Unsupervised Image Registration for Video SAR

Xuejun Huang, Jinshan Ding, and Qinghua Guo, *Senior Member, IEEE*

***Abstract*—Existing approaches for SAR image registration focus on the global transformation correction between SAR images. However, there are often local deformations between images. Due to the time-changing viewpoint of video SAR, the images suffer a lot from local deformations, which can result in false alarms in moving target detection. This article presents an unsupervised image registration approach for the use of video SAR moving target detection, which has good registration performance and acceptable processing efficiency. The designed unsupervised learning-based framework is a cascade of two convolutional neural networks. The first network directly predicts the parameters of the rigid transformation between the reference and unregistered images, and recovers the global transformation between them. Then, the second network uses the reference image and the registered image from the first network as input and then predicts a displacement field. After that, we put a limitation on the predicted displacement field to prevent moving target shadows from being aligned. Finally, the displacement field with limitation is used to compensate local deformations between the two images. Processing results of real video SAR images have shown good performance of the proposed approach with convincing generation ability.**

***Index Terms*—Image registration, local deformations, moving target detection, moving target detection, unsupervised learning, video synthetic aperture radar (SAR).**

I. Introduction

VIDEO synthetic aperture radar (SAR) has received a lot of research attention [1]–[3] recently, which provides a persistent view of a scene of interest by forming high frame rate sequential images [4]. It allows for effective detection and tracking of moving targets [5], [6], where target shadows in sequential radar images can be used to detect moving targets [7]. Some methods have been developed for moving target detection in video SAR [8], [9], which use the information contained in successive frames. Image registration is always used to compensate for background change between the frames [8], [9], which plays an important role in the video SAR moving target detection.

The conventional image registration methods can be roughly grouped into intensity-based methods [10]–[14] and feature-based methods [15]–[18]. Intensity-based methods recover the transformation between two images by maximizing an image similarity, such as cross correlation or mutual information. However, these methods are generally associated with high computational loads. Alternatively, feature-based methods show higher precision and effectiveness, which basically consist of three steps: feature extraction, feature matching, and transformation parameters estimation. First, salient and distinctive features are extracted from two images. Then, the corresponding features are identified by a matching technique. Finally, the geometric transformation parameters are estimated by using the correct feature correspondences. Besides, some methods combine the advantages of intensity-based and feature-based methods for robust image registration [19], [20].

Most of the conventional registration methods just recover the global transformation between SAR images by estimating the parameters of a transformation model, such as rigid transformation, similarity transformation, or affine transformation. Unfortunately, although the global transformation is corrected, there are often local deformations between two SAR images from different viewpoints. Since the video SAR system often works in the spotlight mode, its viewpoint varies with frame continuously [21]. As a result, video SAR images suffer a lot from local deformations, significantly adding to the difficulty in moving target detection.

Recently, convolutional neural networks (CNNs) have been successfully used in image registration. In [22], image registration is regarded as a regression task, and a CNN is trained to predict a transformation matrix for the rigid registration of synthetic images. In [23], a CNN is used to learn the mapping from a pair of input images to an output deformation filed. However, all these methods, which are based on supervised learning, have to rely on ground truths that come from simulation or the conventional registration methods. Most recently, some unsupervised CNN-based methods have been developed for image registration [24]–[26]. They estimate the deformation filed between images by optimizing an image similarity often combined with a smoothing constraint. However, these methods focus on medical image registration.

Video SAR provides image sequences of a region of interest at a high frame rate, indicating that the radiometric difference between video SAR images is very small. Therefore, we can maximize an image similarity to estimate the deformations between video SAR images via a displacement field. However, the motion caused by different viewpoints results in a relatively large displacement field between video SAR images, and it is difficult to accurately estimate the displacement field by a CNN. Furthermore, when the CNN is used to estimate the displacement field directly, moving target shadows in two images will be aligned, which leads to missing alarms in moving target detection.

Manuscript received August 10, 2020; revised October 7, 2020; accepted October 16, 2020. Date of publication October 21, 2020; date of current version January 6, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFE0200400. *(Corresponding author: Jinshan Ding.)*

Xuejun Huang and Jinshan Ding are with the National Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China (e-mail: xjhuang@stu.xidian.edu.cn; ding@xidian.edu.cn).

Qinghua Guo is with the School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong 2522, Australia (e-mail: qguo@uow.edu.au).

Digital Object Identifier 10.1109/JSTARS.2020.3032464

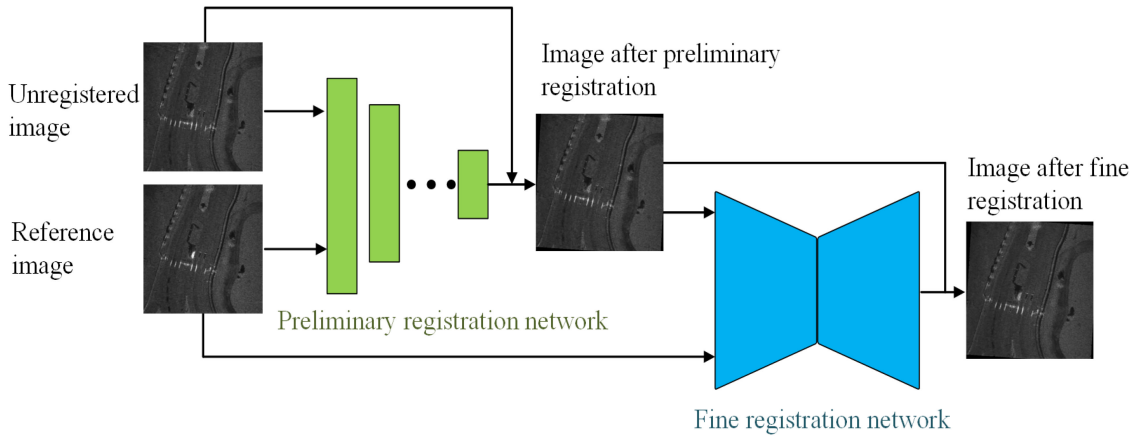


Fig. 1. Unsupervised framework developed for video SAR image registration.

