Joint Exploitation of Coherent Change Detection and Global-Context Capturing Network for Subtle Changed Track Detection With Airborne SAR

Jinsong Zhang, *Member, IEEE*, Mengdao Xing, *Fellow, IEEE*, Wenkang Liu, *Member, IEEE*, and Guangcai Sun, *Senior Member, IEEE*

***Abstract*—Change detection is a crucial remote sensing (RS) application because it can locate the interesting changed regions and provide corresponding time-series information with multitemporal RS images acquired in the same region. Synthetic aperture radar (SAR), with the advantages of all-day and all-weather conditions, can achieve high-resolution imaging from long operation distances. Moreover, the coherence imaging characteristics of the SAR system cause the attained complex-value images to be more sensitive to ground surface features. Traditional intensity-based change detection methods merely use the intensity difference between multitemporal images; thus, the results only reflect the significant landmark changes, such as seismic disasters and flood disasters. In this article, we use the coherent imaging characteristics of the SAR complex images to detect very subtle changes, such as vehicle tracks and footprints. Specifically, coherent change detection based on amplitude and phase generates a difference image utilizing the repeat-pass repeat-geometry SAR images, while the global attention-based convolutional neural network achieves automatic subtle change track extraction from the difference images. The experimental results based on our measured airborne SAR data demonstrate that our proposed method reduces the sensitivity of the detectable changed region from the meter level to the centimeter level, further providing our method with the ability to detect very subtle changes, such as vehicle tire tracks or footprints by human activities. This subtle change detection capability can be used for the search and rescue of vehicles and personnel lost in the field.**

***Index Terms*—Coherence change detection (CCD), convolutional neural network (CNN), synthetic aperture radar (SAR), UNet.**

I. Introduction

Remote sensing (RS) is a long-range and contactless sensing approach that can be used to observe interesting regions based on airborne or spaceborne systems. Introducing popular data-driven deep-learning (DL) methods into RS image analysis has achieved intelligent and efficient processing. The typical DL methods for RS image analysis include object detection, scene classification, cloud detection, and especially change detection [1], [2].

Different from optical or spectrum measurements [3], [4], the result of synthetic aperture radar (SAR) is a complex-value image that can be decomposed as the intensity fraction and phase fraction; here, the intensity fraction reflects the reflectivity of landmarks and the phase fraction represents the targets' electromagnetic scattering characteristics [5]. The SAR-based change detection method can be divided into incoherent change detection (ICCD) and coherent change detection (CCD). The ICCD methods use the intensity component of SAR complex images as input, and the input images need to be the standard Level-2 images within radiometric and geometric correction, such as the Sentinel-1 L2 product [6].

Unlike ICCD methods, CCD methods use both intensity and phase fractions to generate the differences [7], [8]. However, this requires strict repeat-pass repeat-geometry imaging. Specifically, the reference image for the region of interest is acquired during the first pass. Later, after a few hours or days of subtle changes to the track, a slave image is acquired for the same region using the same imaging geometry. Once the repeat-pass image pair is acquired, preprocessing strategies, such as resolution harmonization and geometric correction, are implemented before utilizing a complex-valued coherence estimator, such as maximum likelihood estimation, to generate the difference image between image pairs. The coherence value of the generated difference image falls within the range [0, 1]. A coherence value of 0 indicates whole decorrelation, while a coherence value of 1 indicates whole correlation. The detectable sensitivity of the CCD method to subtle changes is related to the working frequency and image resolution. The Sentinel-1 system, with a wavelength of 5.6 cm in the C-band, can utilize the CCD method for urban building monitoring with meter deformation [9]. In contrast, an airborne millimeter-wave SAR system can detect subtle tracks with centimeter deformations [10]. While the spaceborne CCD

Manuscript received 21 December 2023; revised 14 March 2024; accepted 23 March 2024. Date of publication 2 April 2024; date of current version 19 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62301403 and Grant 62101404, and in part by the Key Program of National Natural Science Foundation of China under Grant 62331020. *(Corresponding authors: Jinsong Zhang; Wenkang Liu.)*

Jinsong Zhang and Wenkang Liu are with the Academy of Advanced Interdisciplinary Research, Xidian University, Xi'an 710071, China (e-mail: jinsongxd@163.com; wkliu@stu.xidian.edu.cn).

Mengdao Xing is with the National Key Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China, also with the Collaborative Innovation Center of Information Sensing and Understanding, Xidian University, Xi'an 710071, China, and also with the Academy of Advanced Interdisciplinary Research, Xidian University, Xi'an 710071, China (e-mail: xmd@xidian.edu.cn).

Guangcai Sun is with the National Key Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China, and also with the Collaborative Innovation Center of Information Sensing and Understanding, Xidian University, Xi'an 710071, China.

Digital Object Identifier 10.1109/JSTARS.2024.3382634

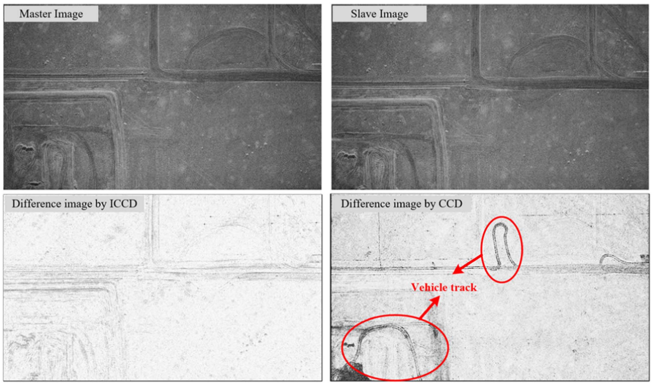


Fig. 1. Difference image by ICCD and CCD for the Sandia airborne data.

can share a stable imaging orbit and imaging parameters, the airborne CCD can detect subtle changes but is easily affected by imaging geometry. The purpose of this article is to demonstrate the ability of an airborne millimeter system to detect vehicle tracks or footprints resulting from human activities [11]. We selected a Sandia SAR dataset comprising two complex-valued images [10]. Fig. 1 depicts the difference images generated using the intensity-based ICCD method and the complex-valued-based CCD method. Interestingly, the CCD method detected vehicle tracks that were overlooked by the ICCD, which demonstrates that CCD can extract weaker changes compared to ICCD.

SAR CCD is generally divided into three stages: image registration, coherent estimation, and differential image extraction. The main purpose of the image registration stage is to generate repeat-pass SAR images with good coherence. The coherent estimation stage utilizes a complex coherent estimation operator to invert and extract the difference between amplitude and phase information through energy accumulation. The difference map analysis stage uses image analysis or postprocessing to extract regions of interest changes from the difference map affected by clutter and interference and achieves automatic target recognition.

The registration process of SAR CCD is similar to that of interferometric SAR (InSAR) image registration: both are given complex image pairs, and the pixel offset between the secondary image and the main image is obtained through two steps of coarse registration and fine registration. The subpixel-level registration results are achieved through interpolation of the secondary image. According to the different data used, SAR complex image registration algorithms can be divided into two categories: geometric registration and image registration [12]. Geometric registration is based on the geometric relationship between the sensors and ground points during InSAR imaging [13]. The image registration is based on the data information contained in the SAR complex images and usually uses the method of image window matching to calculate the registration offset of control points.

CCD coherence estimation needs to consider both the intensity and phase information because the track deformation is usually weak and cannot be extracted from the intensity difference [14]. Modeling the paired images as a jointly circular, zero mean Gaussian vector, the most classical coherence estimator is defined as follows [15], [16]:

$$\hat{\gamma}_c = \frac{\left|\sum_{k=1}^{N} f_k^* g_k\right|}{\sqrt{\sum_{k=1}^{N} f_k^* f_k \cdot \sum_{k=1}^{N} g_k^* g_k}} \tag{1}$$

where f_k and g_k are the complex-valued pixels of the master image and slave image, respectively, $*$ means the conjugate operation, and N is the scale of the local window, usually 7×7 or 9×9. Replacing the multiplication operation in the denominator of (1) with an addition operation, the Berger estimator is more stable and demonstrates better performance in a certain condition [17]. Focusing on improving the contrast of the changed regions and background regions, the complex reflectance change detection (CRCD) metric is proposed to distinguish the change regions from low clutter-to-noise ratio areas [18]. In addition, the estimator-fused coherent and noncoherent estimation can effectively distinguish significant changes and subtle changes, further improving the visibility of the coherence image [19].

The change pixel extraction from a difference image can be viewed as a semantic segmentation task that classifies each pixel as interesting changes or cluttered change pixels [20]. Different from ICCD methods, the interesting areas, such as track regions in the CCD task, are usually weak and small-scale, which drastically increases the difficulty of track detection [21]. Quach et al. [7] used an explicit objective function based on the Bayesian information criterion to find the curves in the difference images. Kuny et al. [22] analyzed the vehicle track detection performance of different data augmentation strategies based on the classical UNet structure. Based on these methods, we also proposed the multiple statics contributing to the compressed UNet method focusing on the few track samples [23], and the spatial feature enhanced UNet and adaptive augmentation to improve the track detection performance [24]. In summary, the track detection network also needs to be improved by focusing on the special features of different tracks, such as footprints and vehicle tracks.

