion

Image fusion is one of the most important applications of image registration technology [13,478]. Formally, image fusion extracts the most meaningful information from images acquired by different sensors or under different shooting settings, and combines the information to generate a single image, which contains more abundant information and is more conducive to subsequent applications [13]. However, different sensors or shooting settings may cause some degradation, such as different resolutions and motion blur, which will make the captured images not aligned. Therefore, in almost all current image fusion methods, the source images are already strictly registered by default, which is also a prerequisite for subsequent feature extraction, fusion, and reconstruction. According to the sensor or shooting settings, image fusion can be divided into multiple scenarios, including IR and visible image fusion, medical image fusion, pansharpening, multiexposure image fusion, and multifocus image fusion. We introduce these fusion scenarios and provide representative works.

The purpose of IR and visible image fusion is to preserve the significant contrast in IR images and the texture in visible images to generate a single image with significant contrast and rich texture details. Ma et al. [479] proposed an end-to-end model called FusionGAN, which generates a fused image with a dominant IR intensity and an additional visible gradient on the basis of GAN. This is further improved in [480] by preserving better details. Subsequently, they introduced a dual discriminator to exploit the structure information in the source images [481]. Zhang et al. [482] adopted the generative adversarial network to measure the effectiveness of information at the pixel level, so as to adaptively guiding the model to produce the full-focused image. Medical image fusion is dedicated to combining the body's metabolic function information and organizational structure information to generate a composite image that is more conducive to the diagnosis of lesions. Hou et al. [483] designed a CT and MRI fusion scheme based on CNNs and a dual-channel spiking cortical model, which can preserve the salient features and details of the source images. Pansharpening aims to fuse a low-resolution multispectral image and a high-resolution panchromatic image to produce a high-resolution multispectral image. Yang et al. [484] proposed the PanNet, in which the network is trained in the high-pass filtering domain rather than the image domain. They added upsampled multispectral images to the residuals learned by the network to generate the final results to strengthen the preservation of the spatial structure. Ma et al. [485] used GANs to realize pansharpening without supervision for the first time. Their model consists of a generator, a spectral discriminator, and a spatial discriminator, which work together to preserve spectral and spatial information.

The above three scenarios are all fusions of images captured by different sensors. For fusing images captured under different shooting settings, the two most typical scenarios are multiexposure image fusion and multifocus image fusion. Multiexposure image fusion involves fusing an overexposed image and an underexposed image to generate a properly illuminated image. Some methods can produce promising results. For example, a ghost-free multiexposure image fusion technique using the dense SIFT descriptor and guided filter was proposed by Hayat et al. [486], which can produce high-quality images without the ghosting artifact by using ordinary cameras. Prabhakar et al. [487] proposed an unsupervised deep learning framework that utilizes a no-reference quality metric as a loss function to assess the quality of exposure and can produce satisfactory fusion results. Multifocus image fusion as an image enhancement method can fuse images with different focused regions to obtain a single full-clear image. Innovatively, Guo et al. [488] proposed to use the conditional GAN for multifocus image fusion. Unlike these methods for a single fusion scene, some works can realize multiple fusion tasks uniformly. Zhang et al. [489] unified multiple image fusion tasks into the extraction and reconstruction of intensity information and gradient information, and designed a loss function in a unified form, which can produce results with good visual perception. Similarly, Xu et al. [490] used continuous learning technology to maintain the memory capacity of the network, thus realizing unified image fusion.

All the above methods perform feature extraction in alignment, followed by feature fusion and image reconstruction. Therefore, the spatial registration directly determines the fusion performance. In this regard, research on high-accuracy image registration technology is of great significance to the field of image fusion.

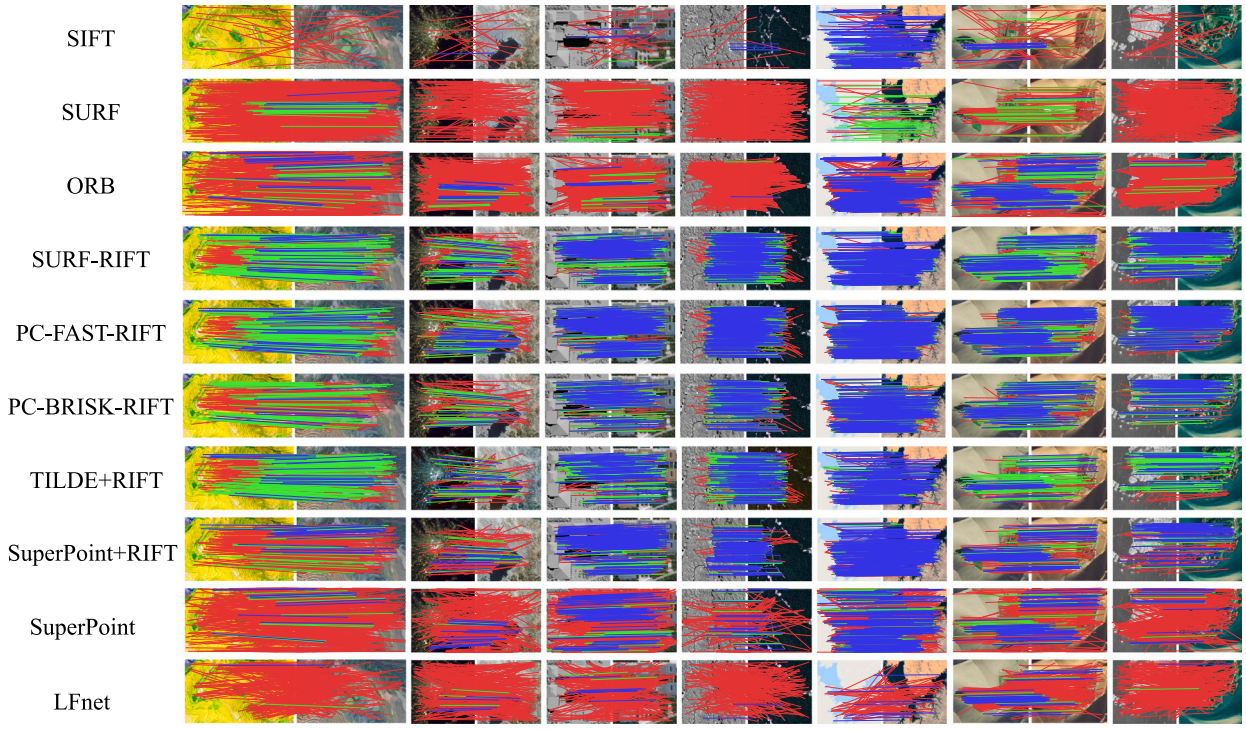

Fig. 9. Qualitative descriptor matching results of 10 methods on typical multimodal image pairs in the remote sensing research area. (blue = correct matches with threshold of 3 pixels, green = correct matches with threshold of 5 pixels, red = incorrect matches). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

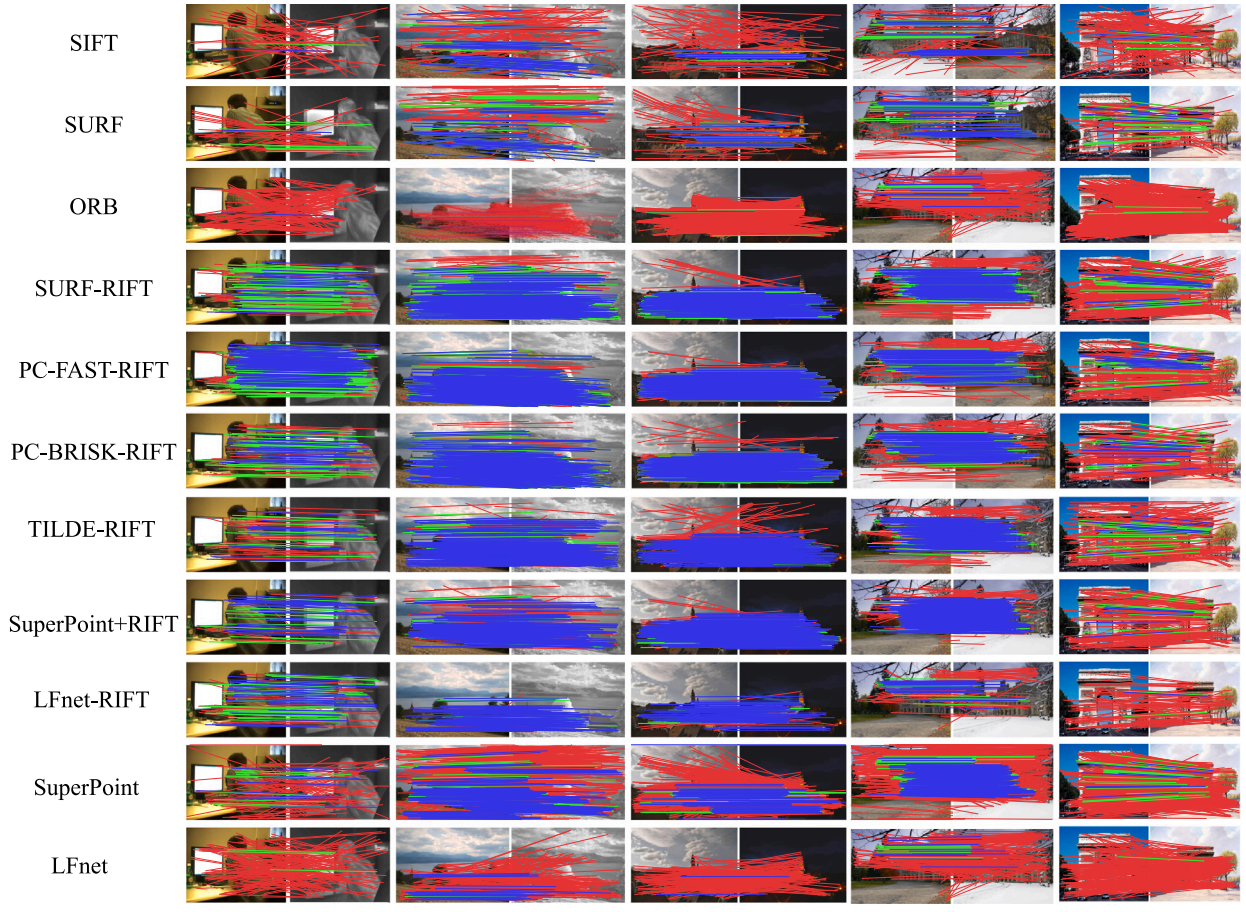

Fig. 10. Qualitative descriptor matching results of 11 methods on typical multimodal image pairs in the computer vision research area. (blue = correct matches with threshold of 3 pixels, green = correct matches with threshold of 5 pixels, red = incorrect matches). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

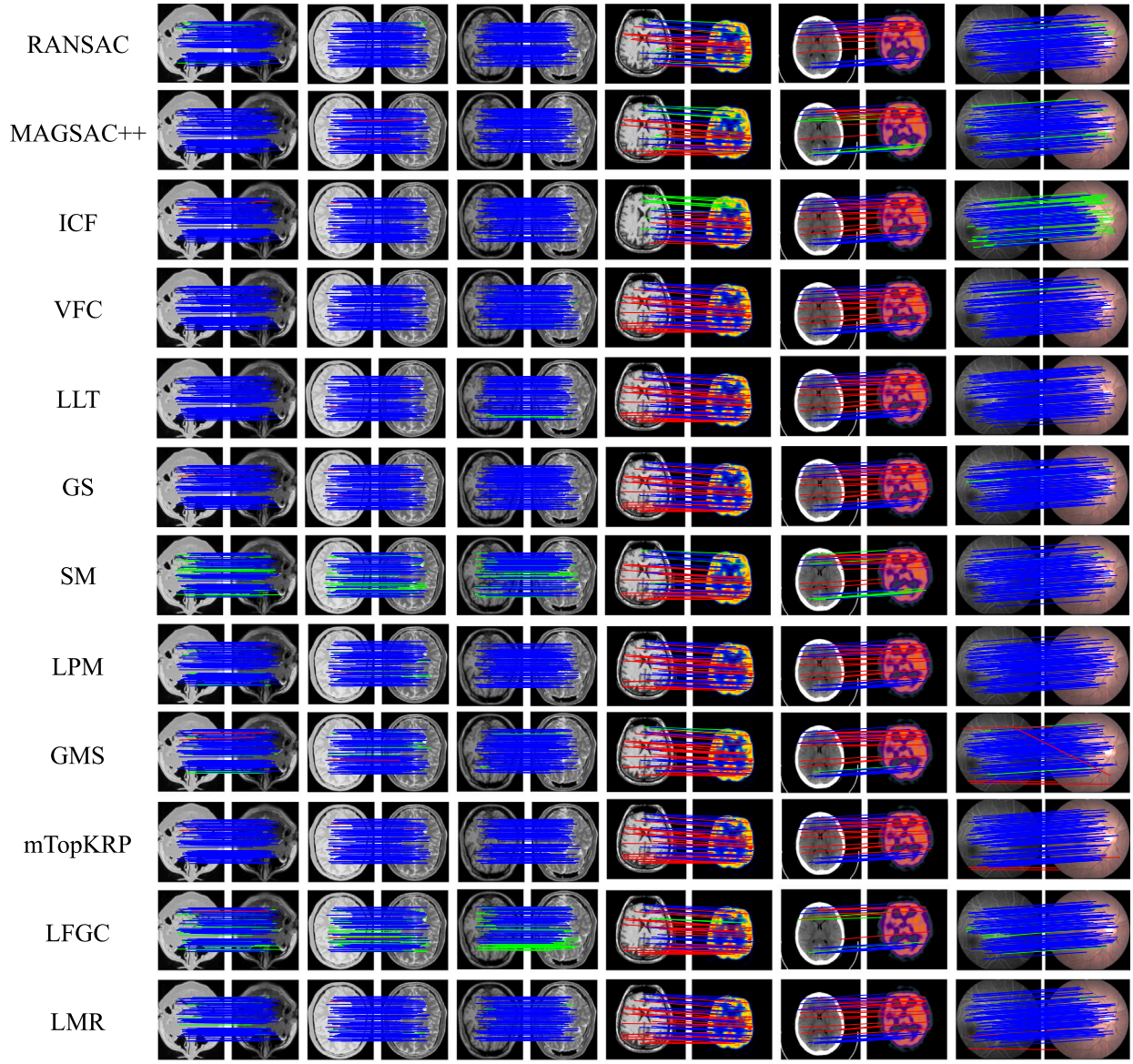

Fig. 11. Qualitative results of 12 mismatch removal algorithms on typical medical multimodal image pairs. For visibility, in each image pairs, at most 100 randomly selected matches are presented, and the true negatives are not shown. (blue = true positive, green = false negative, red = false positive). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.2. Change detection

Change detection, which refers to determining the discrepancy between image pairs of the same scene taken at different time, has aroused increasing interest in computer vision, medical, and remote sensing [491]. Generally, change detection is widely and successfully applied for environment monitoring [492], damage assessment [493], and so on.

