An evaluation of conventional and deep learning-based image-matching methods on diverse datasets

Article *in* The Photogrammetric Record · April 2023

DOI: 10.1111/phor.12445

Shunping Ji
Wuhan University

DOI: 10.1111/phor.12445

ORIGINAL ARTICLE

An evaluation of conventional and deep learning-based image-matching methods on diverse datasets

Shunping Ji[1] | Chang Zeng[1] | Yongjun Zhang[1] | Yulin Duan[2]

[1]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China

[2]Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing, China

Correspondence
Shunping Ji, School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan 430079, China.
Email: jishunping@whu.edu.cn

Funding information
National Key Research and Development Program of China, Grant/Award Number: 2019YFE0125300; National Natural Science Foundation of China, Grant/Award Number: 42171430; State Key Program of the National Natural Science Foundation of China, Grant/Award Number: 42030102

Abstract

Image matching plays an important role in photogrammetry, computer vision and remote sensing. Modern deep learning-based methods have been proposed for image matching; however, whether they will surpass and take the place of the conventional handcrafted methods in the remote sensing field still remains unclear. A comprehensive evaluation on stereo remote sensing images is also lacking. This paper comprehensively evaluates the performance of conventional and deep learning-based image-matching methods by dividing the matching process into feature point extraction, description and similarity measure on various datasets, including images captured from close-range indoor and outdoor scenarios, unmanned aerial vehicles (UAVs) and satellite platforms. Different combinations of the three steps are evaluated. The experimental results reveal that, first, the performance of the different combinations varies between individual datasets, and it is difficult to determine the best combination. Second, by using more comprehensive indicators on all of the datasets, that is, the average rank and absolute rank, the combination of scale-invariant feature transform (SIFT), ContextDesc and the nearest neighbour distance ratio (NNDR), and also the original SIFT, achieve the best results, and are recommended for use in remote sensing. Third, the deep learning-based Sub-SuperPoint extractor obtains a good performance, and is second only to SIFT. The learning based ContextDesc descriptor is as effective as the SIFT descriptor, and the learning based SuperGlue matcher is not as stable as NNDR, but leads to a few top-performing combinations. Finally, the handcrafted methods are generally faster than the deep learning-based methods, but the efficiency of the latter is acceptable. We conclude that although a full deep learning-based method/combination has not yet beaten the conventional methods, there is still much room for improvement with the deep learning-based methods because large-scale aerial and satellite training datasets remain to be constructed, and specific methods for remote sensing images remain to be developed. The performance of the different combinations of feature extractor, descriptor and similarity measure varies between individual datasets. The combination of SIFT, ContextDesc and NNDR, and also the original SIFT, achieve the best results when using more comprehensive indicators on all the datasets. For extractor, the learning based Sub-SuperPoint is second only to SIFT; for descriptor, learning-based ContextDesc is as effective as the SIFT descriptor; and for matcher, learning-based SuperGlue is not as stable as NNDR.

KEYWORDS
deep learning, description, feature point extraction, image matching, similarity measure

INTRODUCTION

Discovering corresponding points in overlapping images from different viewpoints, which is called “image matching”, is a necessary technique for downstream applications such as image stitching, camera calibration, bundle adjustment, structure from motion (SfM), simultaneous localisation and mapping (SLAM) and three-dimensional (3D) reconstruction. Therefore, the image-matching performance greatly affects these tasks. The mainstream image-matching methods are feature-based methods and consist of three main steps: (1) extracting repeatable feature points from stereo images; (2) calculating a distinctive descriptor for the feature points; and (3) establishing the point correspondence between the two feature point sets based on the similarity measure of the descriptors.

In the past few decades, many feature-matching algorithms have been proposed, including both conventional and deep learning-based methods. Among the handcrafted conventional methods, scale-invariant feature transform (SIFT) (Lowe, 2004) and Oriented FAST and Rotated BRIEF (ORB) (Rublee et al., 2011) are the most widely used in many visual applications (Parente et al., 2021). There have been many subsequent attempts to improve these methods, such as affine scale-invariant feature transform (ASIFT) (Morel & Yu, 2009) and speeded up robust features (SURF) (Bay et al., 2008) and rotation-invariant self-similarty (Mohammadi et al., 2022), which were developed based on SIFT detector or descriptor. In recent years, more and more researchers have been shifting to learnable methods. Some of these methods focus only on feature extraction, such as the Temporally Invariant Learned DEtector (TILDE) (Verdie et al., 2015) and Quad-networks (Savinov et al., 2017); some focus on the descriptor, such as DeepDesc (Simo-Serra et al., 2015), L2Net (Tian et al., 2017), group invariant feature transform (GIFT) (Liu et al., 2019), and ContextDesc (Luo et al., 2019); and some focus on the similarity measure, such as SuperGlue (Sarlin et al., 2020). Other studies have focused on two or three of these steps at the same time, such as SuperPoint (Detone et al., 2018), UnSuperPoint (Christiansen et al., 2019), D2-Net (Dusmanu et al., 2019), LF-Net (Ono et al., 2018), and learned invariant feature transform (LIFT) (Yi et al., 2016). Currently, the conventional methods, such as SIFT, are still the most widely used, although some deep learning-based methods claim to have surpassed the handcrafted algorithms, most of which, such as HPatches (Balntas et al., 2017) and ScanNet (Dai et al., 2017), have in fact only been trained and tested on specific close-range datasets. There have also been some studies in the computer vision field that have pointed out that deep learning-based algorithms have not outperformed the handcrafted methods on certain datasets or tasks (Fan et al., 2019; Schonberger et al., 2017). However, there is a lack of comprehensive evaluations on remote sensing datasets, particularly aerial and satellite datasets. Can these deep learning-based methods perform better than the handcrafted methods in various types of images, especially remote sensing images? This paper attempts to give the answer by systematically and comprehensively evaluating not only the existing handcrafted and deep learning-based methods, but also different combinations of feature extractor, descriptor, and matcher, on various close-range and remote sensing datasets.

We selected representative feature point extraction (extractor), feature description (descriptor) and similarity measure (matcher) algorithms, combined them, and evaluated their combinations on various datasets. Specifically, we chose SIFT, ORB and ASIFT as representative conventional extractors, and SuperPoint, UnSuperPoint and ASLFeat as representative deep learning-based extractors. As for the descriptor, in addition to the three deep learning methods mentioned above, we also added GIFT and ContextDesc. As for the matcher, we chose the nearest neighbour distance ratio (NNDR) algorithm as representing the conventional methods and SuperGlue as representing the deep learning-based methods. We also selected an end-to-end learning method—the Local Feature TRansformer (LoFTR)—for the comparison. In order to cover the various situations using the image matching technique, we used different experimental datasets, including the Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) dataset (Geiger et al., 2012), the HPatches dataset (Balntas et al., 2017), the ScanNet dataset (Dai et al., 2017), and unmanned aerial vehicle (UAV) and ZY3 satellite data. We then evaluated the performance of the image-matching algorithms on these datasets, both separately and generally.

The main contribution of this paper is that we provide a much more comprehensive and general evaluation of modern image-matching methods than ever before. We also fill in the gap of specific evaluations on remote sensing datasets being lacking. The second contribution is that, through a comprehensive between-method comparison, we identify the stable and high-performance combinations that are suitable for all types of images, especially for remote sensing images, with the components of the combinations including both conventional and deep learning-based methods.

The remainder of the paper is structured as follows. The next section reviews the related image-matching methods and the previous studies of between-method evaluation. The third section briefly describes the methods of feature extraction, description and similarity measure. The fourth section describes the datasets used in the experiments and presents the results for each dataset. The fifth section gives an overall analysis considering all the datasets. The sixth section concludes the paper.

RELATED WORK

Feature point extraction

Feature point extraction involves detecting a distinctive point that appears on stereo or multi-view images. The conventional methods are designed to detect corners or intersections with shapes such as "X" and "T", or distinct regions in which the pixels are similar to each other and different from their neighbours. The early Moravec operator (Moravec, 1977) calculates the intensity variation of each pixel in four or eight directions, and the points of interest are then detected according to a given threshold and non-maximum suppression. Förstner and Gülch (1987) used the Robert gradient to find points of interest with the error ellipse as small and close to a circle as possible. Harris and Stephens (1988) detected the points of interest by calculating the eigenvalues of the autocorrelation matrix. The adaptive and generic accelerated segment test (AGAST) (Mair et al., 2010) and features from accelerated segment test (FAST) (Trajković & Hedley, 1998) methods have been widely used due to their computational efficiency. The ORB method (Rublee et al., 2011) is an improved method based on FAST and binary robust independent elementary features (BRIEF). Blob features have also been widely used, such as Laplacian of Gaussian (LoG) (Lindeberg, 1998) and the Hessian matrix. The difference of Gaussians (DoG) is used to approximate and speed up LoG. SIFT (Lowe, 2004) detects keypoints as the local extrema in a DoG pyramid. Affine-invariant SIFT (ASIFT) (Morel & Yu, 2009) simulates multiple affine changes of the image. SURF accelerates SIFT by using the Haar wavelet and an integral image strategy to approximate a Hessian matrix.

