Hindawi
Scientific Programming
Volume 2021, Article ID 8509164, 25 pages
https://doi.org/10.1155/2021/8509164

Review Article

A Review of Keypoints' Detection and Feature Description in Image Registration

Cuiyin Liu,[1] **Jishang Xu**,[1] **and Feng Wang**[2,3]

[1]*Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China*
[2]*School of Physics and Electronic Engineering, Guangzhou University, Guangzhou 51006, China*
[3]*Astrophysics Centre of Guangzhou University, Guangzhou 51006, China*

Correspondence should be addressed to Cuiyin Liu; liucuiyin@163.com and Feng Wang; fengwang@gzhu.edu.cn

Received 16 May 2021; Revised 21 October 2021; Accepted 28 October 2021; Published 1 December 2021

Academic Editor: Cristian Mateos

Copyright © 2021 Cuiyin Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

For image registration, feature detection and description are critical steps that identify the keypoints and describe them for the subsequent matching to estimate the geometric transformation parameters between two images. Recently, there has been a large increase in the research methods of detection operators and description operators, from traditional methods to deep learning methods. To solve the problem, that is, which operator is suitable for specific application problems under different imaging conditions, the paper systematically reviewed commonly used descriptors and detectors from artificial methods to deep learning methods, and the corresponding principle, analysis, and comparative experiments are given as well. We introduce the handcrafted detectors including FAST, BRISK, ORB, SURF, SIFT, and KAZE and the handcrafted descriptors including BRISK, FREAK, BRIEF, SURF, ORB, SIFT, KAZE. At the same time, we review detectors based on deep learning technology including DetNet, TILDE, LIFT, multiscale detector, SuperPoint, and descriptors based on deep learning including pretrained descriptor, Siamese descriptor, LIFT, triplet network, and SuperPoint. Two group of comparison experiments are compared comprehensively and objectively on representative datasets. Finally, we concluded with insightful discussions and conclusions of descriptor and detector selection for specific application problem and hope this survey can be a reference for researchers and engineers in image registration and related fields.

1. Introduction

Image registration is an important process that is used to align two or more images of the same scene taken by different sensors, at different times, from different viewpoints, and with different illuminations. It provides the probability of the fusion of various visual data for further research. As the critical and fundamental problem in computer vision, its direct task is to identify and align a wide range of visual information from multisensors, thereby yielding richer visual representation for subsequent research and application [1, 2]. This technique is used for image fusion enhancing image quality [3], image mosaic creating a seamless panorama image from multiple images (which may be obtained from different time, different perspectives, or different sensors) [4], image segmentation dividing an image into several specific regions with unique properties [5], and object tracking constructing the complete motion trajectory of an object [4] as well as object detection finding out all interested targets in the image and determining their positions and categories [6].

The existing approaches of registration are categorized into intensity-based and feature-based pipelines [7, 8]. Intensity-based image registration aligns two images through an iterative process with a specified metric, an optimizer, and a transformation matrix. The accuracy of the registration is determined by the similarity measurement that describes the accuracy of image alignment and decides when to terminate the optimization process. The ultimate goal is to warp the sensed image to the coordinates of the reference image according to the obtained transformation matrix and align common area pixel-to-pixel. The feature-based method starts with feature detection, feature description, and then feature matching, followed by a transformation matrix

estimation, and finished with image warping. The local features used in image registration, include points, lines, contours, and polygons, etc. [9–23]. It is difficult to describe and locate line, contour, and polygon structure, so the keypoint is used as the primary feature in image registration. An algorithm of a keypoint detection, called as a detector, is designed to extract the local distinctive region, and an algorithm of keypoint description, called as a descriptor, is designed to represent the detected local region with invariance to the deformation of geometric, illumination, etc., called as the descriptor. During the last decades, detectors and descriptors have been developed extensively. These existing methods have been defined and proposed in [24–33]. Recently, because of the ability to obtain deep features and the representation from linear and nonlinear space through multilayer networks, deep learning techniques have increasingly grown up and have been used in every process of image registration to replace the traditional algorithm.

Even though numerous image registration methods have been developed, it is still difficult to construct a universal scheme for actual images with different sensors under different imaging condition. To the area-based method, it is available for image pairs with small deformation and is time-consuming. To the keypoint-based pipeline, a certain kind of detection operator is invariant to a certain kind of distorted image, and so is the detection operator. For example, FAST would be invalid to noise, blur, and compression because the scale space and denoising are not considered. The BRIEF descriptor is invalid to the rotation because the orientation is not considered in the process of constructing description vector [26]. The deep learning has achieved great success in computer vision such as image recognition, classification, enhancement, and segmentation [23, 34–43]. However, it still faces enormous challenges. Directly aligning image pairs through deep learning network would be hindered by the lack of training data with abundant geometrical deformations. On the contrary, it is still a great challenge to realize the end-to-end learning of feature detection, description, and matching using a deep network and then directly output transformation matrix.

Most of the existing surveys review handcrafted methods, and there is relatively little review on deep learning techniques that are applied and developed in image registration [44–47]. Recently, some reviews involved machine learning and deep learning used for one process of registration [48–51]. These papers have not surveyed comprehensively from handcrafted methods to the up-to-date methods, only involving one part of them. There are some survey papers comprehensively reviewed from the handcrafted to the deep learning [52, 53]. However, these papers briefly describe, summarize, and assess existing image registration methods rather than sketching the principle and analysis related to the algorithm.

The motivations of this survey include considering the precision of the image registration; we present recent developments of keypoints' detectors and descriptors, especially the deep learning techniques. For image pairs to be aligned, the question is how to choose the most appropriate algorithm to find a local feature and how to select the most suitable descriptor to represent, resulting in subsequent successful match. Compared to previous work, binary methods and deep learning methods are involved, the relative analyses of algorithms are included, and more variety of scenes are used in the experimental data [24–33].

We introduce the handcrafted detectors include FAST (features from accelerated segment test) [26], BRISK (binary robust invariant scalable keypoints) [28], ORB (oriented BRIEF) [30], SURF (speeded up robust features) [32], SIFT (the scale invariant feature transform keypoint) [31], and KAZE [33] and the handcrafted descriptors include BRISK (binary robust invariant scalable keypoints) [28], FREAK (fast retina keypoint) [29], BRIEF (binary robust independent elementary features) [27], SURF, ORB, SIFT, and KAZE. At the same time, we review detectors based on deep learning technology including DetNet [23], TILDE [42], LIFT [43], multiscale detector [36], and SuperPoint [39] and descriptors based on deep learning including pretrained descriptor [34], Siamese descriptor [37], LIFT, triplet network [35], and SuperPoint. These methods are all listed in Table 1, and the scale space, invariance to rotation, and illumination are listed as well. Harris and FAST are only detectors; BRISK, ORB, SIFT, SURF, and KAZE are both detectors and descriptors; BRIEF and FREAK are only descriptors. In application, the relative combination can be adopted for specific needs; for example, the combination of FAST with BRIEF [12].

The reminders of this paper are organized as follows. Section 2 introduces the methods of feature detection. Section 3 gives the methods of feature description. Section 4 gives the experimental datasets, evaluation indexes, and matching approaches. In Section 5, we carry out comparison experiments on datasets that include a large variety of scene types and transformations, present the evaluation, and discuss the applicability of various methods in light of evaluation values. Finally, Section 6 summarizes and discusses the future directions for detectors and descriptors in image registration.

2. Feature Detection

Image registration based on features is low-level processing that serves as the essential part for computer vision applications in remote-sensing observation, aided aviation, and astronomy observation. The image registration consists of detecting local feature regions describing them, matching them, estimating the transform model, and aligning two images. The step of detection and description is critical because it determines whether subsequent processing can be performed or not. The local features that include points, lines, curves, edges, and contours have been proposed and used in [12, 15, 18]. However, points (a.k.a. keypoints or interest points), as the most popular features, are used for image registration because they are easy to locate and descript compared with the other features. Accordingly, detectors and descriptors of points are excessively researched [25–33].

Table 1: Detectors and descriptors of the handcrafted feature point.

Detectors	Scale space	Invariance to rotation	Invariance to illumination	Descriptor	Invariance to scale	Invariance to rotation	Invariance to illumination
Harris	—	✓	✓	—	—	—	—
FAST	—	—	✓	—	—	—	—
BRISK	✓	—	✓	BRISK	✓	✓	✓
ORB	—	✓	✓	ORB	×	✓	✓
—	—	—	—	BRIEF	×	×	✓
—	—	—	—	FREAK	×	×	✓
KAZE	✓	✓	✓	KAZE	✓	✓	✓
SIFT	✓	✓	✓	SIFT	✓	✓	✓
SURF	✓	✓	✓	SURF	✓	✓	✓
DetNet	×	×	×	Pre-Net	✓	×	✓
TILDE	×	×	✓	Siamese-Net	×	✓	✓
LIFT	×	✓	✓	Triplet-Net	×	✓	✓
Multiscale	✓	✓	✓	LIFT	×	✓	✓
SuperPointer	✓	✓	✓	SuperPointer	✓	✓	✓

A good detector is designed to find stable and distinct local regions from images. Furthermore, these detectors still find or identify the local regions; even they have been transformed by viewpoint, illumination, scale, blur, and compression. Traditional methods are designed according to the prior mathematical theory, which is called a handcrafted operator [16, 25, 33, 44, 45] and classified into corner, binary corner, and blob. Corner points, defined as the intersecting location of two edge lines, are implemented by gradient computation (Harris) or comparison of pixels within the template [25, 26, 28, 30]. For the FAST, the calculations are replaced with a binary pixel comparison [25–30]. The blob detector is to find a maximum in a scale space constructed by difference-of-Gaussian filters [31] or nonlinear filters [33]. Recently, machine learning and deep learning techniques are extended to detect feature points [34, 54–59]. Some work uses a pretrained network to decompose images into feature blocks as keypoints [44]. Meanwhile, a triplet network is trained to identify whether a local region is an interesting point [35, 60]. In the following, we introduce these feature detectors from the handcrafted to the trainable. Different from other works, we not only investigated the related algorithms but also gave interpretation of principle of related algorithms and comparative experiments and insights of their strengths and weaknesses [49, 50].

2.1. Corner Detector: Harris. Common algorithms of detection are mainly classified into gradient-based detectors and intensity-based detectors such as Harris and FAST with introduction as follows. Harris corner detector proposed by Harris C. and *M* Shi and Tomasi identifies a candidate point as a corner point by gradients' computation. It is invariant to the image rotation, affine transformations, intensity, and viewpoint change in matching features [24]. The method consists of three steps. The first is to calculate the corresponding gradient according to the mathematic definition. The second work is to select the candidate corner point with a specified threshold. Finally, a region suppression is adopted to eliminate outliers. The Harris detector is defined as

$$M_{\text{Harris}} = \begin{bmatrix} A & C \\ C & B \end{bmatrix}, A = (I_x)^2 \oplus w, B = (I_y)^2 \oplus w, C = (I_x I_y)^2 \oplus w, \tag{1}$$

where M is a squared difference matrix, $\oplus$ denotes a convolution operation, w denotes a convolution kernel called filter window, and I_x and I_y are the gradient images of the input image. For obtaining computational efficiency, R response of equation (2) is used to replace the computation of eigenvalues of the matrix M:

$$R = AB - C^2 - k(A + B)^2, \tag{2}$$

where k is the sensitivity factor and constant typically 0.04. When the response R is higher than a specified threshold, a corner is detected. The Harris detector is invariant to rotation and illumination but sensitive to noise, as the gradient calculation is easily affected by noise [61]. It is not suitable for image pairs degraded by the noise.

2.2. FAST: Features from Accelerated Segment Test. FAST corner detector, proposed by E. Rosten and T. Drummond, is an improved detector of Harris by binary calculation and machine learning [26]. It is remarkable in the efficiency. The candidate Ip is identified as a corner point if there is a set of contiguous pixels on the circle template all brighter than the intensity of the candidate pixel Ip plus a threshold t, or all darker than $Ip - t$ [32, 62]. The radius of the circle template can be of any size or scale in theory. The original algorithm used *3* as the value of radius, checking 12 pixels on this circumference. For FAST, the check only proceeds on the pixels at the location of 1, 5, 9, and 13, as shown in Figure 1. If three of the absolute differences between Ip and each 1, 5, 9, and 13 are less than $Ip - t$ or more than $Ip_+ t$, Ip is a candidate corner. Otherwise, it is excluded.

With the application of machine learning in computer vision, FAST combines with decision trees to distinguish corners. To increase the stability, FAST-ER [62] increases the thickness of a circular template. For generalization in

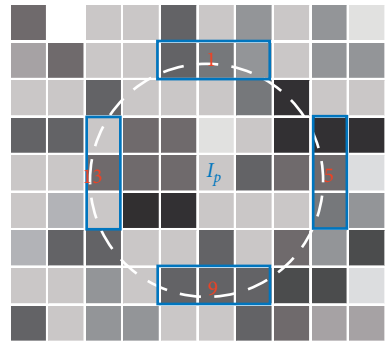

Figure 1: The principal of FAST detecting. For FAST, the check only proceeds on the pixels at the location of 1, 5, 9, and 13.

different environments, the collection of decision trees is adopted in the FAST detector [63]. Although using machine learning to speed up the detection accelerates the FAST detection, database-dependent problems exist in the FAST. As the detector, without considering the multiscale space, FAST is sensitive to noise, blur, compression, etc.

