Remote Sensing Image Registration Using Convolutional Neural Network Features

Famao Ye, Yanfei Su, Hui Xiao, Xuqing Zhao, and Weidong Min

***Abstract*— Successful remote sensing image registration is an important step for many remote sensing applications. The scale-invariant feature transform (SIFT) is a well-known method for remote sensing image registration, with many variants of SIFT proposed. However, it only uses local low-level information, and loses much middle- or high-level information to register. Image features extracted by a convolutional neural network (CNN) have achieved the state-of-the-art performance for image classification and retrieval problems, and can provide much middle- and high-level information for remote sensing image registration. Hence, in this letter, we investigate how to calculate the CNN feature, and study the way to fuse SIFT and CNN features for remote sensing image registration. The experimental results demonstrate that the proposed method yields a better registration performance in terms of both the aligning accuracy and the number of correct correspondences.**

***Index Terms*— Convolutional neural network (CNN), remote sensing image registration, scale-invariant feature transform (SIFT).**

I. Introduction

FOR many remote sensing tasks such as image fusion or change detection, image registration is an indispensable part. In general, remote sensing image registration aims to estimate an optimal geometric transformation between two images of the same scene captured at different times, different viewpoints, and with different sensors.

Generally speaking, remote sensing image registration methods can be coarsely classified into two categories: intensity-based methods and feature-based methods [1]. The intensity-based methods find the geometric transformation by calculating the maximum similarity of pixel intensities between the sensed image and the reference image with a similarity metric. The most commonly used similarity measures are cross correlation and mutual information [2]. But the intensity-based methods suffer from the flatness of the similarity measure maxima, monotonous textures, and high computational complexity [1]. The feature-based methods attempt to estimate the geometric transformation between images through identifying matching features, where features must be salient and stable. Suitable features for this purpose may be point, edge, contour, region, and so on. Features from the sensed image and the reference image with the most similar invariant descriptions are paired as the matching ones. A number of feature descriptions have been developed, for example, the image intensity values in their close neighborhood of the feature, the moment-based descriptor, and the scale-invariant feature transform (SIFT) descriptor. Among these feature descriptors, SIFT descriptor is one of the popular low-level features [3], [4]. Speeded up robust features (SURFs), a variant of SIFT, are faster and more robust than SIFT [5]. Wu *et al.* [6] have proposed FSC-SIFT, in which a fast sample consensus algorithm was used to find the initial correct result, and an iterative selection of correct matches' algorithm was used to increase the correct matches. Dellinger *et al.* [7] presented an SAR-SIFT algorithm to register the SAR images. Paul and Pati [8] presented the modified uniform R-SIFT, which effectively generates enough robust, reliable, and uniformly distributed aligned key points. Ma *et al.* [9] proposed the position scale orientation (PSO)-SIFT by using a new gradient definition and a feature matching method by combining the position, scale, and orientation of each key point.

With the development of deep learning methodology, a convolutional neural network (CNN) [10] has recently enjoyed a great success in various fields such as image retrieval [11], image classification [10], [12], object recognition [13], and so on. Current state-of-the-art performances obtained by the CNN in these fields show the powerful feature representation and discriminative ability of CNN for various vision tasks. From the CNN trained based on a large-scale data set ImageNet, the middle- and high-level features can be extracted which are of better performance than the low-level features. Moreover, CNN features are also transferable among different domains [14], [15]. However, whether the CNN features can be used for remote sensing image registration still remains an open question. Therefore, in this letter, we aim to find out how to use CNN features to register remote sensing image. First, we study the way to extract the CNN feature and then generate the combined feature of CNN and SIFT. Because PSO-SIFT has achieved the state-of-the-art performance in remote sensing image registration, finally, the combined feature of CNN and SIFT is integrated into PSO-SIFT algorithm to register.

The remainder of this letter is organized as follows. In Section II, the PSO-SIFT algorithm is briefly described. In Section III, our image registration methods based on CNN features are described. Experimental results and analysis are presented in Section IV. We conclude this letter and discuss the future work in Section V.