In addition to the above handcrafted features, more recent algorithms have been based on deep learning. FAST (Trajković & Hedley, 1998) is the first algorithm using classic machine learning methods to obtain reliable feature points. More recently, supervised deep learning-based methods have been proposed. Nevertheless, there are no large-scale feature point datasets currently available. Therefore, these methods take the results of the conventional feature point extraction methods as the labels. For example, TILDE (Verdie et al., 2015) and D2-Net (Dusmanu et al., 2019), which have been used in SfM learn from SIFT keypoints (Maiwald & Maas, 2021). TILDE trains on the positions of the SIFT points of single images and their counterparts under various weather and brightness changes, so as to increase the robustness to illumination changes. D2-Net and LF-Net (Ono et al., 2018) use depth maps constructed by Colmap (Schonberger & Frahm, 2016) or similar methods, to eliminate the errors of the sparse points obtained from SIFT-based SfM, where the remaining SIFT points are used as training samples. Detone et al. (2018) proposed SuperPoint, and they constructed a synthetic dataset composed of basic graphics such as triangles, rectangles, and their changes (fuzzy, affine, etc.) for pre-training. The pretrained model was then used to predict and generate the pseudo-labels of a natural image from the common objects in context (COCO) dataset (Lin et al., 2014) and its counterparts generated with different homographic transformation parameters, and the pseudo-labels were then used for the second training. Differing from D2-Net and LF-Net, which use sparse correspondences, ASLFeat (Luo et al., 2020) obtains dense correspondences and poses of stereo images from the multi-view stereo (MVS) algorithm and 3D mesh models in GL3D. A convolutional neural network (CNN) is then applied to learn the dense correspondence relationships, where the local maxima of the feature maps are treated as feature points. Like ASLFeat, UnSuperPoint (Christiansen et al., 2019) learns feature points from the dense correspondences, but by simulating the correspondences by the homographic counterparts of single images.

Descriptors

Corresponding feature points can be matched according to the similarity metric between their descriptors. Therefore, the corresponding descriptors need to be discriminative with regard to the other points and robust to changes of illumination and viewpoints. The histogram of oriented gradients from Gaussian pyramids is used in SIFT. DSP-SIFT (Dong & Soatto, 2015) improves SIFT by pooling the gradient orientation across various domain sizes, and CSIFT (Abdel-Hakim & Farag, 2006) enhances the distinctiveness of SIFT by using more colour information, whereas DAISY (Tola et al., 2009) uses different Gaussian kernels to convolve the gradient histogram to speed up the calculation. BRIEF (Calonder et al., 2010) creates binary descriptors by directly comparing the pixel intensity, and is very fast as a result. Rotated BRIEF (Calonder et al., 2010) further combines the improved FAST

(Trajković & Hedley, 1998) to form the ORB descriptor (Rublee et al., 2011). Among the deep learning methods, MatchNet (Han et al., 2015) uses a Siamese network to learn the descriptor and metric. Kumar Bg et al. (2016) introduced a global loss to minimise the mean value and variance of the distance between features of the same class, and to maximise the mean value of the distance between different classes. TFeat (Balntas et al., 2016) use triplets of samples for training, and HardNet (Mishchuk et al., 2017) introduces hinge triplet loss. Apart from metric learning, BinGAN (Zieba et al., 2018) introduces a regularisation generative adversarial network (GAN) to generate binary descriptors. GIFT (Liu et al., 2019) simulates multiple rotations and scales of an image, and then extracts and fuses the corresponding descriptors under various changes. ContextDesc (Luo et al., 2019) fuses the geometric context and the visual context of keypoints to form a discriminative descriptor. Both SuperPoint (Detone et al., 2018) and UnSuperPoint (Christiansen et al., 2019) use an encoder and decoder network structure. LIFT (Yi et al., 2016), RF-Net (Shen et al., 2019), and LF-Net (Ono et al., 2018) first learn the main direction of the keypoints and then create descriptors with deep features.

Similarity measure

The similarity measure determines the one-to-one correspondence between two sets of feature descriptors, which are expressed as floating or integer vectors. The judgement on the difference or cross-correlation of two vectors with a fixed threshold is the simplest strategy, but it prunes to one-to-many matching and introduces many wrong matches. NNDR assumes that there should be a large difference between the nearest distance and the next nearest distance of a reference descriptor and the descriptor set to be matched, to guarantee the distinctiveness of a descriptor so as to avoid one-to-many problems. The mutual strategy mutually matches the two sets of descriptors so as to eliminate points with a low matching quality. Bas and Ok (2021) propose a similarity measure pipeline which contains NNDR, cross check and an iterative construction of the tentative matches based on AKAZE features. There are a few deep learning methods that can be used as a similarity measure. For example, SuperGlue use a graph neural network (GNN) to fuse the localisation and description of keypoints, and then Sinkhorn's algorithm (Cuturi, 2013) is used for matching the fused features.

There are also many deep learning-based methods containing both extraction and description steps. For example, LIFT, RF-Net and LF-Net detect keypoints, calculate their main orientation, and compute the local descriptors, whereas D2-Net and ASLFeat extract feature points and descriptors at the same time.

Between-method evaluation studies

There have been some studies that have focused on comparing conventional image-matching methods since the early 21st century. For example, Zitova and Flusser (Zitova & Flusser, 2003) provided several evaluation indicators for image registration, including localisation error and matching error in 2003. Mikolajczyk and Schmid (2005) compared the matching performance of different local descriptors, and they also compared the performance of affine-invariant extractors in various imaging conditions in another work (Mikolajczyk et al., 2005). Moreels and Perona (2007) compared the ability of different extractors and descriptors in matching 3D objects from different viewpoints and light conditions. Aanæs et al. (2012) evaluated various detectors on a large close-range dataset (Aanæs et al., 2016) with known camera poses, intrinsic parameters, and controlled illumination conditions. Heinly et al. (2012) compared various binary feature descriptors and inter-combined extractors and descriptors on different datasets. Similarly, Mukherjee et al. (2015) also used the inter-combination method to evaluate and analyse various matching methods on different datasets. However, these studies all used close-range indoor and outdoor datasets, while remote sensing datasets have rarely been applied.

Recently, reviews and evaluations of deep learning-based methods have gradually appeared. Balntas et al. (2017) proposed HPatches, a large benchmark dataset for evaluating handcrafted and deep learning-based descriptors; however, this dataset uses only close-range planar images, that is, objects appearing on the same plane. Schonberger et al. (2017) discovered that some deep learning-based local features do not perform well in the conditions of large viewpoint and illumination changes. A few papers have evaluated image-matching methods in various tasks. For example, Fan et al. (2019) evaluated the performance of conventional and learning based descriptors in 3D reconstruction. Jin et al. (2021) introduced a benchmark for local features and robust estimation algorithms, focusing on the pose recovery. Komorowski et al. (2018) compared handcrafted and deep learning-based feature detection algorithms on a street view dataset for driverless technology. Ma et al. (2021) reviewed the handcrafted and learning-based matching algorithms and evaluate the algorithms on pose estimation and loop-closure detection.

In summary, these studies have focused on matching close-range images, and there have been few studies that have evaluated handcrafted and deep learning-based methods on different remote sensing images. Moreover, most of these studies have only evaluated one or two parts of the image matching, that is, the extractor or descriptor. In this study, various feature extractors, descriptors and similarity measure algorithms, including newly proposed examples, were combined and evaluated on five datasets covering close-range indoor and outdoor scenarios, and also UAV and satellite scenes, providing a much more comprehensive and general evaluation.

METHOD

Feature point extraction

Three classical and most widely used methods—SIFT, ORB and ASIFT—and three representative deep learning-based methods—SuperPoint, UnSuperPoint and ASLFeat—were selected for the experiments. The three learning based methods use different ways to obtain the labels of keypoints. SuperPoint obtains pseudo keypoint labels by a self-supervised strategy. ASLFeat obtains the dense correspondence via an MVS dense matching algorithm. UnSuperPoint obtains the dense correspondence by homography matrix. We summarise their principal ideas and processes below.

SIFT

The DoG pyramid is constructed by filtering the original image with different Gaussian kernels and subtracting the neighbouring filtered images. A keypoint is a local maximum in the DoG pyramid. More accurate positions of the keypoints are determined by fitting a 3D quadratic function, and the points with a low response are eliminated by a Hessian matrix. The orientation of the SIFT feature points is computed by the gradient histogram of the local patch of the feature points, which have been Gaussian weighted. The peak value is taken as the orientation of the feature point.

ORB

ORB uses oriented FAST for feature detection. FAST considers 16 pixels in a circle around a pixel. The 16 pixels are divided into three subsets—darker, similar and brighter—depending on the pixel intensity. The ID3 algorithm (Quinlan, 1986) is used to train the decision tree and generate optimised feature points. ORB utilises an image pyramid to achieve scale invariance and computes the direction of the intensity centroid to realise rotation invariance.

ASIFT

Based on SIFT, ASIFT improves the robustness to affine deformation between viewpoint changes. ASIFT constructs a hemispherical space, which uniquely determines the spatial position of the camera through the longitude and latitude in the sphere. At these spatial positions, all possible affine deformations in the image can be simulated. SIFT is used to detect the keypoints in dozens of such simulated images, so as to realise affine invariance.

SuperPoint

A CNN can extract deep features for image classification. However, for image matching, how to define a point of interest with learnable deep features was not previously clear. In a pioneering study, Detone et al. (2018) constructed a synthetic shapes dataset, which included some simple geometric shapes, triangles, quads, stars, etc. The intersections of the “L”, “Y” and “T” shapes were considered as feature points. The designed network consists of one encoder and two decoders, which are, respectively, the interest point decoder for calculating the response of pixels and the descriptor decoder for semi-dense descriptors. SuperPoint first trains the encoder and the interest point decoder on the synthetic shapes dataset with some introduced noise, and the trained extractor is referred to as MagicPoint. MagicPoint is then used for detecting the pseudo-labels of natural images and their counterparts with random homography, with which it is further trained.

UnSuperPoint

Inspired by SuperPoint, UnSuperPoint consists of one encoder and three decoders representing the pixel scores, keypoint locations and descriptors, respectively, so as to share most of the computation. UnSuperPoint utilises a Siamese framework to train the network. The original image and its homographic counterpart are used as the input, so that each pixel in the original image can be matched with the corresponding location in the wrapped image. The loss functions are designed to ensure that the original image and the wrapped image have the same keypoints and descriptors, and to make the distribution of the keypoints more uniform.

ASLFeat

Both D2-Net and ASLFeat train with depth maps produced by MVS algorithms. D2-Net attempts to obtain the correspondence between stereo images from a sparse SfM 3D point cloud, while ASLFeat wraps the input image according to the ground-truth depths and camera poses. Inspired by D2-Net, ASLFeat uses the detect-and-describe strategy to calculate the keypoints and descriptors, which means that the extractors also serve as descriptors. In order to utilise the local patch information (i.e., scale and orientation) of feature points and solve the insufficient localisation accuracy of keypoints in D2-Net, a deformable convolutional network (DCN) is used to achieve affine invariance, and a feature pyramid is used to restore the spatial resolution and fine-level details, so as to refine the keypoint localisation.

