
Deep Learning Solutions for Persistent Scatterers Detection in Multitemporal Differential Interferometric Synthetic Aperture Radar Monitoring Techniques

Doctorate School in
Information and Communication Technology and Engineering
Dipartimento di Ingegneria
Università degli Studi di Napoli Parthenope
Doctorate Institute in
Institute for electromagnetic sensing of the environment (IREA)

by

Weili Tang



Advisor: Prof. Gianfranco Fornaro
Co-Advisor: Prof. Simona Verde
Coordinator: Prof. Agostino Iadicicco
Year: 2022–2026

Abstract

Multitemporal Differential Interferometric Synthetic Aperture Radar (MT-DInSAR) enables precise long-term monitoring of surface deformation by analyzing the phase evolution of Persistent Scatterers (PS). Conventional PS detection approaches typically rely on statistically rigorous hypothesis testing frameworks, such as the Generalized Likelihood Ratio Test (GLRT), to strictly control the Probability of False Alarm (PFA). However, such conservative, threshold-based strategies often discard valid scatterers that exhibit spatial consistency but reduced local coherence at full resolution. This leads to insufficient density PS, which degrades the spatial continuity and interpretability of the deformation field inversion.

To address this limitation, this study investigates the integration of Deep-Learning (DL) techniques into the PS detection stage and proposes a Three-Stream U-Net (3U-Net) architecture to replace conventional threshold-based detection schemes. Such architecture is able to exploit multi-source features, including statistical indicator for the temporal coherence, the mean deformation velocity (rate), and a local characteristic measuring the signal spatial variability, thus enabling the network to learn discriminative representations in a data-driven manner and reducing dependence on empirically selected thresholds. Proper filtering is also incorporated into the framework based on rate-consistency constraints. Proper loss functions are identified to solve the typical data imbalance.

Experiments are conducted on high-resolution X-band Synthetic Aperture Radar (SAR) data acquired by the COSMO-SkyMed (CSK) constellation and its COSMO-SkyMed Second Generation (CSG). The results demonstrate that the proposed DL-based approach significantly outperforms the conventional GLRT method in terms of PS detection density, structural integrity preservation, and noise suppression, particularly in complex scattering environments.

Furthermore, the proposed framework is extended to dual-polarization Sentinel-1A data (VV/VH), where PS detection is performed through joint exploitation of multi-polarization information. By leveraging the complementary scattering characteristics of different polarization channels, the method is capable of identifying additional stable scatterers in low-coherence and heterogeneous environments, thereby improving detection completeness and robustness.

Experiments are also conducted on simulated data. Together with the results of the analysis of real datasets pertinent to different areas over the Italian territory, we demonstrate that the proposed framework exhibits strong generalization capability across different sensors, polarization modes, and surface conditions. Compared with conventional approaches, the proposed method significantly increases PS density, while effectively preserving spatial consistency and suppressing noise.

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to all my friends and family who have supported me throughout my doctoral studies. I am especially grateful to:

- **Prof. Gianfranco Fornaro**, my supervisor. Over the past three years, he has provided me with invaluable support in both my academic research and daily life. During my time studying under Prof. Fornaro, I have truly experienced the charm of scientific research. He has always answered my questions with great patience, enabling me to gain profound insights each time. His forward-looking guidance has shaped the direction of my academic development. Moreover, his rigorous scientific attitude has set a benchmark for my current and future research career. Beyond academia, he has also shown genuine care for my well-being, allowing me to feel a sense of family even while living abroad. I am always grateful for the opportunity to be his student. Although I may not be his best student, he is the most respected professor in my heart.
- **Prof. Vito Pascazio**, whose warm smile in response to my greetings has always brought me comfort. I am deeply thankful for your kindness and for offering help whenever I needed it, like a caring elder.
- **Prof. Simona Verde**, my co-supervisor. She has patiently guided me in my research and consistently devoted time before every presentation to carefully review my work and listen to my reports. Beyond academia, as fellow women, we have shared many personal moments like close friends. I deeply cherish these wonderful memories. May she rest in peace.
- **Prof. Agostino Iadicicco**, for being an excellent coordinator. Thank you for always responding kindly and patiently to my numerous questions regarding conferences, courses, academic matters, and residence permits.
- **Prof. Diego Reale**, you were the first person I met in the lab and you showed me great care during my induction. Your witty and humorous personality, coupled with your serious attitude towards research, has always been an inspiration to me.
- **Prof. Sergio Vitale**, for his guidance throughout my research journey. He introduced me to the field of deep learning and provided invaluable insights. He has also been extremely supportive in helping me deal with various practical matters, which had been especially important given my limited Italian.

In addition, I would like to express my sincere gratitude to my co-supervisor **Gabriel Vasile**. During my exchange period in France, he shared his expertise with great professionalism. Thanks to his guidance, I was able to successfully and enjoyably complete both my studies and life in France.

I would like to thank my ab colleagues: **Stefano Franceschini**, **Maddalena Autorino**, **Edinson Andres Solarte Casanova**, **Wenyu Yang**, **Yunpeng Zhang**, **Gianpaolo Passarello**, **Francesca**

Razzano, Qiang Han and Zhiqiang Tian. Thank you for the time we shared together, which made my three-year doctoral journey abroad much more colorful and fulfilling. I am equally grateful to all other members of the lab; your support has been a constant source of warmth.

Finally, I would like to give special thanks to my family and friends, who have always supported and encouraged me from behind the scenes. Our conversations during difficult times, your guidance when I felt confused, and your celebrations of my achievements have all been essential sources of strength along the way. Last but not least, I would like to thank myself. I have always believed that courage is one of the most admirable qualities a person can possess. Three years ago, I made a courageous decision, and I hope that in the face of future choices, I will continue to move forward with the same courage.

List of Abbreviations

List of Figures

1	SAR imaging geometry.	17
2	Principles of synthetic aperture radar operation.	18
3	Schematic of InSAR geometric principles.	19
4	Schematic of InSAR geometric principles with deformation.	21
5	Differential Interferometric Synthetic Aperture Radar (DInSAR) Processing flowchart.	22
6	Schematic diagram of a space closed loop.	26
7	Scattering characteristics and phase diagrams of PS and Distributed Scatterers (DS) points. (Left) PS point; (Right) DS point.	28
8	Persistent Scatterers Interferometric Synthetic Aperture Radar (PS-InSAR) processing workflow and baseline network.	29
9	PS-InSAR processing workflow and baseline network.	32
10	Hybrid time-series Interferometric Synthetic Aperture Radar (InSAR) workflow.	34
11	Architecture of the Artificial Neural Network (ANN).	36
12	Convolutional layer computation without padding.	37
13	An example of the max pooling process chain.	39
14	Sigmoid activation function curve and its gradient curve.	40
15	Tanh activation function curve and its gradient curve.	41
16	ReLU activation function and its derivative (curves overlap in the negative region).	41
17	Schematic diagram of the upsampling process.	41
18	Architecture of the Fully Convolutional Network (FCN) model.	43
19	Architecture of the U-Net model.	43
20	Architecture of the Channel Attention Module.	46
21	Architecture of the Spatial Attention Module.	46
22	Architecture of the Convolutional Block Attention Module (CBAM) Module.	46
23	Structure of the proposed 3U-Net algorithm based on the use of the GLRT statistic and the corresponding Displacement Mean Velocity (DMV) (upper image). U-Net architecture (lower image).	56
24	Structure of the Mean Square Error (MSE) module.	57
25	Structure of the Isolated and Outliers Filtering (IOF) module.	58
26	Study areas: Lungro and Salerno.	63
27	PS for the Salerno dataset: Low-threshold with GLRT after IOF filtering (leftmost column) and 3U-Net (rightmost column) results. Zooms on selected areas are shown in the mid columns with GLRT after IOF on the left and 3U-Net on the right.	65
28	PS for the Salerno dataset: Standard threshold with GLRT (mid column), low-threshold with GLRT after IOF filtering (leftmost column) and 3U-Net (rightmost column) results.	66
29	PS for the Lungro dataset: Standard-threshold with GLRT (left column), low-threshold with GLRT after IOF filtering (mid column) and 3U-Net (right column) results.	67
30	Initial label generated with a low threshold (a). Label after initial label with IOF filtering (b). PS velocity corresponding for the initial label (c), for the label after the IOF filtering (d), for the label predicted by 3U-Net (e) and detected with classical GLRT thresholding, I.e. with a high threshold with low PFA (f).	68
31	Probabilities associated with the three-streams outputs: displacement mean velocity DMV, GLRT statistics (b) and entropy (c). The probability to 3U-Net after Weighted Similarity Merging (WSM) module merging is shown in (d).	69
32	PS detection of the proposed 3U-Net architecture without (a) and with the IOF filtering module (b).	70

33	PS corresponding with the DMV detected by a Single-Stream U-Net (1U-Net) (a), by 3U-Net (b).	71
34	DMV (a), Residual Topography (RT) (b), Signal to Noise Ratio (SNR) (c) scenario for the simulated datasets.	72
35	PS results on the Salerno simulated dataset; Subgraphs (a) and (b) shows the PS corresponding DMV provided by the GLRT before and after the detection based on thresholding. Subgraphs (c) and (d) provide the estimated DMV of the PS detected by 3U-Net model trained on Lungro and Salerno real data, respectively.	73
36	Quantitative analysis of the PS detection for the Salerno simulation dataset achieved with the GLRT method based on thresholding and the 3U-Net models trained by Salerno and Lungro real data, separately. PS Probability of Detection vs the SNR (left); Displacement Mean Velocity vs the SNR (right).	74
37	As for Fig. 35 but for the Lungro simulated dataset.	74
38	As in Fig. 36 but for the Lungro simulation dataset.	75
39	Schematic of Different Polarization States	77
40	Study areas: Val d’Agri and Moncenisio areas in Italy. In both images, standard high-resolution DMV product provided by GLRT overlaid to Google Map images.	81
41	PS detection in two areas (top and bottom rows) of Val d’Agri using VV-channel data. The left column shows the results obtained using the GLRT detector, the middle column shows the results obtained using the GLRT with IOF detector, and the right column shows the results obtained using the 3U-Net detector.	83
42	PS detection of Val d’Agri data with VH channel. The left column is for the GLRT detector, the middle column is for the GLRT with IOF detector and the right column is for the 3U-Net detector.	85
44	Differences in PS Point Distribution Between GLRT and 3U-Net Methods Across VV and VH Channels. (a) and (c) show the PS points that can be detected using only the GLRT method for the VH and VV channels, respectively. (b) and (d) show the PS points that can be detected using only the 3U-Net method for the VH and VV channels, respectively.	87
45	PS detection results for the Moncenisio dataset using a model trained on the Val d’Agri dataset for VV and VH channels. (a) GLRT-based PS detection for the VH channel; (b) 3U-Net-based PS detection for the VH channel; (c) 3U-Net-based PS detection only for the VH channel compared to VV channel; (d) GLRT-based PS detection for the VV channel; (e) 3U-Net-based PS detection for the VV channel; (f) 3U-Net-based PS detection only for the VV channel compared to VH channel.	89

