This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

StateNet: Deep State Learning for Robust Feature Matching of Remote Sensing Images

Jiaxuan Chen, Shuang Chen, Xiaoxian Chen, Yang Yang, *Member, IEEE*, and Yujing Rao

***Abstract*— Seeking good correspondences between two images is a fundamental and challenging problem in the remote sensing (RS) community, and it is a critical prerequisite in a wide range of feature-based visual tasks. In this article, we propose a flexible and general deep state learning network for both rigid and nonrigid feature matching, which provides a mechanism to change the state of matches into latent canonical forms, thereby weakening the degree of randomness in matching patterns. Different from the current conventional strategies (i.e., imposing a global geometric constraint or designing additional handcrafted descriptor), the proposed StateNet is designed to perform alternating two steps: 1) recalibrates matchwise feature responses in the spatial domain and 2) leverages the spatially local correlation across two sets of feature points for transformation update. For this purpose, our network contains two novel operations: adaptive dual-aggregation convolution (ADAConv) and point rendering layer (PRL). These two operations are differentiable, so our network can be inserted into the existing classification architecture to reduce the cost of establishing reliable correspondences. To demonstrate the robustness and universality of our approach, extensive experiments on various real image pairs for feature matching are conducted. Experiments reveal the superiority of our StateNet significantly over the state-of-the-art alternatives.**

***Index Terms*— Adaptive state learning (ASL), deep learning, feature matching, image matching, image registration.**

I. Introduction

As A fundamental and pivotal research in computer vision, establishing reliable point correspondences between two images of the same or similar scenes is the core of many vision-based tasks. Feature matching is a very important basic tool for image matching and plays a prominent role in many important areas, such as remote sensing (RS), photogrammetry, and medical imaging. Particularly in RS community, matching-based tasks, including image registration and fusion, 3-D reconstruction, panorama production, object identification, and tracking, start by assuming that the high-quality correspondences have been successfully established.

Feature matching possesses a combinatorial nature, making the matching space huge, which causes highly computation complexity in non-Pareto criterion complex optimization. Even without considering outliers, matching N points to another N points may lead to a total of $N!$ permutations [1], [2]. To relieve the computation complexity, a popular strategy operates on sets of matches that are estimated by nearest neighbor (NN) search, i.e., ignoring the assignment structure. Under the circumstances, seeking good correspondences task boils down to detecting and rejecting the outliers from the given putative point correspondences. Therefore, this article focuses on determining the correctness of each match in arbitrary putative matches.

Over the past few decades, a variety of robust estimators have been proposed for eliminating or alleviating the undue influence of mismatches. More interestingly, due to the unordered and irregular structure of sparse points, point-based learning for outlier rejection has only been introduced in recent years. With the advent of deep learning on point sets (PointNet) [3], leveraging learning techniques for establishing good correspondences has also been revisited and manifests promising results. The key here is that training the network in an end-to-end fashion to infer the putative matches as inliers or outliers and regress the relative pose encoded by the essential matrix [4], [5]. Nevertheless, if the underlying image transformation does not satisfy the parametric geometric model (e.g., epipolar geometry), these methods also tend to severely degrade. In order to accommodate complex scenarios, another recent trend has been toward constructing match representation by exploiting the local geometrical relationship. The key of these methods is to construct a proper match representation. However, the geometric transformation models between two images are various, making it difficult to design a general and robust handcrafted descriptor. More importantly, RS images often involve local distortions, which results in complex spatial relationships caused by the imaging viewpoint changes and ground relief variations. Furthermore, due to the vast radiometric differences, multimodal images captured by different sensors usually suffer from a high number of false matches. On the whole, multiple factors, including the lack of obvious context structure in matching space, the randomness of outliers distribution, and unknown nonrigid spatial transformations, make the general outlier rejection a challenging problem.

Contributions: To address the above challenges, we propose a unified deep learning architecture that directly takes sparse

Manuscript received March 11, 2021; revised October 10, 2021; accepted October 13, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 41971392, in part by the Graduate Research Innovation Fund of Yunnan Normal University under Grant YJSJJ21-B73 and Grant YJSJJ21-B75, and in part by the Yunnan Ten-Thousand Talents Program. *(Jiaxuan Chen and Shuang Chen contributed equally to this work.) (Corresponding author: Yang Yang.)*

Jiaxuan Chen, Shuang Chen, Yang Yang, and Yujing Rao are with the Laboratory of Pattern Recognition and Artificial Intelligence, School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China (e-mail: jrbook_chen@foxmail.com; chenshuang283@163.com; yyang_ynu@163.com; yujing_rao@163.com).

Xiaoxian Chen is with the Laboratory of Pattern Recognition and Artificial Intelligence, School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China, and also with JD.com Inc., Beijing 100000, China (e-mail: xxianchen@foxmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3120768.

Digital Object Identifier 10.1109/TNNLS.2021.3120768

2162-237X © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on February 10,2023 at 14:55:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

putative match sets as input and outputs matching results. Without consideration for the data representation of sparse matches, the putative matches often contain (many) false-positive (FP) pairs (also known as outliers), and they are randomly scattered. Moreover, unknown deformations make the spatial distribution of the two feature points may change significantly. These two inevitable factors can result in complex matching patterns, which sharply increases the cost of screening out the outliers. Thus, from a novel perspective, this work provides a mechanism to change the state of matches into latent canonical forms for weakening the degree of randomness in matching patterns, with no explicit loss or constraint in enforcing the canonicalization. More specifically, the main contributions of this article are as follows.

1) The concept of regularity for evaluating the degree of randomness in matching patterns is presented, namely, state entropy. Under this regularity condition, target loss function with the entropy regularization can guide the searching of reliable correspondences, thus implementing consistent performance gains throughout the training process.
2) We provide a specific cascading architecture for seeking the consensus of matching patterns, termed deep state learning. This network model alternately performs outlier suppression and transformation update by two modules: matching-neighbors selector and spatial aligner.
3) We introduce an adaptive dual-aggregation convolution (ADAConv) works on the local sparse matches directly, which encodes the global geometric features of unordered sparse matches in a learnable manner. ADAConv is the key technique for matchwise dependencies decoding in matching-neighbors selector.
4) We design a point rendering layer (PRL), which allows us to cast the arbitrary permutation of feature points into a potentially canonical order, thus leveraging the spatially local correlation of sparse matches under a unified network framework for capturing structural cues. By the collaborative use of matching-neighbors selector, the PRL also provides a spatial gating mechanism for suppressing outlier response.

The remainder of this article is organized as follows. Section II describes background material and related work. In Section III, we detail the architecture and design of the adaptive state learning (ASL) mechanism, which makes up the proposed end-to-end outlier rejection framework. Section IV illustrates the matching and registration performance of our method on various types of RS image pairs (as well as other computer vision images) with comparisons to other approaches, followed by some concluding remarks in Section V.

II. Related Work

Here, we briefly review the background material applied as a reference for the current research. In the literature, this material includes two kinds of paradigms: the first type aims to assign correspondences between two point sets and to recover the spatial transformation, whereas the second type constructs a set of putative correspondence and then filters false matches.

A. Point Set Registration

Point set registration (PSR) is the process of finding one-to-one correspondence of two point sets, which aims to determine the underlying global transformation that optimally aligns two point sets. In general, PSR formulates the matching problem as the estimation of a mixture of densities utilizing Gaussian mixture models, which is solved within the maximum-likelihood framework and expectation–maximization algorithm. One of the best known point matching approaches for rigid registration, the iterative closest point (ICP) [6] algorithm, was proposed to handle PSR with least-squares estimation of transformation parameters. In order to deal with nonrigid point sets, Chui and Rangarajan [7] established a general framework for nonrigid registration, which named robust point matching with thin-plate spline (TPS-RPM). In a continuous optimization framework that involves deterministic annealing, TPS-RPM replaces the nearest point strategy of ICP with soft assignments. Yang *et al.* [8] further introduced two distance features for measuring the global and local structure of point sets called global and local mixture distance (GLMD) and have shown satisfying results. In recent years, PSR has commonly been solved by probabilistic methods, such as coherent point drift (CPD) [9], Gaussian mixture model-based registration (GMMREG) [10], Student's-t mixture model with prior probability modeling (DSMM) [11], combinative strategy with regression and clustering (VBPSM) [12], global–local correspondence, and transformation estimation (GL-CATE) [13]. However, since these methods are completely independent of the abundant information of local image descriptors, their matching performance very likely degrades, especially when the image pair involves complex nonrigid deformations [2].

B. Outlier Rejection

Indirect matching formulates the matching task as a two-stage problem, which commonly starts with establishing preliminary correspondences. Common strategies for constructing putative matches include fixed threshold (FT), NN, mutual NN (MNN), and NN distance ratio (NNDR) [14]. Given a candidate matching set, feature matching reduces to an outlier detection and removal problem. The existing techniques for eliminating or alleviating the undue influence of outliers can be broadly classified into five categories: resampling methods, nonparametric methods, graph matching (GM) methods, relaxed methods, and learning-based methods.

Random sample consensus (RANSAC) [15] is the most classical resampling approach. Basically, two images are assumed to be coupled by a certain parametric geometric relation, such as epipolar geometry and homography. This type of method aims to find the smallest possible outlier-free subset to estimate the predefined transformation model parameters by repeatedly sampling. Various variants have been proposed to improve the performance of RANSAC, such as maximum-likelihood estimation sample consensus (MLESAC) [16], locally optimized RANSAC (LO-RANSAC) [17], progressive sample consensus (PROSAC) [18], spatially consistent random sample consensus (SCRAMSAC) [19], marginalizing sample consensus (MAGSAC) [20], and MAGSAC++ [21]. However,

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on February 10,2023 at 14:55:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

the resampling methods rely on a specific parametric model, which leads to the degradation of matching performance when the image transformation is nonrigid. In addition, as the outlier proportion increases, matching performance tends to severely degrade. Several proposed nonparametric methods can alleviate the above problems, including identifying correspondence function (ICF) [22], vector field consensus (VFC) [23], and manifold regularization-based robust point matching (MR-RPM) [24]. In general, these methods interpolate a nonparametric function by applying the prior condition, in which the motion field associated with the feature correspondence is slow-and-smooth. GM is another alternative for overcoming matching problem, with dual decomposition [25], spectral matching [26], graph shift (GS) [27], mode seeking [28], and multigraph matching [29]–[31] as representatives. However, nonparametric techniques typically have cubic complexities, and GM-based pipeline also suffers from similar defects, namely, its nonpolynomial-hard nature.

