IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 28, 2018, accepted June 19, 2018, date of publication July 5, 2018, date of current version July 30, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2853100

Multi-Temporal Remote Sensing Image Registration Using Deep Convolutional Features

ZHUOQIAN YANG[1], TINGTING DAN[2], AND YANG YANG[2,3]

[1]College of Software, Beihang University, Beijing 100083, China
[2]School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China
[3]The Engineering Research Center of GIS Technology in Western China, Ministry of Education, Yunnan Normal University, Kunming 650500, China

Corresponding authors: Tingting Dan (dandycn721@gmail.com) and Yang Yang (yyang_ynu@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 41661080 and in part by the National Basic Research Program of China (973 Program) under Grant 2010CB434803.

ABSTRACT Registration of multi-temporal remote sensing images has been widely applied in military and civilian fields, such as ground target identification, urban development assessment, and geographic change assessment. Ground surface change challenges feature point detection in amount and quality, which is a common dilemma faced by feature-based registration algorithms. Under severe appearance variation, detected feature points may contain a large proportion of outliers, whereas inliers may be inadequate and unevenly distributed. This paper presents a convolutional neural network (CNN) feature-based multi-temporal remote sensing image registration method with two key contributions: (i) we use a CNN to generate robust multi-scale feature descriptors and (ii) we design a gradually increasing selection of inliers to improve the robustness of feature point registration. Extensive experiments on feature matching and image registration are performed over a multi-temporal satellite image data set and a multi-temporal unmanned aerial vehicle image dataset. Our method outperforms four state-of-the-art methods in most scenarios.

INDEX TERMS Remote sensing, feature matching, image registration, convolutional feature.

I. INTRODUCTION

Image registration is the process of finding the optimal alignment between images. It is a fundamental task in order to be able to integrate and compare images captured from different viewpoints, at different times or with different sensors. Registration of multi-temporal remote sensing images has been widely applied in military and civilian fields such as ground target identification, urban development assessment, geographic change assessment.

Approaches of image registration can be classified into two major categories: (i) area based methods and (ii) feature based methods [1]–[5]. Instead of working directly with image intensity values (area-based methods), feature-based methods employs feature descriptors that represent high-level information, thus is more preferable in multi-temporal analysis where appearance variation is expected [6]. Since we mainly focus on developing a feature based method in this work, we introduce and discuss current methods amongst (ii).

The majority of feature based methods rely on SIFT [7] or its improved version to detect feature points due to its outstanding invariance against scale and rotation [8]–[11]. Nevertheless, in multi-temporal or multi-sensor image registration where certain extent of appearance difference exists, feature points detected by SIFT may contain severe outliers. In worse cases, SIFT cannot detect sufficient number of feature points. Such issues limit the application of image registration.

In this work, we propose a novel non-rigid image registration method. Two of our essential contributions can be summarized as follows. (i) We generate a multi-scale feature descriptor using layers from a pretrained VGG [12] network. Aiming at the effective utilization of convolutional neural networks in image registration, our feature utilizes high level convolutional information while preserving some localization capabilities. (ii) We design a point-set registration that works in accordance with the proposed feature. Instead of using a stationary distinction of inliers and outliers, we design a gradually increasing selection of inliers. At the early stage of registration, the rough transformation is rapidly determined by the most reliable feature points. After which the registration details are optimized by increasing the number of feature points while restricting the mismatches simultaneously. The point-wise correspondence is evaluated by both the convolutional features and geometric structural information.

We compare our feature detection method against SIFT. Our image registration method is tested on multi-temporal satellite images and UAV images, compared against four state-of-the-art works. We compare feature detection methods by measuring the precision of feature prematching. The performance of registration is evaluated by measuring the distance between corresponding pixels.

The rest of this paper is organized as follows. Section 2 reviews classic and cutting-edge researches regarding feature-based registration problem. Section 3 presents the detailed methodology of our work. Section 4 demonstrates our experiments. Conclusion is drawn in Section 5.

II. RELATED WORKS

Feature-based image registration methods typically consists of four stages.

1) A sufficient number of feature points are detected in a pair of images (i.e., the sensed and reference image) using feature descriptors like SIFT [7].
2) Estimate a preliminary point-wise correspondence by finding the nearest neighbors in a feature space, which we call feature prematching.
3) The non-rigid point set registration [13]–[17] which searches for optimal transformation parameters.
4) Image transformation, which resamples the sensed image according to the recovered transformation.

The majority of feature-based methods rely on SIFT or its improved version to detect feature points. A variety of SIFT based methods were developed over the last few years, they introduce different techniques to enhance feature point matching. Random sample consensus (RANSAC) [18] and its variants [19]–[22] are widely used for remote sensing registration, propose to use a hypothesize-and-verify framework to eliminate false correspondences. Point set registration by preserving global and local structures (PR-GLS) [13] uses local features such as shape context to assign the membership probabilities of the mixture model, so that both global and local structures can be preserved during the matching process. In Wang et al.'s robust registration using spatially constrained gaussian fields (SCGF), intrinsic manifold is considered and used to preserve the geometrical structure. Also a priori knowledge of the point set is extracted and applied. Context-aware Gaussian fields (CA-LapGF) [23] proposes a Laplacian regularized term which is added to preserve the essential geometry of the transformed set. Ma *et al.* [24] employs a local geometric constraint which applies well on retinal images. Yang et al.'s global local feature distance (GLMD) [25] considers global and local structural differences of two point sets as a linear assignment problem and recovered correspondences by using mixture features. Zhang *et al.* [26] developed an effective method that maintains a high matching ratio on inliers while taking advantage of outliers for varying the warping grids. Bilateral KNN spatial orders around geometric centers (Bi-SOGC) [27] is a graph matching approach based on bilateral K nearest neighbors spatial orders around geometric centers, which enhances feature matching with spatial relations. Optimization and iterative logistic regression matching (OILRM) [28] combines optimization model and logistic regression to improve linear object matching. Wu *et al.* [29] employed a weighted total least squares (WTLS) based estimator to cope with control point coordinates that are of unequal accuracies. Zhao *et al.* [30], [31] propose to achieve robust feature point matching by removing outliers and reserve sufficient inliers for remote sensing images. Yeand *et al.* [32] uses support-line voting and affine-invariant ratios to serve the same purpose. These SIFT based methods suffer from the problems of insufficient feature points and high outlier ratio under severe appearance change.