Contents

1	Multi-temporal DInSAR on persistent scatterers detection.	16
1.1	Interferometry	16
1.1.1	SAR image	16
1.1.2	Interferometric Phase	18
1.2	Principles of Differential Interferometry	22
1.2.1	DInSAR technology	22
1.2.2	Limitations of DInSAR	26
1.3	Time-series InSAR	27
1.3.1	PS and DS Scatterers in SAR Observations	27
1.3.2	PS-InSAR Technique	28
1.3.3	Small Baselines Subsets (SBAS) Technique	31
1.3.4	Hybrid PS-SBAS Time Series Interferometric Synthetic Aperture Radar (TS-InSAR) Method	34
1.3.5	Chapter Summary	35
2	Theoretical Foundations of Deep Learning-based Semantic Segmentation	36
2.1	Fundamental Components of Convolutional Neural Networks	36
2.2	Basic Components of Convolutional Neural Networks	37
2.3	Research on Deep Learning-Based Semantic Segmentation Methods for Remote Sensing Images	42
2.4	multi-source data fusion based semantic segmentation method	47
2.5	Chapter Summary	48
3	Proposed DL-based PS Detection	49
3.1	The Limitation of Conventional PS Detection	49
3.2	DL Approaches for SAR Data Processing	50
3.3	DL-based PS detection	51
3.4	GLRT method	51
3.5	U-Net	53
3.6	Proposed Solution	54
3.7	Experimental Results	60
3.7.1	Overview of the CSK Satellite System and Data	60
3.7.2	Experiments on Real Data	61
3.7.3	Experiments on Simulated Data	71
3.7.4	Chapter Summary	74
4	Extension of the proposed DL based PS detector on dual-pol sentinel-1 data	76
4.1	Fundamentals of Polarimetric Synthetic Aperture Radar (PolSAR)	76
4.2	Experimental results	79
4.2.1	Overview of the Sentinel-1A Satellite System and Data	79
4.2.2	Experiments on Dual-pol SAR Data	79
4.2.3	Chapter Summary	90
5	Conclusion and Outlook	91
5.1	Conclusion	91
5.2	Outlook	92

Bibliography	93
List of Publications	102

Introduction

SAR is a high-resolution image radar that can produce radar images with optical-level resolution, even in conditions of extremely low visibility. It synthesizes a large aperture through the digital processing of the data acquired by the real small antenna available to the sensing instrument, leveraging of the presence of a relative motion between the radar and its target.

Due to its high-resolution imaging capability and its ability to operate anytime and in all weather conditions, as well as to its effectiveness in identifying camouflage and penetrating cover, SAR is exploited many different areas of the risk monitoring and security. In earth observation, for instance, it is employed for topographic mapping, land use surveys, and geological exploration. It allows terrain features, structures and road networks to be clearly identified. In environmental monitoring, SAR technology tracks environmental changes, such as variations in vegetation coverage and water distribution, and assists in the analysis of pollution and climate change. In military reconnaissance, SAR technology provides high-resolution ground imagery for monitoring enemy deployments and operations. In disaster warning systems, it monitors and assesses events such as earthquakes and floods, providing precise information for rescue and reconstruction efforts. In scientific research, SAR technology supports studies of changes to the Earth’s surface through high-resolution imagery.

DInSAR exploits the temporal baseline of single or multiple interferograms to retrieve phase contributions related to ground motion. However, its accuracy is often constrained by undesired phase terms, such as the Atmospheric Phase Screen (APS), and by signal decorrelation induced by temporal and geometric factors. Persistent Scatterers Interferometry (PSI), an advanced extension of DInSAR, mitigates these limitations by analyzing the phase evolution of stable scatterers across large interferometric datasets. PS correspond to dominant point targets exhibiting high temporal coherence and consistent visibility throughout the interferogram stack. Owing to their robustness against decorrelation effects, their displacement parameters can be estimated with improved reliability. A sufficiently dense distribution of PS points is generally required to effectively characterize the underlying geophysical processes. Consequently, PSI methodologies have been widely applied for deformation monitoring in urban environments as well as in mountainous regions.

Traditional PS-InSAR techniques (PSI, Stanford Method for Persistent Scatterers (StaMPS) and Interferometric Point Target Analysis (IPTA), for example) identify PS based on physical and statistical features such as amplitude and phase stability and temporal coherence. These techniques offer several key advantages, including clear physical mechanisms, strong interpretability, and the ability to reliably and consistently extract large numbers of PS points from highly coherent areas, such as buildings in urban areas and bare rock. This can be achieved without the need for external training data. This has led to the accumulation of extensive engineering experience. However, traditional methods still face limitations in complex terrain environments. PS point selection relies heavily on empirical thresholds, necessitating repeated parameter tuning across different regions and imaging modes. In vegetated or mixed-scatter areas, single statistical indicators struggle to fully characterize long-term scatterer stability, resulting in sparse or missing PS points. Furthermore, traditional methods typically utilize only single pixel or local temporal sequence features, neglecting spatial context and nonlinear

relationship among multi-source features. This makes them susceptible to noise and outlier interference, thereby compromising the spatial continuity and accuracy of deformation field inversion. In contrast, DL methods automatically learn multi-dimensional, high-order features through end-to-end training. They fully integrate spatial neighborhood and temporal sequence information, leveraging spatial homogeneity and local consistency to identify 'quasi-stable scatterers' in complex environments that traditional methods struggle to capture. This significantly improves the density and robustness of PS point distribution, reducing the need for manual thresholds and prior knowledge. Therefore, DL methods offer a novel approach to high-precision PS-InSAR deformation monitoring in complex environments.

This thesis explores the feasibility of using DL techniques to identify PS in multi channel SAR imagery. Representative traditional approaches are also implemented for a comparative analysis, enabling a thorough and objective evaluation.

The structure of this thesis is as follows. Chapter 1 provides an overview of SAR imaging principles, outlining its key characteristics and offering a detailed introduction to SAR interferometry, including its mathematical formulation and practical applications. The subsequent discussion will focus on the utilization of advanced InSAR and DInSAR techniques, with particular emphasis on signal processing strategies and feature representation. In Chapter 2, the fundamental components and architectures of Convolutional Neural Network (CNN) are introduced. The chapter covers commonly used models for classification and segmentation tasks, as well as the associated loss functions. Chapter 3 explores the efficacy of DL-based methodologies for PS detection, encompassing the network architectures, dataset, and experimental evaluations. In conclusion, Chapter 4 delves into the implementation of DL methodologies for deformation measurement through the utilization of cross-polarization SAR datasets.

1 Multi-temporal DInSAR on persistent scatterers detection.

Radar systems estimate the distance to targets by transmitting electromagnetic signals and analyzing the corresponding echoes. In SAR systems, the relative motion of the sensor platform is used to integrate signals acquired from multiple spatial positions coherently, typically through range-Doppler-based processing schemes. As a single SAR acquisition is insufficient to retrieve distance dependent information, conventional InSAR methods use paired SAR images to construct differential interferograms and derive surface deformation, elevation and related geophysical parameters. PS-InSAR builds on this by analyzing long-term sequences of SAR images of the same area, focusing on scatterers that consistently exhibit strong and stable backscattering properties. By monitoring these persistent targets, PS-InSAR enables reliable long-term deformation estimation. Compared to traditional DInSAR approaches, this technique effectively suppresses atmospheric phase delays and mitigates spatio-temporal decorrelation effects included by geometric variations and abrupt surface changes. This substantially improves the precision of InSAR-based deformation measurements.

1.1 Interferometry

1.1.1 SAR image

Radar operates as an active microwave sensing system, emitting electromagnetic waves and recording the echoes reflected from the Earth's surface to obtain terrain-related information. In SAR, signal transmission and reception are associated with the effective antenna length. The antenna utilizes a side-looking geometry to illuminate the ground, thereby generating a narrow radiation beam on the surface. The time delay of the returned echoes is determined by the distance between the ground targets and the antenna; the signals received at different times correspond to the range dimension in the resulting radar image. Concurrently, as the platform progresses along its designated flight path, successive narrow swaths are acquired at regular intervals [Oli91]. The sequential arrangement of these observations is illustrated in Fig.1. In this figure, P_{rg} denotes the range resolution, P_{az} represents the azimuth resolution, β is the beamwidth, D indicates the antenna aperture, and R corresponds to the slant range. The following formulas are used to calculate the azimuth and range resolutions within the resolution call:

$$P_{rg} = \frac{c\tau_{rg}}{2} \quad (1.1)$$

$$P_{az} = \beta R = \left(\frac{\lambda}{D}\right) R \quad (1.2)$$

In the above formula, c represents the speed of light, τ_{rg} represents the pulse width and λ represents the wavelength. Therefore, range resolution is related to pulse width: higher range resolution requires a smaller pulse width. However, the azimuth resolution is related to the aperture of the transmitter antenna. High-resolution azimuth imaging requires a larger antenna aperture. However,

the size of the antenna aperture can be limited by many practical factors, which is why the synthetic aperture radar was developed.

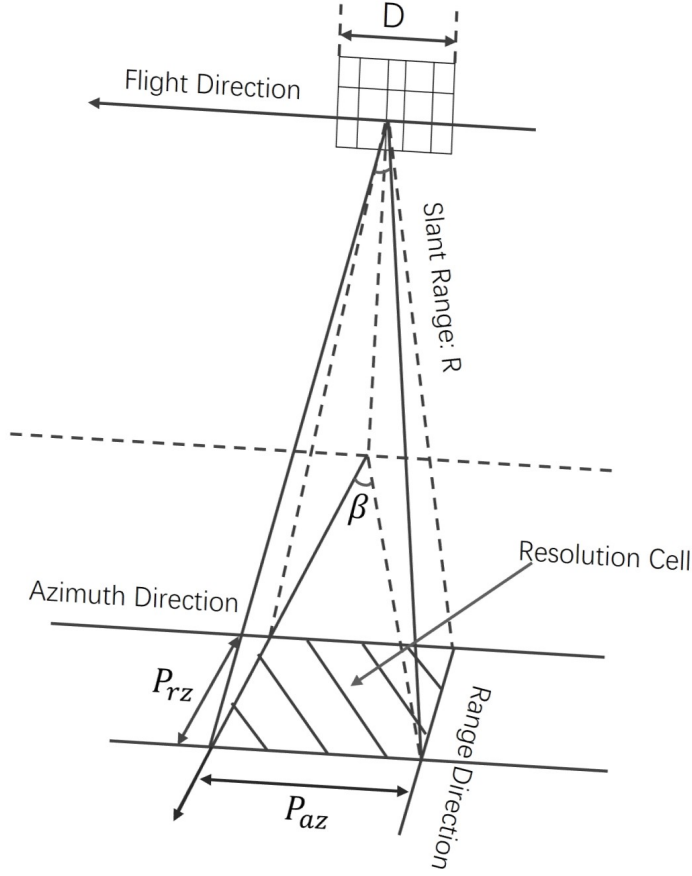


Figure 1: SAR imaging geometry.

SAR simulates the detection capabilities of a large-aperture antenna by moving the radar rapidly, thus increasing the amount of information received and improve azimuth resolution.

For target point P , the antenna begins receiving reflected signals from point P at position X_1 and continues until position X_N . It is evident from the figure's geometry that:

$$L = \beta R = \left(\frac{\lambda}{D} \right) R \quad (1.3)$$

At this point, the equivalent beam coverage is $2L$, and the angle of the side lobe β_s can be expressed as:

$$\beta_s = \frac{\lambda}{2L} = \frac{D}{2R} \quad (1.4)$$

The azimuth resolution of the synthetic aperture radar is therefore as follows:

$$P_{az} = \beta_s R = \frac{D}{2} \quad (1.5)$$

Consequently, the azimuth resolution is improved from D to $D/2$.