In remote sensing applications over the past decade, an increasing number of change detection techniques have been proposed for environment monitoring under an all-around observation of the earth's surface. Existing methods can be roughly classified into two categories based on whether the used images are of the same modality: *i.e.*, single-sensor-based and multisensor-based change detection. Different shoot times (*e.g.*, day and night, cross-season) and different sensors (*e.g.,* optical and SAR) may induce differences in imaging, such as color and illumination variations and different resolutions. These variances may lead to great challenges in registering captured image pairs. Hence, high-accuracy registration of multitemporal or multisensor images is urgently required to avoid generating significantly spurious change detection results. In the following, we will deliver representative works in response to the two aforementioned.categories

The first type of change detection is typically performed on the image series of a single sensor. In accordance with the used unit for analysis, this type of change detection it can be further divided into pixel-based change detection (PBCD) and object-based change detection (OBCD). The former utilizes the pixel as the basic unit of image analysis whose spectral characteristics are used to detect and measure changes almost without taking the spatial context into consideration. In general, statistical operations are applied to gauge the individual pixel. Celik et al. [494] applied the PCA technique to map local neighborhoods in different images to a higher-dimensional space, which is performed on several defined non-overlapping image blocks. Quin et al. [495] proposed a change detection method called MIMOSA, which was especially designed for SAR time series. Wang et al. [496] introduced an unsupervised change detection approach for multitemporal SAR images by means of a triplet Markov field. In contrast to PBCD, OBCD methods aim to extract objects from source images by

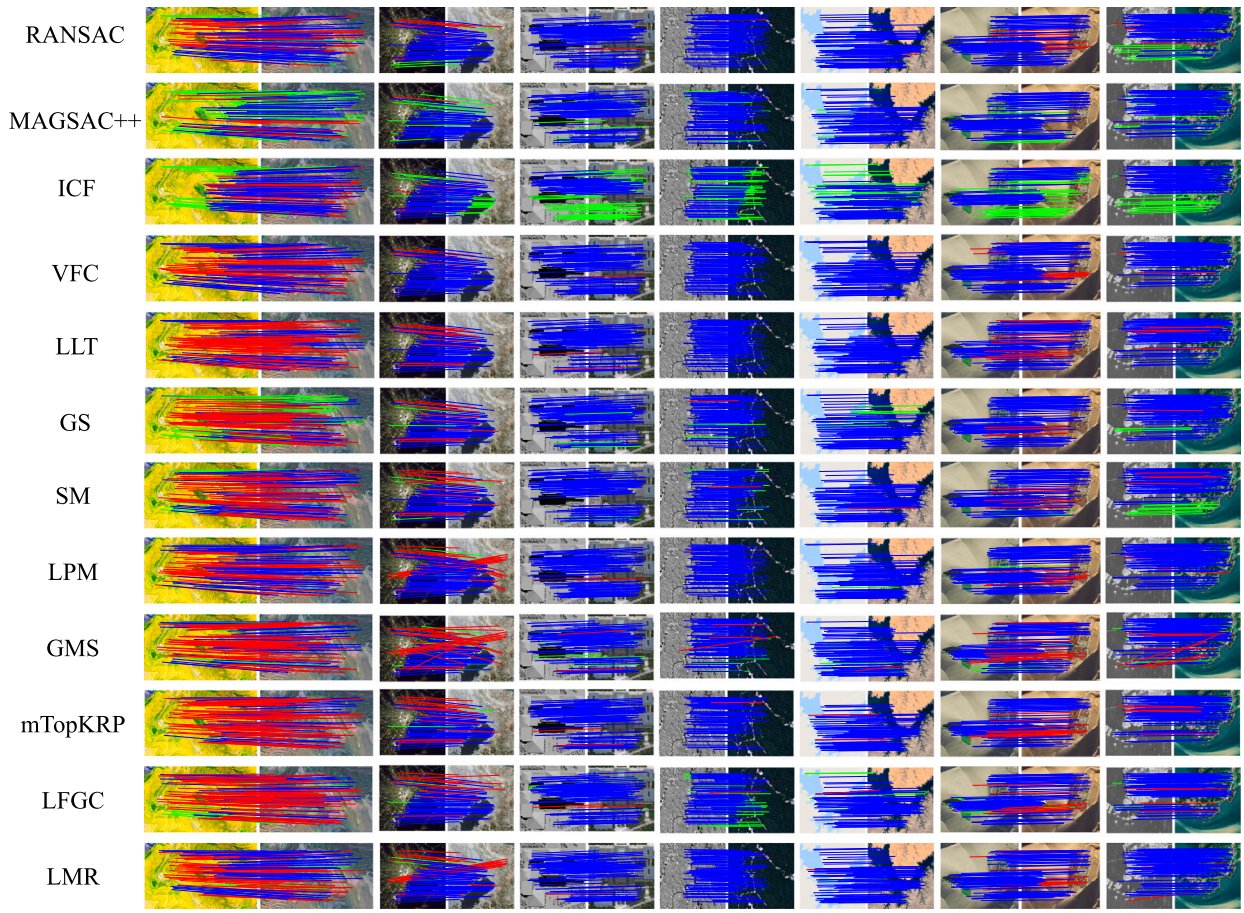

Fig. 12. Qualitative results of 12 mismatch removal algorithms on typical remote sensing multimodal image pairs. For visibility, in each image pair, at most 100 randomly selected matches are presented, and the true negatives are not shown. (blue = true positive, green = false negative, red = false positive). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

using image segmentation and other feature extraction algorithms, then determine the changes between corresponding objects. For example, Miller et al. [497] presented an OBCD algorithm to detect the changes of significant blobs (*i.e.*, objects) between a pair of gray-level images. Another method introduced in et al. [498] performed change detection by using correlation image analysis and image segmentation.

As a result of the significant breakthrough of the imaging sensors, images captured by different sensors have stimulated increasing research in related applications, including change detection. Existing approaches of this type can be roughly divided into three categories: *difference map-based*, *deep learning-based*, and *classification-based* approaches. The core idea of difference map-based methods is to produce difference maps under a thresholding strategy, then discover the changed regions on these maps. Alberga et al. [499] introduced a similarity measure that has been merely applied to image co-registration to compute the difference maps for multisensor remote sensing images. Mercier et al. [500] proposed a semi-supervised method based on the assumption that some dependence may commonly exist in these two images in unchanged areas. The dependence is modeled by quantile regression according to the copula theory. Meanwhile, the symmetrical Kullback–Leibler distance is used to acquire the change indices. Prendes et al. [501] presented another semi-supervised method. They modeled the objects contained in an analysis window by using mixtures of distributions. Subsequently in [502], they introduced a Bayesian nonparametric model based on their previous work, which successfully overcomes the drawbacks of demanding a prior knowledge of the number of objects in the analysis window. Deep learning-based methods aim to learn feature maps by using deep neural networks, thus guiding the change detection with these deep features for better performance. For instance, a multispatial-resolution change detection framework was proposed by Zhang et al. [503], which constructs a mapping neural network to exploit the inner relationships between multisensor images. Zhao et al. [504] introduced an approximately symmetric deep neural network to detect changes between multisensor images with two sides containing the same number of coupled layers. As for classification-based approaches, the common strategy is post-classification comparison, which is capable of minimizing the influence of different sensors. A more typical method would be [505], in which Mubea et al. applied maximum likelihood and support vector machine to classify remote sensing images. Land-use change monitoring was achieved by comparing the change between two corresponding years.

As a prerequisite procedure, image registration for multitemporal and/or multisensor images plays a significant role in image-based change detection. More emphasis on research into MMIR is necessary to the field of change detection.

5.3. Image localization

Image localization aims to recognize a known location by comparing the current place image captured by the visual sensor with previous images. Image localization mainly consists of the image processing module, the mapping framework, and the belief generation module [506]. With the changes in the environment (*e.g.*, season, weather, lighting), and the adjustment of the shooting angle, improving the accuracy of image localization is an essential but challenging issue [507]. Reliable registration is the cornerstone of alleviating environmental and seasonal changes, and for long-term localization [508].

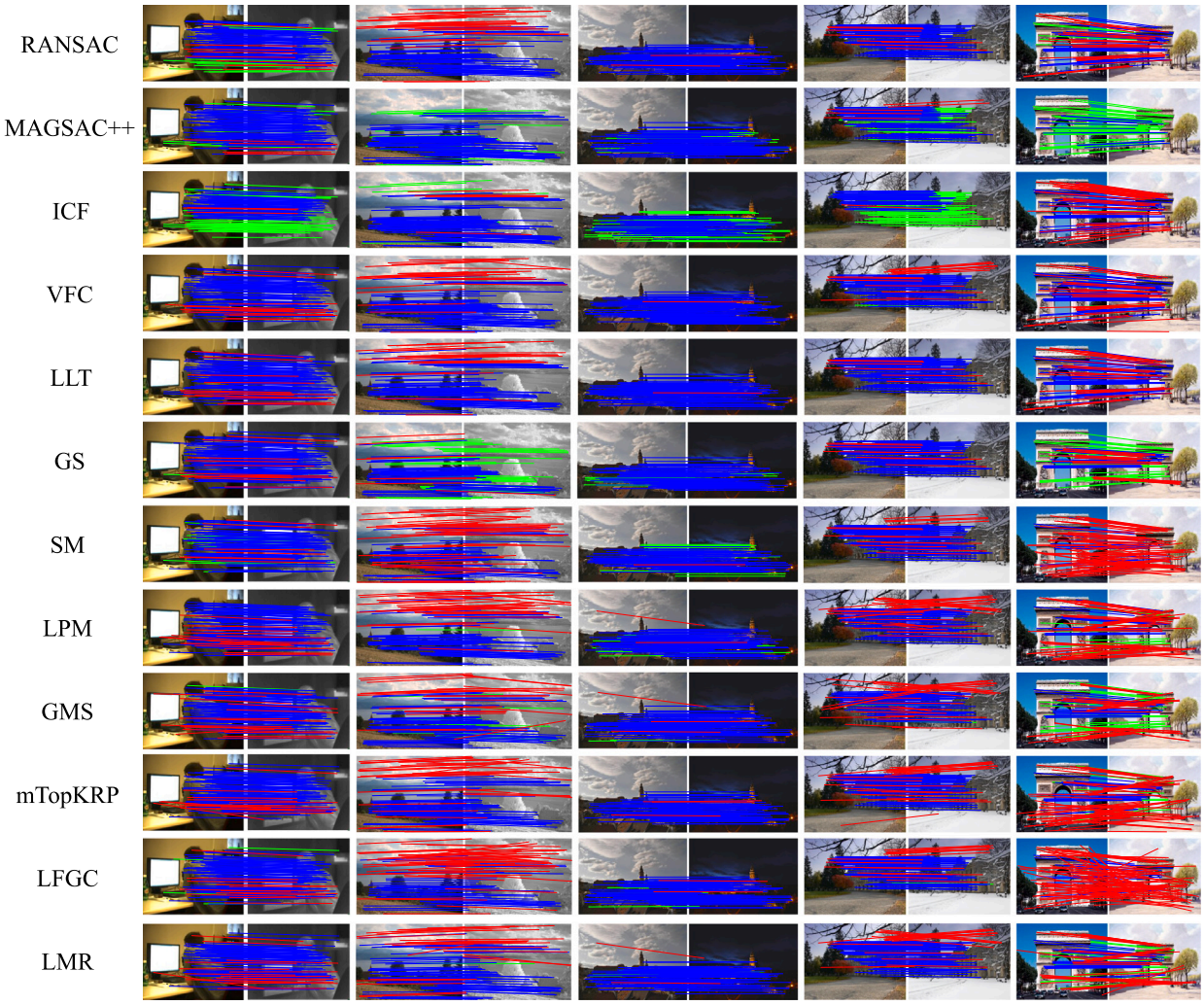

Fig. 13. Qualitative results of 12 mismatch removal algorithms on typical multimodal image pairs in the field of computer vision. For visibility, in each image pair, at most 100 randomly selected matches are presented, and the true negatives are not shown. (blue = true positive, green = false negative, red = false positive). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Next, we will provide a brief description of some of the representative image localization methods.

To enable robots to create maps of the environment while simultaneously using those maps to work out where they are (localize), Milford et al. [24] proposed a method called SeqSLAM, which is used to visualize navigation under changing conditions. They calculated the best candidate matching location within every local navigation sequence, and localization is achieved by recognizing coherent sequences of these local best matches, thus adapting to environmental changes. To reduce the effects of various seasonal differences and to handle revisits of places and loop closures, an approach to visual localization of mobile robots in outdoor environments was presented by Naseer et al. [7]. They formulated image matching as a minimum cost flow issue in a data association graph to efficiently use sequence information and deal with non-matching image sequences resulting from temporal occlusions or from visiting new places. On this basis, they introduced a semi-dense image description based on HOG features and global descriptors from deep CNNs for robust localization [509]. They also utilized image sequences to address the issue of visual localization under massive perceptual changes. Schönberger et al. [510] devised an approach based on a joint 3D geometric and semantic understanding of the world, which can be adapted to some extreme situations. They leveraged a generative model for descriptor learning and trained with semantic scene completion as an auxiliary task. In addition to outdoor scene localization, Taira et al. [511] presented InLoc for indoor visual localization. With three steps, namely, efficient retrieval of candidate poses, pose estimation using dense matching, and pose verification by virtual view synthesis, InLoc can alleviate indoor localization challenges such as lack of texture, significant changes in viewpoint, scene layout, and occluders. Liu et al. [512] proposed a method that can quickly and precisely pinpoint the location and perspective of image shots according to a prestored large-scale 3D point-cloud map. Their proposed method utilizes global contextual information exhibited both within the query image and among all the 3D points in the map, which takes account of not only visual similarities between individual 2D–3D matches but also their global compatibilities among all matching pairs.

In conclusion, image matching plays a pivotal role in different localization scenarios, and high-precision registration results commonly lead to reliable localization. Therefore, the study of excellent matching techniques is of great importance in the field of image localization.

5.4. Target recognition and tracking

The purpose of recognition is to interpret the scene or to distinguish different objects from an image, and the goal of tracking is to detect moving objects and to pursue the objectives of interest by estimating their motion parameters [513]. In addition, automatic target recognition (ATR) is a technique to identify objects or targets by comparing

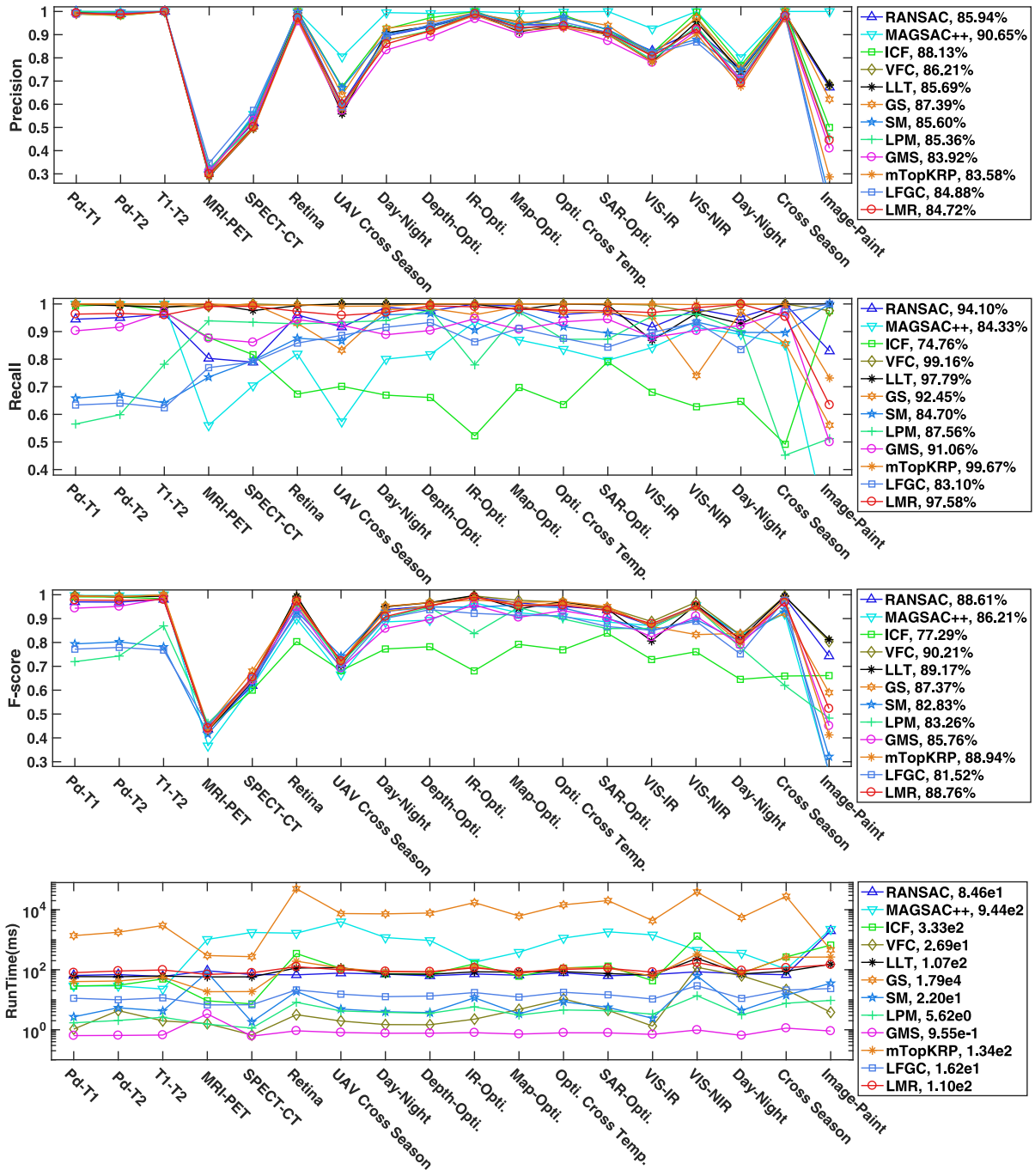

Fig. 14. Mismatch removal results about average *Precision, Recall, F-score* metrics, and *RunTime* of 12 representative methods on our constructed 18 multimodal image datasets.

images acquired in real time with data stored in the database [458]. Target recognition and tracking are commonly used in many practical scenarios, such as video surveillance and traffic control [514,515]. As for the ATR, it is typically used in disposable applications that can enhance missile guidance capabilities. In multisource image ATR tasks, different source images need to be registered given the apparent differences in the images captured by various sensors that prevent the target from being aligned, thus affecting the accuracy of target recognition [516]. In target recognition and tracking missions, image matching is a crucial step to compensate for background motion [517]. Subsequently, we introduce typical target recognition and tracking methods, and the application of image matching therein.