II. PSO-SIFT

Here, we briefly review the PSO-SIFT [9] method to detect and match the SIFT features in remote sensing images. The algorithm consists of three main modules: feature detection, description, and matching.

Manuscript received July 15, 2017; revised October 19, 2017 and November 18, 2017; accepted November 20, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 41261091, Grant 61762061, and Grant 61662044 and in part by the Natural Science Foundation of Jiangxi Province, China, under Grant 20161ACB20004. *(Corresponding author: Weidong Min.)*

The authors are with the School of Information Engineering, Nanchang University, Nanchang 330031, China (e-mail: minweidong@ncu.edu.cn).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2017.2781741

A. Feature Detection

A difference of Gaussian scale space is first constructed by convolving the image with a Gaussian filter. Second, candidate features (or key points) are detected by searching the extrema of the responses in each scale layer in the scale space. Candidates with a low response or those located on edges are filtered by a threshold on the response.

B. Feature Description

A feature descriptor is characterized by a histogram computed from gradient magnitudes and orientations sampled from multiple regions around the feature location. A new gradient definition for each pixel in Gaussian scale space is proposed, and the gradient orientation and gradient magnitude of Gaussian scale-space image are computed by means of Sobel filters as

$$R_\sigma = \arctan\left(\frac{G_{y,\sigma}}{G_{x,\sigma}}\right), \quad G_\sigma^2 = \sqrt{(G_{x,\sigma})^2 + (G_{y,\sigma})^2} \quad (1)$$

where σ is the scale of Gaussian scale space and $G_{x,\sigma}$ and $G_{y,\sigma}$ are the horizontal and vertical derivatives of the gradient magnitude image with scale σ. A feature descriptor with the 136-D feature vector is created by a GLOH-like circular neighborhood (radius of 12σ) and log-polar sectors (17 location bins).

C. Feature Matching

After the feature candidates have been selected, the candidates from the sensed image have to be matched with the candidates from the reference image. The matching algorithm is given in the following.

1) *Initial Matching:* The feature candidates are matched by the ratio of the Euclidean distance between the nearest neighbor and the second nearest neighbor. The threshold of the ratio is represented by d_{ratio}. The FSC algorithm [5] is used to calculate the initial transformation parameter.
2) *Rematching:* The PSO-SIFT takes a similarity transformation model with three parameters: translation, scale, and rotation, which is widely used in the remote sensing images registration. It matches the feature candidates again by a more robust joint distance, PSO Euclidean distance (PSOED). PSOED is defined as

$$\text{PSOED}(i) = (1 + e_p(i))(1+e_s(i))(1+e_o(i))\text{ED}(i) \quad (2)$$

where $e_p(i)$ denotes the position transformation error of the i pair of feature candidates, $e_s(i)$ denotes the scale error of the i pair of feature candidates, $e_o(i)$ denotes the relative main orientation error of the i pair of feature candidates, and $\text{ED}(i)$ is the Euclidean distance of the feature descriptors corresponding to the i pair of feature candidates. The threshold of the ratio between the distance of the nearest neighbor and that of the second nearest neighbor is denoted as d_r.
3) *Outlier Removal:* After twice matching, there are some outliers. Hence, the method proposed in MS-SIFT [16] is used to filter out most of the outliers.

III. Methodology

A. Convolutional Neural Networks

The CNN is a category of deep learning architecture, which is a multilevel feedforward artificial neural network. The CNN is widely used in speech recognition, image recognition, video analysis, and natural language processing. A typical CNN usually consists of many layers, alternating between convolution and pooling, and one or more fully connected layers as in a regular neural network. The convolution layer convolves the output feature maps by producing the feature maps of the previous layer with a bank of filters. The pooling layer takes small rectangular blocks from the convolution layer's output and subsamples it to produce a single output from that block. There are several types of pooling such as max pooling and average pooling. In max pooling, the maximum value from the values observable in the window is sent to the next layer, while average pooling takes the average of all values in the window. The fully connected layer has full connections to all neurons in the previous layer, while each connection has its own individually weight.