2.3. SIFT: The Scale Invariant Feature Transform Keypoint. SIFT, a blob feature point, is proposed by Lowe to solve the image rotation, affine transformations, intensity, and viewpoint change in matching features [31]. The SIFT algorithm is composed of four steps. Firstly, using multi-difference-of-Gaussian, we construct a scale space in which the subimages are produced by subtracting adjacent images filtered by Gaussian filter. The operation is defined as

$$\begin{aligned} L(x,y,\sigma) &= G(x,y,\sigma) * I(x,y), G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-x^2+y^2/2\sigma^2} \\ D(x,y,\sigma) &= (G(x,y,k\sigma) - G(x,y,\sigma)) * I(x,y) \\ &= L(x,y,k\sigma) - L(x,y,\sigma). \end{aligned} \tag{3}$$

Secondly, after the keypoint has been identified by comparing a pixel to its neighbours, the accurate localization by exploiting the 3D quadratic function is defined as

$$\begin{aligned} D(X) &= D + \frac{\partial \mathrm{D}^T}{\partial \mathrm{X}} X + \frac{1}{2} X^T \frac{\partial^2 \mathrm{D}}{\partial \mathrm{X}^2} x, \\ \hat{\mathrm{x}} &= -\frac{\partial^2 \mathrm{D}^{-1}}{\partial \mathrm{x}^2} \frac{\partial \mathrm{D}}{\partial \mathrm{x}}, \\ D(\hat{\mathrm{x}}) &= D + \frac{1}{2} \frac{\partial \mathrm{D}^T}{\partial \mathrm{x}} \hat{\mathrm{x}}. \end{aligned} \tag{4}$$

When the offset calculated $\hat{x}$ is larger than 0.5, the final keypoint location needs to correct and the offset is added. Eliminating edge responses and low contrast keypoint are also implemented in this section. Thirdly, a keypoint orientation $\theta(x, y)$ is defined as

$$\theta(x,y) = \tan^{-1} \frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)}. \tag{5}$$

An orientation histogram with 36 bins covering the 360 degrees range is formed from the gradient orientation within a region around the point. For the matching stability, only about 15% of points are assigned, and multiple orientations taking the histogram value are within 80% of the highest peak as a necessity. Lastly, the local area of a keypoint is divided into $4 * 4 = 16$ subareas with 8 values of the orientation histogram in each. Therefore, the descriptor vector of a keypoint is composed of $16 * 8 = 128$ feature values.

2.4. SURF: Speeded-Up Robust Features. SURF is modified from SIFT. However, SURF employs integral images to calculate the second-order of the Hessian matrix. The scale space is constructed by upscaling the filter size rather than iteratively downsampling the filtered images. The keypoint is determined in the $3 * 3 * 3$ neighbourhood in the successive layer of scale space, and the maximum is retained for further nonmaximum. A circular neighbourhood of radius $6s'$ around the keypoint is used as the area in which the cumulative operation of the wavelet response values is performed every 60 degrees. The maximum summed response of the subarea defines the orientation of a keypoint. The last step is to construct the description vector. A local square area with the same orientation of the keypoint and the size of $20s$ (where s is the scale of this keypoint) is used to generate the description and split into a $4 * 4$ subsquare region. $\sum \mathrm{d}x$, $\sum |\mathrm{d}x|$, $\sum \mathrm{d}y$, and $\sum |\mathrm{d}y|$ filtered by Gaussian ($\sigma = 3.3\,\mathrm{s}$) are summed separately in each subregion. Therefore, the descriptor vector of a keypoint is a $16 * 4 = 64$ dimension which is recorded SURF-16. Several extended versions of SURF have been developed according to the size of the neighbourhood, such as SURF-72 and SRUF-144. SURF is invariant to scale, rotation, illumination, and faster than SIFT [32].

2.5. KAZE Features. KAZE is an improved version of SIFT proposed by Alcantarilla et al. Aiming to avoid blurring edges and details lost in linear scale space that is constructed by the Gaussian filter, nonlinear scale space is created by the nonlinear diffusion filter. To FAST, AOS (additive operator splitting) is exploited to accelerate calculation [54]. The nonlinear diffusion filter is described by

$$\frac{\partial L}{\partial t} = \operatorname{div}(c(x, y, t)\nabla L), \tag{6}$$

where t is a scale factor and div is the conduction function corresponding to the Gaussian filter used in SIFT. Three conduction functions have been used in different details and remaining are present as

$$c(x, y, t) = g(|\nabla L\sigma(x, y, t)|)$$

$$g1 = \exp\left(-\frac{|\nabla L\sigma|^2}{k^2}\right),$$

$$g2 = \frac{1}{1 + |\nabla L\sigma|^2/k^2},$$

$$g3 = \begin{cases} 1, & |\nabla L\sigma|^2 = 0 \\ 1 - \exp\left(-\dfrac{-3.115}{\left(|\nabla L\sigma|^2/k\right)^8}\right), & |\nabla L\sigma|^2 > 0 \end{cases}, \tag{7}$$

where $g1$ reserves the edge of high contrast first, $g2$ reserves the area with large width first, and g3 can effectively smooth the interior of the area and retain the boundary information. KAZE can detect more keypoints than SIFR from theory. An improvement vision of KAZE uses advanced numerical schemes called fast explicit diffusion (FED) embedded in a pyramidal framework to highly speed up feature detection in the nonlinear scale spaces [64].

2.6. Learnable Detectors

2.6.1. The Background of Learning-Based Detectors. Deep learning has achieved rapid developments in computer vision and image processing such as object detection, image identification, image classification, and image enhancement. Deep learning can be used in the method of image registration that is classified into the intensity-based method and the feature-based method. In the classical deep neural intensity-based method, a general solution is that the deep learning is used as an iterator to optimize the loss function between the reference image and the floating image to estimate the transformation function. When the loss value reaches the required range, the transformation matrix is obtained [19, 65–70]. For FAST, reinforcement learning and supervised transformation have also been adopted to speed up the convergence [6, 60, 71–74]. To improve the invariance to deformation, semi- and self-supervised learning is also attempted using GANs and autoencoder [75, 76]. However, intensity-based methods are unsuitable for large displacement problems which are handled by feature-based methods.

Learning schemes have been used in feature-based image registration to detect features, describe features, and to estimate transformation between images. The FAST detector firstly uses machine learning techniques to classify a pixel point into a corner point or not without constructing descriptors [26]. For the high repeatability, the simulated annealing algorithm has been adopted in optimization. It is worth noting that the corners detected using this learning algorithm depends on the training data, which could not cover all possible corners. With the developments of deep learning, Ma et al. have reviewed and proved that CNNs are the mostly used deep net architecture in feature detection, description, and matching in comparison with other models [49, 61, 77, 78]. The principle of the deep learning-based detector is to construct a response map and then search keypoints in it, which is trained in a differentiable manner and under the geometric transformation constraints between images [49, 61, 78]. This type of method can be classified into supervised, self-supervised [42, 43], or unsupervised methods [39]. In this section, we introduce representative learning-based detectors and sketch their main principles.

2.6.2. DetNet: Learning Feature Covariant. Lenc and Vedali propose the method of unsupervised learning local covariant feature detectors [23]. It claims that all common and many uncommon detectors can be derived theoretically and can be automatically learned with a covariance constraint under geometric transformations by the regressor. This paper shows that different detectors can be characterized by which transformation they are covariant with. This work learned two complementary types of detectors: a corner detector and an orientation, corresponding to the variance constraint of equations (8) and (9):

$$\min_{\psi} \frac{1}{n} \sum_{i=1}^{n} \left\| \psi(T_i x_i) - (x_i) - T_i \right\|^2, \tag{8}$$

where (x_i, T_i) are example patches and transformations and the optimization is over the parameters of a deep neural network ψ.

$$R_2^T R_1 = R, \tag{9}$$

where $h_i = (R_i, 0)$ are the rotations estimated by a deep neural network ψ.

This paper compares the learned detectors to FAST [26], the difference-of-Gaussian detector, SIFT [31], Harris corner detector [25], and Hessian point detector [32]. The author claimed that the trained "corner detector" network called DetNet clearly outperformed the other methods at one scale, and the rotation ROTNET was sensibly better than the SIFT orientation detector. Although, the work only accomplished the training network of single and primitive model detector, such as translation and rotation, that performed well, for the actual application, detecting feature of complex transformation, and detecting multiple features in a patch, a lot of work needs to be done to this basic method. Another unsupervised learning method is proposed by Luo et al. and named ASFeat that explores local shape information to learn to detect the feature points accurately [79].

2.6.3. TILDE: A Temporally Invariant Learned Detector. Veredie et al. proposed the temporally invariant learned detector (TILDE), which was designed to detect repeatable keypoints in images with drastic illumination changes, as the

imaging condition is different in times of day, weathers, and seasons [42]. First, images are used to create the training set which was collected by capturing a series of images from outdoor webcams captured at different times of days, over a long period. Then, SIFT was used to detect and locate the position of keypoints that is detected repeatedly in the same position. At last, the training set consists of positive and negative samples. The positive samples were made of the patches from all captured images, simultaneously including the ones where the keypoint was detected or undetected, and centered on the average of the detections, and the negative samples were created by extracting matches far away from the keypoints.

A piecewise linear regressor is trained to predict a score map, whose value is greater than a thresholding which can be identified as a keypoint. In order to distinguish locations close to or far from the keypoint and enforce the repeatability of the regressor over time, the objective function of three terms is defined with classification-like loss L_c, shape regularization loss L_s, and temporal regularization loss L_t. The objective function L is minimized over three terms with parameter ω of the regressor that is written as follows:

$$\underset{\omega}{\text{minimize}}\, L_c(w) + L_s(w) + L_t(w). \tag{10}$$

The results showed that using the piecewise-linear functions' regressor gave consistently more reliable keypoints than the alternative regressor and then known keypoint detectors such as SURF and MSER [32, 80]. However, TILDE only remains a state-of-the-art approach to detection in the presence of illumination changes, but it is limited to situations where only keypoints with a common scale are present.

2.6.4. LIFT: Learn Invariant Feature Transform. Yi et al. attempted to learn detection, orientation estimation, and feature description in a unified pipeline that consists of three convolutional neural networks (CNNs) trained individually in the reverse order but performed well in order as the different CNNs try to optimize for different objectives [43]. They claimed that LIFT can be regarded as a trainable SIFT and outperforms the state of the art with good generalization properties. The training procedure learned the descriptor first, then the orientation estimator for the descriptor, and finally the detector. In this section, we first introduce the learning approach for the detector, and the learning for orientation estimation and descriptor will be sketched in Section 3. LIFT is an improvement of TILDE that is learned to robustly detect features in spite of illumination changes. However, the learning only carries on a dataset without viewpoint and scale changes. The first improvement involves creating the training dataset by collecting image sets with viewpoint changes such as Piccadilly Circus in London and the Roman Forum in Rome from [81]. The second improvement involves adopting the softargmax function, defined in equation (11), to locate the feature point in score map S, which lets maximum be found other than fixed by SfM (structure-from-motion). Then, the patch $p = \text{Crop}(P, x)$ is cropped and then used as an input to the orientation estimator:

$$x = \text{softargmax}(S) = \frac{\sum_y \exp(\beta S(y))\, y}{\sum_y \exp(\beta S(y))}. \tag{11}$$

The third improvement involves adopting Siamese training architecture with four branches, which takes as an input a quadruplet of patches (P^1, P^2, P^3, P^4) and minimizes the redefined loss function that is the sum of $L_{\text{class}}(P^1, P^2, P^3, P^4)$ and $L_{\text{pair}}(P^1, P^2)$, where P^1 and P^2 are the positive samples, P^3 is a negative sample with different salient features, and P^4 is only used as a negative example with no salient features to train the detector.

The method does not learn the scale invariance of the detector in the training process. In runtime, the repeatability of the detector to multiscale is attained by applying at different resolutions of the image to obtain score maps in scale space. Although the method proposed an effective strategy to train each component individually, resulting in running jointly, the further objection is to look into performing the method over the whole image instead of pre-extracted patches. The scale invariance of the detector does not learn in training.

2.6.5. Multiscale Detector: A Learning-based Method to Detect Multiscale Keypoints. Hani A [35] proposed a method composed of two independent networks: detection network and description network. Two networks are trained independently in advance. The detection network is trained to detect keypoints, and the description network is trained to match the keypoint and give the description. The process is similar to the traditional image registration method based on keypoint. In this section, we only reviewed the part of network trained for the detector.

Unlike the LIFT, the ability of detecting multiscale keypoints is learned in the network instead of runtime. Two main tasks have been done, one is to establish a multiscale database and the other is to establish a multiscale learning-based detection network. A large set of datasets are extracted from the 3D model using Structure-from-Motion (SfM) [82] to identify good keypoints and to generate matching patches in five scales: $S = \{64, 96, 128, 192, 256\}$. The generated set of patches P is denoted as $P = \{p_i: (x_i, s_i, k_i); k \in \{-1, 1\}\}$, where p_i is a patch with x_i as the raw pixels, s_i indicates the scale of the patch, and p_i is the label. Detection network learns a nonlinear function $f(x)$ that is capable of identifying whether a patch contains a keypoint. The framework of this network is presented in Figure 2 which consists of a sequence of convolutional and pooling layers, a scale-dependent branching mechanism that is shown in Figure 2 is indicated by blue arrows, and then followed by two fully connected layers for classification. Furthermore, in this work, a large-scale dataset was created by extracting patches from a 3D model using structure-from-motion.