This article presents an unsupervised learning-based framework for video SAR image registration, where a coarse-to-fine strategy is adopted to register two video SAR images. Specifically, image registration consists of the global transformation correction and the local deformation compensation, and the designed framework is a cascade of two CNNs. We correct the global transformation to compensate for the motion by the first CNN, and accurately estimate the residual displacement field for the local deformation compensation by the second CNN. Since the remained displacement of stationary targets is smaller than that of moving target shadows in two images, we put a limitation on the estimated displacement field to protect moving target shadows from being aligned.

The rest of this article is organized as follows. Section II details the complete framework for fine image registration. Experimental results are presented in Section III, finally Section IV concludes this article.

II. Methodology

An unsupervised learning-based framework has been developed for both global transformation correction and local deformation compensation, which is a cascade of a preliminary registration network and a fine registration network, as shown in Fig. 1. First, given a reference image I_r and an unregistered image I_u, the preliminary registration network estimates the transformation model parameters between the two images. Second, according to the parameters, the image after preliminary registration I_p can be obtained by warping the unregistered image to the reference image. Third, the fine registration network takes a pair of the reference image I_r and the image I_p as input and predicts a displacement field, which represents the displacement of the corresponding pixels in two images. Then, we put a limitation on the predicted displacement field to prevent moving target shadows from being aligned. Finally, based on the displacement field, we obtain the image after fine registration I_f by wrapping the image I_p to the reference image I_r.

A. Global Transformation Correction

The pixel displacement is usually large due to the global transformation between the reference and unregistered images, and thus, it is difficult to predict the displacement for compensating local deformations directly by using a CNN. Therefore, the global transformation between images should be corrected before local deformation compensation.

The preliminary registration network is similar to the conventional image registration. It estimates the parameters of the transformation model between images, and then recovers the global transformation according to these parameters. Different from the conventional image registration methods, we use the CNN to estimate the transformation model parameters directly. Since the sizes of different images in video SAR image sequences are the same, the motion between two images can be seen as a rigid transformation. Hence, the preliminary registration network only needs to predict the horizontal translation Δx, vertical translation Δy and rotation angle θ between two images. Then, the image after preliminary registration can be obtained by bilinear interpolation, which is given as

$$\begin{cases} x_2 = x_1 \cos\theta + y_1 \sin\theta + \Delta x \\ y_2 = -x_1 \sin\theta + y_1 \cos\theta + \Delta y \end{cases} \tag{1}$$

where (x_1, y_1) and (x_2, y_2) represent 2-D location in the planes of the unregistered image and the image after preliminary registration, respectively. It should be pointed out that the bilinear interpolation can make the preliminary registration network fully differentiable. Finally, the preliminary registration network is trained by minimizing the image difference between the image after preliminary registration and the reference image.

The preliminary registration network takes a concatenated pairs of the reference and unregistered images as input and outputs the parameters of the transformation model between the two images, as shown in Fig. 2. The preliminary registration network is a fully convolutional neural network, which consists of eight convolution layers and a global average pooling layer. The kernel size of the first two convolution layers are 7×7 and 5×5, respectively, and all the others are 3×3. The filter numbers of eight convolution layers are 32, 64, 128, 256, 256, 256, 256, and 3, respectively. The stride of the last convolution layer is 1, and all the others are 2. Throughout the network, a rectified linear unit (ReLU) is used for activation, except for the final convolution layer, which has a linear output.

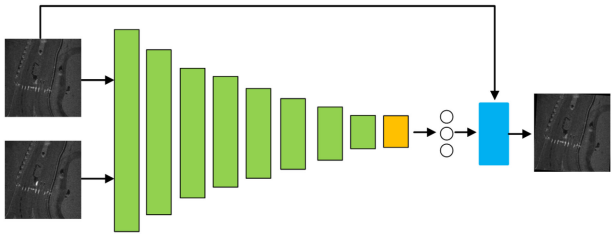


Fig. 2. Structure of the preliminary registration network, where the green, orange, and blue rectangle represents the convolution layer, the global average pooling layer, and bilinear interpolation, respectively. The circles denote the estimated parameters of the transformation model.

The cross correlation (CC) is frequently used to measure image similarity. A higher CC between the reference image I_r and the image after preliminary registration I_p indicates a better alignment, and hence, the loss function of the preliminary registration network can be defined as

$$Loss_1 = 1 - CC(I_r, I_p). \tag{2}$$

B. Local Deformation Compensation

After global transformation correction, there are still some local deformations remained between images. For local deformation compensation, we first use the fine registration network to learn the complex nonlinear mapping from a pair of the reference image and the image after preliminary registration to a displacement field D. The displacement field D consists of two 2-D matrix D_x and D_y, which represent the horizontal and vertical displacement of pixels corresponding to the same target in two images, respectively. Then, according to the displacement field D, the image after preliminary registration I_p is deformed to match the reference image I_r by bilinear interpolation, and the image after fine registration I_f is obtained, which can be expressed as

$$I_f(x, y) = I_p(x + D_x(x, y), y + D_y(x, y)) \approx I_r(x, y) \tag{3}$$

where (x, y) represents a 2-D location in the image plane. Finally, the fine registration network is optimized by minimizing the image difference between the image after fine registration and the reference images.