A novel framework that fuses SAR CCD and DL models is proposed to detect subtle tracks in SAR images. First, we create an airborne SAR track detection dataset that includes six large-scene image pairs covering diverse terrains and tracks. Each image pair comprises two complex-valued repeat-pass repeat-geometry images, and tire tracks and footprints are arranged in some locations during data acquisition. Next, to preprocess the images, we consider the azimuth resolution difference between paired images caused by natural factors and execute resolution correction. We then represent the two-step coregistration to achieve subpixel-level image registration, and a difference image is generated based on the complex-value coherence estimator. After, we propose a global-attention-based detection network that uses the UNet structure as the basic detection network and involves the transformer structure and group spatial layers to capture special track features, such as parallel distribution and long continuity. Finally, we compare our proposed method with popular detection networks to validate its effectiveness, and the experimental results demonstrate that our proposed method can extract not only vehicle tracks and footprints from images of

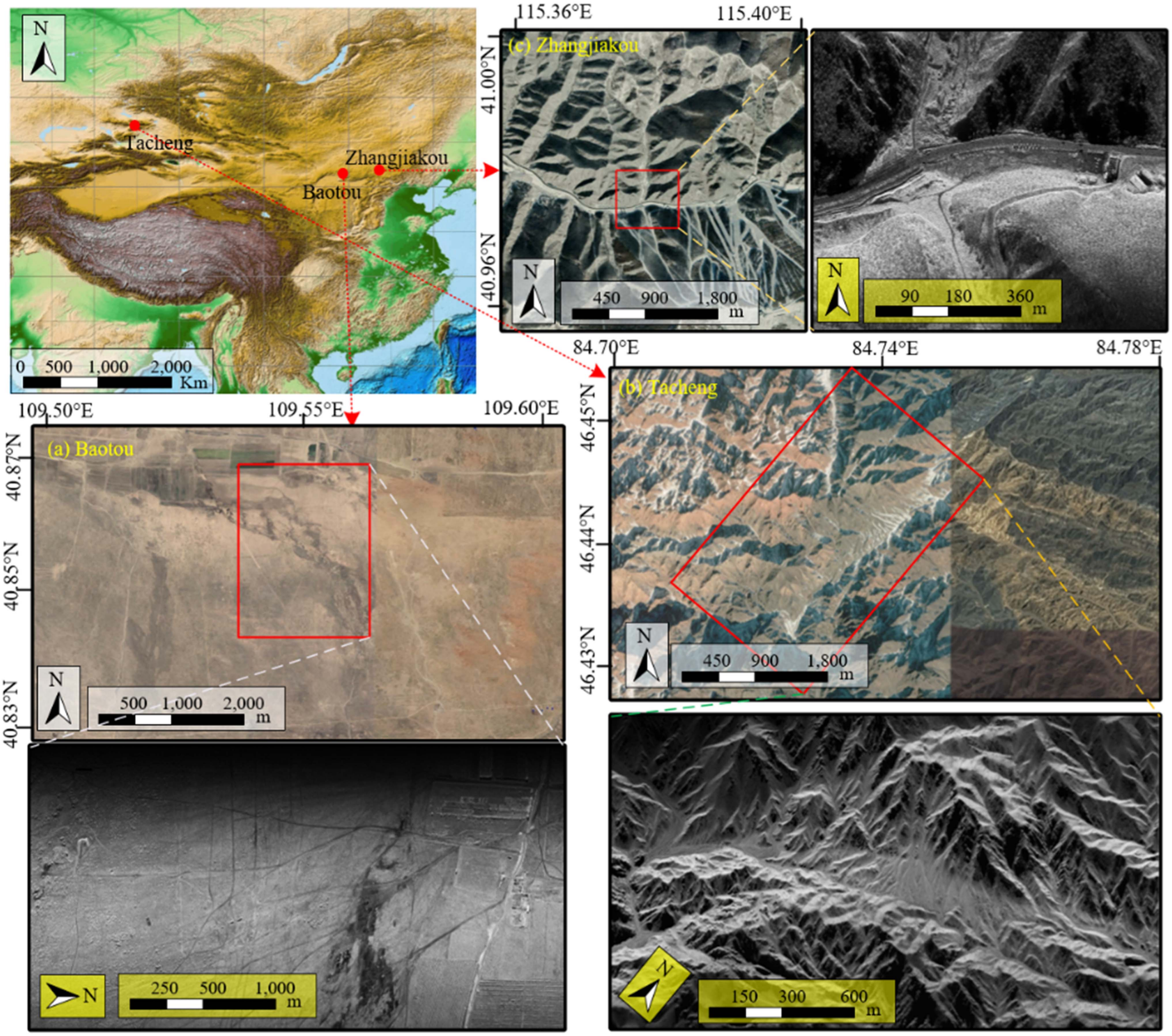


Fig. 2. Three study areas and SAR images acquired by our airborne CCD system.

various terrains but also reflect the time sequences of human activities. In summary, our main contributions are as follows.

1) We create a subtle track detection dataset from our airborne SAR system, containing six large-scene complex-value image pairs of different human activities in multiple terrain backgrounds.
2) We construct a track detection framework by fusing SAR CCD and a convolutional network. The CCD in our framework includes resolution correction, two-step coregistration, and complex coherence estimation, and can steadily generate the track pixels between the paired images under complex flight conditions and terrain backgrounds.
3) We develop a global context capturing network (GCCN) by integrating our proposed group spatial convolution module (GSCM) into the transformer-based pixelwise TransUNet, which can achieve subtle track detection under complex backgrounds with high recall and precision.

The rest of this article is organized as follows. The measured experimental data are provided in Section II. The processing diagram or proposed track detection is shown in Section III. Section IV contains the experimental results, while Section V provides the extended results. Finally, Section VI concludes this article.

II. Study Area and Data

Three regions of China of Baotou, Tacheng, and Zhangjiakou were selected to validate the proposed track detection method. As shown in Fig. 2, these regions cover different topographies. The Baotou region has soft and hard sand as its main terrain, with sparsely distributed shrubs. We arranged some vehicle tracks and footprints in this region to validate the track detection performance. The Tacheng region has intermittent mountainous terrain with a gravel ground surface. We also arranged some

TABLE I
WORKING PARAMETERS OF OUR AIRBORNE CCD SYSTEM

Parameter	Value
Carrier frequency	35 GHz
Chirp bandwidth	900 MHz
Pulse repetition frequency	1250 Hz
Flight altitude	Nearly 4,000 m
Imaging mode	Stripmap
Theoretical azimuth resolution	0.3 m
Theoretical range resolution	0.3 m

vehicle tracks and footprints in this area. Since gravel is usually hardening, the left tracks were even weaker, which could better validate the detection method. Finally, the Zhangjiakou region hosted the snow events of the 2022 Beijing Winter Olympic Games, with snow land as its main ground surface; the result from this area could validate the detection performance on this special surface. In summary, we arranged different tracks in these regions with various backgrounds. Due to the differences in topography and scene arrangement, the tracks left human activity would show different characteristics, which could better demonstrate the superiority and robustness of the different methods.

Table I presents the working parameters of our airborne CCD system. The system operates at a carrier frequency of 35 GHz (Ka band) with a wavelength of 8.6 mm, which maximizes the phase sensitivity of the CCD methodology [8]. The system is capable of achieving an azimuth and range resolution of 0.3 m, which remains relatively constant across different imaging geometries and scenes.

Fig. 2 shows the imaging results for each region, with one image selected for each. Optical images confirm that the SAR single look complex (SLC) images accurately reflect the real landcover of these regions, covering an area ranging from 3 km^2 to nearly 10 km^2. The images were acquired in 2021 and 2022. Table II provides a detailed description of the flight route and image acquisition. For region (a) of Baotou city, we arranged two flight routes, and each route was acquired in two passes. For example, the two passes of the north-to-south route were acquired at 19:27:36 and 20:17:51 on September 10th, 2021. Since these two images were acquired from the same route, they shared the same imaging geometry and were combined into an image pair with a time interval of 50 min for CCD processing. During this interval, two offload vehicles were driven across the sand with complicated routes, such as curved paths and intersections, leaving double-row tire tracks in the sandy ground. The axle width of the vehicles was approximately 1.8 m, and the wheel width was approximately 0.25 m. Moreover, two people also left foot tracks in the imaging region with careless walking routes. Photos of the left vehicle track and foot track are shown in Fig. 3. Both the vehicle track and foot track were continuous, with the vehicle track usually double-row and the foot track usually single-row. According to our rough statistics, the depth of these tracks was approximately a few centimeters, which was a very low deformation for SAR images and could better validate the effectiveness of our detection method. Notably, during the track arranging process, we used a handheld GPS to record the location of the tracks and recorded them as the ground truth in the subsequent experimental results. We also arranged the flight route from east to west, and the acquired images formed the image pair with index 2.

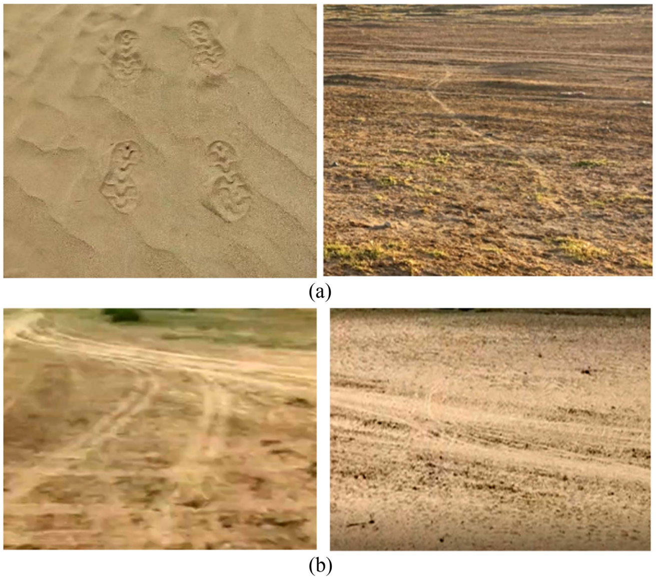


Fig. 3. Footprint and vehicle tracks in our experiments. (a) Foot tracks. (b) Vehicle tracks.

For region (b) of Tacheng city, under the flight route from southwest to northeast, we acquired three images from three passes, maintaining the same imaging geometry. Since every two images could form an image pair, we obtained three image pairs with the indices of 3, 4, and 5 in this region for track detection. We also arranged some vehicle tracks and footprint tracks in this region during each pass interval. For region (c) of Zhangjiakou city with a snow surface, we arranged the flight route from east to west and only arranged the vehicle tracks in this region. Two-pass images were acquired in region (c) and combined into image pair 6.

Notably, the temporal interval between these image pairs was a few hours, which ensured entire coherence between the image pairs while highlighting the decorrelation caused by track deformation due to human activities [8], [15]. Our proposed network-based detection method belonged to supervised learning, which required training data to optimize the detection model and testing data to validate the detection performance. We used the image pair with indices 1 and 5 as the training samples, while the image pairs with indices 2, 3, 4, and 6 were used as the testing samples. More details regarding the data samples are provided in the experimental section.

III. PROPOSED METHOD

In this section, divided into the training and testing stage, the complete flow is proposed starting from the preprocessing of the SAR CCD data and ending with the detection network, as shown in Fig. 4.