2.6.6. SuperPoint: Self-Supervised Interest Point Detection and Description. Unlike LIFT, SuperPoint constructs a self-supervised fully convolutional framework that implements the full interest point matching pipeline, that is, detection,

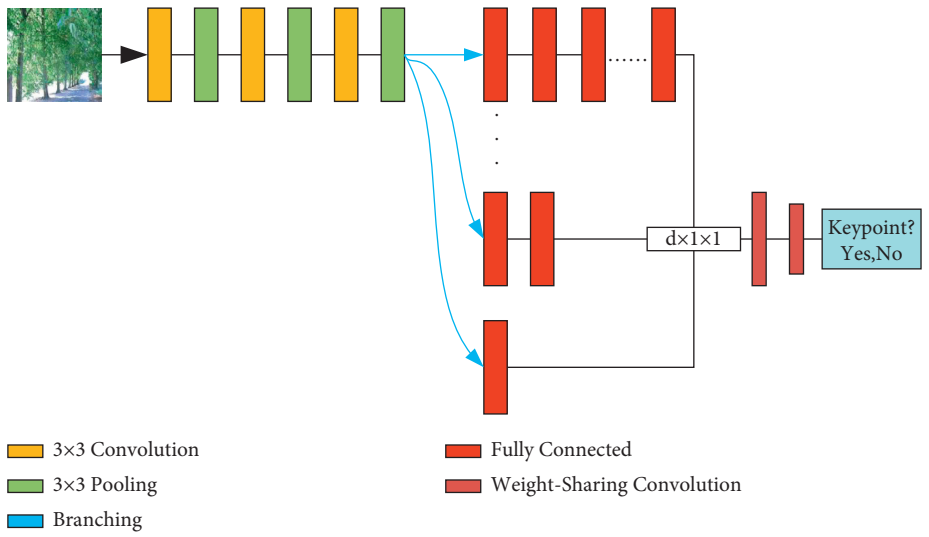

Figure 2: The feature point detection network [83].

description, and matching [7]. Different from LIFT, this method performs on full-sized images to computer interest point at pixel level and associated descriptors in one forward pass instead of relying on preextracted patches. Furthermore, a large dataset with pseudoground truth interest points is defined and supervised by the detector itself instead of human annotation. First, to overcome the ambiguity in the location of interest point, synthetic dataset is created from simple geometric shape with accurate interest point locations such as Y-junctions, L-junctions, T-junctions, centres of tiny ellipses, and the end of line segments. The synthetic dataset is used to train the base detector. Then, to boost the detector repeatability on the natural image with large viewpoint changes, the homographic adaptation is designed to enhance the geometric of the detector, which denotes that a keypoint can be detected in images undergoing various geometrical transform. The final training dataset is created on COCO images through the technique of homographic adaptation [8]. The image and corresponding keypoints are transformed by the homographic adaptation and then aggregated to generate the needed dataset. The last jointly training is gone on a fully convolutional neural network with two branches that compute the location and description vector of interest points in a single forward pass with image pair as the input. It is presented in Figure 3.

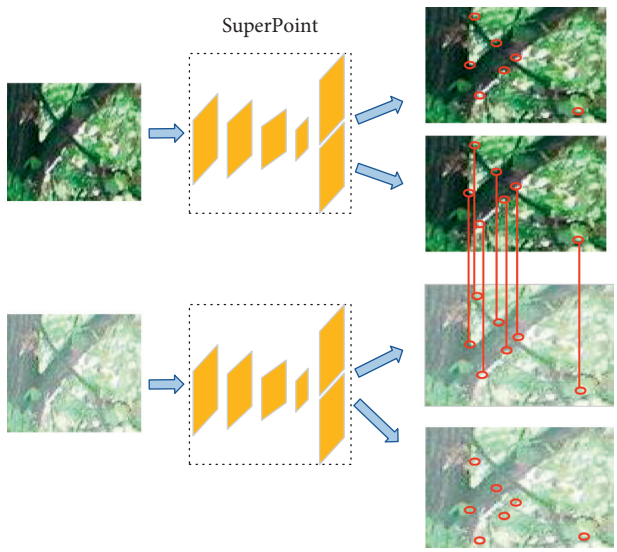

Figure 3: SuperPoint joint training. Train a fully convolutional network that jointly extracts interest points and descriptors from an image.

3. Feature Description

After detecting, the remaining constructs an appropriate descriptor, characterizing and discriminating the detected region. Many techniques have been developed for the task. The simplest descriptor is the numerical vector of the local feature region. However, it is time-consuming and sensitive to view transformation. Histograms of pixel intensity, gradient, and orientations have been used to construct descriptor vectors resulting in the invariance to geometric deformation and illumination [16, 25, 27–32]. To speed it, the binary descriptor is explored by comparing pixel pairs sampled from the local feature region [26–30]. In this section, we review descriptors from the handcrafted to the learnable and sketch the basic principles.

3.1. Local Gradient-Based Descriptor: SIFT, SRUF, and KAZE. Gradient-based descriptors have been widely used in image registration as their effectiveness and invariance to the variance of lighting, rotation, and scale. The most representative gradient-based descriptors are SIFT, SURF, and KAZE and improvements associated with them, which also are designed jointly with detectors at the same time [32, 33, 44].

In SIFT, descriptor vectors are constructed on the same scale space to the detector, and the local area of detected keypoint, within a local circular region, is divided into

$4 * 4 = 16$ nonoverlapping subareas which support formation of the final descriptor. In each subarea, each orientation for 8 pixels is calculated, and a histogram of gradient orientation is constructed according to 16 subareas with 8-bit bins; then, normalization is performed for the invariance to illumination. SIFT has achieved remarkable performance from the detection and the description compare with other handcrafted methods. However, the whole processing is time-consuming. An improvement vison, namely, SURF, explores integral images to replace the Hessian matrix and use the wavelet response to approximate gradient computation, for saving computing. The descriptor, a 64-dimension vector, constructed a square local area of a detected keypoint that is divided into 16 subsquare regions independently. There are various versions of SURF according to the size of the neighbourhood such as SURF-72 and SRUF-144. SURF is not only invariant to scale, rotation, and illumination but also faster than SIFT in detecting and matching by using integral images in convolution and using wavelet responses to accomplish orientation assignment, regardless of the scale [30].

KAZE is an improved vision of SIFT that constructs the nonlinear scale space to detect the keypoints avoiding blurring of the edges in the process of the filter. For a detected feature point, the descriptor is constructed on a rectangular area of $24\sigma_i \times 24\sigma_i$ centered on the subregion that is divided into 4×4 subregions of size $9\sigma_i \times 9\sigma_i$ with an overlap $2\sigma_i$. For each subregion, $dv = (\sum L_x, \sum L_y, \sum |L_x|, \sum |L_y|)$ is calculated, then weighted using a Gaussian ($\sigma_2 = 1.5\sigma_i$), and then rotated according to the dominant orientation. Finally, the descriptor vector of length 64 is normalized into a unit vector to get the invariance to illumination [32]. A modified-local difference binary (M-LDB) descriptor is proposed to highly speed up feature description in nonlinear scale space. It is a scale and rotation invariant and has low storage requirements [84].

3.2. BRIEF: Binary Robust Independent Elementary Features. Inspired by the FAST detectors, the BRIEF descriptor proposed by Michael Calonder aims to speed up matching and reduce memory consumption. Firstly, it uses a short binary string to present the local feature region. The descriptor is simple that matches fast for calculating the Hamming distance [27]. Many approaches speed up the feature description and matching by reducing dimensionalities such as PCA (principal component analysis) [85] and LDA (linear discriminant embedding) [86], designing a short descriptor to replace the original such as SURF for SIFT [32] or directly binarizing descriptors such as the GIST binarizing an entire image [24]. Compared with these methods, BRIEF builds short descriptors by comparing the intensity of pixel pairs sampled from the neighbourhood around the candidate corner.

It contains sampling patterns, smoothing patches, and testing the response of sampling-point pairs to build a binary vector. For a corner point, a local patch centered on itself is used as the random sampling space to sample point pairs p_1 and p_2. A binary descriptor vector is encoded by comparing the integrity of two points according to the following equation:

$$\tau(p; x, y) = \begin{cases} 1, & p_1(x, y) < p_2(x, y), \\ 0, & \text{others}, \end{cases} \tag{12}$$

where $p_1(x, y)$ and $p_2(x, y)$ are the sampling. The binary descriptor is constructed by n_d comparing results according to equation (4). In the original work of the proposer, experiments showed only 256 bits or 128 bits are good enough:

$$f_{n_d}(p) = \sum_{1<i<n_d} 2^{i-1}\, \tau(p; p_1, p_2). \tag{13}$$

As the first binary descriptor proposed, BRIEF does not involve a detailed pattern and compensation measures so that the detector is simple and sensitive to noise, orientation, and scale [62].

3.3. BRISK: Binary Robust Invariant Scalable Keypoints. Precision and speed are the eternal pursuits of state-of-the-art feature detection and description. The BRISK constructs multiscale space for detecting and design sampling patterns for orientation, which gets the invariance to scale and rotation [28]. First, the scale-space pyramid is constructed by downsampling the original image c_0 with the 4 octaves c_i and 4 intraoctaves d_i each of which locates in-between layers c_i and c_{i+1} (shown in Figure 4). FAST 9-16 detectors are applied on each octave and intraoctaves separately to identify potential corner points. To find continuous image saliency not only in-plane but also in the scale dimension, using 2D quadratic to interpolate in patches of three-layer resulting in subpixel position and using 1D parabola to interpolate along scale axis resulting in scale refinement. Nonmaxima suppression is conducted in local regions to achieve the invariance to noises.

Second, the sampling pattern is different from the BRIEF where sampling-point pairs are located equally on circles concentric with the keypoint and smoothed by Gaussian kernel to avoid aliasing effects. Figure 5 shows the position of 60 sampling points of a corner point. Furthermore, to achieve invariance to rotation, the pattern direction of the corner point is defined as equation (14). $I(p_i, \sigma_i)$ and $I(p_j, \sigma_j)$ represent the intensity of sampling-point pairs:

$$g = \begin{pmatrix} g_x \\ g_y \end{pmatrix} = \frac{1}{L} \cdot \sum_{(p_i, p_j) \in L} g\left(p_i, p_j\right), \tag{14}$$

where $g(p_i, p_j)$ is the local gradient and L is the subset of long-distance point pairs, which is defined as equations (15)–(17):

$$g\left(p_i, p_j\right) = \left(p_i - p_j\right) \frac{I\left(p_j, \sigma_j\right) - I\left(p_i, \sigma_i\right)}{\left\| p_i - p_j \right\|^2}, \tag{15}$$

$$L = \left\{ \left(p_i, p_j \in A\right) \middle| \left\| p_i - p_j \right\| > \delta_{\min} \right\}, \tag{16}$$

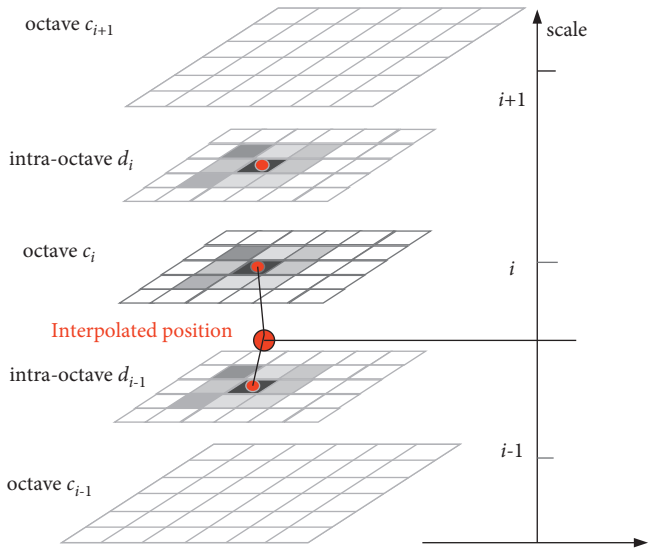

FIGURE 4: An interesting point is identified in the octave by comparing 8 pixels of a neighbourhood c_i as well as the corresponding patches of the immediately adjacent layers above and below.

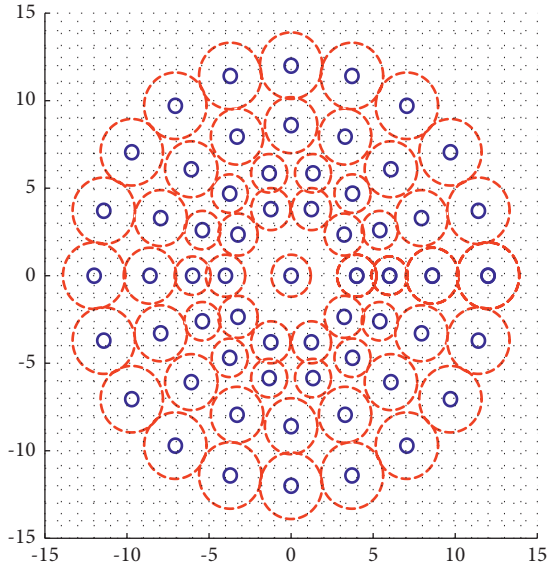

FIGURE 5: The BRISK sampling pattern with $N = 60$ points. The small blue circles denote the sampling locations. The red dashed circles are the Gaussian kernels that are used to smooth the neighbourhood.

$$S = \left\{ \left(p_i, p_j \in A \right) \middle| \left\| p_i - p_j \right\| < \delta_{\min} \right\}. \tag{17}$$

While the intensity difference between p_i and p_j is greater than $\delta_{\max}$, the pair of p_i and p_j is classified into the long-distance subset L. Correspondingly, the short-distance subset is defined as $S = \left\{ (p_i, p_j \in A) \middle| \left\| p_i - p_j \right\| < \delta_{\min} \right\}$. A is the set of all sampling-point pairs. Rotating the sampling pattern by $\alpha = \arctan2(g)$ around the corner point, each bit b constructs a binary descriptor generated by comparing the intensity difference of point pairs $(p_i^\alpha, p_j^\alpha) \in s$ just as follows:

$$b = \begin{cases} 1, & I\left(p_j^\alpha, \sigma_i\right) > I\left(p_i^\alpha, \sigma_i\right), \\ 0, & \text{otherwise}, \end{cases} \quad \forall \left(p_i, p_j \right) \in S \tag{18}$$

Finally, the BRISK descriptor encoded as 512 bits is an improved BRIEF [27]. In the actual registration work, the FAST detector and the BRIEF descriptor are used jointly. However, the performance of the invariance to noise is not considered in the original paper, which will be demonstrated in experiments afterward.