On the basis of the assumption that visible images of the target are available as a priori, Cheng et al. [458] proposed an algorithm for object recognition between IR and visible images under various conditions. Edge detection and binary template matching were exploited to initialize IR and visible images, followed by a local fuzzy threshold to recognize highly similar objects. Yoon et al. [518] presented a method for automatic airborne target recognition and tracking in forward-looking IR images in backgrounds. They employed an image segmentation and merging technique to detect reliable targets from complex environments, followed by training a Bayesian classifier to complete the classification using normalized inertial matrix features. Eventually, they combined joint detection and classification with the joint integrated highest probability to achieve multiobject tracking in clutter. To detect and identify moving targets from multiangle surveillance videos, Zhou et al. [519] investigated a multiview foreground matching model based on HOG detection and system clustering using

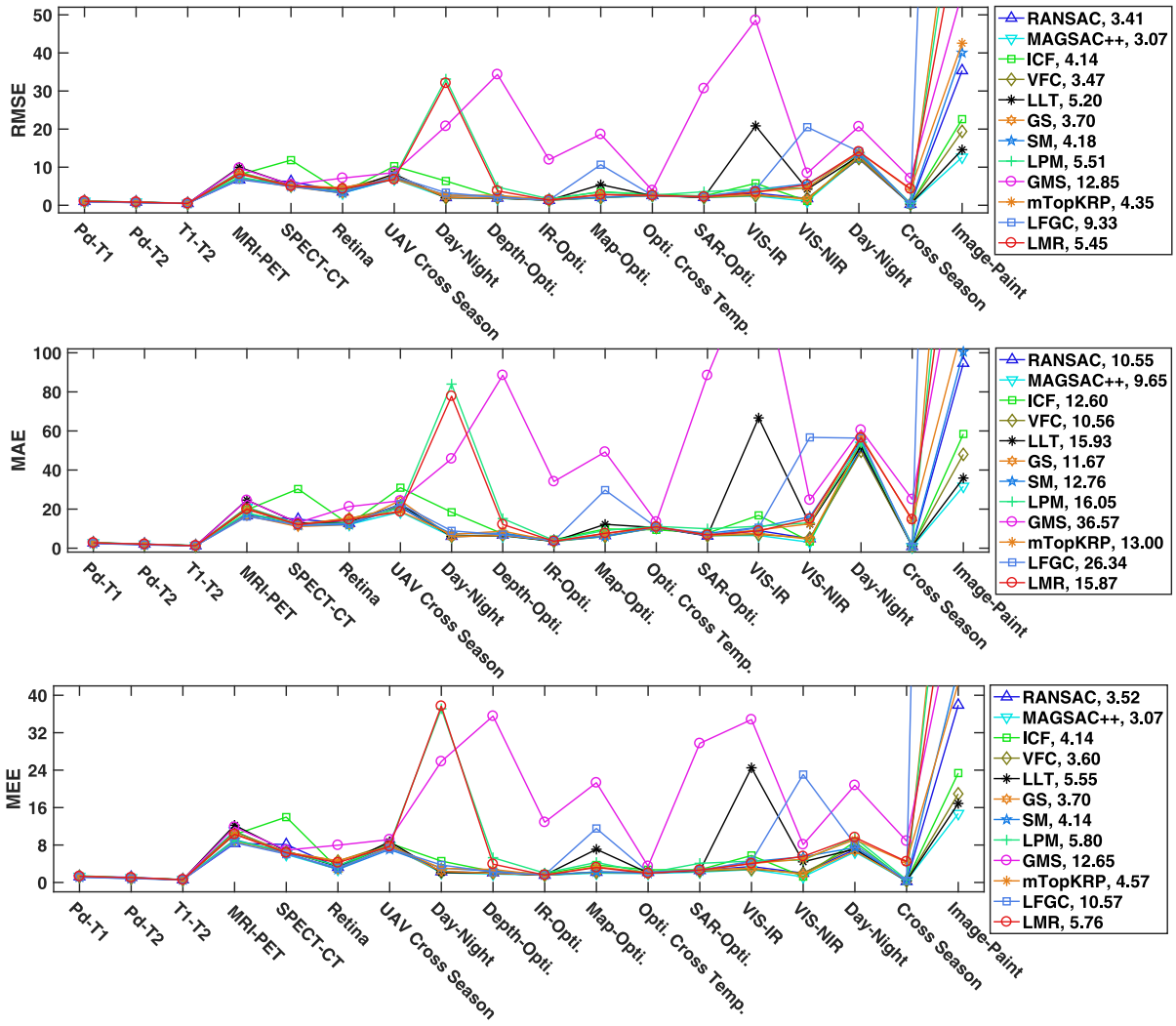

Fig. 15. Registration results for average RMSE, MSE, MEE metrics of 12 representative mismatch removal methods on our constructed 18 multimodal image datasets. A lower value is better.

background subtraction, HOG feature detection, and least squares of deviations in system clustering. Given the inlier points chosen by RANSAC, excessively concentrating on the image and on the feature points in the object regions may hinder accurate registration and affect background compensation. Xu et al. [520] devised a unique framework for moving object detection, in which feature point selection and registration accuracy prediction are utilized to boost detection accuracy. A practical approach for detecting and tracking multiple moving objects from a video sequence was proposed by Hu et al. [521]. In this method, the authors addressed the challenge of mixing camera motion and object movement in moving object detection and tracking. They used a Kalman filter based on the center of gravity of a moving object region in the minimum bounding box for tracking multiple moving objects.

The above-mentioned methods are just the tip of the iceberg in the field of target recognition and tracking. Devoting efforts in MMIM to achieve advanced performance of this application is significant for researchers.

6. Conclusion and feature trends

Image matching or registration for multimodal cases has played a critical role in various fields, including medical diagnosis, remote sensing (change detection, map updating, data fusion), and computer vision (image fusion, target recognition and tracking, image localization, or place recognition). Over the past decades, an increasing number and diversity of techniques have been proposed to improve performance. To provide a significant reference and understanding of MMIM for researchers and engineers in related areas, we provide a comprehensive and systematic review of the methods and their applications, which covers the typical cases in medical, remote sensing, and computer vision research. In addition, to perform experimental evaluations and provide a public standard for future research, we provide a complete database that includes 18 modality pairs (164 image pairs) together with their ground truths, which we collected and annotated. To better understand the significance of this problem, we introduce several typical applications related to MMIM.

In spite of the great progress that has been achieved in both theory and practice, MMIM remains an open problem with the following challenges for future developments:

- The insufficient and unavailable image data of different modalities are a significant limitation in general MMIM tasks.
- Area-based methods are largely limited by the small overlap and large deformation between two images. Computational consumption is another weakness due to their iterative optimization strategy, particularly for high-resolution images.
- Feature-based framework still greatly suffers from nonlinear intensity variance in feature detection and description, which would create few or even no accurate correspondences in some scenarios.

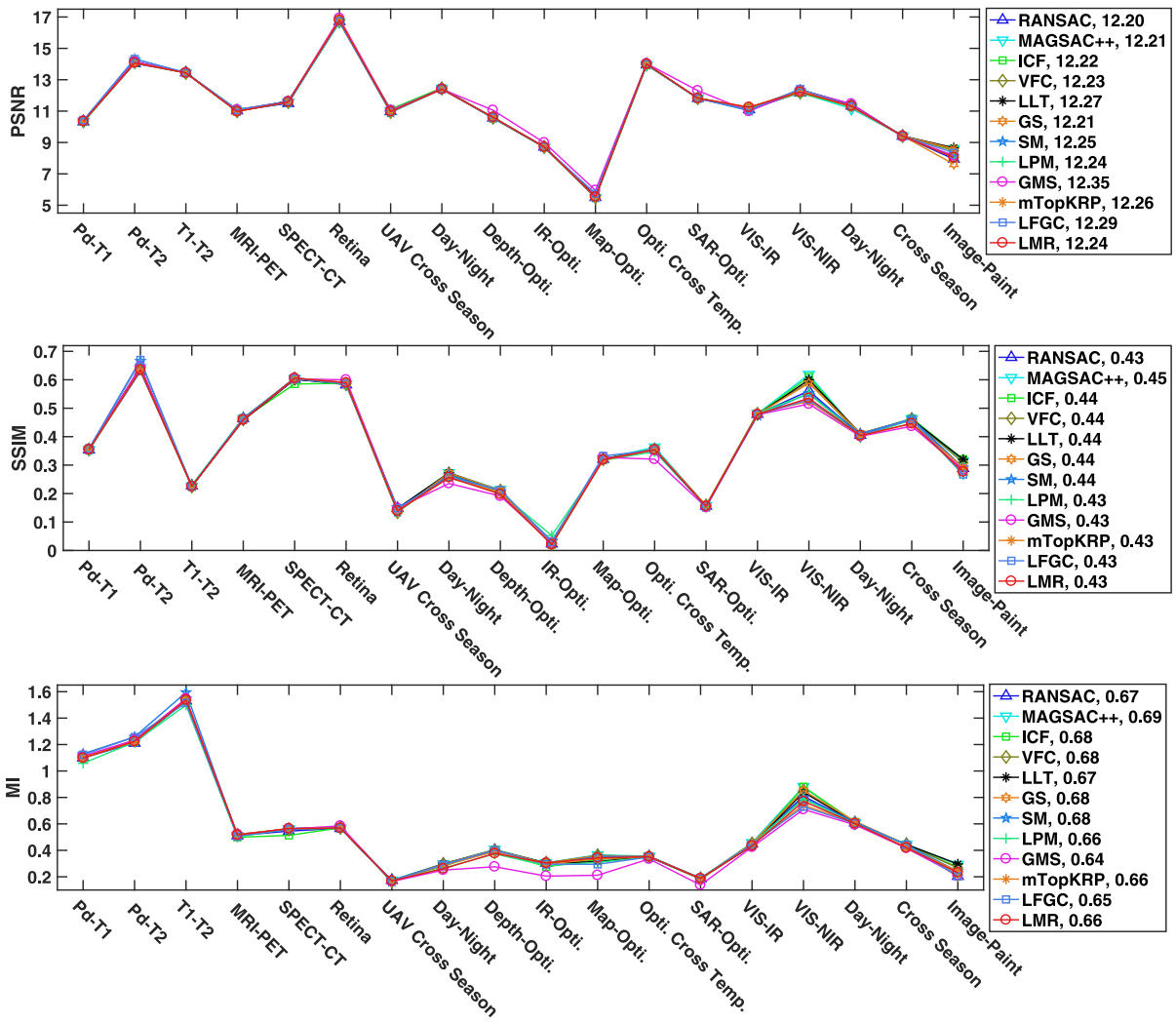

Fig. 16. Registration results for average PSNR, SSIM, MI metrics of 12 representative mismatch removal methods on our constructed 18 multimodal image datasets. A higher value is better.

- Learning from images for transformation parameter estimation is typically limited by complex geometrical deformations and image content.
- The use of only a local descriptor to construct putative match set would create a high number and ratio of mismatches. Therefore, an accurate, robust, and efficient mismatch removal method is required to achieve improved classification performance between inliers and outliers from a putative match set, thus maintaining accurate estimation of transformation parameters or deformation fields.
- A general matching method that can handle all types of multimodal images is still an open problem for both handcrafted and deep methods.
- Existing methods usually perform image registration and subsequent task in a successive manner. A worthwhile task is to attempt to integrate the matching problem into a high-level task for combinational optimization [522,523] or to directly perform the final task and bypass the matching or registration requirement, such as from unaligned images to image fusion.

CRediT authorship contribution statement

Xingyu Jiang: Conceived and designed thesurvey, Performed the experiments, Analyzed the results, Wrote the manuscript. **Jiayi Ma:** Conceived and designed the research, Provided insightful advices to this work, Revised the manuscript. **Guobao Xiao:** Provided insightful advices to this work, Revised the manuscript. **Zhenfeng Shao:** Analyzed the results, Provided insightful advices to this work, Revised the manuscript. **Xiaojie Guo:** Provided insightful advices to this work, Revised the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was sponsored in part by the National Natural Science Foundation of China (61773295), in part by the Key Research and Development Program of Hubei Province, China (2020BAB113), and in part by the Natural Science Foundation of Hubei Province, China (2019CFA037).

References

[1] B. Zitova, J. Flusser, Image registration methods: a survey, Image Vis. Comput. 21 (11) (2003) 977–1000.
[2] J. Ma, X. Jiang, A. Fan, J. Jiang, J. Yan, Image matching from handcrafted to deep features: A survey, Int. J. Comput. Vis. (2020) 1–57.

[3] H. Zhou, T. Sattler, D.W. Jacobs, Evaluating local features for day-night matching, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 724–736.

[4] A. Rana, G. Valenzise, F. Dufaux, Learning-based tone mapping operator for efficient image matching, IEEE Trans. Multimed. 21 (1) (2018) 256–268.

[5] Z. Luo, T. Shen, L. Zhou, J. Zhang, Y. Yao, S. Li, T. Fang, L. Quan, ContextDesc: Local descriptor augmentation with cross-modality context, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2527–2536.

[6] H. Zhou, J. Ma, C.C. Tan, Y. Zhang, H. Ling, Cross-weather image alignment via latent generative model with intensity consistency, IEEE Trans. Image Process. 29 (2020) 5216–5228.

[7] T. Naseer, L. Spinello, W. Burgard, C. Stachniss, Robust visual robot localization across seasons using network flows, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2014, pp. 2564–2570.

[8] A. Shrivastava, T. Malisiewicz, A. Gupta, A.A. Efros, Data-driven visual similarity for cross-domain image matching, in: Proceedings of the 2011 SIGGRAPH Asia Conference, 2011, pp. 1–10.

[9] M. Aubry, B.C. Russell, J. Sivic, Painting-to-3D model alignment via discriminative visual elements, ACM Trans. Graph. 33 (2) (2014) 1–14.

[10] X. Wei, T. Zhang, Y. Li, Y. Zhang, F. Wu, Multi-modality cross attention network for image and sentence matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 10941–10950.

[11] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W.Y. Wang, L. Zhang, Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6629–6638.

[12] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, Y. Zhang, Graph structured network for image-text matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 10921–10930.

[13] J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: A survey, Inf. Fusion 45 (2019) 153–178.

[14] A.P. James, B.V. Dasarathy, Medical image fusion: A survey of the state of the art, Inf. Fusion 19 (2014) 4–19.

[15] L. Wan, Y. Xiang, H. You, A post-classification comparison method for SAR and optical images change detection, IEEE Geosci. Remote Sens. Lett. 16 (7) (2019) 1026–1030.

[16] Y. Chen, J. Dai, X. Mao, Y. Liu, X. Jiang, Image registration between visible and infrared images for electrical equipment inspection robots based on quadrilateral features, in: Proceedings of the International Conference on Robotics and Automation Engineering, IEEE, 2017, pp. 126–130.

[17] Z. Wei, Y. Han, M. Li, K. Yang, Y. Yang, Y. Luo, S.-H. Ong, A small UAV based multi-temporal image registration for dynamic agricultural terrace monitoring, Remote Sens. 9 (9) (2017) 904.

[18] Y. Li, J. Ma, Y. Zhang, Image retrieval from remote sensing big data: A survey, Inf. Fusion 67 (2021) 94–115.

[19] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, J. Lai, Rgb-infrared cross-modality person re-identification, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5380–5389.

[20] Z. Feng, J. Lai, X. Xie, Learning modality-specific representations for visible-infrared person re-identification, IEEE Trans. Image Process. 29 (2019) 579–590.

[21] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, Z. Hou, Rgb-infrared cross-modality person reidentification via joint pixel and feature alignment, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3623–3632.

[22] S. Choi, S. Lee, Y. Kim, T. Kim, C. Kim, Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 10257–10266.

[23] Y. Ma, Y. Wang, X. Mei, C. Liu, X. Dai, F. Fan, J. Huang, Visible/infrared combined 3D reconstruction scheme based on nonrigid registration of multi-modality images with mixed features, IEEE Access 7 (2019) 19199–19211.

[24] M.J. Milford, G.F. Wyeth, Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights, in: Proceedings of the IEEE International Conference on Robotics and Automation, IEEE, 2012, pp. 1643–1649.

[25] D. Iwaszczuk, U. Stilla, Camera pose refinement by matching uncertain 3D building models with thermal infrared image sequences for high quality texture extraction, ISPRS J. Photogramm. Remote Sens. 132 (2017) 33–47.

[26] S. Oh, S. Kim, Deformable image registration in radiation therapy, Radiat. Oncol. J. 35 (2) (2017) 101.

[27] M.A. Schmidt, G.S. Payne, Radiotherapy planning using MRI, Phys. Med. Biol. 60 (22) (2015) R323.

[28] G. Haskins, U. Kruger, P. Yan, Deep learning in medical image registration: a survey, Mach. Vis. Appl. 31 (1) (2020) 8.

[29] A. Sotiras, C. Davatzikos, N. Paragios, Deformable medical image registration: A survey, IEEE Trans. Med. Imaging 32 (7) (2013) 1153–1190.

[30] Y. Fu, Y. Lei, T. Wang, W.J. Curran, T. Liu, X. Yang, Deep learning in medical image registration: a review, Phys. Med. Biol. (2020).

[31] P. Markelj, D. Tomaževič, B. Likar, F. Pernuš, A review of 3D/2d registration methods for image-guided interventions, Med. Image Anal. 16 (3) (2012) 642–661.

[32] E. Ferrante, N. Paragios, Slice-to-volume medical image registration: A survey, Med. Image Anal. 39 (2017) 101–123.

[33] Y. Guo, R. Sivaramakrishna, C.-C. Lu, J.S. Suri, S. Laxminarayan, Breast image registration techniques: a survey, Med. Biol. Eng. Comput. 44 (1–2) (2006) 15–26.

[34] J.H. Hipwell, V. Vavourakis, L. Han, T. Mertzanidou, B. Eiben, D.J. Hawkes, A review of biomechanically informed breast image registration, Phys. Med. Biol. 61 (2) (2016) R1.

[35] A. Gholipour, N. Kehtarnavaz, R. Briggs, M. Devous, K. Gopinath, Brain functional localization: a survey of image registration techniques, IEEE Trans. Med. Imaging 26 (4) (2007) 427–451.

[36] S. Matl, R. Brosig, M. Baust, N. Navab, S. Demirci, Vascular image registration techniques: A living review, Med. Image Anal. 35 (2017) 1–17.

