Optical and SAR Image Dense Registration Using a Robust Deep Optical Flow Framework

Han Zhang, Lin Lei, Weiping Ni, Xiaoliang Yang, Tao Tang, Kenan Cheng, Deliang Xiang, *Member, IEEE*, and Gangyao Kuang, *Senior Member, IEEE*

***Abstract*—The coregistration of optical and synthetic aperture radar (SAR) imageries is the bottleneck in exploring the complementary information from the two multimodal datasets. The difficulties lie in not only the complex radiometric relationship between them, but also the distinct geometrical models of the optical and SAR imaging systems, which cause it nontrivial to explicitly depict the spatial relationship between the corresponding image regions when elevation fluctuations exist. This article aims to investigate the optical flow technique for the pixelwise dense registration of the high-resolution optical and SAR images, so as to get rid of the outlier removal and geometric mapping procedures, which have to be conducted in the classical image registration approaches that are based on sparse feature point matching. Herein, a deep optical flow framework is designed. First, a dilated feature concatenation method is proposed to enhance the discriminability of the pixelwise features for similarity measurement. An effective network training strategy is used, based on a smoothed flow loss, and also a training dataset that contains simulated elevation fluctuations. Second, we propose a self-supervised optical flow fine-tuning strategy. It incorporates the strength of the blockwise matching approach, which produces better matching precision, into the proposed pixelwise matching procedure. In this way, the accuracy of the optical-SAR dense registration is substantially improved. Extensive experiments conducted on the 1-m resolution optical-SAR image pairs of different land-cover types and distinct topographic conditions indicate that the proposed optical-SAR optical flow network -Ft framework is quite robust, and has the potential to perform the optical-SAR image dense registration in practical applications. The Python code of the proposed deep optical flow network will be made available.**

***Index Terms*—Convolutional neural networks (CNNs), dense registration, multimodal, optical flow, optical image, self-supervised finetuning, synthetic aperture radar (SAR), topographic relief.**

I. Introduction

OPTICAL remote sensing imageries reveal very detailed information of the earth surface, which is a critical resource for monitoring landscapes, natural disasters, structural changes, or even smaller objects, such as vehicles, vessels, and airplanes. However, optical imaging relies on the external circumstances including lighting conditions and weather. On the other hand, synthetic aperture radar (SAR) can image at both day and night, in almost all-weather conditions. Although SAR imageries only contain measures of polarization and intensity of backscatters, they are sensitive to structure and material of the target surface [1], which offers us more insight information. The combined use of the multimodal optical and SAR imageries is able to present more robust interpretation of image scenes or specific objects. For example, in [2], the land cover classification accuracy on the PoDelta1 Dataset [3] increases from 90.86% if using only the optical image, or 80.17% if using use only the SAR image, to 97.86% when combining the information from both the optical and the SAR images.

With the burst of the number of optical and SAR satellites on orbit, we now have at our disposal, a regular time series of both optical and SAR data. Images captured by high-resolution SAR satellites like TerraSAR-X [4] exhibit an absolute geolocalization accuracy in the order of a few decimeters. On the other hand, the geolocation error of optical imageries ranges from several tens to hundred meters due to the inaccurate measurements of the attitude angles in space [5]. In order to fully explore the complementary information from optical and SAR images, they need to be geometrically coregistered robustly and with high precision.

Image registration is the process of transforming two different imageries, the reference image and the sensed image, into one coordinate system with matched contents. As for land surface with negligible elevation variations, the affine or projective transform is sufficient to modal the geometric relationship between the reference and sensed image pair. Therefore, the majority of remote sensing image registration approaches try to first identify sparse correspondences distributed across the input imagery pair. Then, the affine [5], [6] or projective transform [7], [8], [9] can be estimated based on the geometric locations of the sparsely distributed corresponding feature points.

However, when topographic relief exists, which is quite common for the earth surface, the geometric relationship between image pairs captured by different imaging sensors or in different

Manuscript received 15 October 2022; revised 7 December 2022; accepted 27 December 2022. Date of publication 9 January 2023; date of current version 19 January 2023. *(Corresponding author: Deliang Xiang.)*

Han Zhang is with the College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China, and also with the Northwest Institute of Nuclear Technology, Xi'an 710024, China (e-mail: zhang.han@aliyun.com).

Lin Lei, Tao Tang, and Gangyao Kuang are with the College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China (e-mail: alaleilin@163.com; tangtaonudt@gmail.com; kuangyeats@hotmail.com).

Weiping Ni, Xiaoliang Yang, and Kenan Cheng are with the Northwest Institute of Nuclear Technology, Xi'an 710024, China (e-mail: niweiping@nint.ac.cn; yangxiaoliang@nint.ac.cn; chengkenan@nint.ac.cn).

Deliang Xiang is with the Beijing Advanced Innovation Center for Soft Matter Science and Engineering, Beijing University of Chemical Technology, Beijing 100029, China, and also with the Interdisciplinary Research Center for Artificial Intelligence, Beijing University of Chemical Technology, Beijing 100029, China (e-mail: xiangdeliang@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2023.3235535

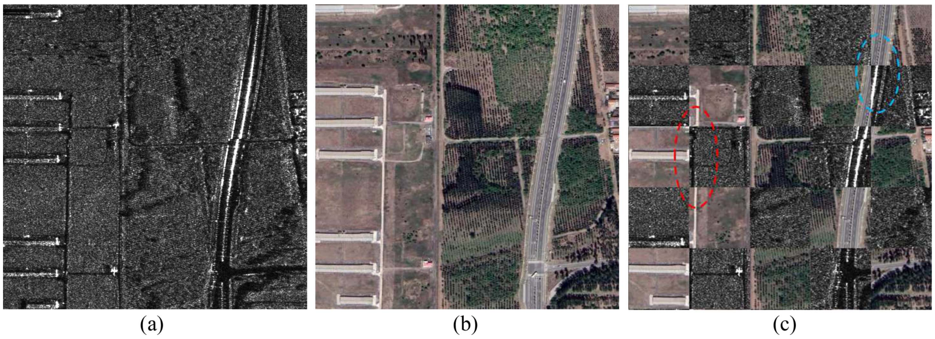


Fig. 1. Example of optical and SAR image registration result for areas with elevation variation. (a) Reference SAR image (GaoFen-3 with 1m spatial resolution, 500 × 500 pixels). (b) Sensed optical image (Google Earth, 500 × 500 pixels) that has been coregistered to (a) using the global projective transformation. (c) Mosaic image of (a) and (b), with the blue ellipse indicating the correctly matched region, while the red ellipse indicating the failure region.

viewpoints no longer obeys the affine or projective transform. Specifically, for high-resolution optical and SAR images of hilly area, the side-ways-looking acquisition of SAR sensors causes typical geometric distortion effects, termed as layover and foreshortening [10]. These effects further make the geometric relationship between optical and SAR images not able to be explicitly depicted. With the spatial resolution of remote sensing image becomes finer, even slight elevation variations would lead to non-neglectable local geometric distortions. For example, in Fig. 1, the global projective transformation successfully coregisters the optical-SAR image pair on the right part [as shown in the blue ellipse of Fig. 1(c)], while fails in the left part (as shown in the red ellipse).

A feasible solution for this issue would be to acquire much more corresponding feature points that spread evenly and densely across the whole image, and then estimate the geometric formula within each small local area. This may be realizable for optical–optical image registration. However, for optical-SAR image pairs, especially the high-resolution ones, a high ratio of mismatches is always obtained, caused by the vast modal disparity in-between. Therefore, in recent years, numerous researches have been conducted in the field of optical and SAR sparse feature point matching problem so as to increase the correctly matching rate, including the traditional handcrafted approaches [11], [12], [13], [14], [15], [16], [17], [18], [19], as well as the learning-based methods [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34].

Since the classical SIFT [35] or SURF [36] like feature point detector applied separately on optical and SAR images is not able to identify adequate candidate correspondences of high repeatability [7], most of the current feature point detecting and matching methods tend to first apply the Harris operator [37] (or other similar feature point detectors) on the optical image only. Then, for each feature point on the optical image, its correspondence on the SAR image is identified by the local searching strategy [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34]. The searching range can be narrowed down by a precoarse-registration procedure. Although with respectable amount of researches, the outlier ratio is still quite high. Besides for the local searching approach, a large proportion of mismatches would present a much smaller displacement error, compared to the SIFT-like feature point matching procedure, which seeks for the correspondence feature points from across the whole input image. The high outlier ratio and small displacement error both make it nontrivial for the proceeding mismatching removal process. Moreover, when the geometric consistency across the input image pair does not hold, the outlier removal would be even more challenging.

These issues make it necessary to explore the dense registration strategy, so as to find the displacement vector of each pixel location, which is mostly conducted by the optical flow technique [38]. Optical flow is the pattern of apparent motion of objects in a visual scene caused by the relative motion between the sensor and the scene [39]. It is a core computer vision problem and has many applications, such as moving object detection, object tracking [40], action recognition, autonomous driving, and video editing [41]. Optical flow methods have also been introduced into the remote sensing registration field to deal with the topographic relief problem in the mountainous area [9], [42], [43], [44], [45], [46], [47]. However, the researches on the remote sensing image dense registration are quite few, while much more efforts have been put in the sparse feature point matching approach, considering its time efficiency for remote sensing imageries that exhibit huge frame size.

The traditional optical flow algorithms always require brightness constancy assumption, which apparently does not hold for high-resolution optical and SAR image pairs. Fortunately, the deep-learning technique has shown great potential to learn homogenous features from heterogeneous image pairs [5], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34]. Furthermore, the deep-learning-based optical flow methods have recently outperformed the best elaborately designed traditional methods, and also been significantly faster at inference time. Here, in this article, we try to explore the potential of deep-learning-based optical flow approach to deal with the high-resolution optical and SAR image dense registration problem. As far as we know, this is the first work to introduce the deep flow network to the remote sensing image registration issue. The novelties lie in two aspects:

First, an effective pseudo-Siamese network architecture is designed as the feature extractor. By incorporating a novel dilated feature concatenation strategy, the pixelwise features collected from a very limited neighborhood become more representative for the pixelwise similarity measurement. The feature extractor is followed by a gate recurrent unit (GRU) [48] that mimics the iterative optical flow optimization procedure. For network training, a smoothed flow loss is used to impose more penalty on the image areas with larger displacement error. Besides, considering it is almost impossible to acquire the ground truth optical flow value of optical and SAR image pairs with topographic relief, we decide to simulate the elevation fluctuation using the two-dimensional (2-D) Gaussian random surface (GRS) [49].

Second, we propose to further improve the matching precision of the trained optical-SAR optical flow network (OSFlowNet) during inference time in a self-supervised way. In this procedure, a set of sparse feature point correspondences of high confidence is first obtained using the blockwise deep feature matching method proposed in the previous article [23], which is considered as the pseudo-ground truth matches, and then used for the self-supervised finetuning of the GRU part of the OSFlowNet, termed as OSFlowNet-Ft.

By properly incorporating the feature learning ability of deep CNN network with the stepwise optimizing ability of the GRU network, and also an effective network training and self-supervised finetuning strategy, the accuracy of the optical-SAR dense registration result is remarkably improved, compared with the existing dense matching approaches. Extensive experiment is conducted on the 1-m resolution optical and SAR image pairs of different land-cover types and distinct topographic conditions. The Python code of the proposed deep optical flow network will be made available at out Github page (https//github.com/zhanghan9718/).

The remainder of this article is organized as follows. Section II presents the related works. In Section III, the proposed deep optical flow based optical and SAR image registration method is depicted in detail. Comparative experiments and discussions are conducted in Section IV, and finally the conclusions are made in Section V.

II. Related Works

A. Deep-Learning-Based Sparse Feature Point Matching for Optical and SAR Image Registration

As mentioned previously, almost all the current applicable remote sensing image registration methods first try to identify sparse feature point correspondences that are evenly distributed across the input images, and then calculate a globally or locally unified geometric mapping formula for image registration. For these approaches, the essential problem is to increase the matching accuracy of the sparse correspondences.

Since the work of [5], the deep-learning-based sparse feature point matching approaches have been extensively explored [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34] for optical and SAR image registration. In most of these researches, the sparse correspondence feature points are identified by extracting the deep feature vector from the local image patch (mostly sized of 100 × 100 to 200 × 200 pixels), and then the best matching is located through local searching. The matching accuracy has been raised gradually by making efforts on the network architecture designing [5], [21], [23], loss function definition [21], [23], [31], hard negative sample mining [20], [22], [26], training data augmentation [27], and so on. Researches in [21] and [23] have demonstrated that the deep-learning technique is much more effective to extract the homogeneous features from heterogeneous optical and SAR image pairs, compared to traditionally handcrafted methods. In the previous articles [22], [23], we prove that not only the high-level semantic features, but also the low-level fine-grained features are essential for the feature point matching issue. Besides, the loss function definition is also crucial for the image matching network training. In [32], a residual denoising network is first applied on the SAR images, followed by a deep CNN feature extractor and template feature matching procedure. Similar approach is applied in [33], which also first denoise the SAR images for the subsequent optical-SAR patch matching. In order to deal with the lack of training data problem, the authors of [34] propose to transfer the deep matching models trained with annotated source domains to nonannotated target domains, so as to increase the generalization of the learned models.

However, these deep sparse feature point matching approaches are not able to solve the problem of spatially varying geometric projection problem for image areas with topographic relief. Furthermore, the use of only local appearance information will unavoidably result in a large number of false matches [50], which are not easy to be identified and removed, especially for images with irregular topography [51], [52], [53].

B. Optical Flow Methods for Pixelwise Dense Matching

In order to deal with the topographic relief problem, the ideal registration should be that, for each pixel on the sensed image, find its corresponding pixel location on the reference image. The optical flow technique is an effective tool to perform pixelwise dense image registration. The traditional optical flow methods can be divided into two categories [38]: the local method (e.g., the LK algorithm [53]), and the global method (e.g., the HS algorithm [54]), which both formulate the optical flow estimation as an optimization problem, by maximizing the similarity between the matched pixels, and also the smoothness of the flow field.

Sophisticated optical flow approaches [40], [55], [56], [57] mainly focus on the challenges of large displacement, occlusion, illumination varying, and noise. For example, the LDOF method [55] proposes a coarse-to-fine warping strategy to deal with the large displacement problem. The EpicFlow [56] conducts the dense matching by edge-preserving interpolation from a sparse set of matches, and initializes the variational energy minimization with the dense matches. The FlowFields [57] uses the approximate nearest neighbor fields to find most inliers. The SIFT-Flow [40] adopts the computational framework of the traditional HS or LK algorithm, but by matching SIFT descriptors instead of raw pixels. Although with all these refinements, the basic formulation has changed little since the LK and HS algorithms.

The deep-learning methods have outperformed the traditional handcrafted methods in most computer vision fields. However, for the first few years, the deep optical flow networks, such as the representative FlowNet [58], SpyNet [59], and PWC-Net [41], are not on par with the top handcrafted methods. Besides, the size of the network parameters is large, and has to be trained elaborately in multiple stages.

The RAFT method [60] proposed in the year of 2020 is the first deep flow architecture that noticeably outperforms the delicately designed top handcrafted optical flow methods. Besides a CNN feature encoder that extracts the feature vector for each pixel, the RAFT architecture also contains a correlation layer that produces a 4-D correlation volume, as well as a recurrent GRU-based update operator that iteratively updates the flow field based on correlation volume values. The subsequent GOCor method [61] proposes a differentiable neural network module that acts as a replacement to the feature correlation layer, but brings limited performance increase. The COTR [62] is the first that brings the transformers into the deep flow field. It proposes a functional correspondence architecture to combine the strengths of dense and sparse methods. However, the COTR network requires fixed input image size, considering that the output is produced by a fully connected layer.

C. Optical Flow Methods for Remote Sensing Image Dense Registration

The previous optical flow approaches are mostly evaluated and applied in the object tracking, action recognition, and video processing domains. Some classical approaches have also been introduced into the remote sensing image dense registration problems.

Back in 2000, the HS and LK methods have been incorporated into the image processing system of the Landsat, JERS-1, and CBERS-1 satellite imageries to conduct the image registration procedure [63]. In [64], the SIFT features are added as a constraint into the classical LK model, so that the affine coefficients are calculated as the initial value of the optical flow field. In [65], the SIFT-Flow method, which has been proved to be more robust to drastic appearance changes due to changes of seasons, or variation of imaging condition, is applied on the unmanned aerial vehicle image registration problem, which often contains nonrigid transform between images captured by different view-points. In order to deal with the land surface change problem, the authors of [47] proposed to first find a sparse set of feature correspondence, and then obtain the pixelwise offset map by using bilinear interpolation based on the sparse correspondences. Also, in [9], the abnormal optical flow results caused by land cover changes are detected and corrected by the weighted Taylor expansion of the nearby displacement values.

Specifically, for multimodal image registration issue, the GeFolki method is proposed in [42], which modifies the classical Lucas–Kanade algorithm with a rolling guidance filter, a rank filter, and a local contrast inversion strategy, so as to deal with the texture and contrast difference of multimodal images. The GeFolki has been proved to be the most appropriate approach for the registration of Sentinel-2 and Sentinel-3 multimodal imageries [44], compared to the Phase-Only Correlation [66] and a deep-learning-based approach [67]. A double-U-net architecture is designed in [68] specifically for the pixelwise alignment of the OSM building map with optical remote sensing images. Mimic to the SIFT-Flow method, the authors of [43] proposed the OS-Flow dense registration method. It extracts the pixelwise optical and SAR image features using the optical-gradient location and orientation histogram (GLOH) descriptor and the SAR–GLOH descriptor, respectively, and therefore narrows down the radiometric and geometric gap between image pairs of the two different modal types.