There are also diverse solutions in the phase of point set registration. A classical approach is by the means of probability optimization, measuring the degree of alignment with a Gaussian mixture model (GMM). One representative GMM based method is the Coherent Point Drift (CPD) [33], which places a Gaussian distribution centroid on each of the sensed feature points and then iteratively update point locations under the expectation maximization framework. Wang *et al.* [16] proposes to use a mixture of asymmetric Gaussian models (MoAG) and apply soft assignment technique to recover the correspondences. This method uses a correlation-based method to estimate the transformation parameters. Ma *et al.* [34] introduced vector field consensus (VFC) which can efficiently establish robust point correspondences between two sets of points. Their successive work SparseVFC [35] demonstrates significant speed advantage over the original VFC algorithm without much performance degradation. Locally linear transforming (LLT) [14] starts by creating a set of putative correspondences based on the feature similarity and then focus on removing outliers from the putative set and estimating the transformation as well. They formulate this as a maximum-likelihood estimation of a Bayesian model with hidden/latent variables indicating whether matches in the putative set are outliers or inliers. Zhang *et al.* [36] proposed a dual-feature approach that preserves the structure at both global and local scales during the registration. Manifold regularized coherent vector field (MRCVF) [37] also approach by removing outliers. It learns coherent vector fields fitting for the inliers with graph Laplacian constraint. An optimization strategy which instead minimizes a residual complexity was introduced in Myronenko et al.'s paper [38], it derives the similarity measure by analytically solving for the intensity correction field and its adaptive regularization. A simulated, astrophysics based Gravitational Approach (GA) [39] formulates registration as a modified N-body problem. It mimics a template point set moving in a viscous medium under gravitational forces induced by a reference point set. This method currently registers rigid point sets. Vongkulbhisal *et al.* [40] proposed the Discriminative Optimization (DO), which learns search

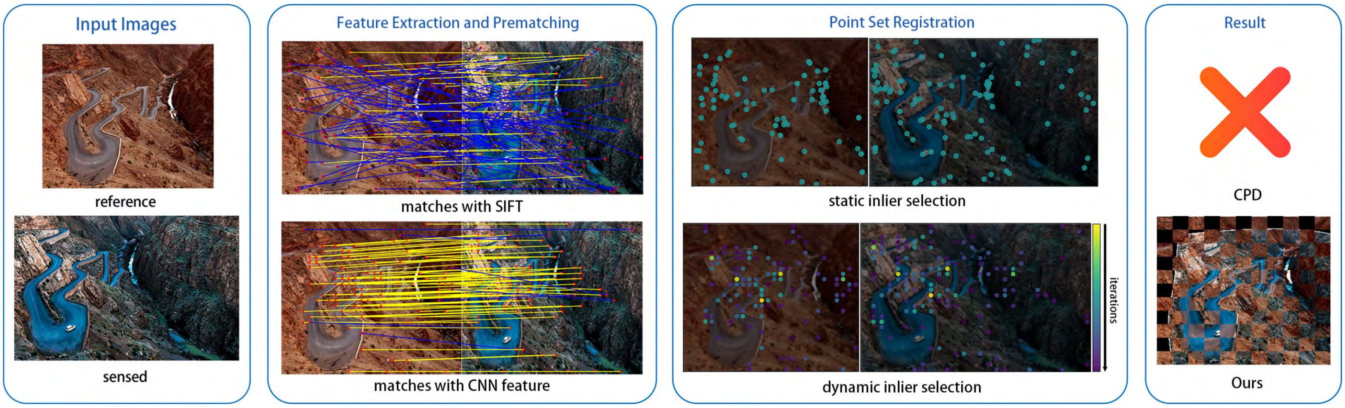


FIGURE 1. Process diagram of our method with comparison to CPD. In feature extraction and prematching, equal number of feature points are generated using SIFT and our CNN feature. Correct feature point matches are denoted by yellow lines, incorrect ones are denoted by blue lines. In the point set registration stage, we use an increasing selection of inliers. Only prominent feature point pairs are used to estimate the transformation at early iterations, other feature points are moved coherently.

directions from data without the need of a cost function. Specifically, DO explicitly learns a sequence of updates in the search space that leads to stationary points that correspond to desired solutions.

In the past few years, convolutional neural networks (CNNs) have been studied for processing remote sensing data [41]. Zhang *et al.* developed a set of effective CNN based methods for extracting features [42], classify scenes [43] and detect specific ground targets [44]. A number of works use CNNs to learn feature description. MatchNet [45] presents a unified approach to learn feature representations and learn feature comparison through which improved computational efficiency is achieved. It consists of a deep convolutional network that extracts features from patches and a network of three fully connected layers that computes a similarity between the extracted features. In order to overcome the shortage of labeled data, Du *et al.* [46] proposes a general active learning framework that effectively fuses the representativeness and informativeness of data and an unsupervised deep network called the stacked convolutional denoising auto-encoders [47], which can map images to hierarchical representations without any label information. Žbontar and LeCun [48] train a convolutional neural network to predict how well two image patches match and use it to compute the stereo matching cost. These methods use relatively large image patches and focus on computing certain metric from the image patches, whereas localization is not demanded. In our approach, we attempt to utilize high level convolutional information while preserving some localization capabilities.

Several methods have been developed to approach category level registration. These works attempt to train specific networks for feature extraction and registration. Kanazawa *et al.* [49] constructed a Siamese network to predict transformations and trains it on fine-grained datasets. Rocco *et al.* [50] propose an architecture based on three main components that mimic the standard steps of feature extraction, matching and model parameter estimation, each one is trainable network. In our proposed approach, CNN is only utilized for feature extraction, for point set registration we build a novel method on traditional frameworks. The reason for such choice is that neural networks can only yield a limited, constant number of transformation parameters, thus is incapable of rectifying complicated distortion and unsuitable for remote sensing registration.

III. METHOD

A. SOLUTION FRAMEWORK

The objective of the algorithm is to transform a sensed image I_Y so that it is aligned to a reference image I_X. We detect a feature point set X from the reference image and a feature point set Y from the sensed image. Next we use a expectation maximization (EM) based procedure to obtain the transformed locations of Y, namely Z. Y and Z are then used to solve a thin plate spline (TPS) interpolation for image transformation. The main process of our method is shown in Fig.1

Throughout the paper we use the following notations:

- $X_{N\times 2}$, $Y_{M\times 2}$ - feature point set extracted from the reference image and the sensed image, respectively.
- Z - transformed locations of Y.
- N, M - the number of points in X and Y, respectively.
- x_n, y_m - point at index n in point set X; point at index m in point set Y.