[37] R. Shams, P. Sadeghi, R.A. Kennedy, R.I. Hartley, A survey of medical image registration on multicore and the GPU, IEEE Signal Process. Mag. 27 (2) (2010) 50–60.

[38] L. Shi, W. Liu, H. Zhang, Y. Xie, D. Wang, A survey of GPU-based medical image computing techniques, Quant. Imaging Med. Surg. 2 (3) (2012) 188.

[39] S. Dawn, V. Saxena, B. Sharma, Remote sensing image registration techniques: A survey, in: Proceedings of the International Conference on Image and Signal Processing, 2010, pp. 103–112.

[40] M. Gesto-Diaz, F. Tombari, D. Gonzalez-Aguilera, L. Lopez-Fernandez, P. Rodriguez-Gonzalvez, Feature matching evaluation for multimodal correspondence, ISPRS J. Photogramm. Remote Sens. 129 (2017) 179–188.

[41] A. Roche, G. Malandain, X. Pennec, N. Ayache, The correlation ratio as a new similarity measure for multimodal image registration, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 1998, pp. 1115–1124.

[42] J. Le Moigne, W.J. Campbell, R.F. Cromp, An automated parallel image registration technique based on the correlation of wavelet features, IEEE Trans. Geosci. Remote Sens. 40 (8) (2002) 1849–1864.

[43] B.B. Avants, C.L. Epstein, M. Grossman, J.C. Gee, Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain, Med. Image Anal. 12 (1) (2008) 26–41.

[44] F. Zhao, Q. Huang, W. Gao, Image matching by normalized cross-correlation, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, IEEE, 2006, p. II.

[45] J. Luo, E.E. Konofagou, A fast normalized cross-correlation calculation method for motion estimation, IEEE Trans. Ultrason. Ferroelectr. Freq. Control 57 (6) (2010) 1347–1357.

[46] W.M. Wells III, P. Viola, H. Atsumi, S. Nakajima, R. Kikinis, Multi-modal volume registration by maximization of mutual information, Med. Image Anal. 1 (1) (1996) 35–51.

[47] P. Viola, W.M. Wells III, Alignment by maximization of mutual information, Int. J. Comput. Vis. 24 (2) (1997) 137–154.

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

[49] M.B. Skouson, Q. Guo, Z.-P. Liang, A bound on mutual information for image registration, IEEE Trans. Med. Imaging 20 (8) (2001) 843–846.

[50] D. Loeckx, P. Slagmolen, F. Maes, D. Vandermeulen, P. Suetens, Nonrigid image registration using conditional mutual information, IEEE Trans. Med. Imaging 29 (1) (2009) 19–29.

[51] C. Studholme, C. Drapaca, B. Iordanova, V. Cardenas, Deformation-based mapping of volume change from serial brain MRI in the presence of local tissue contrast change, IEEE Trans. Med. Imaging 25 (5) (2006) 626–639.

[52] A.C. Chung, W.M. Wells, A. Norbash, W.E.L. Grimson, Multi-modal image registration by minimising kullback-leibler distance, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2002, pp. 525–532.

[53] C. Guetter, C. Xu, F. Sauer, J. Hornegger, Learning based non-rigid multi-modal image registration using Kullback-Leibler divergence, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2005, pp. 255–262.

[54] R. Liao, C. Guetter, C. Xu, Y. Sun, A. Khamene, F. Sauer, Learning-based 2d/3D rigid registration using jensen-Shannon divergence for image-guided surgery, in: International Workshop on Medical Imaging and Virtual Reality, Springer, 2006, pp. 228–235.

[55] J.-P. Thirion, Image matching as a diffusion process: an analogy with maxwell's demons, 1998.

[56] S. Henn, A full curvature based algorithm for image registration, J. Math. Imaging Vision 24 (2) (2006) 195–208.

[57] B. Glocker, N. Komodakis, N. Paragios, N. Navab, Approximated curvature penalty in non-rigid registration using pairwise mrfs, in: International Symposium on Visual Computing, Springer, 2009, pp. 1101–1109.

[58] G.E. Christensen, R.D. Rabbitt, M.I. Miller, Deformable templates using large deformation kinematics, IEEE Trans. Image Process. 5 (10) (1996) 1435–1447.

[59] A. Trouvé, Diffeomorphisms groups and pattern matching in image analysis, Int. J. Comput. Vis. 28 (3) (1998) 213–221.

[60] S.C. Joshi, M.I. Miller, Landmark matching via large deformation diffeomorphisms, IEEE Trans. Image Process. 9 (8) (2000) 1357–1370.

[61] S. Marsland, C.J. Twining, Constructing diffeomorphic representations for the groupwise analysis of nonrigid registrations of medical images, IEEE Trans. Med. Imaging 23 (8) (2004) 1006–1020.

[62] L. Zagorchev, A. Goshtasby, A comparative study of transformation functions for nonrigid image registration, IEEE Trans. Image Process. 15 (3) (2006) 529–538.

[63] F.L. Bookstein, Principal warps: Thin-plate splines and the decomposition of deformations, IEEE Trans. Pattern Anal. Mach. Intell. 11 (6) (1989) 567–585.

[64] F.L. Bookstein, Thin-plate splines and the atlas problem for biomedical images, in: Biennial International Conference on Information Processing in Medical Imaging, Springer, 1991, pp. 326–342.

[65] T.W. Sederberg, S.R. Parry, Free-form deformation of solid geometric models, in: Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques, 1986, pp. 151–160.

[66] J. Declerck, J. Feldmar, M.L. Goris, F. Betting, Automatic registration and alignment on a template of cardiac stress and rest reoriented SPECT images, IEEE Trans. Med. Imaging 16 (6) (1997) 727–737.

[67] D. Rueckert, L.I. Sonoda, C. Hayes, D.L. Hill, M.O. Leach, D.J. Hawkes, Nonrigid registration using free-form deformations: application to breast MR images, IEEE Trans. Med. Imaging 18 (8) (1999) 712–721.

[68] J. Kybic, M. Unser, Fast parametric elastic image registration, IEEE Trans. Image Process. 12 (11) (2003) 1427–1442.

[69] M. Sdika, A fast nonrigid image registration with constraints on the Jacobian using large scale constrained optimization, IEEE Trans. Med. Imaging 27 (2) (2008) 271–281.

[70] M.H. Davis, A. Khotanzad, D.P. Flamig, S.E. Harms, A physics-based coordinate transformation for 3-d image matching, IEEE Trans. Med. Imaging 16 (3) (1997) 317–328.

[71] G.E. Christensen, H.J. Johnson, Consistent image registration, IEEE Trans. Med. Imaging 20 (7) (2001) 568–582.

[72] J. Ashburner, K.J. Friston, Nonlinear spatial normalization using basis functions, Human Brain Mapp. 7 (4) (1999) 254–266.

[73] P. Hellier, C. Barillot, E. Mémin, P. Pérez, Hierarchical estimation of a dense deformation field for 3-d robust registration, IEEE Trans. Med. Imaging 20 (5) (2001) 388–402.

[74] V. Arsigny, X. Pennec, N. Ayache, Polyrigid and polyaffine transformations: a novel geometrical tool to deal with non-rigid deformations–application to the registration of histological slices, Med. Image Anal. 9 (6) (2005) 507–523.

[75] L. Younes, F. Arrate, M.I. Miller, Evolutions equations in computational anatomy, NeuroImage 45 (1) (2009) S40–S50.

[76] V.R.S. Mani, S. Arivazhagan, Survey of medical image registration, J. Biomed. Eng. Technol. 1 (2) (2013) 8–25.

[77] L.R. Ford Jr, D.R. Fulkerson, Flows in Networks, Princeton university press, 2015.

[78] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Elsevier, 2014.

[79] N. Komodakis, G. Tziritas, Approximate labeling via graph cuts based on linear programming, IEEE Trans. Pattern Anal. Mach. Intell. 29 (8) (2007) 1436–1453.

[80] N. Komodakis, G. Tziritas, N. Paragios, Performance vs computational efficiency for optimizing single and dynamic MRFs: Setting the state of the art with primal-dual strategies, Comput. Vis. Image Underst. 112 (1) (2008) 14–29.

[81] D. Shen, C. Davatzikos, HAMMER: hierarchical attribute matching mechanism for elastic registration, IEEE Trans. Med. Imaging 21 (11) (2002) 1421–1439.

[82] T. Liu, D. Shen, C. Davatzikos, Deformable registration of cortical structures via hybrid volumetric and surface warping, NeuroImage 22 (4) (2004) 1790–1801.

[83] M.P. Wachowiak, R. Smolíková, Y. Zheng, J.M. Zurada, A.S. Elmaghraby, An approach to multimodal biomedical image registration utilizing particle swarm optimization, IEEE Trans. Evol. Comput. 8 (3) (2004) 289–301.

[84] J. Santamaría, O. Cordón, S. Damas, A comparative study of state-of-the-art evolutionary image registration methods for 3D modeling, Comput. Vis. Image Underst. 115 (9) (2011) 1340–1354.

[85] G. Wu, M. Kim, Q. Wang, Y. Gao, S. Liao, D. Shen, Unsupervised deep feature learning for deformable registration of mr brain images, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2013, pp. 649–656.

[86] G. Wu, M. Kim, Q. Wang, B.C. Munsell, D. Shen, Scalable high-performance image registration framework by unsupervised deep feature representations learning, IEEE Trans. Biomed. Eng. 63 (7) (2015) 1505–1516.

[87] X. Cheng, L. Zhang, Y. Zheng, Deep similarity learning for multimodal medical images, Comput. Methods Biomech. Biomed. Eng. Imaging Vis. 6 (3) (2018) 248–252.

[88] G. Haskins, J. Kruecker, U. Kruger, S. Xu, P.A. Pinto, B.J. Wood, P. Yan, Learning deep similarity metric for 3D MR–TRUS image registration, Int. J. Comput. Assist. Radiol. Surg. 14 (3) (2019) 417–425.

[89] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, N. Navab, N. Komodakis, A deep metric for multimodal registration, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 10–18.

[90] M. Blendowski, M.P. Heinrich, Combining MRF-based deformable registration and deep binary 3D-cnn descriptors for large lung motion estimation in COPD patients, Int. J. Comput. Assist. Radiol. Surg. 14 (1) (2019) 43–52.

[91] R. Liao, S. Miao, P. de Tournemire, S. Grbic, A. Kamen, T. Mansi, D. Comaniciu, An artificial agent for robust image registration, 2016, arXiv preprint arXiv:1611.10336.

[92] K. Ma, J. Wang, V. Singh, B. Tamersoy, Y.-J. Chang, A. Wimmer, T. Chen, Multimodal image registration with deep context reinforcement learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 240–248.

[93] J. Krebs, T. Mansi, H. Delingette, L. Zhang, F.C. Ghesu, S. Miao, A.K. Maier, N. Ayache, R. Liao, A. Kamen, Robust non-rigid registration through agent-based action learning, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2017, pp. 344–352.

[94] S. Miao, S. Piat, P. Fischer, A. Tuysuzoglu, P. Mewes, T. Mansi, R. Liao, Dilated fcn for multi-agent 2d/3d medical image registration, 2017, arXiv preprint arXiv:1712.01651.

[95] X. Yang, R. Kwitt, M. Niethammer, Fast predictive image registration, in: Deep Learning and Data Labeling for Medical Applications, Springer, 2016, pp. 48–57.

[96] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, X. Pennec, Svf-net: Learning deformable image registration using shape matching, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 266–274.

[97] S.S.M. Salehi, S. Khan, D. Erdogmus, A. Gholipour, Real-time deep registration with geodesic loss, 2018, arXiv preprint arXiv:1803.05982.

[98] M. Ito, F. Ino, An automated method for generating training sets for deep learning based image registration, in: BIOIMAGING, 2018, pp. 140–147.

[99] J. Zheng, S. Miao, Z.J. Wang, R. Liao, Pairwise domain adaptation module for CNN-based 2-d/3-d registration, J. Med. Imaging 5 (2) (2018) 021204.

[100] H. Uzunova, M. Wilms, H. Handels, J. Ehrhardt, Training CNNs for image registration from few samples with model-based data augmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 223–231.

[101] X. Yang, Uncertainty quantification, image synthesis and deformation prediction for image registration, 2017.

[102] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M.K. Kalra, Y. Zhang, L. Sun, G. Wang, Low-dose CT image denoising using a generative adversarial network with wasserstein distance and perceptual loss, IEEE Trans. Med. Imaging 37 (6) (2018) 1348–1357.

[103] A. Hering, S. Kuckertz, S. Heldmann, M.P. Heinrich, Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking, in: Bildverarbeitung Für Die Medizin 2019, Springer, 2019, pp. 309–314.

[104] J. Fan, X. Cao, Q. Wang, P.-T. Yap, D. Shen, Adversarial learning for mono-or multi-modal registration, Med. Image Anal. 58 (2019) 101545.

[105] G. Haskins, U. Kruger, P. Yan, Deep learning in medical image registration: A survey, 2019, arXiv preprint arXiv:1903.02026.

[106] Y. Hu, E. Gibson, N. Ghavami, E. Bonmati, C.M. Moore, M. Emberton, T. Vercauteren, J.A. Noble, D.C. Barratt, Adversarial deformation regularization for training image registration neural networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 774–782.

[107] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[108] P. Yan, S. Xu, A.R. Rastinehad, B.J. Wood, Adversarial image registration with application for MR and TRUS image fusion, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2018, pp. 197–204.

[109] C. Qin, B. Shi, R. Liao, T. Mansi, D. Rueckert, A. Kamen, Unsupervised deformable registration for multi-modal images via disentangled representations, in: Proceedings of the International Conference on Information Processing in Medical Imaging, Springer, 2019, pp. 249–261.

[110] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in: Advances in Neural Information Processing Systems, 2015, pp. 2017–2025.

[111] X. Cao, J. Yang, L. Wang, Z. Xue, Q. Wang, D. Shen, Deep learning based inter-modality image registration supervised by intra-modality similarity, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2018, pp. 55–63.

[112] D. Mahapatra, B. Antony, S. Sedai, R. Garnavi, Deformable medical image registration using generative adversarial networks, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, 2018, pp. 1449–1453.

[113] L. Sun, S. Zhang, Deformable mri-ultrasound registration using 3d convolutional neural network, in: Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation, Springer, 2018, pp. 152–158.

[114] J. Zhang, Inverse-consistent deep networks for unsupervised deformable image registration, 2018, arXiv preprint arXiv:1809.03443.

[115] J. Fan, X. Cao, Z. Xue, P.-T. Yap, D. Shen, Adversarial similarity network for evaluating image alignment in deep learning based registration, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 739–746.

[116] I. Yoo, D.G. Hildebrand, W.F. Tobin, W.-C.A. Lee, W.-K. Jeong, Ssemnet: Serial-section electron microscopy image registration using a spatial transformer network with learned features, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2017, pp. 249–257.

[117] A. Kori, G. Krishnamurthi, Zero shot learning for multi-modal real time image registration, 2019, arXiv preprint arXiv:1908.06213.

[118] C.G. Harris, M. Stephens, et al., A combined corner and edge detector, in: Proceedings of the Alvey Vision Conference, 1988, pp. 147–151.

[119] S.M. Smith, J.M. Brady, SUSAN—A new approach to low level image processing, Int. J. Comput. Vis. 23 (1) (1997) 45–78.

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

[121] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[122] H. Bay, T. Tuytelaars, L. Van Gool, Surf: Speeded up robust features, in: Proceedings of the European Conference on Computer Vision, 2006, pp. 404–417.

[123] K.M. Yi, E. Trulls, V. Lepetit, P. Fua, Lift: Learned invariant feature transform, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 467–483.

[124] J. Canny, A computational approach to edge detection, in: Readings in Computer Vision, Elsevier, 1987, pp. 184–203.

[125] P. Perona, J. Malik, Scale-space and edge detection using anisotropic diffusion, IEEE Trans. Pattern Anal. Mach. Intell. 12 (7) (1990) 629–639.

[126] S. Xie, Z. Tu, Holistically-nested edge detection, in: ProceedingsoftheIEEEInternationalConferenceon ComputerVision, 2015, pp. 1395–1403.

[127] J. He, S. Zhang, M. Yang, Y. Shan, T. Huang, Bi-directional cascade network for perceptual edge detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3828–3837.

[128] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide-baseline stereo from maximally stable extremal regions, Image Vis. Comput. 22 (10) (2004) 761–767.

[129] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. Van Gool, A comparison of affine region detectors, Int. J. Comput. Vis. 65 (1–2) (2005) 43–72.

[130] R. Kimmel, C. Zhang, A. Bronstein, M. Bronstein, Are MSER features really interesting?, IEEE Trans. Pattern Anal. Mach. Intell. 33 (11) (2011) 2316–2320.

[131] S. Belongie, J. Malik, J. Puzicha, Shape context: A new descriptor for shape matching and object recognition, in: Advances in Neural Information Processing Systems, 2001, pp. 831–837.

[132] H. Goncalves, L. Corte-Real, J.A. Goncalves, Automatic image registration through image segmentation and SIFT, IEEE Trans. Geosci. Remote Sens. 49 (7) (2011) 2589–2600.

[133] J. Ma, J. Zhao, Y. Ma, J. Tian, Non-rigid visible and infrared face registration via regularized Gaussian fields criterion, Pattern Recognit. 48 (3) (2015) 772–784.

[134] Y. Ye, L. Shen, Hopc: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching, ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 3 (2016) 9.