Feature description

In addition to the descriptors corresponding to the extractors above, we also add GIFT and ContextDesc descriptors, which are proposed more recently.

SIFT

SIFT rotates the scale-normalised 16×16 image patch around the keypoints to the main orientation and divides them into 4×4 grids. SIFT then calculates the histogram of oriented gradients, with a Gaussian weight assigned to each pixel in eight directions in each grid. The histogram is then concatenated and normalised to form the SIFT descriptor.

ORB

ORB uses rotated BRIEF (rBRIEF) to compute the descriptors, in which 256 pairs of points around the feature points are selected and their intensity is compared with generate binary descriptors. In order to overcome the high discreteness in each dimension of the descriptor vector and the strong correlation among dimensions, rBRIEF chooses 256 pairs of points with the following strategy: for a dataset with 300k keypoints, each described by 31×31 local windows, rBRIEF chooses all of the 5×5 pixel patches from each local window and obtains M pairs of them. A 300k×M matrix is then calculated by comparing the intensity of each pair. The first column vector is then selected depending on the minimum distance from their mean value to 0.5. The remaining column vectors are chosen according to their correlations with the selected vector, until 256 column vectors and their corresponding point pairs are selected.

SuperPoint (UnSuperPoint)/ASLFeat:

SuperPoint uses the encoder and decoder structure, where a decoder branch computes the semi-dense features (1/8 of the original image size), which are aggregated to obtain the 256-dimension descriptors at each feature point. ASLFeat computes the feature maps of 128 dimensions at a quarter of the resolution of the input image, and then aggregates them to a 128-dimension descriptor.

GIFT

CNN-based descriptors are usually sensitive to viewpoint changes because convolutions are not invariant to those geometric transformations beyond translation. Group convolutions are used in GIFT to extract features in an image group, to make the descriptors more invariant and discriminative. The image group is constructed by simulating five rotations and five scale transformations, so that each keypoint has 5×5 descriptors. Finally, the descriptors are fused by convolutions to output a GIFT descriptor.

ContextDesc

Instead of rotating and scaling the input image to obtain descriptors that are invariant to viewpoints, ContextDesc takes another approach. The visual context from the whole image (a 32×downscaled feature map) and a local patch is cropped according to the location, scale, and orientation of the keypoints, and the geometric contexts from the keypoint locations are fused together to compute a 128-dimension descriptor.

Similarity measure

We selected the widely used NNDR for the traditional methods and SuperGlue, which is used in the best solution in the Image Matching Challenge 2022, for the deep learning-based methods as the matchers for recognising corresponding features between two sets of descriptors.

NNDR

Given a reference descriptor, NNDR considers that the distance between it and the closest descriptor in the searched descriptor set should differ greatly from that between it and the second-closest descriptor, so as to obtain robust matching.

SuperGlue

Inspired by the success of transformer models, SuperGlue utilises the relationship between the keypoint locations and their descriptors with self- and cross-attention. A multilayer perceptron (MLP) is then used to fuse the coordinates and descriptors of the keypoints. SuperGlue then constructs a graph convolutional network (GCN) which has two kinds of edges: self-edges that connect the keypoint to all the other keypoints within the same image, and cross edges that connect the keypoint to all the keypoints in the other image. Sinkhorn's algorithm is then used to determine the optimal corresponding points in the graph.

End-to-end learning method

LoFTR

LoFTR, which is also used in the best solution in the Image Matching Challenge 2022, considers that the use of a feature extractor may fail to extract enough repeatable points of interest between stereo images due to various factors, such as poor texture. Inspired by SuperGlue, LoFTR uses a coarse-to-fine strategy and applies a transformer to process the dense local features extracted from the convolutional backbone. Dense matches are extracted between two sets of transformer features at a low resolution and later refined with a cross-correlation-based approach.

Implementation

All the above methods, except UnSuperPoint, were provided with source code. Therefore, we retrained UnSuperPoint with the model and empirical parameters provided by the authors. For the traditional handcrafted extractors, in addition to using their own descriptors, we also used ContextDesc and GIFT as the descriptors, respectively. For SuperPoint, UnSuperPoint, and ASLFeat, in addition to their own descriptors, we also used SIFT, ORB, ContextDesc, and GIFT as the extractors. We used default settings of different methods. For example, the default parameters of SIFT in OpenCV (sigma is 1.6, the contrast threshold is 0.04 and the edge threshold is 10) were applied. In the matching, the threshold of NNDR was set to 0.8. The SuperGlue network was originally trained with the SuperPoint descriptors, and in order to evaluate the performance of other descriptors, we retrained SuperGlue on the other descriptors. SuperPoint detects keypoints at the pixel level, but in many applications, we need a sub-pixel accuracy. We obtained the sub-pixel SuperPoint extractor (called SPSub) by calculating the intensity centroid of the local features. In addition, in many studies, the algorithms were evaluated on images that had been resized to a certain fixed resolution. However, we did not resize the images, to guarantee fairness, except for the UAV dataset. The images from this dataset were too large for the deep learning methods, so we resized the images to a smaller uniform size. Considering the efficiency and practicability, we extracted up to 3500 keypoints in a single image.

DATASETS AND EXPERIMENTS

Datasets

We selected the open-source KITTI, HPatches, and ScanNet datasets, as well as ZY3 satellite images and UAV data. Thus, the datasets covered close-range indoor and outdoor scenarios, and also aerial and satellite stereo images. The image numbers, resolutions, and other characteristics of the datasets are listed in Table 1.

The KITTI dataset is a famous outdoor dataset widely used in SLAM and other visual applications. We used the visual odometry data, which consist of greyscale stereo image pairs and their poses. The stereo camera has a fixed viewpoint, and shares almost the same illumination conditions. However, there are some moving objects in the imagery, such as vehicles and pedestrians, which affect the estimation of poses. The HPatches dataset is a planar dataset, where all the scenes (e.g., murals) are planar scenes, and has been widely used in image matching evaluation, including SuperPoint, RF-Net, ASLFeat, etc. There are 57 scenes (subset A) with only illumination variations and 59 scenes (subset B) are real stereo images with viewpoint changes attached with the homography matrices, with respect to the first image of each scene. The indoor ScanNet dataset is composed of monocular sequences attached with ground-truth poses and depth maps. Therefore, the baseline of two frames is short and the illumination changes are small.

The ZY3 satellite three-line camera (TLC) has a forward view, a nadir view, and a backward view, with 22° intersection angles. The altitude of the satellite is 506 km, and the ground sampling distance (GSD) of the nadir view is 2.1 m, and that of the forward and backward views is 2.5 m. The images are cropped into patches of 600 × 600 pixels (nadir) and 500 × 550 pixels (backward). The UAV dataset was collected in the city of Meitan, Guizhou province, China, in October 2013, by the use of an oblique five-view camera rig, with one nadir view and four 40° oblique view angles, mounted on a UAV. There are dense and tall buildings, sparse factories, mountains with forest, bare ground and rivers in the covered region. The poses of each image and the digital surface model (DSM) have been computed by photogrammetric

TABLE 1 Details of the evaluation datasets.

Dataset	No. of images	Resolution	Property
KITTI	5520	370 × 1226	Street views; little viewpoint change
HPatches	696	Varied	2D scenes: Case A: illumination change Case B: viewpoint change
ScanNet	5578	968 × 1296	Indoor scenes; little viewpoint change
ZY3	3200	Nadir: 600 × 600 Backward: 500 × 550	Satellite; 22° intersection angle
UAV	7767	936 × 624	Aerial: Case A (normal photography): nadir–nadir matching Case B (oblique photography): nadir–oblique matching (about 40° intersection angle)

software. First, we carried out an experiment on the adjacent nadir view images (subset A). We then matched the nadir view images with the backward and forward views, respectively (subset B), for the subsequent experiments.

Metrics

In this paper, three widely used indicators are used to measure the performance of the different methods. The first metric is the pose recovery accuracy, which is measured by the area under the cumulative error curve (AUC), and is computed according to the number of images whose pose error is under a given threshold. We first calculate the pose error of each image pair, and then, we calculate the recall rate, which is the ratio between the number of image pairs whose pose error is under a given threshold x and the number of the image pairs. By changing the threshold, we can draw a recall-error curve as below. AUC@x is then calculated as a/x where a is the area under the curve at pose error threshold x, as shown in Figure 1.

The second metric is the epipolar accuracy, which is the ratio between the number of corresponding points whose epipolar accuracy is under a given threshold and the number of all the matched corresponding points after random sample consensus (RANSAC). The epipolar accuracy of a reference point m_1 is the distance between its corresponding point m_2 and the epipolar line in the right image $l=\mathrm{F}m_1$ where F is the ground truth fundamental matrix. The epipolar accuracies of each image in a dataset are summed and averaged as the final epipolar accuracy of the different methods. The third metric is the matched point number.