However, when applied on the high-resolution optical and SAR image pairs, the previous dense registration approaches either tend to produce prominent local distortion results or even totally fail the registration. The major reason is that it is very challenging to depict the similarity between each optical and SAR corresponding pixel pair using local features drawn from only the nearby neighborhood. For example, the Gefolki method uses the raw pixel values of the local 17×17 pixels surrounding the central pixel for pixelwise similarity measurement. The SIFT-Flow and the OS-Flow methods collect the pixelwise SIFT or SIFT-like features from the local 16×16 pixels. On the other hand, for the sparse feature point matching approaches, no matter the handcrafted approaches or the learning-based ones, local image block larger than 100×100 pixels is used to identify each corresponding feature point pair, which is much more discriminative.

III. Methodology

Our approach is driven by the RAFT deep optical flow architecture [60], but with a novel deep feature extraction network and a refined flow loss definition, which both make it more effective to learn pixelwise homogenous features from the close neighborhood of the heterogeneous optical and SAR images. Furthermore, we propose a self-supervised network fine-tuning procedure, which combines the advantages of both dense and sparse feature point matching techniques, and brings obvious performance increase.

A. Deep Close Neighborhood Feature Extractor

The deep optical flow network produces a pixelwise displacement map for the input image pair to be coregistered. The displacement value of each pixel location is estimated based on the local feature similarity, which is mostly measured by the correlation value between the pixelwise image features for recent deep optical flow frameworks [59], [60], [61], [62]. Therefore, the essential problem for the learning-based optical flow estimation is to let the network to learn to capture the distinguishing pixelwise image features, which is especially important for optical and SAR image registration problem, considering their vast radiometric and geometric gaps.

However, it is nontrivial to measure the similarity between a very small local patch of optical and SAR images, let alone using only the pixelwise local features. Fig. 2 shows an experiment conducted in [23], which presents the matching accuracy of the corresponding optical and SAR image patches of different sizes. We can see that the matching accuracy declines dramatically as the template size decreases. Considering that the optical flow framework is intended to find the correspondence pixel on the sensed image for each pixel on the reference image, it has to

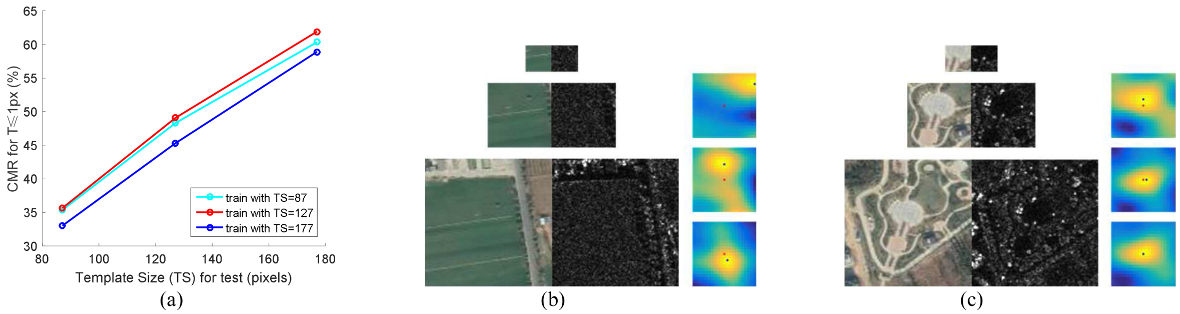


Fig. 2. Illustration of optical and SAR patch matching results of different template sizes. (a) Correctly matching rate (CMR) with different training and testing template size. This experiment has been conducted in [23]. (b) Feature similarity score maps of the first example. (c) Feature similarity score maps of the second example. For both (b) and (c), the three template image pairs from top to bottom are sized of 21 × 21 pixels, 51 × 51 pixels, and 101 × 101 pixels, respectively. In each score map image, the blue dot represents for the matching location that presents the best feature similarity, while the red dot is the ground truth matching location.

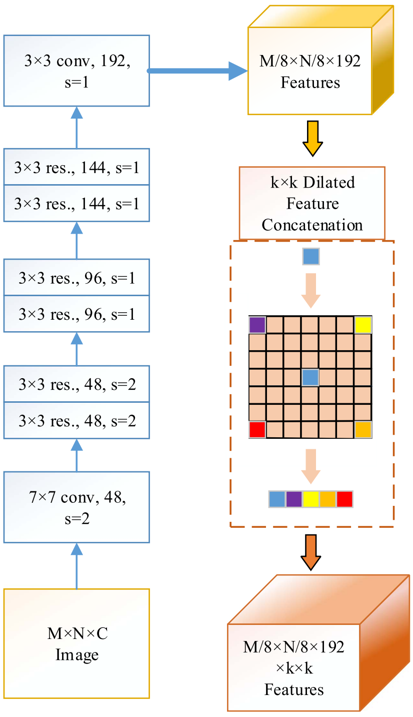


Fig. 3. Proposed deep close neighborhood feature extractor with the dilated feature concatenation strategy.

make a compromise between the feature discriminability and the time efficiency. Therefore, current optical flow methods use only the information from a very close neighborhood of the central pixel to estimate the pixelwise displacement value, which is much more challenging than using all the features collected from a local optical and SAR image patches. Herein, we carefully design a pseudo-Siamese CNN-based close neighborhood feature extractor with distinct weights for the optical and SAR branches, which enriches the pixelwise image features with a limited computation expense, so as to increase the discriminability of the pixelwise features between optical and SAR images.

The proposed deep feature extractor is shown in Fig. 3. The network consists of one 7 × 7 convolutional layer, six 3 × 3 residual blocks, and one 3 × 3 convolutional layer. Each residual block contains two 3 × 3 convolutional layers. In total, the feature extractor is composed of 14 parametric convolutional layers, which is quite light weighted. Except for the last 3 × 3 convolutional layer, all the previous ones are followed by a batch normalization layer and a rectified linear unit.

The pooling or striding operations are effective tools to enlarge the receptive field size of the pixelwise deep features. However, they would at the same time reduce the features' localization precision. Similar to the original RAFT network, we decide to produce a dense feature volume with a downsampling factor of 8. Different from the RAFT architecture, which applies the striding operation on the first, fourth, and sixth convolutional layers, we put the three stride $= 2$ operations on the first 7 × 7 convolutional layer and the first two residual layers of the feature extractor network, which are the first, second, and fourth convolutional layers. In this way, the receptive field size of the output features would increase from the original 163 × 163 pixels to 195 × 195 pixels. Furthermore, we propose to apply a dilated feature concatenation operation to mimic the patch matching approach. Instead of collecting all the pixelwise features within the local image patch, which will lead to unbearable dimension explosion, we collect the local features only in sparse locations, as shown in the orange dashed rectangle in Fig. 3. In this way, we cannot only enrich the feature diversity for pixel matching, but also further enlarge the receptive field size of the pixelwise features. Specifically, in our follow-up experiments, we use a 3 × 3 kernel with the dilation value set as 6. Therefore, the output feature channel number would be $192 \times 3 \times 3 = 1728$. Also, the receptive field size would be 207 × 207 pixels.

B. Multiscale Correlation Similarity Measurement and Supervised Network Training With Smoothed Flow Loss

The correlation cost volume has been repeatedly proven to be more discriminative for visual similarity measurement than raw images or features [59], [60], [61], [62]. Herein, we also use the cost volume between the downsampled optical and SAR pixelwise deep features for displacement field estimation, as shown in Fig. 4. For a pair of optical and SAR images $I_O, I_S \in \mathbb{R}^{H\times W}$ to be coregistered, the corresponding deep features would be $F_O, F_S \in \mathbb{R}^{K\times L\times D}$, where $K = H/8$,

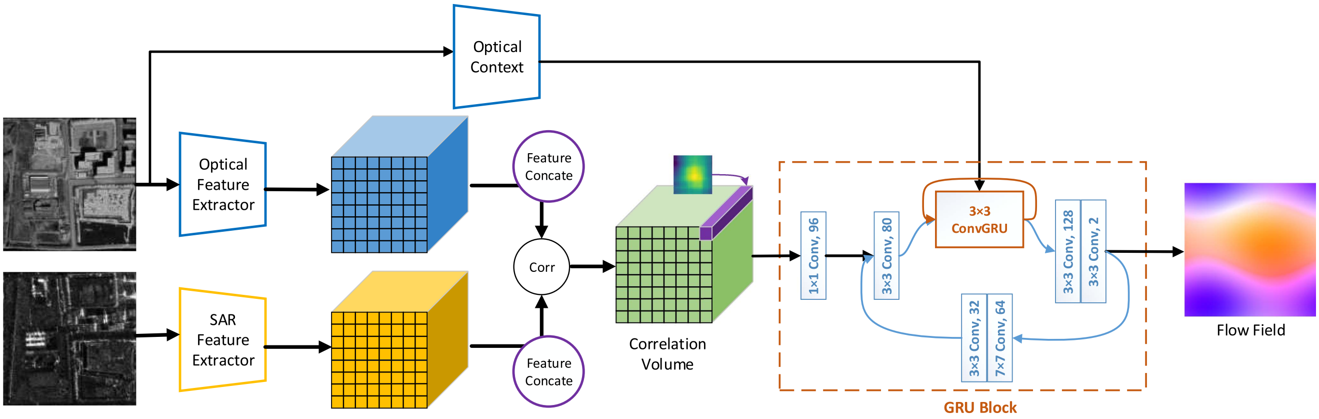


Fig. 4. Overall architecture of the proposed OSFlowNet. The optical and SAR feature extractor branches, as well as the optical context branch, are all composed of the 14 convolutional layers as shown in Fig. 3, but with unshared weights. The 4-D correlation volume is illustrated in 3-D by reshaping the last two dimensions into one. The 4-D correlation volume and the features produced by the optical context branch are fed into the GRU block, which contains not only the iterative ConvGRU, but also several head encoders in terms of convolutional layers.

$L = W/8$. For each pixel location (i,j) on F_O (or otherwise F_S), its correlation value with pixel (k,l) on F_S would be $\sum_{d=1}^{D} F_O(i,j,d) \cdot F_S(k,l,d)$. Therefore, a full correlation cost volume $C \in \mathbb{R}^{K\times L\times K\times L}$ can be obtained more efficiently and straightforwardly using the matrix multiplication operation

$$C_{ijkl} = \sum_{d=1}^{D} F_O(i,j,d) \cdot F_S(k,l,d). \tag{1}$$

Considering the difficulty to illustrate the 4-D cost volume, a 3-D sketch of C is presented in Fig. 4 by reshaping the last two dimensions, which is a 2-D correlation similarity score map into one dimension. Subsequently, the correlation cost volume is fed into a GRU block to iteratively update a flow field by emulating the first-order optimization procedure. Note that the GRU block is composed of not only the iterative GRU, but also several convolutional head encoders, as shown in the orange dashed rectangle of Fig. 4.

In order to simultaneously deal with the small and large displacement, the correlation pyramid strategy proposed in [60] is also applied. Specifically, pooling operation with kernel size 1, 2, 4, and 8 are performed on the last two dimensions of C, respectively, resulting in a four-scale correlation pyramid set $C_{\rm MS} = \{C^1, C^2, C^4, C^8\}$, where $C^s \in \mathbb{R}^{K\times L\times K/s\times L/s}$, with each pyramid level reveals the similarity measurement of different displacement range.

It is apparently unnecessary and also harmful to put the full correlation cost volume into the GRU block for optical flow estimation. Therefore, we collect the correlation information from $C_{\rm MS}$ with a predefined range of r pixels. Herein, for each pixel on the reference image, a correlation similarity feature vector sized of $4r^2$ is obtained. The similarity feature volume F_c that fed into the GRU is actually sized of $K \times L \times 4r^2$. Due to the four-scale correlation pyramid strategy, the maximum displacement value would be $r \times 2^3 \times 8$. In our subsequent experiments, we set $r = 3$, which means that the largest displacement of the estimated flow field would theoretically reaches to 192 pixels.

Besides the correlation similarity feature vector, the context information $F_O^c \in \mathbb{R}^{K\times L\times D}$ learned from the optical image is also incorporated as the input of the GRU block. The architecture of context network is identical with the feature extractor network as shown in Fig. 3, but with distinct network parameters. Therefore, the input of the GRU is composed of three parts:$C_{\rm MS}$, F_O^c , and also the estimated flow field in the previous iteration f^{t-1}, with f^0 initialized as 0 for all the pixel locations. Note that the spatial resolution of the estimated flow image is only 1/8 of the input image pair. The convex based upsampling operation proposed in [60] is applied to upsample the flow image to the same spatial resolution of the input images.

The network is trained in a supervised way, based on the latest N sequence of the estimated flow field $\{f^1, f^2, \ldots f^N\}$ with exponentially increasing weights. Different from the common practice of the supervised flow learning that directly uses the l_1 normed distance between the estimated flow and the ground truth flow $f_{\rm gt}$, we propose to use a smoothed flow loss. Assuming $d_f^t = |f^t - f_{\rm gt}|$, the smoothed flow error is defined as

$$sd_f^t = \begin{cases} \left(d_f^t\right)^2/4, & d_f^t \leq 2 \\ \left(d_f^t - 1\right)^a, & d_f^t > 2. \end{cases} \tag{2}$$

Then, the smoothed flow error would be

$$L = \sum_{t=1}^{N} \omega^{N-t} \cdot \left\| sd_f^t \right\|_1. \tag{3}$$

Here, we set a= 1.2 and ω= 0.8. The smoothed flow loss would impose more penalty on the image areas with larger displacement error, which mimics the hard negative mining process, and brings flexibility into the optimization process and makes noticeable performance improvement.

C. Self-Supervised Optical Flow Finetuning Based on Sparse Matching Results of Higher Precision

The estimation of the pixelwise optical flow field relies on two priors: 1) the similarity between corresponding feature

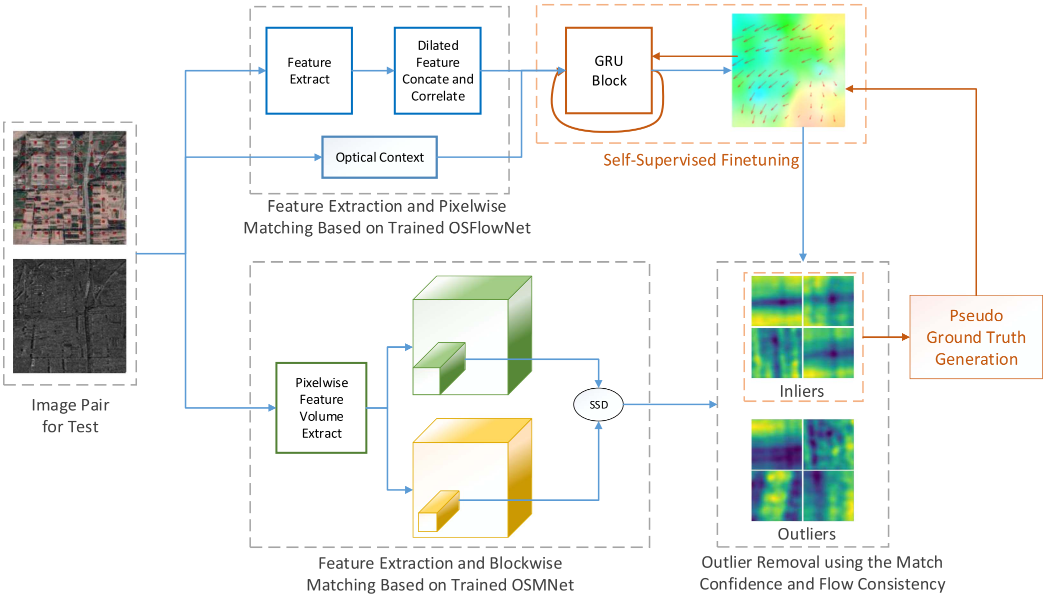

Feature Extract
Dilated Feature Concate and Correlate
Optical Context
Feature Extraction and Pixelwise Matching Based on Trained OSFlowNet
GRU Block
Self-Supervised Finetuning
Image Pair for Test
Pixelwise Feature Volume Extract
SSD
Feature Extraction and Blockwise Matching Based on Trained OSMNet
Inliers
Outliers
Outlier Removal using the Match Confidence and Flow Consistency
Pseudo Ground Truth Generation

Fig. 5. Workflow of the self-supervised optical flow fine-tuning method. The top branch presents the OSFlowNet architecture as in Fig. 4. The bottom branch shows the sparse feature points matching procedure based on the OSMNet proposed in [23]. After the inliers of the sparse matches are identified, they are considered as the pseudo-ground truth flow values that are used as the supervision of the subsequent optical flow fine-tuning step. The fine-tuning procedure is only conducted on the GRU block shown in the dashed orange rectangle, while the network parameters shown in the dashed gray rectangles are fixed.

points and 2) the smoothness of the flow field. In the common computer vision applications, the biggest obstacles for optical flow estimation include the large displacement and occlusion, which both fail to meet the smoothness prior conditions. On the contrary, for the optical-SAR image registration issue, it is the feature disparity between the input image pair of different modal types that makes the similarity between correspondences does not always hold. Herein, we propose a self-supervised flow network finetuning approach, as shown in Fig. 5, which takes advantage of sparse correspondences that are matched with higher precision, so as to make up the deficiency of the low confidence of pixelwise feature point matching.