TABLE II
DESCRIPTION OF THE FLIGHT ROUTE AND IMAGE ACQUISITION

Region	Flight route	Data index	Acquiring time (yyyy-MM-dd) HH-mm-ss	Image pair (interval)	Track type	Training/testing
(a) Baotou	North south	pass 1	(2021-09-10) 19-27-36	pass1, pass2 **1** (50 min)	Vehicle, footprint	Training
		pass 2	(2021-09-10) 20-17-51			
	East west	pass 1	(2021-09-25) 18-49-13	pass1, pass2 **2** (53 min)	Vehicle, footprint	Testing
		pass 2	(2021-09-25) 19-42-15			
(b) Tacheng	Southwest to northeast	pass 1	(2022-07-21) 14-56-02	pass1, pass2 **3**(2 h 38 min) pass1, pass3 **4**(4 h 16 min)	Vehicle, footprint	Testing
		pass 2	(2022-07-21) 17-34-36			
		pass 3	(2022-07-21) 19-22-21	pass2, pass3 **5**(1 h 38 min)	Vehicle, footprint	Training
(c)Zhang-jiakou	East west	pass 1	(2021-12-25) 12-19-33	pass1, pass2 **6**(1 h 43 min)	Vehicle	Testing
		pass 2	(2021-12-25) 14-02-23			

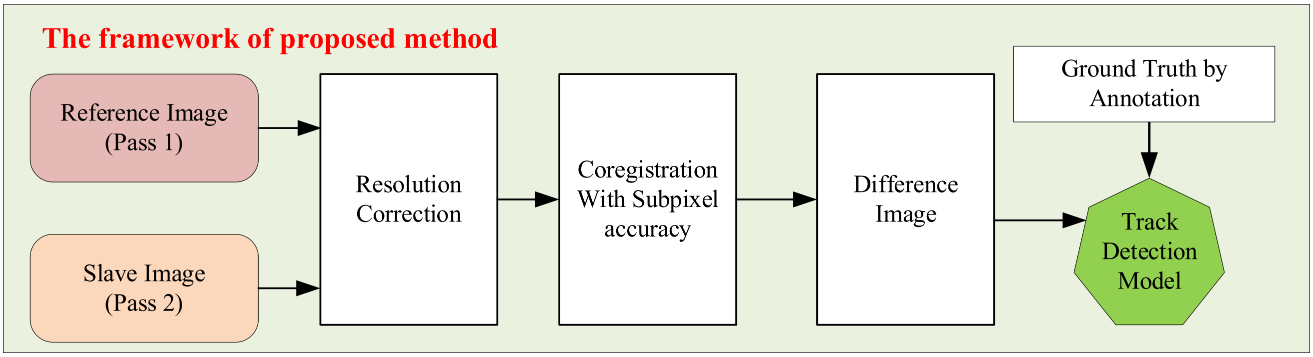


Fig. 4. Processing diagram or proposed track detection method.

A. Azimuth Resolution Correction for Improving Coherence

In contrast to traditional ICCD methods that use standard SAR L2 data as input [9], our method utilizes SLC data for subtle track detection. In a spaceborne-based CCD system, the imaging system has a stable running orbit and flight speed, which are not easily affected by the surrounding environment. As a result, the acquired image pairs from different times have the same azimuth and range resolution. However, for an airborne system, the key problem is that the airplane velocity can be affected by weather conditions during the imaging period, resulting in different azimuth resolutions for images acquired at different times [25].

Specifically, the imaging geometry of the spotlight mode is shown in Fig. 5, where h is the flight altitude and v is the average flight velocity along the preset route. During the acquisition time, the track region is covered by the antenna footprint with the aircraft moving. The theoretical imaging slant range resolution ρ_r is defined as $c/2B_r$, where c is the light speed and B_r is the devised bandwidth. For the SLC image in CCD processing, the range spacing resolution is $c/2F_r$, where F_r is the frequency sampling rate higher than B_r. Thus, the range resolution of the repeat-pass images always remains the same under the identified

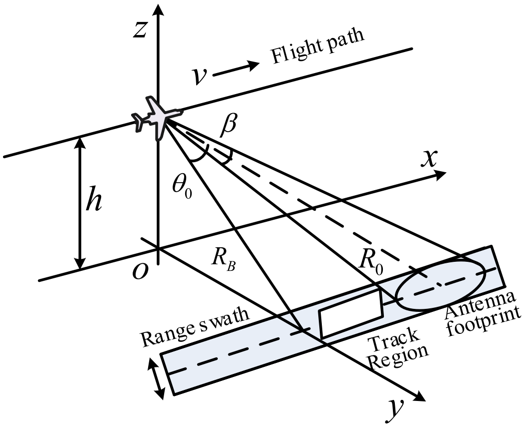


Fig. 5. Geometric model of the airborne SAR CCD system.

flight path. The theoretical azimuth resolution ρ_a is $\lambda/2\Delta\theta$ for our CCD system in stripmap mode, where λ is the wavelength and $\Delta\theta$ is the beamwidth. According to the definition of ρ_a, the azimuth resolution is not affected by the velocity. Specifically, the azimuth spacing resolution that determines the characteristics of CCD images is v/PRF, where PRF is the pulse repetition period. Thus, when the average flight velocity changes due to weather factors, the azimuth resolution also varies between SLC repeat-pass images, resulting in decreased coherence between the paired images and heavily affecting the final detection result [25], [26]. Under these circumstances, we need to unify the azimuth resolution of repeat-pass images according to the respective average flight speed$v$.

More specifically, assuming that the average speed velocity for the first pass and the second pass is v_1 and v_2, respectively, and the corresponding azimuth resolution is ρ_r^1 and ρ_r^2, we choose the low values of ρ_r^1 and ρ_r^2 as the baseline resolution and then use the ratio v_1/v_2 to adjust the image with higher resolution to the baseline resolution using the frequency-domain interpolation method. With this processing, the azimuth resolution difference can be suppressed, and the entire coherence between paired images is also improved [8]. Notably, the topographical phase caused by radar position changes needs to be removed with a DEM or accurate topography for spaceborne CCD systems. However, for an airborne CCD method, the flight path of repeat-pass images can be kept consistent by controlling the airplane platform. Thus, in this article, we do not consider the topographical phase.

B. Image Coregistration and Coherence Estimation

Similar to the procedure in InSAR processing, accurate coregistration of image pairs is essential for generating the difference image in CCD methods [27]. Although the identical imaging geometry of repeat-pass images can be ensured by strict flight routes, slight horizontal position deviations can lead to subtle deformations between the image pairs [15], [28].

According to the matching accuracy, the coregistration method can be divided into two processes: coarse coregistration at the pixel level and fine coregistration at the subpixel level. The coarse coregistration attempts to achieve matching of repeat-pass images at one or two pixel accuracy, which includes offset searching and image shifting. Specifically, in our difference image generation process, the SLC image acquired at the first pass is assigned as the master image, and the other image is assigned as the slave image. Then, offset searching is implemented based on the cross-correlation of the image's intensity fraction, and the peak of the correlation map reflects the offset [28], [29]. By utilizing the global offset to shift the slave image, the master image and slave image are coregistered at the pixel level.

Since the phase identity between repeat-pass images highly affects the complex coherence, fine coregistration for subpixel-level accuracy is essential. One constant offset is not adequate for resampling the slave image due to the topographic relief; thus, the polynomial transformation is usually utilized to fit the offset variants between selected tie points when the orbits and digital elevation model are not available [8], [30]. However, when the landmark backgrounds are complicated, the coregistration accuracy is easily affected by tie points, further affecting the track detection performance. To achieve accurate coregistration with subpixel accuracy, the extended flow optical Lucas–Kanade (LK) iterative (eFolki) method is utilized in this article; this method does not require tie points or polynomial regression in the registration process [31], [32]. More specifically, given the master image I_1 and slave image I_2 defined on a 2-D support $S \in R^2$ with coarse coregistration, the dense optical flow, which is the displacement between paired images, is defined by $u : x \to u(x) \in R^2$. The local-based LK approach tries to minimize $u(x)$ over a local window centered on x as follows:

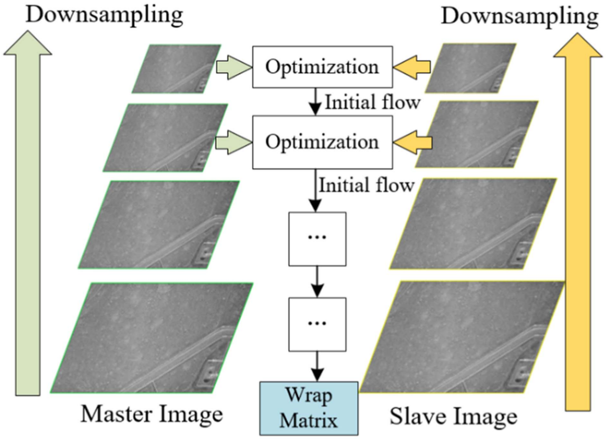


Fig. 6. Downsampling pyramid for improving coregistration accuracy.

$$J(u,x) = \sum_{x' \in S} \omega(x'-x)\left(R(I_1)(x') - R(I_2)(x'+u(x))\right)^2 \tag{2}$$

where ω is usually a square $(2r+1) \times (2r+1)$ window parameterized by radius r. $R(I)$ is a rank function as follows:

$$R(I)(x) = \#\{x' : x' \in S_R(x) \text{ with } |I(x)| > |I(x')|\} \tag{3}$$

where $S_R(x)$ is a neighborhood of pixel x. Here, the rank function can significantly compress signal dynamics and enhance the robustness to the repeat-pass flight route. In addition to the introduction of the rank function for improving the coregistration accuracy, multiple local windows with scales from 8 to 32 at an interval of 4 are introduced to increase the convergence in the optimization process.