First, each combination of an extractor, descriptor, and matcher was used to match the points, and RANSAC was used for the outlier rejection. We use the default parameters of RANSAC in OpenCV on most datasets. In the KITTI and HPatches datasets, the threshold is set to 3 pixels when calculating the fundamental matrix and homography matrix. In the ScanNet dataset, the threshold is set to 1 pixel when calculating the essential matrix. As for the UAV dataset, we first convert the coordinates of the feature points to the normalised coordinate system, so the threshold is set to the reciprocal of the average focal length of the left and right images. We then evaluated the performance of the combinations for specific datasets. For the ZY3 satellite dataset, which has no homography

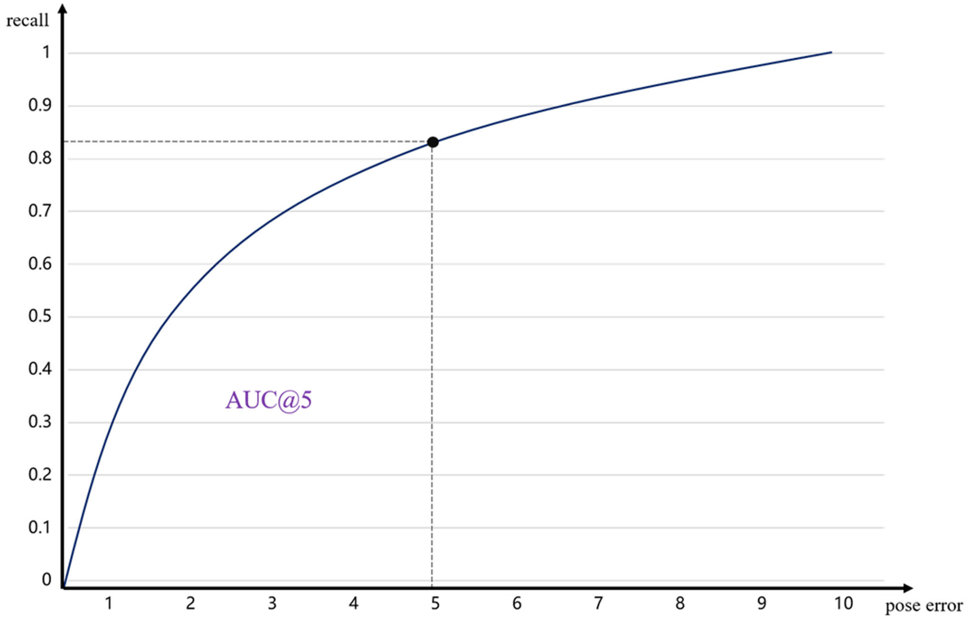

FIGURE 1 AUC@x is the ratio of the area under the curve at pose error threshold x and x (here $x=5$).

relationships, we evaluated only the epipolar accuracy. For the planar HPatches dataset, we evaluated only the AUC because homography estimation is adequate to evaluate this type of imagery. For the other datasets, we evaluated both the AUC and the epipolar accuracy. Matched point numbers are reported for all the datasets.

The calculation of the AUC was different for the specific datasets. For the KITTI dataset, where the fundamental matrix is provided as the ground truth, we calculated the symmetric geometry distance (SGD) as the pose error, according to (Bian et al., 2019). First, we randomly chose *N* (here 400) points in the left image, and calculated the epipolar lines in the right image using the calculated and the ground-truth fundamental matrices, respectively. One point in each calculated epipolar line is randomly selected and its distance to the ground truth epipolar line is calculated. We then calculated the arithmetic mean of all the *N* distances. Second, we repeated the first step, but using the epipolar lines in the left image. Third, we repeated the first and second steps by swapping the ground truth and the predicted fundamental matrix. The SGD was then calculated as the average of these four distances. Finally, the AUC could be calculated from the SGD, and we report it at thresholds of 1, 3, 5 and 10 pixels. For the HPatches dataset, we used the homography matrix to evaluate the pose recovery accuracy. The four corners of the left image were projected to the right image by the estimated and ground-truth homography, respectively. The average distance of the two point sets was regarded as the pose error. For the ScanNet and UAV datasets, we report the AUC of the pose error (the maximum of the angular errors in rotation and translation), according to (Sarlin et al., 2020).

Results for the KITTI dataset

The results for the KITTI dataset are shown in Figure 2. We first analyse the matcher. Generally speaking, in both pose recovery and epipolar accuracy, the traditional NNDR performs better than the deep learning-based SuperGlue, but the difference is not that significant. When using ORB, ASLFeat, SuperPoint and UnSuperPoint as descriptors, SuperGlue performs worse than the handcrafted matcher. Meanwhile, when using SIFT and

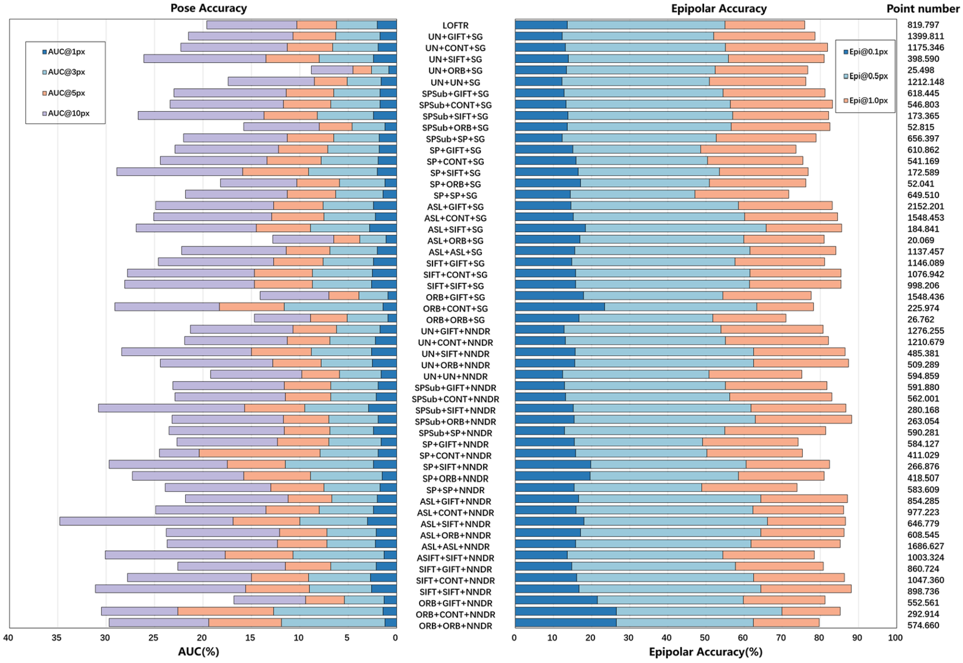

FIGURE 2 Results of the pose recovery and epipolar accuracy for the KITTI dataset. We report the AUC score for the pose recovery accuracy, epipolar accuracy and matching points. ASL, ASLFeat; CONT, ContextDesc; SG, SuperGlue; SP, SuperPoint; SPSub, Sub-SuperPoint; Un, UnSuperPoint.

ContextDesc as the descriptors, the matching performance is similar. This indicates that SuperGlue is sensitive to the descriptor. For the descriptors, an interesting observation is, for any extractor, the SIFT descriptor is always the best choice for pose recovery in the KITTI dataset. Among the deep learning-based descriptors, ContextDesc performs well and is stable, and is second only to SIFT. For the extractors, it is difficult to decide which one is the best, except that the ORB extractor shows significant advantages in pose recovery and epipolar accuracy. This can be explained by the fact that the ORB extractor is specially designed for such scenes.

Overall, the combination of ORB+ContextDesc+NNDR obtains the best AUC score at a 5-pixel accuracy, ASL+SIFT+NNDR obtains the best 10-pixel accuracy, while many combinations, such as ALS+SIFT+SuperGlue, SPSub+ORB+NNDR, and SIFT+SIFT+NNDR, obtain similar high performances in 1-pixel epipolar accuracy.

Results for the HPatches dataset

The result for HPatches dataset are shown in the Figure 3. For the scenes with illumination changes, the SuperGlue matcher is sensitive to the descriptor, as with the KITTI dataset, while NNDR is more stable for different combinations. As for the descriptors, GIFT and ContextDesc perform relatively well, followed by SIFT. For the extractors, it is difficult to say which one is the best, but using the ORB extractor, in some cases, yields extremely poor results, and Sub-SuperPoint performs better than SuperPoint in most cases.

For the viewpoint change scenes, the use of SuperGlue results in many extremely poor results, compared with NNDR, especially when using ORB as the descriptor. In most cases, GIFT and ContextDesc rank first and second, respectively, among the descriptors. For the extractors, UnSuperPoint, ASLFeat and ORB are not recommended as they result in relatively poor results, compared with the other extractors.

Overall, the combination of Sub-SuperPoint+GIFT+SuperGlue performs the best for the HPatches dataset (including both the illumination change and viewpoint change scenes). However, many other combinations can approach this performance.

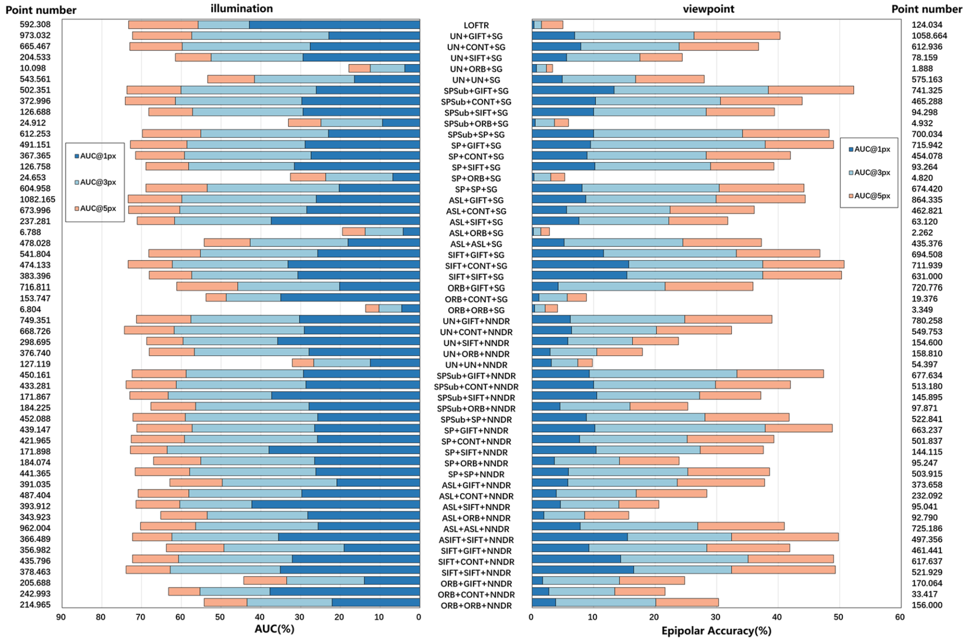

FIGURE 3 AUC scores for the planar HPatches datasets. ASL, ASLFeat; CONT, ContextDesc; SG, SuperGlue; SP, SuperPoint; SPSub, Sub-SuperPoint; Un, UnSuperPoint.