The fine registration network takes a concatenated pair of the reference image and the image after preliminary registration as input and outputs a displacement field, as shown in Fig. 4, which consists of an encoder module and a decoder module. The encoder extracts features from the input image. It has the basic structure of ResNet-50 [28], which has been found to have a good performance in feature representation. Since the original ResNet-50 is designed for image recognition, we slightly modify it for image registration. We remove the ending average pooling and the full connection layer, and the filters of all convolution layers are reduced by half, and thus the parameters and computational cost decrease. The decoder module outputs the predicted displacement field from features, which consists of six deconvolution layers to enlarge the spatial feature maps to full scale as input. The kernel size of the last deconvolution layers is 7×7, and the others are 3×3. The filter numbers of six deconvolution layers are 512, 256, 128, 64, 32, and 2, respectively. To combine both high-level and low-level features, we use skip connections between the encoder and the decoder module at different resolutions.

The loss function of the fine registration network includes local intensity loss L_i, structural loss L_s, and smoothness loss L_f, which can be written as

$$\begin{aligned} Loss_2 &= L_i + \alpha L_s + \beta L_d \\ L_i &= 1 - LCC(M_p(I_r), M_p(I_f)) \\ L_s &= \|Sobel(M_p(I_r)) - Sobel(M_p(I_f))\| \\ L_f &= \|\partial_x D\| e^{-\|\partial_x M_p(I_r)\|} + \|\partial_y D\| e^{-\|\partial_y M_p(I_r)\|} \end{aligned} \tag{4}$$

where LCC denotes local cross correlation [25] which indicates local image similarity between the reference image I_r and the image after fine registration I_f. $Sobel$ represents the edge detection by using Sobel operator. α and β are hyper-parameters, which are set to 10 and 0.1 by default. M_p denotes the mean filter with 3×3 kernel size, which can mitigate the effect of the speckle noise. ∂_x and ∂_y represent partial derivatives along horizontal and vertical directions. The structural loss and the local intensity loss encourage the fine registration image to appear similar to the reference image. The smoothness loss L_f is generally needed to encourage the estimated displacement field to be locally smooth. Since local deformations often occur at the edge, there is large displacement in the image with a large gradient. As a result, in the smoothness loss, the gradients of the displacement field are weighted by image gradients.

To facilitate moving target detection, moving target shadows in two images should not be aligned. After global transformation correction, the remained displacement of stationary targets is smaller than that of moving target shadows in two images. Therefore, we put a limitation on the displacement field D_1 predicted by the fine registration network, which makes the displacement of dark areas small, and the final displacement field D is given as

$$D(x, y) = \begin{cases} D_1(x, y), & I_p(x, y) > I_p^m \text{ or } \|D_1(x, y)\| < \gamma \\ 0, & \text{others} \end{cases} \tag{5}$$

where I_p^m represents the mean value of the image I_p. γ is a threshold of displacement, which is set to 3 by default.

III. Experimental Results

The proposed unsupervised image registration approach has been used to process the real video SAR data released by Sandia National Laboratory. The generalization ability of the proposed approach is discussed, which becomes a concern on deep learning algorithms when applied in radar applications.

A. Datasets and Training Strategy

The datasets that we built based on the released Sandia data include a training set and two testing sets, and one of the test sets corresponds to the same scenario as the training set and the other is different from the training set. The two testing sets

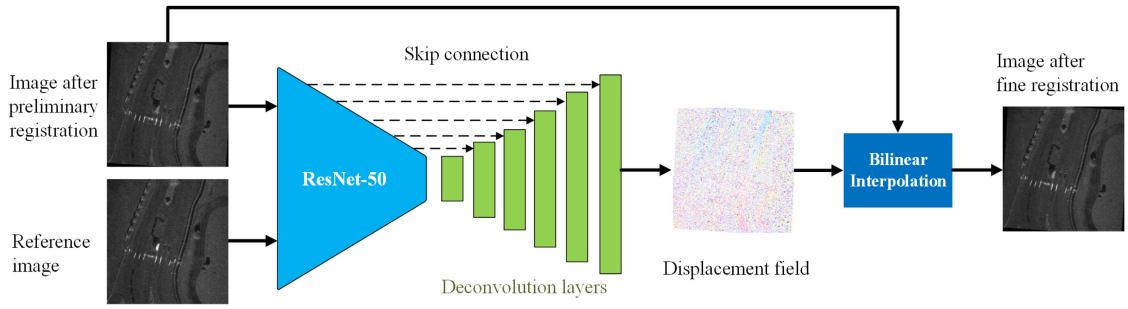


Fig. 3. Structure of the fine registration network.