Moreover, to find the displacements with various scales, downsampling with a Gaussian filter of input image pairs is performed with different sampling rates, as shown in Fig. 6. In the optimization process, the pyramid with the lowest resolution is initially optimized with an initialization flow of 0, and the other pyramids utilize both the previous estimation result and the corresponding image pairs. There are six levels in our pyramid; thus, the lowest resolution is 1/64 of the original images. Using the final wrap matrix as input, the frequency interpolation method is performed to transform the slave image to the value

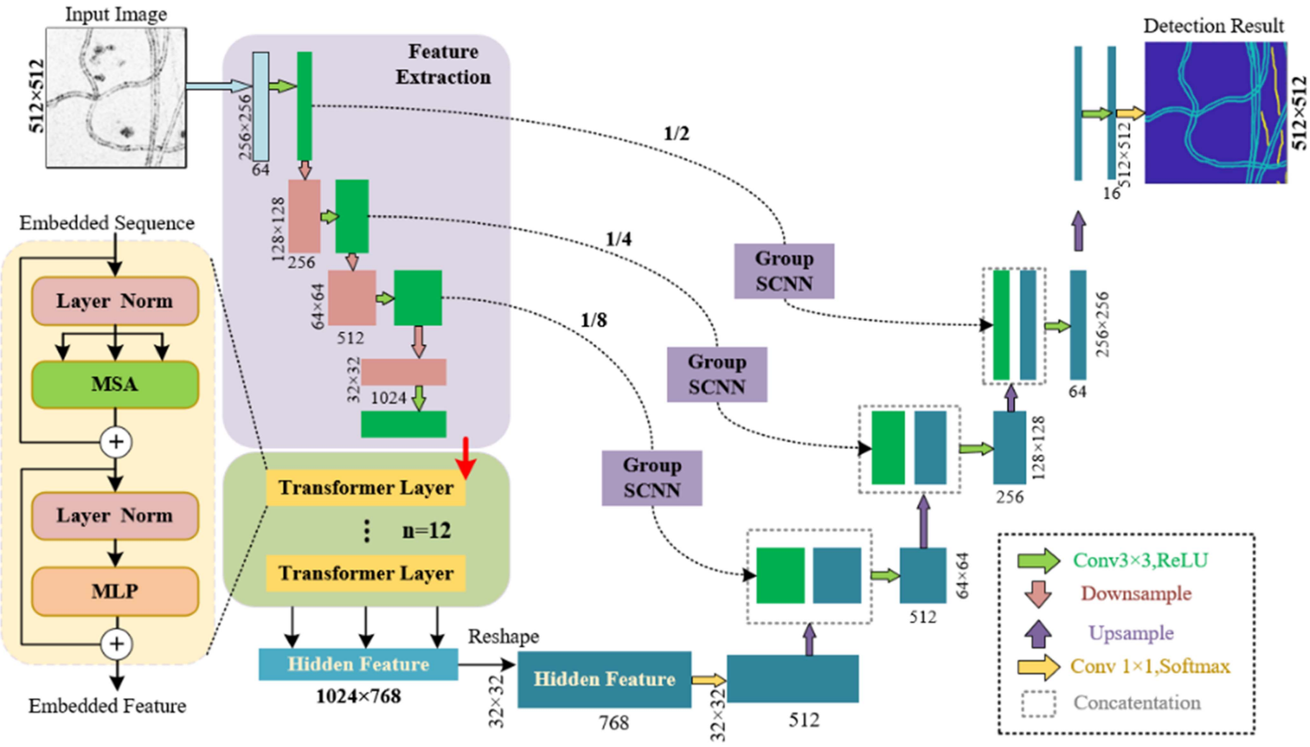


Fig. 7. TransUNet-based global context capturing network for track detection.

of the master image. After, we can obtain the final paired images with subpixel coregistration [31].

In addition to the estimator in (1), there are also other typical estimators, such as the Berger estimator and CRCD estimator [17], [18]. These estimators can improve the coherence accuracy under specific conditions, such as the Berger estimator used for images with identical intensity [17] and the CRCD estimator considering the clutter-to-noise ratio; however, the estimator in (1) still performs with the best stability for different backgrounds. Therefore, we still use the estimator in (1) to attain the coherence estimation.

C. Subtle Track Extraction With Global Context Capturing Network

Given the difference image patch $x \in \mathbb{R}^{H\times W}$ with a scale of $H \times W$, the task in this section is to predict the pixelwise track detection result with size $H \times W$ [33]. Both H and W are set to 512 pixels. Among the CNN-based detection methods, the UNet structure consisting of an encoder module and decoder module has been widely used, especially for pixelwise classification tasks [34]. However, convolution layers with a local receptive field cannot effectively model the global features and long-range relation of the input image. Fig. 7 shows one difference image patch of our track detection, and the vehicle tracks and footprints have long continuity that may span the entire image. Under these circumstances, if the original CNN-based UNet is still used as the detection network, the global context is disregarded, and the final result may be interrupted and considered to be unsatisfactory. Recently, the transformer structure has drawn much attention for its powerful global feature modeling capacity and superior transferability for downstream tasks [35]. By utilizing the CNN structure to capture high-resolution information semantic features, the transformer structure can be used to leverage global context, and TransUNet has succeeded in the semantic segmentation field [36]. Thus, in this article, we also use TransUNet as the main detection framework and utilize the GSCM to further extract the global context of subtle track to improve the detection performance. The entire framework of our GCCN is shown in Fig. 7.

1) CNN-Transformer as an Encoder for Global Feature Extraction: For the input image, our GCCN initially uses four cascaded convolution modules and downsampling layers to extract features on different levels [36]. The size of the final feature maps is 32 × 32, and the feature map number is 1024. By performing image sequentialization and patch embedding with a convolution layer with a kernel size of 1 × 1, we can obtain the embedded patch vectors as follows:

$$\mathbf{z}_0 = \left[\mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}\right] + \mathbf{E}_{\text{pos}} \tag{4}$$

where $\mathbf{x}$ is the image patch, $\mathbf{E}$ is the patch embedding projection, p is the patch size, $\mathbf{E}_{\text{pos}}$ is the position embedding vector to utilize the patch information, and N is the patch number, setting and set as 1024 in this article. Then, the transformer module is combined with L layers of multihead self-attention and multilayer perceptron and blocks [35], [36]. The lth layer is defined as follows:

$$\mathbf{z'}_l = \text{MSA}\left(\text{LN}\left(\mathbf{z}_{l-1}\right)\right) + \mathbf{z}_{l-1} \tag{5}$$

$$\mathbf{z}_l = \text{MLP}\left(\text{LN}\left(\mathbf{z'}_l\right)\right) + \mathbf{z'}_l \tag{6}$$

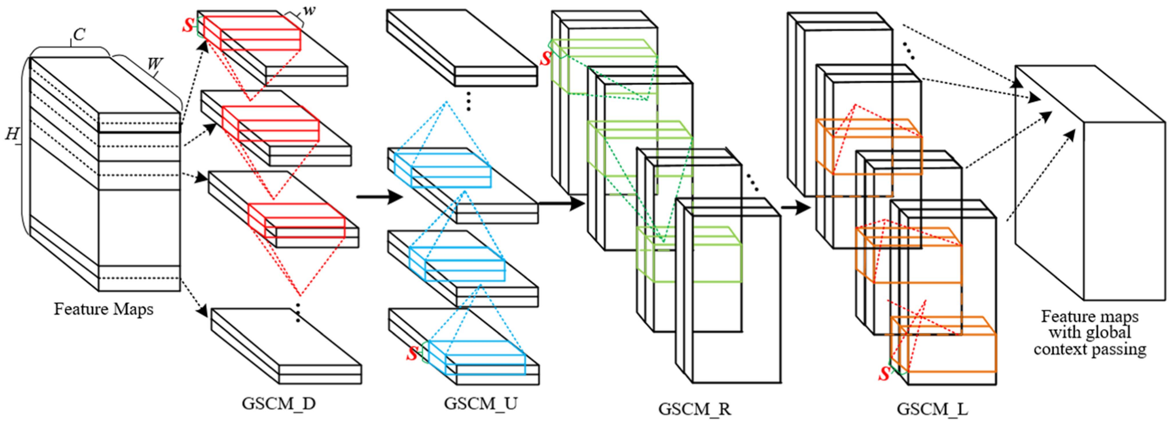


Fig. 8. Proposed GSCM for capturing global context passing.

where LN means the layer normalization layer and $\mathbf{z}_L$ with a size of 1024 × 768 is the final embedded feature of the input image. Afterward, the embedded feature is reshaped and convoluted with a kernel size of 1 × 1, which generates 512 feature maps with 32 × 32. The final embedded feature maps not only cover the high-dimensional semantic features of the CNN but also absorb the global context of the transformer module.

2) Group Spatial Convolution for the Message Passing of the Feature Map: Through the feature extraction of the CNN and transformer modules, feature maps with multiple levels are effectively extracted. The decoder module is shown in Fig. 7, and the skip-connection combined with different upsampling layers gradually recovers the size of the feature maps to the original image size. For the last conversion layer, the convolution with 1×1 transforms the feature maps into final detection results. In the decoder module, we introduce the GSCM into the feature mapping layer to improve the global context capturing ability. The detailed structure of our GSCM is shown as follows.

Our proposed GSCM consists of four cascaded modules: GSCM_D, GSCM_U, GSCM_R, and GSCM_L, according to the convolution direction of downward, upward, rightward, and leftward, respectively. The original spatial-CNN (SCNN) can improve the detection performance by achieving global feature passing with four cascaded convolution modules [37]. However, the SCNN computation complexity is strongly related to the size of the feature maps, such as W and H in Fig. 8; thus, it is more appropriate for the information to pass in the last feature maps of the encoder module. In regard to the feature passing of intermediate feature maps, which are usually large scale, such as 128 × 128 or 256 × 256 pixels, too many duplicate convolution layers lead to considerable computational resource waste.

The main goal of our GSCM is to maintain effective computing speed on the basis of increasing global information passing. The original SCNN utilizes the convolution kernel with $w \times 1$ or $1 \times h$ to transform the slice with $W \times 1$ or $1 \times H$, where w and h are the width and height of the kernels, respectively, and W and H are the width and height of the input slices, respectively. In our GSCM, the kernel size is set as $w \times s$ or $s \times h$, while the slice size is $W \times S$ or $S \times H$, as shown in Fig. 8. More specifically, for the GSCM_D module, the slice of the input 3-D feature map $\mathbf{X}$ is denoted as $\mathbf{X}_k$ with a size of $S \times W$, where $k = 1, 2, \ldots, H/S$, and the convolution kernel K has a size of $s \times h$. The convolution results of $\mathbf{X}_k$ are as follows:

$$\mathbf{X}'_k = \mathbf{X}_k + f\left(\mathbf{X}_{k-1} * \mathbf{K}\right) \tag{7}$$

where f is the rectified linear unit and $*$ means the convolution operation. $\mathbf{X}'$(with superscript ') denotes the slice that has been updated. Notably, all convolution kernels in one GSCM_D shared the same weights across all slices. With the above steps, for the feature maps with a size of 256 × 256 and number of 128, the convolution layers decrease from 128 to 128/S, thus improving the entire computational efficiency. For the three groups of feature maps in Fig. 8 with sizes of 256 × 256, 128 × 128, and 64 × 64, the kernel dimension s and slice dimension S are set to [1], [3], [5] and [1], [4], [8], respectively. Utilizing the above settings, the global context can be captured while maintaining a low computational complexity. With proposed GSCM network, the particularity of subtle tracks like long continuity of human footprint and parallel distribution of vehicle tracks can be effectively captured, and the detection performance can also be improved.