B. FEATURE DESCRIPTION AND PREMATCHING

1) GENERATING FEATURE DESCRIPTORS

Our convolutional feature descriptor is constructed using the output of certain layers in a pretrained VGG-16 [12] network, which is a image classification network that classifies 1000 categories. VGG is selected for this task due to some of its desirable characteristics: (i) Its remarkable performance on image classification proves its resolving

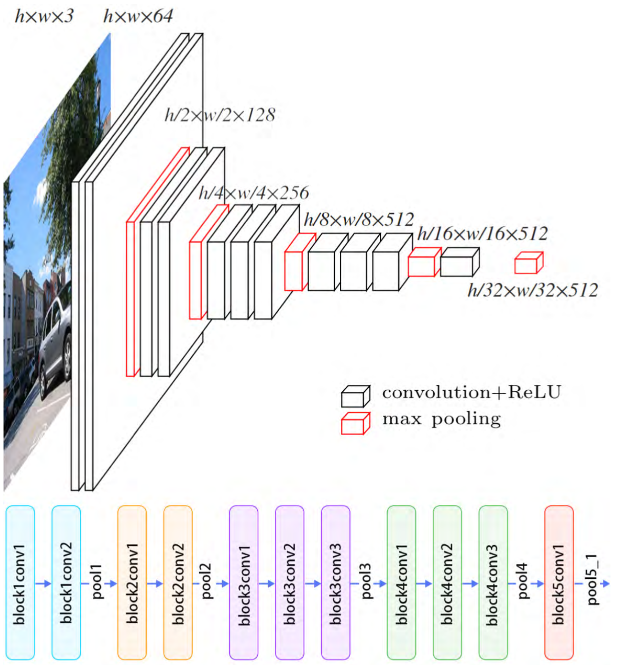


FIGURE 2. Slightly modified VGG-16 Network Architecture. *h* and *w* respectively represents the height and width of the input image. Since we only use convolution layers to extract features, the input image can be of any size as long as *h* and *w* are multiples of 32.

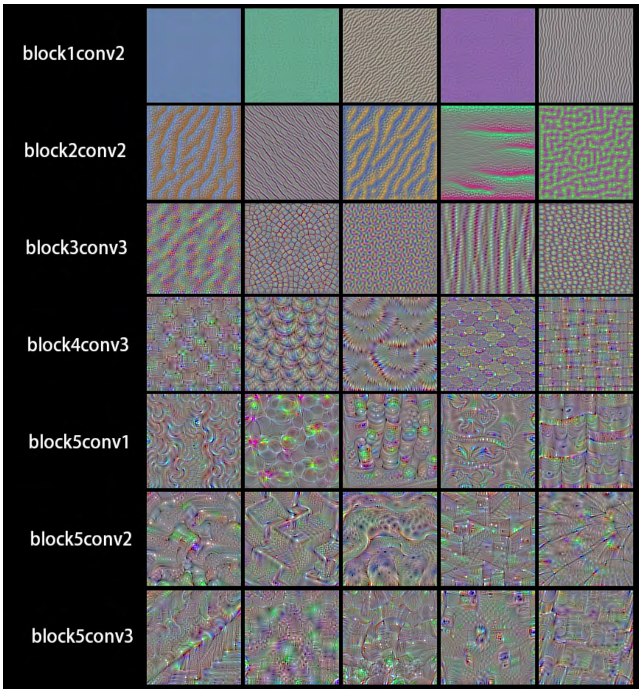


FIGURE 3. Visualization of Convolution Filters. With the increase of depth in the convolutional neural network using convolution layers and pooling layers, the pattern searched by convolution filters becomes larger in scale and tend to be more sophisticated.

power. (ii) It is concise in structure, constructed simply by stacking convolution, pooling and fully connected layers while no branches or shortcut connections are employed to reinforce gradient flow. Such design made adapting this network for different purposes practicable. (iii) It is extremely deep, trained on enormous and diversified image data. As a result, its convolution filters searches universal patterns and generalizes very well. VGG is frequently used for feature extraction in many computer vision solutions, such as faster-RCNN [51] object detector and super-resolution generative adversarial network (SRGAN) [52].

Based on visualization of convolution filters and trial-and-error experiments using single layer output as feature, several network layers have been selected to build our feature descriptor. We mainly consider the generalizability of convolution filters and the receptive field size when selecting the layers. A convolution layer in a neural network contains various small filters and each searches for a specific pattern in the input image. The filters in each convolution layer of VGG-16 are visualized by applying gradient ascent [53] on an input image generated using random values. We choose to use the VGG network trained on Imagenet dataset [54] so that our feature descriptor searches for common, universal patterns. Fig.3 shows representative visualized filters. The pool5 layer is not used for feature because it is affected by specific classification objects thus not suitable for detecting general features.

Since we only use convolution layers to extract features, the input image can be of any size as long as the height and the width are multiples of 32. However, the size of the input image can have two aspects of influence: (i) The receptive field of each descriptor would be different and affect the performance. (ii) Larger input images require more computation. We resize input images to 224 × 224 before propagating them through the network in order to have properly sized receptive fields and reduced computation. The outputs of three layers are used to build our feature: pool3, pool4 and a max-pooling layer added after block5conv1, namely pool5_1. These layers searches for a universal set of patterns and yield feature response values that well cover different sizes of receptive fields.

As shown in Fig.2, VGG-16 contains 5 blocks of convolution computation, each with 2-3 convolution layers and a max-pooling layer at the end of each block. We lay a 28 × 28 grid over the input image dividing our patches, each corresponding to a 256-d vector in the pool3 output, a descriptor is generated in every 8 × 8 square. The center of each patch is regarded as a feature point. The 256-d vector is defined as the pool3 feature descriptor. The pool3 layer output directly forms our pool3 feature map F_1, which is of size 28 × 28 × 256. The pool4 layer output, which is of size 14 × 14 × 512, is handled slightly differently. In every 16 × 16 area we generate a pool4 descriptor, therefore it is shared by 4 feature points. As shown in Eq.1, pool4 feature map F_2 is acquired using Kronecker product (denoted by $\bigotimes$):

$$F_2 = O_{\text{pool4}} \bigotimes I_{2\times2\times1} \tag{1}$$

O_{pool4} stands for the output of pool4 layer. I denotes a tensor of subscripted shape and filled with 1s.

The pool5_1 layer output is of size 7 × 7 × 512. Similarly, every pool5_1 descriptor is shared by 16 feature points.

$$F_3 = O_{\text{pool5\_1}} \bigotimes I_{4\times4\times1} \tag{2}$$

The distribution of feature descriptors is shown is Fig.4.

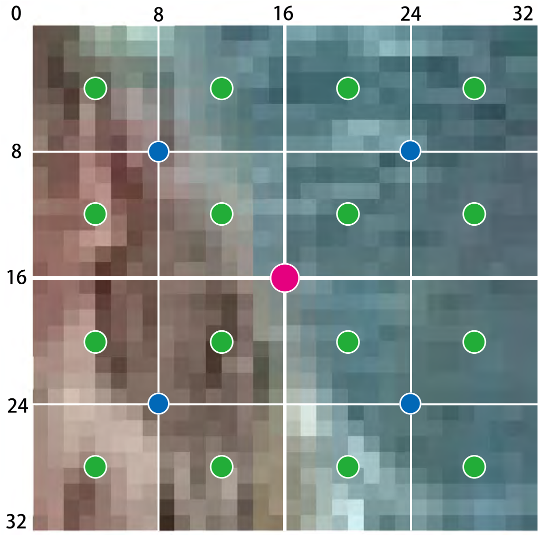


FIGURE 4. Distribution of feature descriptors. This figure shows the distribution of feature descriptors in a 32 × 32 squared region. Green dots denote pool3 descriptors, generated in a 8 × 8 squared region. Blue dots denote pool4 descriptors, each shared by 4 feature points. The cyan dot denote a pool5_1 descriptor, shared by 16 feature points.