[135] J. Fan, Y. Wu, M. Li, W. Liang, Y. Cao, Sar and optical image registration using nonlinear diffusion and phase congruency structural descriptor, IEEE Trans. Geosci. Remote Sens. 56 (9) (2018) 5368–5379.

[136] E. Rosten, R. Porter, T. Drummond, Faster and better: A machine learning approach to corner detection, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2010) 105–119.

[137] Z. Li, D. Mahapatra, J.A. Tielbeek, J. Stoker, L.J. van Vliet, F.M. Vos, Image registration based on autocorrelation of local structure, IEEE Trans. Med. Imaging 35 (1) (2015) 63–75.

[138] H.P. Moravec, Techniques towards automatic visual obstacle avoidance, 1977.

[139] J. Shi, C. Tomasi, Good Features to Track, Tech. Rep., Cornell University, 1993.

[140] J. Chen, J. Tian, N. Lee, J. Zheng, R.T. Smith, A.F. Laine, A partial intensity invariant feature descriptor for multimodal retinal image registration, IEEE Trans. Biomed. Eng. 57 (7) (2010) 1707–1718.

[141] Q. Du, A. Fan, Y. Ma, F. Fan, J. Huang, X. Mei, Infrared and visible image registration based on scale-invariant piifd feature and locality preserving matching, IEEE Access 6 (2018) 64107–64121.

[142] L. Huang, Z. Li, R. Zhang, Sar and optical images registration using shape context, in: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Ieee, 2010, pp. 1007–1010.

[143] Y. Ye, L. Bruzzone, J. Shan, F. Bovolo, Q. Zhu, Fast and robust matching for multimodal remote sensing image registration, IEEE Trans. Geosci. Remote Sens. 57 (11) (2019) 9059–9070.

[144] E. Rosten, T. Drummond, Machine learning for high-speed corner detection, in: Proceedings of the European Conference on Computer Vision, 2006, pp. 430–443.

[145] E. Mair, G.D. Hager, D. Burschka, M. Suppa, G. Hirzinger, Adaptive and generic corner detection based on the accelerated segment test, in: Proceedings of the European Conference on Computer Vision, 2010, pp. 183–196.

[146] J. Aldana-Iuit, D. Mishkin, O. Chum, J. Matas, In the saddle: chasing fast and repeatable features, in: Proceedings of the International Conference on Pattern Recognition, 2016, pp. 675–680.

[147] X. Zhang, Q. Hu, M. Ai, X. Ren, A multitemporal uav images registration approach using phase congruency, in: Proceedings of the IEEE International Conference on Geoinformatics, 2018, pp. 1–6.

[148] J. Li, Q. Hu, M. Ai, Rift: Multi-modal image matching based on radiation-variation insensitive feature transform, IEEE Trans. Image Process. 29 (2019) 3296–3310.

[149] B. Zhao, T. Xu, Y. Chen, T. Li, X. Sun, Automatic and robust infrared-visible image sequence registration via spatio-temporal association, Sensors 19 (5) (2019) 997.

[150] M. Awrangjeb, G. Lu, C.S. Fraser, Performance comparisons of contour-based corner detectors, IEEE Trans. Image Process. 21 (9) (2012) 4167–4179.

[151] F. Mokhtarian, R. Suomela, Robust image corner detection through curvature scale space, IEEE Trans. Pattern Anal. Mach. Intell. 20 (12) (1998) 1376–1381.

[152] A.M.G. Pinheiro, M. Ghanbari, Piecewise approximation of contours through scale-space selection of dominant points, IEEE Trans. Image Process. 19 (6) (2010) 1442–1450.

[153] U. Ramer, An iterative procedure for the polygonal approximation of plane curves, Comput. Graph. Image Process. 1 (3) (1972) 244–256.

[154] M. Awrangjeb, G. Lu, Robust image corner detection based on the chord-to-point distance accumulation technique, IEEE Trans. Multimed. 10 (6) (2008) 1059–1072.

[155] A. Rosenfeld, J.S. Weszka, An improved method of angle detection on digital curves, IEEE Trans. Comput. 100 (9) (1975) 940–941.

[156] A. Masood, M. Sarfraz, Corner detection by sliding rectangles along planar curves, Comput. Graph. 31 (3) (2007) 440–448.

[157] X. Zhang, H. Wang, A.W. Smith, X. Ling, B.C. Lovell, D. Yang, Corner detection based on gradient correlation matrices of planar curves, Pattern Recognit. 43 (4) (2010) 1207–1223.

[158] X. Zhang, Y. Qu, D. Yang, H. Wang, J. Kymer, Laplacian scale-space behavior of planar curve corners, IEEE Trans. Pattern Anal. Mach. Intell. 37 (11) (2015) 2207–2217.

[159] T. Lindeberg, Feature detection with automatic scale selection, Int. J. Comput. Vis. 30 (2) (1998) 79–116.

[160] D.G. Lowe, et al., Object recognition from local scale-invariant features. in: Proceedings of the IEEE International Conference on Computer Vision, 1999, pp. 1150–1157.

[161] J.-M. Morel, G. Yu, ASIFT: A new framework for fully affine invariant image comparison, SIAM J. Imaging Sci. 2 (2) (2009) 438–469.

[162] M. Agrawal, K. Konolige, M.R. Blas, Censure: Center surround extremas for realtime feature detection and matching, in: Proceedings of the European Conference on Computer Vision, 2008, pp. 102–115.

[163] P. Mainali, G. Lafruit, Q. Yang, B. Geelen, L. Van Gool, R. Lauwereins, Sifer: scale-invariant feature detector with error resilience, Int. J. Comput. Vis. 104 (2) (2013) 172–197.

[164] H. Deng, W. Zhang, E. Mortensen, T. Dietterich, L. Shapiro, Principal curvature-based region detector for object recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[165] L. Ferraz, X. Binefa, A sparse curvature-based detector of affine invariant blobs, Comput. Vis. Image Underst. 116 (4) (2012) 524–537.

[166] P.-E. Forssén, Maximally stable colour regions for recognition and matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1-8.

[167] T. Tuytelaars, L. Van Gool, Matching widely separated views based on affine invariant regions, Int. J. Comput. Vis. 59 (1) (2004) 61–85.

[168] A. Mustafa, H. Kim, A. Hilton, Msfd: Multi-scale segmentation-based feature detection for wide-baseline scene reconstruction, IEEE Trans. Image Process. 28 (3) (2018) 1118–1132.

[169] Y. Li, S. Wang, Q. Tian, X. Ding, A survey of recent advances in visual feature detection, Neurocomputing 149 (2015) 736–751.

[170] Z. Ghassabi, J. Shanbehzadeh, A. Sedaghat, E. Fatemizadeh, An efficient approach for robust multimodal retinal image registration based on UR-SIFT features and PIIFD descriptors, EURASIP J. Image Video Process. 2013 (1) (2013) 25.

[171] E.P. Ong, J.A. Lee, J. Cheng, G. Xu, B.H. Lee, A. Laude, S. Teoh, T.H. Lim, D.W. Wong, J. Liu, A robust outlier elimination approach for multimodal retina image registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 329–337.

[172] L. Yu, D. Zhang, E.-J. Holden, A fast and fully automatic registration approach based on point features for multi-source remote-sensing images, Comput. Geosci. 34 (7) (2008) 838–848.

[173] A. Sedaghat, M. Mokhtarzade, H. Ebadi, Uniform robust scale-invariant feature matching for optical remote sensing images, IEEE Trans. Geosci. Remote Sens. 49 (11) (2011) 4516–4527.

[174] H. Sun, L. Lei, H. Zou, C. Wang, Multimodal remote sensing image registration using multiscale self-similarities, in: Proceedings of the IEEE International Conference on Computer Vision in Remote Sensing, 2012, pp. 199–202.

[175] B. Fan, C. Huo, C. Pan, Q. Kong, Registration of optical and SAR satellite images by exploring the spatial relationship of the improved SIFT, IEEE Geosci. Remote Sens. Lett. 10 (4) (2012) 657–661.
[176] Y. Ye, J. Shan, A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences, ISPRS J. Photogramm. Remote Sens. 90 (2014) 83–95.
[177] C. Xu, H. Sui, H. Li, J. Liu, An automatic optical and SAR image registration method with iterative level set segmentation and SIFT, Int. J. Remote Sens. 36 (15) (2015) 3997–4017.
[178] Q. Zeng, J. Adu, J. Liu, J. Yang, Y. Xu, M. Gong, Real-time adaptive visible and infrared image registration based on morphological gradient and C_SIFT, J. Real-Time Image Process. (2019) 1–13.
[179] M. Trajković, M. Hedley, Fast corner detection, Image Vis. Comput. 16 (2) (1998) 75–87.
[180] C. Strecha, A. Lindner, K. Ali, P. Fua, Training for task specific keypoint detection, in: Joint Pattern Recognition Symposium, Springer, 2009, pp. 151–160.
[181] W. Hartmann, M. Havlena, K. Schindler, Predicting matchability, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 9–16.
[182] Y. Verdie, K. Yi, P. Fua, V. Lepetit, Tilde: a temporally invariant learned detector, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5279–5288.
[183] X. Zhang, F.X. Yu, S. Karaman, S.-F. Chang, Learning discriminative and transformation covariant local feature detectors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6818–6826.
[184] L. Zhang, S. Rusinkiewicz, Learning to detect features in texture images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6325–6333.
[185] D. DeTone, T. Malisiewicz, A. Rabinovich, Superpoint: Self-supervised interest point detection and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 224–236.
[186] K. Lenc, A. Vedaldi, Learning covariant feature detectors, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 100–117.
[187] N. Savinov, A. Seki, L. Ladicky, T. Sattler, M. Pollefeys, Quad-networks: unsupervised learning to rank for interest point detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1822–1830.
[188] Y. Ono, E. Trulls, P. Fua, K.M. Yi, Lf-net: learning local features from images, in: Advances in Neural Information Processing Systems, 2018, pp. 6234–6244.
[189] G. Georgakis, S. Karanam, Z. Wu, J. Ernst, J. Kosecká, End-to-end learning of keypoint detector and descriptor for pose invariant 3d matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1965–1973.
[190] A.B. Laguna, E. Riba, D. Ponsa, K. Mikolajczyk, Key. Net: Keypoint detection by handcrafted and learned cnn filters, 2019, arXiv preprint arXiv:1904.00889.
[191] X. Shen, C. Wang, X. Li, Z. Yu, J. Li, C. Wen, M. Cheng, Z. He, Rf-net: An end-to-end image matching network based on receptive field, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8132–8140.
[192] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, T. Sattler, D2-net: A trainable cnn for joint description and detection of local features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8092–8101.
[193] J. Revaud, P. Weinzaepfel, C. De Souza, N. Pion, G. Csurka, Y. Cabon, M. Humenberger, R2d2: Repeatable and reliable detector and descriptor, 2019, arXiv preprint arXiv:1906.06195.
[194] K. Lenc, A. Vedaldi, Large scale evaluation of local image feature detectors on homography datasets, 2018, arXiv preprint arXiv:1807.07939.
[195] J. Komorowski, K. Czarnota, T. Trzcinski, L. Dabala, S. Lynen, Interest point detectors stability evaluation on apolloscape dataset, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 1–13.
[196] F. Ye, Y. Su, H. Xiao, X. Zhao, W. Min, Remote sensing image registration using convolutional neural network features, IEEE Geosci. Remote Sens. Lett. 15 (2) (2018) 232–236.
[197] W. Ma, J. Zhang, Y. Wu, L. Jiao, H. Zhu, W. Zhao, A novel two-step registration method for remote sensing images based on deep and local features, IEEE Trans. Geosci. Remote Sens. 57 (7) (2019) 4834–4843.
[198] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.
[199] A.E. Abdel-Hakim, A.A. Farag, Csift: A sift descriptor with color invariant characteristics, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 1978–1983.
[200] R. Arandjelović, A. Zisserman, Three things everyone should know to improve object retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2911–2918.
[201] J. Dong, S. Soatto, Domain-size pooling in local descriptors: Dsp-sift, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5097–5106.
[202] F. Tang, S.H. Lim, N.L. Chang, H. Tao, A novel feature descriptor invariant to complex brightness changes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2631–2638.
[203] M. Toews, W. Wells, Sift-rank: Ordinal description for invariant feature correspondence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 172–177.
[204] B. Fan, F. Wu, Z. Hu, Rotationally invariant descriptors using intensity order pooling, IEEE Trans. Pattern Anal. Mach. Intell. 34 (10) (2011) 2031–2045.
[205] J. Chen, J. Tian, Real-time multi-modal rigid registration based on a novel symmetric-SIFT descriptor, Prog. Nat. Sci. 19 (5) (2009) 643–651.
[206] M.T. Hossain, G. Lv, S.W. Teng, G. Lu, M. Lackmann, Improved symmetric-sift for multi-modal image registration, in: 2011 International Conference on Digital Image Computing: Techniques and Applications, IEEE, 2011, pp. 197–202.
[207] A. Sedaghat, H. Ebadi, Distinctive order based self-similarity descriptor for multi-sensor remote sensing image matching, ISPRS J. Photogramm. Remote Sens. 108 (2015) 62–71.
[208] M. Calonder, V. Lepetit, C. Strecha, P. Fua, Brief: Binary robust independent elementary features, in: Proceedings of the European Conference on Computer Vision, 2010, pp. 778–792.
[209] S. Leutenegger, M. Chli, R. Siegwart, Brisk: Binary robust invariant scalable keypoints, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 2548–2555.
[210] A. Alahi, R. Ortiz, P. Vandergheynst, Freak: Fast retina keypoint, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 510–517.
[211] Y. Ke, R. Sukthankar, et al., Pca-sift: A more distinctive representation for local image descriptors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2004, pp. 506–513.
[212] H. Cai, K. Mikolajczyk, J. Matas, Learning linear discriminant projections for dimensionality reduction of image descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2010) 338–352.
[213] M. Brown, G. Hua, S. Winder, Discriminative learning of local image descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 33 (1) (2010) 43–57.
[214] B. Kulis, K. Grauman, Kernelized locality-sensitive hashing for scalable image search, in: Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 2130–2137.
[215] Y. Weiss, A. Torralba, R. Fergus, Spectral hashing, in: Advances in Neural Information Processing Systems, 2009, pp. 1753–1760.
[216] Y. Gong, S. Lazebnik, A. Gordo, F. Perronnin, Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2012) 2916–2929.
[217] T. Trzcinski, V. Lepetit, Efficient discriminative projections for compact binary descriptors, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 228–242.
[218] T. Trzcinski, M. Christoudias, P. Fua, V. Lepetit, Boosting binary keypoint descriptors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2874–2881.
[219] J.L. Schonberger, H. Hardmeier, T. Sattler, M. Pollefeys, Comparative evaluation of hand-crafted and learned local features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1482–1491.
[220] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (Feb) (2009) 207–244.
[221] S. Zagoruyko, N. Komodakis, Learning to compare image patches via convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4353–4361.
[222] X. Han, T. Leung, Y. Jia, R. Sukthankar, A.C. Berg, Matchnet: Unifying feature and metric learning for patch-based matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3279–3286.
[223] J. Wang, F. Zhou, S. Wen, X. Liu, Y. Lin, Deep metric learning with angular loss, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2593–2601.
[224] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, F. Moreno-Noguer, Discriminative learning of deep convolutional feature point descriptors, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 118–126.
[225] V. Balntas, E. Riba, D. Ponsa, K. Mikolajczyk, Learning local feature descriptors with triplets and shallow convolutional neural networks, in: Proceedings of the British Machine Vision Conference, 2016, pp. 1–11.
[226] V. Balntas, E. Johns, L. Tang, K. Mikolajczyk, PN-Net: Conjoined triple deep network for learning local image descriptors, 2016, arXiv preprint arXiv:1601.05030.
[227] A. Mishchuk, D. Mishkin, F. Radenovic, J. Matas, Working hard to know your neighbor's margins: Local descriptor learning loss, in: Advances in Neural Information Processing Systems, 2017, pp. 4826–4837.
[228] X. Wei, Y. Zhang, Y. Gong, N. Zheng, Kernelized subspace pooling for deep local descriptors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1867–1875.
[229] Y. Tian, X. Yu, B. Fan, F. Wu, H. Heijnen, V. Balntas, Sosnet: Second order similarity regularization for local descriptor learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 11016–11025.