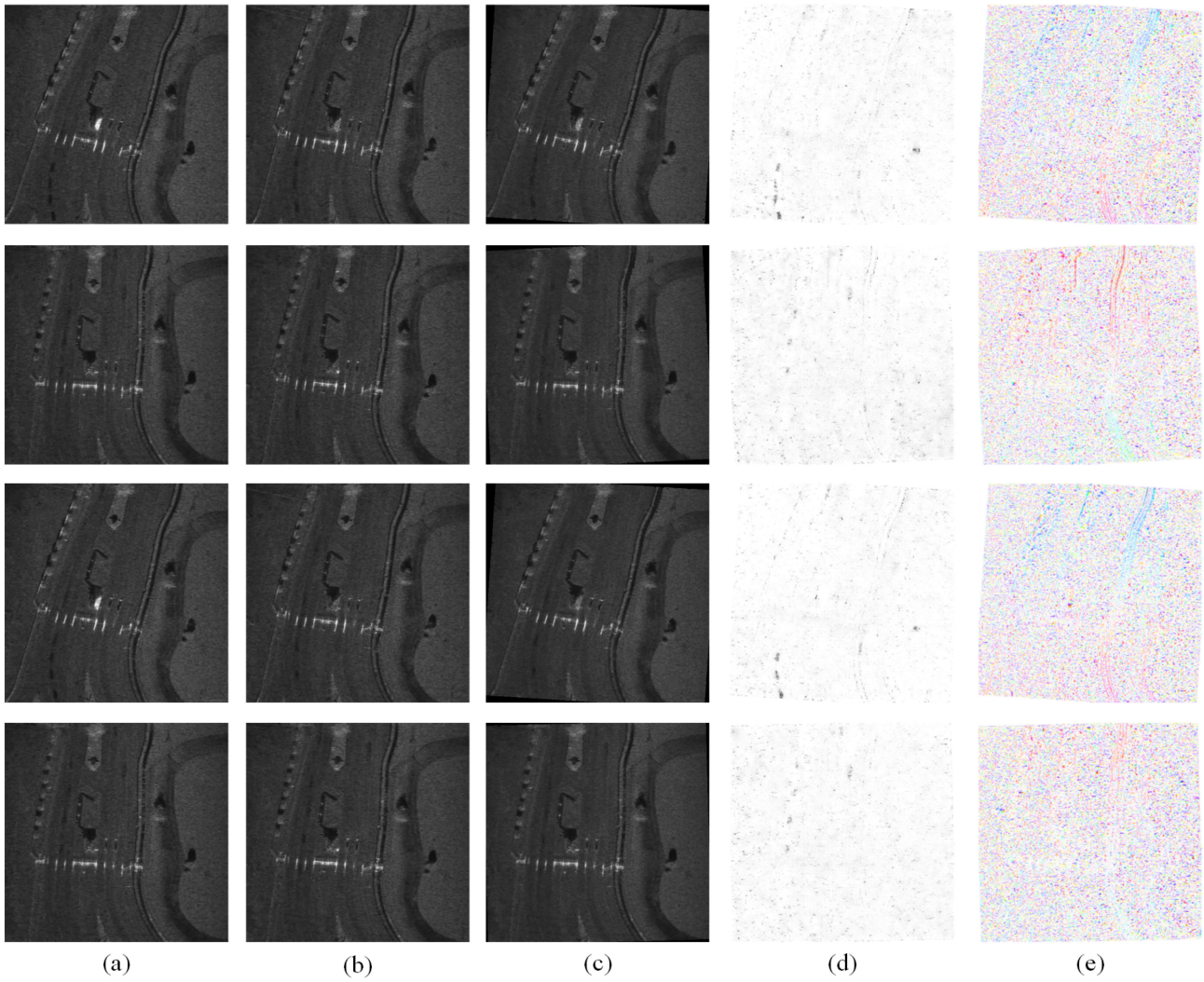


Fig. 4. Processing results of the proposed approach on four representative image pairs from the SAR video of Eubank Gate. (a) Reference images. (b) Unregistered images. (c) Registered images were obtained by the proposed approach. (d) Ratio images between the reference images and the registered images, which are calculated as a pixel to pixel ratio between the two images. (e) Estimated displacement fields which are visualized using the standard optical-flow visualization [29].

are used to verify the image registration performance of the proposed approach, under the condition that the training scenario is consistent and inconsistent with the test scenario, respectively.

Two different SAR videos are used. The first radar video of Eubank Gate contains 900 frames with a size of 720×640 pixels, and the first 100 frames are for testing and the remaining 800 frames are used to build the training set. The original training set is augmented by cutting and rotating as we usually do in preparing the datasets for deep learning applications, providing total 80 000 images with a size of 512×512 pixels. The original 100 frames for testing are cut into 100 images with a size of 512×512 pixels. The first testing set are composed

of the 100 cropped images. The second testing set consists of 55 consecutive frames from the second SAR video. In the 55 consecutive frames, each frame contains a SAR image and a range-Doppler spectrum, and moving targets in the SAR image are marked by green squares near them. Therefore, these frames cannot be directly used for testing, and an image preprocessing has been done to remove the green square markings in the 55 frames and the range-Doppler spectrum. All images in datasets are resized to 512×512.

The proposed deep learning framework is trained in two stages. In the first place, the preliminary registration network is optimized by Adam optimizer with a batch size of 16. The learning rate is set to 2×10^{-4} and 2×10^{-5} in the first 80 000 and the next 20 000 iterations, respectively. After that, the weights in the preliminary registration network is fixed, and the fine registration network is trained by Adam optimizer with a batch size of 2. The learning rate is set to 2×10^{-4} and 2×10^{-5} in the first 80 000 and the next 20 000 iterations, respectively.

B. Evaluation Metrics

We use the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) [27] to quantitatively assess the performance of the proposed approach. The PSNR and the SSIM measure the similarity between the reference image and registered image, and higher PSNR or SSIM indicates better performance in image registration. It should be pointed out that the reference and registered images are first denoised by the Lee filter [30], a common despeckling algorithm for SAR images, to mitigate the effect of the speckle noise on the performance assessment.

Additionally, a better method for video SAR image registration should be more helpful for moving target detection. Therefore, we employ a moving target detection method to process the registration results, and evaluate the detection performance via false alarms and missing alarms. More false alarms and missing alarms reveal worse performance in moving target detection, namely, worse image alignment. The detection method reported in [9] is used, which consists of image registration, constant false alarm rate (CFAR) detection, and morphological processing. More specifically, given a current image and its nearby images in a video SAR image sequence, the nearby images are aligned to the current frame by an image registration method. Subsequently, the registered nearby images are used to calculate a reference image which represents the static background. After that, the current image is divided by the reference image to yield a ratio image, which highlights moving target shadows and suppresses the static scene. Then, the CFAR is performed on the ratio image to detect moving targets. Finally, morphological processing is employed to suppress the missing and false alarms.

C. Registration Results

We compare the proposed approach with three state-of-the-art image registration approaches including SAR-SIFT, locally linear transforming (LLT) [31], and locality preserving matching (LPM) [32]. The SAR-SIFT is a classical and well-known algorithm for SAR image registration. The LPM and the LLT focus on feature matching, which remove mismatches from given putative image feature correspondences for robust image registration.