Note that the fine-tuning procedure is conducted during the test stage, instead of the network training procedure. Specifically, for an optical-SAR test instance $I_O, I_S \in \mathbb{R}^{H\times W}$ to be coregistered, it is fed into the OSFlowNet that has been trained supervisingly, resulting in the initially estimated optical flow fields $f \in \mathbb{R}^{H\times W\times 2}$. At the same time, the four-scaled correlation cost volume C_{MS} and the downscaled optical context features $F_O^c \in \mathbb{R}^{H/8\times W/8\times D}$ are also recorded. In the proceeding flow field fine-tuning step, the optical and SAR image feature extractor networks are fixed, while only the GRU block is updated. It means that the C_{MS} and F_O^c would act as the fixed inputs of the GRU block, which is then finetuned in a self-supervised manner.

As revealed by Fig. 2, it is reasonable to assume that the sparse corresponding feature points generated by the block matching method would present higher matching accuracy and precision, compared with the pixelwise matching result. Herein, dozens or hundreds of correspondences that sparsely distributed across the input image pairs are identified, using the blockwise deep feature matching method OSMNet, which is proposed in the previous article [23]. These sparse matches are considered as the pseudo-ground truth matches, and then used for the self-supervised finetuning of the GRU block of the OSFlowNet.

As for the sparse feature point matching step, first of all, the test image pair $I_O, I_S \in \mathbb{R}^{H\times W}$ is fed into the OSMNet, and produces the pixelwise dense feature volumes $V_O, V_S \in \mathbb{R}^{H\times W\times 9}$. Second, the evenly distributed sparse interest point set $P^O = \{p_1^o, p_2^o, \dots p_K^o\}$ is collected from the optical image using the block-based Harris operator or simply using a fixed step. Third, for each interest point $p_i^o \in P^O$, its initial correspondence on the SAR image would be $p_i^s = p_i^o + f(p_i^o)$. Then, the local dense feature volumes surrounding the central points $p_i^o$and p_i^s are extracted from V^O, V^S with fixed extent size e and $e + sr$, respectively, therefore $V_p^O \in \mathbb{R}^{(2e+1)\times(2e+1)}$ and $V_p^S \in \mathbb{R}^{(2(e+sr)+1)\times(2(e+sr)+1)}$. Here, e stands for the extent size of the extracted local feature volume and sr stands for the search radius of the correspondence. Hereafter, a sum of squared differences (SSD) score map $S_p \in \mathbb{R}^{(2\times sr+1)\times(2\times sr+1)}$ between the local optical and SAR feature volumes is obtained

$$S_p(x) = \sum_i \left[V_p^O(x) - V_p^O(x-u)\,T(u)\right]^2 \tag{4}$$

where $T(u)$ represents for the sliding window on location u [22]. If the initial flow vector calculated by the OSFlowNet is precise, the minimum value of S would locate at the center of the score map. Otherwise, we consider the block-matching result produced by (4) as the pseudo-ground truth matching location, and then the flow vector would be updated by the offset vector $v = (v_x, v_y)$ of the minimum location. In this way, we are able to obtain an initial sparsely distributed correspondence point set $\{(p_i^o, q_i^s)\}_{i\in[1,K]}$, where $q_i^s = p_i^s + v$.

Although the previous blockwise matching results obtained by the OSMNet are assumed to be more reliable, there still inevitably exist some outliers that have to be identified and removed. Herein, we propose a straightforward but quite effective outlier removal method. First, the candidate correspondences with the SSD values smaller than a strict threshold T_0 are assumed to be high confidential, and therefore all considered to be inliers. As for the candidate correspondences with the SSD values larger than T_0 but smaller than a loose threshold T_1, the ones are taken as inliers if the l_1 norm of the offset vector v is smaller than a third threshold T_v, which means that there is a good consistency between the blockwise matching and pixelwise matching results. In this way, the inlier restraint is defined as

$$\text{inliers}_1 = \{\min(S_p) \le T_0\}_{p \in P^O} \tag{5}$$

$$\text{inliers}_2 = \left\{T_0 < \min(S_p) \le T_1 \text{ and } \|v_p\|_1 \le T_v\right\}_{p \in P^O}. \tag{6}$$

Then, we obtain a correspondence sparse feature point set with high matching confidence: $cp = \{(p_i^o, q_i^s), v_i\}_{i \in [1, K_c]}$, where $v_i = q_i^s - p_i^s$ and $K_c \le K$. Therefore, the pseudo-ground truth optical flow value for p_i^o would be

$$f_{p^{gt}}(p_i^o) = f(p_i^o) + v_i. \tag{7}$$

Hereafter, we conduct the self-supervised optical flow finetuning by taking the cp set as the pseudo-ground truth sparse optical flow values. As mentioned previously, only the GRU block is fine-tuned, while the parameters of the feature extractors and image context branches are fixed. The loss function is defined as the optical flow error of the sparse feature points between the calculated flow vectors and the pseudo-ground truth, termed as L_{pt}

$$L_{pt} = \frac{1}{K_c}\sum_{i=1}^{K_c} |f(p_i^o) - f_{p^{gt}}(p_i^o)|. \tag{8}$$

Besides, mimic to the traditional optical flow technique, we bring in the l_1 smoothness regularization termed as L_s

$$L_s(f) = 0.5 \cdot (df_x + df_y) \tag{9}$$

where df_x and df_y are the first-order gradient of the estimated flow field f. Then, the final loss function for the self-supervised flow fine-tuning is defined as

$$L = L_{pt} + \lambda \cdot L_s(f) \tag{10}$$

where λ is set as a number that makes the sparse optical flow error L_{pt} and the smoothness term L_s on the same order of magnitude. In the follow-up experiments, we set λ= 50.

IV. Experiments and Results

This section evaluates the performance of our proposed deep optical flow network. In Section IV-A, datasets used for network training and testing are described. The details of network configuration are also presented here. In Section IV-B, we conduct an ablation study to evaluate the effectiveness of the proposed OSFlowNet. Section IV-C evaluates the self-supervised network fine-tuning strategy. Section IV-D compares the proposed deep framework OSFlowNet-Ft with the existing optical flow methods that are also specifically designed for optical-SAR image dense registration. The comparative experiments are conducted on various scenes to test the robustness and generalization of our approach.

A. Dataset Construction and Network Configuration

The lack of training dataset is the first obstacle to apply learning to the optical flow issue. For areas with noticeable topographic relief, the accurate coregistration between the corresponding optical and SAR image pair would require not only the accurate geolocation and attitude of the satellite sensors, but also precise digital surface model with high spatial resolution, which are unavailable in most circumstances. Therefore, it is almost impossible to acquire the ground truth optical flow maps for optical and SAR image pairs of fluctuating surface.

Herein, we construct the training dataset using images of plain areas, which can be coregistered with few, but accurate and reliable sparsely distributed matching points [5], using the affine or projective transformation. We propose to simulate the sensors' geolocation and attitude errors using random affine transforms, and further imitate the elevation fluctuation by generating a smooth, and spatially varying 2-D optical flow fields using the 2-D GRS, which can be easily produced in the Fourier domain

$$H = \exp\left(-0.5 \times \left(x^2 + y^2\right)/\sigma^2\right) \tag{11}$$

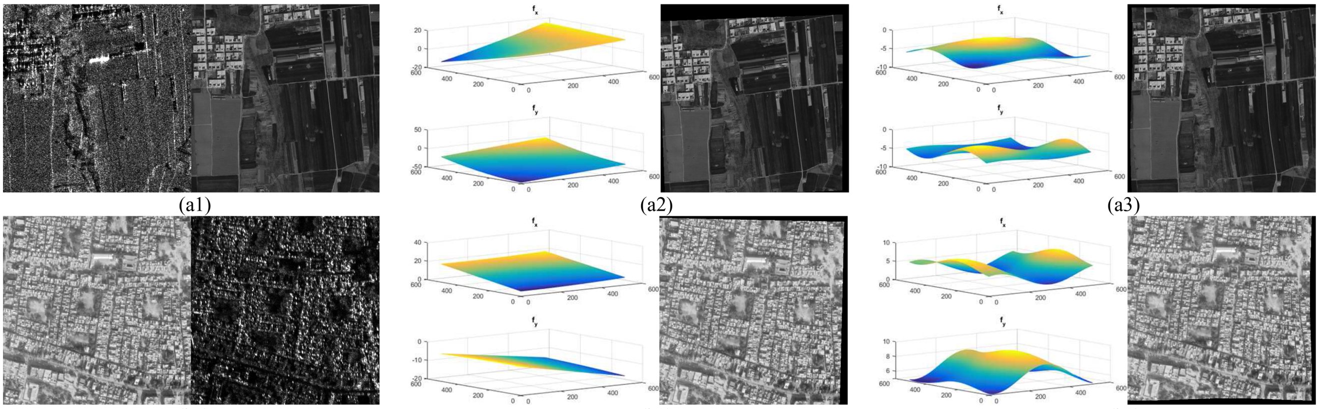
$$R_{im} = \text{randn}(M, N) \tag{12}$$

$$f_{\text{GRS}} = \alpha \cdot \text{real}\left(F^{-1}\left(H.* F\left(R_{im}\right)\right)\right) \tag{13}$$

where H is the low-pass filter in frequency domain with σ as the cut-off frequency that controls the fluctuating frequency. R_{im} is a random image sized of $M \times N$, with each pixel value sampled from the standard normal distribution. α restrains the amplitude of the generated 2-D GRS. In this article, we set $\sigma = 0.5$ and $\alpha = 0.5e^{-4}$.

We use the OSdataset [69] that is publicly available to train the proposed OSFlowNet, so as to ensure the reproducibility of our experiment results. The OSdataset is collected from 20 different scenes located at different cities around the world, with the SAR images produced by the Chinese GaoFen-3 satellite and the optical images collected from the Google Earth platform. In total, the OSdataset contains 2673 pairs of coregistered optical-SAR patch pairs sized of 512 × 512 pixels, all with 1-m spatial resolution. The dataset is divided into the training, validation, and test sets, containing 2011, 238, and 424 patch pairs, respectively.

For each instance in the training set of the OSdataset, the optical patch is warped by two different transformations separately: a random affine transformation and a GRS transformation which is simulated using the random optical flow field produced by (13). In this way, a training dataset containing 2011 × 2 instances is constructed, termed as warped OSdataset (wOSdataset), which is quite small. The validation and test sets of OSdataset are also converted in the same way for the network evaluation. Two examples are shown in Fig. 6. The left column shows two coregistered optical and SAR image pairs randomly selected from the OSdataset. The middle and the right columns present



(a1) (a2) (a3)

(b1) (b2) (b3)

Fig. 6. Two examples of the simulated training dataset. (a1) and (b1) show two input image pairs that have been precisely coregistered. (a2) and (b2) show the flow values in the *x* and *y* directions generated by random affine transforms, as well as the corresponding warped optical images. (a3) and (b3) show the flow values in the *x* and *y* directions generated by Gauss random surface, as well as the corresponding warped optical images.

the flow values of the random affine and the GRS transformation, respectively, as well as the transformed optical images.

The PyTorch framework is used to implement the OSFlowNet, which is initialized with random weights. The AdamW optimizer is adopted for network training, with the initial learning rate set as 0.0003, and the maximum training iteration set as 100 000. During training, the batch size is set as 6, and the input optical and SAR image pairs sized of 512 × 512 pixels are randomly cropped into 384 × 384. For each training iteration, the GRU block is unrolled for 12 times. Using 1 Tesla V100 GPU, the network training procedure would take about 10 h.

Since the proposed OSFlowNet is fully convolutional, the input image size for network inference can be arbitrary. In the self-supervised optical flow finetuning procedure, the sparse feature points are sampled from the reference image by a fixed step size of 128 pixels and the block size for feature matching is 193 × 193 pixels. After obtaining the sparse feature point correspondences with higher precision, 50 updates are performed on the GRU block, with each update unrolled for 6 times.

B. Ablation Study of the Deep Optical-SAR Flow Network Architecture

We perform an ablation study to evaluate the performance of the OSFlowNet for optical and SAR image dense registration. Specifically, we test the effectiveness of the proposed dilated feature concatenation strategy, and then the modified smoothed loss function. Considering that our proposed optical-SAR dense registration method is inspired from the RAFT network, the registration results of the primary RAFT network after trained with the same dataset are also presented here, as shown in Tables I and II.

Note that the network is trained using the AdamW optimizer, which gradually decays the initial learning rate by a fixed factor. The training procedure is terminated until the learning rate is decayed to 0. Therefore, neither the validation set nor the test set of the wOSdataset is seen during the training phase. Herein, we use both of them for the network evaluation and comparison. As presented previously, each optical and SAR image pair from the validation or the test set is warped by a random affine transform, and then a random flow field generated using the GRS transform. The network is evaluated, respectively, based on these two types of geometric transforms.

TABLE I
RESULTS ON OSDATASET VALIDATION SET

Methods		OSFlowNet	No dilated concate	No smoothed loss	RAFT
Affine transform	T≤1px	13.77	11.23	11.43	9.38
	T≤3px	63.68	62.00	62.40	56.63
	T≤5px	90.80	91.05	90.57	89.80
	EPE	2.71	2.80	2.83	3.08
Gaussian random surface transform	T≤1px	9.84	7.94	7.41	7.53
	T≤3px	51.80	48.77	48.87	45.02
	T≤5px	83.79	81.56	82.89	78.40
	EPE	3.22	3.42	3.39	3.69

TABLE II
RESULTS ON OSDATASET TEST SET

Methods		OSFlowNet	No dilated concate	No smoothed loss	RAFT
Affine transform	T≤1px	17.00	16.02	16.11	17.41
	T≤3px	72.73	70.91	72.36	70.96
	T≤5px	91.93	91.94	91.65	91.19
	EPE	2.45	2.50	2.47	2.47
Gaussian random surface transform	T≤1px	12.31	10.68	12.02	11.86
	T≤3px	61.91	59.76	61.67	62.09
	T≤5px	87.81	87.14	87.40	87.17
	EPE	2.87	2.97	2.92	2.92

We use the matching accuracy and the end-point error (EPE) to measure the precision of the predicted optical flow field.

The matching accuracy is the percentage of matches whose l_2 distance to the ground truth correspondence is smaller than a predefined threshold, such as 1, 3, or 5 pixels as used here. The EPE is the standard error measure between the predicted flow vector and the ground truth, averaged over all pixels [55]. In order to avoid the disturbing of boundary pixels, the experiment results shown in Tables I and II are calculated using only the central parts of the predicted flow maps, sized of 312 $\times$ 312 pixels.

As we can see, for both the validation and the test set, the registration accuracies of the GRS transform are lower than the random affine transform, which is much simpler in calculation and the flow field presents a linear variation as shown in Fig. 6. Compared with the test set, the validation set is more difficult to be coregistered, considering its lower matching accuracy and higher EPE. It is caused by that the test set is mostly composed by image patches of rural areas, while the validation set contains more image patches of urban areas, which comprise plenty of man-made objects that present sharp height variations and distinctly higher radar reflection intensity, compared with its surroundings. Therefore, the optical and SAR image pairs of urban area exhibit wider radiometric and geometric gap between the corresponding optical and SAR images.

Compared with the registration results when the network is trained without the dilated feature concatenation procedure, the proposed OSFlowNet presents higher matching accuracy and also lower EPE for all the four different test trials. We speculate that the performance improvement is brought by the enriched the pixelwise features and the enlarged receptive field size of the proposed feature extraction network. The feature discriminability is enhanced with very limited additional computation. When the network is trained straightforwardly using the l_1 normed distance between the estimated flow and the ground truth flow, other than the proposed smoothed flow error, the performance is noticeably reduced. It verifies that paying more attention on “hard negative” image areas with larger displacement error during network training is helpful. Besides, the previous performance superiorities are more distinct for the GRS transforms, where the geometric relationship between the input image pair is more complex.

We also evaluate the performance of the primary RAFT optical flow network, which uses shared weights for the feature extraction of the input image pair, and a wider network that doubles the training time. Although a comparable image registration result is obtained on the test set, its performance on the validation set decreases significantly. It is likely because that, compared with the rural images, sharing weights is more detrimental for the matching of the optical and SAR images of urban areas, which exhibit larger appearance difference in-between.

C. Parameter Sensitivity Analysis

The proposed OSFlowNet framework consists of three key parameters as follows:

1) The predefined pixel range r, which is used to collect the local correlation similarity feature vectors from the multiscale full correlation cost volumes. Our default set is $r = 3$.