The preprocessed radar signal generates a Signal Look Complex (SLC) image, where each pixel can be represented by a complex value P .

$$P = \Re\{P\} + j\Im\{P\} = Ae^{j\psi} \quad (1.6)$$

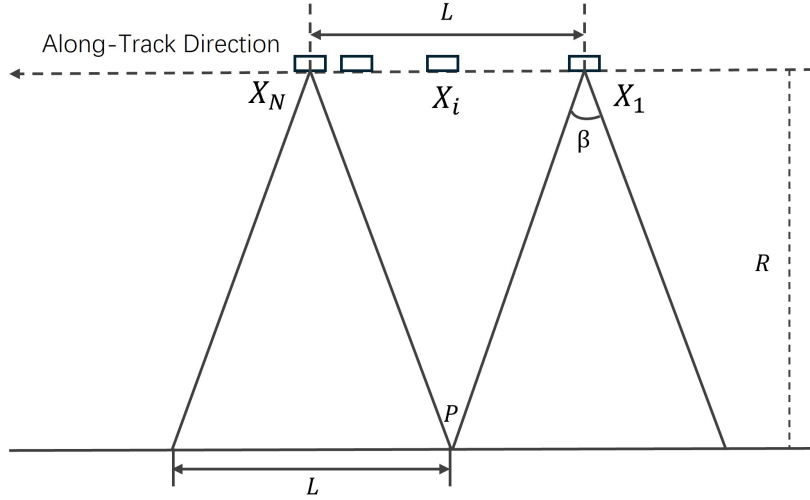


Figure 2: Principles of synthetic aperture radar operation.

In this context, j represents the imaginary unit, the amplitude A is defined as the square root of I , where I denotes the intensity, and ψ signifies the phase information. Intensity and phase information is shown in Fig.2. The symbols $Re(\cdot)$ and $Im(\cdot)$ denote the real and imaginary parts of a complex number, respectively. It is evident that each pixel is capable of containing both the amplitude and phase information of the echo signal. The formation of the final two dimensional radar image is the result of these two components being collectively represented. The interrelationship between the complex value P , A and ψ is as follows[Han01, FPP⁺24]:

$$\Re(P) = A \cos(\psi) \quad (1.7)$$

$$\Im(P) = A \sin(\psi) \quad (1.8)$$

$$A = \sqrt{\Re(P)^2 + \Im(P)^2} \quad (1.9)$$

$$\psi = \arctan\left(\frac{\Re(P)}{\Im(P)}\right) \quad (1.10)$$

Each resolution cell contains the sum of all reflected signals within its range. The intensity of these reflected signals is contingent on the physical characteristics of the ground surface(such as slope, roughness, and surface morphology) and its electronic(dielectric) properties [Ela88].

1.1.2 Interferometric Phase

In the domain of synthetic aperture radar interferometry, the extraction of target information is predominantly reliant on the processing and analysis of interferometric phases at each. Following the accurate registration of two radar images of the same area, processing involves the generation of an interferometric phase map for that region. In practical measurements, the electromagnetic waves from radar sensors are subject to interference from a number of factors, including atmospheric delay, ionospheric disturbances, and decorrelation caused by both angular and temporal diversity between acquisition pairs. Furthermore, the combined effects of surface morphology(reference ellipsoid, terrain undulations) and surface deformation contribute to the interference phase under complex conditions. Consequently, in subsequent processing, to extract target information of interest, it is necessary to remove as much interference phase as possible, thereby achieving the objective of accurately extracting wanted target phase information.

The InSAR geometry of ground target P is illustrated in Fig.3. The interferometric process entails an hermitian product on two registered images Signal Look Complex (SLCs), I.e. the product of one image by the complex conjugate of the other one, expressed as follows:

$$P^{12} = P^1 P^{2*} = A^1 A^2 \exp(j(\psi^1 - \psi^2)) \quad (1.11)$$

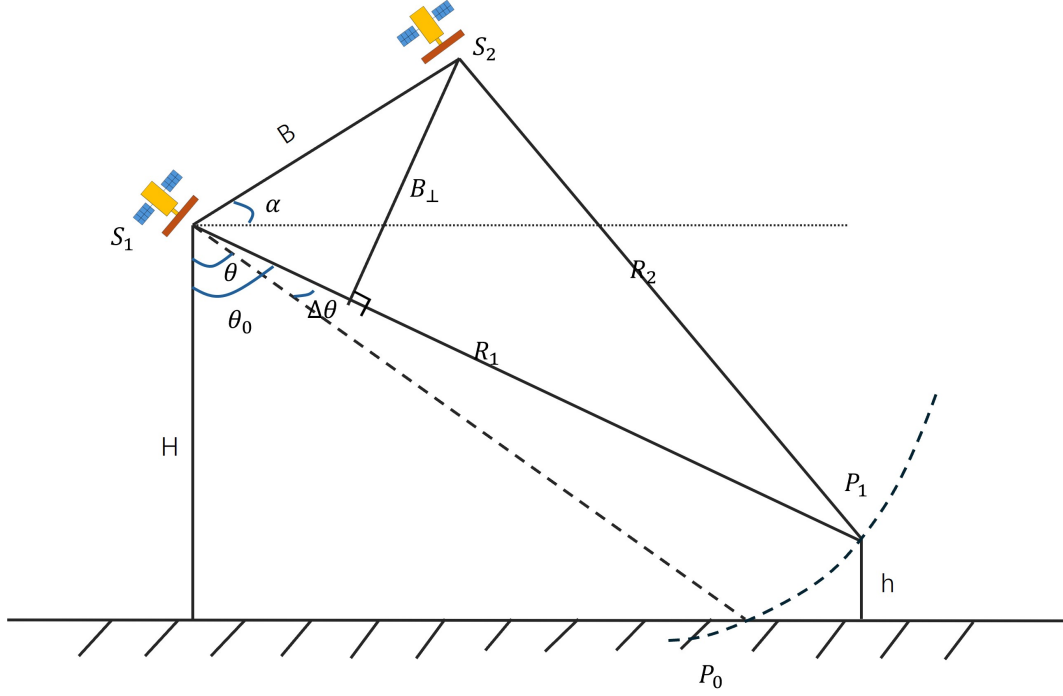


Figure 3: Schematic of InSAR geometric principles.

$$\varphi = \arctan\left(\frac{\Re(P^{12})}{\Im(P^{12})}\right) = W\{\psi^1 - \psi^2\} \quad (1.12)$$

In this context, the superscript '*' denotes the complex conjugate, the superscript represents the image number, φ is the interferometric phase, and $W\{\cdot\}$ denotes the wrapping operation, that is the restriction in the $[-\pi, \pi]$ by adding or subtracting multiples of 2π .

The interferometric phase can be decomposed into several physically meaningful components, including the flat-earth phase, topographic phase, deformation phase, atmospheric phase, and noise term. These components originate from different physical mechanisms, such as the imaging geometry between the radar and the target, atmospheric propagation delays along the signal path, and the electromagnetic scattering properties of the observed surface.

1. Distance-related quantities

As demonstrated in Fig.3, S_1 denotes the first SAR sensor(SAR1) imaging position, S_2 denotes the second SAR sensor(SAR2) imaging position, P represents a ground target, and P_0 is a point on the reference ellipsoid. It is evident that the slant range from S_1 to P is congruent with that from S_1 to P_0 . H denotes the altitude at SAR1 position, h represents the elevation corresponding to P , θ is the incidence angle towards point P_0 and $\delta\theta$ is the difference in incidence angles. The distance from S_1 to P_2 , termed the baseline, is represented by B . The angle between the baseline (B) and designated R_1 and R_2 , respectively. The projection of the baseline B along the slant range R_1 direction is known as the parallel baseline, while the projection perpendicular to the slant range direction is referred to as the perpendicular baseline $B_\perp$.

The interferometric phase φ for this segment can be expressed as:

$$\psi = -\frac{4\pi}{\lambda}(R_1 - R_2) = -\frac{4\pi}{\lambda}\Delta R \quad (1.13)$$

Among these, ΔR represents the difference in the one-way propagation distance of the radar wave. According to the Far-field assumption [ZMLZ10], the slant range difference can be expressed as:

$$\Delta R \approx B \sin(\theta - \alpha) \quad (1.14)$$

Substituting (1.14) into (1.13) yields:

$$\phi = -\frac{4\pi}{\lambda} B \sin(\theta - \alpha) \quad (1.15)$$

As shown in the Fig.3, ΔR consists of two components: the baseline and the elevation at point P . To separate these components, we expand the phase at the reference plane, i.e., $h = 0$, yielding:

$$\phi = -\frac{4\pi}{\lambda} B [\sin(\theta_0 - \alpha) + \cos(\theta_0 - \alpha)(\theta - \theta_0)] \quad (1.16)$$

Based on the distance from the sensor to the reference plane, the geometric relationship shown in the figure yields:

$$h = H - R_1 \cos \theta \quad (1.17)$$

$$\frac{\partial h}{\partial \theta} = R_1 \sin \theta \implies \frac{\partial \theta}{\partial h} = \frac{1}{R_1 \sin \theta} \quad (1.18)$$

$$\theta - \theta_0 = \frac{h}{R_1 \sin \theta} \quad (1.19)$$

It follows that the distance-related phase can be expressed as:

$$\phi = -\frac{4\pi}{\lambda} \left(B \sin(\theta - \alpha) + \frac{B \cos(\theta - \alpha)}{R_1 \sin \theta} h \right) \quad (1.20)$$

(I) Flat-Earth Phase

The first term in (1.20) is denoted as the flat-Earth phase ϕ_f . This component arises due to the change over the range of the off-nadir angle, also referred to as look angle or, in case curvature is negligible, incidence angle. It can be expressed as:

$$\phi_f = -\frac{4\pi}{\lambda} B \sin(\theta - \alpha) = -\frac{4\pi}{\lambda} B_{\parallel} \quad (1.21)$$

(II) Topographic Phase

The second term in (1.20) is referred to as the topographic phase ϕ_h . This component is induced by topographic relief and is a function of both the incidence angle and the terrain elevation. It is given by:

$$\phi_h = -\frac{4\pi}{\lambda} \frac{B \cos(\theta - \alpha)}{R_1 \sin \theta} h = -\frac{4\pi}{\lambda} \frac{B_{\perp}}{R_1 \sin \theta} h \quad (1.22)$$

(III) Deformation Phase If surface deformation occurs between the two acquisitions, the resulting phase component, denoted as the deformation phase ϕ_d , can be written as:

$$\phi_d = -\frac{4\pi}{\lambda} d \quad (1.23)$$

where $d = \delta r$ represents the Line-of-Sight (LOS) displacement, as shown in Fig. 4.