TABLE I
PSNR AND SSIM RESULTS ON 500 IMAGE PAIRS FORM THE FIRST TESTING SET

Methods	PSNR	SSIM
SAR-SIFT	31.1174	0.8485
LPM	31.8002	0.8597
LLT	30.2555	0.8473
Ours	39.8345	0.9771

TABLE II
FALSE ALARM AND MISSING ALARMS ON THE IMAGES FROM THE FIRST TESTING SET

Methods	False Alarms	Missing Alarms
SAR-SIFT	263	72
LPM	193	52
LLT	159	65
Ours	91	47

Two images are randomly selected from the testing set that corresponds to the same scenario as the training set, which repeats 500 times to yield 500 image pairs. The proposed approach is used to process the 500 image pairs, and the registration results on four representative image pairs are given in Fig. 4. One of the representative results is compared with those of the SAR-SIFT, the LLT, and the LPM, as shown in Fig. 5. It is observable in Fig. 5 that the proposed approach has the best ability to compensate for local deformations compared to these conventional methods. Furthermore, Table I lists the assessment results of different methods on these 500 image pairs. As revealed by the PSNR and SSIM values, the SAR-SIFT and the LPM perform poorly in video SAR image registration because they focus on correcting the global affine transformation. Meanwhile, the LLT is not suitable for video SAR image registration although it can estimate the displacement field between images by a displacement function. The LLT assumes that the displacement function lies within a specific functional space. However, this assumption cannot hold in some cases. By contrast, the proposed approach shows the best performance in video SAR image registration.

We detect moving targets on the 100 sequential images in testing set that corresponds to the same scenario as the training set. During the detection, different registration approaches are used to align each image and its adjacent images. The detection results are given in Table II. It can be seen that the proposed approach has the fewest false alarms and missing alarms. Therefore, compared to the conventional image registration methods, the proposed approach is more helpful for moving target detection. However, there are still a few missing alarms and false alarms in the detection results using the registration results of the proposed approach, which will resort to advanced detection methods.

Additionally, we compare the proposed approach with the conventional methods in terms of processing efficiency, as shown in Table III. The conventional methods and the proposed

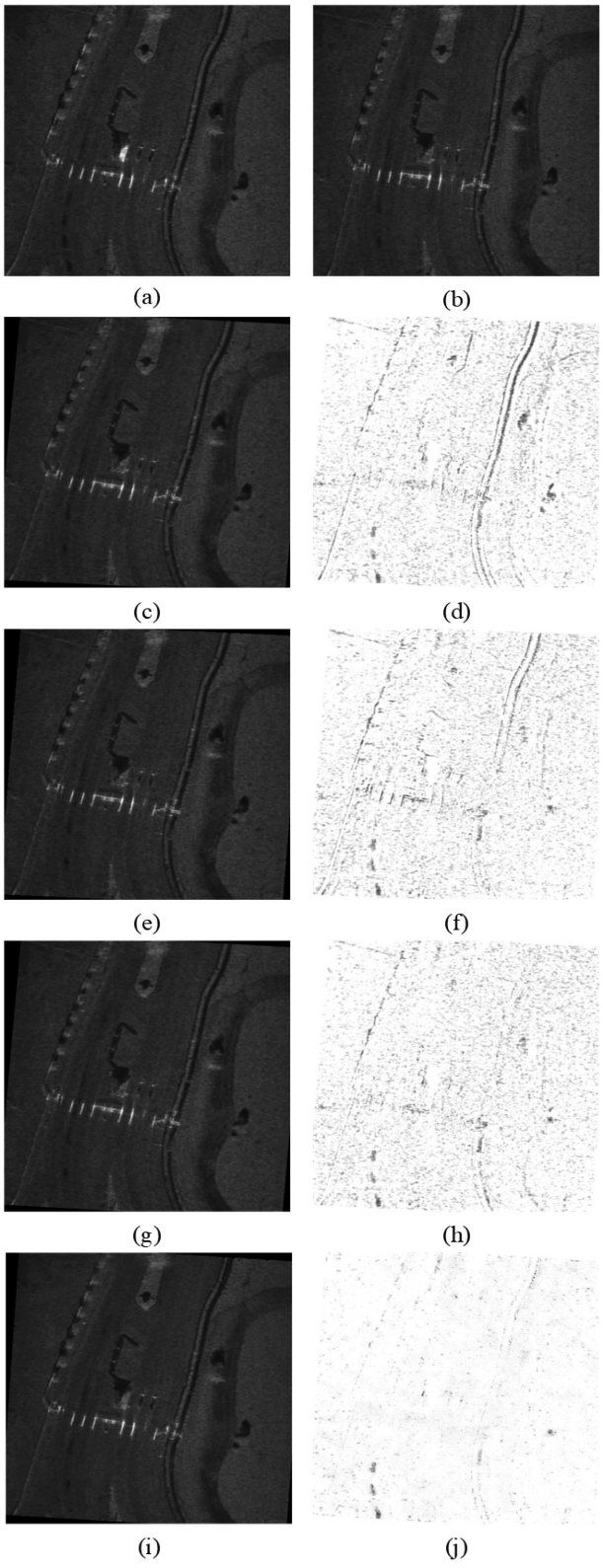


Fig. 5. Registration results of a representative image pair from the SAR video of Eubank Gate by using different approaches. (a) Reference image. (b) Unregistered image. (c), (e), (g), and (i) are registered images obtained by the SAR-SIFT, LLT, LPM, and the proposed approach, respectively. (d), (f), (h), and (j) are ratio images between the reference image and the registered images obtained by the SAR-SIFT, LLT, LPM, and the proposed approach, respectively.