After acquiring F_1, F_2 and F_3, the feature maps are normalized to unit variance using

$$F_i \leftarrow \frac{F_i}{\sigma(F_i)}, \quad i = 1, 2, 3 \tag{3}$$

where $\sigma(\cdot)$ computes the standard deviation of elements in a matrix. The pool3, pool4 and pool5_1 descriptors of point x are denoted by $D_1(x)$, $D_2(x)$ and $D_3(x)$ respectively.

2) FEATURE PREMATCHING

We first define the distance metric of our feature. The feature distance between two feature points x and y is a weighted sum of three distance values

$$d(x, y) = \sqrt{2}d_1(x, y) + d_2(x, y) + d_3(x, y) \tag{4}$$

and each component distance value is the Euclidean distance between the respective feature descriptors

$$d_i(x, y) = \text{Euclidean-distance}(D_i(x), D_i(y)) \tag{5}$$

The distance computed with pool3 descriptors $d_1(x, y)$ is compensated with a weight $\sqrt{2}$ because D_1 is 256-d whereas D_2 and D_3 are 512-d.

Feature point x is matched to y if the following conditions are satisfied:

1) $d(x, y)$ is the smallest of all $d(\cdot, y)$
2) There does not exist a $d(z, y)$ such that $d(z, y) < \theta \cdot d(x, y)$. θ is a parameter valued greater than 1 and is called the matching threshold.

This matching method does not guarantee bijection.

C. DYNAMIC INLIER SELECTION

Our feature points are generated at the center of square shaped image patches. Under circumstances of deformation, corresponding feature points may have their image patches overlapping partly or completely. Therefore, to achieve more accurate registration, feature points with larger overlapping ratios should have a better degree of alignment, where as partly overlapping patches should have a small distance between their centers. The degree of alignment is determined using our dynamic inlier selection.

While using EM algorithm to iteratively solve Z (the transformed locations of Y in every iteration), we update the selection of inliers in every k iterations. Points selected as inliers guide the movement of point locations whereas outliers are moved coherently. At the feature prematching stage, a large number of feature points are selected using a low threshold θ_0 to filter out irrelevant points. Then we designate a large starting threshold $\hat{\theta}$ that only confident inliers (feature points with overlapping patches) satisfy. In the rest of registration process, threshold θ is subtracted by a step-length δ in every k iterations, allowing a few more feature points to affect the transformation. Such practice enables strongly matched feature points to determine the overall transformation while other feature points optimize registration accuracy.

Inlier selection produces a $M \times N$ prior probability matrix P_R which is then taken by our Gaussian mixture model (GMM) based transformation solver. The entry $P_R[m, n]$ of this matrix, is the putative probability of x_n and y_m to be corresponding. Supposing that x_n is corresponding to y_m, we obtain a large putative probability $P_R[m, n]$. And a large probability would further lead to a conspicuous transformation over y_m by which the corresponding pair can be aligned.

The putative probabilities are determined using both the convolutional feature and geometric structural information. Prior probability matrix P_R is obtained through the following procedure:

1) Prepare the $M \times N$ convolutional feature cost matrix C_θ^{conv} by

$$C_\theta^{\text{conv}}[m, n] = \begin{cases} \dfrac{d(y_m, x_n)}{d_\theta^{\max}}, & \text{condition 1} \\ 1, & \text{otherwise.} \end{cases} \tag{6}$$

Condition 1 is when y_m and x_n are a valid match under threshold θ. $d(\cdot, \cdot)$ is the previously defined distance metric of our convolutional feature. $d_\theta^{\max}$ is the maximum distance of all matched feature point pairs under threshold θ.

2) Compute a geometric structure cost matrix C^{geo} using shape context [55], which is a histogram based

descriptor that profiles neighborhood structure of a point. The descriptor places the profiled point at the center of a polar coordinate system and records the number of points fell in arc-shaped bins. C^{geo} is acquired by performing a χ^2 test

$$C^{\text{geo}}[m,n] = \frac{1}{2}\sum_{b=1}^{B}\frac{[h_m^y(b) - h_n^x(b)]^2}{h_m^y(b) + h_n^x(b)}, \tag{7}$$

where $h_m^y(b)$ and $h_n^x(b)$ denotes the number of points fell in the b^{th} bin surrounding y_m and x_n, respectively.

3) Both C_θ^{conv} and C^{geo} are valued in $[0, 1]$. We compute a integrated cost matrix C using a element-wise Hadamard product (denoted by $\odot$):

$$C = C_\theta^{\text{conv}} \odot C^{\text{geo}} \tag{8}$$

4) We apply Jonker-Volgenant algorithm [56] to solve the linear assignment on cost matrix C. Assigned point pairs are regarded as putatively corresponding. Finally, we compute the prior probability matrix using

$$P_R[m,n] = \begin{cases} 1, & \text{if } y_m \text{ and } x_n \text{ are corresponding} \\ \dfrac{1-\epsilon}{N}, & \text{otherwise.} \end{cases} \tag{9}$$

ϵ is a hyper-parameter valued in $[0, 1]$ which should be designated according to our confidence of the inlier selection to be accurate. Prior probability matrix requires normalization:

$$P_R[m,n] := \frac{P_R[m,n]}{\sum_{k=1}^{N} P_R[m,k]} \tag{10}$$

The step length of threshold is determined by $\delta = \frac{\hat{\theta}-\theta_0}{10}$.

D. MAIN PROCESS

We consider point set Y as Gaussian mixture model (GMM) centroids. The GMM probability density function is defined as:

$$p(x) = \omega\frac{1}{N} + (1-\omega)\sum_{m=1}^{M} g_m(x) \tag{11}$$

$g_m(x)$ is a normal distribution density function:

$$g_m(x) = \frac{1}{2\pi\sigma^2}\exp(-\frac{1}{2\sigma^2}\,\|x - y_m\|^2) \tag{12}$$

The model uses isotropic variances σ^2 for every single Gaussian centroid in the mixture. An additional uniform distribution term $\frac{1}{N}$ is added to account for outliers with a weighting parameter ω, $0 < \omega < 1$.