3.4. ORB: Oriented FAST and Rotated BRIEF Features. ORB is the combination of the FAST detector and the improved BRIEF descriptor [30], which is viewed as the replacement of SIFT and attain better efficiency by binary description and can be used in real-time application [30]. For image pairs in good image conditions, it also shows remarkable performance. The improvement is the addition of orientation invariance in descriptions. ORB exploits the intensity centroid to measure corner orientation. The centroid of a feature point patch Ip is calculated by the moments presented as follows:

$$\forall p, 1 \in \{0, 1\}: m_{pq} = \sum_{x,y} x^p y^q I(x, y), \tag{19}$$

$$c = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\right), \tag{20}$$

$$\theta = \arctan2(m_{01}, m_{10}). \tag{21}$$

Equation (19) is the moments of a feature patch. Equation (21) is the centroid of a feature patch. ORB makes the coordinates (x, y) of a keypoint multiplied with the patch's orientation to attain the invariance to rotation:

$$(x, y)_\theta = \theta * \begin{pmatrix} x \\ y \end{pmatrix}. \tag{22}$$

The low variance makes saliency indistinct in the matching. To recover the loss of variance and reduce correlation, a greedy search has been used to de-correlating which leads to better performance in matching [29]. ORB has fast computation speed and high matching accuracy, which make it an alternative to SIFT and SURF. Compared with the BRIEF, ORB uses the statics moments to attain the invariance to the rotation, not the sampling pixel pairs.

3.5. FREAK: Fast Retina Keypoint. Proposed by Alexandre Alahi and Raphael Ortiz and inspired by the HUS (human visual system), FREAK is a binary descriptor involving retinal sampling pattern similar to retinal ganglion cells, different sizes of overlapping receptive fields, coarse-to-fine descriptor, and orientation mechanism that is similar to BRISK [29].

Firstly, FREAK is different from BRIEF and BRISK in sampling patterns. BRIEF uses a circular sampling pattern randomly to sample point pairs by an isotropic Gaussian distribution, resulting in point pairs which are fairly distributed from circles concentric. However, FREAK uses the retinal sampling grid to the sample that mimics the mechanism of the human visual system. Meanwhile, the different sizes of the Gaussian kernel are used to smooth the neighbourhood of the sampling point. The sampling pattern is shown in Figure 6. The red circle indicates a receptive field. Secondly, FREAK selects the point pairs with low correlation and high variance from the support region. 512 pairs are enough to describe a keypoint, which has been verified by the author. Finally, the mimic HVS saccadic searching discards 90% of the candidate points representing coarse

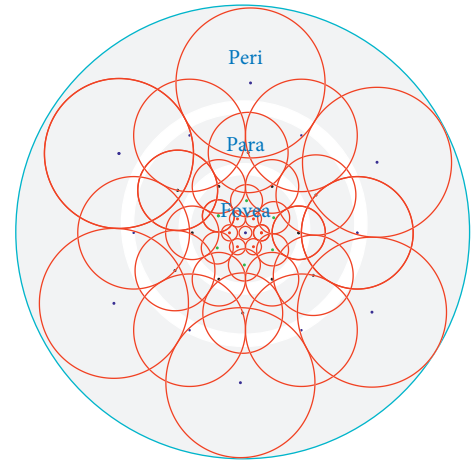

Figure 6: The FREAK sampling pattern of a keypoint that simulates the retinal receptive structures.

information, which dramatically improved matching efficiency. The orientation of FREAK is estimated by local gradients selected from receptive symmetrically, which is different from BRISK and BRIEF. FREAK is suitable for image alignment with the difference in scales, orientations, and noise to some extent.

3.6. Learnable Descriptors

3.6.1. Descriptor Based on a Pretrained Network. When the loss value reaches the required range, the transformation matrix is obtained [12, 13, 65, 70, 87]. For the FAST, reinforcement learning and supervised transformation have also been adopted to speed up the convergence [60, 71–73]. In these methods, some methods use pretrained networks to extract the feature point and then use conventional methods to match. Other methods train a specified network to attain the deformation filed between image pairs [34, 54, 59, 88]. However, these methods resolve the image registration from the global integrity similarity or partly from feature points, not totally from detectors and descriptors.

In this section, we introduce deep learning techniques that are used to detect keypoints and describe them in the image registration. In [45], Yang et al. proposed a model based on the pretrained deep network VGG16. It extracts patches from layers of pretrained VGG network as feature vectors and uses the MLESC algorithm to match patches between two images. The vector of the patch is as the description without considering illumination, rotation, and noise. So, the approach is only feasible for the image pairs with differences in translation. The author's experiment has proved this. We experimented on image pairs different in angle from the public dataset and presented results in Figure 7. As shown in Figure 7(c), the method fails, so it is not suitable to align the image pairs with the variance of rotation.

3.6.2. Siamese Descriptor: Discriminative Learning of Deep Convolutional Feature Point Descriptors. Simo-Serra et al. [37] have also proposed a simple schematic of a Siamese

(a) (b) (c)

Figure 7: Yang's method fails to align the image pairs with a rotational difference: (a) reference image, (b) float image, and (c) the failure result.

network consisting of two same branches of convolutional neural network, to learn the discriminating representation of a local patch. The work claimed that the learned network can generate a 128-D vector to describe corresponding feature point discriminatively, which can be used as credible alternative to SIFT. However, the structure of the Siamese network is very simple, which is a three-layer network, as shown in Figure 8. Using the L2 norm as the similarity metric between two vectors, the hinge loss $l(x_1, x_2)$ is defined as follows:

$$l(x_1, x_2) = \begin{cases} \|D(x_1) - D(x_2)\|_2, & p_1 = p_2 \\ \max\left(0, C - \|D(x_1) - D(x_2)\|_2\right), & p_1 \neq p_2 \end{cases} \tag{23}$$

To achieve a descriptor making noncorresponding patches far apart and corresponding patches close together, the major innovation of this work is reflected in the sample selection in the training process, and a method of difficult sample mining is proposed. Randomly selected negative samples easily make $l(x_1, x_2)$ equal to zero, which leads the training become ineffective. At each epoch, after the patch pairs from a set of s_n through the network and computing their loss, a subset of them with the loss littler than the specified hard threshold s_n^H is backpropagated through the network to update the weights. Similarly, for positive samples, difficult samples are those that are similar blocks, but the distance is large. In the training process, when the distance is larger than the hard threshold s_p^H, the data are retained to continue training the network. The proposed novel method on mining of both positive and negative obtained large performance in patch description.

3.6.3. Triplet Network: Descriptor Consisting of the Siamese Network. Hani A proposed a network architecture, for multiscale detection, and another triplet Siamese network architecture for keypoints' description [35]. The learning-based detector has been described in Section 2.6.5. The descriptor network consisting of a triplet Siamese network learns a function that can decide whether feature pairs match or not. The framework of this network is shown in Figure 9. The anchor patch, the positive patch, and the negative patch are fed through the same convolutional network to compute its embedding feature vector, respectively; then, two Euclidean distances are computed among three vectors. To train the keypoint description network, loss functions are defined as equation (24) to decide whether two patches can be matched:

$$L_T = \frac{1}{N} \sum_j \left[\max\left(0, \|f(p_a) - f(p_b)\|_2 - \|f(p_a) - f(p_b)\|_2 + h\right)\right]. \tag{24}$$

Corresponding experiments proved the effectiveness of Hani A's method. His experimental results show that it outperforms DeepCopare [89], MatchNet [38], and method with deep learning as an iterator. For image pairs with large viewpoint differences, this method becomes more and more ineffective with the increase of the difference. The explanation of author is that this type of images differs largely to the training dataset. However, this work still shows the promising results of the deep network in descriptors for image registration.

3.6.4. Descriptor of LIFT: Learned Invariant Feature Transform. LIFT is a novel deep architecture that involves the detector, the orientation estimator, and the descriptor. In practice, it is impossible to train a full architecture for each component with different objectives. As the learning-based detector has been introduced above, the remaining two items are introduced in this section. The descriptor of LIFT is an improved vision of [43]. To achieve invariance to different perspectives, the sample patch pairs including the patch p and the rotated patch p_θ are added in training dataset. p_θ is generated by the method of structure of motion. Siamese architecture trained for the descriptor consists of three branches each of whom is the same to [43], which takes as an input a triplet patches: p_1 and p_2 are positive samples which are from the same physical point and p_3 is the negative sample that is from a different 3D point. The parameters ρ of network are learned by minimizing the sum of loss for patch pairs (p_θ^k, p_θ^l) is defined as

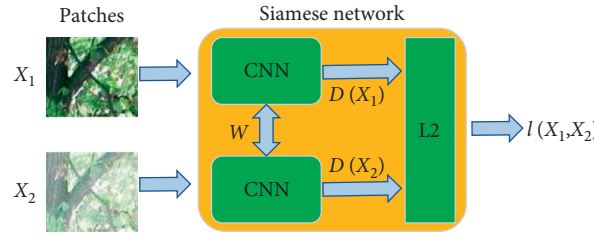

Figure 8: Schematic of a Siamese network, where pairs of input patches are processed by two copies of the same CNN.

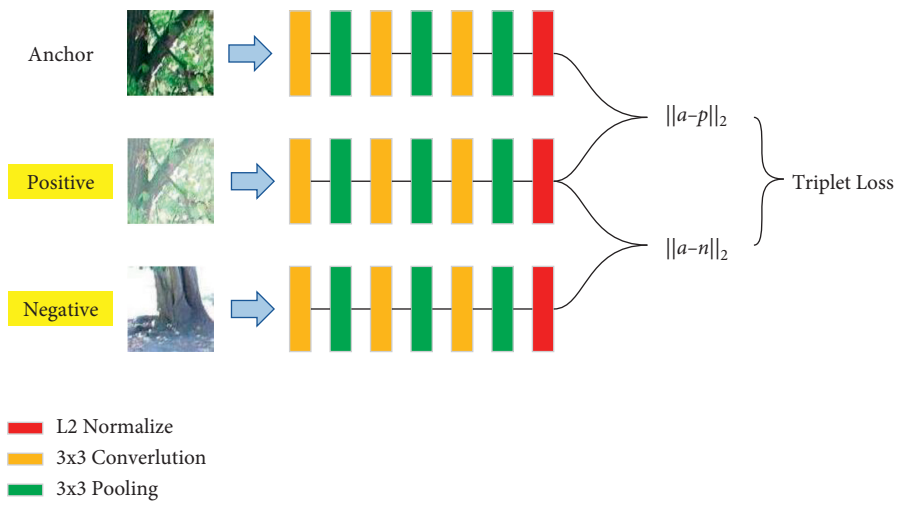

Figure 9: Deep network architecture of the keypoint description triplet network. Three patches are passed through channels that share weights to rank their Euclidean distance in the feature space [83].

$$L_{desc}\left(p_{\theta_1}^k, x_\theta^l\right) = \begin{cases} \left\| h_\rho\left(p_\theta^k\right) - h_\rho\left(p_\theta^l\right)\right\|_2, & \text{positive pairs,} \\ \max\left(0, C - h_\rho\left(p_\theta^k\right) - h_\rho\left(p_\theta^l\right)_2\right), & \text{negative pairs.} \end{cases} \tag{25}$$

After training descriptor network, training the orientation estimator to provide the orientation by minimize the distance $\text{Lorientation}(p^1, x^1, p^2, x^2)$ between description x^1, x^2 vectors for different views of the same 3D points:

$$\text{Lorientation}\left(p^1, x^1, p^2, x^2\right) = \left\| h_\rho\left(G\left(p^1, x^1\right) - h_\rho\left(G\left(p^2, x^2\right)\right)\right)\right\|_2. \tag{26}$$

Finally, the trained descriptor and orientation estimator are used to train the detector for further performance. Therefore, this work proposed an effective strategy to mesh them into a unified network that can be trained end-to-end at the last step.

3.6.5. SuperPoint: Self-Supervised Interest Point Detection and Description. SuperPoint [38] constructs a self-supervised framework to train interest point detectors and descriptors, which is suitable for a large number of multiple-view geometry problems in computer vision [39]. Constructing synthetic dataset to train the base detector network called Magic point and adopting homographic adaptation to boost the generality of Magic point on real images, unlike block-based neural networks such as LIFT, this model operates on full-sized image pairs and extracts the locations of interest points, associated descriptors, and matching results at the time. The last jointly training is gone on double branches' fully convolutional neural network with an image pair as the input. The descriptor network is shown in Figure 10.

The descriptor consists of learned part and nonlearned part. The learned part first outputs a semidense grid of descriptors, and then, in the nonlearned part, it performs interpolation and then L2-normalization to be unit length. The loss is the sum of two intermediate losses: one for the interest point detector, L_p, and one for the descriptor, L_d.

Recently, there have been many research studies on description operators. Most of these methods consist of Siamese, triple, or multibranched convolutional network to learn a nonlinear mapping is represented by a CNN that is optimized to distinguish pairs of corresponding or non-corresponding patches, such as LF-Net [40] and RF-Net [41]. LF-Net exploits train a network to learn a local feature pipeline from scratch in a two-branch setup by confining it

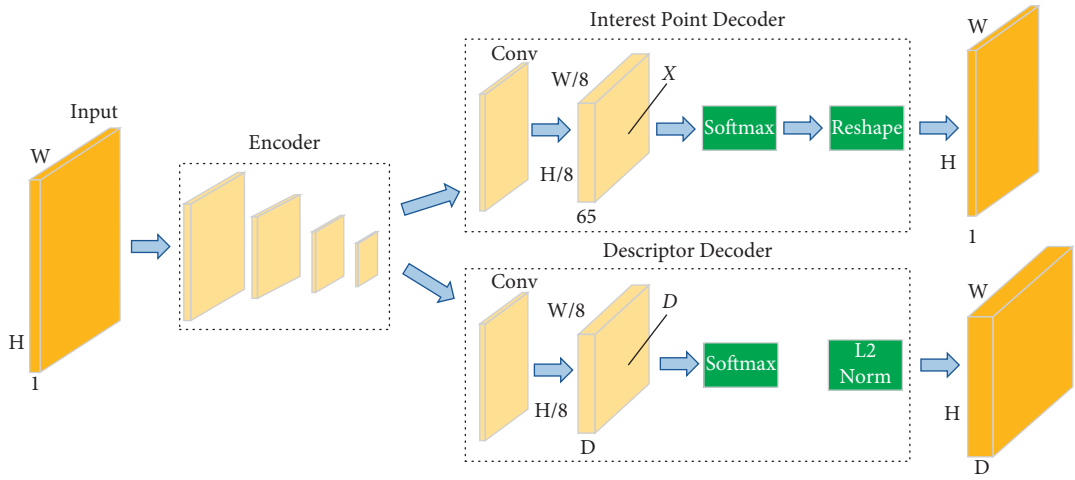

Figure 10: SuperPoint decoders. Both decoders operate on a shared and spatially reduced representation of the input. To keep the model fast and easy to train, both decoders use nonlearned upsampling to bring the representation back to $R^{H\times W}$.

to one branch. The method is trained on both indoor and outdoor dataset, and superior performance is attained than the state of the art on sparse feature matching on both datasets. RF-Net is an improvement of the LF-Net, and the work concentrates on two aspects. First, receptive feature maps preserving low-level scale and high-level scale are constructed to detect keypoints more effectively. Second, introducing a loss function term, neighbour mask facilitates training patch selection.