2. Path Dependent Parameters.

The phenomenon of path dependent errors occurs when microwaves emitted by radar sensors penetrate the atmosphere, reach ground targets, and are received as surface reflection waves. The various effects encountered during signal propagation are termed path dependent errors. These errors are primarily constituted by atmospheric phase delay errors caused by ionospheric and tropospheric effects, denoted as ϕ_a . Specifically, when signals cross the ionosphere, electromagnetic waves of different frequencies are affected to varying degrees by the ionospheric medium. This medium exerts a substantial influence on the phase delay, but also as dispersion, of long-wave signals and signals in high-latitude regions. The troposphere, distinguished by its intricate and mutable composition, is subject to the influence of meteorological parameters such as pressure, temperature, and humidity. It is possible to categorize this phenomenon based on its spatial distribution characteristics and physical mechanisms. The

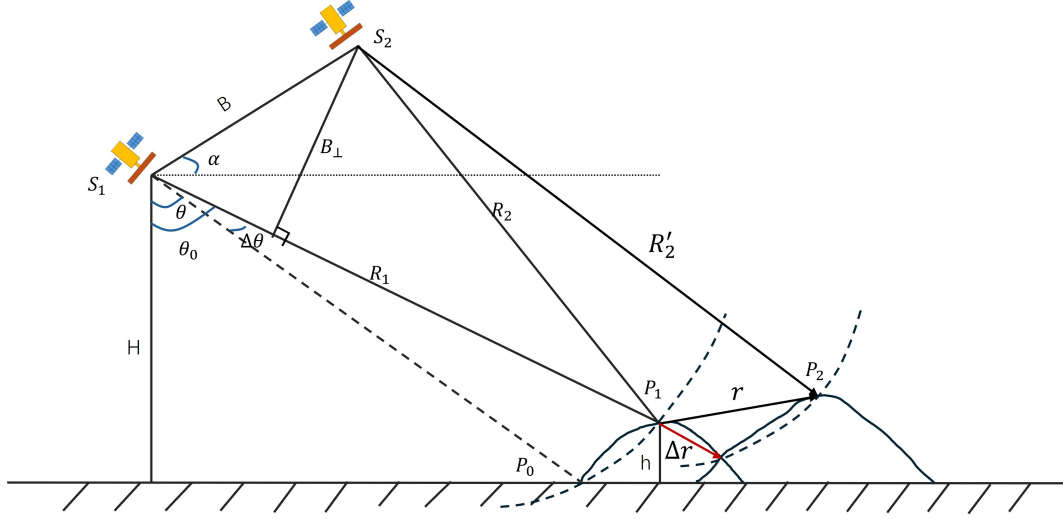


Figure 4: Schematic of InSAR geometric principles with deformation.

categorization of the phenomenon can be subdivided into two distinct types: first, structures that are stratified vertically, and second, turbulent disturbances.

3. Surface Scattering Correlation The component of the interferometric phase relating to surface scattering is derived principally from differences in the scattering characteristics of two echoes, denoted as ϕ_s . It is evident that, owing to variations in observation conditions across different imaging periods and differences in the physical properties of the target terrain, this component cannot be expressed through deterministic parameters. Consequently, the modulus of the complex correlation coefficient $|\gamma| \in [0, 1]$ of the interferometric pair is generally used to quantify the magnitude of this effect. When $\gamma = 1$, and $\phi = 0$, this indicates that the ground scattering characteristics does not change during the first and second imaging frames. In the absence of these conditions, the interferometric pair exhibits varying degrees of coherence decay.

4. Sensor Related Quantities The noise phase ϕ_n of the observation system refers to the thermal noise of the system, jointly determined by the system noise temperature and noise equivalent bandwidth. Similarly to the phase of the scatterer, the manifestation of the noise phase is one cause of phase decoherence in interferometry. This component of error is typically treated as a random error. Combined with errors arising during the interferometric processing, a random error model is employed for compensation.

Concluding, the final interferometric phase can be expressed as:

$$\phi = \phi_f + \phi_h + \phi_d + \phi_a + \phi_s + \phi_n + 2n\pi \quad (1.24)$$

ϕ_f represents the flat-earth phase, ϕ_a denotes the terrain phase, ϕ_d indicates the deformation phase, ϕ_a signifies atmospheric backscatter noise and ϕ_n reflects system noise. n denotes an unknown integer. The above formula represents the general expression for the interferometric phase, with specific variants applied for different target applications. To recover ground elevation, the following formula is applicable:

$$h = -\frac{\lambda}{4\pi} \frac{R_1 \sin \theta}{B_{\perp}} (\phi - \phi_f - \phi_d - \phi_a - \phi_s - \phi_n - 2n\pi) \quad (1.25)$$

In this context, the flat phase ϕ_f is corrected via (1.21), while the wrapping phase $2n\pi$ can be obtained through a non-linear inversion which can be critical in the presence of steep phase variations or too high noise. It can be concluded that $\phi_d = \phi_a = \phi_s = \phi_n = 0$, under the assumption that no deformation occurs between imaging times and disregarding atmospheric and decorrelation noise. In such a scenario, surface elevation information can be obtained directly using the differential interferometric phase ϕ . In the event that the objective is to invert deformation, then:

$$d = -\frac{\lambda}{4\pi} (\phi - \phi_f - \phi_h - \phi_a - \phi_s - \phi_n - 2n\pi) \quad (1.26)$$

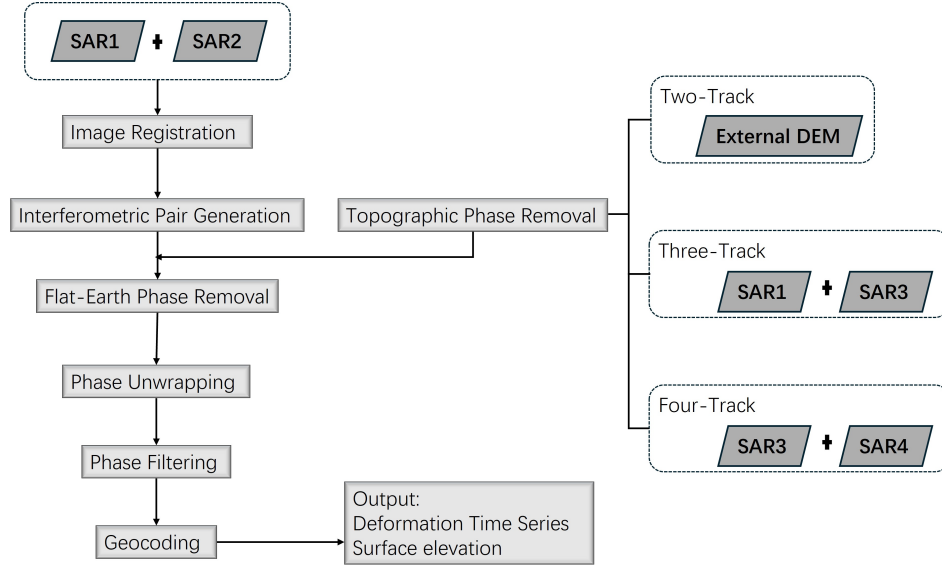


Figure 5: DInSAR Processing flowchart.

As demonstrated by the aforementioned equations, the process of extracting deformation necessitates the separate resolution of each component. A range of InSAR deformation monitoring techniques have been developed, each employing distinct approaches to address the non-deformation phases (noise phases). These approaches encompass traditional differential interferometry, time-series differential interferometry algorithms, and multi-channel data time-series algorithms. The underlying objective of each technological advancement is to eliminate specific noise components, thereby enabling high precision extraction of surface deformation information.

1.2 Principles of Differential Interferometry

1.2.1 DInSAR technology

Differential interferometry can be defined as a process that generates differential interferometric pairs through repeated pass interferometry. The purpose of the process is to reconstruct the deformation information of the target areas. The extraction of SLC imagery is a process which involves a series of processing steps, including image registration, the generation of interferometric pairs, the elimination of terrain phase, phase decoupling, filtering, and geocoding. The flow chart illustrating these processes is presented in Fig. 5. The construction method of differential interferometric pairs forms the basis of the categorization of DInSAR, which is divided into the following methods: The terms "two-track", "three-track" and "four-track" are employed to denote the number of tracks on a record. The primary processing workflows of these three methods are consistent, differing only in the number of images used and the approach to handling the terrain phase, denoted by the symbol ϕ_h . In the course of the terrain phase elimination step, the "two-track" method employs an external Digital Elevation Model (DEM) to obtain the elevation phase for each pixel in the target area involves the subtraction of the elevation phase from the interferometric phase derived from the SAR1 and SAR2 images. The calculation formula for the elevation phase is as follows:

$$\phi_h = -\frac{4\pi}{\lambda} \frac{B_{\perp}}{R_1 \sin \theta} h_0 \quad (1.27)$$

The value of h_0 for each pixel is derived from an external DEM, whose elevation accuracy directly determines the precision of the terrain phase and ultimately impacts the accuracy of deformation inversion.

The implementation of the "two-track" method is subject to certain limitations, primarily due to its reliance on external DEM. The following limitations can be enumerated: Firstly, the accuracy of external DEM frequently falls short of the required standard, with errors subsequently propagating into deformation calculations. Secondly, external DEM data is unavailable for certain study areas, rendering

deformation component monitoring in these regions an urgent issue to address. In order to address this issue, the "three-track" method incorporates a third SAR3 image into the deformation interferogram pair formed by the primary SAR1 and secondary SAR2 images. The approach under discussion involves the calculation of corresponding elevation phases using the interferogram pair formed by SAR1 and SAR3. This process eliminates reliance on external DEM data. Following the processes of terrain processing and phase ambiguity resolution, the subsequent calculation of elevation information is as follows:

$$h_0 = \frac{\lambda R_1 \sin \theta}{4\pi} \frac{1}{B_{\perp}^{13}} (\phi^{13} - \phi_a^{13} - \phi_s^{13} - \phi_n^{13} - \phi_d^{13}) \quad (1.28)$$

$$\phi_h = -\frac{4\pi}{\lambda} \frac{B_{\perp}^{12}}{R_1 \sin \theta} h_0 \quad (1.29)$$

Here, ϕ^{13} represents the interferometric phase formed by SAR1 and SAR3. Assuming no deformation exists in this terrain interferogram and neglecting atmospheric noise phase, we have $\phi_d^{13} = \phi_a^{13} = \phi_s^{13} = \phi_n^{13} = 0$. At this point, ϕ_h is calculated via the equation, subtracted from the deformed interferometric phase, and the resulting deformation is obtained. The "three-track" method operates under the assumption that terrain interferograms contain no deformation information and disregards phase components such as atmospheric delay. In order to satisfy these prerequisites, it is imperative that the observation quality of the two images which constitute the terrain interferogram is of the highest standard. In the event that SAR3 is unable to form a valid interferogram exhibits severe decoherence, it is necessary to acquire an additional terrain interferogram combination. When sufficient data is available, the "four-track" method can be employed to extract surface deformation. The method involves the construction of deformation interferograms using SAR1 and SAR2, whilst simultaneously selecting qualified SAR3 and SAR4 to form terrain interferograms. The principle of terrain elevation calculation remains identical to that of the "three-track" method and is based on independent images:

$$h_0 = \frac{\lambda R_1 \sin \theta}{4\pi} \frac{1}{B_{\perp}^{34}} (\phi^{34} - \phi_a^{34} - \phi_s^{34} - \phi_n^{34} - \phi_d^{34}) \quad (1.30)$$

$$\phi_h = -\frac{4\pi}{\lambda} \frac{B_{\perp}^{12}}{R_1 \sin \theta} h_0 \quad (1.31)$$

Among these, the assumption for terrain interference pairs $\phi_d^{34} = \phi_a^{34} = \phi_s^{34} = \phi_n^{34} = 0$ is more readily satisfied. Therefore, the "three-track" or "four-track" method can be selected based on the actual image quality and quantity to obtain deformation. The aforementioned content differentiates between various methodologies for processing DInSAR elevation phases, namely "two-track", "three-track", and "four-track" processing. Beyond these distinctions, all subsequent steps follow the same processing workflow, as illustrated in Fig.5.