The CNN model employed in this letter is the VGG16 model. The VGG16 model [17] is a simple but competitive CNN architecture, which won the second place on the ILSVRC2014 classification task. The VGG16 model consists of 13 convolutional layers, five max pooling layers, and three last fully connection (FC) layers.

B. Features From FC Layers

The strength of CNN is in learning multiple layers of feature representation, corresponding to different levels of extraction. The lower layers of CNN are mainly selective to simple low-level features, while the higher layers tend to detect more complex high-level features. The 1000-D features in the FC8 layer are the posterior probability scores corresponding to the 1000 classes from the data set ImageNet. The outputs of FC7 and FC6 layers with 4096-D features are commonly used as pretrained feature representations [18], [19] in the domain-adaptation literature for traditional machine learning approaches. Therefore, we discard the FC8 layer and consider the output of FC7 and FC6 layers as CNN-based features for remote image registration.

C. Fine-Tuning With Pretrained CNN Model

The VGG16 model pretrained on ImageNet was fine-tuned using a custom data set to adjust the trained parameters for remote sensing image registration. We use 21 remote sensing image pairs, which have been registered, and select $N = 2096$ matching "seed" patch pairs of size 64×64 pixels extracted from these image pairs. Each of these "seed" patch pairs is declared to represent a class of its own. To expand these classes, $K = 210$ random transformations are applied to each of the "seed" patch pairs. Each transformation is a composition of three random elementary transformations, which include rotation, scale, and brightness. Therefore, we build the custom labeled data set with N classes containing $2K$ sample each, and randomly split it into training and testing data sets with about an 80%/20% split.

Regarding the fine-tuning process, we initialize the weights from the pretrained VGG16 model. We use learning rate equal to 0.001, momentum 0.9, and weight decay 0.0005.

D. Feature Fusion and Similarity Measurement

For the PSO-SIFT feature description, it only used local low-level features such as gradient magnitude. Hence, we fuse the low-level SIFT feature and the high-level CNN feature in order to achieve accurate correspondences matching. For

a specific candidate key point $P(x, y)$, first, we calculate the SIFT descriptor f_s. Then, we extract an image patch with 64×64 pixels around $P(x, y)$, and feed it into the fine-tuning CNN model to get the CNN descriptor f_c. Finally, we transform the two vectors to one vector to represent the candidate key point, as formula: $F_{P(x,y)} = \{f_s, f_c\}$. As the large difference between the SIFT descriptor f_s and the CNN descriptor f_c, which is directly obtained, the normalization operation is necessary.

We use Euclidean distance to compute similarity between the candidate key point from the reference image, P^r, and that from the sensed image, P^s, denoted as

$$D(P^r, P^s) = w_1 \sqrt{\sum_{i=1}^{L_1} \left(f_i^r - f_i^s\right)^2} + w_2 \sqrt{\sum_{j=1}^{L_2} \left(f_j^r - f_j^s\right)^2} \tag{3}$$

where w_1 and w_2 are the two weight parameters, $w_1,\ w_2 \in [0,\ 1]$, and $w_1 + w_2 = 1$. L_1 and L_2 are separately the dimension of the SIFT and CNN descriptors.

E. Image Registration Procedure Using CNN Feature

The fine-tuning VGG16 model with the custom data set can be used to register remote sensing image. The proposed registration flowchart is similar to that of PSO-SIFT. The procedures of feature detection and SIFT descriptors extraction are the same as the PSO-SIFT. To calculate CNN feature, it extracts the images of size 64×64 pixels according to the feature points detected from the sensed image and the reference image, and then feeds these feature images into the VGG16 model. It modifies the step of initial matching and uses (3) to compute similarity. The rest of the procedure is the same as the procedure of PSO-SIFT.

IV. Experimental Results

Here, different types of remote sensing images pairs were used to discuss the quantitative performance and robustness of the proposed algorithm.