In order to accommodate more complex scenarios, a variety of approaches exploiting neighborhood consensus, also known as relaxed methods, have been investigated to improve generality and efficiency, which has become a very important trend in feature matching. Technically, relaxed methods focus on the locality of each correspondence rather than the global image transformation due to the stable local geometrical relationship of the potential true correspondences. These strategies usually start by building an appropriate handcrafted descriptor describing the similarity of two local structures and then removing the outliers whose local structures are sufficiently different via various detecting technologies. Several representative studies include grid-based motion statistics (GMS) [32], [33], locality preserving matching (LPM) [2], its improved version guided locality preserving feature matching (GLPM) [34], and multiscale locality and rank preservation (mTopKRP) [1]. GMS assumes that neighboring pixels of image would move together and incorporates the smoothness constraint into a statistic framework based on the number of neighboring matches for separation outliers. LPM explores the local topology structure of surrounding feature points, which makes more restrictive assumptions than GMS, thus further filtering out those matches with different spatial neighborhood structures among feature points. To weaken the matching cost, GLPM employs a small putative set with a lower outlier ratio (i.e., building initial inlier pool) to guide the matching on a large putative set, where the initial inlier pool can be easily constructed by the NNDR algorithm. Complex matching patterns (e.g., low inlier ratio), however, would lead to the initial inlier pool being unreliable. While mTopKRP transforms the feature points from the feature space into the ranking list space, i.e., a more strict measuring criterion for local structure preservation. Obviously, relaxed methods suffer from the inherent drawback of handcrafted techniques, the abundant information of local context is discarded, and the matching performance of these methods very likely degrades seriously when failing to capture valid feature representations or extract intricate feature interactions. Additionally, to the methods mentioned above, Jiang *et al.* [35] considered outlier rejection as a spatial clustering problem via density-based spatial clustering of applications with noise algorithm (DBSCAN). The principle is to adaptively cluster the putative matches into several motion-consistent clusters together with an outlier cluster, but the results are sensitive to the fluctuation of density and the variation of the distance between clusters.

Recently, applying deep learning to a wide range of complex computer vision tasks has proven to be very useful, such as image classification, object detection and tracking, image segmentation, and GM [36]–[40]. However, learning from points is not as popular as those in images for feature extraction, representation, and similarity evaluation, particularly for seeking good correspondences, because of the unordered, irregular structure and dispersed nature of point sets. Point-based learning for feature matching has only been introduced in recent years. Qi *et al.* [3] proposed the first network backbone for handling 3-D point clouds called PointNet, which uses shared multilayer perceptrons (MLPs) and max-pooling layers to obtain features of point clouds. On the basis of PointNet, Yi *et al.* [4] designed a novel normalization technique (PointCN) and first attempted to introduce a deep learning framework operating on coordinates of putative matches for finding good correspondences. PointCN aims to train an MLP-based pipeline under parametric geometrical constraint. Zhang *et al.* [5] proposed an order-aware network (OANet) to further improve PointCN, which leverages DiffPool and DiffUnpool layers to capture the local context of unordered sparse correspondences and takes regression loss over the essential matrix as the object function of the optimum design. OANet has gained satisfactory matching results for wide-baseline stereo. The correspondence set, however, is assumed to be coupled by a certain geometric relation, and its versatility is greatly limited. To address this issue, Ma *et al.* [41] proposed learning a two-class classifier for mismatch removal (LMR), in which key idea is to construct a set of match representations measuring the consensus of local neighborhood elements and topology and then find the optimal classifier by function approximation ability of learning model (e.g., MLP and random forest). Substantially, LMR relies on additional handcrafted descriptors for detecting outliers, instead of learning from point sets directly. Due to the focus on local feature learning without additional geometric constraints, LMR reveals significant superiority in dealing with general image problems, but the consensus of neighborhood topology based on the ratio of length and angle is sensitive to large viewpoint change. In order to fully exploit the raw local structure information, Chen *et al.* [42] leveraged visual representation (LSV) and two-channel convolutional neural networks (CNNs) to deal with the feature matching of RS. However, such visual representation still relies on nondifferentiable handcrafted features to solve mismatch removal.

III. Deep State Learning Network

In this section, we will present the proposed StateNet, which alternately cascades two novel modules: matching-neighbors selector and spatial aligner. The formulation of problem is first

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on February 10,2023 at 14:55:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

introduced and then the key technologies existing in these two modules successively.

A. Problem Formulation

Suppose that feature point sets X and Y are extracted from two images, and then, construct a set of putative matches $S = \{(x_i, y_i)\}_{i=1}^{n}$ (x_i and y_i denote the spatial positions of feature points) by matching each feature point in X to the nearest one in Y based on the local descriptor distances, e.g., scale-invariant feature transform (SIFT) [43] and radiation-variation insensitive feature transform (RIFT) [44]. Our target is to seek high-quality correspondences by removing the outliers contained in S, and this task varied in difficulty according to the proportion of outliers and the degree of deformation. Therefore, we focus instead on regularizing and weakening such randomness in matching patterns, shown in Fig. 1, thereby reducing the cost for establishing good correspondences in a unified deep learning framework.

1) Formulation for State Entropy: Due to the physical constraints in a small region around a point, the local spatial distribution of most feature points will not change significantly after transformation, relative to overall spatial relationship [45]. Thus, for capturing the randomness of S mathematically, we consider the small region around each point and quantize the degree of randomness into two parts: the relative distances of local inliers and the number of local outliers. The state entropy is designed to work on local regions, and the space coordinates should not be dependent on the absolute position in the original image but on their relative positions. Toward that end, we position local coordinate systems at the geometric center of neighboring points and normalize the coordinates. Finally, the state entropy can be formulated as

$$H(S) = \sum_{i=1}^{n}\left[\left(\sum_{x_i^j \in \mathcal{N}_{x_i}} t_{x_i}^j \left\| x_i^j - y^j \right\|_2^2 + \sum_{y_i^j \in \mathcal{N}_{y_i}} t_{y_i}^j \left\| y_i^j - x^j \right\|_2^2\right) + \left(|\mathcal{N}_{x_i}| - \sum_{x_i^j \in \mathcal{N}_{x_i}} t_{x_i}^j + |\mathcal{N}_{y_i}| - \sum_{y_i^j \in \mathcal{N}_{y_i}} t_{y_i}^j\right)\right] \quad (1)$$

where $\mathcal{N}_{x_i}$ is used to represent the neighborhood of point x_i, $|\cdot|$ denotes the cardinality of a set, and $t_{x_i}^j$ is the ground truth label of putative match (x_i^j, y^j). Specifically, the ground truth label is represented as 1 for an inlier and 0 for an outlier. The first term penalizes inliers that do not preserve the distance of a point pair due to geometric deformations, and the second term penalizes the number of outliers existing in the neighborhood. The coordinate is normalized to a range of 0–1, and therefore, we do not consider the tradeoff parameters, that is, each outlier contributes a unit of state entropy. Intuitively, as the inlier ratio increases, the degree of deformation will make a larger and larger effect on the state entropy and at last become the main factor.

In (1), feature points have no well-defined neighbors. Our strategy is to add a smoothing constraint in the k NNs under the Euclidean distance, as shown in Fig. 2. Concretely, motion smoothness causes neighboring pixels and features move

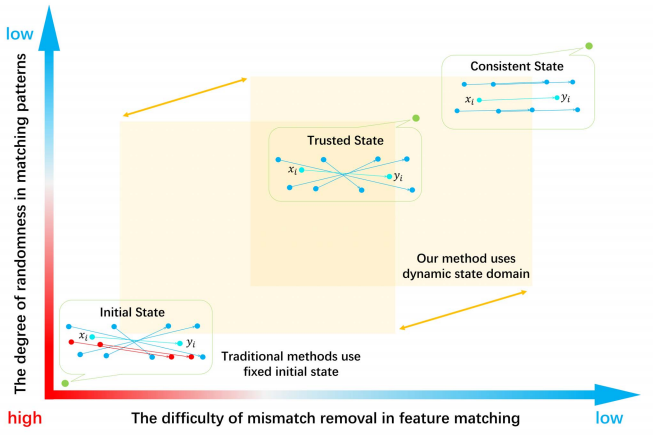


Fig. 1. Schematic of the relationship between the degree of randomness in matching patterns and the difficulty of mismatch removal. For brevity, we leverage three representative states to demonstrate the degree of randomness (initial state: the original spatial distributions of two sets of feature points; trusted state: the outliers have been screened out; and consistent state: a high level of topology consensus between two feature point sets).

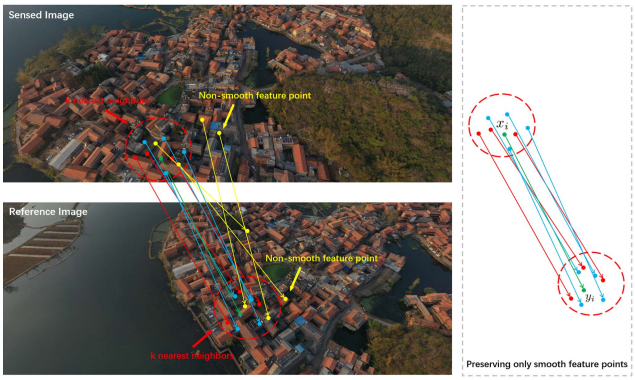


Fig. 2. Schematic of the matching neighbors for putative match (x_i, y_i). • denotes nonsmooth feature point (green arrow: putative match (x_i, y_i), red arrow: outlier, blue arrow: inlier, and yellow arrow: nonsmooth putative match).

together, which implies that neighborhoods of true match share many similar features across both images. In other words, if two key points that form a correspondence simultaneously appear in the local neighborhoods $\mathcal{N}_{x_i}$ and $\mathcal{N}_{x_i}$, then this correspondence have a greater probability of being an inlier. Formally, the neighbors of x_i and y_i can be defined as

$$\mathcal{M}_{x_i} = \left\{x_i^j | x_i^j \in \mathcal{N}_{x_i}^k, y^j \in \mathcal{N}_{y_i}^k\right\} \quad (2)$$

$$\mathcal{M}_{y_i} = \left\{y_i^j | y_i^j \in \mathcal{N}_{y_i}^k, x^j \in \mathcal{N}_{x_i}^k\right\} \quad (3)$$

where $\mathcal{N}_{x_i}^k$ denotes the k NNs of x_i searched from $\{x_i\}_{i=1}^n$. Obviously, the elements in $\mathcal{M}_{x_i}$ and $\mathcal{M}_{y_i}$ can constitute m_i putative matches, named matching neighbors. With the neighborhood definition, the ground truth label in (1) turns out to be

$$t_{x_i}^j = \begin{cases} t_{x_i}^j, & \text{if } y^j \in \mathcal{N}_{y_i}^k \\ 0, & \text{if } y^j \notin \mathcal{N}_{y_i}^k \end{cases} \quad (4)$$

$$t_{y_i}^j = \begin{cases} t_{y_i}^j, & \text{if } x^j \in \mathcal{N}_{x_i}^k \\ 0, & \text{if } x^j \notin \mathcal{N}_{x_i}^k. \end{cases} \quad (5)$$

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on February 10,2023 at 14:55:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

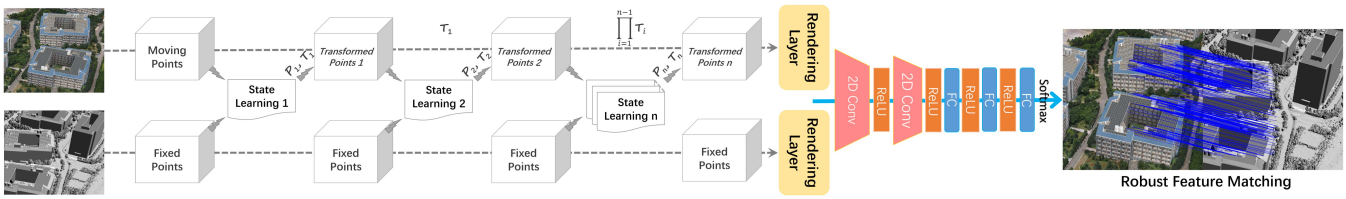


Fig. 3. Deep state network (StateNet) architecture for feature matching. Given image pairs, the goal of our StateNet is to update the initial geometrical structure by cascading state learning blocks so that the initial structure can gradually become more similar to the target structure. Then, we convert the confidence of each putative correspondence into a dynamic visual similarity evaluation by PRL and two-channel CNN. In summary, such cascade architecture explicitly allows the stepwise spatial manipulation of putative match set, with no explicit loss or constraint in enforcing the canonicalization, thus deducing optimal state for solving matching task.