1. Image Registration and Resampling

The generation of interference pairs is achieved by the multiplication of complex conjugates, a process that necessitates processing of the primary and secondary images at identical pixel locations. Consequently, both images must be established within a unified reference frame. In order to achieve this objective, it is customary to first perform image registration in order to estimate the geometric offset between the primary and secondary images. Subsequently, the secondary image is remapped to the primary image coordinated system through resampling, thereby eliminating geometric distortions caused by factors such as incident angle, orbital position, and sampling rate differences. The process of image registration is typically comprised of two distinct stages: coarse registration and fine registration. The initial offsets are estimated using precise orbit information, which forms the basis of coarse registration. Conversely, fine registration is dependent on image intensity data, which is utilized to calculate pixel offsets in both the range and azimuth directions via cross-correlation methods. In practice, a windowing strategy is commonly employed to estimate offsets within local windows, with a two dimensional functional model then fitting the offsets across the entire region [Bro92, SLM⁺03]. In regions characterized by substantial terrain undulations, polynomial models have demonstrated to be incapable of accurately delineating pixel shifts. In such cases, DEM assisted registration methods can be employed, combining external elevation data with orbital information to calculate pixel offsets [AvLGH07, NHR⁺08, NHR⁺10, SBM⁺06]. The incorporation of terrain information has been demonstrated to enhance the registration accuracy of the method, particularly in acenarios where the number

of cross-correlation windows is limited. This enhancement is especially apparent in high-resolution SAR platforms such as TerraSAR-X and CSK. In the context of SLC imagery in swath mode resolution, it is imperative that both distance and azimuth registration accuracy reach approximately 0.1 pixels to circumvent substantial loss of coherence [Han01, JB94]. For scan mode and Sentinel-1's Terrain Observation by Progressive Scans (TOPS) imaging mode, azimuth registration accuracy must reach approximately 0.001 pixels to ensure interferometric coherence [HB02, PISM+12, SJPI+14, FPP+24]. Subsequent to the registration process, sub-images are subject to resampling. The process is initiated by the reconstruction of the continuous signal of the sub-image using a resampling kernel function. Subsequent to this, the sub-image is resampling within the coordinate system of the main image, based on estimated displacement. Common interpolation methods include nearest neighbor interpolation, linear or bilinear interpolation, and high order interpolation functions. It is evident that four-point bilinear interpolation and six-point kernel functions generally meet the practical accuracy requirements.

2. Interferometric Pair Generation

Prior to interferometric processing, SLC images generally necessitate pre-filtering in order to minimize decorrelation noise. A common approach is frequency shift compensation filtering, which corrects the spectrum in both range and elevation directions to mitigate the effects of frequency shifts. This method has been demonstrated to be effective in the suppression of noise from distributed scatterers. However, it has been observed that this process may result in a reduction in image coherence and the introduction of additional noise in resolution cells containing strong scatter points. Interferograms are generated by multiplying the complex conjugates of then registered and resampled primary and secondary SAR images, forming interferometric pairs with interference fringes. It has been demonstrated that complex conjugation in the frequency domain is equivalent to convolution; therefore, the frequency range of interferograms is approximately twice that of single scene SLC images. This can result in the occurrence of spectral aliasing, which manifests as phase noise, particularly evident in short wavelength radar data. Consequently, primary and secondary images should be oversampled by a factor of two prior to interferometric computation. This process serves to enhance the initial resolution of the interferogram, thereby achieving a level that is at most twice that of the original images. Subsequently, low pass filtering is employed to restore the resolution to match that of the SLC images.

3. Terrain and Flat Phase Correction

In order to extract surface deformation information, it is necessary to correct both the flat phase and the terrain phase in the interferograms. The principal objective of this process is the elimination of phase components caused by Earth flat or curvature component and the surface elevation. Subsequently, a reduction in interferometric fringe density will be observed, thus establishing favorable conditions for phase unwrapping and deformation inversion. The flat phase is conventionally linked to baseline geometric parameters and can be calculated directly based on imaging geometry. The terrain phase, however, necessitates estimation and elimination through an external DEM or through techniques such as "two-track", "three-track", or "four-track" phase correction methods.

4. Phase Filtering

Interferograms are susceptible to low SNR due to registration errors, decoherence effects, and system noise. Consequently, the necessity arises for the filtration of the interferometric phase. In contrast to intensity images, the interferometric phase representation is characterized by a wrapped phase, thereby hindering the direct applicability of conventional filtering methods. Complex multi-look averaging is a common approach that suppresses noise by averaging complex signals within local windows, but this simultaneously reduces spatial resolution. The effectiveness of multi-look processing is closely tied to the size of the window; therefore, the adaptive selection of filtering windows based on pixel coherence is employed to balance noise suppression and detail preservation. The complex value resulting from multi-look processing for a single point x can be calculated using the following equation:

$$\langle P_x^{12} \rangle = \frac{1}{L} \sum_i^L P_i^1 P_i^{2*} \quad (1.32)$$

$$\phi_x^{12} = \arctan \left(\frac{\sum_i^L \text{Im} (P_i^1 P_i^{2*})}{\sum_i^L \text{Re} (P_i^1 P_i^{2*})} \right) \quad (1.33)$$

In this context, the symbol $\langle \cdot \rangle$ denotes spatial averaging or multi-look operations, where i represents the index of neighboring points utilized in the averaging process. It has been demonstrated that, for rectangular windows, L is equal to $L_r L_a$, where L_r and L_a denote the range vector and

azimuth vector, respectively. The premise of complex signal multi-look filtering is predicated on two fundamental tenets: Firstly, it is assumed that phase stability is present, meaning phase variations remain relatively smooth over specified time intervals. Consequently, multi-look processing typically occurs subsequent to the cancellation of the flat and terrain phases. In instances where substantial spatial phase gradients are present, phase trend compensation (phase de-trending) must be performed beforehand. Secondly, statistical consistency is assumed, meaning pixels within a window exhibit similar statistical characteristics. In order to satisfy these conditions, the utilization of spatial adaptive windows for the purpose of filtration a common approach. It is evident that the prevailing methodologies encompass techniques founded upon the Speckle-Lee filter (SLf), in conjunction with other adaptive filtering technologies that have been meticulously developed for synthetic aperture radar image sequences [FFN⁺11, JDL13, GA13].

5. Phase Unwrapping

Due to the periodic nature of radar signals, range variations differing by an integer multiple of the wavelength produce identical phase changes. Consequently, SAR observations measure the phase only within a principal interval typically $[-\pi, \pi)$. The process of recovering the continuous phase from the wrapped phase measurements is known as phase unwrapping. The relationship between the true phase and the wrapped phase can be expressed as:

$$\phi^u = \phi + 2n\pi \quad (1.34)$$

Where ϕ^u denotes the true(unwrapped) phase, ϕ represents the wrapped phase, and n is an integer ambiguity that must be determined. (1.34) represents an underdetermined problem with multiple possible solutions. Therefore, additional assumption is that the absolute difference between the true phases of neighboring pixels is smaller than π ,

$$\phi_1^u - \phi_2^u \in [-\pi, \pi). \quad (1.35)$$

which is referred to as Itoh condition.

Under this assumption, the phase difference between adjacent pixels can be used to estimate the unwrapped phase. The relationship between the true phase difference and the wrapped phase difference can then be written as:

$$\phi_1^u - \phi_2^u = W\{\phi_1 - \phi_2\} \quad (1.36)$$

$$\begin{aligned} & (\phi_{i+1,j}^u - \phi_{i,j}^u) + (\phi_{i,j+1}^u - \phi_{i,j}^u) + (\phi_{i+1,j+1}^u - \phi_{i,j+1}^u) + (\phi_{i+1,j}^u - \phi_{i+1,j+1}^u) \\ &= W\{\phi_{i+1,j} - \phi_{i,j}\} + W\{\phi_{i,j+1} - \phi_{i,j}\} + W\{\phi_{i+1,j+1} - \phi_{i,j+1}\} \\ &+ W\{\phi_{i+1,j} - \phi_{i+1,j+1}\} = 0 \end{aligned} \quad (1.37)$$

The simplest phase unwrapping approach is the one-dimensional path-following method, which starts from a reference pixel and integrates the phase differences sequentially along a predefined path. In practice, if the wrapped phase difference exceeds π , 2π is subtracted; if it is smaller than $-\pi$, 2π is added. However, when errors exist in the phase difference estimates, the one dimensional integration method tends to accumulate errors along the integration path, which may significantly degrade the unwrapping accuracy across the entire image. To overcome this limitation, two dimensional phase unwrapping algorithms are commonly adopted. These methods consider the spatial relationships among neighboring pixels and enforce the Zero Closed Loop (ZCL) condition of the true phase field on the field estimated through (1.36), thereby improving the robustness and reliability of phase reconstruction. An illustration of the closed loop phase consistency is shown in 6.

If the closed loop residual calculated from the wrapped phase in (1.36) is non-zero, it suggests that there may be significant phase inconsistencies within the four-pixel loop. Such inconsistencies are typically by strong noise, terrain variations, or surface deformations, all of which render direct path integration unreliable. In such cases, the affected region must be treated separately to prevent error propagation during phase unwrapping. The development of various two-dimensional phase unwrapping algorithms has been driven by different strategies to include Minimum Cost Flow (MCF) [CMM11], branch cutting [GZW88], Least Squares Method (LSM) [GR94], Maximum Likelihood (ML) [SC94], Maximum a Posteriori Estimation (MAP) [GL94], and Bayesian estimation [NPD02].

6. Deformation Estimation and Geocoding

The unfolded phase derived from inversion allows the corresponding surface deformation to be estimated. Initially, the deformation results are expressed in radar coordinate systems(slant range and

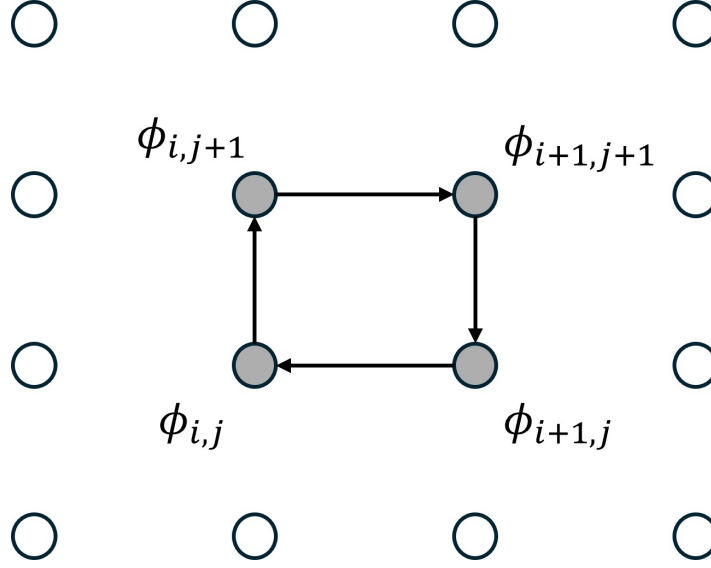


Figure 6: Schematic diagram of a space closed loop.

azimuth coordinates), before being converted to geographic coordinate systems via geocoding. This process yields deformation measurements along the radar LOS. For further analysis, deformation along the LOS can be decomposed into horizontal and vertical displacement components.