We then use expectation maximization (EM) algorithm to find the optimal transformation parameters (W, σ^2, ω). The objective of such approach is to maximize a likelihood function, or equivalently minimize the negative log-likelihood function:

$$L(W,\sigma^2,\omega) = -\sum_{n=1}^{N} log \sum_{m=1}^{M+1} P_R[m,n]g_m(x_n), \tag{13}$$

from which we cannot directly compute gradients due to the existence of unobservable variable m. Alternatively, EM algorithm minimizes the expectation of the negative log-likelihood function:

$$Q = -\sum_{n=1}^{N}\sum_{m=1}^{M+1} P^{\text{old}}(m|x_n)\log(P_R[m,n]g_m(x_n)), \tag{14}$$

$P^{\text{old}}(m|x_n)$ denotes a posterior probability term computed using parameters from the last iteration. After expanding this equation and omitting derivative-redundant terms, the equation can be rewritten as:

$$\begin{aligned} Q(W,\sigma^2,\omega) = {} & \frac{1}{2\sigma^2}\sum_{n=1}^{N}\sum_{m=1}^{M} P^{\text{old}}(m|x_n)\,\|x_n - \tau(y_m, W)\|^2 \\ & - \frac{1}{2}N_P\log(\frac{\sigma^2\omega}{1-\omega}) - N\log(\omega), \end{aligned} \tag{15}$$

where $N_P = \sum_{n=1}^{N}\sum_{m=1}^{M} P^{\text{old}}(m|x_n)$ and $\tau(y_m, W)$ denotes the transformed location of y_m.

The non-rigid transformation is defined as:

$$Z = Y + GW \tag{16}$$

in which G is the matrix generated by Gaussian radial basis function (GRBF) and W contains the transformation parameters to be learned.

$$G[i,j] = \exp\left(-\frac{\|x_j - y_i\|^2}{2\beta^2}\right) \tag{17}$$

Added a regularization term based on Motion Coherence Theory (MCT) [57], we obtain

$$Q_r = Q + \frac{\lambda}{2}\text{tr}(W^TGW), \tag{18}$$

where $tr(\cdot)$ represents trace operation.

EM algorithm iteratively computes the expectation and the minimizing gradients until convergence.

E-step: computing the posterior probability matrix P_O with parameters from the last iteration.

$$P_O[m,n] = P^{\text{old}}(m|x_n) = \frac{P_R[m,n]g_m(x_n)}{p(x_n)} \tag{19}$$

M-step: solving the derivatives and updating parameters.

$$W := (G + \lambda\sigma^2P_d^{-1})^{-1}\cdot(P_d^{-1}PX - Y) \tag{20}$$

$$\sigma^2 := \frac{1}{2N_P}(\text{tr}(X^TP_dX) - 2\text{tr}(X^TP^TZ) + \text{tr}(Z^TP_dZ)) \tag{21}$$

$$\omega := 1 - \frac{N_P}{N} \tag{22}$$

$P_d = diag(P\mathbf{1})$. $\mathbf{1}$ is a column vector of filled with 1s.

E. IMPLEMENTATION DETAILS

- **Parameter Setting**
  In the feature prematching stage, threshold θ_0 is automatically determined by selecting the most reliable 128 pairs of feature points. Similarly, $\hat{\theta}$ is determined by selecting the most reliable 64 pairs of feature points. In the inlier selection stage, the step-length δ is found by $\delta = (\hat{\theta} - \theta_0)/10$; confidence parameter ϵ is set 0.5; shape context uses 5 bins on the radial direction and 12 bins on the tangential direction. In the point set registration stage, the annealing constant α is set 0.95; Gaussian radial basis variance β is set 2.
- **Initialization**
  Input images are resized to 224 × 224 before feature extraction. Outlier balancing weight ω is initialized as 0.5. λ is initialized as 2. Transformation coefficient W is initialized to a matrix of all zeros. GMM variance σ^2 is initialized using:

$$\sigma^2 \leftarrow \frac{1}{2MN} \sum_{m=1,n=1}^{M,N} \|x_n - y_m\|^2 \tag{23}$$

- **Computational Cost**
  Feature computation on single 224 × 224 image takes 13.45B FLOPs. On a 2.9GHz dual core Intel i5 CPU this costs 1.2s. On solving matrix P_O, we obtain the worst-cost time $O(N^3)$. The weight matrix W has $N \times N$ entries, each of which requires N iterations to compute, hence, the complexity is $O(N^3)$. Overall, point set registration has $O(N^3)$ complexity.

F. PSEUDOCODE

We summarize our method for multi-temporal remote sensing image registration in Algorithm 1.

IV. EXPERIMENT

The performance of our work is tested on a multi-temporal satellite image dataset and a multi-temporal UAV image dataset. We compare our feature descriptor with SIFT. Our image registration method is tested against four SIFT-based state-of-the-art methods: CPD [33], GLMDTPS [25], GL-CATE [17] and PRGLS [13].

A. EXPERIMENT SETTING

Two types of experiments are conducted.

1) FEATURE PREMATCHING PRECISION TEST

Feature prematching is an important intermediate stage of image registration, we compare our convolutional feature with SIFT. In each pair of test images, we extract and prematch feature points using both methods. We then use the most reliable 95-105 pairs of matches and measure precision by $Precision = \frac{TP}{TP+FP}$. Pairs of feature matches are selected by controlling the threshold.

Algorithm 1 Image Registration Using Deep Convolutional Features and Dynamic Inlier Selection (DeepIRDI)