Results for the ScanNet dataset

The performance of the different methods is shown in Figure 4. For the matcher, the performance of SuperGlue varies. It creates more bad results than NNDR, for example, and there are 10 relatively short bars in the AUC scores when using SuperGlue, but only five short bars when using NNDR.

In pose recovery, the deep learning-based methods, such as ASLFeat + NNDR and (Sub-) SuperPoint+SuperGlue, perform well. The GIFT descriptor can be combined well with many extractors. GIFT performs the best for the UnSuperPoint extractor. As for the epipolar accuracy, the SIFT extractor always obtains a high performance, for all the descriptors and matchers. The Sub-SuperPoint extractor performs slightly better in pose recovery and epipolar accuracy than SuperPoint.

Overall, the combination of Sub-SuperPoint + SuperPoint + SuperGlue performs the best in pose recovery (20°), and that of SIFT + ContextDesc + NNDR performs the best in epipolar accuracy (1 pixel).

Results for the ZY3 satellite dataset

The result for ZY satellite dataset are shown in the Figure 5. The combination of Sub-SuperPoint + SIFT + SuperGlue achieves the best 1-pixel epipolar accuracy; however, NNDR performs better than SuperGlue in general. The three combinations of Sub-SuperPoint + ContextDesc + SuperGlue, SuperPoint + SIFT + SuperGlue, and Sub-SuperPoint + ContextDesc + NNDR perform well and are second only to the top combination. It should be noted that all the deep learning models were trained on the close-range datasets, and we directly applied the pretrained model to the satellite data. This indicates that most of the deep learning-based methods, including extractors,

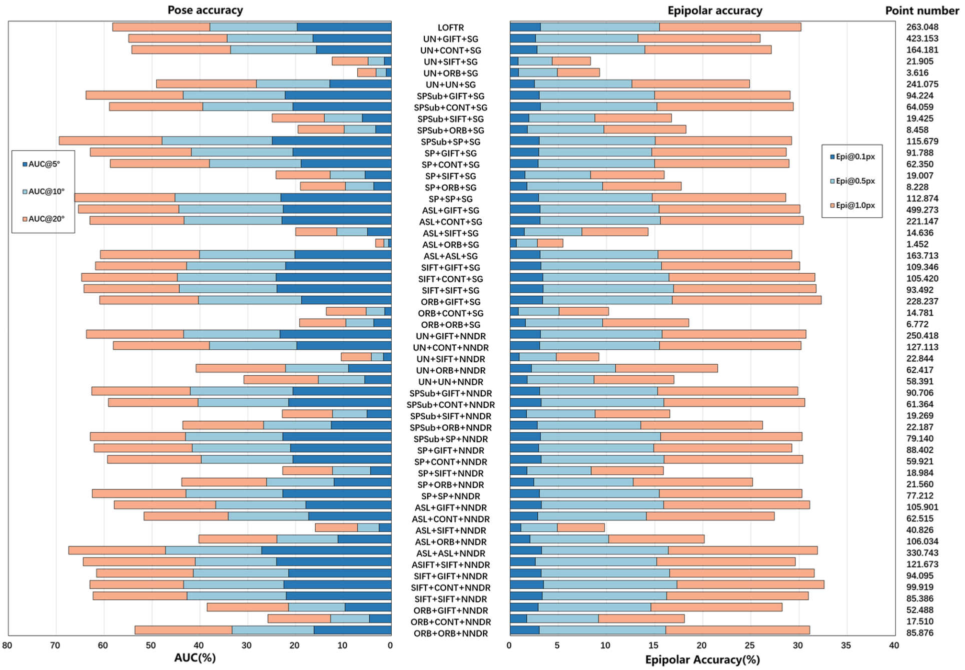

FIGURE 4 AUC scores for different combinations with the ScanNet dataset. ASL, ASLFeat; CONT, ContextDesc; SG, SuperGlue; SP, SuperPoint; SPSub, Sub-SuperPoint; Un, UnSuperPoint.

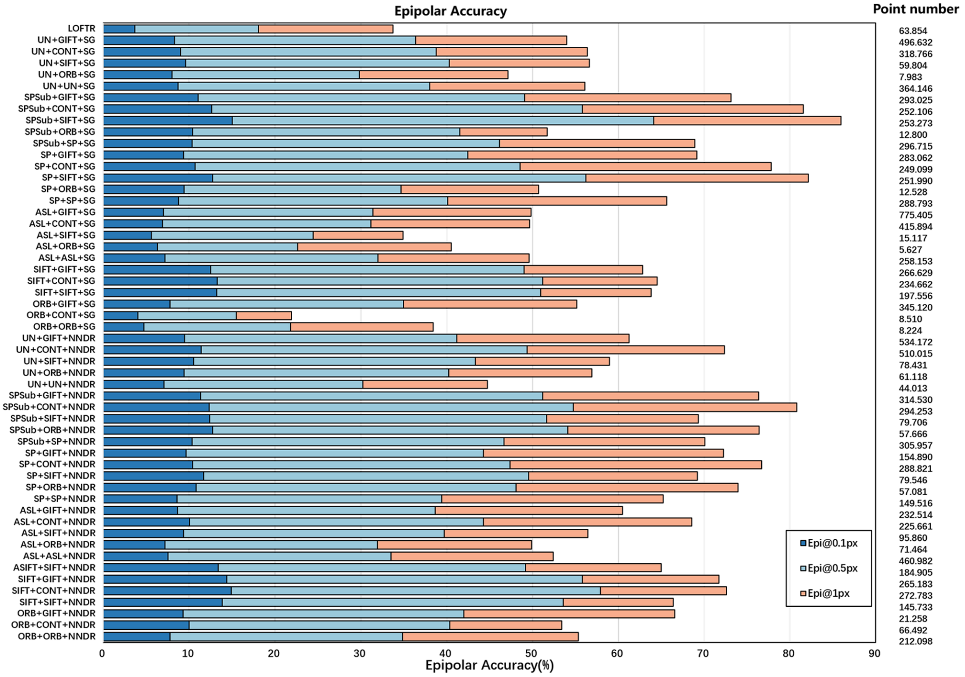

FIGURE 5 Epipolar accuracy of different combinations with the ZY3 dataset. ASL, ASLFeat; CONT, ContextDesc; SG, SuperGlue; SP, SuperPoint; SPSub, Sub-SuperPoint; Un, UnSuperPoint.

descriptors, and matchers, are robust for satellite image matching, despite the significant difference between close-range images and satellite images.

Results for the UAV dataset

The oblique UAV dataset is one of the key datasets for the analysis. We give the results for the nadir-to-nadir, nadir-to-backward, and nadir-to-forward matching, respectively, in Figures 6–8.

One observation from Figure 6 is that the use of SuperGlue introduces some extremely poor results, such as its combination with the ORB extractor or ORB descriptor. ORB is actually not the only factor resulting in the poor results because, with NNDR, it performs relatively well. Nevertheless, if ORB is removed, SuperGlue performs almost as well as NNDR, and most of the combinations are effective.

For the cases of different view angles, as shown in Figure 7 (nadir-backward) and Figure 8 (nadir-forward), many combinations perform very poorly. In most cases, the SIFT and ORB descriptors result in poor results, but the original SIFT extractor and descriptor combination is among the best. However, the deep learning-based ContextDesc is relatively robust. Many combinations perform well using ContextDesc as the descriptor. The ORB extractor performs relatively and naturally poorly since the large view angle change cases are not suitable for the ORB extractor. The combination of UnSuperPoint extractor and GIFT or the ContextDesc descriptor functions as well as the best combinations in recovering poses, but in other cases, the UnSuperPoint extractor performs poorly.

Overall, the results for the three cases demonstrate that many combinations, such as SPSub+GIFT+SuperGlue/NNDR and SIFT+SIFT/ContextDesc+NNDR/SuperGlue, can obtain top performances. It is also suggested that we should avoid those combinations that obtain very poor results.

SUMMARY AND ANALYSIS

Discovering effective combinations for all cases

The performances of the different extractor/descriptor/matcher combinations vary between the different datasets. What we wish to identify is those combinations which are effective and robust in various situations. In the following, we quantitively evaluate the performance of all the different combinations, taking account of all the datasets, with simple and clear metrics.

First, for a combination in a dataset, for example, the UAV dataset, it has a unique rank score among all the combinations in eight evaluation indicators: AUC@0.5, AUC@1, AUC@1.5, AUC@2.0, epi@0.1, epi@0.5, epi@1, and matched number, respectively. Second, the performance of this combination is ranked according to the mean value of the eight rank scores. Finally, according to the rank score of each combination in each dataset, we calculate two comprehensive metrics. The first is called the "average rank", which is calculated by averaging the ranks of a combination in all of the datasets. The second is called the "absolute rank", which means that a combination must be ranked in at least the top *x*% for all of the datasets.