[230] S. Wang, D. Quan, X. Liang, M. Ning, Y. Guo, L. Jiao, A deep learning framework for remote sensing image registration, ISPRS J. Photogramm. Remote Sens. 145 (2018) 148–164.
[231] C.B. Choy, J. Gwak, S. Savarese, M. Chandraker, Universal correspondence network, in: Advances in Neural Information Processing Systems, 2016, pp. 2414–2422.
[232] K. Han, R.S. Rezende, B. Ham, K.-Y.K. Wong, M. Cho, C. Schmid, J. Ponce, Scnet: Learning semantic correspondence, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1831–1840.
[233] T. Plötz, S. Roth, Neural nearest neighbors networks, in: Advances in Neural Information Processing Systems, 2018, pp. 1087–1098.
[234] Y.-C. Chen, P.-H. Huang, L.-Y. Yu, J.-B. Huang, M.-H. Yang, Y.-Y. Lin, Deep semantic matching with foreground detection and cycle-consistency, in: Proceedings of the Asian Conference on Computer Vision, 2018, pp. 347–362.
[235] S. Kim, S. Lin, S.R. JEON, D. Min, K. Sohn, Recurrent transformer networks for semantic correspondence, in: Advances in Neural Information Processing Systems, 2018, pp. 6126–6136.
[236] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 27 (10) (2005) 1615–1630.
[237] E.M. Loiola, N.M.M. de Abreu, P.O. Boaventura-Netto, P. Hahn, T. Querido, A survey for the quadratic assignment problem, European J. Oper. Res. 176 (2) (2007) 657–690.
[238] G. Levi, A note on the derivation of maximal common subgraphs of two directed or undirected graphs, Calcolo 9 (4) (1973) 341.
[239] D.J. Cook, L.B. Holder, Mining Graph Data, John Wiley & Sons, 2006.
[240] L. Babai, Groups, graphs, algorithms: The graph isomorphism problem, Proc. Internat. Congr. Math. (2018).
[241] M. Leordeanu, M. Hebert, A spectral technique for correspondence problems using pairwise constraints, in: Proceedings of the IEEE International Conference on Computer Vision, 2005, pp. 1482–1489.
[242] T. Cour, P. Srinivasan, J. Shi, Balanced graph matching, in: Advances in Neural Information Processing Systems, 2007, pp. 313–320.
[243] H. Liu, S. Yan, Common visual pattern discovery via spatially coherent correspondences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 1609–1616.
[244] B. Jiang, H. Zhao, J. Tang, B. Luo, A sparse nonnegative matrix factorization technique for graph matching problems, Pattern Recognit. 47 (2) (2014) 736–747.
[245] A. Egozi, Y. Keller, H. Guterman, A probabilistic approach to spectral graph matching, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2012) 18–27.
[246] S. Umeyama, An eigendecomposition approach to weighted graph matching problems, IEEE Trans. Pattern Anal. Mach. Intell. 10 (5) (1988) 695–703.
[247] T. Caelli, S. Kosinov, An eigenspace projection clustering method for inexact graph matching, IEEE Trans. Pattern Anal. Mach. Intell. 26 (4) (2004) 515–519.
[248] Q. Zhao, S.E. Karisch, F. Rendl, H. Wolkowicz, Semidefinite programming relaxations for the quadratic assignment problem, J. Comb. Optim. 2 (1) (1998) 71–109.
[249] C. Schellewald, C. Schnörr, Probabilistic subgraph matching based on convex relaxation, in: Proceedings of the International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, 2005, pp. 171–186.
[250] I. Kezurer, S.Z. Kovalsky, R. Basri, Y. Lipman, Tight relaxation of quadratic matching, in: Computer Graphics Forum, Vol. 34, Wiley Online Library, 2015, pp. 115–128.
[251] H.A. Almohamad, S.O. Duffuaa, A linear programming approach for the weighted graph matching problem, IEEE Trans. Pattern Anal. Mach. Intell. 15 (5) (1993) 522–525.
[252] W.P. Adams, T.A. Johnson, Improved linear programming-based lower bounds for the quadratic assignment problem, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 16 (1994) 43–77.
[253] P. Swoboda, J. Kuske, B. Savchynskyy, A dual ascent framework for lagrangean decomposition of combinatorial problems, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1596–1606.
[254] Q. Chen, V. Koltun, Robust nonrigid registration by convex optimization, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2039–2047.
[255] P. Swoboda, C. Rother, H. Abu Alhaija, D. Kainmuller, B. Savchynskyy, A study of lagrangean decompositions and dual ascent solvers for graph matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1607–1616.
[256] L. Torresani, V. Kolmogorov, C. Rother, A dual decomposition approach to feature correspondence, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2) (2012) 259–271.
[257] Z. Zhang, Q. Shi, J. McAuley, W. Wei, Y. Zhang, A. Van Den Hengel, Pairwise matching through max-weight bipartite belief propagation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1202–1210.
[258] M. Zaslavskiy, F. Bach, J.-P. Vert, A path following algorithm for the graph matching problem, IEEE Trans. Pattern Anal. Mach. Intell. 31 (12) (2009) 2227–2242.
[259] Z.-Y. Liu, H. Qiao, Gnccp—Graduated nonconvexityand concavity procedure, IEEE Trans. Pattern Anal. Mach. Intell. 36 (6) (2014) 1258–1267.
[260] S. Gold, A. Rangarajan, A graduated assignment algorithm for graph matching, IEEE Trans. Pattern Anal. Mach. Intell. 18 (4) (1996) 377–388.
[261] Y. Tian, J. Yan, H. Zhang, Y. Zhang, X. Yang, H. Zha, On the convergence of graph matching: Graduated assignment revisited, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 821–835.
[262] B. Jiang, J. Tang, C. Ding, Y. Gong, B. Luo, Graph matching via multiplicative update algorithm, in: Advances in Neural Information Processing Systems, 2017, pp. 3187–3195.
[263] T. Yu, J. Yan, B. Li, Determinant regularization for gradient-efficient graph matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
[264] M. Cho, J. Lee, K.M. Lee, Reweighted random walks for graph matching, in: Proceedings of the European Conference on Computer vision, 2010, pp. 492–505.
[265] J. Lee, M. Cho, K.M. Lee, A graph matching algorithm using data-driven markov chain monte carlo sampling, in: Proceedings of the International Conference on Pattern Recognition, 2010, pp. 2816–2819.
[266] Y. Suh, M. Cho, K.M. Lee, Graph matching via sequential monte carlo, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 624–637.
[267] J. Yan, Y. Tian, H. Zha, X. Yang, Y. Zhang, S.M. Chu, Joint optimization for consistent multiple graph matching, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1649–1656.
[268] J. Yan, M. Cho, H. Zha, X. Yang, S.M. Chu, Multi-graph matching via affinity optimization with graduated consistency regularization, IEEE Trans. Pattern Anal. Mach. Intell. 38 (6) (2015) 1228–1242.
[269] P. Swoboda, A. Mokarian, C. Theobalt, F. Bernard, et al., A convex relaxation for multi-graph matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 11156–11165.
[270] F. Bernard, J. Thunberg, P. Swoboda, C. Theobalt, Hippi: Higher-order projected power iterations for scalable multi-matching, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10284–10293.
[271] Z. Jiang, T. Wang, J. Yan, Unifying offline and online multi-graph matching via finding shortest paths on supergraph, IEEE Trans. Pattern Anal. Mach. Intell. (2020).
[272] J. Lee, M. Cho, K.M. Lee, Hyper-graph matching via reweighted random walks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1633–1640.
[273] R. Zass, A. Shashua, Probabilistic graph and hypergraph matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[274] J. Yan, C. Zhang, H. Zha, W. Liu, X. Yang, S.M. Chu, Discrete hyper-graph matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1520–1528.
[275] T.S. Caetano, J.J. McAuley, L. Cheng, Q.V. Le, A.J. Smola, Learning graph matching, IEEE Trans. Pattern Anal. Mach. Intell. 31 (6) (2009) 1048–1058.
[276] M. Leordeanu, R. Sukthankar, M. Hebert, Unsupervised learning for graph matching, Int. J. Comput. Vis. 96 (1) (2012) 28–45.
[277] A. Zanfir, C. Sminchisescu, Deep learning of graph matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2684–2693.
[278] T. Wang, H. Liu, Y. Li, Y. Jin, X. Hou, H. Ling, Learning combinatorial solver for graph matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
[279] P.J. Besl, N.D. McKay, Method for registration of 3-d shapes, in: Sensor Fusion IV: Control Paradigms and Data Structures, Vol. 1611, 1992, pp. 586–607.
[280] S. Granger, X. Pennec, Multi-scale em-icp: A fast and robust approach for surface registration, in: Proceedings of the European Conference on Computer Vision, 2002, pp. 418–432.
[281] A.W. Fitzgibbon, Robust registration of 2D and 3D point sets, Image Vis. Comput. 21 (13–14) (2003) 1145–1153.
[282] D. Chetverikov, D. Stepanov, P. Krsek, Robust euclidean alignment of 3D point sets: the trimmed iterative closest point algorithm, Image Vis. Comput. 23 (3) (2005) 299–309.
[283] F. Pomerleau, F. Colas, R. Siegwart, S. Magnenat, Comparing ICP variants on real-world data sets, Auton. Robots 34 (3) (2013) 133–148.
[284] S. Gold, A. Rangarajan, C.-P. Lu, S. Pappu, E. Mjolsness, New algorithms for 2D and 3D point matching: Pose estimation and correspondence, Pattern Recognit. 31 (8) (1998) 1019–1031.
[285] H. Chui, A. Rangarajan, A new point matching algorithm for non-rigid registration, Comput. Vis. Image Underst. 89 (2–3) (2003) 114–141.
[286] M. Sofka, G. Yang, C.V. Stewart, Simultaneous covariance driven correspondence (cdc) and transformation estimation in the expectation maximization framework, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[287] A. Myronenko, X. Song, Point set registration: Coherent point drift, IEEE Trans. Pattern Anal. Mach. Intell. 32 (12) (2010) 2262–2275.
[288] J. Ma, J. Zhao, A.L. Yuille, Non-rigid point set registration by preserving global and local structures, IEEE Trans. Image Process. 25 (1) (2016) 53–64.

[289] S. Zhang, Y. Yang, K. Yang, Y. Luo, S.-H. Ong, Point set registration with global-local correspondence and transformation estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2669–2677.

[290] Y. Tsin, T. Kanade, A correlation-based approach to robust point set registration, in: Proceedings of the European Conference on Computer Vision, 2004, pp. 558–569.

[291] D. Campbell, L. Petersson, An adaptive data representation for robust point-set registration and merging, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4292–4300.

[292] Q. Liao, D. Sun, H. Andreasson, Point set registration for 3D range scans using fuzzy cluster-based metric and efficient global optimization, IEEE Trans. Pattern Anal. Mach. Intell. (2020).

[293] L. Silva, O.R.P. Bellon, K.L. Boyer, Precision range image registration using a robust surface interpenetration measure and enhanced genetic algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 762–776.

[294] C. Papazov, D. Burschka, Stochastic global optimization for robust point set registration, Comput. Vis. Image Underst. 115 (12) (2011) 1598–1609.

[295] H. Li, R. Hartley, The 3d-3d registration problem revisited, in: Proceedings of the IEEE International Conference on Computer Vision, 2007, pp. 1–8.

[296] A. Parra Bustos, T.-J. Chin, D. Suter, Fast rotation search with stereographic projections for 3d registration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3930–3937.

[297] D. Campbell, L. Petersson, Gogma: Globally-optimal gaussian mixture alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5685–5694.

[298] J. Yang, H. Li, D. Campbell, Y. Jia, Go-ICP: A globally optimal solution to 3D ICP point-set registration, IEEE Trans. Pattern Anal. Mach. Intell. 38 (11) (2016) 2241–2254.

[299] Y. Liu, C. Wang, Z. Song, M. Wang, Efficient global point cloud registration by matching rotation invariant features through translation search, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 448–463.

[300] H. Maron, N. Dym, I. Kezurer, S. Kovalsky, Y. Lipman, Point registration via efficient convex relaxation, ACM Trans. Graph. 35 (4) (2016) 73.

[301] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.

[302] J.-C. Bazin, Y. Seo, M. Pollefeys, Globally optimal consensus set maximization through rotation search, in: Proceedings of the Asian Conference on Computer Vision, 2012, pp. 539–551.

[303] J. Ma, J. Zhao, J. Tian, Z. Tu, A.L. Yuille, Robust estimation of nonrigid transformation for point set registration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2147–2154.

[304] A.E. Johnson, M. Hebert, Using spin images for efficient object recognition in cluttered 3D scenes, IEEE Trans. Pattern Anal. Mach. Intell. 21 (5) (1999) 433–449.

[305] N. Gelfand, N.J. Mitra, L.J. Guibas, H. Pottmann, Robust global registration, in: Symposium on Geometry Processing, Vol. 2, Vienna, Austria, 2005, p. 5.

[306] R.B. Rusu, N. Blodow, M. Beetz, Fast point feature histograms (fpfh) for 3d registration, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2009, pp. 3212–3217.

[307] P. Torr, A. Zisserman, Robust computation and parametrization of multiple view relations, in: Proceedings of the International Conference on Computer Vision, 1998, pp. 727–732.

[308] P.H.S. Torr, A. Zisserman, MLESAC: A new robust estimator with application to estimating image geometry, Comput. Vis. Image Underst. 78 (1) (2000) 138–156.

[309] K. Ni, H. Jin, F. Dellaert, Groupsac: Efficient consensus in the presence of groupings, in: Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 2193–2200.

[310] R. Raguram, O. Chum, M. Pollefeys, J. Matas, J.-M. Frahm, Usac: a universal framework for random sample consensus, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2012) 2022–2038.

[311] D. Barath, J. Matas, J. Noskova, Magsac: marginalizing sample consensus, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10197–10205.

[312] D. Nasuto, J.M.B.R. Craddock, Napsac: High noise, high dimensional robust estimation-its in the bag, in: Proc. Brit. Mach. Vision Conf. 2002, pp. 458–467.

[313] O. Chum, J. Matas, Matching with prosac-progressive sample consensus, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 220–226.

[314] O. Chum, J. Matas, J. Kittler, Locally optimized RANSAC, in: Joint Pattern Recognition Symposium, Springer, 2003, pp. 236–243.

[315] K. Lebeda, J. Matas, O. Chum, Fixing the locally optimized ransac-ful experimental evaluation, in: Proceedings of the British Machine Vision Conference, 2012, pp. 1–11.

[316] D. Barath, M. Ivashechkin, J. Matas, Progressive NAPSAC: sampling from gradually growing neighborhoods, 2019, arXiv preprint arXiv:1906.02295.

[317] D. Barath, J. Noskova, M. Ivashechkin, J. Matas, MAGSAC++, a fast, reliable and accurate robust estimator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 1304–1312.

[318] J. Pilet, V. Lepetit, P. Fua, Fast non-rigid surface detection, registration and realistic augmentation, Int. J. Comput. Vis. 76 (2) (2008) 109–122.

[319] J. Ma, J. Zhao, J. Tian, A.L. Yuille, Z. Tu, Robust point matching via vector field consensus, IEEE Trans. Image Process. 23 (4) (2014) 1706–1721.

[320] J. Ma, W. Qiu, J. Zhao, Y. Ma, A.L. Yuille, Z. Tu, Robust L_2E estimation of transformation for non-rigid registration, IEEE Trans. Signal Process. 63 (5) (2015) 1115–1129.

[321] J. Ma, J. Jiang, C. Liu, Y. Li, Feature guided Gaussian mixture model with semi-supervised EM and local geometric constraint for retinal image registration, Inform. Sci. 417 (2017) 128–142.

[322] V. Gay-Bellile, A. Bartoli, P. Sayd, Direct estimation of nonrigid registrations with image-based self-occlusion reasoning, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2008) 87–104.

[323] X. Li, Z. Hu, Rejecting mismatches by correspondence function, Int. J. Comput. Vis. 89 (1) (2010) 1–17.

[324] J. Ma, J. Zhao, J. Tian, X. Bai, Z. Tu, Regularized vector field learning with sparse approximation for mismatch removal, Pattern Recognit. 46 (12) (2013) 3519–3532.

[325] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, J. Tian, Robust feature matching for remote sensing image registration via locally linear transforming, IEEE Trans. Geosci. Remote Sens. 53 (12) (2015) 6469–6481.

[326] J. Ma, J. Wu, J. Zhao, J. Jiang, H. Zhou, Q.Z. Sheng, Nonrigid point set registration with robust transformation learning under manifold regularization, IEEE Trans. Neural Netw. Learn. Syst. (2019).

[327] J. Ma, X. Jiang, J. Jiang, Y. Gao, Feature-guided Gaussian mixture model for image matching, Pattern Recognit. 92 (2019) 231–245.

[328] Y. Lipman, S. Yagev, R. Poranne, D.W. Jacobs, R. Basri, Feature matching with bounded distortion, ACM Trans. Graph. 33 (3) (2014) 26.

[329] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T.-D. Nguyen, M.-M. Cheng, Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4181–4190.

[330] J. Ma, J. Zhao, J. Jiang, H. Zhou, X. Guo, Locality preserving matching, Int. J. Comput. Vis. 127 (5) (2019) 512–531.

[331] J. Ma, J. Jiang, H. Zhou, J. Zhao, X. Guo, Guided locality preserving feature matching for remote sensing image registration, IEEE Trans. Geosci. Remote Sens. 56 (8) (2018) 4435–4447.

[332] X. Jiang, J. Jiang, A. Fan, Z. Wang, J. Ma, Multiscale locality and rank preservation for robust feature matching of remote sensing images, IEEE Trans. Geosci. Remote Sens. 57 (9) (2019) 6462–6472.

[333] X. Jiang, J. Ma, J. Jiang, X. Guo, Robust feature matching using spatial clustering with heavy outliers, IEEE Trans. Image Process. 29 (2019) 736–746.

[334] W.-Y. Lin, F. Wang, M.-M. Cheng, S.-K. Yeung, P.H. Torr, M.N. Do, J. Lu, Code: Coherence based decision boundaries for feature correspondence, IEEE Trans. Pattern Anal. Mach. Intell. 40 (1) (2017) 34–47.

[335] X. Jiang, J. Ma, J. Chen, Progressive filtering for feature matching, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2019, pp. 2217–2221.