1.2.2 Limitations of DInSAR

In DInSAR, surface deformation is retrieved by separating the deformation signal from the total interferometric phase. From a signal processing perspective, this involves isolating the deformation-related components from the mixed phase observations. From a geodetic viewpoint, this can be considered as solving an underdetermined inverse problem. Conventional DInSAR processing typically assumes that sources of error such as atmospheric delay, decorrelation noise, and phase unwrapping errors are negligible. In practice, however, these factors often introduce significant uncertainties into deformation estimates.

Atmospheric delay ϕ_a is one of the dominant sources of error, particularly in regions with complex atmospheric conditions. Since SAR acquisitions are usually separated by several days, variations in water vapor and atmospheric pressure can generate substantial phase artifacts that obscure the true deformation signal. Although multi-pass interferometric strategies can mitigate the effects of the atmosphere to some extent, it is difficult to ensure consistent atmospheric conditions between acquisitions in real observations.

Decorrelation noise ϕ_s is another major limitation. Temporal decorrelation arises from changes in surface scattering properties between acquisitions, whereas spatial decorrelation relates to the interferometric baseline. Large temporal or spatial baselines can significantly reduce interferometric coherence and may even lead to the complete loss of phase information. Furthermore, extensive low-coherence areas complicate phase unwrapping and may introduce errors in deformation retrieval.

Errors in the estimation of the topographic phase ϕ_h can affect deformation measurements. In "two-track" interferometry, external DEM are used to eliminate the topographic contribution. However, residual errors may arise due to the limited resolution and accuracy of DEMs. "three-track" and "four-track" approaches estimate the topographic phase directly from SAR data. However, the reliability to these approaches depends strongly on the quality of the interferograms used for elevation estimation. Additionally, errors in phase unwrapping can propagate into the final deformation results.

Overall, atmospheric artifacts, decorrelation effects, and topographic uncertainties are fundamental limitations of conventional DInSAR. Although careful selection of interferometric pairs can mitigate some of these issues, it is often challenging in practice to fulfill the requirements of short spatial baselines, suitable temporal baselines, and favorable imaging conditions simultaneously, thereby restricting the applicability and accuracy of traditional DInSAR techniques.

1.3 Time-series InSAR

TS-InSAR techniques have been developed to overcome the limitations of conventional DInSAR. As SAR observations become more widely available, these approaches use long sequences of interferometric data over time to suppress various sources of error, including residual topographic errors, atmospheric delays, and decorrelation noise. TS-InSAR enables the more reliable and accurate retrieval of surface deformation by jointly analyzing multi-temporal interferometric measurements. In practical applications, an external DEM is usually used to eliminate the topographic phase, and the associated DEM errors and orbital inaccuracies are explicitly considered. After correcting for the flat earth and topographic components, the interferometric phase can be expressed as:

$$\phi = \phi_\tau + \phi_d + \phi_o + \phi_a + \phi_s + \phi_n + 2n\pi \quad (1.38)$$

Here, ϕ_τ and ϕ_o represent the residual topographic phase and orbital error, respectively, ϕ_d denotes the deformation phase, ϕ_a corresponds to the atmospheric delay, ϕ_s describes decorrelation noise and ϕ_n accounts for system noise. To mitigate decorrelation effects, TS-InSAR generally adopts two complementary strategies. The first focuses on persistent scatterers(PS), which exhibit high-coherence and low decorrelation noise due to their stable backscattering characteristics over long time periods. The second strategy exploits DS, which initially have lower coherence but can achieve improved phase stability through spatial or temporal filtering. This enables the extraction of deformation signals from a larger set of pixels. According to these processing strategies, time-series InSAR methods can be broadly classified into three categories: single master approaches based on PS points, multi master approaches based on DS point, and hybrid techniques combining both scattering mechanisms.

1.3.1 PS and DS Scatterers in SAR Observations

In SAR imagery, the phase measurement of each pixel represents the coherent superposition of the signals backscattered from all the scatterers within a resolution cell. Based on the scattering characteristics within the resolution cell, radar targets can be categorized as either PSs [FPR01b] or DSs [FFN⁺11].

PSs are pixels whose backscattered signal is dominated by a single strong, stable scatterer. In these cases, the echo intensity of the dominant scatterer is significantly higher than that of the surrounding scatterers, and the scattering model can be described by a rice distribution consisting of a deterministic component. Since the phase center of the dominant scatterer remains stable over long temporal baselines, PS pixels exhibit stable phase and amplitude characteristics in SAR image time series. Typical PS targets include artificial structures such as buildings, bridges, and dams, which often act as strong corner reflectors for radar signals. In nature environments, some large, exposed rocks oriented towards the radar's direction of observation may also behave as PS targets. Due to their high interferometric coherence over long temporal and spatial baselines, PS points are particularly well-suited to deformation monitoring in urban environment, forming the theoretical basis of single master TS-InSAR processing frameworks.

By contrast, DSs correspond to resolution cells comprising numerous scattering elements, none of which dominates the radar return individually. The backscattered signal from a DS pixel is the vector sum of echoes from multiple scatterers with similar statistical properties and generally follows a complex Gaussian distribution. DS targets are commonly associated with natural surfaces such as bare soil, desert areas, grasslands, and sparsely vegetated terrain. Compared with PS pixels, DS pixels typically exhibit stronger temporal variability in phase and amplitude due to changes in the distribution of scattering elements. This leads to decorrelation effects. Nevertheless, DS pixels may still maintain relatively high-coherence in an interferometric pair with short spatial and temporal baselines. Consequently, TS-InSAR approaches based on DS pixels usually adopt multi master interferometric networks and employ spatial filtering or statistical estimation techniques to improve phase stability.

Fig. 7 illustrates the variations in signal and phase characteristics between pixels dominated by a PS and pixels associated with DSs. When a dominant scatterer is present within a resolution cell, its strong reflection dominates the radar return and other contributions are considered background clutter. In this case, the phase center remains stable, with only slight fluctuations around the dominant component observed. By contrast, DSs consist of numerous scattering elements with comparable strengths. This results in a radar return formed by the coherent superposition of multiple signals. Consequently, the

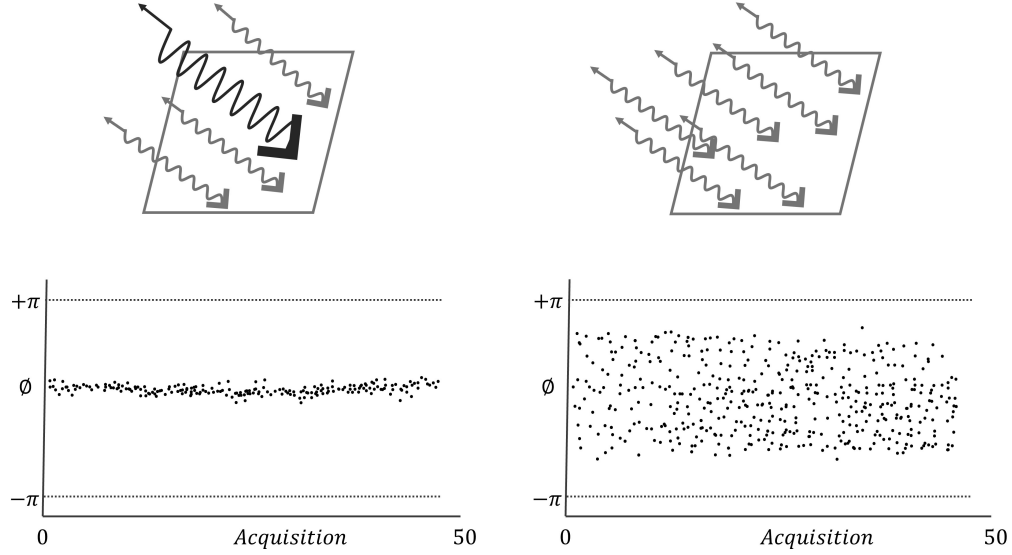


Figure 7: Scattering characteristics and phase diagrams of PS and DS points. (Left) PS point; (Right) DS point.

interferometric phase exhibits strong randomness and is distributed approximately within the interval $[-\pi, \pi]$.

Overall, PS and DS targets exhibit complementary characteristics in deformation monitoring. PS points provide highly precise measurements, but are mainly concentrated in urban areas. In contrast, DS pixels offer broader spatial coverage in nature terrain, but require additional statistical processing to ensure reliable phase estimation. Modern time-series InSAR techniques increasingly exploit both types of scatterer to simultaneously achieve high measurement accuracy and dense spatial sampling of surface deformation.

1.3.2 PS-InSAR Technique

PS-InSAR provides a robust approach for high-precision monitoring of ground deformation by targeting point-like scatterers that remain stable throughout a temporal SAR sequence. These PSs possess strong radar backscatter and preserve elevated coherence over extended observation periods, thereby mitigating the effects of decorrelation noise. Unlike conventional DInSAR methods, PS-InSAR overcomes limitations imposed by spatial and temporal baselines and reduces dependence on strict interferogram selection criteria. Typically, one SAR acquisition is designated as the global reference, and differential interferograms are generated relative to other images in the time series, enhancing data reliability and utilization. Compared to traditional DInSAR, PS-InSAR's key innovation is the systematic identification of permanent scatterers coupled with a temporal analysis framework. Originally introduced by [FPR01b], subsequent developments have focused on optimizing coherent point identification and temporal phase unwrapping strategies. Although implementation details may vary, the essential interferometric processing workflow and solution framework remain largely consistent. Fig.8 depicts the complete PS-InSAR processing sequence and baseline network, with major steps including:

1. Generating Interferometric Pairs

PS-InSAR processing is typically performed using a single-master interferometric configuration. In this configuration, one SAR acquisition is selected as the reference image and the others are used to generate interferometric pairs. The choice of master image is crucial for determining the quality of the interferometric dataset and the reliability of subsequent parameter estimation. In practice, the selection of the master image generally follows three main considerations:

(1) Multitemporal Interferometric Coherence (MIC): PSs exhibit stable backscattering properties over long observation periods, suppressing decorrelation effects in time-series analysis to a large extent. However, an appropriate reference image can further improve the coherence level of the generated interferograms. A common strategy is to select an acquisition located near the temporal midpoint of the dataset, thereby minimizing the temporal baselines between interferometric pairs can be evaluated

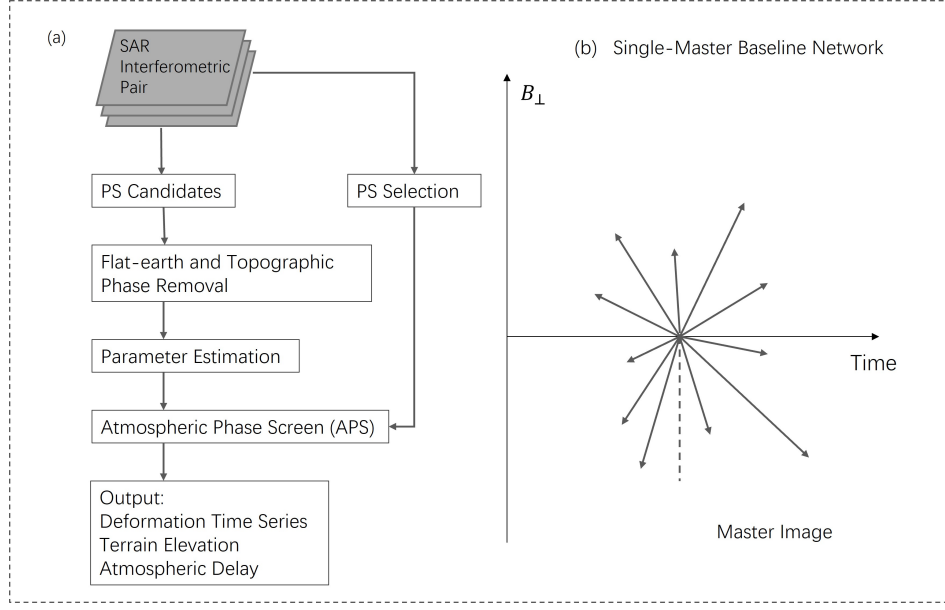


Figure 8: PS-InSAR processing workflow and baseline network.

quantitatively using the empirical relationship proposed by [FPR01a].

$$\gamma_t = \frac{1}{N-1} \sum_{i=1}^{N-1} \left(1 - \frac{|B_{\perp}|}{B_{\perp,c}}\right) \left(1 - \frac{|T|}{T_c}\right) \left(1 - \frac{|f_{dc}|}{f_{dc,c}}\right) \quad (1.39)$$

γ_t represents the temporal coherence coefficient, while $B_{\perp}$, T , and f_{dc} represent the perpendicular baseline, the temporal baseline, and the difference in Doppler centroid frequency, respectively. $B_{\perp,c}$, T_c , and $f_{dc,c}$ correspond to their critical values.