TABLE III
Parameter Settings for Sensitivity Analysis

Parameters	Variable	Fixed parameters
r	r = 2, 3, 4, 5	a = 1.2, ω = 0.8
a	a = 1.0, 1.2, 1.4	r = 3, ω = 0.8
ω	ω = 0.6, 0.8, 1.0, 1.2	r = 3, a = 1.2

TABLE IV
Parameter Sensitivity of r

Metrics	r, a = 1.2, ω = 0.8			
	2	3	4	5
T ≤ 1	6.72	9.84	8.79	5.00
T ≤ 3	46.35	51.80	52.76	36.13
T ≤ 5	80.88	83.79	83.67	71.67
EPE	3.54	3.22	3.25	4.19

TABLE V
Parameter Sensitivity of a

Metrics	a, r = 3, ω = 0.8		
	1.0	1.2	1.4
T ≤ 1	8.33	9.84	8.13
T ≤ 3	50.39	51.80	50.05
T ≤ 5	82.85	83.79	81.66
EPE	3.33	3.22	3.42

TABLE VI
Parameter Sensitivity of ω

Metrics	ω, r = 3, a = 1.2			
	0.6	0.8	1.0	1.2
T ≤ 1	8.03	9.84	9.21	8.60
T ≤ 3	51.53	51.80	51.19	49.08
T ≤ 5	82.73	83.79	84.01	83.29
EPE	3.35	3.22	3.24	3.32

2) a from (2), which determines the loss penalty exponent on image areas with larger displacement errors during network training. We set $a = 1.2$.

3) ω from (3), which is the exponentially weight of different GRU unrolled iterations. The default set is $\omega = 0.8$. In order to inspect our parameter settings, herein we conduct three independent experiments for parameter sensitivity analysis.

As shown in Table III, in each independent comparative experiment, only one parameter is variable, while the other two are set as the default values. Tables IV–VI present the dense matching results on the validation set of the OSdataset with GRS transforms, after the network is retrained and evaluated under different parameter settings. We can see that within the three parameters, the network is most sensitive to the predefined pixel range r for the correlation similarity feature vector collection, as shown in Table IV. The default value $r = 3$ presents the best matching accuracy. Setting $r = 4$ is also acceptable. However, the matching performance would significantly decrease if r is too small or too large. We assume that when r is set as only two pixels, the maximum search range of the correlation similarity features collected from the first and the second correlation volumes C^1, C^2 would be too small. On the other hand, setting

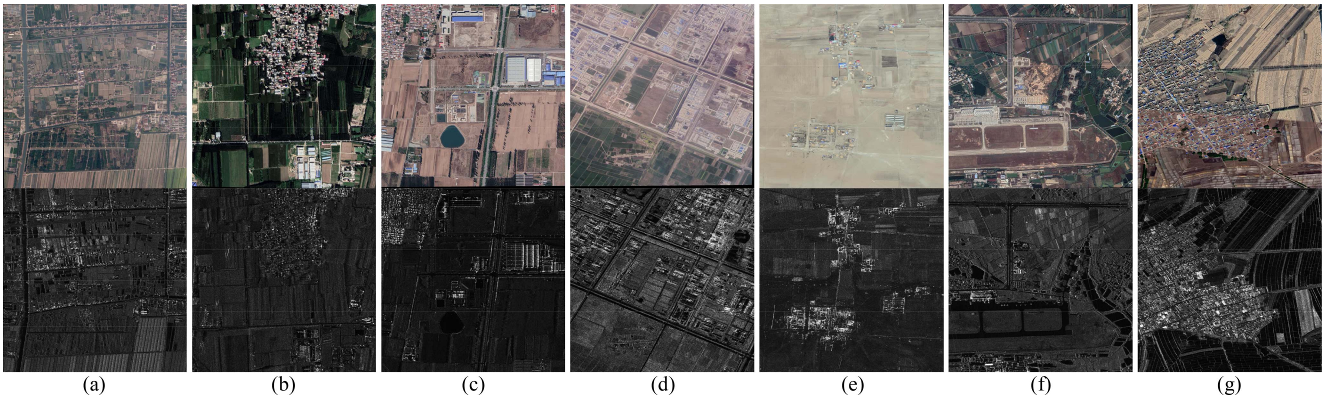


Fig. 7. Optical and SAR image pairs of seven different scenarios used for comparative test, with each image sized of 1600 × 1600 pixels. The top row shows the optical images that have been warped by the simulated spatially varying geometric distortions, and the bottom row shows the corresponding SAR images. Similar to [21], the optical-SAR pairs shown in subfigure (a), (b), (c), (d), (e), (f), and (g) are, respectively, denoted as 1.SH, 2.ZZ, 3.BJ, 4.XS, 5.MG, 6.HZ, and 7.SHJ, which are short for 1-ShangHai, 2-ZhuoZhou, 3-BeiJing, 4-XiangShui, 5-MingGan, 6-HuiZhou, and 7-SongHuaJiang.

$r = 5$ would bring in too many distractions for each scale level, which would also impair the matching performance. As for the loss penalty exponent value$a$, as well as the exponentially weight of different GRU unrolled iterations ω, a smaller or bigger value also decreases the matching accuracy, but with a quite small margin.

TABLE VII
MATCHING ERROR COMPARISON BETWEEN PIXELWISE AND BLOCKWISE APPROACHES

Data	NPT	Pixelwise match			Blockwise match		
		E_fx	E_fy	EPE	E_fx	E_fy	EPE
1.SH	75	1.14	0.80	1.39	0.79	0.62	1.00
2.ZZ	36	1.16	1.30	1.74	1.27	1.14	1.71
3.BJ	44	1.27	0.80	1.50	1.09	0.58	1.24
4.XS	52	1.34	1.56	2.06	0.86	1.53	1.76
5.MG	53	1.06	1.48	1.82	0.83	0.97	1.28
6.HZ	91	0.93	1.02	1.38	0.78	0.63	1.00
7.SHJ	78	1.17	0.76	1.39	0.76	0.55	0.94

D. Evaluation of the Self-Supervised Network Fine-Tuning Procedure

As can be seen from Tables I and II, the EPE of the predicted optical flow vectors averages about three pixels. Still, over 30% pixels present a mismatching error of more than three pixels. Further performance improvement is expected. Here, in this subsection, we conduct an experiment to verify the effectiveness of the proposed self-supervised network finetuning strategy.

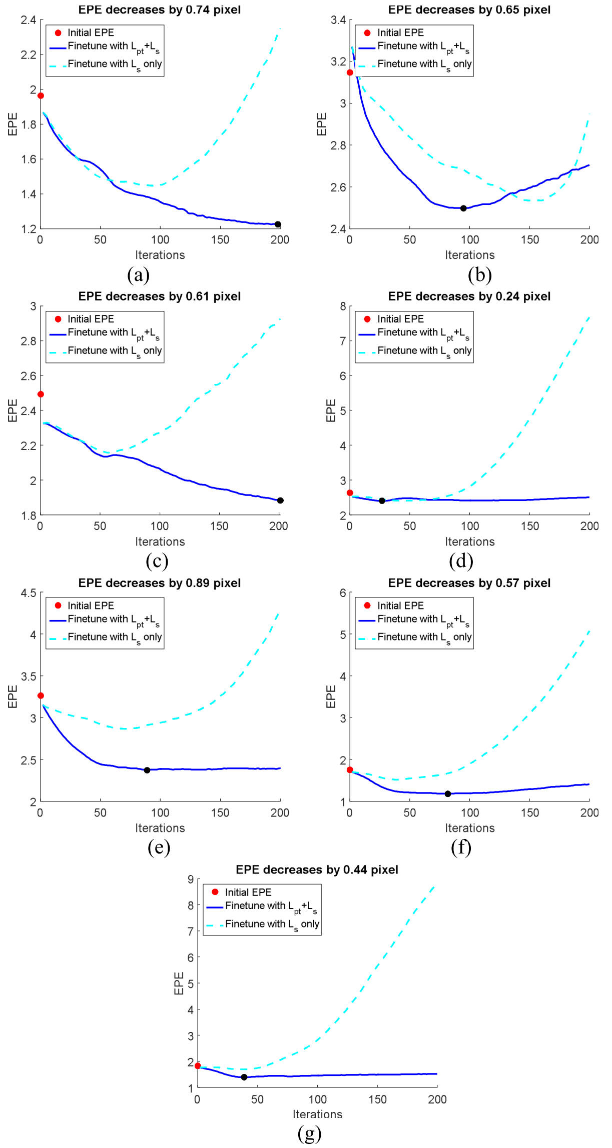
1) Experiment Setup: Different from the previous ablation study section that uses the small image patches sized of 512 × 512 pixels, here we prepare seven pairs of large optical and SAR image pairs in various scenarios, all sized of 1600 × 1600 pixels with 1-m spatial resolution, as shown in Fig. 7. They are images of the GaoFen-3 SAR and Google Earth optical pairs of seven different cities from China, termed as 1.SH, 2.ZZ, 3.BJ, 4.XS, 5.MG, 6.HZ, and 7.SHJ, respectively. These imageries have been used and described in detail in the previous article on optical and SAR sparse feature point matching [23].

Note that these are all images of flat areas, and have been coregistered using a half-manual strategy [20], [23]. Herein, a random geometric transformation is conducted on each large image pair, which is the combination of a random affine transform and a GRS transform. In this way, an optical-SAR dense registration test set is acquired, which exhibits complex spatially varying geometric transforms, and also the ground-truth optical flow vectors are preknown.

2) Comparison Between the Pixelwise Matching and Blockwise Matching Results: First of all, we would like to check if the blockwise matching method is able to produce better matching precision than the pixelwise matching result. For each test pair sized of 1600 × 1600 pixels, we collect the fixed feature points from the optical image using a fixed step of 128 pixels, along both the horizontal and vertical directions, shown as the red dots on the optical input image in Fig. 5. In this way, 100 sparse points are obtained for the optical image. The corresponding moving feature points on the SAR image can be located using the initial flow vector estimated by the pretrained OSFlowNet. Next, for each candidate point pair, block matching is conducted to refine the initial flow vector, based on the blockwise feature volumes extracted by the OSMNet [23]. Here, in the subsequent experiment, we set the local searching radius as 32 pixels, with the optical template features sized of 193 × 193 × 9, and SAR search features sized of 257 × 257 × 9, resulting in a SSD score map sized of 65 × 65 pixels.

Since the ground truth optical flow fields are preknown for the seven test instances, we are able to calculate the matching error of all the sparse locations, for both the initial pixelwise matching results obtained by the OSFlowNet and the blockwise matching results obtained by the OSMNet, as shown in Table VII. Following (5) and (6), dozens of the sparse correspondences are considered as outliers, and only the remaining ones are used as the pseudo-ground truth for the subsequent optical flow fine-tuning procedure. For example, 75 out of 100 corresponding feature points are assumed to be high confidential for the 1.SH test instance, while only 36 for the 2.ZZ test instance.

Table VII presents the matching error of both the horizontal and the vertical directions, termed as E_fx and E_fy, respectively. The EPE values are also calculated. Compared with



(a) (b) (c) (d) (e) (f) (g)

Fig. 8. EPE plots of the self-supervised optical flow finetuning procedure on the seven simulated test instances shown in Fig. 7. In each subfigure, the red dot represents for the initial EPE value of the optical flow map estimated by the pretrained OSFlowNet. The blue curves illustrate the variation of the EPE value with increased finetuning iterations. The dashed cyan curves show the EPE results when the sparse matching loss is ignored, and only the smoothness term is used for the supervision of flow fine-tuning.

the initial pixelwise matching results, the blockwise matching method decreases the EPE values by a significant margin, about 0.3 to 0.5 pixel for six out of the seven test instances. It verifies that the block matching method is able to obtain higher matching precision for the sparse feature points, and the matched correspondences are indeed feasible to be considered as the pseudo-ground truth, and be used to further refine the dense optical flow field.

3) Self-Supervised Fine-Tuning Results: By taking the sparse correspondences calculated by the blockwise matching network as the supervision, the deep flow network can be fine-tuned specifically for each test instance. Fig. 8 presents the EPE plots of the self-supervised optical flow fine-tuning procedure on the seven test instances. In each subfigure, the red dot represents for the initial EPE value of the optical flow map estimated by the pretrained OSFlowNet. The blue curve illustrates the variation of the EPE value with increased finetuning iterations.

TABLE VIII
EPE AND CMR (%, T≤1PX, 3PX, AND 5 PX) RESULTS BEFORE AND AFTER SELF-SUPERVISED NETWORK FINE-TUNING

Data	OSFLOWNET				OSFLOWNET-FT			
	EPE	1PX	3PX	5PX	EPE	1PX	3PX	5PX
1.SH	1.96	20.1	84.1	98.3	1.55	35.7	92.2	99.8
2.ZZ	3.15	8.4	54.0	88.1	2.64	13.0	71.6	91.7
3.BJ	2.49	17.2	64.5	95.3	2.15	22.9	72.8	97.4
4.XS	2.64	9.8	61.7	97.8	2.48	8.4	67.6	99.8
5.MG	3.26	10.7	57.0	83.9	2.45	22.9	76.0	88.4
6.HZ	1.75	19.9	89.6	99.9	1.21	49.5	97.7	100
7.SHJ	1.83	30.2	83.2	96.1	1.42	40.5	91.6	99.7

We can see that the proposed self-supervised finetuning method, termed as OSFlowNet-Ft, leads to a significant improvement in terms of EPE, ranging from 0.24 to 0.89 pixels. The black dot denotes the position of the best EPE result, whose location varies for different test instances. Therefore, in practical application, it is implausible to preknow the optimal number of fine-tuning iterations. On the other hand, we can see that for all the seven instances, the EPE values decline noticeably for the first 50 to 100 iterations. Accordingly, we record the optical flow estimation result after 50 fine-tuning iterations, as shown in Table VIII. Averagely on the seven different test instances, the EPE value decreases by 20%, the CMR with T $\leq$ 1 pixel increases by 64%, and the CMR with T $\leq$ 3 pixel increases by 17%.

Furthermore, we try to conduct the fine-tuning procedure by ignoring the sparse matching loss L_{pt}, and using only the smoothness regularization term L_s. The EPE results are also presented in Fig. 8 with the dashed cyan curves. For the first three datasets, noticeable performance improvement can still be acquired. However, for the last four datasets, the EPE values only decrease by a quite slight margin. Besides, when using the smoothness term only, the EPE values blow up with increased finetuning iterations for all the test instances. This phenomenon further verifies the effectiveness of the proposed sparse matching supervision strategy.

E. Comparison With Representative Optical Flow Based Optical-SAR Dense Registration Approaches

Here in this subsection, we compare the proposed OSFlowNet-FT framework with three representative optical flow based optical-SAR image registration methods, including the SIFT-Flow [40], GeFolki [42], and OS-Flow [43] algorithms, which have been described in detail in Section II-C. They are all handcrafted methods specifically designed for the optical and SAR image pixelwise dense matching. Until recently, we have not found any deep-learning-based approach in the particular research field of optical-SAR dense registration.

Two different test sets are used in this comparative experiment. The first one is the previously used seven optical-SAR pairs with simulated ground surface fluctuations, as shown in

TABLE IX
EPE/CMR (%, T≤3px) Results of Different Registration Methods

Methods	SIFT-Flow	GeFolki	OS-Flow	OSFlowNet-Ft
1.SH	5.59/26.7	4.21/41.6	6.45/55.5	1.55/92.2
2.ZZ	5.74/22.4	11.12/5.9	5.71/26.6	2.64/71.6
3.BJ	6.78/14.8	124.0/0.0	3.64/51.4	2.15/72.8
4.XS	8.01/18.0	5.38/44.5	5.30/46.1	2.48/67.6
5.MG	515/0.0	44.53/0.0	6.64/44.2	2.45/76.0
6.HZ	9.10/16.9	6.00/41.4	2.54/73.3	1.21/97.7
7.SHJ	6.45/15.1	8.25/29.5	2.26/76.9	1.42/91.6

Fig. 7. Since the ground truth optical flow maps are preknown, we can evaluate the performance of different registration methods in both quantitative and qualitative ways. The second test set consists of six optical-SAR pairs of real hilly land areas for which the registration results are checked through visual inspection.

As for the parameter settings of the three handcrafted optical-SAR image dense registration methods, we follow the instructions of the OS-Flow paper [43], where the authors have conducted extensive comparative experiments to find the proper parameter configurations of all the three comparative algorithms. For both the SIFT-Flow and OS-Flow methods, we choose to conduct the optical flow optimization procedure using only the global HS framework, considering that in the LK framework, the interpolation and convolution operators on high-dimensional feature descriptors are unbearably time-consuming, and also the HS and LK approaches produce similar registration accuracy, as presented in [43].

1) Experimental Results on the Simulated Test Set: The three handcrafted methods and our proposed OSFlowNet-Ft method are first applied on the seven optical-SAR image pairs that present simulated locally varying geometric distortions. The registration results are evaluated quantitatively in terms of registration precision using both the EPE and the matching accuracy indicators, as shown in Table IX. Also, they are compared qualitatively in terms of the visual inspection of the estimated optical flow field images, as shown in Fig. 9. The visualizing results of the flow fields shown in Fig. 9 are produced by following the instructions in [39], which takes the magnitude and orientation of the corresponding flow vector as the hue and saturation of the optical flow image.

From Table IX, we can see that the GeFolki method fails in registration of both the 3.BJ and 5.MG test instances. We assume it is caused by that both of the instances exhibit large amount of repeated textures or even textureless areas, as shown in Fig. 7(c) and (e). In this circumstance, the iterative optical flow optimization procedure is easily to fall into the wrong local optimal position, considering that the GeFolki algorithm measures the pixelwise similarity merely based on the ranked raw pixel intensities within a small local window.