3) Loss Function for Network Optimization: For the optical image segmentation task, the cross-entropy (CE) loss is usually used as the optimization target, which is defined as follows:

$$L_{\mathrm{CE}} = -\left[y \log \hat{y} + (1-y) \log (1-\hat{y})\right] \tag{8}$$

where y is the label for one pixel and $\hat{y}$ is the corresponding predicted value, which is usually activated by the sigmoid function for two-class classification and the softmax function for multiclass classification. The CE loss can provide a very stable optimization convergence in the training process for the difference segmentation task. However, the track detection in this article has the problems of fewer training samples and unbalanced sample distributions of different classes, while the traditional CE loss may not be able to suppress them [38]. Thus, the dice loss is introduced as follows:

$$L_{\mathrm{dice}} = 1 - \frac{2\left|X \cap Y\right|}{|X| + |Y|} \tag{9}$$

where X and Y are the classification result and label, respectively, and $|\cdot|$ is the element number. Dice loss computing the union and intersection cannot only suppress the unbalanced sample

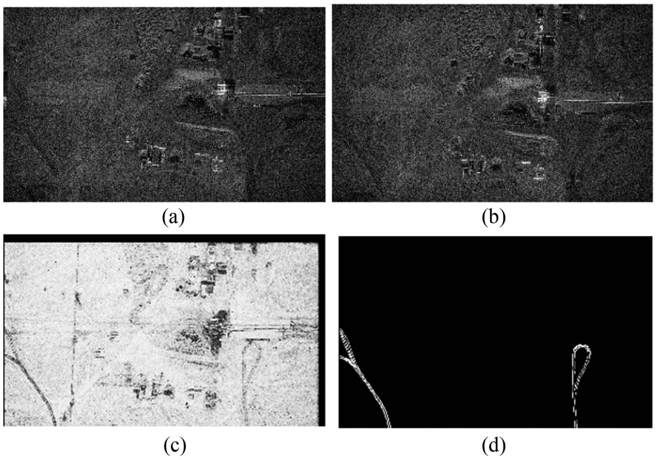


Fig. 9. One result including (a) master image, (b) slave image, (c) difference image, (d) and labeled images.

distribution but also improve the global classification performance. According to the above analysis, the combination of CE loss and dice loss is taken as the final optimization loss

$$L_{\text{total}} = L_{\text{CE}} + L_{\text{dice}}. \tag{10}$$

IV. Experiments

A. Training Dataset Description and Accuracy Assessment

1) Sandia Dataset for the Pretraining of Detection Model: The Sandia National Laboratory released a public SAR CCD dataset for the first time [10]. This dataset was collected by a Sandia test platform on February 14, 2006, in New Mexico. It consists of two flights, each containing 32 spotlight SAR images, providing 32 image pairs with a size of 1754 × 3000 pixels that can be used for CCD applications. During the acquisition interval, a car was driven in the region, leaving parallel tire tracks on the ground [39]. The image pairs in this Sandia SAR dataset are uncalibrated and not coregistered, which can be used to validate the effectiveness of our coregistration method.

Since Sandia does not provide the ground truth for these images, we manually annotated the vehicle track label based on our measured CCD data. We utilized our proposed method to process these images, including coregistration and coherence estimation. The two processing results are shown in Fig. 9. The master image and slave image in each pair show nearly identical intensity and surface distribution, making it impossible to find the subtle track directly from the intensity difference. However, within the coregistration and coherence estimation, the generated difference image highlights the vehicle track region and suppresses the background region. In addition, the vehicle track in the difference image shows characteristics, such as continuity and parallelism, which remain consistent with those in our measured CCD data. The labeled images show the track region based on our manual annotation, where 1 indicates the vehicle track and 0 indicates an unrelated background region.

Although the Sandia dataset provides vehicle track data similar to our experimental settings, it only contains vehicle tracks and not footprints. Furthermore, the system parameters and working modes of Sandia differ from those of our airborne system, which may impact the final track detection results. Therefore, we utilized the Sandia data as a pretraining dataset to optimize the GCCN for vehicle track detection [40]. Specifically, we adjusted the final classification layer of GCCN to one channel for vehicle track detection and utilized the Sandia data to obtain a well-trained detection model. We then utilized the well-trained weighting parameters, with the exception of the last layers, to initialize the final detection model with three channels in the final layers of GCCN for the detection of both vehicle tracks and footprints. Compared to the random initialization or transfer from other detection tasks, this pretraining strategy not only utilizes the track feature in the Sandia dataset but also maximizes the dependencies between the two detection tasks.

2) Entire Procedure for Our Airborne CCD Experiments: To fully learn the structure and features of subtle tracks in various regions and terrains, two of our CCD image pairs (pair 1 and pair 5) were selected as the training data. Table II shows the details of the selected pairs. The unprocessed image pairs were used to test the entire procedure of our method. The primary image pairs and corresponding difference images are shown in Fig. 10. The intensity difference between paired images was very weak, and it was difficult to directly detect the track. By using our proposed resolution correction, coregistration, and coherence estimation methods, the difference images were accurately generated, as shown in the figure. Notably, the vehicle tracks showed continuity and parallelism, while the footprints had continuity and a more random direction; this result indicated that our method attained the accurate coregistration of the images and coherence estimation. However, some decorrelation pixels existed in the generated difference images due to terrain occlusion and temporal position changes caused by random vegetation swing [15]. In addition, the tracks were discontinuous and inhomogeneous from a local perspective due to differences in track shape, depth, and surface characteristics. Moreover, we provided annotated ground truth for each image pair based on the track distribution in the difference images and our experimental records, as shown in Fig. 10; the blue and yellow regions indicate the vehicle track and footprint, respectively.

Due to the limitations of computational memory and resources, the original large images cannot be directly used as input for the detection networks. Therefore, the difference images and their corresponding annotation images were cropped into slices of 512 × 512 pixels with a 50-pixel overlap. This overlap helps the network learn the integrity of the track region. With this processing, we obtained 1603 images and their corresponding labeled images for network training.

3) Model Implementation Details and Evaluation Metrics: In this article, all experiments were conducted on a local server with an Intel(R) Xeon (R) Gold 6130 CPU, 1 NVIDIA Titan RTX, and Windows 10 operating system. The PyTorch framework was utilized to implement the detection network. During the training process, stochastic gradient descent was used with an initial learning rate of 0.01 and momentum of 0.9 to optimize the combination loss (11). The batch size was set to 6, and the epoch number was set to 150. To improve the training dataset's diversity, the images and corresponding labels were augmented

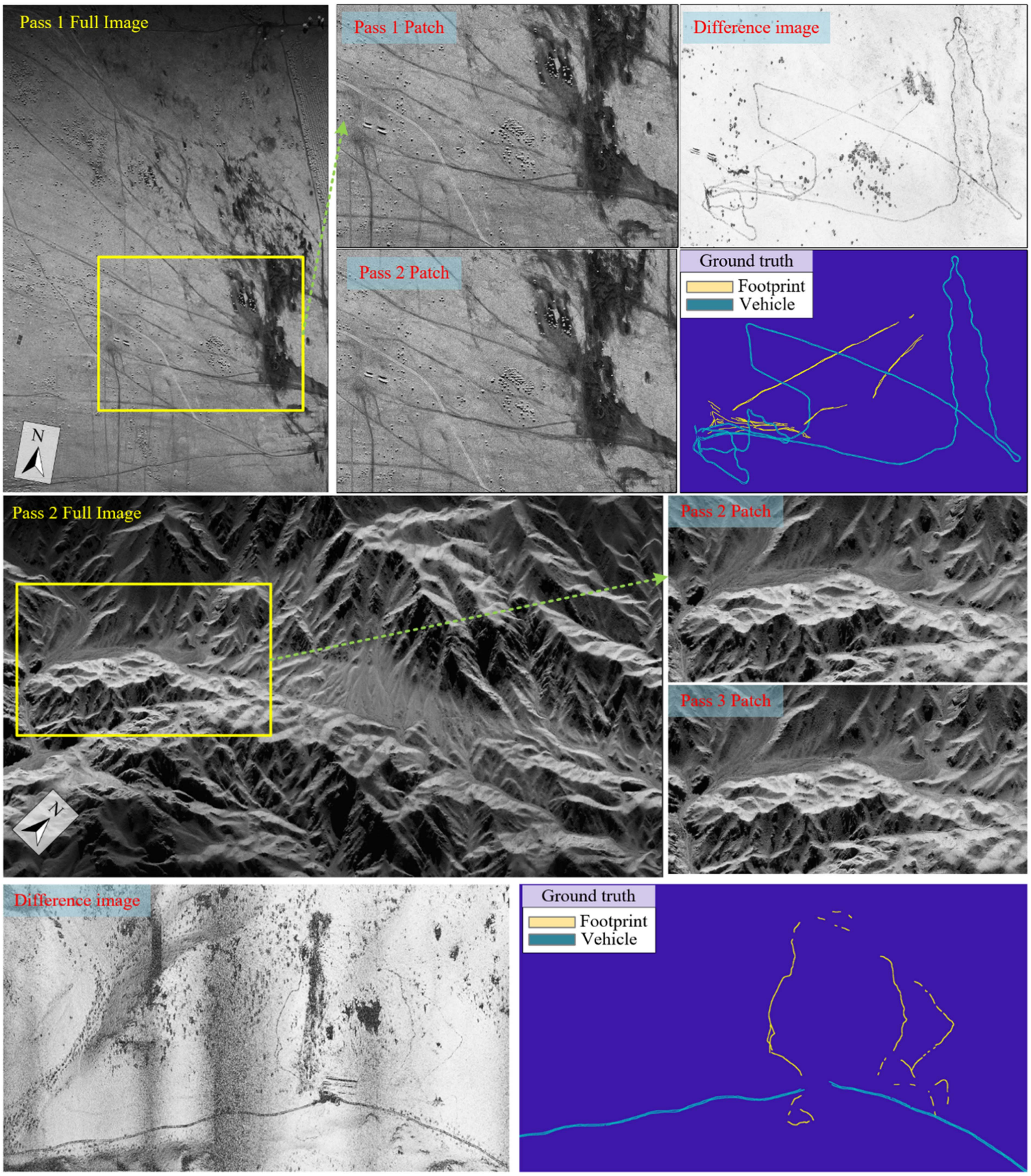


Fig. 10. Training data generation result and corresponding ground truth.