4. Experiment Settings

4.1. Datasets. We evaluate these detectors and descriptors on two datasets. Namely, the publicly available natural light collection dataset and the specialized multimodal data. Figure 11 shows examples of the public dataset owned by Mikolajczyk [36, 90] (https://www.robots.ox.ac.uk/~vgg/). Five kinds of image transformations are included: compression (Figure 11(a)), illumination (Figure 11(b)), image blur (Figures 11(c) and 11(d)), scale change (Figures 11(e) and Figure 11(f)), and viewpoint change (Figures 11(g) and 11(h)). Each test image sequence contains 6 images. The first image and each subsequent image form a pair of images that is to be aligned. The variance of the image pair is gradually increasing.

The second datasets consist of three multimodal image pairs. The first from Dronehub (https://medium.com/dronehub/datasets-96fc4f9a92e5) is composed of low-altitude visual and thermal aerial images captured by small-scale UAV. The second from Landsat (https://serc.carleton.edu/eyesinthesky2/week11/get_to_know_multispectral_imaging.html) comprises different band images taken by satellite and used to investigate deforestation. The third from SDO (solar dynamics observatory) (https://www.nasa.gov/mission_pages/sdo/main/index.html) and NVST (new vacuum solar telescope) (http://english.ynao.cas.cn/ti/nvst/), composed of different band images, is taken by SDO observatory and NVST observatory, with differences in resolution, rotation, and scale. Figure 12 shows examples from three sets. Figure 12(a) is a sample pair of infrared and visible light images with rotation transformation. Figure 12(b) is a pair of remote sensing images of different wavebands. Figure 12(c) is a pair of heterogeneous multimodal astronomical images.

4.2. The Evaluations. To compare the performance of various detectors and descriptors, we test how well the keypoint can be correctly matched between two images. The accurate matching ratio called precision is used as the evaluation that is defined by equation (27) and described in [85]:

$$\text{precision} = \frac{n_c}{N} = \frac{\text{num of correct matches}}{\text{num of correspondences}}, \tag{27}$$

where N is the number of corresponding point pairs attained in rough matching and n_c is the number of correctly matched point pairs. This precision is calculated by two items. First, matching features N are found from the two input feature sets [31, 47, 90, 91], which is named rough matching. Second, correct matchings n_c are selected from the result of the first constrained with a specified transformation, and outliers are excluded by MSAC (M-estimator sample consensus) algorithm [30, 83].

4.3. Matching Approaches. We divide the experiment into two parts. First, we carried experiment on the same descriptor with different detectors. It is to compare the performance of detectors. Second, we carried experiments on the same detector with different descriptors. It is to compare the performance of descriptors. There are three matching methods, namely, the threshold, the nearest neighbour, and the nearest neighbour distance ratio. The threshold method determines a matching pair only if the distance between them is below a threshold. With this approach, a descriptor has multiple matches. The nearest neighbour approach determines a matching pair only if the distance between them is smaller than a specified threshold and one descriptor is the nearest neighbour to the other. With this approach, a descriptor has one match. The third method determines a

Figure 11: Sample images in the datasets. (a) JPEG compression change, (b) illumination change, (c, d) blur change, (e, f) zoom + rotation, and (g, h) viewpoint change.

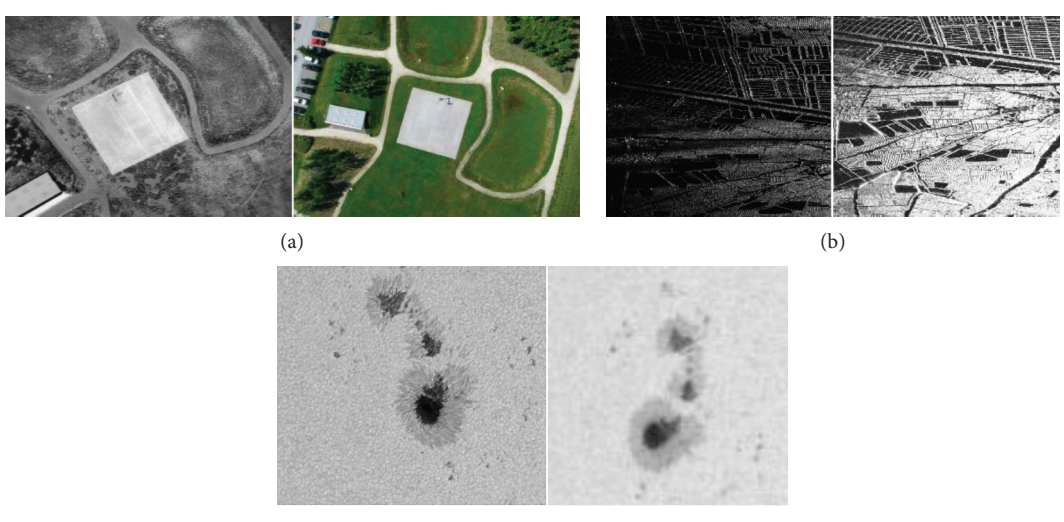

Figure 12: Multimodel image pairs. (a) UAV image pairs. (b) LandSat image pair. (c) Astronomical image pair: NVST and SDO.

matching if the distance ratio between two vectors is low at a threshold, resulting in a descriptor which has multiple matchings. Because of the distinctiveness of descriptors and the application in actuality, we select the nearest neighbour as the correspondence measure in the experiment.

Although various approaches are available for putative feature correspondences matching the representation vector of local area, a simple matching strategy may result in a large number of incorrect matches. Therefore, a robust, accurate, and efficient mismatch elimination method is required to eliminate as many mismatches as possible under a specific geometric constraint. The removing of mismatches is the last significant task in the entire image matching pipeline. Generally, mismatch removal methods can be divided into resampling-based [92, 93], nonparametric model-based [94–96] and relaxed methods [30] and learning for matching [64, 97].

RANSAC and MSAC, as representative resampling-based methods, are often used to eliminate outliers in image registration [30, 83]. Recently, further research on resampling-based methods has been going on [98, 99]. However, few of them have been widely used in image registration, perhaps because of stability, runtime, and other factors [100]. It is the same to other two types of methods. As the paper concentrates on the detectors and descriptors, the introduction of these methods is not developed.

5. Experiment Result

5.1. Comparison Experiment of Detectors. The detectors determine whether a local region is an alternative keypoint. We compare their performances by testing the correct matching precision of their detected feature points. For the fairness of comparison, we use SURF as a descriptor for all detecting experiments and use RANSAC algorithm to eliminate outliers under the specified geometric constraint. The detectors included the corner (Harris), the binary detector (FAST and BRISK), the rotated binary detector (ORB), the scale-invariant detector of the linear scale space (SIFT and SURF), the scale-invariant detector of the nonlinear scale-space (KAZE), DetNet, TILDE, LIFT, multiscale detector, and SuperPoint. We experiment on two sets of datasets. One is the public dataset shown in Figure 11, and the other is the multimodel dataset that is shown in Figure 12.

5.1.1. Results of the Experiment on Public Dataset. Figure 13(a) shows how the matching precision of each detector varies with compression. The curves go down slightly when the compression increasing, which means detectors are all affected by the extent of compression. Similarly, Figure 13(b) shows how the illumination changes affect detectors. As shown in Figure 13(a) and Figure 13(b), the ORB and KAZE perform better than other handcrafted ones in identifying keypoints to the image pairs with variance of illumination changes and compression changes. The learnable detectors including DetNet, LIFT, multiscale, TILDE, and SuperPoint score highly with their curves locating at the top of the graph. TILDE is proposed to detect keypoints reliably under drastic changes of weather and lighting condition. LIFT implements the detector that is similar to the TILDE, orientation estimation, and descriptor in a pipeline. Continuously, the SuperPoint is the further improved version of LIFT by involving synthetic training dataset and homographic adaptation techniques to increase detection accuracy. So, they all perform well and closely.

Figures 13(c) and 13(d) show the precision of each detector varying with blur. When the blur increases, the curves descend rapidly, which means these detectors hardly process the images with a high degree of blur. Figure 13(c) shows that ORB gets precisions that are more than 50% for the textured image pair with blur. Figure 13(d) shows KAZE gets precisions that more than 60% for the structured image pair. We can conclude ORB and KAZE more professional in detecting key point to the changes of blur between images. It is obviously that learned detectors significantly outperform the handcrafted detectors as multiple convolution layers provide detection with more feature space. Especially, the SuperPoint scores outperform other methods as they are trained on hundreds of thousands of images that consist of indoors, outdoors, and synthetic dataset. DetNet, multiscale, TILDE, and LIFT show more stable and invariant than handcrafted detectors to the variance of the blur.

Figures 13(e) and 13(f) show the results of detectors varying with rotation and scale. The curve descends when the rotation and scale increase, which means this detector can hardly deal with images different in scale and rotation. For the structured scene shown in Figure 13(f), the curves go down rapidly and less than 30% at the last pair of images. The worst serious case is Harris which fails to detect local feature regions in the third image pair (the first image and the fourth image is the third pair). In Figure 13(e), the situation is better for the textured scene except for the Harris that fails for all. From Figures 13(e) and 13(f), we conclude that ORB, SIFT, and SURF are more stable and robust than other detectors to changes of scale or rotation. However, the learnable detectors outperform the handcrafted detectors in precession and stability. Only the multiscale detector, it does not consider the rotation in learning process, which results in a low precision compared to other learnable detectors.

Figures 13(g) and 13(h) show the results of different results of detectors vary with viewpoints. The curve shows precisions are all low and descend rapidly. In Figure 10(g), ORB gets the highest score is only 24%. The result is worse in Figure 13(h). The matching fails in dealing with the fourth image pairs in Figure 13(g) and the third image pairs in Figure 13(h). So, it is a big challenge for a single detector to identify keypoints that can be matched correctly from image pairs with viewpoint changes. SuperPoint, LLFT, and multiscale detectors outperform other learnable detectors lying in a synthetic training dataset that can transfer knowledge from a synthetic dataset to real images, using homography adaptation to boost the capability of detectors.From the results in Figure 13, it is obvious that the learned detectors score strongly in matching precision, which confirms findings from both Choy et al. [101] and Yi et al. [32] which show that learnable detectors outperform the handcrafted method.

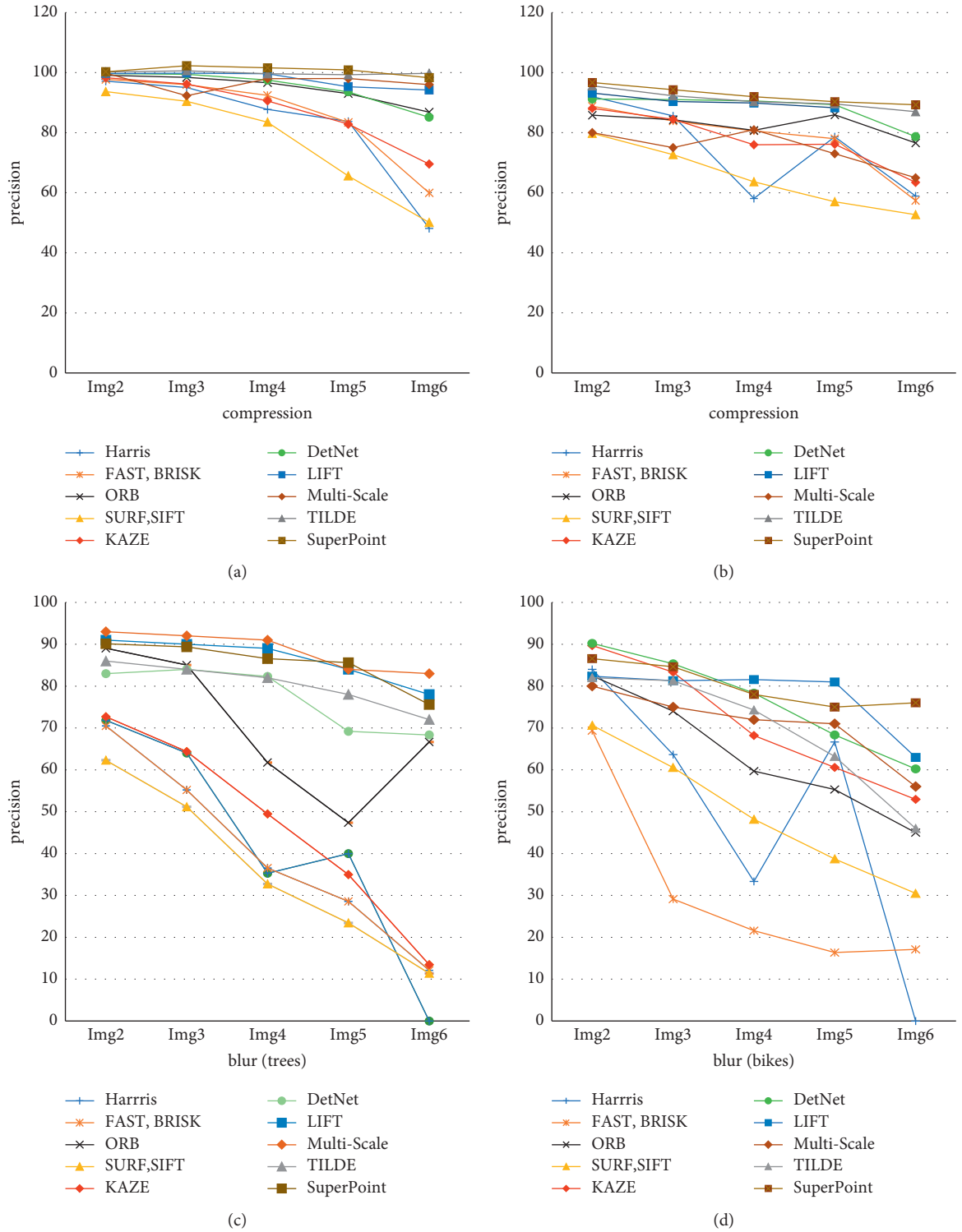

Figure 13: Continued.