A. Experimental Settings

Three improved SIFT image registration algorithms, including SURF, FSC-SIFT, and PSO-SIFT, were used as comparison groups to verify the performance of the proposed registration algorithm. The following parameter values were used consistently by our proposed algorithm for all evaluations: $d_{\text{ratio}} = 0.9$, $d_r = 0.9$, $w_1 = 0.7$, and $w_2 = 0.3$. The other algorithms used the default parameters as follows: SURF with $d_{\text{ratio}} = 0.6$, FSC-SIFT with $d_{\text{ratio}} = 0.9$, and PSO-SIFT with $d_{\text{ratio}} = 0.9$, $d_r = 0.9$.

The four methods are all implemented under MATLAB R2014a from a PC with an Intel CPU ES-2630 v3 at 2.40 GHz $\times$ 8, 16 GB of physical memory and a graphics card Quadro M4000, 8 GB of RAM. The machine is run on Ubuntu 14.04.

B. Evaluation Criterion

In order to quantitatively validate the effectiveness of the proposed algorithm, we take three evaluation criterions to evaluate the performance of the proposed method as follows.

1) RMS_{all} is the root-mean-square error (RMSE) calculation taking all matched points. We manually selected N corresponding point pairs (x_i, y_i), (x_i', y_i') from the reference and sensed images to test the precision of the transformation model parameters. The RMSE is calculated as follows, where (x_i'', y_i'') denotes the transformed coordinates of (x_i', y_i'):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\left(x_i' - x_i''\right)^2 + \left(y_i' - y_i''\right)^2\right)}. \tag{4}$$

2) Root-mean-square leave one out (RMS_{LOO}) denotes the RMSE calculation of the matched points residuals, based on the leave-one-out method [20].
3) N_{red} is the number of redundant control points and equals the number of correct correspondences detected by the match method.

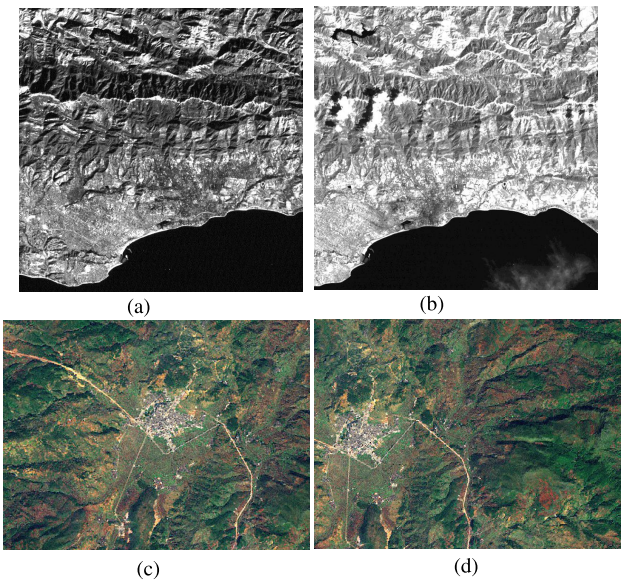

Fig. 1. Image pairs used in Experiment 1. (a) Reference image of P-A. (b) Sensed image of P-A. (c) Reference image of P-B. (d) Sensed image of P-B.

C. Experiment 1

Most of the image pairs used in fine-tuning the pre-trained VGG16 model are from Landsat Thematic Mapper, and the others are the color remote sensing images. The same type image pairs are used to evaluate the proposed method, as shown in Fig. 1. The first pair P-A consists of multiple-band images with the size of 600×600 from Landsat Thematic Mapper. The second pair P-B is the two color images of size 506×378. The results of RMS_{all}, RMS_{LOO}, N_{red}, and the average running time of five tests for two test pairs are shown in Table I. It can be observed that SURF, FSC-SIFT, and PSO-SIFT can accurately match the two image pairs, which indicates that those methods are invariant to monotonic intensity transform or a simple geometric transformation. However, the proposed methods also achieve subpixel accuracy and perform better than other methods, which has higher number of matched features N_{red} and lower values of RMS_{all} and RMS_{LOO}. Because of computing additional CNN feature, the proposed methods need more computation times than other methods.