with flipping and rotation. To quantitatively assess the performance of our proposed detection network and other comparison methods, the detection precision, recall, and $F1$-score were used as metrics [41], which are defined as follows:

$$\text{precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{11}$$

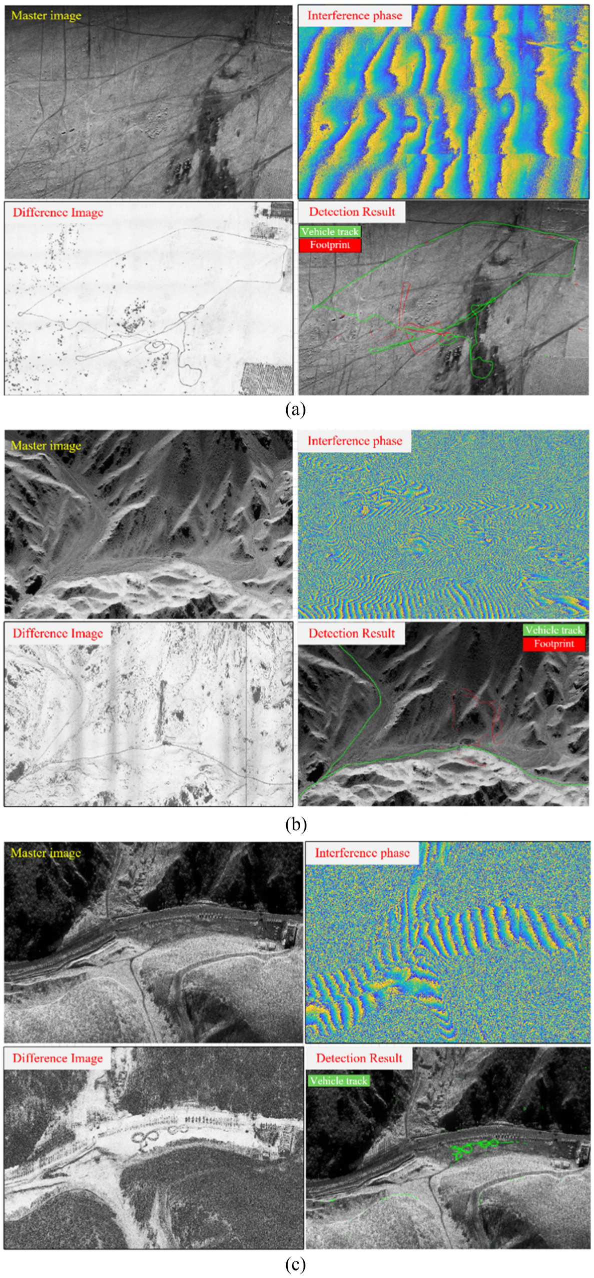
$$\text{recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{12}$$

$$F_1 = 2 \cdot \text{precision} \cdot \text{recall}/(\text{precision} + \text{recall}) \tag{13}$$

where TP and TN are the correctly detected track pixels and background pixels, respectively, and FP and FN are the incorrectly detected track pixels and background pixels, respectively. Since there are two classes of tracks to detect, the above evaluation metrics are computed for each class, and the average metric area is also described in the experimental results.

B. Results of the Proposed Entire Procedure

Utilizing our proposed method to process the testing images shown in Table II, we can attain the corresponding detection results, as shown in Fig. 11. To validate the effectiveness of resolution correction and image coregistration, we initially provide the interference phase image between the repeat-pass images. The interference phase images contain the interference phase fringes between the paired images, and a clearer interference phase fringes correlate to a better coregistration accuracy [42]. We also provide the difference image generated by the coherence estimation, and a brighter difference image correlates to a better coregistration accuracy and coherence estimator. Moreover, we further provide the track detection results obtained by our proposed detection method.



(a)

(b)

(c)

Fig. 11. Detection results from our proposed method on the testing dataset. (a) Track detection results in the sandy region of Baotou city. (b) Detection result in the mountainous region of Tacheng. (c) Detection results in the snowy region of Zhangjiakou.

TABLE III
Detection Results on the Training Dataset

Track	Vehicle			Footprint			Mean
Metric	P	R	F1	P	R	F1	F1
UNet	86.21	61.17	71.50	80.07	48.23	59.67	65.59
UNet++	89.18	63.01	73.76	80.72	55.64	65.65	69.71
LinkNet	83.91	66.62	74.18	67.08	47.34	55.31	64.75
DeepLab V3	71.37	71.17	71.20	71.30	49.63	58.27	64.74
CUNet	82.34	58.56	68.44	65.24	48.07	55.33	61.89
SCNN	88.75	62.08	73.06	81.27	48.96	61.10	67.08
TransUNet	74.98	75.95	73.77	74.89	69.39	71.83	72.80
Proposed	76.06	74.99	74.04	76.72	70.67	73.28	73.66

From the results shown in Fig. 11, the proposed method successfully corrects the resolution and performs accurate coregistration and coherence estimation for all tested regions; this is demonstrated by the clear interference phase images and bright difference images. It is noticeable that the phase jumps in the interference phase of Fig. 11(a) caused by cropping and splicing in the registration process does not affect the track detection result. Moreover, our proposed detection network effectively extracts vehicle and footprint tracks with high precision and recall rates for all regions. Specifically, for the sandy region of Baotou city, the interference fringes in the phase image are very clear, and the brightness of the difference image is very high, indicating the significant preservation of coherence. The detection results for the mountainous region of Tacheng city show that our proposed method can accurately detect footprint tracks on hillsides and vehicle tracks on flat roads between valleys. The results for the snowy region of Zhangjiakou city indicate that despite the heavy interference of vegetation, our proposed detection network can still extract the vehicle track in the flat snowy region. Overall, these results demonstrate the effectiveness and robustness of our proposed method for detecting tracks in different regions and terrains.

C. Detection Results Comparison of the Different Methods

After validating the effectiveness of our proposed preprocessing strategies, such as resolution correction, image coregistration, and coherence estimation, the key problem is to extract track regions with high precision. To compare the detection results from different network-based methods, we evaluated the performance of some typical UNet-based detection methods, including UNet [34], UNet++ [43], LinkNet [44], DeepLabv3 [45], CUnet [23], SCNN [24], and TransUNet [36]. LinkNet consists of encoder and decoder layers, while the UNet structure has an effective encoder and decoder module that is widely used in semantic segmentation tasks. The UNet++ method adds more flexible skip connection layers between the encoder and decoder, increasing the information passing ability. DeepLabv3 uses the atrous spatial pyramid pooling module to capture multiscale context information. Cunet utilizes a compressed Unet to decrease the weighting parameters and SCNN constructs the spatial convolution to improve the global information fitting ability. Finally, TransUNet combines CNN and transformer structures to leverage the global context for semantic segmentation. We evaluated the precision, recall, and F_1 score of each method on the training and testing datasets to compare their performance.

1) Quantitative Results From the Different Methods: Table III shows the detection results obtained by the different methods on the training dataset. LinkNet achieves the highest $F1$ score for vehicle tracks, while our proposed method attains the highest $F1$ score for footprints and the highest average $F1$ score. These results indicate that our proposed method has a strong capability to fit well to the track detection dataset.

Table IV provides the detection results obtained by the different methods on the testing dataset. The detection precision, recall, and F_1 score of all comparison methods decreased with

TABLE IV
DETECTION RESULTS ON THE TESTING DATASET

Track	Vehicle			Footprint			Mean
Metric	P	R	F1	P	R	F1	F1
UNet	75.63	58.07	64.93	58.33	44.26	49.37	56.70
UNet++	78.56	59.43	67.14	58.61	49.08	52.21	58.99
LinkNet	73.26	62.06	66.54	46.13	45.69	44.76	55.86
DeepLabV3	61.96	66.30	63.39	47.15	44.23	44.52	54.67
CUNet	59.60	52.37	55.75	44.68	42.09	43.34	49.55
SCNN	77.54	57.61	66.10	58.50	45.36	51.09	58.59
TransUNet	67.60	66.06	66.55	61.61	57.21	58.76	59.54
Proposed	69.23	69.51	69.13	61.22	58.83	59.39	62.06

TABLE V
DETECTION RESULTS BY THE PROPOSED GCCN ON EACH PAIR OF THE TESTING DATASET

Region	Terrain	Index	Pair	Track	P (%)	R (%)	F_1 (%)
Baotou	Sandy	Pass 1, 2	2	Vehicle	92.89	80.04	85.94
				Footprint	52.93	66.93	59.06
Tacheng	Hilly	Pass 1,2	3	Vehicle	76.10	81.68	78.74
				Footprint	64.72	55.44	59.67
	Hilly	Pass 1,3	4	Vehicle	68.52	71.62	69.99
				Footprint	66.01	54.12	59.42
Zhangjiakou	Snowy	Pass 1, 2	6	Vehicle	39.40	44.70	41.84

different content on the testing dataset than on the training dataset. TransUNet achieves a higher average F_1 score of 59.54% compared to other methods. Among these methods, our proposed GCCN achieves the highest F_1 score for vehicle tracks, footprints, and average score. In particular, the average F_1 score reaches 62.06%, which is 2.52% higher than that of the TransUNet method. Thus, our proposed method exhibits better detection performance than the other methods.

In addition to the results on the entire testing dataset presented in Table IV, we also provide the results obtained by our proposed GCCN method on each image pair of the testing data in Table V.

As shown in Table V, our proposed method achieves a higher detection accuracy and precision for Baotou city, which is covered by sandy regions, than for Tacheng with hilly terrain and Zhangjiakou with snowy conditions. These results demonstrate that tracks in sandy regions are more easily detected. Moreover, the performance of vehicle track detection is much higher than that of the footprints. This occurs because the changed pixels of vehicles are visually prominent and easier to distinguish than those of footprints.

Besides the entire detection results of our proposed detection framework, we also give the results of each component in Table VI. For the baseline method without any azimuth resolution correction, the precision, recall rate, and F_1 score are very poor. With the resolution correction, the average F_1 score is improved from 39.25% to 55.77%, which demonstrates the resolution correction is much an essential step for airborne SAR system under unideal flight conditions. Meanwhile, the image coregistration with optical flow also improves the detection performance. Especially, the proposed GSCM incorporated with TransUNet achieves the highest F_1 score for vehicle track and footprint. Through above analysis, the effectiveness of each components in our detection framework have been verified.

TABLE VI
EXPERIMENTAL RESULTS OF EACH COMPONENT OF THE PROPOSED METHOD (%)

			Baseline	Baseline+different settings		
Azimuth resolution correction			–	✓	✓	✓
Image coregistration with optical flow			Pixelwise		✓	✓
Group spatial convolution module (GSCM)			TransUNet			✓
Metric	vehicle	P	43.12	58.90	67.60	69.23
		R	40.48	60.21	66.06	69.51
		F_1	41.76	59.54	66.55	69.13
	Footprint	P	39.67	52.34	61.61	61.22
		R	34.20	51.66	57.21	58.83
		F_1	36.73	51.99	58.76	59.39
	average	F_1	39.25	55.77	59.54	62.06

2) Visualization Results From the Different Methods: To further understand the detection performance of different methods, we show the results of image pair 2 in Fig. 12. For the master image, the ground truth reflects the vehicle tracks and footprints in the difference image. Although UNet, UNet++, LinkNet, and DeepLabv3 can accurately extract most tracks, some decorrelation regions are incorrectly classified as track regions, and some vehicle tracks are also misclassified as footprints. In comparison, TransUNet detects these tracks with high accuracy. Furthermore, our proposed GCCN method not only extracts more tracks with higher accuracy but also effectively distinguishes between vehicle tracks and footprint tracks. Therefore, the visualization results demonstrate that the well-trained GCCN can effectively extract high-dimensional features from the input images.