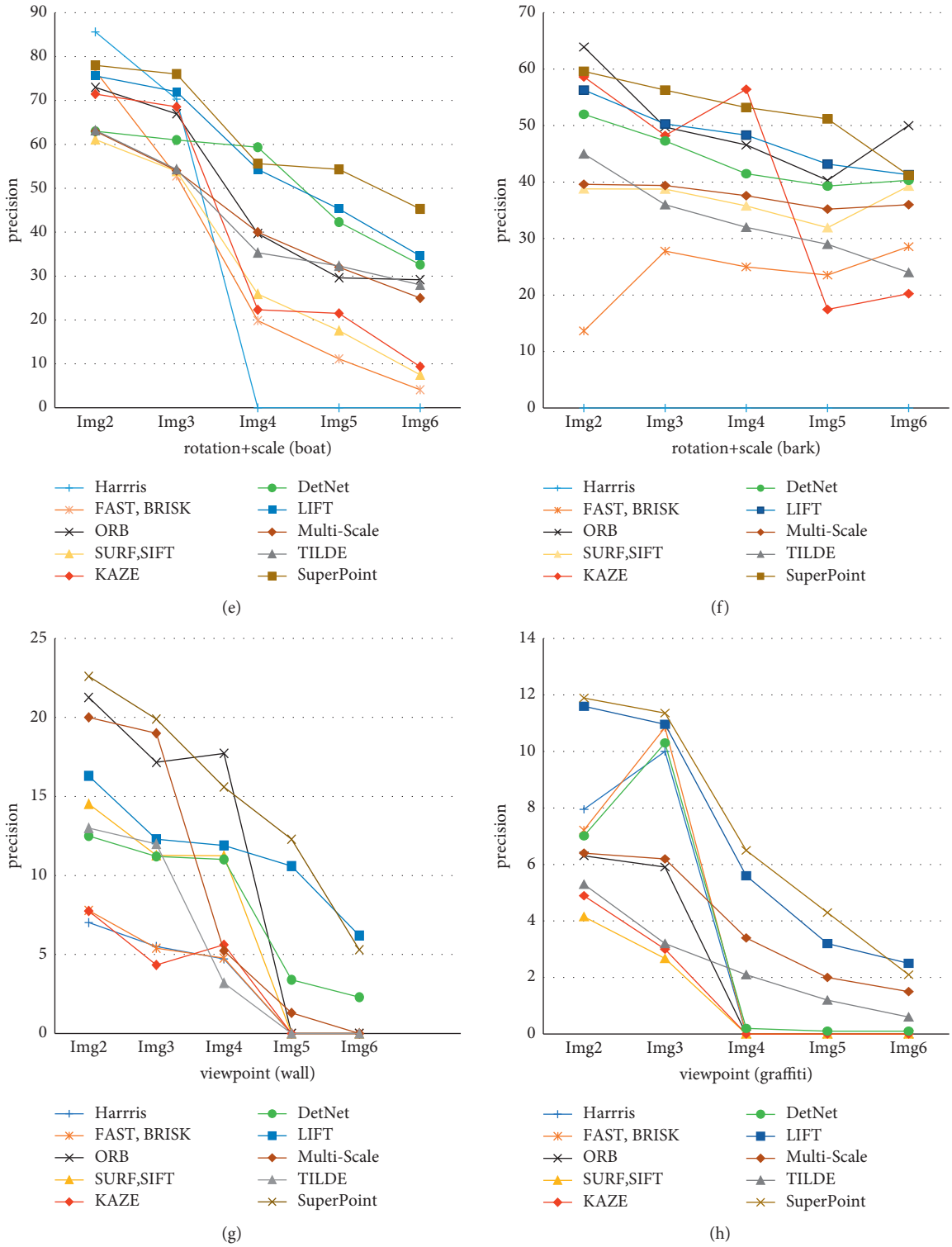

Figure 13: Comparison of different detectors for different geometrical transformations. (a) Results of compression change. (b) Results of illumination change. (c) Results of blur change of the structured scene. (d) Results of blur change of the textured scene. (e, f) Results of rotation and scale change. (g, h) Results of viewpoint change.

5.1.2. Results of Experiments on Multimodal Image. We experiment on three pairs of multimodal images from UAV, Landsat, and SDO and NVST and record precisions presented in Table 2. The differences in angle and scale exist in the UAV image pairs. Except for KAZE, other handcrafted methods are all a fail. The "Fail" in Table 2 means the correct matching keypoints are less than 4, that is, the minimum that is required to estimate the transformation matrix.

Table 2: The comparison of detectors.

Detectors	UAV images	Landsat images	Astronomical images
Harris	Fail	3.70	Fail
FAST	Fail	3.85	Fail
BRISK	Fail	4.05	Fail
ORB	Fail	5.00	Fail
SURF	Fail	10.31	Fail
SIFT	Fail	28.49	2.80
KAZE	5.66	7.42	3.20
DeNet	4.3	23.3	5.4
LIFT	19.3	29.46	17.6
Multiscale	0.6	28.76	3.5
TILDE	2.25	28.36	3.4
SuperPoint	29.3	34.56	21.6

For Landsat Images, all detectors have accomplished enough feature points for image registration. Because the two images of the astronomical image pair have a huge difference in resolution and rotation angle. Except for KAZE and SIFT, other hand-crafted detectors all fail. From Table 2, we conclude that KAZE is the most robust handcrafted detector for multimodel images.

However, all learnable detectors have achieved excellent detection results. For three groups of image pairs, the learnable detectors are all effective to the detector enough feature points to finish the registration. From Table 2, the multiscale detector has a worse performance in detecting image pairs with rotation variance. LIFT applying the rotation estimation and SuperPoint using homography show good generation in detecting actual data.

5.2. Comparison Experiment of Descriptors

5.2.1. Results of the Experiment on Public Dataset. In this section, we evaluate the performance of descriptors to geometrical transformations of different scene types. As reviewed above, BRIEF, BRISK, FREAK, and ORB belong to binary descriptors, SIFT and SURF belong to linear multiscale space, and KAZE belongs to the nonlinear multiscale space. This experiment covers compression, illumination, blur, and viewpoint changes to illustrate the applicability of each descriptor. According to the previous experiment and/or effectively detecting and fairly comparing, we select the ORB as the detector for all descriptors in the matching experiment. So, BRISK, FREAK, BRIEF, SURF, ORG, KAZE, SIFT, Pre-Net, Siamese-Net, Triplet-Net, LIFT, and SuperPoint are evaluated and compared on the ORB's detecting results.

Figure 14(a) shows the matching precision of 7 handcrafted descriptors for JPEG compression changes. The curves gradually decrease with increasing compression, i.e., all descriptors are affected by artifacts. The FREAK obtains the best precision score with the increasing compression extent. The ORB and BRISK get a considerable performance as well. The curve of SIFT drops the largest, which shows this descriptor is even more influenced by artifacts than others. As shown in Figure 14(a), all descriptors can represent the feature points effectively. For the learnable descriptors, only Pre-Net scores the lowest as it is only the vector of a local region from the feature map. SuperPoint, LIFT, Siamese-Net, and Triplet-Net perform well both in image degraded with compression and illumination.

Figure 14(b) shows the results for illumination changes. The image pair is present in Figure 14(b). The curves are very close and decrease slowly, which means all descriptors have a high level of invariance to illumination changes. BRIEF, BRISK, ORB, and FREAK use binary string to represent a local feature region. BRIEF is the first binary descriptor that is simple. BRISK, ORB, and FREAK improved on BRIEF are more invariant to scale and rotation. There is no change of rotation in the image pair shown in Figure 14(b). The descriptors do not need to describe it, so the matching results of BRIEF, BRISK, ORB, and FREAK are close. As SuperPoint and LIFT have trained on an outdoor image sequence that exhibit a huge light change, they score highly and stability to the change in illumination. For learnable descriptors, the performance is largely affected by the amount of training data and their framework. Learnable descriptors all perform well to the illumination change for training on a large number dataset.

Figures 14(c) and 14(d) show results of different descriptors that vary with the blur. The curves of matching precision are all decreased with the increasing blur because blurring low the local feature region distinctive. The binary descriptors get a close result except for ORB with abrupt mutations between the first and the fifth images. For the textured scene, the BRISK obtains the best matching score as its sample pattern can represent more saliency of the local region in Figure 14(c). For the structured scene, FREAK obtains the best matching score because the sample pattern mimicking the human visual system helps to distinguish the local regions. The learnable descriptors all achieved higher scores as their multilevel convolutions that can extract features from multiscale, which helps to identify feature points.

Figures 14(e) and 14(f) show the matching results of descriptors that vary with rotation and scale. We compare all descriptors on the ORB regions. As shown in Figure 14(e), BRIEF gets the lowest score and fails in the second image pair. In Figure 14(f), BRIEF fails too. The results are consistent with the principle of BRIEF because it is the primitive binary string without considering invariance to the rotation

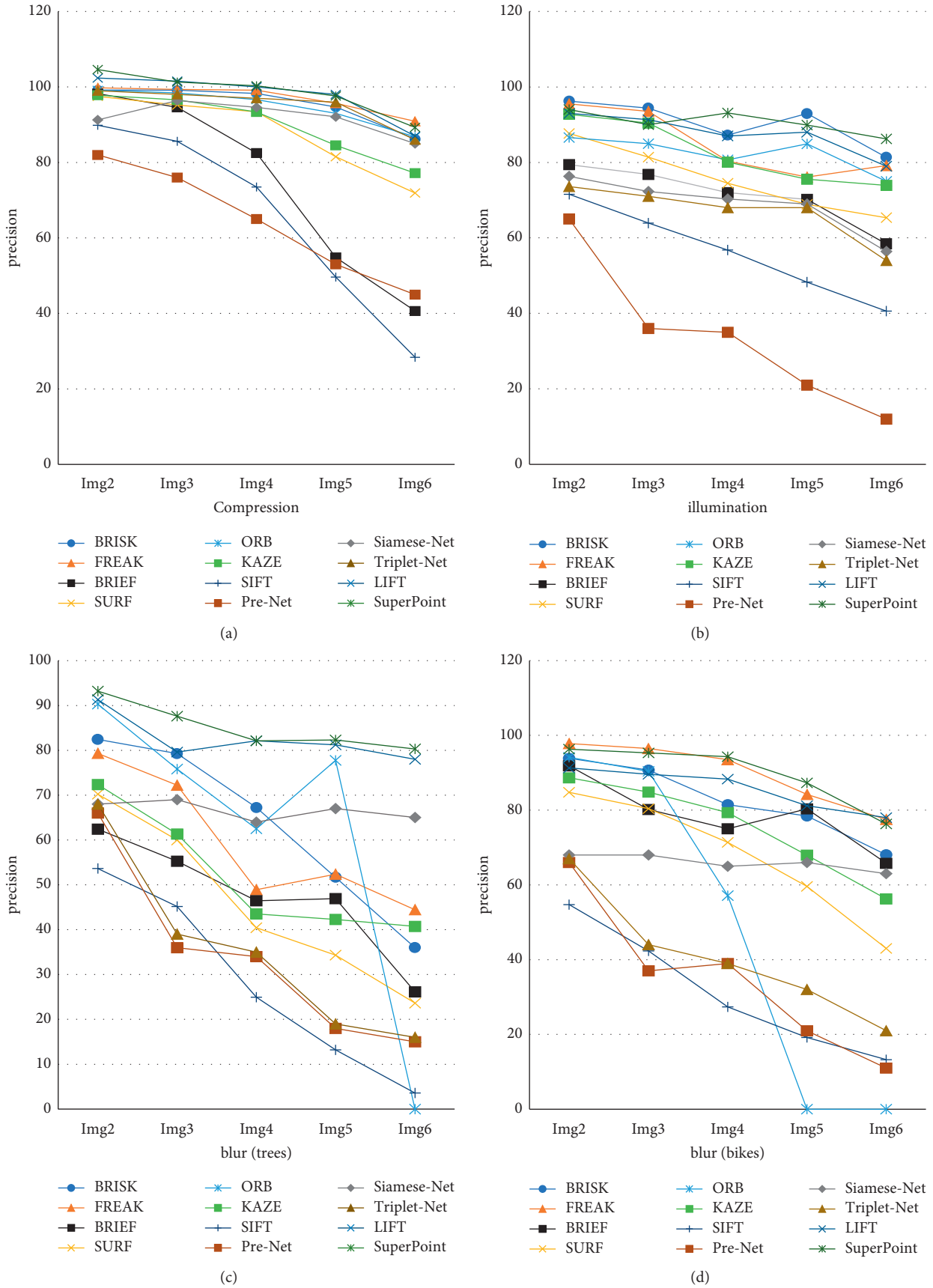

Figure 14: Continued.