[336] X. Jiang, J. Ma, A. Fan, H. Xu, G. Lin, T. Lu, X. Tian, Robust feature matching for remote sensing image registration via linear adaptive filtering, IEEE Trans. Geosci. Remote Sens. (2020).

[337] S. Lee, J. Lim, I.H. Suh, Progressive feature matching: Incremental graph construction and optimization, IEEE Trans. Image Process. (2020).

[338] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, C. Rother, Dsac-differentiable ransac for camera localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6684–6692.

[339] E. Brachmann, C. Rother, Neural-guided ransac: Learning where to sample model hypotheses, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4322–4331.

[340] F. Kluger, E. Brachmann, H. Ackermann, C. Rother, M.Y. Yang, B. Rosenhahn, Consac: Robust multi-model fitting by conditional sample consensus, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 4634–4643.

[341] K. Moo Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, P. Fua, Learning to find good correspondences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2666–2674.

[342] J. Ma, X. Jiang, J. Jiang, J. Zhao, X. Guo, Lmr: Learning a two-class classifier for mismatch removal, IEEE Trans. Image Process. 28 (8) (2019) 4045–4059.

[343] C. Zhao, Z. Cao, C. Li, X. Li, J. Yang, Nm-net: Mining reliable neighbors for robust feature correspondences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 215–224.

[344] R. Ranftl, V. Koltun, Deep fundamental matrix estimation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 284–299.

[345] J. Zhang, D. Sun, Z. Luo, A. Yao, L. Zhou, T. Shen, Y. Chen, L. Quan, H. Liao, Learning two-view correspondences and geometry using order-aware network, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5845–5854.

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

[347] F. Maes, D. Vandermeulen, P. Suetens, Medical image registration using mutual information, Proc. IEEE 91 (10) (2003) 1699–1722.
[348] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, G. Marchal, Automated multi-modality image registration based on information theory, in: Information Processing in Medical Imaging, Vol. 3, Citeseer, 1995, pp. 263–274.
[349] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, P. Suetens, Multimodality image registration by maximization of mutual information, IEEE Trans. Med. Imaging 16 (2) (1997) 187–198.
[350] P.S. Reel, L.S. Dooley, K.C.P. Wong, A. Börner, Multimodal retinal image registration using a fast principal component analysis hybrid-based similarity measure, in: Proceedings of the IEEE International Conference on Image Processing, 2013, pp. 1428–1432.
[351] A. Rényi, et al., On measures of entropy and information, in: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, The Regents of the University of California, 1961.
[352] Y. He, A.B. Hamza, H. Krim, A generalized divergence measure for robust image registration, IEEE Trans. Signal Process. 51 (5) (2003) 1211–1220.
[353] M.R. Sabuncu, P. Ramadge, Using spanning graphs for efficient image registration, IEEE Trans. Image Process. 17 (5) (2008) 788–797.
[354] S. Martin, T.S. Durrani, A new divergence measure for medical image registration, IEEE Trans. Image Process. 16 (4) (2007) 957–966.
[355] H. Sundar, D. Shen, G. Biros, C. Xu, C. Davatzikos, Robust computation of mutual information using spatially adaptive meshes, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2007, pp. 950–958.
[356] X. Zhuang, S. Arridge, D.J. Hawkes, S. Ourselin, A nonrigid registration framework using spatially encoded mutual information and free-form deformations, IEEE Trans. Med. Imaging 30 (10) (2011) 1819–1828.
[357] E.R. Arce-Santana, A. Alba, Image registration using Markov random coefficient and geometric transformation fields, Pattern Recognit. 42 (8) (2009) 1660–1671.
[358] M. Freiman, M. Werman, L. Joskowicz, A curvelet-based patient-specific prior for accurate multi-modal brain image rigid registration, Med. Image Anal. 15 (1) (2011) 125–132.
[359] F. Yang, M. Ding, X. Zhang, W. Hou, C. Zhong, Non-rigid multi-modal medical image registration by combining L-BFGS-b with cat swarm optimization, Inf. Sci. 316 (2015) 440–456.
[360] P.A. Legg, P.L. Rosin, D. Marshall, J.E. Morgan, Feature neighbourhood mutual information for multi-modal image registration: an application to eye fundus imaging, Pattern Recognit. 48 (6) (2015) 1937–1946.
[361] A. Roche, X. Pennec, G. Malandain, N. Ayache, Rigid registration of 3-d ultrasound with MR images: a new approach combining intensity and gradient information, IEEE Trans. Med. Imaging 20 (10) (2001) 1038–1049.
[362] W. Wein, S. Brunke, A. Khamene, M.R. Callstrom, N. Navab, Automatic CT-ultrasound registration for diagnostic imaging and image-guided intervention, Med. Image Anal. 12 (5) (2008) 577–585.
[363] A. Andronache, M. von Siebenthal, G. Székely, P. Cattin, Non-rigid registration of multi-modal images using both mutual information and cross-correlation, Med. Image Anal. 12 (1) (2008) 3–15.
[364] A.K. Jain, F. Farrokhnia, Unsupervised texture segmentation using gabor filters, in: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics Conference Proceedings, 1990, pp. 14–19.
[365] B. Jian, B.C. Vemuri, J.L. Marroquin, Robust nonrigid multimodal image registration using local frequency maps, in: Proceedings of the Biennial International Conference on Information Processing in Medical Imaging, Springer, 2005, pp. 504–515.
[366] Y. Ou, A. Sotiras, N. Paragios, C. Davatzikos, Dramms: Deformable registration via attribute matching and mutual-saliency weighting, Med. Image Anal. 15 (4) (2011) 622–639.
[367] E. Shechtman, M. Irani, Matching local self-similarities across images and videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.
[368] M. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. Gleeson, M. Brady, J. Schnabel, MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration, Med. Image Anal. 16 (2012) 1423–1435, http://dx.doi.org/10.1016/j.media.2012.05.008.
[369] M.P. Heinrich, M. Jenkinson, B.W. Papież, M. Brady, J.A. Schnabel, Towards realtime multimodal fusion for image-guided interventions using self-similarities, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2013, pp. 187–194.
[370] C. Wachinger, N. Navab, Entropy and Laplacian images: Structural representations for multi-modal registration, Med. Image Anal. 16 (1) (2012) 1–17.
[371] F. Yang, M. Ding, X. Zhang, Y. Wu, J. Hu, Two phase non-rigid multi-modal image registration using Weber local descriptor-based similarity metrics and normalized mutual information, Sensors 13 (6) (2013) 7599–7617.
[372] G. Piella, Diffusion maps for multimodal registration, Sensors 14 (6) (2014) 10562–10577.
[373] H. Rivaz, Z. Karimaghaloo, D.L. Collins, Self-similarity weighted mutual information: a new nonrigid image registration metric, Med. Image Anal. 18 (2) (2014) 343–358.
[374] H. Rivaz, Z. Karimaghaloo, V.S. Fonov, D.L. Collins, Nonrigid registration of ultrasound and MRI using contextual conditioned mutual information, IEEE Trans. Med. Imaging 33 (3) (2013) 708–725.
[375] K. Kasiri, P. Fieguth, D.A. Clausi, Self-similarity measure for multi-modal image registration, in: Proceedings of the IEEE International Conference on Image Processing, 2016, pp. 4498–4502.
[376] D. Jiang, Y. Shi, X. Chen, M. Wang, Z. Song, Fast and robust multimodal image registration using a local derivative pattern, Med. Phys. 44 (2) (2017) 497–509.
[377] A. Kelman, M. Sofka, C.V. Stewart, Keypoint descriptors for matching across multiple image modalities and non-linear intensity variations, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–7.
[378] G. Wang, Z. Wang, Y. Chen, W. Zhao, Robust point matching method for multimodal retinal image registration, Biomed. Signal Process. Control 19 (2015) 68–76.
[379] Z. Li, F. Huang, J. Zhang, B. Dashtbozorg, S. Abbasi-Sureshjani, Y. Sun, X. Long, Q. Yu, B. ter Haar Romeny, T. Tan, Multi-modal and multi-vendor retina image registration, Biomed. Opt. Express 9 (2) (2018) 410–422.
[380] D. Lee, M. Hofmann, F. Steinke, Y. Altun, N.D. Cahill, B. Scholkopf, Learning similarity measure for multi-modal 3D image registration, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 186–193.
[381] A. Sedghi, J. Luo, A. Mehrtash, S. Pieper, C.M. Tempany, T. Kapur, P. Mousavi, W.M. Wells III, Semi-supervised deep metrics for image registration, 2018, arXiv preprint arXiv:1804.01565.
[382] R. Wright, B. Khanal, A. Gomez, E. Skelton, J. Matthew, J.V. Hajnal, D. Rueckert, J.A. Schnabel, Lstm spatial co-transformer networks for registration of 3D fetal US and mr brain images, in: Data Driven Treatment Response Assessment and Preterm, Perinatal, and Paediatric Image Analysis, Springer, 2018, pp. 149–159.
[383] J.M. Sloan, K.A. Goatman, J.P. Siebert, Learning Rigid Image Registration-Utilizing Convolutional Neural Networks for Medical Image Registration, SCITEPRESS-Science and Technology Publications, 2018.
[384] Y. Hu, M. Modat, E. Gibson, N. Ghavami, E. Bonmati, C.M. Moore, M. Emberton, J.A. Noble, D.C. Barratt, T. Vercauteren, Label-driven weakly-supervised learning for multimodal deformable image registration, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, 2018, pp. 1070–1074.
[385] 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, Med. Image Anal. 49 (2018) 1–13.
[386] C. Wang, G. Papanastasiou, A. Chartsias, G. Jacenkow, S.A. Tsaftaris, H. Zhang, Fire: unsupervised bi-directional inter-modality registration using deep networks, 2019, arXiv preprint arXiv:1907.05062.
[387] P. Isola, J.-Y. Zhu, T. Zhou, 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.
[388] M.-Y. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, in: Advances in Neural Information Processing Systems, 2017, pp. 700–708.
[389] Z. Yi, H. Zhang, P. Tan, M. Gong, Dualgan: Unsupervised dual learning for image-to-image translation, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2849–2857.
[390] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8789–8797.
[391] B. Gutiérrez-Becker, D. Mateus, L. Peter, N. Navab, Learning optimization updates for multimodal registration, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 19–27.
[392] F.S. Bashiri, A. Baghaie, R. Rostami, Z. Yu, R.M. D'Souza, Multi-modal medical image registration with full or partial data: a manifold learning approach, J. Imaging 5 (1) (2019) 5.
[393] S. Suri, P. Reinartz, Mutual-information-based registration of terrasar-x and ikonos imagery in urban areas, IEEE Trans. Geosci. Remote Sens. 48 (2) (2009) 939–949.
[394] A. Wong, J. Orchard, Efficient FFT-accelerated approach to invariant optical–lidar registration, IEEE Trans. Geosci. Remote Sens. 46 (11) (2008) 3917–3925.
[395] H. Zhang, R. Xu, Exploring the optimal integration levels between SAR and optical data for better urban land cover mapping in the pearl river delta, Int. J. Appl. Earth Obs. Geoinf. 64 (2018) 87–95.
[396] M. Schmitt, F. Tupin, X.X. Zhu, Fusion of SAR and optical remote sensing data—Challenges and recent trends, in: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2017, pp. 5458–5461.

[397] D. Marcos, R. Hamid, D. Tuia, Geospatial correspondences for multimodal registration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5091–5100.
[398] J. Inglada, Similarity measures for multisensor remote sensing images, in: IEEE International Geoscience and Remote Sensing Symposium, Vol. 1, IEEE, 2002, pp. 104–106.
[399] H.M. Chen, M.K. Arora, P.K. Varshney, Mutual information-based image registration for remote sensing data, Int. J. Remote Sens. 24 (18) (2003) 3701–3706.
[400] H.-M. Chen, P.K. Varshney, M.K. Arora, Performance of mutual information similarity measure for registration of multitemporal remote sensing images, IEEE Trans. Geosci. Remote Sens. 41 (11) (2003) 2445–2454.
[401] A.A. Cole-Rhodes, K.L. Johnson, J. LeMoigne, I. Zavorin, Multiresolution registration of remote sensing imagery by optimization of mutual information using a stochastic gradient, IEEE Trans. Image Process. 12 (12) (2003) 1495–1511.
[402] W. Yang, C. Han, H. Sun, Y. Cao, Registration of high resolution SAR and optical images based on multiple features, in: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Vol. 5, IEEE, 2005, pp. 3542–3544.
[403] G. Lehureau, F. Tupin, C. Tison, G. Oller, D. Petit, Registration of metric resolution SAR and optical images in urban areas, in: Proceedings of the European Conference on Synthetic Aperture Radar, VDE, 2008, pp. 1–4.
[404] X. Xu, X. Li, X. Liu, H. Shen, Q. Shi, Multimodal registration of remotely sensed images based on jeffrey's divergence, ISPRS J. Photogramm. Remote Sens. 122 (2016) 97–115.
[405] E. De Castro, C. Morandi, Registration of translated and rotated images using finite fourier transforms, IEEE Trans. Pattern Anal. Mach. Intell. (5) (1987) 700–703.
[406] B.S. Reddy, B.N. Chatterji, An FFT-based technique for translation, rotation, and scale-invariant image registration, IEEE Trans. Image Process. 5 (8) (1996) 1266–1271.
[407] X. Xie, Y. Zhang, X. Ling, X. Wang, A new registration algorithm for multimodal remote sensing images, in: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2018, pp. 7011–7014.
[408] H. Foroosh, J.B. Zerubia, M. Berthod, Extension of phase correlation to subpixel registration, IEEE Trans. Image Process. 11 (3) (2002) 188–200.
[409] D.J. Field, Relations between the statistics of natural images and the response properties of cortical cells, Josa A 4 (12) (1987) 2379–2394.
[410] X. Xie, Y. Zhang, X. Ling, X. Wang, A novel extended phase correlation algorithm based on log-gabor filtering for multimodal remote sensing image registration, Int. J. Remote Sens. 40 (14) (2019) 5429–5453.
[411] J. Zhang, M. Zareapoor, X. He, D. Shen, D. Feng, J. Yang, Mutual information based multi-modal remote sensing image registration using adaptive feature weight, Remote Sens. Lett. 9 (7) (2018) 646–655.
[412] Y. Xiang, F. Wang, L. Wan, N. Jiao, H. You, Os-flow: A robust algorithm for dense optical and sar image registration, IEEE Trans. Geosci. Remote Sens. 57 (9) (2019) 6335–6354.
[413] Y. Xiang, R. Tao, L. Wan, F. Wang, H. You, Os-pc: Combining feature representation and 3-d phase correlation for subpixel optical and SAR image registration, IEEE Trans. Geosci. Remote Sens. (2020).
[414] A. Dame, E. Marchand, Second-order optimization of mutual information for real-time image registration, IEEE Trans. Image Process. 21 (9) (2012) 4190–4203.
[415] M. Hasan, M.R. Pickering, X. Jia, Robust automatic registration of multimodal satellite images using ccre with partial volume interpolation, IEEE Trans. Geosci. Remote Sens. 50 (10) (2012) 4050–4061.
[416] K. Karantzalos, A. Sotiras, N. Paragios, Efficient and automated multimodal satellite data registration through mrfs and linear programming, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 329–336.
[417] M.L. Uss, B. Vozel, V.V. Lukin, K. Chehdi, Multimodal remote sensing image registration with accuracy estimation at local and global scales, IEEE Trans. Geosci. Remote Sens. 54 (11) (2016) 6587–6605.
[418] Z. Yi, C. Zhiguo, X. Yang, Multi-spectral remote image registration based on SIFT, Electron. Lett. 44 (2) (2008) 107–108.
[419] M. Hasan, M.R. Pickering, X. Jia, Modified sift for multi-modal remote sensing image registration, in: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, 2012, pp. 2348–2351.
[420] X. Xiong, Q. Xu, G. Jin, H. Zhang, X. Gao, Rank-based local self-similarity descriptor for optical-to-SAR image matching, IEEE Geosci. Remote Sens. Lett. (2019).
[421] Y. Ye, J. Shan, L. Bruzzone, L. Shen, Robust registration of multimodal remote sensing images based on structural similarity, IEEE Trans. Geosci. Remote Sens. 55 (5) (2017) 2941–2958.
[422] S. Cui, Y. Zhong, Multi-modal remote sensing image registration based on multi-scale phase congruency, in: Proceedings of the IEEE IAPR Workshop on Pattern Recognition in Remote Sensing, IEEE, 2018, pp. 1–5.
[423] S. Cui, Y. Zhong, A. Ma, L. Zhang, A novel robust feature descriptor for multi-source remote sensing image registration, in: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, 2019, pp. 919–922.
[424] H. Sui, C. Xu, J. Liu, F. Hua, Automatic optical-to-SAR image registration by iterative line extraction and voronoi integrated spectral point matching, IEEE Trans. Geosci. Remote Sens. 53 (11) (2015) 6058–6072.
[425] R.G. Von Gioi, J. Jakubowicz, J.-M. Morel, G. Randall, Lsd: A fast line segment detector with a false detection control, IEEE Trans. Pattern Anal. Mach. Intell. 32 (4) (2008) 722–732.
[426] N. Ahuja, Dot pattern processing using voronoi neighborhoods, IEEE Trans. Pattern Anal. Mach. Intell. (3) (1982) 336–343.
[427] M. Carcassoni, E.R. Hancock, Spectral correspondence for point pattern matching, Pattern Recognit. 36 (1) (2003) 193–204.
[428] B. Xiong, W. Li, L. Zhao, J. Lu, X. Zhang, G. Kuang, Registration for SAR and optical images based on straight line features and mutual information, in: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2016, pp. 2582–2585.
[429] P. Kovesi, Phase congruency detects corners and edges, in: The Australian Pattern Recognition Society Conference: DICTA, Vol. 2003, 2003.
[430] Y. Ye, L. Bruzzone, J. Shan, L. Shen, Fast and robust structure-based multimodal geospatial image matching, in: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2017, pp. 5141–5144.
[431] X. Liu, Y. Ai, J. Zhang, Z. Wang, A novel affine and contrast invariant descriptor for infrared and visible image registration, Remote Sens. 10 (4) (2018) 658.
[432] X. Liu, Y. Ai, B. Tian, D. Cao, Robust and fast registration of infrared and visible images for electro-optical pod, IEEE Trans. Ind. Electron. 66 (2) (2018) 1335–1344.
[433] C. Deng, X. Yuan, L. Deng, J. Chen, Detecting matching blunders of multi-source remote sensing images via graph theory, Sensors 20 (13) (2020) 3712.
[434] Z. Yang, T. Dan, Y. Yang, Multi-temporal remote sensing image registration using deep convolutional features, IEEE Access 6 (2018) 38544–38555.
[435] R. Zhu, D. Yu, S. Ji, M. Lu, Matching RGB and infrared remote sensing images with densely-connected convolutional neural networks, Remote Sens. 11 (23) (2019) 2836.
[436] D. Quan, S. Wang, X. Liang, R. Wang, S. Fang, B. Hou, L. Jiao, Deep generative matching network for optical and SAR image registration, in: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2018, pp. 6215–6218.
[437] N. Merkle, S. Auer, R. Müller, P. Reinartz, Exploring the potential of conditional adversarial networks for optical and SAR image matching, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11 (6) (2018) 1811–1820.
[438] A. Zampieri, G. Charpiat, N. Girard, Y. Tarabalka, Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 657–673.
[439] K. Yu, J. Ma, F. Hu, T. Ma, S. Quan, B. Fang, A grayscale weight with window algorithm for infrared and visible image registration, Infrared Phys. Technol. 99 (2019) 178–186.
[440] Y.N. Dwith Chenna, P. Ghassemi, T.J. Pfefer, J. Casamento, Q. Wang, Free-form deformation approach for registration of visible and infrared facial images in fever screening, Sensors 18 (1) (2018) 125.
[441] T. Hrkać, Z. Kalafatić, J. Krapac, Infrared-visual image registration based on corners and hausdorff distance, in: Scandinavian Conference on Image Analysis, Springer, 2007, pp. 383–392.
[442] X. Liu, J.-B. Li, J.-S. Pan, Feature point matching based on distinct wavelength phase congruency and log-gabor filters in infrared and visible images, Sensors 19 (19) (2019) 4244.
[443] F. Wu, B. Wang, X. Yi, M. Li, J. Hao, H. Qin, H. Zhou, Visible and infrared image registration based on visual salient features, J. Electron. Imaging 24 (5) (2015) 053017.
[444] F. Liu, S. Seipel, Infrared-visible image registration for augmented reality-based thermographic building diagnostics, Vis. Eng. 3 (1) (2015) 16.
[445] C. Min, Y. Gu, Y. Li, F. Yang, Non-rigid infrared and visible image registration by enhanced affine transformation, Pattern Recognit. (2020) 107377.
[446] C. Min, Y. Gu, F. Yang, Y. Li, W. Lian, Non-rigid registration for infrared and visible images via Gaussian weighted shape context and enhanced affine transformation, IEEE Access 8 (2020) 42562–42575.
[447] G.-A. Bilodeau, A. Torabi, F. Morin, Visible and infrared image registration using trajectories and composite foreground images, Image Vis. Comput. 29 (1) (2011) 41–50.
[448] J. Han, E.J. Pauwels, P. De Zeeuw, Visible and infrared image registration in man-made environments employing hybrid visual features, Pattern Recognit. Lett. 34 (1) (2013) 42–51.
[449] S. Sonn, G.-A. Bilodeau, P. Galinier, Fast and accurate registration of visible and infrared videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 308–313.
[450] M.Y. Yang, Y. Qiang, B. Rosenhahn, A global-to-local framework for infrared and visible image sequence registration, in: 2015 IEEE Winter Conference on Applications of Computer Vision, IEEE, 2015, pp. 381–388.
[451] X. Sun, T. Xu, J. Zhang, X. Li, A hierarchical framework combining motion and feature information for infrared-visible video registration, Sensors 17 (2) (2017) 384.