The SIFT-Flow method also fails for the 5.MG test images, but the 3.BJ optical and SAR image pair can be preliminarily coregistered. By replacing the intensity value with the SIFT features, the receptive field of the pixelwise descriptor is enlarged, and therefore better discrimination can be obtained for images with repeated textures. Furthermore, the OS-Flow method obtains acceptable registration results for all the seven test images. Especially for the 6.HZ and 7.SHJ test instances, more than 70% pixels are coregistered with the matching error smaller than three pixels. Note that SIFT-Flow and the OS-Flow methods share the same algorithm framework. The difference lies in that the SIFT-Flow method applies the SIFT method for the pixelwise feature extraction of both the optical and SAR images. On the other hand, the OS-Flow method uses the optical-GLOH descriptor on the optical image, while the specifically designed SAR-GLOH descriptor on the SAR image. In this way, the brightness constancy assumption between the optical and SAR images can be satisfied to a certain degree.

As for our proposed OSFlowNet-Ft method, the EPE values for all the seven test instances are smaller than three pixels. Meanwhile, three out of them are even smaller than two pixels, with more than 90% of the image pixels exhibiting a coregister error smaller than three pixels. This way, the OSFlowNet-Ft framework is able to satisfy most of the practical application of the high-resolution optical and SAR image registration problem.

The registration results on the seven simulated test instances in terms of the flow field visualizing images are presented in Fig. 9, where each column presents the registration results of one test instance using the Gefolki, SIFT-Flow, OS-Flow, the proposed OSFlowNet-Ft method, as well as the ground truth flow image, respectively, from top to bottom. It can be observed that the flow field images produced by the OSFlowNet-Ft method are quite alike to the ground truths. On the other hand, the flow images of all the other three methods contain significant color drift, and also unacceptable local distortions. Note that the flow images of the SIFT-Flow and the OS-Flow methods both exhibit similar color patterns with the ground truth, but comprise many erroneous color fragment, indicating incorrect local registration results. On the other hand, the flow images of the Gefolki method are quite smooth, but sometimes present apparently different color compositions with the ground truth image, which means that the registration is totally failed.

2) Experimental Results on Images of Real Hilly Lands: The previous experiments all use optical-SAR pairs with simulated locally varying geometric distortions in between. Herein, the proposed OSFlowNet-Ft framework as well as the other three comparative approaches are applied on six optical and SAR image pairs of real hilly lands, where the geometric relationships between them are even more complex than the combination of the random affine and GRS transforms. The six test instances are shown in Fig. 10, where the optical images are from the Google Earth platform, and the SAR images are also from GaoFen-3, all with 1-m spatial resolution and size of 1600 × 1600 pixels. The first two instances in the top row of Fig. 10 are images of roughly plainness area with slight and smooth land surface fluctuation. The middle two pairs contain a small part of hilly lands, while the majority part of the last two optical-SAR pairs are mountainous areas.

Fig. 11 shows the visualizing results of the flow fields calculated by different methods, where each column presents the flow images of one test instance using the Gefolki, SIFT-Flow, OS-Flow, and the proposed OSFlowNet-Ft methods, respectively, from top to bottom. Similar to the experiment results on the

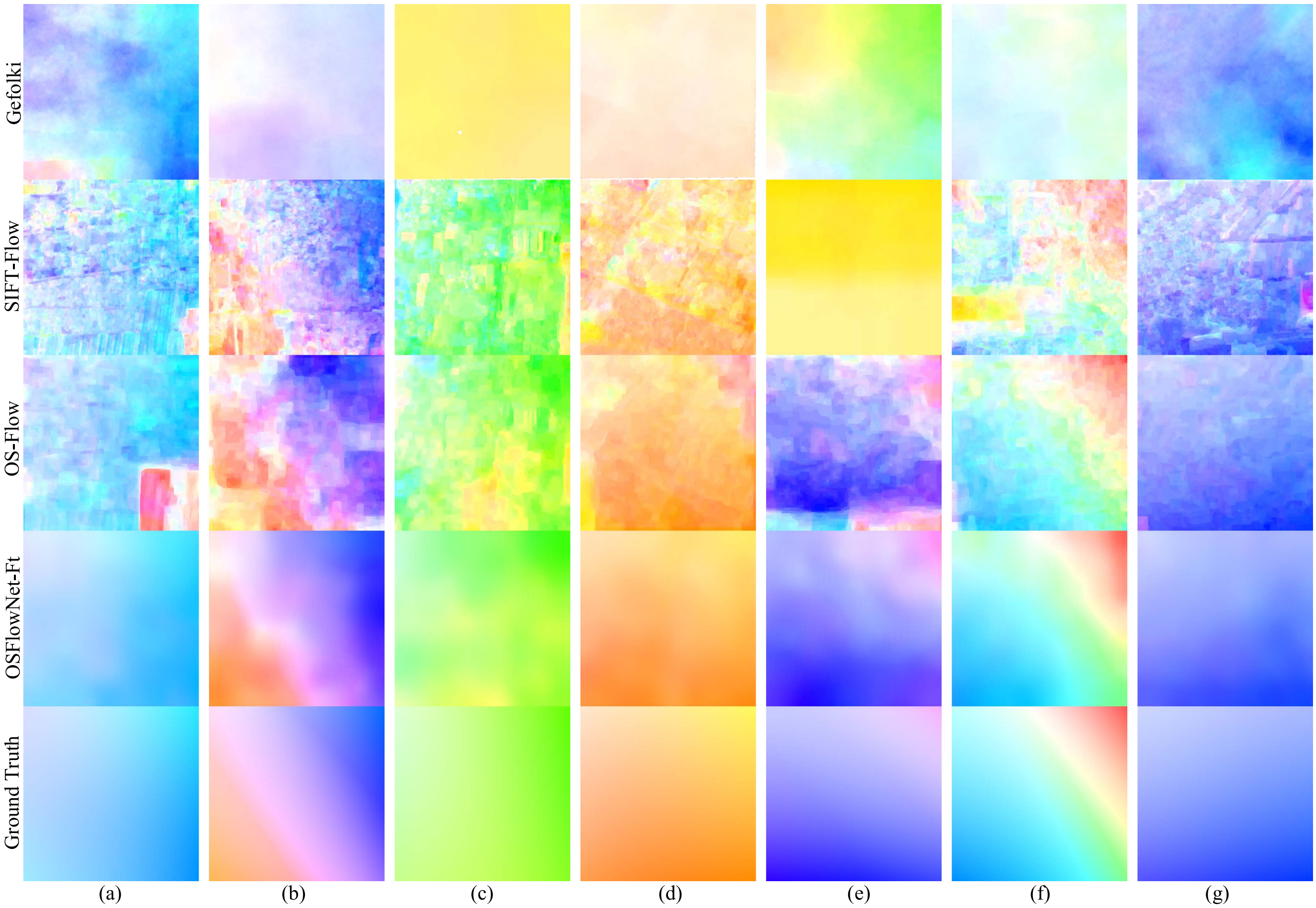


Fig. 9. Visualizing results of the flow fields produced by different optical-SAR dense registration methods, including the Gefolki [42], SIFT-Flow [40], OS-Flow [43], OSFlowNet-Ft, and the ground truth flow images, from top to bottom. In each row, the seven subfigures corresponds to the seven test optical-SAR image pairs shown in Fig. 7, termed as 1.SH, 2.ZZ, 3.BJ, 4.XS, 5.MG, 6.HZ, and 7.SHJ.

simulated test set as shown in Fig. 9, there is a good continuity in each flow field image produced by the OSFlowNet-Ft method. On the other hand, plenty of unrealistic sudden flow changes occur in the flow fields calculated by the other three methods, especially for the Gefolki method.

The registration results in terms of mosaic images of the six test instances are shown in Figs. 12–17. In each figure, the mosaic results of the original geocoding images, the SIFT-Flow method, the OS-Flow method, as well as the proposed OSFlowNet-Ft method are presented in subfigures (a), (b), (c), and (d), respectively. The full mosaic image of the Gefolki method is left out for a better page make-up, considering that it fails for most of the test instances. In addition, for each registration method, enlarged subimages of two randomly selected locations of the full mosaic image, highlighted in a red and a blue rectangle, are shown below, where the results of Gefolki method are also presented as subfigure (e1) and (e2).

As for the first test instance shown in Figs. 10(a) and 12, all the four different algorithms seem to produce acceptable coregistration results. After checking the enlarged subfigures, we can see that both the SIFT-Flow and OS-Flow methods bring in severe image distortions, which make the straight lines of the building and road boundaries wrapped into curves with nonnegligible twists and turns. On the other hand, the proposed OSFlowNet-Ft framework is not only able to produce fine pixelwise matching result, but also keeps the ground objects in good integrality, as shown in Fig. 12(d), (d1), and (d2). The local twist problem seems inconspicuous in the warped image of the Gefolki method. However, the Gefolki fails to coregister the second subimage, as shown in Fig. 12(e2).

Although the Gefolki method has been proven to be effective for the registration of low resolution multimodal image pairs [42], [44], it fails the registration in many test cases used in this article, as shown in the first row of Fig. 11 and the subfigures (e1) and (e2) of Figs. 12–17. It is probably caused by that the radiometric and geometric disparity becomes larger as the spatial resolution goes finer. Therefore, merely using the intensity value is not sufficient to depict the similarity between corresponding optical and SAR pixels of high spatial resolution.

For the second test instance, we can see from Fig. 13(a) that the left 1/3 part of the original geocoded optical-SAR image pair has already been roughly coregistered. Still, the right 2/3 part exhibits apparent location deviation, which is caused by the elevation fluctuation. Note that the roads in the SAR images show relatively low intensities, whereas roads are high-reflectivity lines in the optical images [43]. In this test case, the main roads

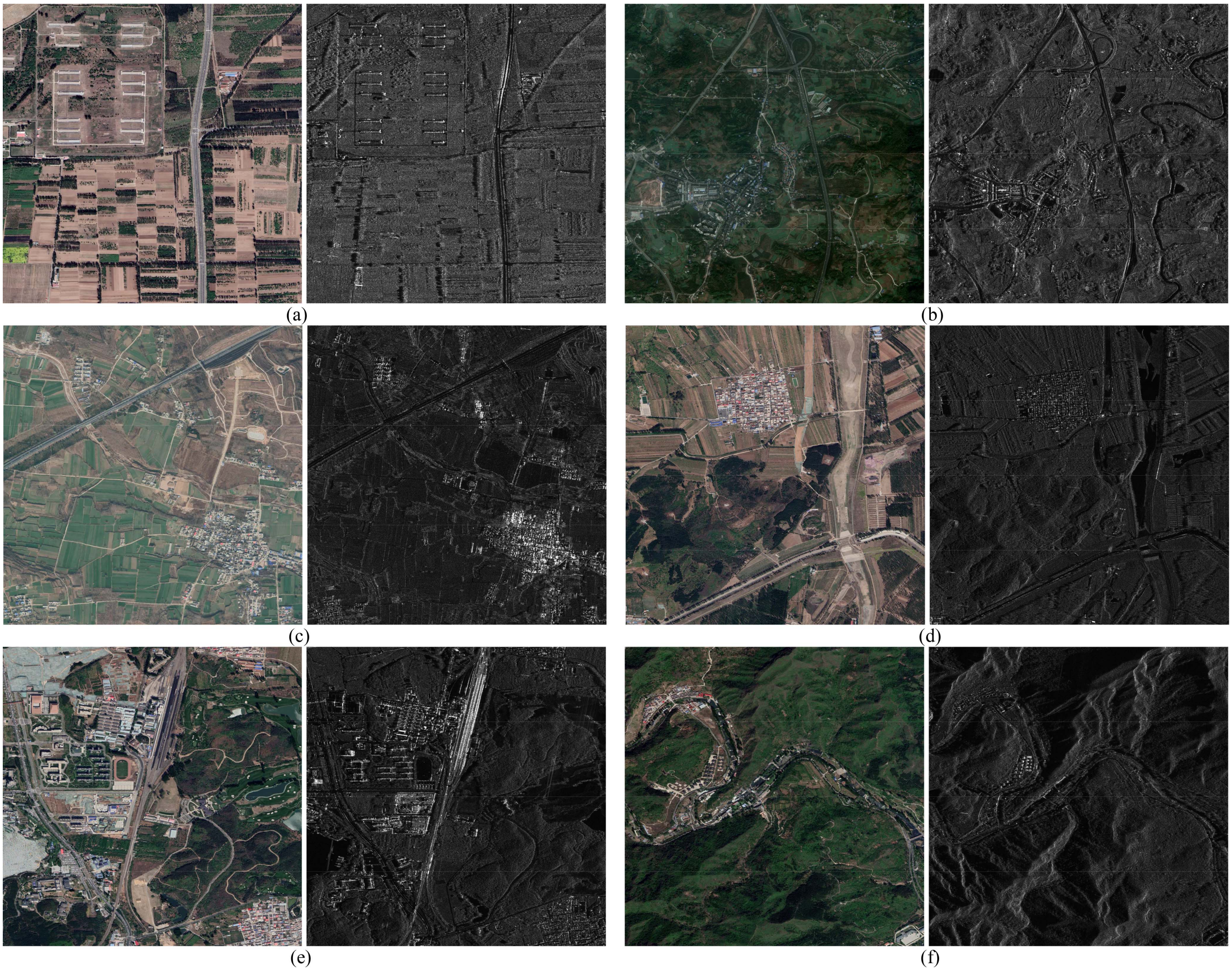


Fig. 10. Optical-SAR image pairs of six different scenarios of real hilly lands used for comparative test, with each image sized of 1600 × 1600 pixels. (a) First test instance. (b) Second test instance. (c) Third test instance. (d) Fourth test instance. (e) Fifth test instance. (f) Sixth test instance.

and the wide river that both north–south cross-cutting seem to be fairly coregistered by all the SIFT-Flow, OS-Flow, and the OSFlowNet-Ft methods. The subroads shown in the first enlarged subimage are also basically coregistered. However, in the second enlarged subimage, the roads are matched to the building rows for both the SIFT-Flow and OS-Flow methods. Moreover, severe local twists are produced, similar with the first test instance. On the other hand, the OSFlowNet-Ft successfully coregister the "dark" SAR roads to the "light" optical roads, as shown in Fig. 13(d2), without any noticeable local distortion.

As can be observed from Fig. 10(c), there are two hills located in the bottom-left corner and the top-right corner of the third instance. In this case, the flow field images of the SIFT-Flow, OS-Flow, and OSFlowNet-Ft methods present similar color patterns, as shown in Fig. 11(c). The mosaic images shown in Fig. 14(b), (c), and (d) indicate that these three methods are all able to roughly coregister the third optical-SAR image pair. However, as shown in the second subimage, only the SIFT-Flow and OSFlowNet-Ft methods successfully matched the main road and the rectangle man-made platform.

In all our experiments, the SAR images are taken as the fixed reference one, while the optical images are warped to the SAR images. Note that there is a very short distance between the two subimages of the third test instance. By comparing the enlarged mosaic images before and after registration, we can see for the first subimages shown in Fig. 14(a1) and (d1), the optical patch is registered to the SAR patch by moving to the top-left direction. While for the second subimages shown in Fig. 14(a2) and (d2), the optical patch has to be moved in the direction of bottom-right. It indicates that the hill located in the top-right corner of the third test instance presents a drastic topographic relief, which is quite challenging for image registration. Still, our proposed OSFlowNet-Ft method is able to present desirable matching result.

For the fourth test instance shown in Figs. 10(d) and 15, the OS-Flow method totally fails the registration. Surprisingly, the SIFT-Flow method is able to obtain acceptable matching result. This circs also happens for the second subimage of the third case. It indicates that the optical-GLOH and SAR-GLOH features are not robust enough for the optical-SAR image matching issue.