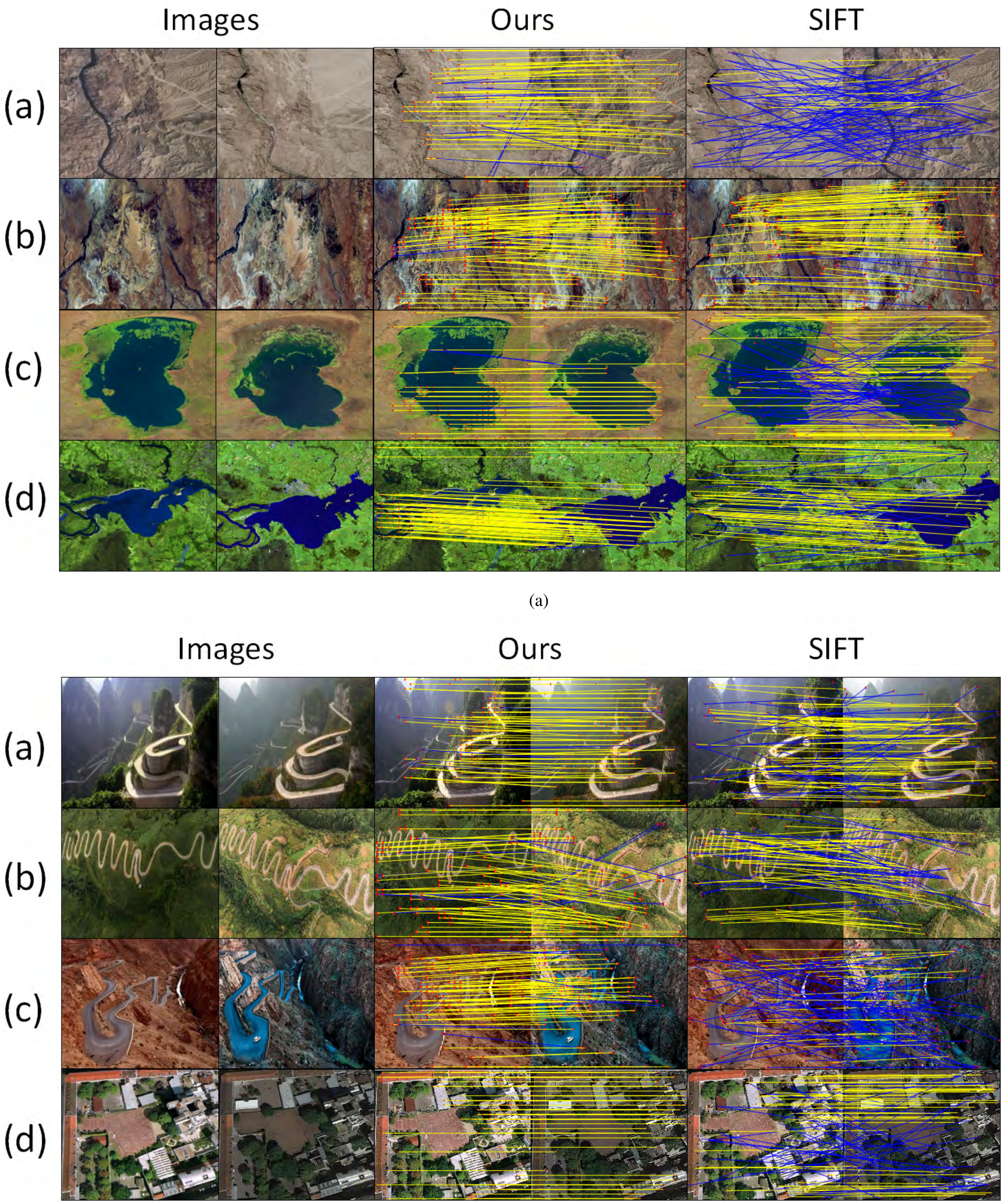
```
input   : $I_X$ and $I_Y$
Initialize parameters $\theta_0, \hat{\theta}, \delta, k, \beta, \epsilon, \omega, \sigma^2, W$ and $\lambda$;
Prematch and select the convolutional feature point
  sets $X$ and $Y$ from $I_X$ and $I_Y$ using threshold $\theta_0$;
Construct the Gaussian kernel $G$ using Equation 17;
Initialize $\theta = \hat{\theta}$;
do
    For every $k$ iterations:
        Compute convolutional feature cost matrix
          $C_\theta^{conv}$ according to Equation 6.
        Compute the geometric structure cost matrix
          $C^{geo}$ using Equation 7;
        Compute the cost matrix $C = C_\theta^{conv} \odot C^{geo}$;
        Employ Jonker-Volgenant [56] algorithm to
          solve the linear assignment on cost matrix $C$.
        Compute the posterior probability matrix $P_R$
          using Equation 9;
        Update the threshold $\theta \leftarrow \theta - \delta$;
    end
    E-Step:
        Compute posterior probability matrix $P_O$ by
          $P_O[m,n] = \frac{P_R[m,n] g_m(x_n)}{p(x_n)}$
    end
    M-Step:
        Update $W$ using Equation 20;
        Compute $Z$ using Equation 16;
        Update $\sigma^2$ and $\omega$ using Equation 21 and
          Equation 22;
    end
while Equation 15 is not convergent;
Compute the transformed image $I_Z$ using thin plate
  spline interpolation.
output  : $I_Z$
```

2) IMAGE REGISTRATION ACCURACY TEST

This type of experiment is conducted using registered images generated by different methods. In each pair of sensed and registered images, 15 pairs of designated landmark points are identified by the tester. The tester records the locations of the landmark points on the image and measure error according to the distance between each pair of landmark points. The error metrics are root mean squared distance (RMSD), mean absolute distance (MAD), median of distance (MED) and the standard deviation of distance (STD).

3) DATASETS

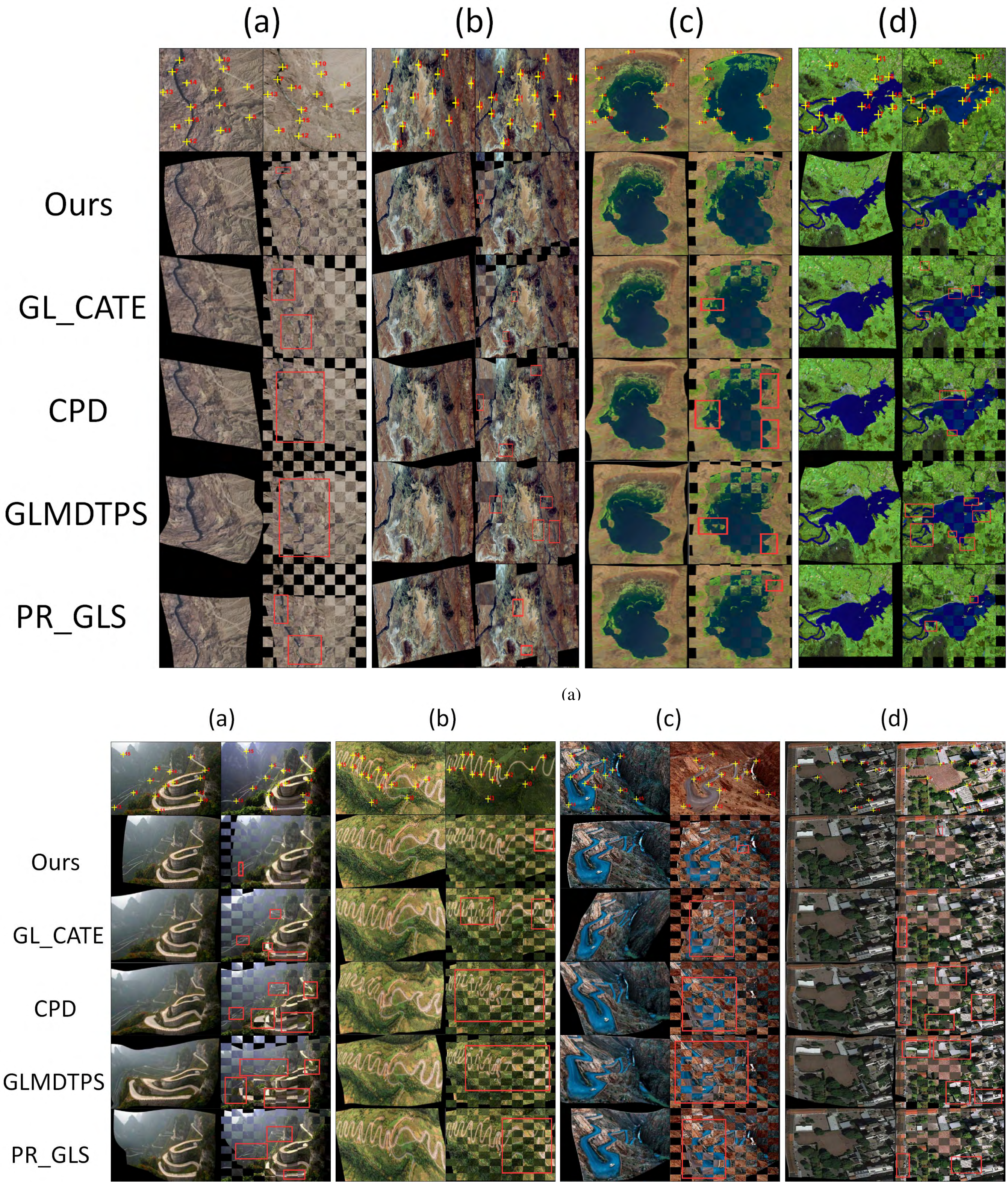
Both types of aforementioned experiments are performed on two datasets: (i) a multi-temporal satellite image dataset acquired from Google Earth; (ii) a multi-temporal UAV image dataset captured using a small UAV (DJI Phantom 4 Pro) with