For definiteness and without loss of generality, we assume that $(x_i^j, y^j) = (x^j, y_i^j) = (x_i^j, y_i^j) \in S$, and $t_{x_i}^j = t_{y_i}^j = t_i^j$. In the sequel, by preserving the matching neighbors, the state entropy [i.e., (1)] can be rewritten as

$$H(S) = 2\sum_{i=1}^{n}\left[\sum_{x_i^j \in \mathcal{M}_{x_i}} t_i^j \left\| x_i^j - y_i^j \right\|_2^2 + \sum_{x_i^j \in \mathcal{M}_{x_i}} \left(1 - t_i^j\right)\right]. \quad (6)$$

2) Formulation for State Learning: Given a set of putative matches S, a mathematical procedure for minimizing (6) is that finding a set of geometric transformation $\{\mathcal{T}_i\}_{i=1}^n$

$$\text{s.t.} \sum_{x_i^j \in \mathcal{M}_{x_i}} t_i^j \left\| x_i^j - \mathcal{T}_i\left(y_i^j\right) \right\|_2^2 \le \sum_{x_i^j \in \mathcal{M}_{x_i}} t_i^j \left\| x_i^j - y_i^j \right\|_2^2.$$

Ideally, the optimal transformations should achieve zero penalty, i.e., the first term of (6) should be zero. The outliers in the matching neighbors, however, prevent us from finding a group of optimum inverse transformations. On the other hand, updating the initial location of $\mathcal{M}_{y_i}$ by geometric transformation to make it closer to $\mathcal{M}_{x_i}$ can further facilitate the identification of outliers within the matching neighbors. Such a strategy has been widely investigated in many traditional iterative PSR algorithms (e.g., ICP and CPD).

According to the analysis above, the deep state learning is designed to perform alternating two steps: match recalibration (namely, gating the responses of outliers) and transformation update, as shown in Fig. 3 (left). Then, our aim is equivalent to designing two modules (i.e., matching-neighbors selector and spatial aligner), which encodes the following mappings:

$$\mathcal{P}_i = \mathbf{F}_{\text{ns}}(\mathcal{M}_i) \quad (7)$$

$$\mathcal{T}_i = \mathbf{F}_{\text{tr}}(\mathcal{M}_i, \mathcal{P}_i) \quad (8)$$

where $\mathcal{M}_i \in \mathbb{R}^{4\times(m_i+1)}$ (including putative match (x_i, y_i) with its matching neighbors) is the only input to our network. $\mathbf{F}_{\text{ns}}$ produces a nonmutually exclusive match score vector $\mathcal{P}_i \in \mathbb{R}^{m_i+1}$, which can be exploited by subsequent transformation estimation for filtering outliers and emphasizing inliers. $\mathbf{F}_{\text{tr}}$ deduces an geometric transformation $\mathcal{T}_i$, which allows us to actively perform spatial manipulation of $\mathcal{M}_{y_i}$ within the network. More importantly, for establishing reliable feature correspondences, both $\mathbf{F}_{\text{ns}}$ and $\mathbf{F}_{\text{tr}}$ are permutation-equivariant and can even be recursively cascaded so that it can be inserted into the existing classifier in an end-to-end manner

$$\mathcal{Z}_i = \text{Softmax}(\text{Classifier}(\mathcal{M}_i; \mathbf{F}_{\text{ns}}, \mathbf{F}_{\text{tr}})). \quad (9)$$

The final softmax layer calculates one weight for each correspondence, encoding its likelihood to be an inlier, i.e., $\mathcal{Z}_i$ is the probability values for classification.

In summary, (7) and (8) constitute the basic submodule of our StateNet (depicted by Fig. 4), in which optimization objective is to minimize a hybrid loss function as follows:

$$\mathcal{L} = \mathcal{L}_c + \alpha \mathcal{L}_s \quad (10)$$

where α is the hyperparameter to balance these two losses. $\mathcal{L}_c$ is a binary cross entropy loss

$$\mathcal{L}_c = -\frac{1}{n}\sum_{i=1}^{n}\left[t_i \cdot \log(\mathcal{Z}_{i0}) + (1 - t_i) \cdot \log(\mathcal{Z}_{i1})\right] \quad (11)$$

where $t_i \in \{0, 1\}$ represents the correctness of putative match (x_i, y_i). $\mathcal{L}_s$ denotes the state entropy loss

$$\mathcal{L}_s = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{x_i^j \in \mathcal{M}_{x_i}} t_i^j \left\| x_i^j - \mathcal{T}_i\left(y_i^j\right) \right\|_2^2\right]. \quad (12)$$

The cross entropy loss can train accurate models by itself, but we observed that using the state entropy as a regularization term can maintain the stability in the training procedure, thus guiding the searching of reliable feature matches.

B. Matching-Neighbors Selector

To explicitly emphasize inliers and suppress outliers within StateNet, the matching-neighbors selector, relying entirely on a set of sparse matches, maps the specific input to a set of match scores for modeling the relationship between matches. Technically, inspired by Hu *et al.* [46], we formulate the match relationship estimation as two steps: global information embedding and matchwise dependencies decoding.

1) Global Information Embedding via ADAConv: Unlike high-dimensional point cloud data, input $\mathcal{M}_i$ is a small subset extracted from two sets of feature points, and thus, it has no obvious context structure. Under the circumstances, for learning global geometric features directly, our main idea is to calculate the global response as a weighted sum of the features at all matches in the input. To this end, we present a nonlocal convolution involving two aggregation suboperations

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on February 10,2023 at 14:55:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

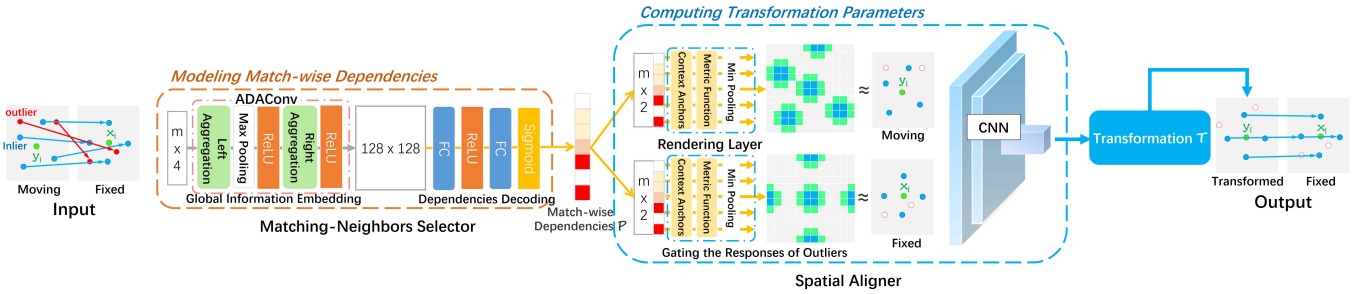


Fig. 4. ASL block. First, the ASL takes sparse putative match set as input, embeds global information via ADAConv, and decodes the matchwise dependencies by a simple MLP. Then, feeding the output $\mathcal{P}$ and the original input to a PRL, which gates the spatial response of outliers and encodes a potentially canonical order (the intensity response of feature map is visualized by gradation of color). Finally, learning a spatial transformation $\mathcal{T}$ by a convolutional regression network.

as a basic building block, called ADAConv. Given a set of sparse matches $\mathcal{M}_i$, dual-aggregation operation can be defined as

$$\text{dualAgg}(\mathcal{M}_i) = \sum_{m=1}^{m_i+1} \alpha_m \sum_{n=1}^{4} \beta_n [\mathcal{M}_i]_{nm} \tag{13}$$

where the operator $[\cdot]_{nm}$ returns the element at position (n, m) in a matrix. Note that the original features of putative match are represented by just its four coordinates in the basic setting, and additional dimensions can be added by computing other local or global features. Two aggregation operations can be regarded as linear projections: the first operation performs adaptive feature embedding for single putative match and the second is used to aggregate global geometrical features by assigning calibration weights to per putative match. Intuitively, in contrast to the progressive behavior of traditional convolution operation, the dual-aggregation mechanism allows StateNet to jointly attend to information from all matches at the outset.

In order to encourage the network learn global geometrical features that cover more details, a simple strategy is to convert the input to the high-dimensional feature vectors by consecutively invoking (13). However, this is computationally inefficient. To reduce the number of parameters, we mathematically express the ADAConv as matrix operations by cross sharing parameters, i.e., (13) can be reformulated as

$$\mathcal{F}_{O\times D} = \text{ADAConv}(\mathcal{M}_i) = \sigma(\sigma(\mathcal{L}\cdot\mathcal{M}_i)\cdot\mathcal{W}) \tag{14}$$

where $\sigma(\cdot)$ represents activation function, and to simplify the notation, the bias terms are omitted. $\mathcal{L} \in \mathbb{R}^{O\times 4}$ is a left aggregation matrix, which realizes the multiple optimization combination of features for each match by using O shared linear combinations. Analogously, $\mathcal{W} \in \mathbb{R}^{(m_i+1)\times D}$ denotes a right (global) aggregation matrix, which learns the global context of $\mathcal{M}_i$ via D shared linear combinations. The row vectors of $\mathcal{L}$ and the column vectors of $\mathcal{W}$ can be considered as convolution kernels with cross-shared weights. It is obvious that in (14), the required parameters for each global feature are reduced to $4/D+(m_i+1)/O$ and decrease with the increasing of the dimensionality of output feature map $\mathcal{F}_{O\times D}$.