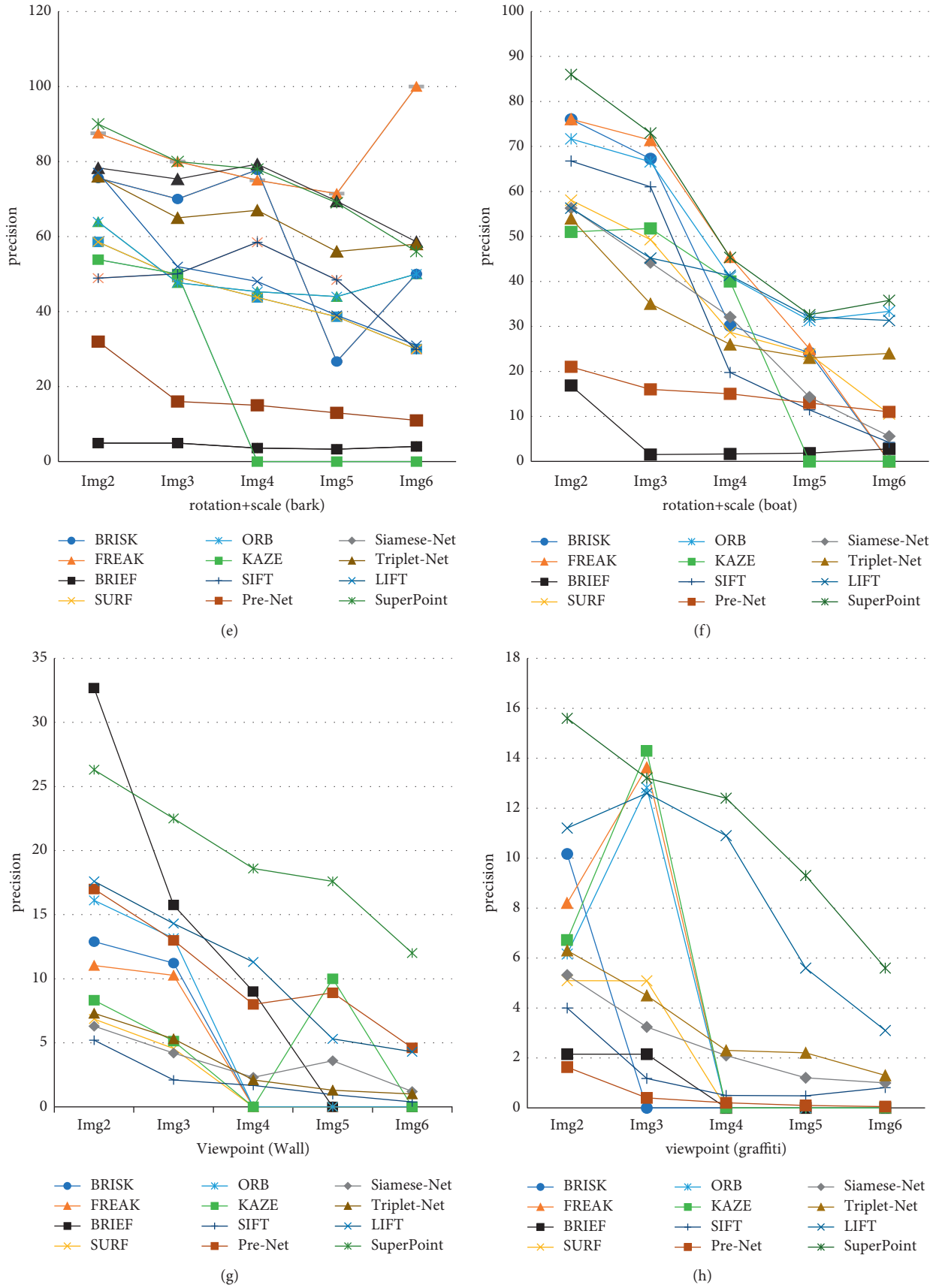

Figure 14: Comparison of different descriptors for different geometrical transformations. (a) Results of compression change. (b) Results of illumination change. (c) Results of blur change of structured scene. (d) Results of blur change of textured scene. (e, f) Results of rotation and scale change. (g, h) Results of viewpoint change.

and scale. However, BRISK, ORB, and FREAK all get a higher value for considering the rotation in sampling. The curves drop sharply in Figure 14(e) than Figure 14(f) because of the considerable scale variance in image pairs in Figure 8(e). Figure 14(f) shows KAZE obtains zero matching score from the second image pair. Figures 14(e) and 14(f) show BRISK, ORB, and FREAK can capture sufficient distinctiveness to change in rotation and scale.

As shown in Figure 14(f), note that LIFT is the best performing competitor on two pairs of images. As discussed above, LIFT applies the rotation estimation in the network to regularize the keypoint patch before generating description vector for it. However, scale invariance is not learnable in the descriptor; as shown in Figure 14(e), LIFT performs worse than some handcrafted descriptors because the evaluated image pair with rotation and scale changes at the same time. Similarly, the SuperPoint whose descriptor learns semidensely rather than densely, and the homographic adaptation attaining invariance of rotation, scale, distort, etc., is only apply in constructing training dataset to self-label, which results in the joint training results having invariance to rotation and scale.

Figures 14(g) and 14(h) show the results of descriptors that vary with viewpoints. The curve shows precisions are all low and descend rapidly. The best matching score is only 34% detected by BRIEF. As shown in Figure 14(h), the fail starts from the third image pair that comprises the first image and the fourth image. It is a big challenge to describe local regions with viewpoint changes. LIFT and SuperPoint show the better performance in descripting keypoints from image with viewpoint change. Triplet-Net is only successful on the first two images, partly correct on the third image, and failure from the fourth image pairs.

5.2.2. Results of Experiments on Multimodal Image. We experiment on three pairs of multimodal images from UAV, Landsat, and SDO and NVST and record the matching precisions in Table 3. The differences in angle and scale exist in the UAV image pairs. For UAV image pairs, all handcrafted descriptors fail in matching. For Landsat Images, there are no significant differences in rotation and scale between the two images, so all descriptors are accomplished. FREAK and ORB had relatively high precision for their designs considering the difference in scale and rotation. For astronomical images, there are significant differences in resolution and rotation between the two images. Except for SIFT, other handcrafted detectors all fail. From Table 3, we conclude that SIFT is the most stable handcrafted descriptors for multimodel images with rotation and size changes. Pre-Net and Siamese-Net show failure in description of the feature points from UAV images with large variance of rotation. These two descriptors do not learn or learn little representation to rotation changes. Other learnable descriptors learn effective representation to changes of rotation, scale, and blur as well.

Table 3: The comparison of descriptors.

Detectors	UAV images	Landsat images	Astronomical images
BRISK	Fail	25.00	Fail
FREAK	Fail	40.00	Fail
BRIEF	Fail	0.05	Fail
SURF	Fail	7.42	Fail
ORB	Fail	35.00	Fail
SIFT	Fail	29.49	10.80
KAZE	Fail	19.33	Fail
Pre-Net	Fail	7.56	Fail
Siamese-Net	Fail	18.78	11.2
Triplet-Net	6.3	41.3	36
LIFT	65.6	56.3	45.6
SuperPoint	35.3	68.7	79.6

6. Conclusion and Feature Trends

Image registration belongs to the basic research of image processing. It provides multiview and multimodal visual information for image fusion, detection, segmentation, and recognition. It is widely used in medicine, aviation, astronomy, transportation, monitoring, and other fields. The method to solve the image registration is mainly based on the feature. Detecting feature points, descripting feature points, matching points, estimating transform matrix, and warping image, the detectors and descriptors are the key procedures that determine whether the subsequent work can be carried out. A lot of significant researching result have been achieved. Therefore, this paper reviewed the popular detectors and descriptors from handcrafted to trainable aiming to provide a reference for the researchers and engineers in the field.

The review provides a detailed introduction of frequently used detectors and descriptors. The handcrafted detectors include Harris, FAST, BRISK, ORB, SURF, SIFT, and KAZE. The learnable detectors include DetNet, LIFT, multiscale, TILDE, and SuperPOINT. The handcrafted descriptors include BRISK, FREAK, BRIEF, SURF, ORB, SIFT, and KAZE. The learnable descriptors include Pre-Net, Siamese-Net, Triplet-Net, LIFT, and SuperPoint. We compared the detectors and descriptors on two datasets that consist of the artificial data with one change and the actual data with complex changes. We also provide the comparison and analyse these classical and deep learning-based techniques through extensive experiments on representative datasets. Our experimental results demonstrate that learnable detectors and descriptor outperform the handcrafted methods as long as the architecture of network is reasonable and the trainable date is enough.

Despite the achieved progresses, the further researcher in the detector and the descriptor will concentrate on the following challenges in the future:

(i) A large training dataset with comprehensive geometric changes needs to be established, which is used to train deep network to learn detectors and descriptors with more generalization.

(ii) Joint training detectors and descriptors in one pipeline to achieve better performance than training separately. Complex network structure and differentiability are a challenge. In pipeline, the ways that keypoint detector and descriptor benefit each other need to investigate carefully.

(iii) The transform matrix is a direct output result. After inputting the image pairs, the network integrates keypoint detection, description, matching in a pipeline and output the transform matrix directly. It is a challenging problem to design an encapsulation network that learns the transformation matrix directly from the image.

(iv) Future work will investigate how to boost the performance of models to the viewpoint changes.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request or from https://www.robots.ox.ac.uk/~vgg/, https://medium.com/dronehub/datasets-96fc4f9a92e5, https://serc.carleton.edu/eyesinthesky2/week11/get_to_know_multispectral_imaging.html, and https://www.nasa.gov/mission_pages/sdo/main/index.html, http://english.ynao.cas.cn/ti/nvst/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2018YFA0404603), the National Natural Science Foundation of China (Title: Research on Multiple Channel Solar Image Registration Method, No. 11773012), the Joint Research Fund in Astronomy (Nos. U1831204, U1931141) under a cooperative agreement between the National Natural Science Foundation of China (NSFC) and Chinese Academy of Sciences (CAS), the National Natural Science Foundation of China (Nos. 11903009 and 11961141001), and the Youth Innovation Promotion Association CAS.

References

[1] V. Balntas, L. Karel, V. Andrea, T. Tinne, M. Jiri, and M. Krystian, "Hpatches: a benchmark and evaluation of handcrafted and learned local descriptors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5173–5182, IEEE, Honolulu, HI, USA, July 2017.

[2] M. H. Hesamian, W. Jia, X. He, and P Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of Digital Imaging*, vol. 32, no. 4, pp. 582–596, 2019.

[3] W. Zhang, X. Li, J. Yu, M. Kumar, and Y. Ma, "Remote sensing image mosaic technology based on SURF algorithm in agriculture," *J Image Video Proc*, vol. 85, pp. 1–9, 2018.

[4] L. Wan, Y. Xiang, and H. You, "A post-classification comparison method for SAR and optical images change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 7, pp. 1026–1030, 2019.

[5] B. Zitová and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, no. 11, pp. 977–1000, 2003.

[6] R. Liao, M. Shun, T. Pierre et al., "An artificial agent for robust image registration," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, February 2017.

[7] D. T. Daniel, M. Tomasz, and R. Andrew, "Superpoint: self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 224–236, IEEE, Salt Lake City, UT, USA, June 2018.

[8] https://roboflow.com/formats/coco-json.

[9] J. D. Watson and F. H. C. Crick, "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid," *Nature*, vol. 171, no. 4356, pp. 737-738, 1953.

[10] D. Zhang and W. Shu, "Two novel characteristics in palmprint verification: datum point invariance and line feature matching," *Pattern Recognition*, vol. 32, no. 4, pp. 691–702, 1999.

[11] F. Tupin, H. Maitre, J.-F. Mangin, J.-M. Nicolas, and E. Pechersky, "Detection of linear features in SAR images: application to road network extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 36, no. 2, pp. 434–453, 1998.

[12] H. Jungong, P. Eric, and M. Z. Paul, "Visible and infrared image registration in man-made environments employing hybrid visual features," *Pattern Recognition Letters*, vol. 34, no. 1, pp. 42–51, 2013.

[13] L. Huang and Z. Li, "Feature-based image registration using the shape context," *International Journal of Remote Sensing*, vol. 31, no. 8, pp. 2169–2177, 2010.

[14] H. Sui, C. Xu, J. Liu, and F. Hua, "Automatic optical-to-SAR image registration by iterative line extraction and voronoi integrated spectral point matching," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 11, pp. 6058–6072, 2015.

[15] V. A. Zimmer, M. Á. González Ballester, and G. Piella, "Multimodal image registration using Laplacian commutators," *Information Fusion*, vol. 49, pp. 130–145, 2019.

[16] A. Okorie and S. Makrogiannis, "Region-based image registration for remote sensing imagery," *Computer Vision and Image Understanding*, vol. 189, no. 02825, pp. 1–15, 2019.

[17] H. Sokooti, G. Saygili, B. Glocker, B. P. F. Lelieveldt, and M. Staring, "Quantitative error prediction of medical image registration using regression forests," *Medical Image Analysis*, vol. 56, pp. 110–121, 2019.

[18] J. Swamidas, C. Kirisits, M. De Brabandere, T. P. Hellebust, F.-A. Siebert, and K. Tanderup, "Image registration, contour propagation and dose accumulation of external beam and brachytherapy in gynecological radiotherapy," *Radiotherapy & Oncology*, vol. 143, pp. 1–11, 2020.

[19] A. Sedghi, J. Luo, A. Mehrtash et al., *Semi-supervised Deep Metrics for Image Registration*, 2018.

[20] K. Mikolajczyk, T. Tuytelaars, C. Schmid et al., "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1, pp. 43–72, 2005.