TABLE III
RUNTIME OF DIFFERENT REGISTRATION APPROACHES APPLIED TO A 512 × 512 RADAR IMAGE PAIR

Methods	SAR-SIFT	LPM	LLT	Ours
Runtime on CPU	2.6 s	0.1 s	0.9 s	1.2 s
Runtime on GPU	–	–	–	0.02 s

TABLE IV
THE ASSESSMENT RESULTS ON THE FIRST TESTING SET IN TWO CASES

Cases	Without Limitation	Under Limitation
PSNR	42.1831	39.8345
SSIM	0.9880	0.9771
False Alarms	65	91
Missing Alarms	183	47

TABLE V
ASSESSMENT RESULTS ON THE FIRST TESTING SET BY USING DIFFERENT REGISTRATION STRATEGIES

Strategies	One-step	Coarse-to-fine
PSNR	38.3475	39.8345
SSIM	0.9747	0.9771
False Alarms	313	91
Missing Alarms	202	47

approach are tested in Matlab R2018a and Tensorflow 1.13.1, respectively, on a platform with an Intel E5-2650 v4 CPU and one Nvidia Titan Rtx GPU. When running on the CPU, the developed CNN-based approach takes 1.2 s to register a 512 × 512 image pair. Although the efficiency of the proposed approach is not the best, it is acceptable.

D. Ablation Study

To examine the influence of the displacement limitation, the proposed deep learning framework is trained without this limitation. We compare the performance of the developed framework under limitation and without limitation, as shown in Table IV. Obviously, the limitation leads to a slight decrease in image similarity between the reference and registered images, but it significantly improves the detection performance.

An experiment reveals the superiority of the adopted coarse-to-fine strategy for video SAR image registration. A CNN is used to directly estimate the displacement field between video SAR images, which only needs one step for image registration as many medical image registration methods. For fair comparison, the CNN shares the same network architecture and loss function with the developed fine registration network. The performance of the CNN is compared with that of the proposed approach, which is given in Table V. In addition, Fig. 6 exhibits the registration results of the CNN and the proposed approach on a representative image pair. Form both the quantitative assessment results and the visual performance, the proposed approach shows a better performance in image registration. Especially in visual results, the CNN brings in the artifacts which make the registered images

TABLE VI
ABLATION STUDY OF HYPER-PARAMETER SETTING

Parameters(default marked by bold)	α			β			γ		
	8	**10**	12	0.05	**0.1**	0.2	2	**3**	4
PSNR	39.8383	39.8345	39.8570	40.0492	39.8345	39.5725	37.9143	39.8345	41.1019
SSIM	0.9769	0.9771	0.9769	0.9775	0.9771	0.9748	0.9616	0.9771	0.9843
False Alarms	101	91	98	99	91	100	112	91	81
Missing Alarms	50	47	48	45	47	49	41	47	59

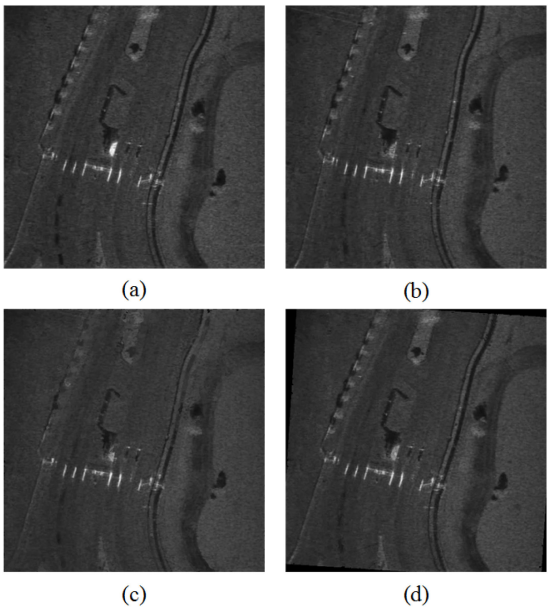


Fig. 6. Registration results of a representative image pair by using two strategies. (a) Reference image. (b) Unregistered image. (c) Registered image obtained by the CNN that directly estimates the displacement field between the reference and unregistered images. (d) Registered image obtained by the proposed approach that uses coarse-to-fine strategy.

unnatural. Actually, unlike the medical image registration, it is unsuitable to use CNN to directly estimate the displacement field between video SAR images due to two factors. On the one hand, the rigid motion between images results in a relatively large displacement field which is more difficult to be accurately estimated. On the other hand, moving target shadows are possible to be aligned when estimating the displacement field directly.

We additionally perform an ablation study to examine the impact of the hyper-parameter α, the hyper-parameter β, and the threshold γ. In the experiment, when changing the value of one parameter to examine its influence, the values of other parameters are default. Table VI shows ablations over these parameter values. Clearly, the performance is found to be relatively stable with respect to the values of hyper-parameters α and β. In addition, with the increase of the threshold γ, the image similarity between the reference and registered images increases, but moving target shadows are more possible to be aligned and thus the detection performance decreases.

E. Generalization Ability

Deep learning has been intensively applied in quite a few related fields, for example, target recognition in electro-optical images or videos. Some works have shown the potentials of deep learning technology in radar, particularly in image-based classification and recognition. Different from the electro-optical images that are easily available, numerous radar images with sufficient diversity cannot be relied on in most cases, which limits the training of any designed networks. There have been some serious concerns about the generalization ability of applying deep learning in radar. We attempt to briefly discuss the generalization ability of the proposed approach for completeness.