2) Matchwise Dependencies Decoding: The output of the dual-aggregation convolutional layer can be interpreted as multiple global signatures, whose statistics are expressive for the whole input. Thus, the dependencies between matches are implicitly embedded in $\mathcal{F}_{O\times D}$ but are entangled with the match presentation captured by the left aggregation matrix. Similar to the way of learning channelwise dependencies in CNNs, we opt to employ a simple MLP as decoder, which calculates a score vector $\mathcal{P}_i \in \mathbb{R}^{m_i\times 1}$ assigned to the matching neighbors of (x_i, y_i) for fully capturing matchwise dependencies

$$\mathcal{P}_i = \text{Sigmoid}(W_2\text{ReLU}(W_1\mathcal{F}_i + b_1) + b_2) \tag{15}$$

where $\mathcal{F}_i$ denotes flattened global geometrical features, W_1 and W_2 are the parameters of two fully connected (FC) layers, and b_1 and b_2 are bias terms. In addition, arbitrary ordering to the matching neighbors should not change the function values, which requires ADAConv to be permutation-equivariant. Sorting seems like a good solution. However, an ordering, which is stable with respect to point perturbations in the general sense, does not exist in high-dimensional space actually. Therefore, sorting cannot essentially solve the problem that consistent mapping from input to output [3]. In ADAConv, each row weight of left aggregation matrix is shared across all the matches, so we can leverage a symmetric function $g(\cdot)$ defined on $\mathcal{L}\cdot\mathcal{M}_i$. In this work, we use mapping $g : \mathbb{R}^{O\times(m_i+1)} \to \mathbb{R}^{O\times 2}$ to perform max-pooling operation for each row vector. Significantly, if the distribution of neighborhood elements is consistent, but $(x_i, y_i) \neq (x_j, y_j)$, the predicted result should be different. Hence, the feature representation of (x_i, y_i) does not participate in max-pooling operations.

C. Spatial Aligner

This is a differentiable module that regresses a relative transformation during a single forward pass, where the transformation is conditioned on the particular state of input. The discrete match data are usually an arbitrary combination without rules, so to fulfill this objective, two problems must be solved: 1) how to leverage the spatially local correlation across two sets of feature points? and 2) how to gate the responses of the outliers in the spatial domain based on the output of $\mathbf{F}_{\text{ns}}$? The proposed PRL intends to conquer issues like these, which is the core operation for building the spatial aligner.

1) Point Rendering Layer: Spatial distribution pattern is a latent and ubiquitous property of various point clouds that is independent of the data representation. On the other hand, for

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on February 10,2023 at 14:55:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

data that are represented in regular grid such as image, leveraging its spatially local correlation is more implementation friendly in deep learning frameworks. This is due to the fact that the order of grid pixels involves explicit spatial information. Inspired by this, PRL encodes latent spatial information of feature points by a set of context anchors. Furthermore, for preserving spatial distribution and gating the responses of outliers concurrently, we extend the spatial information encoded by PRL to the 3-D Euclidean space, where the extra dimension is used to implement feature filtering. Formally, context anchors are defined in the form of an $H \times W$ regular grid

$$G_{hw} = \left(\frac{w-1}{W-1}, \frac{h-1}{H-1}, 1\right), \quad h \in [1, H], \ w \in [1, W]. \tag{16}$$

Then, we design a rendering mapping $\mathbf{F}_{\text{re}} : \mathbb{R}^{2\times(m_i+1)} \to \mathbb{R}^{H\times W}$ to draw out the spatial interaction information from scattered points and context anchors via a metric function and a min pooling function

$$\left[\mathbf{F}_{\text{re}}\left(\mathcal{M}_{x_i}, \mathcal{P}_i\right)\right]_{hw} = \text{MIN}\left\{\left\|\left[x_i^j || p_i^j\right] - G_{hw}\right\|_2^2\right\}_{j=1}^{m_i} \tag{17}$$

where $[\cdot||\cdot]$ is concatenation operator. In this article, we opt the L2 norm squared as feature metric function. Theoretically, any metric function can be used, as long as (sub)gradients can be defined. The partial derivative of (17) can be easily derived

$$\frac{\partial[\mathbf{F}_{\text{re}}]_{hw}}{\partial \tilde{x}_i^j} = \begin{cases} 2\left(\tilde{x}_i^j - G_{hw}\right), & \text{if } \mathcal{C} \text{ is the smallest} \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

where $\mathcal{C} = \|\tilde{x}_i^j - G_{hw}\|_2^2$ and $\tilde{x}_i^j = [x_i^j || p_i^j]$.

The intuition behind is the following. Differentiable mapping $\mathbf{F}_{\text{re}} : \mathbb{R}^{2\times(m_i+1)} \to \mathbb{R}^{H\times W}$ propagates sparse features to a denser one by preserving the metric features for the key point that is closest to the context anchors, thus extracting potential structure representation and ensuring permutation-equivariant. This means that PRL makes the typical CNNs can be generalized to other irregular data, as shown in Fig. 5 (top). From geometrical significance, common feature calibration practice (scalar multiplication) is a scaling transformation in 2-D Euclidean space, which will destroy the raw spatial distribution. On the contrary, matchwise dependencies embedded in the third dimension can control the metric value, thus filtering out the outliers by min pooling and not destroying the original spatial interaction information, as shown in Fig. 5 (bottom).

2) Transformation Estimation: For evaluating the spatially local correlation across two point sets, we use a rendering layer to produce two matrix tensors and then combine them into a two-channel feature map, which provides greater flexibility as it starts by processing the two patches jointly. Specifically, such feature map is directly fed to a convolutional regression network, and then, the convolutional layers analyze spatially corresponding patches and learn transformation parameters. Convolutional regression network should include a final regression layer to produce the transformation parameters θ. The size of parameters θ is determined by the degree of freedom of the transformation model, and we found that using the projection transformation (i.e., homography) works well.

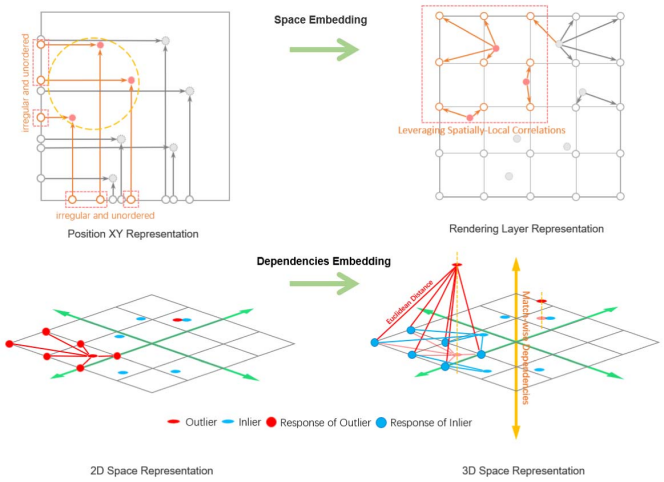


Fig. 5. Schematic of mechanism for PRL. Leveraging the spatially local correlation (top): point sets are irregular and unordered, hence applying convolution for capturing spatial interaction information can be difficult. PRL establishes a potentially canonical spatial order by a set of context anchors, which allows the 2-D convolution to be easily extended to point cloud data. Gating the spatial response of outlier (bottom): if the probability that the feature point being an outlier is high enough, the feature point will leave the original 2-D plane. Then, its distance response feature will be replaced by other feature points by min-pooling operation, thereby achieving the purpose of antioutlier.

D. Alternate Cascading Architecture

Deep learning has been proven to benefit from cascade architectures, such as image super-resolution [47], semantic segmentation [48], object detection [49], and medical image registration [50]. In order to decompose the target state into simple, incremental changes, we recursively perform state learning on the point set. Assuming for c cascades in total, the final inverse transformation is a composition of all predicted transformations, and the final neighborhood selection rule is determined by the output of the last matching-neighbors selector, namely

$$\mathcal{P}_i^c = \mathbf{F}_{\text{ns}}\left(\mathcal{M}_i^c\right) \tag{19}$$

$$\mathcal{T}_i^c = \mathbf{F}_{\text{tr}}\left(\mathcal{M}_i^c, \mathcal{P}_i^c\right) \tag{20}$$

$$\mathcal{M}_i^{c+1} = \mathcal{T}_i^c \circ \mathcal{T}_i^{c-1} \circ \cdots \circ \mathcal{T}_i^1\left(\mathcal{M}_i^1\right). \tag{21}$$

Following this cascaded architecture, the subsequent module is directly dealing with the sparse point set that was optimized by the previous module, which avoids overfitting and information loss due to the increasing number of network layers. In the sense that our StateNet provided a mechanism to change the state of matches into latent canonical forms for being further processed, with no explicit loss or constraint in enforcing the canonicalization. Thus, this recursion can be infinitely applied in theory. The moving point set is transformed successively, enabling the large displacement to be decomposed into cascaded, progressive small displacements, which greatly reduces the learning difficulty of each ASL block.

In the end, for finding good correspondences, our key idea is to transform the confidence of putative correspondences into

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on February 10,2023 at 14:55:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I
SPECIFIC PARAMETER SETTINGS OF THE STATENET

Module	Output size	Parameter settings
$\mathbf{F}_{ns}$	128×25	left aggregation, 128×4
	128×2	max pooling
	128×128	right aggregation, 2×128
	25×1	mlp, [256, 25]
$\mathbf{F}_{tr}$	$20 \times 20 \times 2$	point rendering layer, 20×20
	$18 \times 18 \times 128$	conv, 3×3, 128, stride 1, padding 0
	$16 \times 16 \times 128$	conv, 3×3, 128, stride 1, padding 0
	9×1	mlp, [256, 128, 9]
Clas.	$10 \times 10 \times 2$	point rendering layer, 10×10
	$8 \times 8 \times 64$	conv, 3×3, 64, stride 1, padding 0
	$6 \times 6 \times 64$	conv, 3×3, 64, stride 1, padding 0
	2×1	mlp, [256, 128, 2]

a dynamic consistency evaluation based on the local visual topology of point sets with low state entropy. This is because local visual topology is not greatly susceptible to a small perturbation of point. To fulfill this objective, the rendering layer can be used to build the approximate visual topology after state learning, which is input to the CNN for rejecting outliers. Fig. 3 shows the proposed architecture. In fact, the classification network can take any form (e.g., ADAConv and PointCNN), and we also provide other classification techniques for comparison in our experiments.

E. Implementation Details

In our experiments, the open-source VLFeat toolbox [51] is employed to determine the putative correspondence of SIFT and search for the K NNs by K-D tree [52]. Note that RIFT is used to extract putative matches for multimodal images. In addition, we find empirically that setting $k = 25$ and two cascades can work well. The specific StateNet architecture and parameters are outlined in Table I.

In the training phase, we used minibatch stochastic gradient descent (Adam [53]) for optimizing hybrid loss function $\mathcal{L}$ (where $\alpha = 0.5$). Specifically, the learning rate is $1e^{-3}$, and the batch size is 100. All these training and testing procedures are implemented with Pytorch. The 12 image pairs undergoing different types of transformations are used to construct the training set, which contains 11 659 putative matches in total with 5858 inliers and 5801 outliers. In order to prevent the obvious data tendency in the training process, the number of positive and negative samples should be close to 1:1. Note that the classification (binary cross entropy) loss can train accurate models by itself, and we observed that using the state entropy loss early on can actually harm performance. However, this allows our network to maintain the stability (i.e., avoid the exploding gradient and boost convergence speed) in the training procedure, thus guiding the searching of reliable feature matches and improving relative performance by 1%–3%. In summary, the well-trained StateNet in general has satisfying generalization ability, even of different types of images or with different transformation models.