(2) Mitigating the Effects of Atmospheric Delay: In a single reference interferometric network, the atmospheric delay associated with acquiring the reference image is propagated to all interferograms generated from it. If the reference image is acquired under adverse atmospheric conditions, such as intense precipitation or significant humidity variations, substantial atmospheric phase disturbances may be introduced. These disturbances can degrade the accuracy of deformation parameter estimation and complicate the subsequent retrieval of the atmospheric stability during different acquisitions, enabling images obtained under stable conditions to be selected as the reference image.

(3) Reducing Clutter related Phase Noise: During deformation inversion, the contribution of the clutter phase component ϕ_s should be minimized. This noise term is primarily related to temporal changes in surface scattering characteristics, for example, those caused by vegetation growth or land cover variation. Consequently, SAR acquisitions with relatively high Signal-to-Clutter Ratios (SCRs) are preferred when selecting the reference image. In many cases, images acquired during periods of limited vegetation activity, such as winter, provide more stable scattering conditions and therefore help to reduce clutter induced interferometric noise.

2. PS Selection

PS identification constitutes a fundamental stage in the PS-InSAR processing chain, aiming to identify pixels that exhibit temporally stable scattering behavior and reliable phase measurements within SAR time-series datasets. Existing approaches for PS candidate detection can generally be categorized into three groups: the Amplitude Dispersion Index (ADI) method [EM16], the SCRs method [UBF⁺10], and phase stability based approaches [DZZLD15].

(1) ADI Method: In practical observations, the phase SNR is often affected by atmospheric disturbances and surface dynamics, masking direct estimation difficult. [FPR01b] demonstrated that, within a certain SNR range, the temporal variability of pixel amplitudes is closely related to phase stability. Consequently, the ADI can be employed as an indicator of pixel reliability.

$$\sigma_{\psi} \approx \frac{\sigma_A}{m_A} = D_A \quad (1.40)$$

For each pixel, the mean amplitude m_A and standard deviation σ_A are computed from the SAR amplitude time series to derive the dispersion index D_A . Pixels with D_A values below a predefined threshold are considered potential PS candidates. Empirical studies suggest that an ADI threshold between 0.25 and 0.40 generally provides a reliable balance between point density and stability [Hoo06].

(2) SCRs Method: The SCRs-based approach also relies on amplitude information but emphasizes spatial SNR characteristics rather than temporal amplitude variability [AKE04]. Within a predefined spatial window, the SCRs value is obtained by comparing the amplitude of the central pixel with the average amplitude of surrounding clutter pixels. (1.41), Where S_A denotes the central pixel amplitude and C_A represents the background clutter amplitude.

SCRs provides a quantitative measure of the dominance of the target signal relative to background clutter (1.42) and shows a certain correlation with the amplitude dispersion index D_A . Compared with the ADI method, SCRs-based selection can be performed using a single SAR image, which is advantageous when only a limited number of acquisitions are available. However, both ADI and SCRs methods tend to perform best in areas characterized by strong and stable scatterers, such as urban infrastructures.

In practice, radiometric calibration of SAR images is usually performed before the selection of the candidate PS to reduce the influence of sensor characteristics and variations in incidence angle on amplitude measurements, thus improving the robustness and consistency of PS detection [LBM⁺02, KvLMH05].

$$\text{SCR} = \frac{S_A}{C_A} \quad (1.41)$$

$$\sigma_\psi \approx \frac{1}{\sqrt{2\text{SCR}}} = D_A \quad (1.42)$$

(3) Phase Stability Analysis Method: Amplitude based feature selection methods are typically employed for the initial screening of PS candidate points, necessitating future refinement using phase information. Traditional PS-InSAR workflows typically estimate parameters on preselected pixels and conduct secondary screening of PS points based on residual phase. Hooper [HZSK04, HSZ07] demonstrated that, once spatially correlated signals(e.g. atmospheric delay, deformation and orbital errors) and spatially uncorrelated components(e.g. topography-induced residual phase) have been removed, the residual phase can partially reflect the decorrelated noise level of the pixel. The residual phase can therefore be used to calculate the temporal coherence coefficient γ_t , in order to assess the stability phase of the pixel within the image series:

$$\gamma_t = \frac{1}{N} \sum_{i=1}^N \exp \left\{ j \left(\phi - \bar{\phi} - \hat{\phi}_\tau \right) \right\} \quad (1.43)$$

Where N denotes the number of SAR images, $\bar{\phi}$ represents the mean-filtered phase, and $\hat{\phi}_\tau$ is the estimated residual phase of the terrain. A higher temporal coherence coefficient indicates lower decorrelation noise in pixels. Therefore, pixels with γ_t exceeding a threshold are typically selected as the final PS points. This threshold is usually determined through iteration and filtering, with values between 0.4 and 0.6 typically used in practice [Hoo06].

3. Parameter Estimation

The primary objective of time-series InSAR analysis is to retrieve deformation related phase components. Even after correcting major error sources, the observation model remains nonlinear because of phase ambiguities. In the PS-InSAR framework, parameter estimation is typically performed using time-domain complex coherence as a quality metric. A two dimensional spectral search, commonly implemented through the periodogram approach, is conducted within the parameter space to estimate elevation residuals and deformation rates. Assuming that surface deformation evolves linearly over the observation interval, the corresponding phase model can be expressed as (1.44).

$$\Delta\phi_{\text{mod}}^i = -\frac{4\pi B_\perp^i}{\lambda\rho \sin\theta} \Delta\tau - \frac{4\pi}{\lambda} v \Delta t^i \quad (1.44)$$

In this formulation, $\Delta\phi_{\text{mod}}^i$ denotes the modeled phase of the arc segment in the i th interferometric pair, while $\Delta\tau$ and v represent the elevation error and the deformation rate, respectively. The residual phase can subsequently be used to compute the complex temporal coherence (1.45). The periodogram

method determines the optimal parameter estimates by identifying the parameter combination that maximizes the temporal coherence coefficient γ_t without explicitly resolving phase ambiguities.

$$\gamma_t = \frac{1}{N} \sum_{i=1}^N \exp \{j(\phi - \Delta\phi_{\text{mod}}^i)\} \quad (1.45)$$

Building upon this framework, later studies proposed the spatial-temporal wrapping Network (STUN) algorithm. This approach follows a similar time domain search strategy but incorporates integer LSM estimation, where the lambda method is applied to accelerate the search for integer ambiguities. As a result, elevation residuals, deformation parameters, and integer phase ambiguities can be estimated simultaneously.

4. Atmospheric Delay Mitigation

In TS-InSAR analysis, atmospheric effects are typically separated from interferometric phase observations. This procedure is generally performed on unwrapped phase data and relies on the assumed statistical characteristics of atmospheric signals in both the temporal and spatial domains. After removing topographic contributions and modeled deformation components, the remaining phase usually consists of atmospheric delay, unmodeled deformation, and random noise. If atmospheric turbulence is considered temporally uncorrelated, while deformation signals exhibit temporal coherence, temporal high pass filtering can be applied to isolate atmospheric and noise components. Since atmospheric signals often present spatial correlation at local scales, spatial low pass filtering can further differentiate them from stochastic noise.

It should be noted that the assumption of temporal decorrelation applies mainly to turbulent atmospheric components. In contrast, vertically stratified atmospheric delays frequently display seasonal variability and strong dependence on topographic elevation, and are therefore typically addressed using model-based corrections. Consequently, removing atmospheric delay from residual phases is not always a mandatory step in TS-InSAR processing. [Kam06] pointed out that spatio-temporal filtering may also suppress part of the true deformation signal. The STUN algorithm incorporates atmospheric effects partly through a stochastic model rather than applying explicit filtering to the final estimates. In contrast, the StaMPS framework improves phase unwrapping accuracy by estimating the APS, while subsequently restoring this component in the retrieved deformation signal. Excessive smoothing and insufficient filtering may introduce biases in deformation estimation. For this reason, some studies retain atmospheric delay contributions in the final deformation series and quantify their uncertainties using stochastic modeling.

PS-InSAR is a permanent scatterer based TS-InSAR technique, and the reliability of deformation retrieval strongly depends on the spatial density and stability of PS targets. Urban environments typically provide abundant coherent scatterers due to dense man-made structures. However, mountainous or densely vegetated regions often lack sufficient stable reflectors, which limits the effectiveness of PS based deformation retrieval. In addition, certain high amplitude noise pixels may be incorrectly identified as PS points, producing speckle like artifacts in the derived deformation field. These limitations restrict the applicability of PS-InSAR in complex natural environments. In such situations, the SBAS [BFLS03] can be employed as an alternative or complementary strategy to enhance the spatial coverage and stability of deformation estimates.