V. DISCUSSION

As the sequence of passes is consistent with the sequence of human activities in our outfield experiment in Tacheng city, the track detection results between different passes can reflect the actual human activities during that period. To observe the entire human activity during data acquisition, the automatic track detection results are superimposed onto the generated digital surface model (DSM) image [46], as shown in Fig. 13; this accurately reflects the relationship between the terrain background and human activity tracks. The results are centered on one sheepfold, and the first three images show the detection results, while the last optical image obtained by our optical camera shows the complicated terrain environment, including valleys, peaks, and flat ground in this region.

From the results of passes 1–2, human activity footprint tracks and vehicle movement tracks are observed on the roads between valleys. From the results of passes 2–3, human activity tracks, such as climbing and descending mountains, are present, while vehicles leave unilateral movement tracks. In addition, pass 1–3 result combines the tracks in pass 1–2 and pass 2–3 and includes the complete process of human climbing and descending the mountain. Therefore, through the aforementioned time-series-based DSM detection results, more representations of the human activity process can be obtained. This produces rich and prior information for the determination and prediction

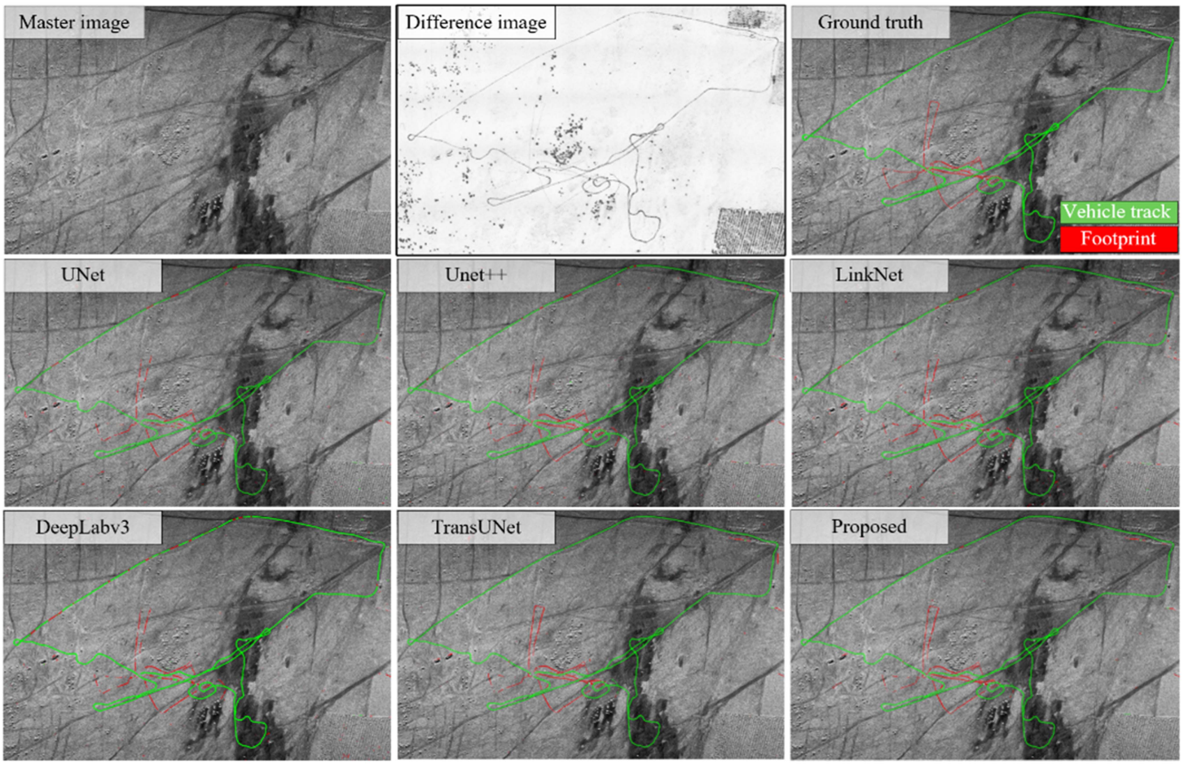


Fig. 12. Detection results from the different methods.

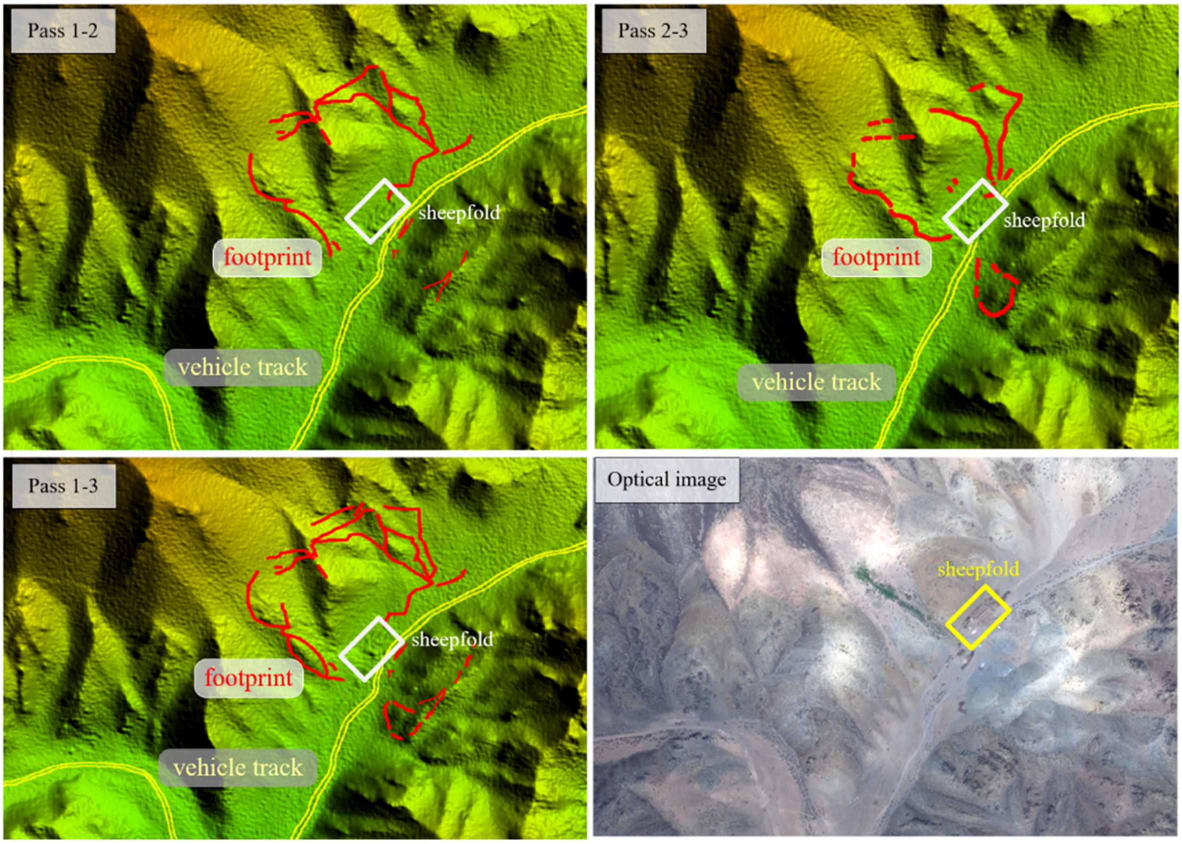


Fig. 13. Detection results over the DSM image of the Tacheng region.

of human activities, such as providing destination references for the search and rescue of lost personnel.

VI. Conclusion

In contrast to traditional RS methods employing image intensity domain change detection for identifying large-scale geophysical alterations, such as floods and earthquakes, our proposed method fully exploits the coherence imaging capabilities of airborne SAR. Utilizing a CCD technique based on a joint amplitude-phase approach, our method effectively enhances the accumulation of subtle differences and constructs a global attention network for the automatic extraction of the subtle change tracks. The process unfolds as follows: initially, a dataset for subtle track detection is assembled across various terrains and temporal phases, utilizing airborne radar systems and ground personnel activities. To tackle the instability issue inherent to repeat-pass data from airborne platforms, a processing flow with the incorporation of the azimuth resolution correction, two-stage subpixel registration, and complex coherence statistics is proposed. This approach facilitates the robust extraction of the difference values from the repeat-pass complex images. In view of the global continuity and locally sparse distribution characteristics of subtle tracks, group spatial convolution is integrated into the transformer-based TransUNet model, enabling the efficient extraction and recognition of subtle change tracks in intricate terrain environments. Experimental results demonstrate that our proposed method exhibits strong detection performance for a range of human activity tracks, such as vehicles and footprints, across diverse terrains, including sandy, mountainous, and snowy areas. Therefore, this method can be extensively applied in environmental monitoring and urban planning contexts, delivering more comprehensive and accurate data support for human activities.

References

[1] J. A. Caraballo-Vega et al., "Optimizing WorldView-2, -3 cloud masking using machine learning approaches," *Remote Sens. Environ.*, vol. 284, Jan. 2023, Art. no. 113332, doi: 10.1016/j.rse.2022.113332.

[2] F. Wu et al., "Built-up area mapping in China from GF-3 SAR imagery based on the framework of deep learning," *Remote Sens. Environ.*, vol. 262, Sep. 2021, Art. no. 112515, doi: 10.1016/j.rse.2021.112515.

[3] B. Adriano et al., "Learning from multimodal and multitemporal earth observation data for building damage mapping," *Int. Soc. Photogrammetry Remote Sens. J. Photogrammetry Remote Sens.*, vol. 175, pp. 132–143, May 2021, doi: 10.1016/j.isprsjprs.2021.02.016.

[4] A. Cerbelaud, L. Roupioz, G. Blanchet, P. Breil, and X. Briottet, "A repeatable change detection approach to map extreme storm-related damages caused by intense surface runoff based on optical and SAR remote sensing: Evidence from three case studies in the South of France," *Int. Soc. Photogrammetry Remote Sens. J. Photogrammetry Remote Sens.*, vol. 182, pp. 153–175, Dec. 2021, doi: 10.1016/j.isprsjprs.2021.10.013.