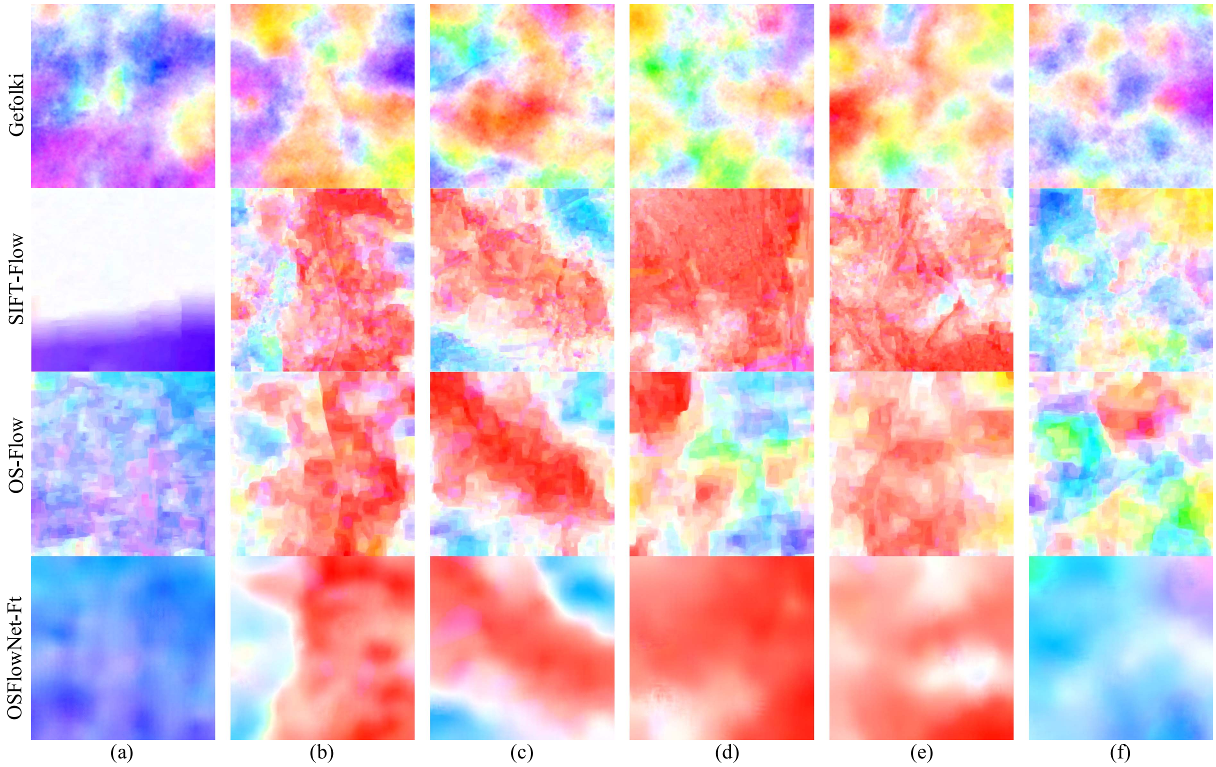


Fig. 11. Visualizing results of the flow fields produced by different optical-SAR dense registration methods, including the Gefolki [42], SIFT-Flow [40], OS-Flow [43], and the proposed OSFlowNet-Ft, from top to bottom. In each row, the six subfigures correspond to the six test optical-SAR image pairs of the real hilly lands shown in Fig. 10.

As always, the proposed OSFlowNet-Ft method obtains excellent coregistration result. Especially as observed from the middle bottom of Fig. 15(d2), the "dark" SAR road is successfully matched to the "bright" optical road, although on the SAR image, there is a confusable "bright" line of field bank lying beside the "dark" road.

The fifth test instance is composed of half urban district and half hilly land. From the flow field color images shown in Fig. 10(e) and the mosaic image shown in Fig. 16(b), (c), and (d), it seems that except for the Gefolki method, all the other three ones are able to generally coregister this optical-SAR image pair. However, as shown in the first subimages shown in Fig. 16(b1) and (c1), which is a dense building area, the local distortions produced by the SIFT-Flow and OS-Flow methods are too abominable to be applied in practice.

As shown in Fig. 10(f), most regions of the sixth test instance are mountainous, for which, the optical-SAR image pair seems impossible to be coregistered, due to the textureless ground surface, and also the severe geometric distortions between them. However, in most of the practical applications of high-resolution optical and SAR images, we are chiefly concerned with image regions that exhibit human activities, other than that of primitive nature. As can be observed from Fig. 17(d) and (d1), the proposed OSFlowNet-Ft method is able to properly coregister the central residence area. Moreover, as shown in Fig. 17(d2), the mountain road is also precisely coregistered. On the other hand, as shown in the enlarged subimages of Fig. 16(b1), (b2), (c1), (c2), and (e1), (e2), all the other three methods fail the registration of this challenging test case.

Above all, the Gefolki method roughly coregister the first and the sixth image pairs, but totally fails for all the other four cases. The SIFT-Flow and OS-Flow methods are able to coregister some local regions, but fail in the other parts of each test instance. Besides, the severe local distortions produced by the two methods make them unacceptable for practical applications. The OS-Flow method is tailor-made for the optical-SAR image dense registration issue, and it has been proved to present better registration performance on the simulated dataset shown in Fig. 7. However, as for the six test instances of real hilly lands, it does not outperform the SIFT-Flow method. Amazingly, despite of the obviously different landcover types and distinct topographic conditions of the six test instances, the proposed deep optical flow framework OSFlowNet-Ft, which is trained on a small dataset with only simulated elevation fluctuations, is always able to produce satisfactory image dense registration results.

Furthermore, we compare the computational time of different dense registration methods on the six optical-SAR image pairs of real hilly lands. Considering that the code of the three handcrafted methods is all running on CPU, we present both the GPU and CPU times of the proposed OSFlowNet and OSFlowNet-Ft frameworks, as shown in Table X. We can see that without the self-supervised fine-tuning procedure, the proposed deep framework takes only 3 s to accomplish the pixelwise dense registration of the input image pair sized of 1600 × 1600 pixels on GPU, and only 12 s on CPU, which makes the proposed framework the most time efficient one when compared with the other three handcrafted methods. When the self-supervised

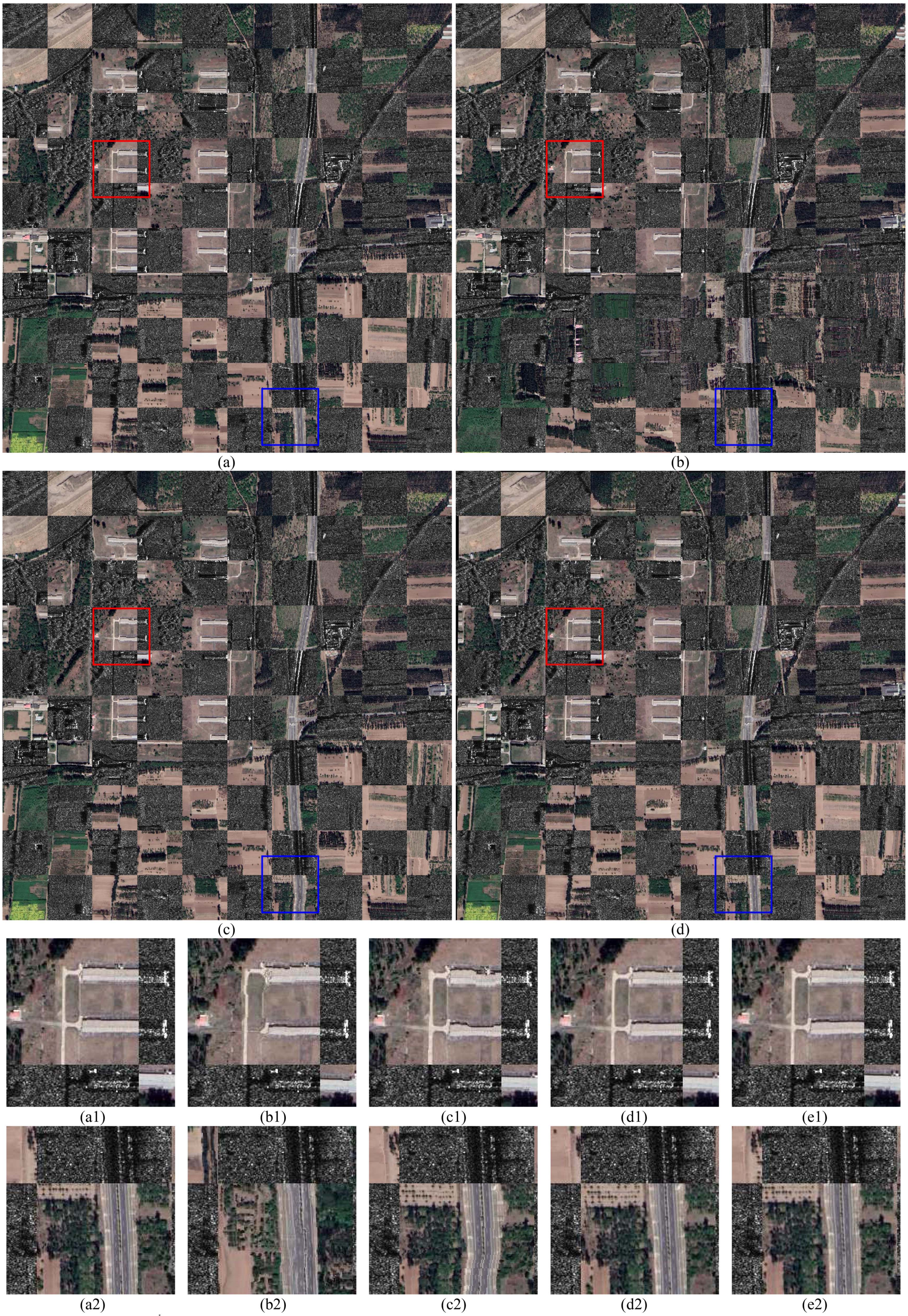
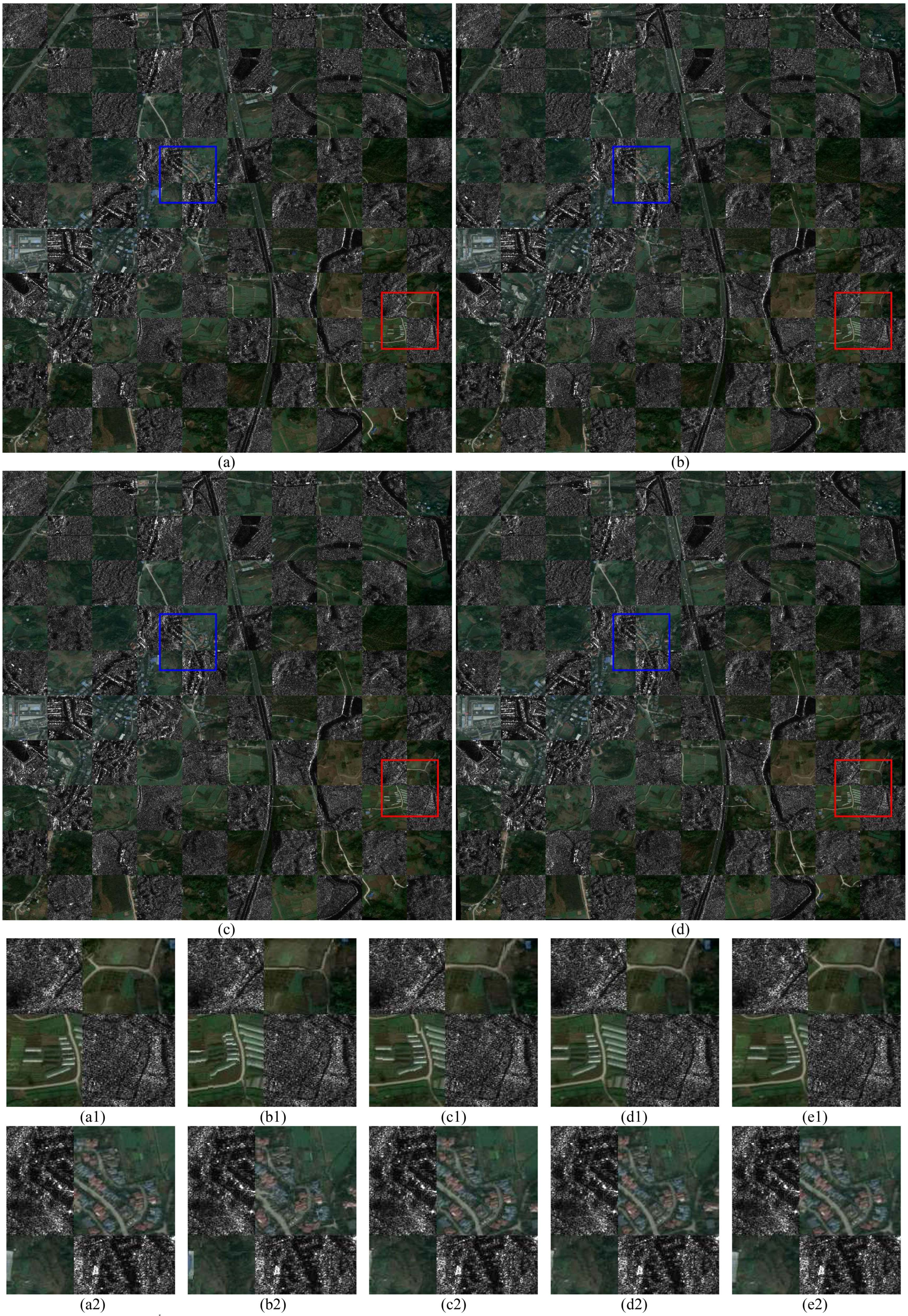


Fig. 12. Mosaic images of the first coregistered optical-SAR pair by different methods. (a) Geocoding. (b) SIFT-Flow. (c) OS-Flow. (d) OSFlowNet-Ft. (a1), (b1), (c1), and (d1) are enlarged subimages from the red rectangle of (a), (b), (c), (d) respectively. (a2), (b2), (c2), (d2) are enlarged subimages from the blue rectangle of (a), (b), (c), and (d), respectively. (e1) and (e2) are the enlarged mosaic subimages of the registration result by the Gefolki method.



(a) (b)

(c) (d)

(a1) (b1) (c1) (d1) (e1)

(a2) (b2) (c2) (d2) (e2)

Fig. 13. Mosaic images of the second coregistered optical-SAR pair by different methods. (a) Geocoding. (b) SIFT-Flow. (c) OS-Flow. (d) OSFlowNet-Ft. (a1), (b1), (c1), and (d1) are enlarged subimages from the red rectangle of (a), (b), (c), and (d), respectively. (a2), (b2), (c2), and (d2) are enlarged subimages from the blue rectangle of (a), (b), (c), and (d), respectively. (e1) and (e2) are the enlarged mosaic subimages of the registration result by the Gefolki method.

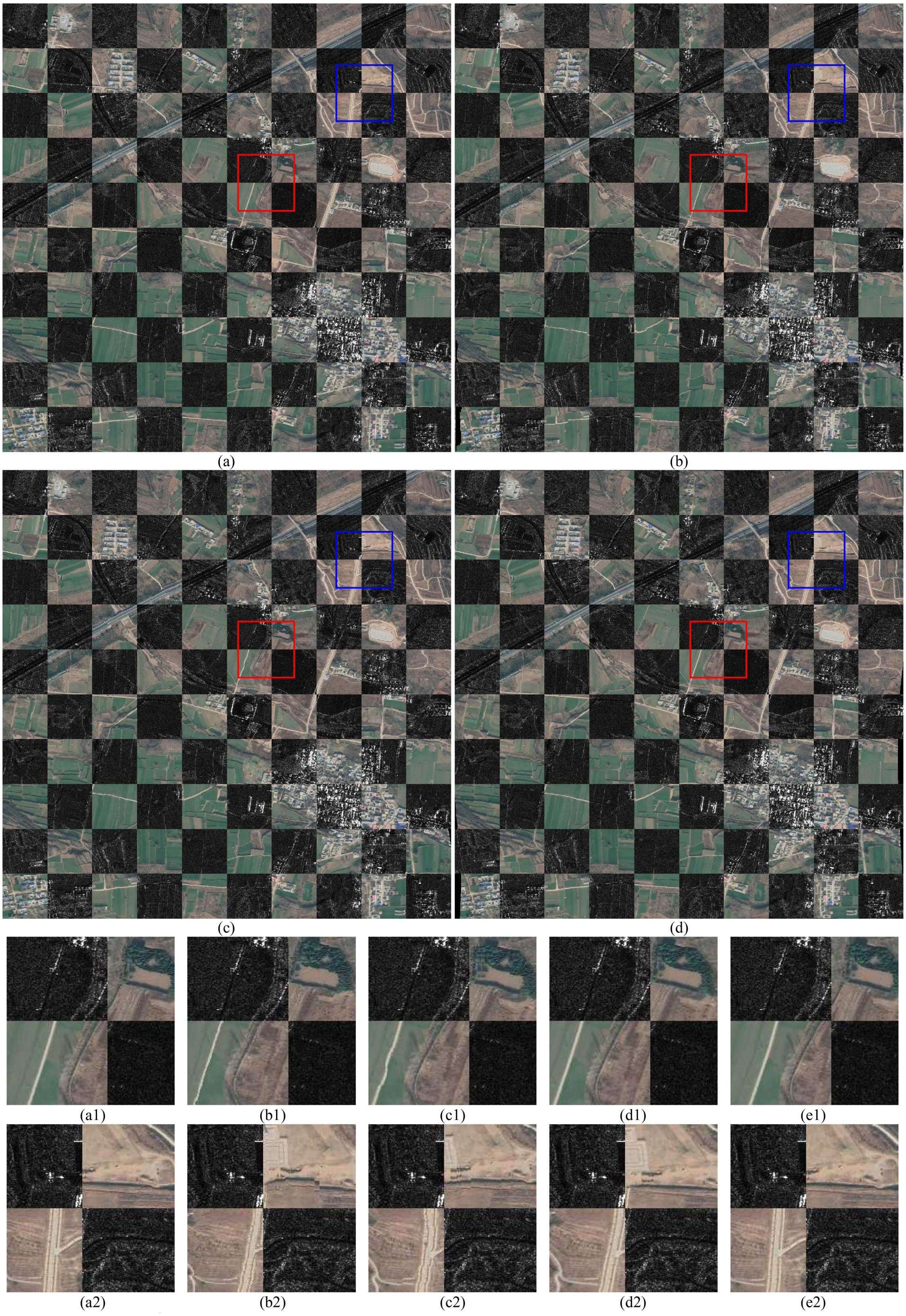


Fig. 14. Mosaic images of the third coregistered optical-SAR pair by different methods. (a) Geocoding. (b) SIFT-Flow. (c) OS-Flow. (d) OSFlowNet-Ft. (a1), (b1), (c1), and (d1) are enlarged subimages from the red rectangle of (a), (b), (c), and (d), respectively. (a2), (b2), (c2), and (d2) are enlarged subimages from the blue rectangle of (a), (b), (c), and (d), respectively. (e1) and (e2) are the enlarged mosaic subimages of the registration result by the Gefolki method.