[21] K. Joshi and M. I. Patel, "Recent advances in local feature detector and descriptor: a literature survey," *International Journal of Multimedia Information Retrieval*, vol. 9, pp. 1–17, 2020.

[22] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 6, pp. 765–781, 2011.

[23] K. Lenc and A. Vedaldi, "Learning covariant feature detectors," in *Proceedings of the European Conference on Computer Vision*, pp. 100–117, Amsterdam, The Netherlands, October2016.

[24] D. Mahapatra and Z. Ge, "Training data independent image registration using generative adversarial networks and domain adaptation," *Pattern Recognition*, vol. 100, Article ID 07109, 2020.

[25] C. Harris and M. Stephens, "A combined corner and edge detector," *Procedings of the Alvey Vision Conference 1988*, vol. 15, no. 50, pp. 10–5244, 1988.

[26] E. Rosten and T. Drummond, "Fusing points and lines for high performance tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1508–1511, IEEE, Beijing, China, October 2015.

[27] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: binary robust independent elementary features," in *Lecture Notes in Computer Science*, K. Daniilidis, P. Maragos, and N. Paragios, Eds., vol. 6314, pp. 778–792, Springer, Berlin, Germany, 2010.

[28] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: binary robust invariant scalable keypoints," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2548–2555, IEEE, Barcelona, Spain, November 2011.

[29] A. Alahi, O. Raphael, and V. Pierre, "FREAK: fast Retina keypoint," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 510–517, IEEE, Providence, RI, USA, June 2012.

[30] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *Proceedings of the 2011 International Conference on Computer Vision*, pp. 2564–2571, IEEE, Barcelona, Spain, November 2011.

[31] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[32] H. Bay, T. Tinne, and V. G. Luc, "SURF: speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[33] P. F. Alcantarilla, A. Bartoli and A. J. Davison, *KAZE Features," ECCV 2012*, Springer, vol. 7577, London, UK, 2012.

[34] S. Klein, S. Marius, M. Keelin, A. V. Max, and P. W. P. Josien, "Elastix: a toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2009.

[35] H. Altwaijry, A. veit, and S. Belongie, "Learning to detect and match keypoints with deep architectures," *Procedings of the British Machine Vision Conference 2016*, BMVA Press, in *Proceedings of the British Machine Vision Conference 2016*, September 2016.

[36] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Foundations and Trends® in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2007.

[37] E. Simo-Serra, T. Eduard, F. Luis, K. Iasonas, F. Pascal, and M. N. Francesc, "Discriminative learning of deep convolutional feature point descriptors," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 118–126, IEEE, Santiago, Chile, December 2015.

[38] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4353–4361, IEEE, Dec 2015.

[39] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, IEEE, Salt Lake City, UT, USA, June 2018.

[40] Y. Ono, T. Eduard, F. Pascal, and M. Y. Kwang, "LF-Net: learning local features from images," *Computer Vision and Pattern Recognition*, 2018.

[41] X. Shen, C. Wang, X. Li et al., "Rf-net: an end-to-end image matching network based on receptive field," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8132–8140, IEEE, Long Beach, CA, USA, June 2019.

[42] Y. Verdie, M. Y. Kwang, F. Pascal, and L. Vincent, "Tilde: a temporally invariant learned detector," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5279–5288, Boston, Massachusetts, USA, 2015.

[43] K. M. Yi, T. Eduard, L. Vincent, and F. Pascal, "Lift: learned invariant feature transform," in *Proceedings of the European Conference on Computer Vision*, pp. 467–483, Springer, Lausanne, Switzerland, September 2016.

[44] Z. Yang, T. Dan, and Y Yang, "Multi-temporal remote sensing image registration using deep convolutional features," *IEEE Access*, vol. 6, pp. 38544–38555, 2018.

[45] Y. Li, S. Wang, Q. Tian, and X. Ding, "A survey of recent advances in visual feature detection," *Neurocomputing*, vol. 149, pp. 736–751, 2015.

[46] A. Torralba, F. Rob, and W. Yair, "Small codes and large databases for recognition," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, Anchorage, AK, USA, June 2008.

[47] M. Muja and D. G. Lowe, "Fast matching of binary features," in *Proceedings of the Conference on Computer and Robot Vision. CRV*, pp. 404–410, IEEE, Toronto, ON, Canada, May 2012.

[48] L. S. Johannes, H. Hans, S. Torsten, and P. Marc, "Comparative evaluation of handcrafted and learned local features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1482–1491, IEEE, Honolulu, HI, USA, July 2017.

[49] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: a survey," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 23–79, 2021.

[50] X. Jiang, J. Ma, G. Xiao, Z. Shao, and X. Guo, "A review of multimodal image matching: methods and applications," *Information Fusion*, vol. 73, pp. 22–71, 2021.

[51] A. P. James and B. V. Dasarathy, "Medical image fusion: a survey of the state of the art," *Information Fusion*, vol. 19, pp. 4–19, 2014.

[52] Y. Li, J. Ma, and Y. Zhang, "Image retrieval from remote sensing big data: a survey," *Information Fusion*, vol. 67, pp. 94–115, 2021.

[53] Z. Qu, "etc., An algorithm of image mosaic based on binary tree and eliminating distortion error," *PLoS One*, vol. 14, no. 1, 2019.

[54] T. De Silva, A. Uneri, M. D. Ketcha et al., "3D-2D image registration for target localization in spine surgery: investigation of similarity metrics providing robustness to content mismatch," *Physics in Medicine and Biology*, vol. 61, no. 8, pp. 3009–3025, 2016.

[55] J. Krebs, M. Tommaso, D. Hervé et al., "Robust non-rigid registration through agent-based action learning," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 344–352, Springer, September 2017.

[56] M. Lorenzi, N. Ayache, G. B. Frisoni, and X. Pennec, "LCC-Demons: a robust and accurate symmetric diffeomorphic registration algorithm," *NeuroImage*, vol. 81, pp. 470–483, 2013.

[57] R. Liao, S. Miao, P. D. Tournemire, S. Grbic, and A. Kamen, "An articial agent for robust image registration," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 4168–4175, AAAI Press, Princeton, NJ, USA, February 2017.

[58] K. Ma, J. Wang, V. Singh et al., "Multimodal image registration with deep context reinforcement learning," *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, pp. 240–248, 2017.

[59] M. Shun, P. Sebastien, F. Peter et al., *Dilated Fcn for Multi-Agent 2d/3d Medical Image Registration*, AAAI, Menlo Park, CA, USA, 2018.

[60] A. Sheikhjafari, M. Noga, K. Punithakumar, and N. Ray, "Unsupervised deformable image registration with fully connected generative neural network," in *Proceedings of the International Conference on Medical Imaging with Deep Learning*, Shenzhen, China, April 2018.

[61] Y. Jin, D. Mishkin, A. Mishchuk et al., "Image matching across wide baselines: from paper to practice," *International Journal of Computer Vision*, vol. 129, no. 2, pp. 517–547, 2021.

[62] Rosten, "Faster and better: a machine learning approach to corner detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 105–119, 2008.

[63] E. Mair, G. D. Hager, D. Burschka, M. Suppa, and G. Hirzinger, "Adaptive and generic corner detection based on the accelerated segment test," in *Proceedings of the Computer Vision—ECCV 2010*, pp. 183–196, Heraklion, Greece, September 2010.

[64] Y. Li, Q. Huang, Y. Liu, Y. Huang, and X. Sun, "Efficient properties-based learning for mismatch removal," *IEEE Access*, vol. 7, pp. 149612–149622, 2019.

[65] G. Wu, K. Minjeong, W. Qian, C. M. Brent, and S. Dinggang, "Scalable high-performance image registration framework by unsupervised deep feature representations learning," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1505–1516, 2015.

[66] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, N. Navab, and N. Komodakis, "A deep metric for multimodal registration," *Lecture Notes in Computer Science*, pp. 10–18, 2016.

[67] R. Wright, K. Bishesh, G. Alberto et al., "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. PIPPI 2018, DATRA 2018*, A. Melbourne, L. Roxane, D. F. Matthew et al., Eds., vol. 11076, Cham, Switzerland, Springer, 2018.

[68] I. Goodfellow, P. A. Jean, M. Mehdi et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, pp. 2672–2680, ACM, Montreal, Canada, 2014.

[69] G. Haskins, J. Kruecker, U. Kruger et al., "Learning deep similarity metric for 3D MR-TRUS image registration," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 3, pp. 417–425, 2019.

[70] M. Heinrich, J. Mark, W. P. Bartlomiej, M. B. Sir, and A. S. Julia, "Towards realtime multimodal fusion for image-guided interventions using self-similarities," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 187–194, Springer, Boston, MA, USA, September 2014.

[71] T. Cao, S. Nikhil, J. Vladimir, and N. Marc, "Semi-coupled dictionary learning for deformation prediction," in *Proceedings of the 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pp. 691–694, Brooklyn, NY, USA, April 2015.

[72] X. Cao, J. Yang, J. Zhang et al., "Deformable image registration based on similarity-steered CNN regression," *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, pp. 300–308, 2017.

[73] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec, "SVF-net: learning deformable image registration using shape matching," *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, pp. 266–274, 2017.

[74] S. Sun, H. Jing, Y. Mingqing et al., "Robust multimodal image registration using deep recurrent reinforcement learning," in *Proceedings of the Asian Conference on Computer Vision*, vol. 30, no. 4.

[75] S. L. Wei, B. H. Jia, and H. Y. Ming, "Semi-supervised learning for optical flow with generative adversaria lnetworks," *Advances in Neural Information Processing Systems*, pp. 353–363, Curran AssociatesInc., LongBeach, CA, USA, 2017.

[76] A. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "EV-FlowNet: self-supervised optical flow estimation for event-based cameras," *Robotics: Science and Systems XIV*, 2018.

[77] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: a meta-analysis and review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 166–177, 2019.

[78] K. Kuppala, S. Banda, and T. R. Barige, "An overview of deep learning methods for image registration with focus on feature-based approaches," *International Journal of Image and Data Fusion*, vol. 11, no. 2, pp. 113–135, 2020.

[79] Z. Luo, Z. Lei, B. Xuyang et al., "Aslfeat: learning local features of accurate shape and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6589–6598, IEEE, Seattle, WA, USA, June 2020.

[80] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[81] K. Wilson and N. Snavely, "Robust global translations with 1D SfM," in *Proceedings of the European Conference on Computer Vision*, September 2014.

[82] S. Saeedi, N. Luigi, J. Edward, B. Bruno, H. J. K. Paul, and J. D. Andrew, "Application-oriented design space exploration for SLAM algorithms," in *Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5716–5723, IEEE, Singapore, 2017.

[83] R. Hartley and A. Zisserman, "Multiple view geometry in computer vision," *Cambridge University Press*, vol. 23, no. 2, p. 271, 2005.

[84] P. F. Alcantarilla, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1281–1298, 2011.

[85] Y. Ke and R. Sukthankar, "PCA-SIFT: a more distinctive representation for local image descriptors," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, vol. 2, July 2004.

[86] G. Hua, B. Matthew, and W. Simon, "Discriminant embedding for local image descriptors," in *Proceedings of the International Conference on Computer Vision*, pp. 1–8, IEEE, Rio de Janeiro, Brazil, Oct 2007.

[87] G. Wu, M. Kim, Q. Wang, Y. Gao, S. Liao, and D. Shen, "Unsupervised deep feature learning for deformable registration of MR brain images," *Advanced Information Systems Engineering*, pp. 649–656, 2013.

[88] W. Ziyu, S. Tom, H. Matteo, V. H. Hado, L. Marc, and D. F. Nando, *Dueling Network Architectures for Deep Reinforcement Learning*, pp. 1995–2003, MLR Press, London, UK, 2016.

[89] X. Han, L. Thomas, J. Yangqing, S. Rahul, and C. B. Alexander, "Matchnet: unifying feature and metric learning for patch-based matching," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 3279–3286, IEEE, Boston, MA, USA, June 2015.

[90] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[91] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *Proceedings of the International Conference on Computer Vision Theory and Applications.VISAPP*, p. 331, DBLP, Lisboa, Portugal, January 2009.

[92] P. Torr and A. Zisserman, "Robust computation and parametrization of multiple view relations," in *Proceedings of the Sixth International Conference on Computer Vision*, pp. 727–732, IEEE, Bombay, India, January 1998.

[93] P. H. S. Torr and A. Zisserman, "MLESAC: a new robust estimator with application to estimating image geometry," *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 138–156, 2000.

[94] C. Yang, M. Zhang, Z. Zhang, L. Wei, R. Chen, and H. Zhou, "Non-rigid point set registration via global and local constraints," *Multimedia Tools and Applications*, vol. 77, no. 24, pp. 31607–31625, 2018.

[95] H. Zhou and J. Jayender, "Smooth deformation field-based mismatch removal in real-time," 2020, https://arxiv.org/abs/2007.08553.

[96] Y. Yang, S. H. Ong, and K. W. C. Foong, "A robust global and local mixture distance based non-rigid point set registration," *Pattern Recognition*, vol. 48, no. 1, pp. 156–173, 2015.

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

[98] D. Barath, M. Ivashechkin, and J. Matas, "Progressive NAPSAC: sampling from gradually growing neighborhoods," 2019, https://arxiv.org/abs/1906.02295.

[99] D. Barath, J. Matas, and J. Noskova, "MAGSAC: marginalizing sample consensus," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10197–10205, IEEE, Long Beach, CA, USA, June 2019.

[100] X. Jiang, M. Jiayi, J. Junjun, and G. Xiaojie, "Robust feature matching using spatial clustering with heavy outliers," *IEEE Transactions on Image Processing*, vol. 29, pp. 736–746, 2019.

[101] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker, "Universal correspondence network," *News in Physiological Sciences*, vol. 2, no. 3, p. 8, 2016.