[452] D.-L. Nguyen, P.-L. St-Charles, G.-A. Bilodeau, Non-planar infrared-visible registration for uncalibrated stereo pairs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 63–71.
[453] L. Wang, C. Gao, Y. Zhao, T. Song, Q. Feng, Infrared and visible image registration using transformer adversarial network, in: Proceedings of the IEEE International Conference on Image Processing, 2018, 1248–1252.
[454] E.B. Baruch, Y. Keller, Multimodal matching using a hybrid convolutional neural network, 2018, arXiv preprint arXiv:1810.12941.
[455] M. Arar, Y. Ginger, D. Danon, A.H. Bermano, D. Cohen-Or, Unsupervised multi-modal image registration via geometry preserving image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 13410–13419.
[456] M. Ghantous, M. Bayoumi, Mirf: a multimodal image registration and fusion module based on DT-CWT, J. Signal Process. Syst. 71 (1) (2013) 41–55.
[457] A. Torabi, G. Massé, G.-A. Bilodeau, An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications, Comput. Vis. Image Underst. 116 (2) (2012) 210–221.
[458] K.-S. Cheng, H.-Y. Lin, Automatic target recognition by infrared and visible image matching, in: Proceedings of the IAPR International Conference on Machine Vision Applications, IEEE, 2015, pp. 312–315.
[459] S.-Y. Cao, H.-L. Shen, S.-J. Chen, C. Li, Boosting structure consistency for multispectral and multimodal image registration, IEEE Trans. Image Process. 29 (2020) 5147–5162.
[460] Z. Wei, C. Jung, C. Su, Reginet: Gradient guided multispectral image registration using convolutional neural networks, Neurocomputing (2020).
[461] D. Yi, S. Liao, Z. Lei, J. Sang, S.Z. Li, Partial face matching between near infrared and visual images in mbgc portal challenge, in: International Conference on Biometrics, Springer, 2009, pp. 733–742.
[462] D. Yi, R. Liu, R. Chu, Z. Lei, S.Z. Li, Face matching between near infrared and visible light images, in: International Conference on Biometrics, Springer, 2007, pp. 523–530.
[463] O. Boiman, M. Irani, Detecting irregularities in images and in video, Int. J. Comput. Vis. 74 (1) (2007) 17–31.
[464] I. Rocco, R. Arandjelovic, J. Sivic, Convolutional neural network architecture for geometric matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6148–6157.
[465] D. Zosso, X. Bresson, J.-P. Thiran, Geodesic active fields—a geometric framework for image registration, IEEE Trans. Image Process. 20 (5) (2010) 1300–1312.
[466] V.A. Zimmer, M.Á.G. Ballester, G. Piella, Multimodal image registration using Laplacian commutators, Inf. Fusion 49 (2019) 130–145.
[467] C.A. Cocosco, V. Kollokian, R.K.-S. Kwan, G.B. Pike, A.C. Evans, Brainweb: Online interface to a 3D MRI simulated brain database, in: NeuroImage, Citeseer, 1997.
[468] M. Brown, S. Süsstrunk, Multi-spectral SIFT for scene category recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 177–184.
[469] D. Mishkin, J. Matas, M. Perdoch, K. Lenc, Wxbs: Wide baseline stereo generalizations, 2015, arXiv preprint arXiv:1504.06603.
[470] D. Olid, J.M. Fácil, J. Civera, Single-view place recognition under seasonal changes, 2018, arXiv preprint arXiv:1808.06516.
[471] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, et al., Benchmarking 6dof outdoor visual localization in changing conditions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8601–8610.
[472] J. Heinly, E. Dunn, J.-M. Frahm, Comparative evaluation of binary features, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 759–773.
[473] C.L. Zitnick, K. Ramnath, Edge foci interest points, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 359–366.
[474] A. Vedaldi, B. Fulkerson, Vlfeat: An open and portable library of computer vision algorithms, 2008, http://www.vlfeat.org/.
[475] G. Bradski, A. Kaehler, Learning OpenCV: Computer vision with the OpenCV library, O'Reilly Media, Inc., 2008.
[476] J.M. Fitzpatrick, J.B. West, C.R. Maurer, Predicting error in rigid-body point-based registration, IEEE Trans. Med. Imaging 17 (5) (1998) 694–702.
[477] Z. Wang, A.C. Bovik, A universal image quality index, IEEE Signal Process. Lett. 9 (3) (2002) 81–84.
[478] H. Ghassemian, A review of remote sensing image fusion methods, Inf. Fusion 32 (2016) 75–89.
[479] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, Fusiongan: A generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (2019) 11–26.
[480] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, J. Jiang, Infrared and visible image fusion via detail preserving adversarial learning, Inf. Fusion 54 (2020) 85–98.
[481] J. Ma, H. Xu, J. Jiang, X. Mei, X.-P. Zhang, Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion, IEEE Trans. Image Process. 29 (2020) 4980–4995.
[482] H. Zhang, Z. Le, Z. Shao, H. Xu, J. Ma, MFF-Gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion, Inf. Fusion 66 (2020) 40–53.
[483] R. Hou, D. Zhou, R. Nie, D. Liu, X. Ruan, Brain CT and MRI medical image fusion using convolutional neural networks and a dual-channel spiking cortical model, Med. Biol. Eng. Comput. 57 (4) (2019) 887–900.
[484] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, J. Paisley, Pannet: A deep network architecture for pan-sharpening, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1753–1761.
[485] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, J. Jiang, Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion, Inf. Fusion 62 (2020) 110–120.
[486] N. Hayat, M. Imran, Ghost-free multi exposure image fusion technique using dense sift descriptor and guided filter, J. Vis. Commun. Image Represent. 62 (2019) 295–308.
[487] K.R. Prabhakar, V.S. Srikar, R.V. Babu, Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4714–4722.
[488] X. Guo, R. Nie, J. Cao, D. Zhou, L. Mei, K. He, Fusegan: Learning to fuse multi-focus image via conditional generative adversarial network, IEEE Trans. Multimed. 21 (8) (2019) 1982–1996.
[489] H. Zhang, H. Xu, Y. Xiao, X. Guo, J. Ma, Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 12797–12804.
[490] H. Xu, J. Ma, J. Jiang, X. Guo, H. Ling, U2fusion: A unified unsupervised image fusion network, IEEE Trans. Pattern Anal. Mach. Intell. (2020).
[491] A. Singh, Review article digital change detection techniques using remotely-sensed data, Int. J. Remote Sens. 10 (6) (1989) 989–1003.
[492] M.K. Ridd, J. Liu, A comparison of four algorithms for change detection in an urban environment, Remote Sens. Environ. 63 (2) (1998) 95–100.
[493] L. Giustarini, R. Hostache, P. Matgen, G.J.-P. Schumann, P.D. Bates, D.C. Mason, A change detection approach to flood mapping in urban areas using terrasar-x, IEEE Trans. Geosci. Remote Sens. 51 (4) (2012) 2417–2430.
[494] T. Celik, Unsupervised change detection in satellite images using principal component analysis and k-means clustering, IEEE Geosci. Remote Sens. Lett. 6 (4) (2009) 772–776.
[495] G. Quin, B. Pinel-Puyssegur, J.-M. Nicolas, P. Loreaux, Mimosa: An automatic change detection method for SAR time series, IEEE Trans. Geosci. Remote Sens. 52 (9) (2013) 5349–5363.
[496] F. Wang, Y. Wu, Q. Zhang, P. Zhang, M. Li, Y. Lu, Unsupervised change detection on SAR images using triplet Markov field model, IEEE Geosci. Remote Sens. Lett. 10 (4) (2012) 697–701.
[497] O. Miller, A. Pikaz, A. Averbuch, Objects based change detection in a pair of gray-level images, Pattern Recognit. 38 (11) (2005) 1976–1992.
[498] J. Im, J.R. Jensen, J. Tullis, Object-based change detection using correlation image analysis and image segmentation, Int. J. Remote Sens. 29 (2) (2008) 399–423.
[499] V. Alberga, Similarity measures of remotely sensed multi-sensor images for change detection applications, Remote Sens. 1 (3) (2009) 122–143.
[500] G. Mercier, G. Moser, S.B. Serpico, Conditional copulas for change detection in heterogeneous remote sensing images, IEEE Trans. Geosci. Remote Sens. 46 (5) (2008) 1428–1441.
[501] J. Prendes, M. Chabert, F. Pascal, A. Giros, J.-Y. Tourneret, A new multivariate statistical model for change detection in images acquired by homogeneous and heterogeneous sensors, IEEE Trans. Image Process. 24 (3) (2014) 799–812.
[502] J. Prendes, M. Chabert, F. Pascal, A. Giros, J.-Y. Tourneret, A Bayesian nonparametric model coupled with a Markov random field for change detection in heterogeneous remote sensing images, SIAM J. Imaging Sci. 9 (4) (2016) 1889–1921.
[503] P. Zhang, M. Gong, L. Su, J. Liu, Z. Li, Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images, ISPRS J. Photogramm. Remote Sens. 116 (2016) 24–41.
[504] W. Zhao, Z. Wang, M. Gong, J. Liu, Discriminative feature learning for unsupervised change detection in heterogeneous images based on a coupled neural network, IEEE Trans. Geosci. Remote Sens. 55 (12) (2017) 7066–7080.
[505] K. Mubea, G. Menz, Monitoring Land-Use Change in Nakuru (Kenya) Using Multi-Sensor Satellite Data, Scientific Research, 2012.
[506] S. Lowry, N. Sünderhauf, P. Newman, J.J. Leonard, D. Cox, P. Corke, M.J. Milford, Visual place recognition: A survey, IEEE Trans. Robot. 32 (1) (2015) 1–19.
[507] P. Mühlfellner, M. Bürki, M. Bosse, W. Derendarz, R. Philippsen, P. Furgale, Summary maps for lifelong visual localization, J. Field Robotics 33 (5) (2016) 561–590.
[508] G. Bresson, Z. Alsayed, L. Yu, S. Glaser, Simultaneous localization and mapping: A survey of current trends in autonomous driving, IEEE Trans. Intell. Veh. 2 (3) (2017) 194–220.
[509] T. Naseer, W. Burgard, C. Stachniss, Robust visual localization across seasons, IEEE Trans. Robot. 34 (2) (2018) 289–302.
[510] J.L. Schönberger, M. Pollefeys, A. Geiger, T. Sattler, Semantic visual localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6896–6906.

[511] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, A. Torii, Inloc: Indoor visual localization with dense matching and view synthesis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7199–7209.

[512] L. Liu, H. Li, Y. Dai, Efficient global 2d-3d matching for camera localization in a large-scale 3d map, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2372–2381.

[513] G.L. Foresti, Object recognition and tracking for remote video surveillance, IEEE Trans. Circuits Syst. Video Technol. 9 (7) (1999) 1045–1062.

[514] T.Y.-H. Chen, L. Ravindranath, S. Deng, P. Bahl, H. Balakrishnan, Glimpse: Continuous, real-time object recognition on mobile devices, in: Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems, 2015, pp. 155–168.

[515] K. Abdulrahim, R.A. Salam, Traffic surveillance: A review of vision based vehicle detection, recognition and tracking, Int. J. Appl. Eng. Res. 11 (1) (2016) 713–726.

[516] Y. Qu, G. Zhang, Z. Zou, Z. Liu, J. Mao, Active multimodal sensor system for target recognition and tracking, Sensors 17 (7) (2017) 1518.

[517] S. Ojha, S. Sakhare, Image processing techniques for object tracking in video surveillance-a survey, in: Proceedings of the International Conference on Pervasive Computing, 2015, pp. 1–6.

[518] S.P. Yoon, T.L. Song, T.H. Kim, Automatic target recognition and tracking in forward-looking infrared image sequences with a complex background, Int. J. Control Autom. Syst. 11 (1) (2013) 21–32.

[519] W. Zhou, Z. Li, P. Gao, Research on moving object detection and matching technology in multi-angle monitoring video, in: Proceedings of the Joint International Information Technology and Artificial Intelligence Conference, 2019, pp. 741–744.

[520] W. Xu, S. Zhong, L. Yan, F. Wu, W. Zhang, Moving object detection in aerial infrared images with registration accuracy prediction and feature points selection, Infrared Phys. Technol. 92 (2018) 318–326.

[521] W.-C. Hu, C.-H. Chen, T.-Y. Chen, D.-Y. Huang, Z.-C. Wu, Moving object detection and tracking from video captured by moving camera, J. Vis. Commun. Image Represent. 30 (2015) 164–180.

[522] J. Ma, C. Chen, C. Li, J. Huang, Infrared and visible image fusion via gradient transfer and total variation minimization, Inf. Fusion 31 (2016) 100–109.

[523] Y. Li, Z. He, H. Zhu, W. Zhang, Y. Wu, Jointly registering and fusing images from multiple sensors, Inf. Fusion 27 (2016) 85–94.