D. Experiment 2

To verify the robustness of the proposed methods, three pairs of remote sensing images with more complicated transforms are tested, as shown in Fig. 2. The third pair P-C

TABLE I

IMAGE REGISTRATION ACCURACY ON TWO TEST IMAGE PAIRS IN EXPERIMENT 1

Method	P-A				P-B			
	RMS_{all}	RMS_{LOO}	N_{red}	Time(s)	RMS_{all}	RMS_{LOO}	N_{red}	Time(s)
SURF	0.8157	0.8555	88	0.52	0.5664	0.59903	98	0.49
FSC-SIFT	0.8323	0.8339	178	53.50	0.4184	0.4186	245	5.20
PSO-SIFT	0.6272	0.6546	239	63.72	0.4158	0.4159	259	6.87
PSO-SIFT +fc6(proposed)	0.5864	0.6109	307	309.40	0.4139	0.4142	268	39.94
PSO-SIFT +fc7(proposed)	0.6013	0.6158	279	311.54	0.4144	0.4147	269	39.70

TABLE II

IMAGE REGISTRATION ACCURACY ON THREE TEST IMAGE PAIRS IN EXPERIMENT 2

Method	P-C				P-D				P-E			
	RMS_{all}	RMS_{LOO}	N_{red}	Time(s)	RMS_{all}	RMS_{LOO}	N_{red}	Time(s)	RMS_{all}	RMS_{LOO}	N_{red}	Time(s)
SURF	0.8039	0.8427	314	0.42	*	*	*	0.31	*	*	*	0.62
FSC-SIFT	0.7601	0.7617	479	24.52	1.1523	1.1567	12	6.91	1.3064	1.3166	11	8.25
PSO-SIFT	0.7385	0.7441	514	31.38	1.0204	1.0510	17	9.87	1.2120	1.2232	17	9.93
PSO-SIFT +fc6	**0.7321**	**0.7389**	**530**	**179.66**	**0.8478**	**0.8611**	**50**	**65.50**	**0.9847**	**0.9914**	**30**	**58.44**
PSO-SIFT +fc7	0.7346	0.7376	523	179.47	1.0137	1.0416	22	63.72	1.0907	1.0935	30	59.01

*: Fails to register the image pair (RMSE>5)