[5] J. Duan, L. Zhang, M. Xing, Y. Wu, and M. Wu, "Polarimetric target decomposition based on attributed scattering center model for synthetic aperture radar targets," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 12, pp. 2095–2099, Dec. 2014, doi: 10.1109/LGRS.2014.2320053.

[6] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016, doi: 10.1109/TNNLS.2015.2435783.

[7] T.-T. Quach, R. Malinas, and M. W. Koch, "A model-based approach to finding tracks in SAR CCD images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 41–47.

[8] A. W. Doerry, "Collection and processing data for high quality CCD images," Sandia National Laboratories (SNL), Albuquerque, NM, USA and Livermore, CA, USA, 2007.

[9] M. Manzoni, A. Monti-Guarnieri, and M. E. Molinari, "Joint exploitation of spaceborne SAR images and GIS techniques for urban coherent change detection," *Remote Sens. Environ.*, vol. 253, Feb. 2021, Art. no. 112152, doi: 10.1016/j.rse.2020.112152.

[10] J. G. Chow and T.-T. Quach, "Scalable track detection in SAR CCD images," Sandia National Lab.(SNL-NM), Albuquerque, NM, USA, 2017.

[11] F. Cigna and D. Tapete, "Tracking human-induced landscape disturbance at the Nasca lines UNESCO world heritage site in Peru with COSMO-SkyMed InSAR," *Remote Sens.*, vol. 10, no. 4, 2018, Art. no. 572, doi: 10.3390/rs10040572.

[12] D. Fang, X. Lv, Y. Yun, and F. Li, "An InSAR fine registration algorithm using uniform tie points based on Voronoi diagram," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1403–1407, Aug. 2017, doi: 10.1109/LGRS.2017.2715189.

[13] D. O. Nitti, R. F. Hanssen, A. Refice, F. Bovenga, and R. Nutricato, "Impact of DEM-assisted coregistration on high-resolution SAR interferometry," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 3, pp. 1127–1143, Mar. 2011, doi: 10.1109/TGRS.2010.2074204.

[14] R. Touzi, A. Lopes, J. Bruniquel, and P. W. Vachon, "Coherence estimation for SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 1, pp. 135–149, Jan. 1999, doi: 10.1109/36.739146.

[15] H. A. Zebker and J. Villasenor, "Decorrelation in interferometric radar echoes," *IEEE Trans. Geosci. Remote Sens.*, vol. 30, no. 5, pp. 950–959, Sep. 1992.

[16] P. A. Rosen et al., "Synthetic aperture radar interferometry," *Proc. IEEE*, vol. 88, no. 3, pp. 333–382, Mar. 2000.

[17] M. Cha, R. D. Phillips, P. J. Wolfe, and C. D. Richmond, "Two-stage change detection for synthetic aperture radar," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6547–6560, Dec. 2015, doi: 10.1109/TGRS.2015.2444092.

[18] D. E. Wahl, D. A. Yocky, C. V. Jakowatz, and K. M. Simonson, "A new maximum-likelihood change estimator for two-pass SAR coherent change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2460–2469, Apr. 2016, doi: 10.1109/TGRS.2015.2502219.

[19] M. Cha, R. Phillips, and P. J. Wolfe, "Test statistics for synthetic aperture radar coherent change detection," in *Proc. IEEE Statist. Signal Process. Workshop*, 2012, pp. 856–859, doi: 10.1109/SSP.2012.6319841.

[20] G. Yang, H.-C. Li, W. Yang, K. Fu, Y.-J. Sun, and W. J. Emery, "Unsupervised change detection of SAR images based on variational multivariate Gaussian mixture model and Shannon entropy," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 826–830, May 2019, doi: 10.1109/LGRS.2018.2879969.

[21] R. D. Phillips, "Activity detection in SAR CCD," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2013, pp. 2998–3001, doi: 10.1109/IGARSS.2013.6723456.

[22] S. Kuny, H. Hammer, and A. Thiele, "CNN based vehicle track detection in coherent SAR imagery: An analysis of data augmentation," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 43, pp. 93–98, 2022.

[23] J. Zhang, M. Xing, G.-C. Sun, and Z. Wang, "Multiple statistics contributing to few-sample deep learning for subtle trace detection in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5210114, doi: 10.1109/TGRS.2021.3078741.

[24] J. Zhang, M. Xing, G.-C. Sun, and X. Shi, "Vehicle trace detection in two-pass SAR coherent change detection images with spatial feature enhanced U-Net and adaptive augmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5232415, doi: 10.1109/TGRS.2022.3194903.

[25] J. Chen, M. Xing, H. Yu, B. Liang, J. Peng, and G.-C. Sun, "Motion compensation/autofocus in airborne synthetic aperture radar: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 185–206, Mar. 2022, doi: 10.1109/MGRS.2021.3113982.

[26] R. K. Raney, H. Runge, R. Bamler, I. G. Cumming, and F. H. Wong, "Precision SAR processing using chirp scaling," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 4, pp. 786–799, Jul. 1994, doi: 10.1109/36.298008.

[27] Z. Li and J. Bethel, "Image coregistration in SAR interferometry," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 37, pp. 433–438, 2008.

[28] A. Plyer, E. Colin-Koeniguer, and F. Weissgerber, "A new coregistration algorithm for recent applications on urban SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2198–2202, Nov. 2015, doi: 10.1109/LGRS.2015.2455071.

[29] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan, "Change detection based on artificial intelligence: State-of-the-art and challenges," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1688, doi: 10.3390/rs12101688.

[30] S. M. Scarborough et al., "A challenge problem for SAR change detection and data compression," in *Proc. Algorithms Synthetic Aperture Radar Imagery XVII*, 2010, pp. 287–291.

[31] 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, doi: 10.1109/JSTARS.2016.2578362.

[32] A. Plyer, G. L. Besnerais, and F. Champagnat, "Massively parallel Lucas Kanade optical flow for real-time video processing applications," *J. Real-Time Image Process.*, vol. 11, pp. 713–730, 2016.

[33] W. Lu, C. Tao, H. Li, J. Qi, and Y. Li, "A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data," *Remote Sens. Environ.*, vol. 270, Mar. 2022, Art. no. 112830, doi: 10.1016/j.rse.2021.112830.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, and N. Houlsby, "An image is worth 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, May 2021.

[36] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[37] X. Pan, J. Shi, P. Luo, X. Wang, and X. Tang, "Spatial as deep: Spatial CNN for traffic scene understanding," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 7276–7283.

[38] C. Huynh, A. T. Tran, K. Luu, and M. Hoai, "Progressive semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16755–16764.

[39] T.-T. Quach, "Convolutional networks for vehicle track segmentation," *J. Appl. Remote Sens.*, vol. 11, no. 4, 2017, Art. no. 042603.

[40] F. Hu, G. S. Xia, J. W. Hu, and L. P. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, Nov. 2015.

[41] A. Ghosh, B. N. Subudhi, and L. Bruzzone, "Integration of Gibbs Markov random field and hopfield-type neural networks for unsupervised change detection in remotely sensed multitemporal images," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3087–3096, Aug. 2013.

[42] R. Abdelfattah and J. M. Nicolas, "Mixture model for the segmentation of the InSAR coherence map," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2007, pp. 4479–4482, doi: 10.1109/IGARSS.2007.4423850.

[43] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Berlin, Germany: Springer-Verlag, 2018, pp. 3–11.

[44] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process.*, 2017, pp. 1–4, doi: 10.1109/VCIP.2017.8305148.

[45] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.

[46] Y. Sadeghi, B. St-Onge, B. Leblon, and M. Simard, "Canopy height model (CHM) derived from a TanDEM-X InSAR DSM and an airborne Lidar DTM in boreal forest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 1, pp. 381–397, Jan. 2016, doi: 10.1109/JSTARS.2015.2512230.

Jinsong Zhang (Member, IEEE) was born in Shandong, China, in 1995. He received the B.S. degree in remote sensing science and technology and the Ph.D. degree in signal and information processing from the Xidian University, Xi'an, China, in 2017 and 2022, respectively.

He is currently a Lecturer with the Academy of Advanced Interdisciplinary Research, Xidian University. His research interests include synthetic aperture radar (SAR) target classification, SAR coherent change detection, and deep neural network for SAR image processing.

Mengdao Xing (Fellow, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from the Xidian University, Xi'an, China, in 1997 and 2002, respectively.

He is currently a Professor with the National Laboratory of Radar Signal Processing, Xidian University, where he holds the appointment of the Associate Dean of the Academy of Advanced Interdisciplinary Research. He has authored or coauthored more than 200 refereed scientific journal articles. He has also authored or coauthored two books about SAR signal processing. The total citation times of his research are greater than 8000. He was rated as most cited Chinese Researchers by Elsevier. He has achieved more than 40 authorized China patents. His research has been supported by various funding programs, such as the National Science Fund for Distinguished Young Scholars. His research interests include synthetic aperture radar, inversed synthetic aperture radar, sparse signal processing, and microwave remote sensing.

Dr. Xing currently serves as an Associate Editor for radar remote sensing of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Wenkang Liu (Member, IEEE) was born in Dancheng, Henan, China, in 1994. He received the B.S. degree in electronic information engineering from the Xidian University (XDU), Xi'an, China, in 2015, and the Ph.D. degree in signal and information processing from the National Laboratory of Radar Signal Processing, XDU, in 2020, with a focus on the imaging techniques of geosynchronous/medium-Earth-orbit SAR.

From 2019 to 2020, he was a Visiting Student with the Naples University of "Parthenope," Naples, Italy. Since 2022, he has been an Associate Professor with the Academy of Advanced Interdisciplinary Research, XDU. He is responsible for the development of the data processing and calibrating modules for the first GEO SAR to be launched. His research interests include novel spaceborne radar system designing, SAR imaging algorithm development, multipass SAR signal processing, and SAR waveform designing.

Dr. Liu was as the recipient of the "Rising Star of Radar" in 2023.

Guangcai Sun (Senior Member, IEEE) received the master's degree in communications engineering from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2006, and the Ph.D. degree in signal and information processing from the Xidian University, Xi'an, China, in 2012.

He is currently a Professor with the National Laboratory of Radar Signal Processing, and also with the Collaborative Innovation Center of Information Sensing and Understanding, Xidian University. He has authored or coauthored one book and published more than 50 papers. His research interests include imaging of several SAR modes, and moving target detection and imaging.