IV. EXPERIMENTAL RESULTS

In order to evaluate the performance of our StateNet, experiments are divided into four parts. First, we benchmark our StateNet against state-of-the-art techniques on several public datasets and then test its robustness and generality. Next, we apply StateNet to solve RS image registration tasks. Finally, we also provide further insights into the components of our StateNet. The experiments implemented in MATLAB and Python on a laptop with 2.80-GHz Intel Core i7-7700HQ CPU and 32-GB RAM.

A. Comparison to the Baselines

Eight state-of-the-art competitors are used for comparison, including resampling-based method: RANSAC [15]; nonparametric methods: VFC [23]; GM method: GS [27]; relaxed constraints methods: LPM [2], GMS [32], [33], and RFM-SCAN [35]; and learning-based methods: LMR [41], and OANet [5]. All the competitors are implemented based on their publicly available codes and their own optimal parameter settings.

B. Datasets

To achieve a direct and fair comparison, we provide the experimental results in the following six datasets.

1) RS Dataset [54]: RS consists of 40 RS image pairs including panchromatic (PAN) photographs, synthetic aperture radar (SAR), color-infrared, and low-altitude RS images captured by small unmanned aerial vehicles (UAVs). The images are of sizes from 600×400 to 800×600, and the ground truth labels are supplied by the dataset.

2) Small UAV Image Registration (SUIR) Dataset [14]: This dataset is provided for image registration/matching research, which includes 60 pairs of low-altitude small UAVs images (800×600). These image pairs contain viewpoint changes in horizontal, vertical, mixture, and extreme patterns, and the ground truth labels are supplied by the dataset.

3) CoFSM *Dataset*[1]*:* CoFSM is a newly published multimodal RS image database, including optical-optical (nine pairs of images), infrared-optical (four pairs of images), depth-optical (eight pairs of images), map-optical (nine pairs of images), SAR-optical (seven pairs of images), and day-night (nine pairs of images). The ground truth labels are checked by the geometrical transform matrix.

4) Oxford Buildings (OxBs) Dataset [55]: The dataset consists of 5062 images collected from Flickr, which were collected by searching for specific Oxford landmarks. The ground-truth geometrical transformations are calculated from 20 matched landmarks.

5) Affine Covariant Features (ACF) Datasets [56]: ACF contains 40 image pairs with changes due to zoom and rotation, viewpoint, blur, light, and JPEG compression. The image pairs are planar scenes or captured by a camera in a fixed position during acquisition. Therefore, they always obey homography, and the ground truth homographies are supplied by the dataset.

[1]CoFSM: https://skyearth.org/publication/project/CoFSM/

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on February 10,2023 at 14:55:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

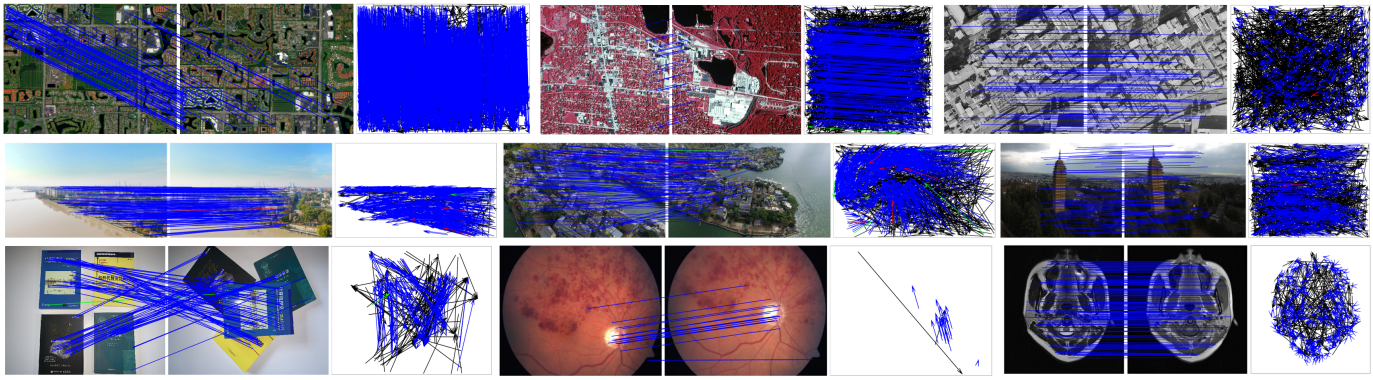


Fig. 6. Feature matching results of our StateNet on nine representative image pairs involving different types of transformations. For each group of results, the right motion field provides the decision correctness of each putative correspondence in the initial set, the inlier ratios in the 9 image pairs are 72.04%, 14.96%, 33.41%, 81.93%, 65.33%, 35.66%, 46.19%, 85.71%, and 66.72%. For clarity, in the image pairs, at most 200 random selected matches are visualized, and the TNs are omitted [blue = TP, black = TN, green = FN, and red = FP].

6) Mixed-Type Image (MTI) Dataset: This dataset was collected by ourselves for experimental evaluation. MTI consists of outdoor scene image pairs, multimodal medical image pairs, and other CV image pairs, which involve different transformations (including affine, homography, piecewise linear, and nonrigid deformation). To ensure objectivity, the ground truth labels are obtained by multiuser manual cross check.

C. Results on Feature Matching

In this section, we focus on the performance of StateNet in establishing feature correspondences and follow the same evaluation metrics in [54] and [14]: recall, precision, and F-score. Given the number of true positives (TPs), true negatives (TNs), FPs, and FNs, the recall is obtained by

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{22}$$

The precision is given as follows:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{23}$$

F-score denotes the matching performance defined as the harmonic mean of recall and precision

$$F = \frac{2 \times P \times R}{P + R}. \tag{24}$$

1) Qualitative Illustration: First, we will present some intuitive results on the matching performance of our StateNet. For this purpose, nine representative image pairs are used for test, as shown in Fig. 6. These test image pairs undergoing different types of transformations, including affine [see Fig. 6(a)–(c)], epipolar geometry or homography [see Fig. 6(d) and (f)], piecewise linear [see Fig. 6(g)], and nonrigid [see Fig. 6(h) and (i)] transformation. As can be seen from Fig. 6, well-trained StateNet has strong generalization ability to handle various types of geometric transformations, and very few putative matches are misjudged. Besides, Table II also provides the quantitative comparison on the nine image pairs with other eight state-of-the-art competitors.

TABLE II

QUANTITATIVE COMPARISON OF NINE METHODS ON TYPICAL IMAGE PAIRS UNDERGOING DIFFERENT TYPES OF TRANSFORMATIONS. TO FACILITATE THE COMPARISON OF PERFORMANCE DIFFERENCES, **BOLD** REPRESENTS THE OPTIMAL INDICATOR VALUE

Method	Recall	Precision	F-score
RANSAC	89.93%	99.46%	93.05%
GS	79.30%	95.12%	85.32%
VFC	94.77%	95.33%	95.00%
LPM	99.37%	83.45%	89.86%
LMR	84.45%	85.27%	84.64%
RFM-SCAN	94.72%	93.36%	93.91%
GMS	66.91%	97.00%	76.72%
OANet	56.08%	88.05%	64.76%
Ours	**99.77%**	**99.84%**	**99.81%**

2) Quantitative Results: In order to provide a comprehensive quantitative comparison with state-of-the-art competitors, we consider four common application scenarios.

1) Satellite RS images involving only linear (e.g., rigid or affine) transformation, which typically arisen in panoramic image mosaic.
2) Low-altitude RS images, which often involve complex spatial relationships due to the ground relief variations, imaging viewpoint changes. It is common for environmental monitoring.
3) Multimodal RS images suffer vast radiometric differences and provide highly complementary information about observed scenes, which frequently happens in image fusion.
4) Outdoor scenes, which play an important role in visual simultaneous localization and mapping (SLAM).

All test image pairs are selected from datasets RS (including satellite RS and low-altitude RS), OxBs, and CoFSM. The recall, precision, and F-score of the nine algorithms are summarized in Fig. 7.

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on February 10,2023 at 14:55:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

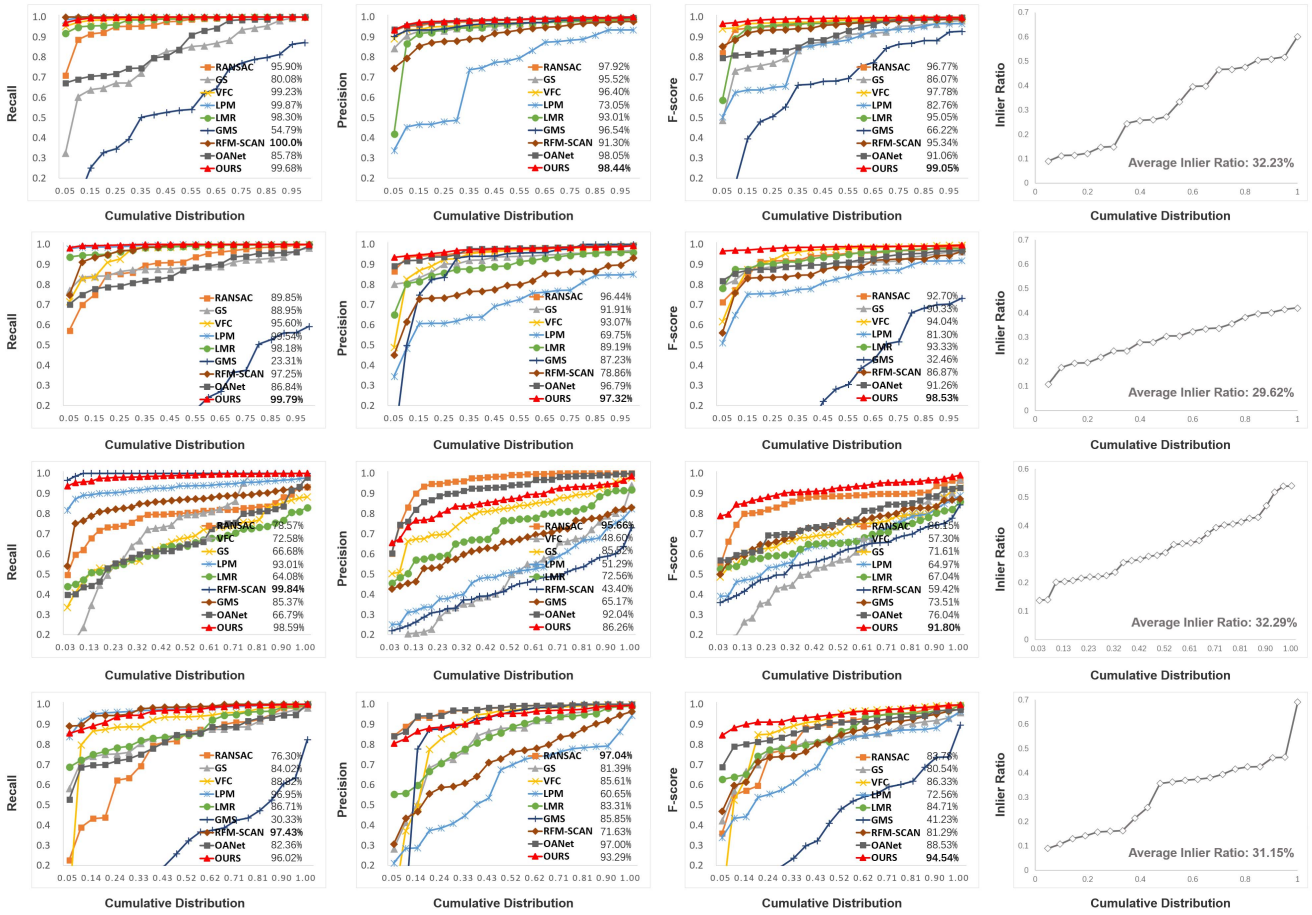


Fig. 7. Quantitative comparisons of our StateNet with eight state-of-the-art matching methods on four common application scenarios. Satellite RS images, low-altitude RS image (RS dataset), multimodal RS images (CoFSM dataset), and outdoor scenes (OxBs dataset) (top to bottom). Recall, precision, F-score, and initial inlier ratio with respect to the cumulative distribution (left to right). **Bold** indicates the best results.