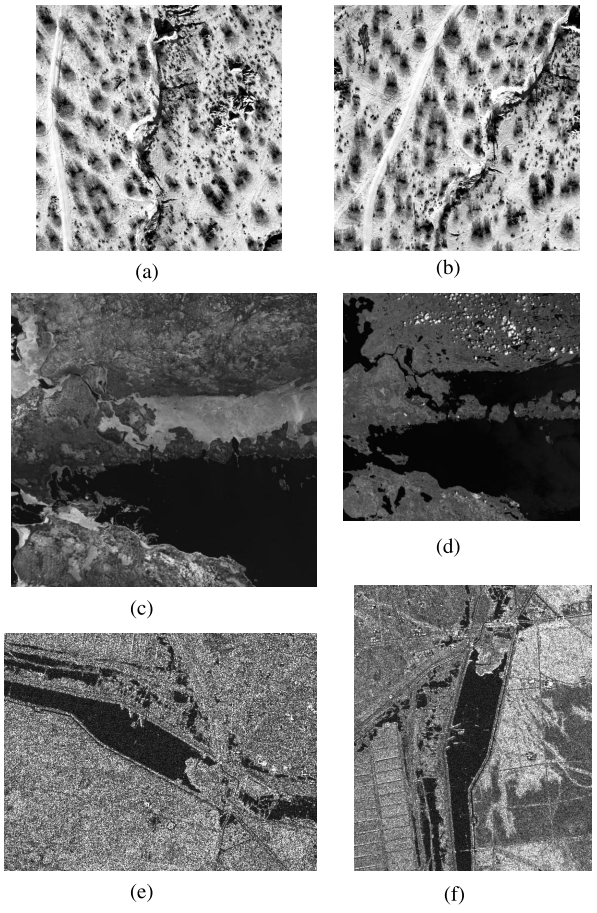

Fig. 2. Image pairs used in Experiment 2. (a) Reference image of P-C. (b) Sensed image of P-C. (c) Reference image of P-D. (d) Sensed image of P-D. (e) Reference image of P-E. (f) Sensed image of P-E.

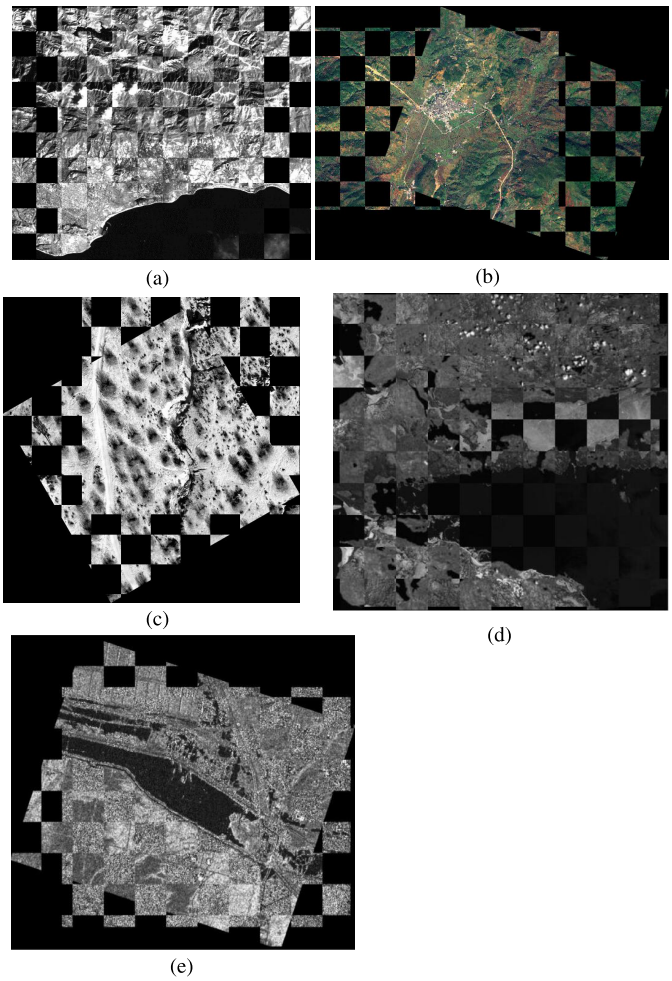

Fig. 3. Checkerboard mosaiced images of the proposed method. (a) Result for the image pair P-A. (b) Result for the image pair P-B. (c) Result for the image pair P-C. (d) Result for the image pair P-D. (e) Result for the image pair P-E.

is two 512 × 512 multiple-date images captured at different imaging time points with changed grays and large geometric transformation. The fourth pair P-D is two multiple-sensor images with size 617 × 593, which captured in two different bands at different imaging time points with different sensors, because different sensors are used to obtain these images, and the gray values are different. To increase the difficulty

of the test data set, the sensed image is transformed with a simulated scale of 0.8. The fifth P-E is two SAR images with the size of 600×500, which inevitably accompanied by speckle noise. The sensed image is with a simulated rotation of 90°. The results of four methods are shown in Table II. For test image pair P-C, all methods can handle the situation with changed grays and large geometric transformation and achieve subpixel accuracy. Due to the different spectral bands with some irregular relations of the intensity mapping in the test image pair P-D, SURF fails to register the image pair, and the proposed method finds more correct matches than the two other methods. For test image pair P-E, because of the speckle noise in SAR image, SURF fails to register the image pair again, and the proposed method performs better than the two other methods too. For the last two data sets, the proposed method with fc6 can achieve subpixel accuracy, but the other methods failed.