Some conclusions can be drawn from Table 2. First, SIFT is still one of the most stable and high-performance descriptors for different types of platforms and illumination/viewpoint changes. As for the deep learning methods, the sub-pixel version of the SuperPoint extractor (i.e., the SPSub extractor we designed) is relatively stable (with four combinations in the top 10–20% interval of the average rank), and approaches SIFT in most cases, but is more sensitive to the descriptor. SPSub outperforms SIFT and the other handcrafted methods mainly on the datasets with large viewpoint changes. It seems that the deep learning methods are more applicable to scenes with larger viewpoint changes. The reason for this may be that the deep learning methods have simulated a large amount of

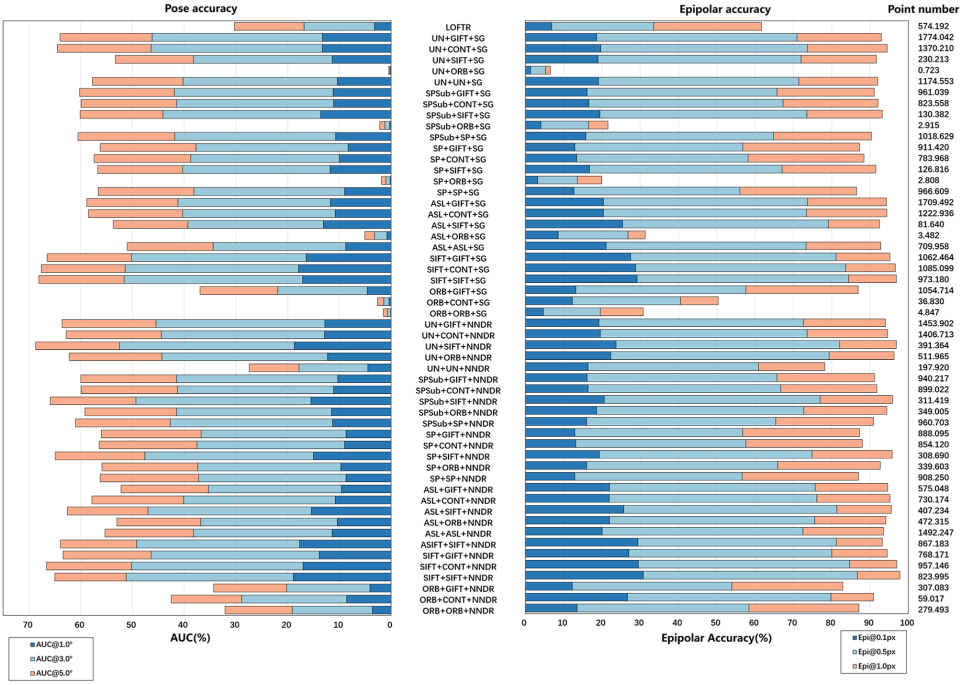

FIGURE 6 Pose recovery and epipolar accuracy of the different combinations for the nadir views of the UAV dataset. ASL, ASLFeat; CONT, ContextDesc; SG, SuperGlue; SP, SuperPoint; SPSub, Sub-SuperPoint; Un, UnSuperPoint.

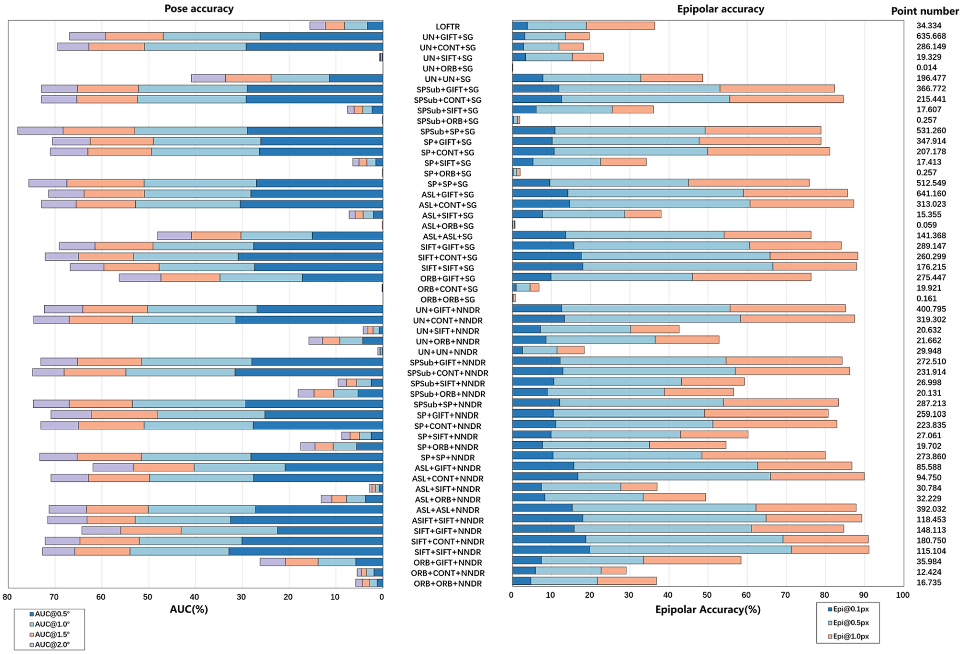

FIGURE 7 Pose recovery and epipolar accuracy of the different combinations for the nadir and backward views of the UAV dataset. ASL, ASLFeat; CONT, ContextDesc; SG, SuperGlue; SP, SuperPoint; SPSub, Sub-SuperPoint; Un, UnSuperPoint.

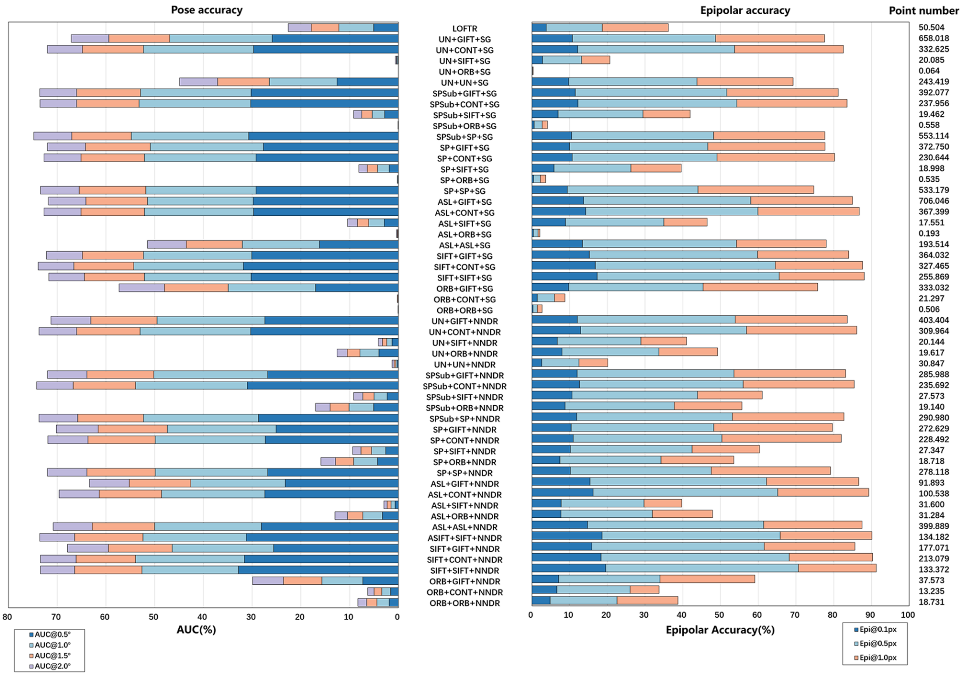

FIGURE 8 Pose recovery and epipolar accuracy of the different combinations for the nadir and forward views of the UAV dataset. ASL, ASLFeat; CONT, ContextDesc; SG, SuperGlue; SP, SuperPoint; SPSub, Sub-SuperPoint; Un, UnSuperPoint.

image pairs with large viewpoint changes, as well as introducing data augmentation, including blurring and illumination changes, and then learn from them to adapt to complex situations.

As for the descriptors, the ContextDesc descriptor achieves a very good performance that is similar to that of the SIFT descriptor. It even ranks number one (with the SIFT extractor) in absolute rank, and beats the original SIFT descriptor.

As for the matcher, the deep learning-based SuperGlue actually performs reasonably well (it appears twice in the top 10% average rank and once in the top 50% absolute rank), but NNDR generally shows a better performance.

As for the combinations, SIFT+ContextDesc+NNDR and SIFT+SIFT+NNDR (i.e., the original setting of SIFT) are the most recommended from Table 2, and they achieve both the best average ranking and the best absolute ranking.

TABLE 2 Ranking of the various combinations.

	Ranking of the different combinations	
Top *x*%	**Average rank**	**Absolute rank**
10%	SIFT+SIFT+NNDR SIFT+CONT+NNDR SIFT+CONT+SG SIFT+SIFT+SG ASIFT+SIFT+NNDR	None
20%	SPSub+CONT+NNDR SPSub+CONT+SG SPSub+GIFT+SG SPSub+SP+NNDR SIFT+GIFT+SG	None
30%	UN+CONT+NNDR ASL+GIFT+SG SPSub+GIFT+NNDR ASL+CONT+SG SPSub+SP+SG ASL+ASL+NNDR	None
40%	SP+CONT+NNDR SIFT+GIFT+NNDR UN+GIFT+NNDR ASL+CONT+NNDR SP+SP+NNDR	SIFT+CONT+NNDR
50%	SP+GIFT+NNDR SPSub+SIFT+NNDR SP+SP+SG SP+SIFT+NNDR UN+CONT+SG ASL+GIFT+NNDR	SIFT+SIFT+NNDR SPSub+CONT+SG

Note: "Average rank" means the average performance of each combination over all the datasets; and "absolute rank" means a combination must be ranked in at least the top *x*% for all the datasets.

ASL, ASLFeat; CONT, ContextDesc; SG, SuperGlue; SP, SuperPoint; SPSub, Sub-SuperPoint; Un, UnSuperPoint.

Nevertheless, there is still likely great potential for the deep learning-based methods. Currently, the deep learning models are only trained on close-range datasets, such as COCO, but if the quantity and types of remote sensing training data (e.g., large-scale satellite and UAV datasets) increases, then the deep learning methods may show significant improvements.

Some examples of matching results of top three combinations of the absolute rank are shown in Figure 9.