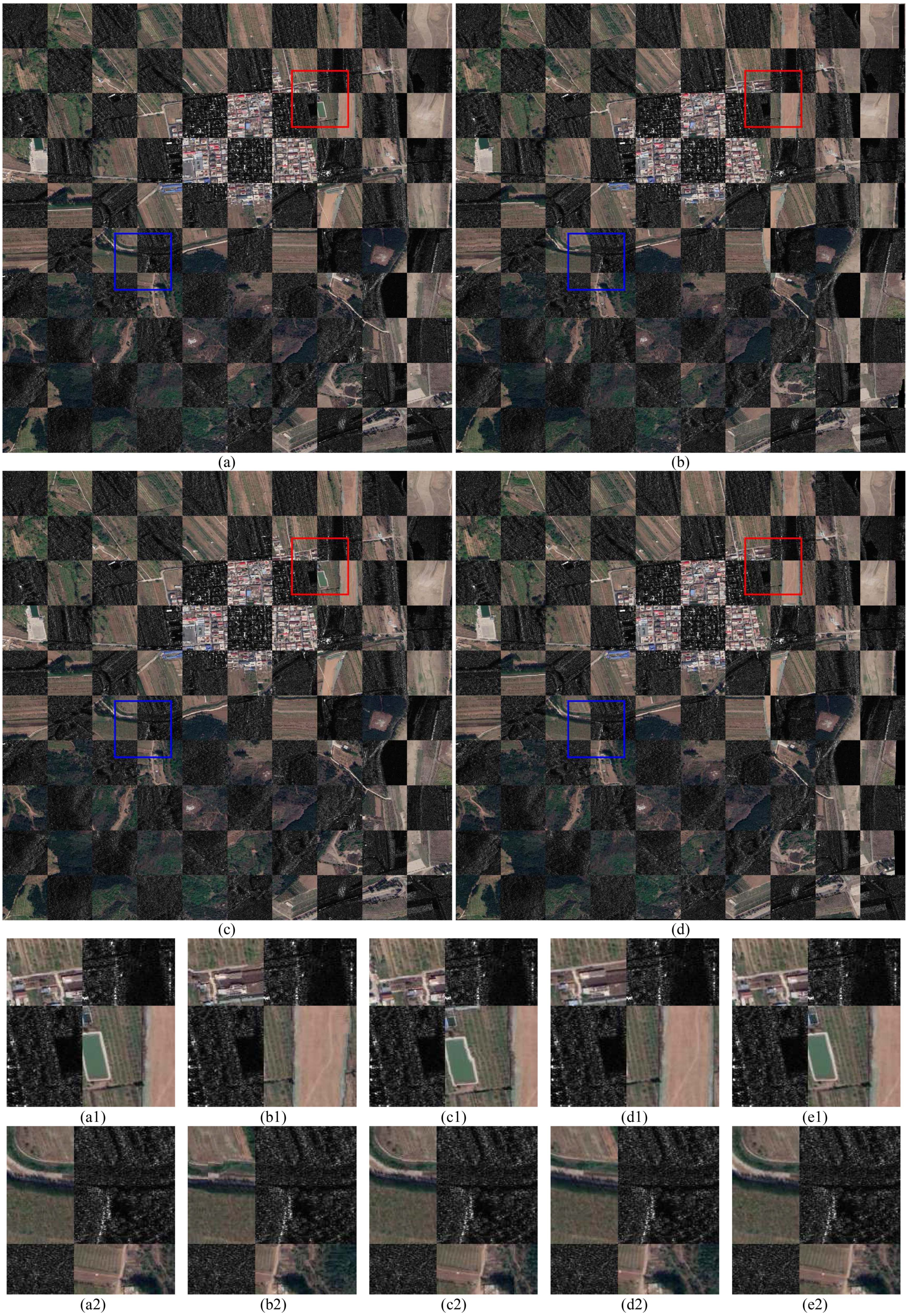


Fig. 15. Mosaic images of the fourth coregistered optical-SAR pair by different methods. (a) Geocoding. (b) SIFT-Flow. (c) OS-Flow. (d) OSFlowNet-Ft. (a1), (b1), (c1), and (d1) are enlarged subimages from the red rectangle of (a), (b), (c), and (d), respectively. (a2), (b2), (c2), and (d2) are enlarged subimages from the blue rectangle of (a), (b), (c), and (d), respectively. (e1) and (e2) are the enlarged mosaic subimages of the registration result by the Gefolki method.

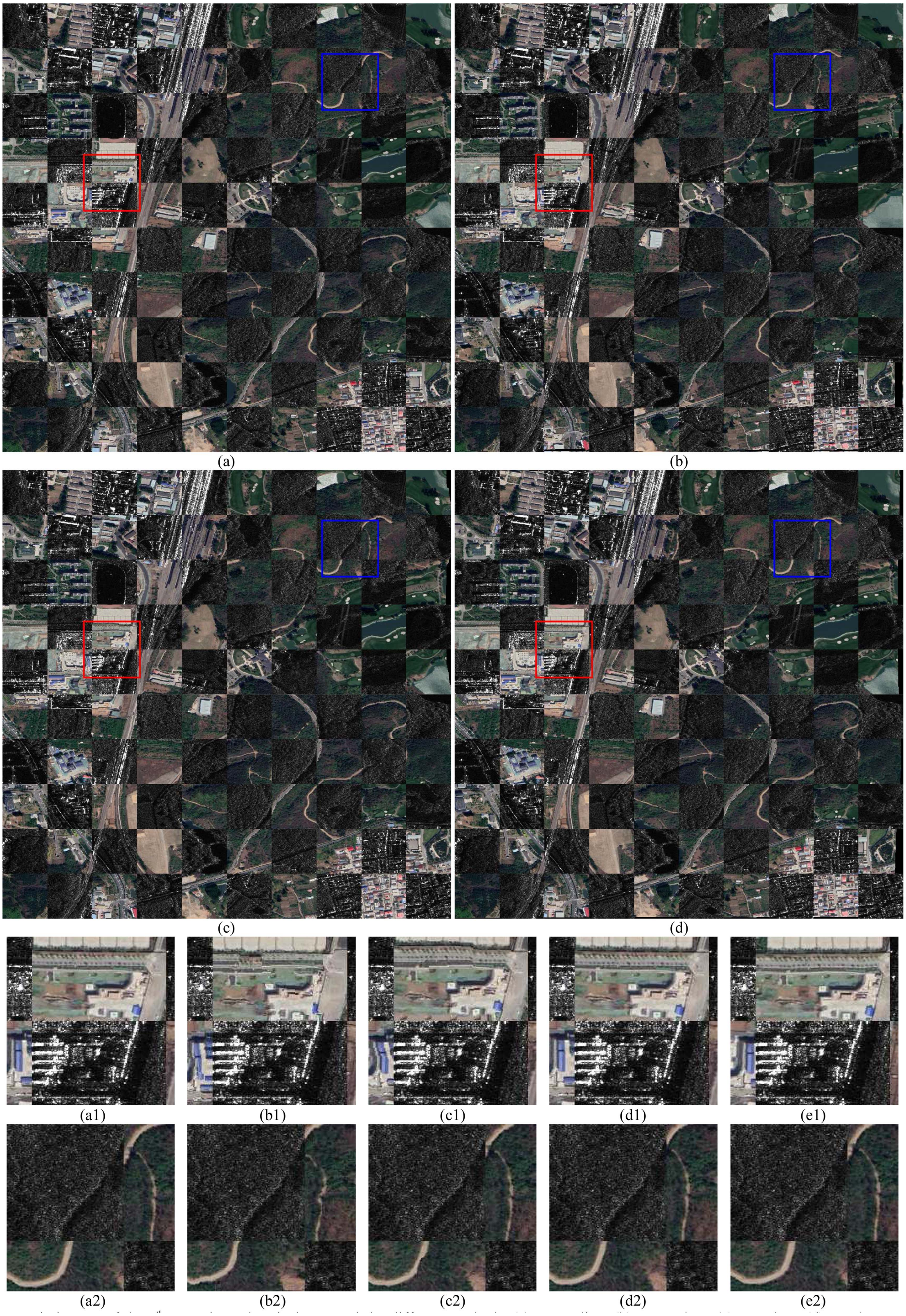
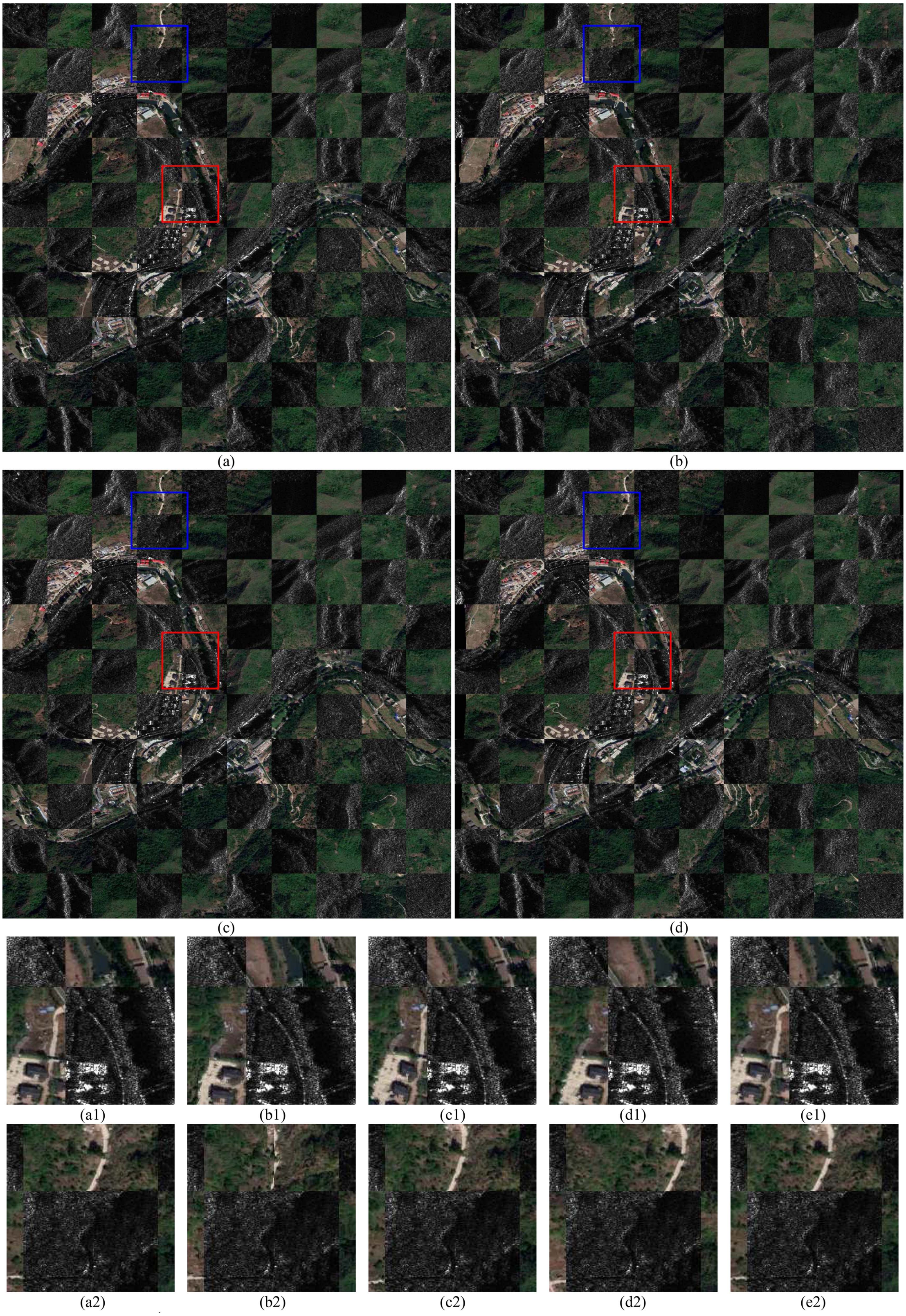


Fig. 16. Mosaic images of the fifth coregistered optical-SAR pair by different methods. (a) Geocoding. (b) SIFT-Flow. (c) OS-Flow. (d) OSFlowNet-Ft. (a1), (b1), (c1), and (d1) are enlarged subimages from the red rectangle of (a), (b), (c), and (d), respectively. (a2), (b2), (c2), and (d2) are enlarged subimages from the blue rectangle of (a), (b), (c), and (d), respectively. (e1) and (e2) are the enlarged mosaic subimages of the registration result by the Gefolki method.



(a) (b)

(c) (d)

(a1) (b1) (c1) (d1) (e1)

(a2) (b2) (c2) (d2) (e2)

Fig. 17. Mosaic images of the sixth coregistered optical-SAR pair by different methods. (a) Geocoding. (b) SIFT-Flow. (c) OS-Flow. (d) OSFlowNet-Ft. (a1), (b1), (c1), and (d1) are enlarged subimages from the red rectangle of (a), (b), (c), and (d), respectively. (a2), (b2), (c2), and (d2) are enlarged subimages from the blue rectangle of (a), (b), (c), and (d), respectively. (e1) and (e2) are the enlarged mosaic subimages of the registration result by the Gefolki method.

TABLE X
Computation Time (s) of Different Registration Methods

Image pair	SIFT-Flow	GeFolki	OS-Flow	OSFlowNet (CPU / GPU)	OSFlowNet-Ft (CPU / GPU)
First	288	40	101	12/3.2	342/58
Second	277	42	104	11/3.3	345/56
Third	286	38	103	11/3.3	367/65
Fourth	280	39	101	11/3.3	347/61
Fifth	276	40	101	12/3.2	348/60
Sixth	280	39	102	12/3.3	330/54

fine-tuning method is applied, the computational time is still acceptable on GPU, where about 50% of the computation time is consumed by the sparse feature point matching procedure, and about 50% consumed by the network finetuning process. While the computation time of the one-pass dense matching process is even neglectable.

V. Discussions

A. Strengths of the Proposed Optical-SAR Dense Registration Framework

The proposed OSFlowNet-Ft method is, as far as we know, the first deep framework that is applied on the optical and SAR image registration problem through pixelwise dense matching, providing an effective and robust way to deal with the problem of nonparametric geometric relationship between image pairs with elevation fluctuation.

Herein, an efficient pseudo-Siamese network architecture is designed, which incorporates a novel dilated feature concatenation strategy. Also, an effective network training strategy is used, based on a smoothed flow loss and a training dataset that contains simulated elevation fluctuations. In this way, the learned pixelwise features present higher discriminative power for similarity measurement. Furthermore, based on the observation that the blockwise matching produces better matching precision than the pixelwise way, we propose an effective self-supervised flow field fine-tuning strategy. By first obtaining a set of sparse feature point correspondences of high confidence based on the blockwise deep feature matching network and taking them as the pseudo-ground truth matches, the GRU part of the optical flow network is fine-tuned and the matching accuracy is noticeably improved.

Extensive experiments on the optical-SAR image pairs with simulated and real ground surface fluctuations validate the effectiveness and robustness of our proposed dense registration architecture. When compared with the existing representative methods, our proposed OSFlowNet-Ft framework not only substantially increases the matching accuracy, but also solves the local distortion problem to a large extent. Though our deep optical flow framework is demonstrated for 1-m resolution optical-SAR image registration, which is already a quite challenging situation, the proposed method can sure be employed for any other remote sensing image dense registration applications, no matter the single modal or multimodal types, considering the required training dataset is not large and the spatially varying optical flow fields can be easily simulated.

B. Limitations and Future Work

There are three apparent limitations of the dense registration framework. First, since the dense registration requires to obtain pixelwise matching result, it would probably fail in image areas with land surface changes. When the change only occurs within a small local region, acceptable pixelwise matching result may still be obtained by taking advantage of the smoothness characteristic of the optical flow field. However, when the change area is large, the local dense registration would be doomed to fail.

The second limitation is that it cannot deal with the rotation and scale variations. Experiments show that, when the rotation difference between the input optical and SAR image pair is larger than 5°, the EPE value would exceed three pixels. We assume that the performance decrease is caused by the enlarged feature disparity between the corresponding feature points when large rotation variance exists. It is even more sensitive to the scale variation. Acceptable registration result can only be obtained when the scale variation is within the range of [0.9, 1.1]. Otherwise, the matching accuracy would significantly decline. It probably due to that when the scale variance occurs between the optical and SAR image pair, the pixelwise features would exhibit different receptive field size, which would significantly deteriorate the similarity measurement result.