From the results, we see that StateNet does not have an obvious advantage in terms of recall compared with RFM-SCAN. However, StateNet can always produce the best precision–recall tradeoff. The results of GMS are not that satisfying since the statistical measures based on the number of neighboring matches are often coarse when the images involve complex spatial transformations. OANet has achieved promising performance on outdoor scenes. The rationale behind is that such matching patterns often obey epipolar geometry, and OANet applies the eight-point algorithm to regress the essential matrix. In conclusion, imposing a geometrical constraint (i.e., RANSAC and OANet) can often get promising matching accuracy. This is because they restrict matches satisfying an underlying image transformation. Meanwhile, this demand severely limits the result in terms of recall, especially for nonrigid deformation. By contrast, RFM-SCAN and LPM usually have high recall, but at the expense of precision, because they do not require a global geometric model between image pairs. Obviously, due to the ASL mechanism, our method has greater advantages in handling different images even if the images are undergoing local deformation.

3) Robustness and Universality Test: In this experiments, to further report the robustness and universality of our StateNet, we consider the following four matching patterns (i.e., different degrees of deformation): 1) affine transformations, which usually occurs in satellite RS; 2) projection transformations, which is often caused by the imaging viewpoint changes; 3) piecewise linear transformations with occlusion, which is often arisen in video retrieval; and 4) nonrigid deformations, which frequently happens in dynamic scene matching and deformable object recognition. The four matching patterns are demonstrated in Fig. 8 (top), and the performances (F-score) are summarized in Fig. 8 (bottom).

We can observe that our StateNet surpasses all the state-of-the-art competitors on all five levels that include different matching patterns. In addition, when the underlying image transformations are nonparametric, the overall matching performances of RANSAC and OANet will suffer degradation significantly, due to the additional parametric model required in these approaches. GMS shows its weakness to seek good correspondence. VFC, LMR, and LPM have relatively high F-score, as they are designed for general (i.e., rigid and nonrigid) feature matching, i.e., the global spa-

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on February 10,2023 at 14:55:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

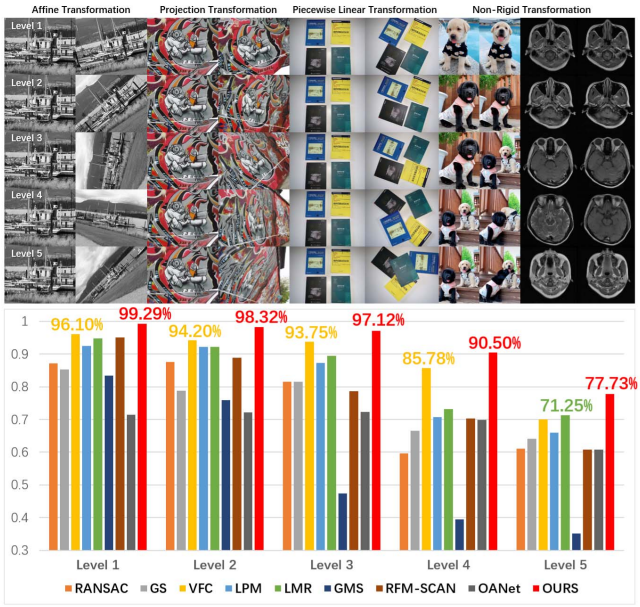


Fig. 8. Quantitative comparisons on different degrees of deformation. Five levels that include different matching patterns (top). The overall performances are reported with the F-score (bottom). Note that the average inlier ratios in five groups are 62.26%, 60.92%, 46.74%, 34.82%, and 34.23%. For convenience, we indicate the best and second best result of each level.

tial constraint is made less strict to accommodate complex scenarios.

D. Results on RS Image Registration

The primary goal of image registration is to geometrically overlay two or more images of the same scene captured from different viewpoints, different sensors, or by different times. Image registration also plays a momentous role in other computer vision tasks. For example, in the image fusion community, the observed images have been successfully registered are a critical prerequisite. Therefore, to exploit the practical value of StateNet, in this experiment, we focus on RS image registration according to the matching results.

1) Qualitative Illustration: First, StateNet processes reference image I^r and sensed image I^s, thus obtaining a reliable set of feature correspondences. Then, the homography matrix is used for geometric transformation. Specifically, the objective function can be defined as

$$\begin{pmatrix} x_i' \\ y_i' \\ 1 \end{pmatrix} = H \begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix} \tag{25}$$

where $\{(x_i, y_i), (x_i', y_i')\}_{i=1}^{c}$ is an inlier set inferred from the putative matches S. Then, we choose direct linear transformation (DLT) algorithm [57] to estimate the homography matrix H. Finally, sampling pixels from sensed image I^s is based on bilinear interpolation to obtain the transformed image I^t.

To demonstrate the effectiveness of our StateNet for different types of RS images, extensive RS registration results, including HRSIs, hyperspectral, color infrared aerial photographs (CIAPs), SAR, PAN aerial photographs, as well as multimodal RS images (optical-map, optical-infrared, optical-depth, optical-SAR, and day-night), are shown in Fig. 9. The visualization results are shown that StateNet can align the overlapped contents of the image pair well, including the challenging edge regions.

2) Quantitative Results: Geometric transformation (especially for nonrigid deformation) is an important factor that affects the registration accuracy. Therefore, we randomly select five representative low-altitude RS images with complex spatial relationships from SUIR to evaluate the robustness of our StateNet for geometric deformations. In order to reflect the robustness of the algorithm more intuitively, we use thin plate spline (TPS) transformation for registration. In this model, the transformation coefficient $\theta_{(n+3)\times 2}$ is found by solving the linear system [58]

$$\theta = \begin{pmatrix} \mathcal{K} & X' \\ X'^T & \mathbf{O}_{3\times 3} \end{pmatrix}^{-1} \begin{pmatrix} Y \\ \mathbf{O}_{3\times 2} \end{pmatrix} \tag{26}$$

where $\mathcal{K}_{n\times n}$ is a radial basis kernel with each entry computed by $\mathcal{K}_{ij} = \|Y_i - Y_j\|^2 \log\|Y_i - Y_j\|$, $X' = (1, X)$ is the $n \times 3$ homogeneous coordinate.

The registration results of our StateNet and eight state-of-the-art methods are shown in Fig. 10, and here are some observations. RANSAC achieves satisfactory registration results, namely, most overlap regions are aligned correctly. In the fourth and fifth image pairs, however, there are some noticeable deviations at the edges. This is due to that local distortions may cause partial inliers to disobey the global transformation, and RANSAC is known to be sensitive to nonrigid deformation, which means that relying solely on the parametric transformation model (e.g., affine and homography) cannot align or restore the local distortions. On the other hand, TPS is a more general transformation model, which also has higher quality requirements of the feature correspondences. Therefore, a very small proportion of false matches can degrade the registration result. LPM, GMS, OANet, and RFM-SCAN usually produce the worst results compared with other methods since a large number of nonsmoothing control points are inferred as true matches, which leads to the generation of intolerable image distortion during the registration process. In contrast, the registration performance of StateNet, without relying on the parametric geometric model, can be compared with or even better than RANSAC and obviously superior to the other seven methods. Finally, following the same evaluation in [59] and [54], the quantitative results of registration are characterized by the root-mean-square error (RMSE), maximum error (MAE), and median error (MEE), as summarized in Table III.

E. Ablation Studies

In order to evaluate the performance boost that state learning mechanism can deliver, we examine the importance of each aspect of our StateNet. Besides the ablation study on the architecture of the ASL, we also provide other three classification techniques (i.e., ADAConv, PointCNN [60], and PointNet [3]) for comparison.

1) Different Neighborhood Sizes: First of all, to reveal the effect of different neighborhood sizes on the performance of our method, we consider five different neighborhood sizes and test the performance on the RS dataset, as summarized

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on February 10,2023 at 14:55:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

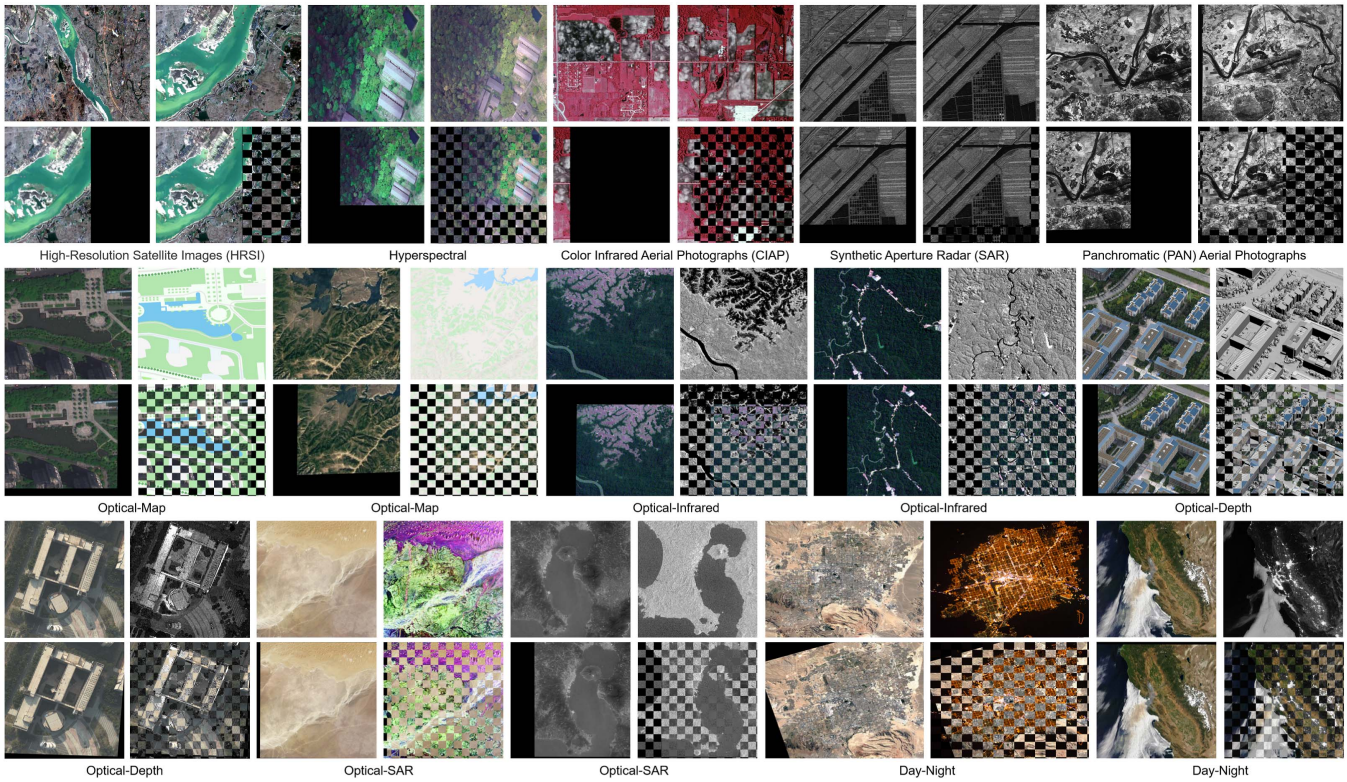


Fig. 9. Representative image registration results of our method. For each group of results, the top left is the sensed image and the top right is the reference image. The first row is single-sensor RS images, including high-resolution satellite images (HRSI), hyperspectral, color infrared aerial photographs (CIAP), SAR, and PAN aerial photographs. The second and third rows are multimodal RS images, including optical-map, optical-infrared, optical-depth, optical-SAR, and day-night.