The proposed approach has been used to process the second SAR video product, while the network is trained by the first SAR video dataset. The 250 image pairs, randomly selected from the second testing set, are fed into the proposed framework for testing. We compare the test results with the registration results of the SAR-SIFT, the LLT, and the LPM. Table VII reports the PSNR and SSIM between the reference and registered images obtained by these approaches. The registration results by using different approach is given in Fig. 7. It is obvious that the proposed approach has better performance in video SAR image registration compared to the conventional methods. In addition, based on different image registration approaches, we perform moving target detection on the second testing set, and the detection results are listed in Table VIII. It should be pointed out that it is easier to detect the moving target on the second SAR video where only a target moves, and hence, all detection results on the second testing set are perfect. As confirmed by the image similarity and the detection results, the proposed approach outperforms the conventional methods when applied

TABLE VII
PSNR AND SSIM RESULTS ON THE OTHER SAR VIDEO DATA THAT ARE NOT USED IN TRAINING

Methods	PSNR	SSIM
SAR-SIFT	30.3973	0.8034
LPM	31.7466	0.8597
LLT	31.5863	0.8374
Ours	38.6903	0.9784

TABLE VIII
FALSE ALARMS AND MISSING ALARMS RESULTS ON THE OTHER SAR VIDEO DATA THAT ARE NOT USED IN TRAINING

Methods	false Alarms	missing Alarms
SAR-SIFT	0	0
LPM	0	0
LLT	0	0
Ours	0	0

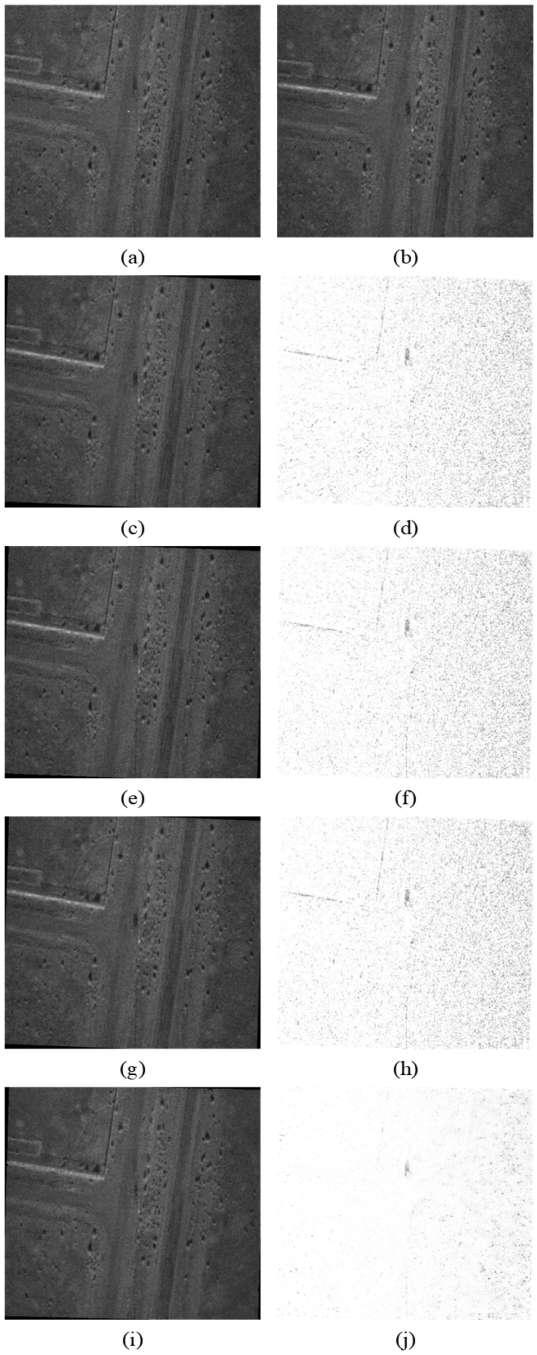


Fig. 7. Registration results of the other SAR video data that are not used in training of the developed framework. (a) Reference image, (b) unregistered image, (c), (e), (g), and (i) are registered images obtained by the SAR-SIFT, LLT, LPM, and the proposed approach, respectively. (d), (f), (h), and (j) are ratio image between the reference image and the registered image obtained by the SAR-SIFT, LLT, LPM, and the proposed approach, respectively.

to an unknown scenario. It can be concluded that the proposed approach has a satisfactory generalization ability, which comes from its unsupervised training strategy.

IV. Conclusion

An unsupervised framework has been developed for image registration of video SAR, which consists of a preliminary registration network and a fine registration network. The preliminary registration network predicts the parameters of the rigid transformation model between two images, and registers the two images accordingly. After that, the remained displacement of stationary targets is smaller than that of moving targets in the two images. Therefore, the fine registration network can accurately estimate the remained displacement of stationary targets to compensate all the differences between two video SAR images except for moving targets shadows. Processing results of real video SAR data have revealed that the proposed unsupervised approach achieves good performance in terms of the image registration. In addition, the proposed approach shows a convincing generalization ability when applied to a different dataset, which is appealing to radar applications.

In future work, we would like to further evaluate the generation ability of the proposed approach on video SAR datasets with much radiometric difference. An effective target detection method based on the proposed image registration framework is highly desired for video SAR.