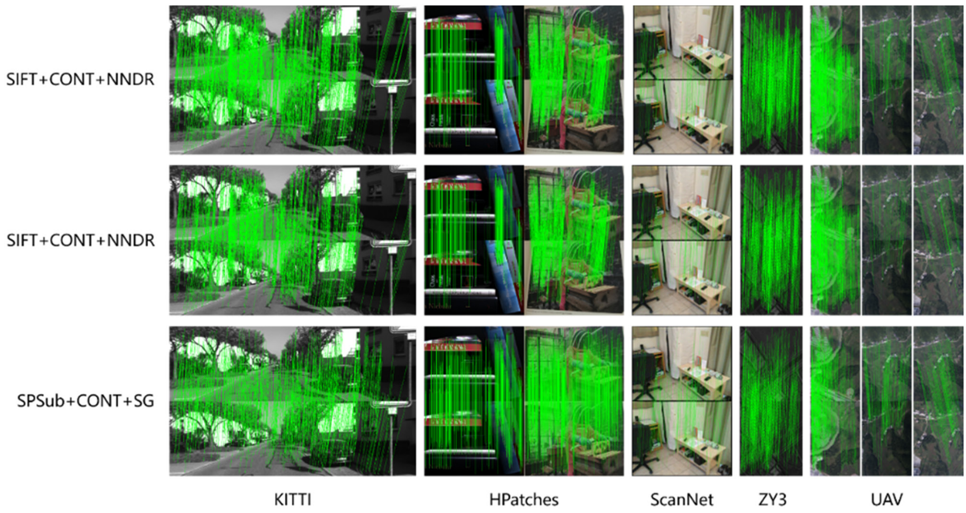

FIGURE 9 Examples of matching results of top combinations on five datasets. HPatches includes illumination (left) and viewpoint (right) scenes. UAV dataset includes nadir views (left), nadir and backward views (medium), and nadir and forward views (right).

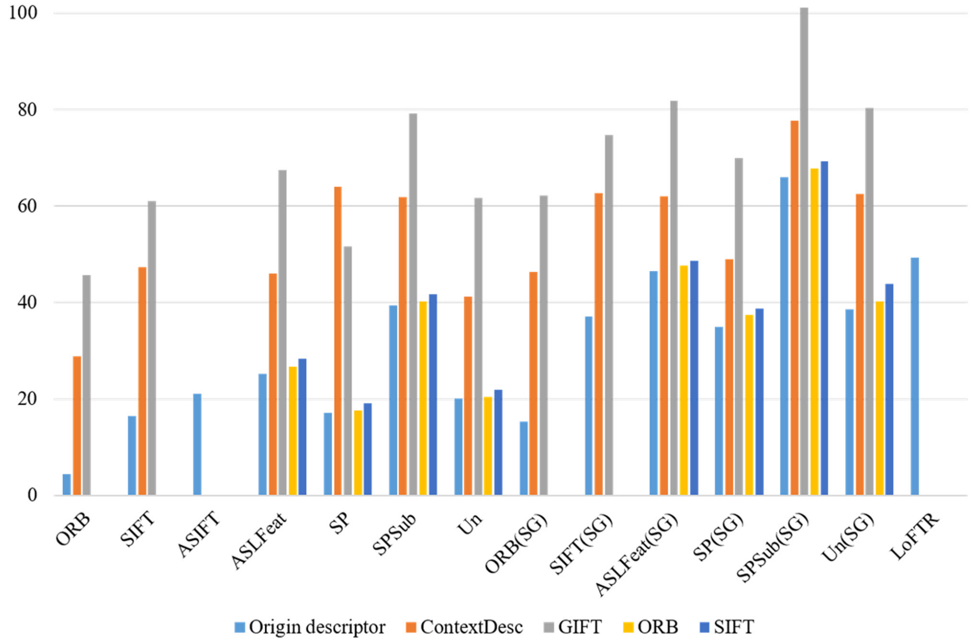

FIGURE 10 Matching time of each method on 100 image pairs from the KITTI dataset. Origin descriptor means using their own descriptor. SG means SuperGlue is used as the matcher; otherwise, NNDR is used.

Efficiency of the matching method

In many applications, such as SLAM, the efficiency of the matching method is vital. We calculated the runtimes for the various methods/combinations on the KITTI dataset by running each method on 100 image pairs. The experiment was carried out on a TITAN RTX GPU and Intel i9-9900KF CPU. Each combination was carried out three times, and the final time is their average, as shown in Figure 10. According to Figure 10, the original ORB is the fastest method. Using the SIFT extractor as the baseline, the deep learning-based SuperPoint and UnSuperPoint are on the same efficiency level, but SPSub and ASLFeat are slower than SIFT. As for the descriptors, ContextDesc and GIFT are slower than the other descriptors. The NNDR matcher is nearly twice as fast as the deep learning-based SuperGlue. It is concluded that, currently, the deep learning-based methods are slower than the conventional methods, but their productivity is acceptable in many applications, as most of them can process 100 images within 1 min.

CONCLUSIONS

In this paper, we have provided a comprehensive evaluation of both the conventional and deep learning-based image-matching methods, covering a wide range of keypoint extraction, description, and matching algorithms, on various datasets from different platforms, to answer the questions whether the deep learning-based methods have comprehensively surpassed the conventional methods, and which are the high-performance and stable combinations of extraction, description, and matching methods on various scenes. The experiments revealed that, first, the performance of the different combinations varies in individual datasets, and it is difficult to say which combination is the best. Second, when using more comprehensive indicators for all the datasets, that is, the average rank and absolute rank, the combination of SIFT + ContextDesc + NNDR and the original SIFT achieve the best results. This means that the SIFT descriptor is still the top matching method, even though deep learning-based methods have been developed for years now. Nevertheless, the deep learning-based ContextDesc can be an effective alternative for the SIFT descriptor, and SPSub is also effective. Third, the handcrafted methods are generally faster than the deep learning-based methods; however, the gap is not huge. Finally, we believe that there is still much room for improvement with the deep learning-based methods as large open-source aerial and satellite training datasets remain to be constructed, and specific methods for remote sensing images remain to be developed.

ACKNOWLEDGEMENTS

This work was supported by the National Key Research and Development Program of China (grant No. 2019YFE0125300), National Natural Science Foundation of China (grant No. 42171430), and the State Key Program of the National Natural Science Foundation of China (grant No. 42030102).

ORCID

Shunping Ji https://orcid.org/0000-0002-3088-1481
Yongjun Zhang https://orcid.org/0000-0001-9845-4251

REFERENCES

Aanæs, H., Dahl, A.L. & Steenstrup, P.K. (2012) Interesting interest points. *International Journal of Computer Vision*, 97(1), 18–35. Available from: https://doi.org/10.1007/s11263-011-0473-8

Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E. & Dahl, A.B. (2016) Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2), 153–168. Available from: https://doi.org/10.1007/s11263-016-0902-9

Abdel-Hakim, A.E. & Farag, A.A. (2006) CSIFT: a SIFT descriptor with color invariant characteristics. In: *Proceedings of the 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*. New York, NY: Ieee. Available from: https://doi.org/10.1109/CVPR.2006.95

Balntas, V., Lenc, K., Vedaldi, A. & Mikolajczyk, K. (2017) HPatches: a benchmark and evaluation of handcrafted and learned local descriptors. Proceedings of the IEEE conference on computer vision and pattern recognition. Available from: https://doi.org/10.1109/CVPR.2017.410

Balntas, V., Riba, E., Ponsa, D. & Mikolajczyk, K. (2016) Learning local feature descriptors with triplets and shallow convolutional neural networks. Proceedings of the Bmvc. Available from: https://doi.org/10.5244/C.30.119

Bas, S. & Ok, A.O. (2021) A new productive framework for point-based matching of oblique aircraft and UAV-based images. *The Photogrammetric Record*, 36(175), 252–284. Available from: https://doi.org/10.1111/phor.12374

Bay, H., Ess, A., Tuytelaars, T. & van Gool, L. (2008) Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3), 346–359. Available from: https://doi.org/10.1016/j.cviu.2007.09.014

Bian, J.-W., Wu, Y.-H., Zhao, J., Liu, Y., Zhang, L., Cheng, M.M. & Reid, I. (2019) An evaluation of feature matchers for fundamental matrix estimation. In: *Proceedings of the British Machine Vision Conference (BMVC)*. Available from: https://doi.org/10.48550/arXiv.1908.09474

Calonder, M., Lepetit, V., Strecha, C. & Fua, P. (2010) Brief: binary robust independent elementary features. In: *Proceedings of the European conference on computer vision*. New York, NY: Springer. Available from: https://doi.org/10.1007/978-3-642-15561-1_56

Christiansen, P.H., Kragh, M.F., Brodskiy, Y. & Karstoft, H. (2019) Unsuperpoint: end-to-end unsupervised interest point detector and descriptor. *arXiv Preprint at arXiv*: 190704011. Available from: https://doi.org/10.48550/arXiv.1907.04011

Cuturi, M. (2013) Lightspeed computation of optimal transportation distances. *Advances in Neural Information Processing Systems*, 26(2), 2292–2300. Available from: https://doi.org/10.48550/arXiv.1306.0895

Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T. & Nießner, M. (2017) Scannet: richly-annotated 3d reconstructions of indoor scenes. Proceedings of the IEEE conference on computer vision and pattern recognition. Available from: https://doi.org/10.1109/CVPR.2017.261

Detone, D., Malisiewicz, T. & Rabinovich, A. (2018) Superpoint: self-supervised interest point detection and description. Proceedings of the IEEE conference on computer vision and pattern recognition workshops. Available from: https://doi.org/10.48550/arXiv.1712.07629

Dong, J. & Soatto, S. (2015) Domain-size pooling in local descriptors: DSP-SIFT. Proceedings of the IEEE conference on computer vision and pattern recognition. Available from: https://doi.org/10.1109/CVPR.2015.7299145

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A. & Sattler, T. (2019) D2-net: a trainable cnn for joint description and detection of local features. Proceedings of the ieee/cvf conference on computer vision and pattern recognition. Available from: https://doi.org/10.1109/CVPR.2019.00828

Fan, B., Kong, Q., Wang, X., Wang, Z., Xiang, S., Pan, C., Pan, C. & Fua, P. (2019) A performance evaluation of local features for image-based 3D reconstruction. *IEEE Transactions on Image Processing*, 28(10), 4774–4789. Available from: https://doi.org/10.1109/TIP.2019.2909640

Förstner, W. & Gülch, E. (1987) A fast operator for detection and precise location of distinct points, corners and centres of circular features. In: *Proceedings of the Proc ISPRS intercommission conference on fast processing of photogrammetric data*. Interlaken, Switzerland: Elsevier.