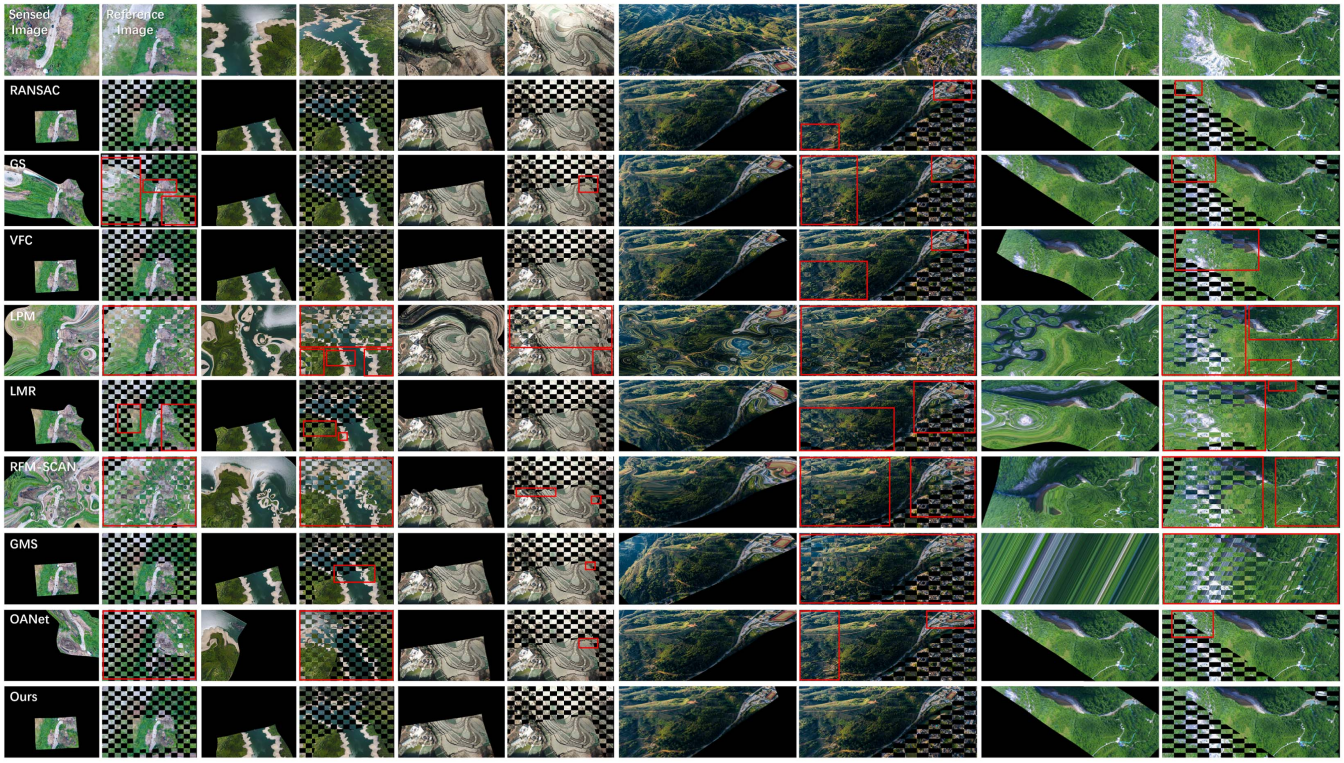


Fig. 10. Quantitative comparisons of RANSAC, GS, VFC, LPM, LMR, RFM-SCAN, GMS, OANet, and StateNet on low altitude RS image registration. The left and right of each group in the first row are the reference and sensed images, respectively. For visibility, we use the red rectangle to indicate the obvious deviation of registration.

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on February 10,2023 at 14:55:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE III

RESULTS OF IMAGE REGISTRATION QUANTITATIVE COMPARISON. THE VALUES IN THE TABLE REPRESENT AVERAGE AND STANDARD DEVIATION. TO FACILITATE THE COMPARISON OF PERFORMANCE VARIATIONS, **BOLD** INDICATES THE OPTIMAL VALUE

Method	RMSE	MAE	MEE
RANSAC	2.54±1.07	5.99±1.79	2.57±1.04
GS	19.94±12.08	34.75±24.82	17.41±16.75
VFC	7.64±8.95	19.87±25.94	3.23±1.68
LPM	42.57±36.71	110.82±88.25	37.36±35.41
LMR	10.08±10.14	24.19±25.57	5.18±2.30
RFM-SCAN	57.01±46.21	131.54±89.57	44.72±41.11
GMS	118.41±121.45	246.21±242.75	119.97±134.01
OANet	111.83±105.78	212.97±225.82	101.99±130.92
Ours	**1.99±0.32**	**4.26±0.62**	**2.48±0.97**

TABLE IV

INFLUENCE OF DIFFERENT NEIGHBORHOOD SIZES ON F-SCORE

Size	k = 15	k = 20	k = 25	k = 30	k = 35
F-score	96.34%	97.94%	98.79%	98.54%	98.26%

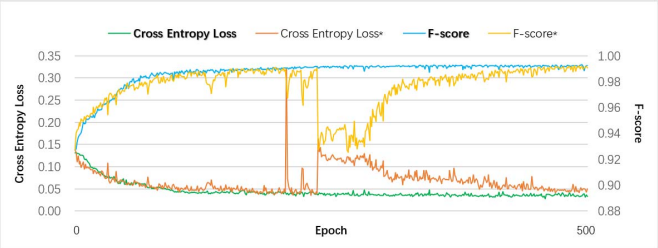


Fig. 11. Influence of state entropy constraint on the optimization of StateNet. The superscript * indicates that state entropy regular term is removed during the training process. State entropy loss actually harms performance in the early stages, but it can effectively avoid the exploding gradient, and the optimization becomes more stable. The data come from the RS test set.

in Table IV. The results show that it is beneficial to expand the scope of the neighborhood appropriately. This is because a larger neighborhood can provide more *a priori* information for correspondence estimation, and matching-neighbors selector can adaptively filter redundant features at the same time.

2) State Entropy Constraint: Second, to provide some insight into the influence of state entropy constraint on the optimization of StateNet, example training curves for runs of the baseline loss function (i.e., cross entropy) and hybrid loss function [i.e., (10)] are shown in Fig. 11. It can be observed from Fig. 11 that state entropy constraint yields a steady improvement throughout the optimization procedure. In fact, although the cross entropy loss can train the model by itself, adding the state entropy regularization term can efficiently alleviate the exploding gradient and overfitting problems.

3) Number of Cascades and Classifier Form: Finally, Table V presents the quantitative results with respect to different numbers of recursive cascades (i.e., the number of ASL blocks) and different forms of classification networks (i.e., PRL + CNN, ADAConv, PointCNN [60], and PointNet [3]). As shown in Table V, recursive cascaded ASL achieves

TABLE V

ABLATION STUDIES OF OUR STATENET ON FOUR COMMON APPLICATION SCENARIOS (**RS**: SATELLITE RS IMAGES AND LOW-ALTITUDE RS IMAGE, **M-RS**: MULTIMODAL RS IMAGES, AND **OS**: OUTDOOR SCENES). CAS. STANDS FOR THE NUMBER OF CASCADES, CLAS. STANDS FOR THE FORM OF THE CLASSIFIER, AND THE VALUES REPRESENT THE AVERAGE F-SCORE

$\mathbf{F}_{ns}$	$\mathbf{F}_{tr}$	Cas.	Clas.	RS	M-RS	OS
×	×	/	PRL(CNN)	93.49%	85.61%	83.98%
×	✓	/	PRL(CNN)	96.29%	85.86%	87.81%
✓	×	/	PRL(CNN)	95.07%	86.94%	87.02%
✓	✓	1	PRL(CNN)	97.37%	89.88%	92.05%
✓	✓	2	PRL(CNN)	98.79%	91.80%	94.54%
✓	✓	3	PRL(CNN)	**99.09%**	92.25%	95.48%
✓	✓	4	PRL(CNN)	99.04%	**92.49%**	95.42%
✓	✓	5	PRL(CNN)	98.49%	92.44%	**96.09%**
×	×	/	ADAConv	92.90%	85.06%	83.71%
×	✓	/	ADAConv	96.01%	85.18%	86.61%
✓	×	/	ADAConv	95.41%	86.79%	86.15%
✓	✓	1	ADAConv	97.09%	88.15%	92.02%
✓	✓	2	ADAConv	97.98%	90.81%	95.01%
✓	✓	3	ADAConv	98.20%	91.23%	**95.47%**
✓	✓	4	ADAConv	**98.46%**	92.12%	94.78%
✓	✓	5	ADAConv	98.42%	**92.14%**	94.80%
×	×	/	PointCNN	91.87%	84.96%	83.81%
×	✓	/	PointCNN	96.12%	84.98%	86.41%
✓	×	/	PointCNN	95.45%	85.89%	86.08%
✓	✓	1	PointCNN	96.94%	87.17%	92.21%
✓	✓	2	PointCNN	97.74%	89.64%	94.83%
✓	✓	3	PointCNN	98.09%	89.88%	94.88%
✓	✓	4	PointCNN	98.17%	**91.51%**	94.76%
✓	✓	5	PointCNN	**98.23%**	90.42%	**95.53%**
×	×	/	PointNet	92.07%	84.66%	83.83%
×	✓	/	PointNet	96.22%	84.49%	86.17%
✓	×	/	PointNet	95.47%	85.69%	85.28%
✓	✓	1	PointNet	96.98%	87.87%	92.91%
✓	✓	2	PointNet	97.81%	**91.24%**	93.93%
✓	✓	3	PointNet	98.01%	90.17%	93.78%
✓	✓	4	PointNet	**98.15%**	90.01%	94.66%
✓	✓	5	PointNet	98.03%	90.14%	**94.73%**

consistent performance gains independently of the classification network. More importantly, when the number of cascades is two, the network already has a mutual promotion mechanism of outliers suppression and transformation estimation. Although the performance may be further improved as the number of cascades increases, it will also cause linear increments to the running times. If more training samples are available or a distributed learning platform is being used, the performance can be further improved by deeper cascades. In addition, we observe that using PRL and CNN provide relatively high matching performance compared with the classifiers directly consuming point cloud data.