(a)

(b)

FIGURE 5. Feature prematching precision test results. Correct matches (true positives) are denoted by yellow lines; wrong matches (false positives) are denoted by blue lines. (a) Feature prematching precision test results on the multi-temporal satellite image dataset. (b) Feature prematching precision test results on the multi-temporal UAV image dataset.



(a)

(b)

FIGURE 6. Image registration accuracy test results. The first row shows the locations of manually selected landmarks. Red frames mark the registration errors. (a) Image registration accuracy test results on multi-temporal satellite image dataset. (b)Image registration accuracy test results on multi-temporal satellite image dataset.

a CMOS camera; Each dataset include 15 pairs of images. The size of the images range from 600 × 400 to 1566 × 874. Image pairs in our datasets contain significant appearance variation and minor dislocation, rotation or viewpoint change.

B. RESULTS OF FEATURE PREMATCHING PRECISION TEST

Numerical results on both datasets are demonstrated in Table.1. Four examples of feature prematching in either dataset are shown in Fig.5, demonstrating detected feature point locations, correct and incorrect matches. The results show that our method generates more correct correspondences than SIFT. Moreover, the feature points detected by our method distributes more evenly over the images, because our method guarantee that only one feature point is exists in every 8 × 8 region.

TABLE 1. Feature prematching precision test result. Unit: %.

dataset	index	Ours	SIFT
Satellite	avg.	95.65	71.71
	min.	88.24	30.21
	max.	100.00	93.48
	std.	3.01	19.03
UAV	avg.	93.37	42.94
	min.	86.27	12.90
	max.	100.00	78.50
	std.	4.166	19.12

C. RESULTS OF IMAGE REGISTRATION ACCURACY TEST

Numerical results on the satellite image dataset are demonstrated in Table.2, the results on the UAV image dataset are demonstrated in Table.3. We demonstrate four examples of registered images generated by different methods in both datasets in Fig.6. The results show that our method produces the best performance in most scenarios, especially when the appearance difference in the image pair is challenging. The reason is that our convolutional feature is more robust to appearance variation than SIFT, which all of the compared methods rely on. Moreover, GLMDTPS performs unsatisfying because this method emphasizes one-to-one correspondence which is vulnerable under the presence of outliers. CPD alleviates this problem by modeling outliers using a uniform distribution. PRGLS performs well, only it suffers from dubious correspondences resulted from similar geometrical neighborhood structures. GL-CATE outruns the other three methods and performs best on the satellite image dataset. Its drawback originates from the extracted feature points that are not sensitive enough to multi-temporal images. The decent accuracy of our method proves that our dynamic inlier selection strategy properly utilizes our patch based feature.

TABLE 2. Image registration accuracy test result on multi-temporal satellite image dataset. Unit: pixels.

Method	RMSD	MAD	MED	STD
Ours	9.88	9.95	7.43	5.23
CPD	10.77	13.58	3.63	6.35
GLMDTPS	22.97	28.12	7.60	10.38
GL-CATE	9.36	11.59	2.95	7.17
PR-GLS	11.82	14.66	3.43	4.92

TABLE 3. Image registration accuracy test result on multi-temporal UAV image dataset. Unit: pixels.

Method	RMSD	MAD	MED	STD
Ours	12.63	13.20	6.78	5.73
CPD	24.01	30.12	7.51	14.56
GLMDTPS	27.40	35.28	8.26	15.26
GL-CATE	15.22	19.23	6.69	13.77
PR-GLS	23.52	29.61	7.75	11.67

V. CONCLUSION

We propose a feature based image registration method with two key contributions: (i) We build a convolutional neural network based feature extraction method using pretrained VGG network. Aiming at the effective utilization of convolutional neural networks in image registration, our feature descriptor utilizes high level convolutional information while preserving some localization capabilities. (ii) We propose a feature point registration procedure that uses a gradually expanding selection of inliers, so that the rough transformation is rapidly determined by the most reliable feature points at the early stage of registration. Afterwards, the registration details are optimized by increasing the number of feature points while restricting the mismatches simultaneously. Performed upon two multi-temporal datasets, the feature prematching test shows considerable accuracy improvement compared to SIFT, the image registration test shows that our method outperform four state-of-the-art methods under most circumstances.

ACKNOWLEDGEMENT

The authors wish to thank David G. Lowe, Herbert Bay, Andriy Myronenko and Jiayi Ma for providing their implementation source codes. This greatly facilitated the comparison experiments.

REFERENCES

[1] L. G. Brown, "A survey of image registration techniques," *ACM Comput. Surv.*, vol. 24, no. 4, pp. 325–376, Dec. 1992.

[2] K. Yang, A. Pan, Y. Yang, S. Zhang, S. H. Ong, and H. Tang, "Remote sensing image registration using multiple image features," *Remote Sens.*, vol. 9, no. 6, p. 581, 2017.

[3] Z. Wei *et al.*, "A small UAV based multi-temporal image registration for dynamic agricultural terrace monitoring," *Remote Sens.*, vol. 9, no. 9, p. 904, 2017.

[4] 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.*, to be published.

[5] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.

[6] B. Zitová and J. Flusser, "Image registration methods: A survey," *Image Vis. Comput.*, vol. 21, pp. 977–1000, Oct. 2003.

[7] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1150–1157.

[8] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun./Jul. 2004, pp. 506–513.

[9] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, and F. Tupin, "SAR-SIFT: A SIFT-like algorithm for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 453–466, Jan. 2015.

[10] A. Sedaghat and H. Ebadi, "Remote sensing image matching based on adaptive binning SIFT descriptor," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5283–5293, Oct. 2015.

[11] X. Li, L. Zheng, and Z. Hu, "Sift based automatic registration of remotely-sensed imagery," *J. Remote Sens.*, vol. 10, pp. 885–892, Jan. 2006.