In addition, the checker board mosaiced images of our method for the five image pairs are shown in Fig. 3, where the image features, such as edges and regions of the images, are perfectly overlapped. From the above results, it can be observed that the proposed method can be used in the image registration not only of the same type images in training but also of other types of remote sensing image.

V. Conclusion

In this letter, we presented an automatic registration algorithm of the remote sensing images using CNN features and SIFT. We built a custom data set to fine-tune the VGG16 model pretrained on ImageNet to obtain CNN features, and constructed a combined feature with CNN feature and SIFT. Then, the combined feature is integrated into PSO-SIFT algorithm to register remote sensing image. The experimental results on multispectral, multisensors, and SAR images demonstrated that the proposed method provides better performance compared with the state-of-the-art methods in terms of both the aligning accuracy and the number of correct correspondences.

In the future, we will examine more CNN features from other models, such as AlexNet [21] and GoogleNet [22], and combine CNN features with other features [23], [24] to register remote sensing images.

References

[1] B. Zitová and J. Flusser, "Image registration methods: A survey," *Image Vis. Comput.*, vol. 21, no. 11, pp. 977–1000, Oct. 2003.

[2] M. Gong, S. Zhao, L. Jiao, D. Tian, and S. Wang, "A novel coarse-to-fine scheme for automatic image registration based on SIFT and mutual information," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 4328–4338, Jul. 2014.

[3] G. Zhu, Q. Wang, Y. Yuan, and P. Yan, "SIFT on manifold: An intrinsic description," *Neurocomputing*, vol. 113, pp. 227–233, Aug. 2013.

[4] Q. Wang, C. Zou, Y. Yuan, H. Lu, and P. Yan, "Image registration by normalized mapping," *Neurocomputing*, vol. 101, no. 4, pp. 181–189, Feb. 2013.

[5] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[6] Y. Wu, W. Ma, M. Gong, L. Su, and L. Jiao, "A novel point-matching algorithm based on fast sample consensus for image registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 1, pp. 43–47, Jan. 2015.

[7] 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.

[8] S. Paul and U. C. Pati, "Remote sensing optical image registration using modified uniform robust SIFT," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 9, pp. 1300–1304, Sep. 2016.

[9] W. Ma *et al.*, "Remote sensing image registration with modified SIFT and enhanced feature matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 1, pp. 3–7, Jan. 2017.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[11] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," *Remote Sens.*, vol. 9, no. 5, p. 489, May 2017, doi: 10.3390/rs9050489.

[12] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[14] G. S. Xie, X.-Y. Zhang, S. Yan, and C.-L. Liu, "Hybrid CNN and dictionary-based models for scene recognition and domain adaptation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1263–1274, Jun. 2017.

[15] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. NIPS*, 2014, pp. 3320–3328.

[16] B. Kufer, N. S. Netanyahu, and I. Shimshoni, "An efficient SIFT-based mode-seeking algorithm for sub-pixel registration of remotely sensed images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 2, pp. 379–383, Feb. 2015.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–29. [Online]. Available: https://arxiv.org/abs/1409.1556

[18] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. AAAI*, 2016, pp. 2058–2065.

[19] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. ICML*, 2014, pp. 647–655.

[20] H. Goncalves, J. A. Goncalves, and L. Corte-Real, "Measures for an objective evaluation of the geometric correction process quality," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 292–296, Apr. 2009.

[21] Q. Wang, J. Gao, and Y. Yuan, "A joint convolutional neural networks and context transfer for street scenes labeling," *IEEE Trans. Intell. Transp. Syst.*, to be published.

[22] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[23] Y. Yuan, J. Wan, and Q. Wang, "Congested scene classification via efficient unsupervised feature learning and density estimation," *Pattern Recognit.*, vol. 56, pp. 159–169, Aug. 2016.

[24] Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1279–1289, Jun. 2016.