V. CONCLUSION

This article introduced a cascading state learning mechanism on sparse points to regularize the matching pattern, which demonstrates the power of end-to-end net for mismatch removal in feature matching. From the perspective

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on February 10,2023 at 14:55:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

of optimizing the state entropy, our network refines initial matches by filtering out the outliers and then computing the relative transformation. In that regard, two problems are solved in a learning fashion: soft correspondence estimation and transformation update. In terms of network architecture, the proposed ADAConv, which encourages the StateNet to better learn the global geometrical feature of sparse points. Furthermore, the PRL casts the arbitrary permutation of the points into a potentially canonical order in space, which means that leveraging the spatially local correlation of point cloud data becomes easier. The experimental results illustrate that StateNet achieves significant improvement over the state-of-the-art competitors.

Our approach focuses on the local consensus learning of each correspondence and ignores the global geometric constraint and the content of source images. In the future work, it bears thinking about how to leverage global context in original data to further improve matching performance. We may be able to introduce an additional context-aware module for encoding global location information, and the state learning framework would remain largely unchanged.

ACKNOWLEDGMENT

The authors would like to thank Su Zhang, Jiayi Ma, Jiahui Zhang, and Jiayuan Li for providing their implementation source codes and experimental datasets, which facilitate the comparison experiments greatly.

REFERENCES

[1] X. Jiang, J. Jiang, A. Fan, Z. Wang, and J. Ma, "Multiscale locality and rank preservation for robust feature matching of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6462–6472, Sep. 2019.

[2] J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo, "Locality preserving matching," *Int. J. Comput. Vis.*, vol. 127, no. 5, pp. 512–531, 2019.

[3] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 652–660.

[4] K. M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua, "Learning to find good correspondences," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2666–2674.

[5] J. Zhang *et al.*, "Learning two-view correspondences and geometry using order-aware network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 5845–5854.

[6] P. J. Besl and D. N. McKay, "A method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[7] H. Chui and A. Rangarajan, "A new algorithm for non-rigid point matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2000, pp. 44–51.

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

[9] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2262–2275, Dec. 2010.

[10] B. Jian and B. C. Vemuri, "Robust point set registration using Gaussian mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1633–1645, Aug. 2010.

[11] Z. Zhou *et al.*, "Accurate and robust non-rigid point set registration using Student's-t mixture model with prior probability modeling," *Sci. Rep.*, vol. 8, no. 1, p. 8742, Dec. 2018.

[12] H.-B. Qu, J.-Q. Wang, B. Li, and M. Yu, "Probabilistic model for robust affine and non-rigid point set matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 371–384, Feb. 2016.

[13] S. Zhang, Y. Yang, K. Yang, Y. Luo, and S. H. Ong, "Point set registration with global-local correspondence and transformation estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2669–2677.

[14] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *Int. J. Comput. Vis.*, vol. 129, pp. 1–57, Jan. 2020.

[15] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[16] P. H. S. Torr and A. Zisserman, "MLESAC: A new robust estimator with application to estimating image geometry," *Comput. Vis. Image Understand.*, vol. 78, no. 1, pp. 138–156, 2000.

[17] O. Chum, J. Matas, and J. Kittler, "Locally optimized RANSAC," in *Proc. Joint Pattern Recognit. Symp.* Berlin, Germany: Springer, 2003, pp. 236–243.

[18] O. Chum and J. Matas, "Matching with PROSAC-progressive sample consensus," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 220–226.

[19] T. Sattler, B. Leibe, and L. Kobbelt, "SCRAMSAC: Improving RANSAC's efficiency with a spatial consistency filter," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2090–2097.

[20] D. Barath, J. Matas, and J. Noskova, "MAGSAC: Marginalizing sample consensus," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 10197–10205.

[21] D. Barath, J. Noskova, M. Ivashechkin, and J. Matas, "MAGSAC++, a fast, reliable and accurate robust estimator," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 1304–1312.

[22] X. Li and Z. Hu, "Rejecting mismatches by correspondence function," *Int. J. Comput. Vis.*, vol. 89, no. 1, pp. 1–17, 2010.

[23] J. Ma, J. Zhao, J. Tian, A. L. Yuille, and Z. Tu, "Robust point matching via vector field consensus," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1706–1721, Apr. 2014.

[24] J. Ma, J. Wu, J. Zhao, J. Jiang, H. Zhou, and Q. Z. Sheng, "Nonrigid point set registration with robust transformation learning under manifold regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 12, pp. 3584–3597, Dec. 2018.

[25] L. Torresani, V. Kolmogorov, and C. Rother, "Feature correspondence via graph matching: Models and global optimization," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2008, pp. 596–609.

[26] M. Leordeanu and M. Hebert, "A spectral technique for correspondence problems using pairwise constraints," in *Proc. IEEE 10th Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Oct. 2005, pp. 1482–1489.

[27] H. Liu and S. Yan, "Common visual pattern discovery via spatially coherent correspondences," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1609–1616.

[28] M. Cho and K. Mu Lee, "Mode-seeking on graphs via random walks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 606–613.

[29] J. Yan, M. Cho, H. Zha, X. Yang, and S. Chu, "Multi-graph matching via affinity optimization with graduated consistency regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1228–1242, Jun. 2015.

[30] J. Yan, J. Wang, H. Zha, X. Yang, and S. Chu, "Consistency-driven alternating optimization for multigraph matching: A unified approach," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 994–1009, Mar. 2015.

[31] J. Yan, C. Li, Y. Li, and G. Cao, "Adaptive discrete hypergraph matching," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 765–779, Feb. 2017.

[32] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T.-D. Nguyen, and M.-M. Cheng, "GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4181–4190.

[33] J.-W. Bian *et al.*, "GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1580–1593, Jun. 2020.

[34] J. Ma, J. Jiang, H. Zhou, J. Zhao, and X. Guo, "Guided locality preserving feature matching for remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4435–4447, Aug. 2018.

[35] X. Jiang, J. Ma, J. Jiang, and X. Guo, "Robust feature matching using spatial clustering with heavy outliers," *IEEE Trans. Image Process.*, vol. 29, pp. 736–746, 2020.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[37] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, 2017, pp. 2961–2969.

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on February 10,2023 at 14:55:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[38] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767

[39] R. Wang, J. Yan, and X. Yang, "Combinatorial learning of robust deep graph matching: An embedding based approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, p. 1, 2020, doi: 10.1109/TPAMI.2020.3005590.

[40] R. Wang, J. Yan, and X. Yang, "Neural graph matching network: Learning Lawler's quadratic assignment problem with extension to hypergraph and multiple-graph matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, p. 1, 2021, doi: 10.1109/TPAMI.2021.3078053.

[41] J. Ma, X. Jiang, J. Jiang, J. Zhao, and X. Guo, "LMR: Learning a two-class classifier for mismatch removal," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4045–4059, Aug. 2019.

[42] J. Chen, S. Chen, Y. Liu, X. Chen, Y. Yang, and Y. Zhang, "Robust local structure visualization for remote sensing image registration," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1895–1908, 2021.

[43] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[44] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Trans. Image Process.*, vol. 29, pp. 3296–3310, 2020.

[45] Y. Zheng and D. Doermann, "Robust point matching for nonrigid shapes by preserving local neighborhood structures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 643–649, Apr. 2006.

[46] J. Hu, L. Shen, and G. Sun, "Squeeze- and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[47] Z. Cui, H. Chang, S. Shan, B. Zhong, and X. Chen, "Deep network cascade for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2014, pp. 49–64.

[48] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3150–3158.

[49] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[50] S. Zhao, Y. Dong, E. Chang, and Y. Xu, "Recursive cascaded networks for unsupervised medical image registration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10600–10610.

[51] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 1469–1472.

[52] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[54] S. Zhang, W. Zhao, X. Hao, Y. Yang, and C. Guan, "A context-aware locality measure for inlier pool enrichment in stepwise image registration," *IEEE Trans. Image Process.*, vol. 29, pp. 4281–4295, 2020.

[55] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[56] K. Mikolajczyk *et al.*, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, no. 1, pp. 43–72, 2005.

[57] Y. Abdel-Aziz, H. Karara, and M. Hauck, "Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry," *Photogram. Eng. Remote Sens.*, vol. 81, no. 2, pp. 103–107, 2015.

[58] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 6, pp. 567–585, Jun. 1989.

[59] Z. Yang, Y. Yang, K. Yang, and Z.-Q. Wei, "Non-rigid image registration with dynamic Gaussian component density and space curvature preservation," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2584–2598, May 2019.

[60] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "PointCNN: Convolution on x-transformed points," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 820–830.

Jiaxuan Chen received the bachelor's degree in information management and information system from Northwest Normal University, Lanzhou, China, in 2018. He is currently pursuing the master's degree with the School of Information Science and Technology, Yunnan Normal University, Kunming, China.

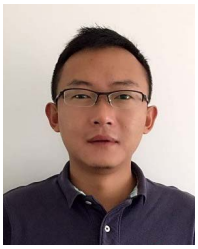
His current research interests include computer vision, remote sensing image processing, and deep learning.

Shuang Chen received the bachelor's degree in information management and information system from China West Normal University, Nanchong, China, in 2019. She is currently pursuing the master's degree with the School of Information Science and Technology, Yunnan Normal University, Kunming, China.

Her current research interests include computer vision, remote sensing image processing, and deep learning.

Xiaoxian Chen received the bachelor's degree in business administration from Guangdong Pharmaceutical University, Guangzhou, China, in 2018, and the master's degree in computer technology from China Agricultural University, Beijing, China, in 2021.

He is currently working with JD.com, Beijing, and also a Visiting Scholar with Yunnan Normal University, Kunming, China. His current research interests include deep learning, deep reinforcement learning, and recommendation systems.



Yang Yang (Member, IEEE) received the master's degree in computer science from Waseda University, Tokyo, Japan, in 2007, and the Ph.D. degree in computer science from the National University of Singapore, Singapore, in 2013.

He is currently a Professor with the School of Information Science and Technology, Yunnan Normal University, Kunming, China. His research interests include computer vision, remote sensing, geography information systems, and medical imaging.

Yujing Rao received the bachelor's degree in computer science from China West Normal University, Nanchong, China, in 2019. She is currently pursuing the master's degree with the School of Information Science and Technology, Yunnan Normal University, Kunming, China.

Her current research interests include image fusion and deep learning.

Authorized licensed use limited to: Xian Jiaotong University. Downloaded on February 10,2023 at 14:55:42 UTC from IEEE Xplore. Restrictions apply.