Third, similar to the other optical flow based dense matching approaches, the large displacement issue is also a significant challenge. Although the proposed OSFlowNet-Ft framework estimates the pixelwise flow vector based on the four-scaled pixelwise correlation cost volumes, the theoretically largest displacement value, which is 192 pixels for our parameter settings, can only be achieved by the coarsest scaled correlation volume C^8. At the same time, the maximum displacement value of the feature vectors collected from C^1, C^2, C^4 are 24, 48, and 96 pixels, respectively. It means that when the initial displacement value is in the range of (96, 192] pixels, only the information collected from C^8 has the potential to identify the correct matching location. At the same time, the correlation features collected from the other three scales C^1, C^2, C^4 are all distractors. In this circumstance, the probability to identify the correct matching location would be very slim. On the other hand, when the displacement value is smaller than 48 pixels, at least 3 out of the 4-scaled correlation cost vectors would contain meaningful matching information, then reliable matching results would be obtained.

Fortunately, since most of the remote sensing images contain geolocation information recently, it becomes unnecessary to deal with the large rotation and scale variations. As for the large displacement issue, it would be essential to conduct a coarse registration procedure beforehand, so as to reduce the initial displacement value to be less than 48 pixels for our proposed dense registration network, which actually is not a quite difficult goal to achieve.

A feasible solution for the general remote sensing image registration would be to properly combine the advantages of the sparse and dense registration approaches. For example, the precoarse registration can be first conducted using the sparse registration approach. Then, the dense registration method can further be applied to produce more accurate pixelwise registration result, especially for image areas with elevation fluctuations. Finally, the sparse registration can further be used to amend the wrong registration results in image areas with ground surface change. In our subsequent research, we would focus on solving the detailed problems of this research direction.

VI. Conclusion

The dense registration approach based on the optical flow estimation is not only able to get rid of the need of geometrically transforming the whole image, but also takes full advantage of the spatial smoothness of the pixelwise flow field, which is at the core of the optical flow technique. Here, in this article, we propose a robust deep optical flow framework for the optical and SAR image dense registration problem. Our main effort is put on two aspects. First, we try to better solve the brightness unconstancy issue between corresponding optical and SAR pixels, so as to meet the requirement of optical flow estimation. Second, we further improve the accuracy of the estimated optical flow field of each test instance using a self-supervised fine-tuning strategy.

The ablation study validates the effectiveness of the proposed network architecture and the network training strategy. Also, substantial progress is made by adopting the proposed self-supervised optical flow fine-tuning method during the network inference phase. Although the network is trained on a small optical-SAR dataset that contains only simulated spatially varying geometric distortions, the experiment results on the optical-SAR image pairs of real hilly or mountainous areas are outstanding and also very robust, indicating that the proposed method has the potential to be used in practical applications. In future, we would like to explore better ways to combine the dense matching and sparse matching approaches, so as to further enhance the registration accuracy, and also deal with the land surface change problem.

References

[1] A. Freeman and S. Durden, “A three-component scattering model for polarimetric SAR data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 3, pp. 963–973, May 1998.

[2] X. Li, L. Lei, Y. Sun, M. Li, and G. Kuang, “Collaborative attention-based heterogeneous gated fusion network for land cover classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 3829–3845, May 2021.

[3] Z. Ren, B. Hou, Z. Wen, and L. Jiao, “Patch-sorted deep feature learning for high resolution SAR image classification,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3113–3126, Sep. 2018.

[4] M. Eineder, C. Minet, P. Steigenberger, X. Cong, and T. Fritz, “Imaging geodesy-toward centimeter-level ranging accuracy with TerraSAR-X,” *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 661–671, Feb. 2011.

[5] N. Merkle, W. Luo, S. Auer, R. Müller, and R. Urtasun, “Exploiting deep matching and SAR data for the geo-localization accuracy improvement of optical satellite images,” *Remote Sens.*, vol. 9, no. 6, Jun. 2017, Art. no. 586.

[6] Y. Bentoutou, N. Taleb, K. Kpalma, and J. Rosin, “An automatic image registration for applications in remote sensing,” *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 9, pp. 2127–2137, Sep. 2005.

[7] Y. Ye, T. Tang, B. Zhu, C. Yang, B. Li, and S. Hao, “A multiscale framework with unsupervised learning for remote sensing image registration,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5622215.

[8] J. Zaragoza, T. Chin, Q. Tran, M. Brown, and D. Suter, “As-projective-as-possible image stitching with moving DLT,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1285–1298, Jul. 2014.

[9] R. Feng, Q. Du, H. Shen, and X. Li, “Region-by-region registration combining feature-based and optical flow methods for remote sensing images,” *Remote Sens.*, vol. 13, no. 8, 2021, Art. no. 1475.

[10] J. Curlander, “Geometric and radiometric distortion in spaceborne SAR imagery,” in *Proc. NASA Workshop Registration Rectification*, 1982, pp. 163–197.

[11] 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, no. 99, pp. 3296–3310, Jan. 2020.

[12] Q. Yu, D. Ni, Y. Jiang, Y. Yan, J. An, and T. Sun, “Universal SAR and optical image registration via a novel SIFT framework based on nonlinear diffusion and a polar spatial-frequency descriptor,” *ISPRS J. Photogramm. Remote Sens.*, vol. 171, pp. 1–17, Jan. 2021.

[13] Y. Ye, L. Shen, M. Hao, J. Wang, and Z. Xu, “Robust optical-to-SAR image matching based on shape properties,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 564–568, Apr. 2017.

[14] S. Suri and P. Reinartz, “Mutual-information-based registration of TerraSAR-X and Ikonos imagery in urban areas,” *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 2, pp. 939–949, Feb. 2010.

[15] D. Konstantinidis, T. Stathaki, and V. Argyriou, “Phase amplified correlation for improved sub-pixel motion estimation,” *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3089–3101, Jun. 2019.

[16] S. Li, X. Lv, J. Ren, and J. Li, “A Robust 3D density descriptor based on histogram of oriented primary edge structure for SAR and optical image co-registration,” *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 630.

[17] Y. Ye, B. Zhua, T. Tang, C. Yang, Q. Xu, and G. Zhang, “A robust multimodal remote sensing image registration method and system using steerable filters with first- and second-order gradients,” *ISPRS J. Photogramm. Remote Sens.*, vol. 188, pp. 331–350, Jun. 2022.

[18] Y. Ye, L. Bruzzone, J. Shan, F. Bovolo, and Q. Zhu, “Fast and robust matching for multimodal remote sensing image registration,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9059–9070, Nov. 2019.

[19] M. P. Heinrich et al., “MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration,” *Med. Image Anal.*, vol. 16, no. 7, pp. 1423–1435, Oct. 2012.

[20] H. Zhang, W. Ni, W. Yan, D. Xiang, and J. Wu, “Registration of multimodal remote sensing image based on deep fully convolutional neural network,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 3018–3042, Aug. 2019.

[21] L. Hughes, S. L. D. Marcos, D. Tuia, and M. Schmitt, “A deep learning framework for matching of SAR and optical imagery,” *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 166–179, Sep. 2020.

[22] H. Zhang et al., “Optical and SAR image matching using pixelwise deep dense features,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Dec. 2020, Art. no. 6000705.

[23] H. Zhang et al., “Explore better network framework for high-resolution optical and SAR image matching,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Nov. 2021, Art. no. 4704418.

[24] T. Bürgmann, W. Koppe, and M. Schmitt, “Matching of TerraSAR-X derived ground control points to optical image patches using deep learning,” *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 241–248, Nov. 2019.

[25] Y. You, C. Li, and W. Zhou, “DRFD-Net: Using dual receptive field descriptors for multitemporal optical remote sensing image registration,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 5610319.

[26] L. Hughes, M. Schmitt, and X. X. Zhu, “Mining hard negative samples for SAR-optical image matching using generative adversarial networks,” *Remote Sens.*, vol. 10, no. 10, Sep. 2018, Art. no. 1552.

[27] N. Merkle, S. Auer, R. Müller, and P. Reinartz, “Exploring the potential of conditional adversarial networks for optical and SAR image matching,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 6, pp. 1811–1820, Jun. 2018.

[28] L. Hughes, N. Merkle, S. Auer, and M. Schmitt, “Deep learning for SAR-optical image matching,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 4877–4880.

[29] S. Hoffmann, C. Brust, M. Shadaydeh, and J. Denzler, “Registration of high resolution SAR and optical satellite imagery using fully convolutional networks,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5152–5155.

[30] Y. Fang, J. Hu, C. Du, Z. Liu, and L. Zhang, “SAR-optical image matching by integrating Siamese U-Net with FFT correlation,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Aug. 2022, Art. no. 4016505.

[31] L. Hughes and M. Schmitt, “Comparative evaluation of deep learning-based SAR-optical image matching approaches,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 1–4.

[32] D. Xiang, Y. Xie, J. Cheng, Y. Xu, H. Zhang, and Y. Zheng, "Optical and SAR image registration based on feature decoupling network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5235913.

[33] W. Wu, Y. Xian, J. Su, and L. Ren, "A Siamese template matching method for SAR and optical image," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Sep. 2022, Art. no. 4017905.

[34] Z. Zhang, Y. Xu, Q. Cui, Q. Zhou, and L. Ma, "Unsupervised SAR and optical image matching using Siamese domain adaptation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Apr. 2022, Art. no. 5227116.

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

[36] H. Bay, A. Ess, and T. Tuytelaars, "SURF: Speeded-up robust features," *Comput. Vis. Image Understanding*, vol. 110, no. 3, pp. 404–417, 2006.

[37] C. Harris and M. Stephens, "A combined corner and edge detector," *Alvey Vis. Conf.*, vol. 23, pp. 147–151, 1988.

[38] A. Bruhn, J. Weickert, and C. Schnorr, "Lucas/Kanade meets Horn/Schunk: Combining local and global optical flow methods," *Int. J. Comput. Vis.*, vol. 61, no. 3, pp. 211–231, 2005.

[39] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2010, pp. 1–31.

[40] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[41] D. Sun, X. Yang, M. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8934–8943.

[42] G. Brigot, E. Colin-Koeniguer, A. Plyer, and F. Janez, "Adaptation and evaluation of an optical flow method applied to coregistration of forest remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 7, pp. 2923–2939, Jul. 2016.

[43] Y. Xiang, F. Wang, L. Wan, N. Jiao, and H. You, "OS-Flow: A robust algorithm for dense optical and SAR image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6335–6354, Sep. 2019.

[44] I. Damian, F. Ruben, and P. Filiberto, "A remote sensing image registration benchmark for operational Sentinel-2 and Sentinel-3 products," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 2246–2249.

[45] C. Elise, "An optical flow method applied to co-registration of remote sensing images: Example for SAR/SAR, SAR/LIDAR, SAR/Optical images of BIOSAR," in *Proc. Living Planet Symp.*, 2016, pp. 1–5.

[46] Z. Petrou and Y. Tian, "High-resolution sea ice motion estimation with optical flow using satellite spectroradiometer data," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1339–1350, Mar. 2017.

[47] M. Zelinski, J. Henderson, and E. Held, "Image registration and change detection for artifact detection in remote sensing imagery," in *Proc. Conf. Defense+Secur. Environ. Sci., Math.*, 2018, pp. 1–18.

[48] K. Cho, B. Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proc. 8th Workshop Syntax, Semantics Struct. Stat. Transl.*, 2014, pp. 103–111.

[49] M. Longuet-Higgins, "The statistical distribution of the curvature of a random Gaussian surface," *Proc. Math. Cambridge Philos. Soc.*, vol. 54, no. 4, pp. 439–453, Oct. 1958.

[50] X. Jiang et al., "Robust feature matching for remote sensing image registration via linear adaptive filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1577–1591, Feb. 2021.

[51] J. Ma, Z. Li, K. Zhang, Z. Shao, and G. Xiao, "Robust feature matching via neighborhood manifold representation consensus," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 196–209, 2022.

[52] J. Bian, W. Lin, Y. Matsushita, S. Yeung, T. Nguyen, and M. Cheng, "GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2828–2837.

[53] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. Joint Conf. Artif. Intell.*, 1981, pp. 674–679.

[54] B. Horn and B. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, pp. 185–203, Aug. 1981.

[55] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2011.

[56] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "EpicFlow: Edge-preserving interpolation of correspondences for optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1164–1172.

[57] C. Bailer, B. Taetz, and D. Stricker, "Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4015–4023.

[58] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2758–2766.

[59] A. Ranjan and M. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4161–4170.

[60] T. Zachary and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 402–419.

[61] P. Truong, M. Danelljan, L. Gool, and R. Timofte, "GOCor: Bringing globally optimized correspondence volumes into your neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1–35.

[62] W. Jiang, E. Trulls, J. Hosang, A. Tagliasacchi, and K. Yi, "COTR: Correspondence transformer for matching across images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 6207–6217.

[63] D. Fedorov, L. Fonseca, C. Kenney, and B. Manjunath, "Automatic registration and mosaicking system for remotely sensed imagery," in *Proc. Int. Symp. Remote Sens.*, 2002, pp. 1–8.

[64] Y. Ma, F. Chen, J. Liu, Y. He, J. Duan, and X. Li, "An automatic procedure for early disaster change mapping based on optical remote sensing," *Remote Sens.*, vol. 8, no. 4, 2016, Art. no. 272.

[65] F. Xu, H. Yu, J. Wang, and W. Yang, "Accurate registration of multitemporal UAV images based on detection of major changes," in *Proc. Int. Conf. Inf. Fusion*, 2018, pp. 1480–1485.

[66] Y. Ri and H. Fujimoto, "Drift-free motion estimation from video images using phase correlation and linear optimization," in *Proc. IEEE Int. Workshop Adv. Motion Control*, 2018, pp. 295–300.

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

[68] A. Zampieri, G. Charpiat, N. Girard, and Y. Tarabalka, "Multimodal image alignment through a multiscale chain of neural networks with application to remote sensing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2018, pp. 657–673.

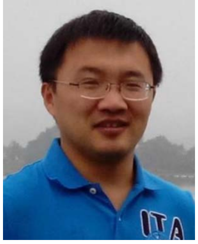
[69] Y. Xiang, R. Tao, F. Wang, H. You, and B. Han, "Automatic registration of optical and SAR images via improved phase congruency model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5847–5861, 2020.

Han Zhang received the B.S. degree in electronics science and technology from the Shanghai Jiao Tong University, Shanghai, China, in 2010, and the M.S. degree in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2012. She is currently working toward the Ph.D. degree in information and communication engineering with the College of Electronic Science, National University of Defense Technology, Changsha, China.

She is currently working as a Research Associate with the Northwest Institute of Nuclear Technology, Xi'an, China. Her research interests include multitemporal and multimodal remote sensing image analysis, pattern recognition, and deep learning.

Lin Lei received the Ph.D. degree in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2008.

She is currently a Professor with the School of Electronic Science, National University of Defense Technology. Her research interests include computer vision, remote sensing image interpretation, and data fusion.



Weiping Ni received the B.S. degree in electronics science and technology from the University of Science and Technology of China, Hefei, China, in 2004, the M.S. degree in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent system from the Xidian University, Xi'an, China, in 2016.

Since 2014, he has been a Research Associate with the Northwest Institute of Nuclear Technology, Xi'an, China. His research interest includes remote sensing image processing, automatic target recognition, and computer vision.

Xiaoliang Yang received the B.S. degree in remote sensing science and technology from the Wuhan University, Wuhan, China, in 2010, and the M.S. and Ph.D. degrees in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2013 and 2017, respectively.

He is currently working as a Research Associate with the Northwest Institute of Nuclear Technology, Xi'an, China. His research interests include remote sensing image analysis and computer vision.

Tao Tang received the B.S., M.S., and Ph.D. degrees in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2002, 2006, and 2016, respectively.

He has been an Associate Professor with the School of Electronic Science, National University of Defense Technology. His research interests include synthetic aperture radar target detection and recognition, and remote sensing image registration and feature extraction.

Kenan Cheng received the B.S. and M.S. degrees in information and communications engineering from the Xi'an Jiaotong University, Xi'an, China, in 2016 and 2019, respectively.

She is currently working as a Research Associate with the Northwest Institute of Nuclear Technology, Xi'an, China. Her research interests include remote sensing image analysis and deep learning.

Deliang Xiang (Member, IEEE) received the B.S. degree in remote sensing science and technology from the Wuhan University, Wuhan, China, in 2010, the M.S. degree in photogrammetry and remote sensing from the National University of Defense Technology, Changsha, China, in 2012, and the Ph.D. degree in geoinformatics from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2016.

In 2019, he was awarded a Humboldt Research Fellowship. Since 2020, he has been a Full Professor with the Interdisciplinary Research Center for Artificial Intelligence, Beijing University of Chemical Technology, Beijing, China. His research interests include urban remote sensing, synthetic aperture radar/polarimetric SAR image processing, artificial intelligence, and pattern recognition.

Prof. Xiang serves as a Reviewer for the *Remote Sensing of Environment*, the *ISPRS Journal of Photogrammetry And Remote Sensing*, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and several other international journals in the remote sensing field.

Gangyao Kuang (Senior Member, IEEE) received the B.S. and M.S. degrees in geophysics from the Central South University of Technology, Changsha, China, in 1988 and 1991, respectively, and the Ph.D. degree in communication and information from the National University of Defense Technology, Changsha, China, in 1995.

He is currently a Professor with the School of Electronic Science, National University of Defense Technology. His research interests include remote sensing, SAR image processing, change detection, SAR ground moving target indication, and classification with SAR images.