Geiger, A., Lenz, P. & Urtasun, R. (2012) Are we ready for autonomous driving? The Kitti vision benchmark suite. In: *Proceedings of the 2012 IEEE conference on computer vision and pattern recognition*. New York, NY: IEEE. Available from: https://doi.org/10.1109/CVPR.2012.6248074

Han, X., Leung, T., Jia, Y., Sukthankar, R. & Berg, A.C. (2015) Matchnet: unifying feature and metric learning for patch-based matching. Proceedings of the IEEE conference on computer vision and pattern recognition. Available from: https://doi.org/10.1109/CVPR.2015.7298948

Harris, C. & Stephens, M. (1988) A combined corner and edge detector. Proceedings of the Alvey vision conference. Manchester, UK. Available from: https://doi.org/10.5244/C.2.23

Heinly, J., Dunn, E. & Frahm, J.-M. (2012) Comparative evaluation of binary features. In: *Proceedings of the European Conference on Computer Vision*. New York, NY: Springer. Available from: https://doi.org/10.1007/978-3-642-33709-3_54

Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K.M. & Trulls, E. (2021) Image matching across wide baselines: from paper to practice. *International Journal of Computer Vision*, 129(2), 517–547. Available from: https://doi.org/10.1007/S11263-020-01385-0

Komorowski, J., Czarnota, K., Trzcinski, T., Dabala, L. & Lynen, S. (2018) Interest point detectors stability evaluation on ApolloScape dataset. Proceedings of the European Conference on Computer Vision (ECCV) Workshops. Available from: https://doi.org/10.1007/978-3-030-11021-5_45

Kumar Bg, V., Carneiro, G. & Reid, I. (2016) Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. Proceedings of the IEEE conference on computer vision and pattern recognition. Available from: https://doi.org/10.1109/CVPR.2016.581

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C.L. (2014) Microsoft coco: common objects in context. In: *Proceedings of the European conference on computer vision*. New York, NY: Springer. Available from: https://doi.org/10.1007/978-3-319-10602-1_48

Lindeberg, T. (1998) Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2), 79–116. Available from: https://doi.org/10.1023/A:1008045108935

Liu, Y., Shen, Z., Lin, Z., Peng, S., Bao, H. & Zhou, X. (2019) Gift: learning transformation-invariant dense visual descriptors via group cnns. *Advances in Neural Information Processing Systems*, 32(11), 6990–7001. Available from: https://doi.org/10.48550/arXiv.1911.05932

Lowe, D.G. (2004) Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91–110. Available from: https://doi.org/10.1023/B:VISI.0000029664.99615.94

Luo, Z., Shen, T., Zhou, L., Zhang, J., Yao, Y., Li, S., Fang, T. & Quan, L. (2019) Contextdesc: local descriptor augmentation with cross-modality context. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Available from: https://doi.org/10.1109/CVPR.2019.00263

Luo, Z., Zhou, L., Bai, X., Chen, H., Zhang, J., Yao, Y., Li, S., Fang, T. & Quan, L. (2020) Aslfeat: learning local features of accurate shape and localization. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Available from: https://doi.org/10.1109/CVPR42600.2020.00662

Ma, J., Jiang, X., Fan, A., Jiang, J. & Yan, J. (2021) Image matching from handcrafted to deep features: a survey. *International Journal of Computer Vision*, 129(1), 23–79. Available from: https://doi.org/10.1007/s11263-020-01359-2

Mair, E., Hager, G.D., Burschka, D., Suppa, M. & Hirzinger, G. (2010) Adaptive and generic corner detection based on the accelerated segment test. In: *Proceedings of the European conference on computer vision*. New York, NY: Springer. Available from: https://doi.org/10.1007/978-3-642-15552-9_14

Maiwald, F. & Maas, H.G. (2021). An automatic workflow for orientation of historical images with large radiometric and geometric differences. *The Photogrammetric Record*, 36(174), 77–103. Available from: https://doi.org/10.1111/phor.12363

Mikolajczyk, K. & Schmid, C. (2005) A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10), 1615–1630. Available from: https://doi.org/10.1109/TPAMI.2005.188

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T. & Van Gool, L. (2005) A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1), 43–72. Available from: https://doi.org/10.1007/s11263-005-3848-x

Mishchuk, A., Mishkin, D., Radenovic, F. & Matas, J. (2017) Working hard to know your neighbor's margins: local descriptor learning loss. *Advances in Neural Information Processing Systems*, 30(12), 4829–4840. Available from: https://doi.org/10.48550/arXiv.1705.10872

Mohammadi, N., Sedaghat, A. & Jodeiri Rad, M. (2022) Rotation-invariant self-similarity descriptor for multi-temporal remote sensing image registration. *The Photogrammetric Record*, 37(177), 6–34. Available from: https://doi.org/10.1111/phor.12402

Moravec, H.P. (1977) Techniques towards automatic visual obstacle avoidance.

Moreels, P. & Perona, P. (2007) Evaluation of features detectors and descriptors based on 3d objects. *International Journal of Computer Vision*, 73(3), 263–284. Available from: https://doi.org/10.1007/s11263-006-9967-1

Morel, J.-M. & Yu, G. (2009) ASIFT: a new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2), 438–469. Available from: https://doi.org/10.1137/080732730

Mukherjee, D., Jonathan, W.Q. & Wang, G. (2015) A comparative experimental study of image feature detectors and descriptors. *Machine Vision and Applications*, 26(4), 443–466. Available from: https://doi.org/10.1007/s00138-015-0679-9

Ono, Y., Trulls, E., Fua, P. & Yi, K.M. (2018) LF-Net: learning local features from images. *Advances in Neural Information Processing Systems*, 31(11), 6237–6247. Available from: https://doi.org/10.48550/arXiv.1805.09662

Parente, L., Chandler, J.H. & Dixon, N. (2021) Automated registration of SfM-MVS multitemporal datasets using terrestrial and oblique aerial images. *The Photogrammetric Record*, 36(173), 12–35. Available from: https://doi.org/10.1111/phor.12346

Quinlan, J.R. (1986) Induction of decision trees. *Machine Learning*, 1(1), 81–106. Available from: https://doi.org/10.1007/BF00116251

Rublee, E., Rabaud, V., Konolige, K. & Bradski, G. (2011) ORB: an efficient alternative to SIFT or SURF. In: *Proceedings of the 2011 International conference on computer vision*. New York, NY: Ieee. Available from: https://doi.org/10.1109/ICCV.2011.6126544

Sarlin, P.-E., Detone, D., Malisiewicz, T. & Rabinovich, A. (2020) Superglue: learning feature matching with graph neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Available from: https://doi.org/10.1109/CVPR42600.2020.00499

Savinov, N., Seki, A., Ladicky, L., Sattler, T. & Pollefeys, M. (2017) Quad-networks: unsupervised learning to rank for interest point detection. Proceedings of the IEEE conference on computer vision and pattern recognition. Available from: https://doi.org/10.1109/CVPR.2017.418

Schonberger, J.L. & Frahm, J.-M. (2016) Structure-from-motion revisited. Proceedings of the IEEE conference on computer vision and pattern recognition. Available from: https://doi.org/10.1109/CVPR.2016.445

Schonberger, J.L., Hardmeier, H., Sattler, T. & Pollefeys, M. (2017) Comparative evaluation of hand-crafted and learned local features. Proceedings of the IEEE conference on computer vision and pattern recognition. Available from: https://doi.org/10.1109/CVPR.2017.736

Shen, X., Wang, C., Li, X., Yu, Z., Li, J., Wen, C., Cheng, M. & He, Z. (2019) Rf-net: an end-to-end image matching network based on receptive field. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Available from: https://doi.org/10.1109/CVPR.2019.00832

Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P. & Moreno-Noguer, F. (2015) Discriminative learning of deep convolutional feature point descriptors. Proceedings of the IEEE international conference on computer vision. Available from: https://doi.org/10.1109/ICCV.2015.22

Tian, Y., Fan, B. & Wu, F. (2017) L2-net: deep learning of discriminative patch descriptor in euclidean space. Proceedings of the IEEE conference on computer vision and pattern recognition. Available from: https://doi.org/10.1109/CVPR.2017.649

Tola, E., Lepetit, V. & Fua, P. (2009) Daisy: an efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5), 815–830. Available from: https://doi.org/10.1109/TPAMI.2009.77

Trajković, M. & Hedley, M. (1998) Fast corner detection. *Image and Vision Computing*, 16(2), 75–87. Available from: https://doi.org/10.1016/S0262-8856(97)00056-5

Verdie, Y., Yi, K., Fua, P. & Lepetit, V. (2015) Tilde: a temporally invariant learned detector. Proceedings of the IEEE conference on computer vision and pattern recognition. Available from: https://doi.org/10.1109/CVPR.2015.7299165

Yi, K.M., Trulls, E., Lepetit, V. & Fua, P. (2016) Lift: learned invariant feature transform. In: *Proceedings of the European conference on computer vision*. New York, NY: Springer. Available from: https://doi.org/10.1007/978-3-319-46466-4_28

Zieba, M., Semberecki, P., El-Gaaly, T. & Trzcinski, T. (2018) Bingan: learning compact binary descriptors with a regularized gan. *Advances in Neural Information Processing Systems*, 31(11), 3612–3622. Available from: https://doi.org/10.48550/arXiv.1806.06778

Zitova, B. & Flusser, J. (2003) Image registration methods: a survey. *Image and Vision Computing*, 21(11), 977–1000. Available from: https://doi.org/10.1016/S0262-8856(03)00137-9

How to cite this article: Ji, S., Zeng, C., Zhang, Y. & Duan, Y. (2023) An evaluation of conventional and deep learning-based image-matching methods on diverse datasets. *The Photogrammetric Record*, 00, 1–23. Available from: https://doi.org/10.1111/phor.12445