References

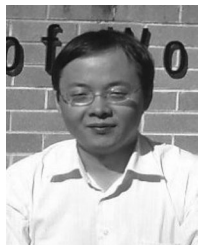
[1] A. Damini, B. Balaji, C. Parry, and V. Mantle, "A video SAR mode for the x-band wideband experimental airborne radar," in *Proc. SPIE*, vol. 7699, pp. 76990E-1–76990E-11, 2010.
[2] J. Miller, E. Bishop, and A. Doerry, "An application of backprojection for video SAR image formation exploiting a subaperature circular shift register," in *Proc. SPIE*, vol. 8746, pp. 874609-1–874609-14, 2013.
[3] B. Wallace, "Development of a video SAR for FMV through clouds," in *Proc. SPIE*, vol. 9479, pp. 94790L-1–94790L-2, 2015.
[4] L. Wells, K. Sorensen, A. Doerry, and B. Remund, "Developments in SAR and IFSAR systems and technologies at Sandia National Laboratories," in *Proc. IEEE Aerosp. Conf.*, 2003, pp. 2_1085–2_1095.
[5] X. Tian, J. Liu, M. Mallick, and K. Huang, "Simultaneous detection and tracking of moving-target shadows in ViSAR imagery," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.2998782.
[6] Y. Zhang, S. Yang, H. Li, and Z. Xu, "Shadow tracking of moving target based on CNN for video SAR system," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 4399–4402.
[7] J. Ding, L. Wen, C. Zhong, and O. Loffeld, "Video SAR moving target indication using deep neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7194–7204, Oct. 2020.
[8] Y. Zhang, D. Zhu, X. Yu, and X. Mao, "Approach to moving target shadow detection for VideoSAR," *J. Electron. Inf. Technol.*, vol. 39, no. 9, pp. 2197–2202, 2017.
[9] H. Wang, Z. Chen, and S. Zheng, "Preliminary research of low-RCS moving target detection based on Ka-Band video SAR," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 6, pp. 811–815, Jun. 2017.
[10] 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.
[11] C. Xing and P. Qiu, "Intensity-based image registration by nonparametric local smoothing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 2081–2092, Oct. 2011.
[12] J. P. Kern and M. S. Pattichis, "Robust multispectral image registration using mutual-information models," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 5, pp. 1494–1505, May 2007.
[13] D. Li and Y. Zhang, "A fast offset estimation approach for InSAR image subpixel registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 2, pp. 267–271, Mar. 2012.
[14] M. Unser and P. Thevenaz, "Optimization of mutual information for multiresolution image registration," *IEEE Trans. Image Process.*, vol. 9, no. 12, pp. 2083–2099, Dec. 2000.
[15] S. Wang, H. You, and K. Fu, "BFSIFT: A novel method to find feature matches for SAR image registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 4, pp. 649–653, Jul. 2012.
[16] A. Sedaghat, M. Mokhtarzade, and H. Ebadi, "Uniform robust scale-invariant feature matching for optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4516–4527, Nov. 2011.

[17] Q. Li, G. Wang, J. Liu, and S. Chen, "Robust scale-invariant feature matching for remote sensing image registration," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 2, pp. 287–291, Apr. 2009.
[18] F. Dellinger, J. Delon, and Y. Gousseau, J. Michel, and F. Tupin, "SAR-SIFT: A SIFT-like algorithm for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 453–466, Jan. 2015.
[19] M. Gong, S. Zhao, L. Jiao, D. Tian, and S. Wang, "A novel coarse-to-fine scheme for automatic image registration based on SIFT and mutual information," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 4328–4338, Jul. 2014.
[20] R. Feng, Q. Du, X. Li, and H. Shen, "Robust registration for remote sensing images by combining and localizing feature- and area-based methods," *ISPRS J. Photogrammetry Remote Sens.*, vol. 151, no. 1, pp. 15–26, 2019.
[21] Y. Zhang and D. Zhu, "Height retrieval in postprocessing-based videoSAR image sequence using shadow information," *IEEE Sensors J.*, vol. 18, no. 19, pp. 8108–8116, Oct. 2018.
[22] S. Miao, Z. J. Wang, and R. Liao, "A CNN regression approach for real-time 2D/3D registration," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1352–1363, 2016.
[23] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, "Quicksilver: Fast predictive image registration-A deep learning approach," *NeuroImage*, vol. 158, pp. 378–396, 2017.
[24] B. D. D. Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Isgum, "End-to-end unsupervised deformable image registration with a convolutional neural network," *Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, 2017, pp. 204–212.
[25] G. Balakrishnan, A. Zhao, M. R. Sabuncu, A. V. Dalca, and J. Guttag, "An unsupervised learning model for deformable medical image registration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9252–9260.
[26] M. Niethammer, R. Kwitt, and F. Vialard, "Metric learning for image registration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8455–8464.
[27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[29] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 1–31, 2011.
[30] J. S. Lee, "Speckle analysis and smoothing of synthetic aperture radar images," *Comput. Graph. Image Process.*, vol. 17, no. 1, pp. 24–32, 1981.
[31] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, and J. Tian, "Robust feature matching for remote sensing image registration via locally linear transforming," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6469–6481, Dec. 2015.
[32] J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo, "Locality Preserving Matching," *Int. J. Comput. Vis.*, vol. 127, no. 5, pp. 512–531, May 2019.

Xuejun Huang received the B.Eng. degree in electronic engineering from Xidian University, Xi'an, China, in 2018, where he is currently pursuing the Ph.D. degree in electronic engineering.

His research interest includes machine learning in radar.

Jinshan Ding is currently a Professor with the School of Electronic Engineering of Xidian University, China. He founded the millimeter-wave and THz research group in Xidian University after his return from Germany. His research interests include millimeter-wave and THz radar, video SAR, and machine learning in radar.



Qinghua Guo (Senior Member, IEEE) received the B.E. degree in electronic engineering and the M.E. degree in signal and information processing from Xidian University, Xian, China, in 2001 and 2004, respectively, and the Ph.D. degree in electronic engineering from the City University of Hong Kong, Kowloon, Hong Kong, in 2008.

He is currently an Associate Professor with the School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Wollongong, NSW, Australia, and an Adjunct Associate Professor with the School of Engineering, The University of Western Australia, Perth, WA, Australia. His research interests include signal processing, machine learning, and telecommunications.

Dr. Guo was a recipient of the Australian Research Councils inaugural Discovery Early Career Researcher Award, in 2012.