[12] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[13] J. Ma, J. Zhao, and A. L. Yuille, "Non-rigid point set registration by preserving global and local structures," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 53–64, Jan. 2016.

[14] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, and J. Tian, "Robust feature matching for remote sensing image registration via locally linear transforming," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6469–6481, Dec. 2015.

[15] G. Wang, Q. Zhou, and Y. Chen, "Robust non-rigid point set registration using spatially constrained Gaussian fields," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1759–1769, Apr. 2017.

[16] G. Wang, Z. Wang, Y. Chen, and W. Zhao, "A robust non-rigid point set registration method based on asymmetric Gaussian representation," *Comput. Vis. Image Understand.*, vol. 141, pp. 67–80, Dec. 2015.

[17] 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.*, Oct. 2017, pp. 2688–2696.

[18] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, pp. 381–395, Jun. 1981.

[19] 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.

[20] 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.

[21] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J.-M. Frahm, "USAC: A universal framework for random sample consensus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 2022–2038, Aug. 2013.

[22] L. Moisan, P. Moulon, and P. Monasse, "Automatic homographic registration of a pair of images, with a contrario elimination of outliers," *Image Process. Line*, vol. 2, pp. 56–73, May 2012.

[23] G. Wang, Z. Wang, Y. Chen, Q. Zhou, and W. Zhao, "Context-aware Gaussian fields for non-rigid point set registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5811–5819.

[24] J. Ma, J. Jiang, C. Liu, and Y. Li, "Feature guided Gaussian mixture model with semi-supervised EM and local geometric constraint for retinal image registration," *Inf. Sci.*, vol. 417, pp. 128–142, Nov. 2017.

[25] 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.

[26] S. Zhang, K. Yang, Y. Yang, and Y. Luo, "Nonrigid image registration for low-altitude SUAV images with large viewpoint changes," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 592–596, Apr. 2018.

[27] M. Zhao, B. An, Y. Wu, and C. Lin, "Bi-SOGC: A graph matching approach based on bilateral KNN spatial orders around geometric centers for remote sensing image registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 6, pp. 1429–1433, Nov. 2013.

[28] X. Tong, D. Liang, and Y. Jin, "A linear road object matching method for conflation based on optimization and logistic regression," *Int. J. Geogr. Inf. Sci.*, vol. 28, no. 4, pp. 824–846, 2014.

[29] T. Wu *et al.*, "A WTLS-based method for remote sensing imagery registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 102–116, Jan. 2015.

[30] M. Zhao, B. An, Y. Wu, B. Chen, and S. Sun, "A robust delaunay triangulation matching for multispectral/multidate remote sensing image registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 711–715, Apr. 2015.

[31] M. Zhao, B. An, Y. Wu, H. Van Luong, and A. Kaup, "RFVTM: A recovery and filtering vertex trichotomy matching for remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 375–391, Jan. 2017.

[32] Y. Yeand and J. Shan, "A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences," *ISPRS J. Photogramm. Remote Sens.*, vol. 90, pp. 83–95, Apr. 2014.

[33] 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.

[34] 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.

[35] J. Ma, J. Zhao, J. Tian, X. Bai, and Z. Tu, "Regularized vector field learning with sparse approximation for mismatch removal," *Pattern Recognit.*, vol. 46, no. 12, pp. 3519–3532, 2013.

[36] S. Zhang, K. Yang, Y. Yang, Y. Luo, and Z. Wei, "Non-rigid point set registration using dual-feature finite mixture model and global-local structural preservation," *Pattern Recognit.*, vol. 80, pp. 183–195, Aug. 2018.

[37] G. Wang, Z. Wang, Y. Chen, X. Liu, Y. Ren, and L. Peng, "Learning coherent vector fields for robust point matching under manifold regularization," *Neurocomputing*, vol. 216, pp. 393–401, Dec. 2016.

[38] A. Myronenko and X. Song, "Intensity-based image registration by minimizing residual complexity," *IEEE Trans. Med. Imag.*, vol. 29, no. 11, pp. 1882–1891, Nov. 2010.

[39] V. Golyanik, S. A. Ali, and D. Stricker, "Gravitational approach for point set registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5802–5810.

[40] J. Vongkulbhisal, F. De la Torre, and J. P. Costeira, "Discriminative optimization: Theory and applications to point cloud registration," in *Proc. IEEE CVPR*, Jul. 2017, pp. 3975–3983.

[41] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[42] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

[43] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1793–1802, Mar. 2016.

[44] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5553–5563, Sep. 2016.

[45] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "MatchNet: Unifying feature and metric learning for patch-based matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3279–3286.

[46] B. Du *et al.*, "Exploring representativeness and informativeness for active learning," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 14–26, Jan. 2017.

[47] B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, and D. Tao, "Stacked convolutional denoising auto-encoders for feature representation," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1017–1027, Apr. 2017.

[48] J. Žbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1592–1599.

[49] A. Kanazawa, D. W. Jacobs, and M. Chandraker, "WarpNet: Weakly supervised matching for single-view reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3253–3261.

[50] I. Rocco, R. Arandjelović, and J. Sivic. (2017). "Convolutional neural network architecture for geometric matching." [Online]. Available: https://arxiv.org/abs/1703.05593

[51] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, p. 91–99.

[52] C. Ledig *et al.* (2016). "Photo-realistic single image super-resolution using a generative adversarial network." [Online]. Available: https://arxiv.org/abs/1609.04802

[53] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. (2015). "Understanding neural networks through deep visualization." [Online]. Available: https://arxiv.org/abs/1506.06579

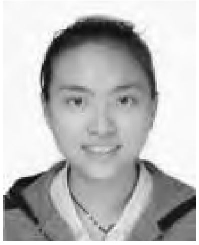
[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[55] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

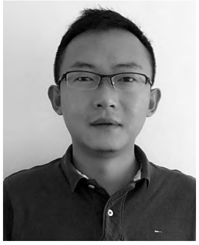
[56] R. Jonker and A. Volgenant, “A shortest augmenting path algorithm for dense and sparse linear assignment problems,” *Computing*, vol. 38, no. 4, pp. 325–340, Nov. 1987.

[57] A. L. Yuille and N. M. Grzywacz, “A mathematical analysis of the motion coherence theory,” *Int. J. Comput. Vis.*, vol. 3, no. 2, pp. 155–175, 1989.

ZHUOQIAN YANG is currently pursuing the B.S. degree with the College of Software, Beihang University. His research interests include computer vision and image registration.



TINGTING DAN is currently pursuing the M.S. degree with the School of Information Science and Technology, Yunnan Normal University. Her current research interests include image registration, point set registration, and change detection.



YANG YANG received the master’s degree from Waseda University, Japan, in 2007, and the Ph.D. degree from the National University of Singapore, Singapore, in 2013. He is currently an Associate Professor with the School of Information Science and Technology, Yunnan Normal University. His research interests include computer vision, remote sensing, geography information system, and medical imaging.