1.3.3 SBAS Technique

In contrast to the PS-InSAR approach, the SBAS technique does not explicitly differentiate between scatterer types during point selection. Instead, spatial coherence is adopted as the principal criterion, and deformation information is retrieved from pixels that maintain relatively high-coherence within selected interferometric pairs. Because DS generally exhibit stable coherence only within short-baseline interferograms, the single-reference interferometric framework commonly employed in PS-InSAR is less suitable for such targets. To reduce the impact of temporal and spatial decorrelation, SBAS constructs an interferometric network from multiple SAR acquisitions and selects interferogram pairs according to predefined spatiotemporal baseline constraints, typically involving short temporal and perpendicular baselines. This strategy preserves interferometric coherence while allowing a larger number of reliable pixels to be included in the analysis. Consequently, compared to PS-InSAR, SBAS usually provides denser spatial coverage of deformation measurements. To further suppress decorrelation noise associated with distributed scatterers, the SBAS processing chain typically begins with interferogram

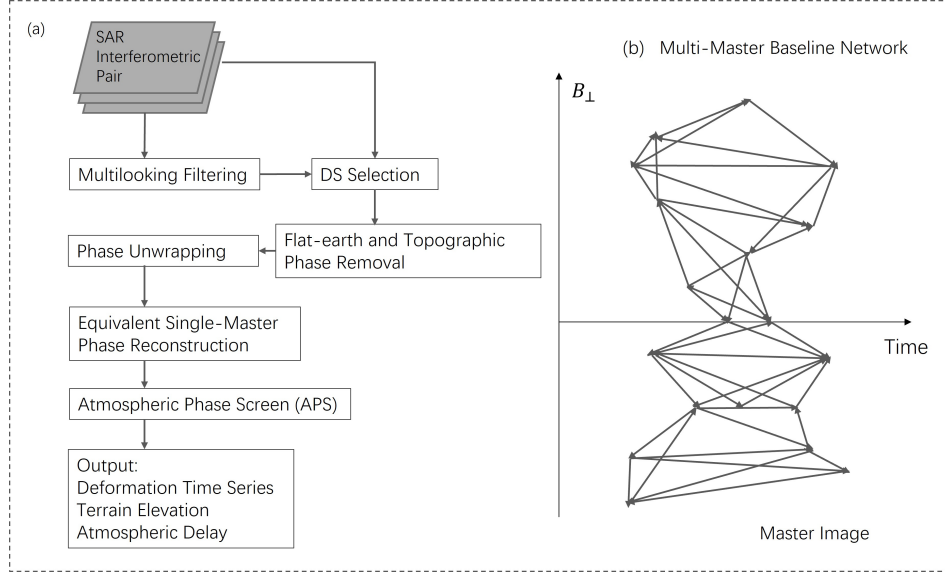


Figure 9: PS-InSAR processing workflow and baseline network.

filtering. In addition to the procedures used in PS-InSAR, the SBAS workflow incorporates an inversion step, which aims to establish temporally consistent phase observations at the different acquisitions starting from the unwrapped interferometric phases. Atmospheric delay effects are generally treated using spatio-temporal filtering or model-based correction approaches similar to those applied in PS-InSAR processing. The overall processing framework and the baseline network configuration of the SBAS method are illustrated in Fig.9, and the principal processing steps are summarized as follows:

1. DS Selection

DSs are commonly identified on the basis of their spatial correlation characteristics. Following the approach proposed by [TLBV02], the spatial correlation coefficient of a pixel, denoted γ_s can be estimated within a local neighborhood window (1.46), where K represents the set of surrounding pixels involved in the calculation. Pixels with spatial correlation values exceeding a predefined threshold are regarded as potential DS candidates.

$$\gamma_s = \frac{\sum_{x \in K} P_x^1 P_x^{2*}}{\sqrt{(\sum_{x \in K} |P_x^1|) (\sum_{x \in K} |P_x^2|)}} \quad (1.46)$$

For multi-temporal interferometric datasets, the identification of DS can also consider the temporal stability of spatial coherence. [Hoo08] introduced a selection strategy based on spatial phase consistency, conceptually similar to the PS identification approach. The main distinction is that this method relies on multi-reference interferograms that have undergone spectral filtering, enabling a more robust evaluation of pixel phase stability.

Additionally, [Hoo08] proposed an amplitude difference dispersion index for preliminary DS candidate screening (1.47). In this formulation, $\sigma_{\Delta A}$ represents the standard deviation of the amplitude differences between the reference and the second images. Compared with the conventional amplitude dispersion index, this metric is derived after spectral filtering of the interferograms, allowing it to better capture the stability of the phase signal. Furthermore, in advanced multi-reference time-series InSAR techniques, such as SqueeSAR, DS candidates are further refined after parameter estimation by analyzing residual phase statistics, thereby enhancing the reliability of the selected targets [FFN⁺11].

$$D_{\Delta A} = \frac{\sigma_{\Delta A}}{m_A} \quad (1.47)$$

2. Interferometric Network Construction

In multi-reference interferometric processing frameworks, the selection and configuration of interferogram pairs play a crucial role in determining the accuracy of deformation retrieval. The primary objective in network construction is to maintain high interferometric coherence while ensuring full connectivity among interferogram pairs, thereby avoiding the formation of isolated subnetworks.

Traditional SBAS implementations typically select interferogram pairs by imposing threshold constraints on temporal and perpendicular baselines, and pairs satisfying these limits are treated as valid observations. Nevertheless, baseline constraints alone cannot fully guaranty interferogram quality. To improve network reliability, [Hoo08] proposed a strategy that evaluates interferogram combinations using integrated coherence metrics, selecting the configuration that maximizes overall network coherence. Alternatively, [RBN06] modeled decorrelation effects using a distance function and constructed an optimized interferometric network through a Minimum Spanning Tree (MST) approach. However additional interferograms are generally added to introduce a redundancy used for checking the unwrapping consistency.

3. Phase Linking

In a multi-reference InSAR processing framework, phase linking is a critical step for achieving a unified phase representation. Its primary objective is to transform the phase information obtained from DSs in a multi-reference interferometric system into a unified single-reference image reference frame, thereby facilitating subsequent parameter estimation and deformation inversion. Depending on the specific algorithm, phase linking can be performed either after phase unwrapping or before it. For example, the SBAS method typically performs phase linking after completing two-dimensional spatial unwrapping, whereas methods such as SqueeSAR can achieve phase reconstruction through phase optimization prior to unwrapping.

In SBAS processing, the interferogram is first subjected to two-dimensional spatial phase de-entanglement. The de-entangled interferometric phase can be expressed as (1.48). Suppose M interferometric pairs observation relationship can be expressed as (1.49) and further rearranged into a linear model from (1.50).

$$E_i = \phi_i^{\text{def}} + \phi_i^{\text{atm}} + \epsilon_i \quad (1.48)$$

$$\mathbf{E} = \begin{bmatrix} \phi_{uw}^1 \\ \vdots \\ \phi_{uw}^M \end{bmatrix} = \begin{bmatrix} 1 & -1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & -1 \\ -1 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} \phi_{uw}^{(1)} \\ \vdots \\ \phi_{uw}^{(N-1)} \end{bmatrix} = \mathbf{B}\mathbf{x} \quad (1.49)$$

where E_i denotes the unwrapped interferometric phase of the i -th interferogram after phase de-entanglement, ϕ_i^{def} and ϕ_i^{atm} represent the deformation phase and atmospheric phase components, respectively, and ϵ_i denotes the residual error term. In (1.49), $\mathbf{E} \in \mathbb{R}^{M \times 1}$ is the observation vector composed of the unwrapped phases of M interferometric pairs, $\mathbf{B} \in \mathbb{R}^{M \times (N-1)}$ is the design matrix describing the interferometric network, and $\mathbf{x} \in \mathbb{R}^{(N-1) \times 1}$ is the vector of unknown phase parameters to be estimated.

$$y = Bx \quad (1.50)$$

Where y is the observation vector, B is the design matrix, and x is the vector of parameters to be estimated. When the design matrix B satisfies the full-rank condition and the interferometric network remains connected, the parameters can be estimated using the LSM. If subnetworks or insufficient connectivity exist among the multi-view observations, leading to a rank-deficient design matrix, Singular Value Decomposition (SVD) can be employed for solution. Furthermore, a weighting matrix can be constructed using interferometric coherence to improve the stability of parameter estimation through weighted LS. It should be noted that phase connection based on de-entangled phases requires high de-entanglement accuracy; therefore, de-entanglement errors may have a certain impact on the final results.

Although the SBAS approach alleviates, to some extent, the limitation of insufficient persistent scatterers encountered by PS-InSAR in complex environments, several constraints remain. First, to reduce spatial decorrelation effects, SBAS commonly applies multi-looking or spatial filtering during interferogram processing. While these operations enhance interferometric coherence, they inevitably lead to a reduction in spatial resolution. Because multi-looking essentially performs spatial averaging, it may suppress the signal contribution of isolated point scatterers. Consequently, SBAS is generally more suitable for detecting broad-area and gradually evolving surface deformation rather than highly localized displacement signals. Second, the reliability of SBAS results strongly depends on the selection and configuration of interferometric pairs. In practice, interferograms are often chosen based

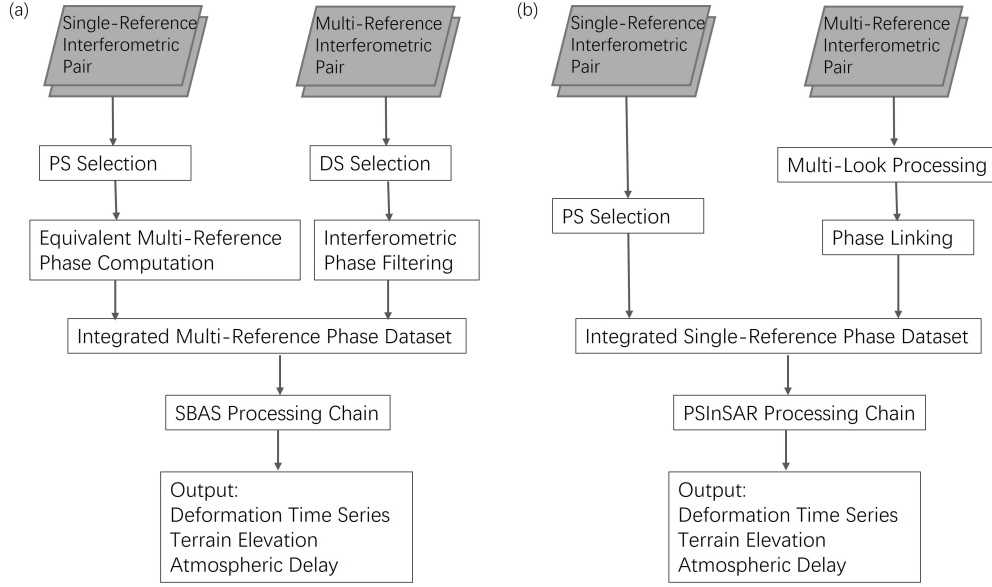


Figure 10: Hybrid time-series InSAR workflow.

on thresholds imposed on temporal and perpendicular baselines. However, such criteria alone cannot always guarantee sufficient interferometric coherence, and manual adjustment or additional quality assessment is frequently required. As the size of the SAR dataset increases, the effort associated with interferogram selection and network optimization can become substantial. Moreover, SBAS processing typically relies on spatial phase unwrapping without explicitly incorporating temporal phase closure constraints. This limitation may introduce discontinuities or abrupt variations in the retrieved deformation time series. Although several improved approaches attempt to identify potential unwrapping errors through temporal loop-closure tests, ensuring the temporal consistency and continuity of phase and deformation estimates remains a fundamental challenge in multi-temporal InSAR analysis.

1.3.4 Hybrid PS–SBAS TS-InSAR Method

In TS-InSAR analysis, the two main approaches PS-InSAR and SBAS offer complementary capabilities for monitoring ground deformation; however, both have certain limitations. PS-InSAR primarily relies on stable PS for high-precision deformation retrieval, making it particularly effective in environments with strong scattering, such as urban areas. However, the limited availability of reliable observations in mountainous or densely vegetated regions is often due to the scarcity of persistent point targets. Conversely, the SBAS approach uses DS to improve spatial coverage, but it is relatively limited in its ability to detect small-scale or highly localized deformation. To overcome these constraints, hybrid time-series InSAR techniques have been proposed that integrate PS and DS observations within a unified inversion framework. This increases the density of valid observation targets while preserving high deformation estimation accuracy. Hybrid time-series InSAR methods can generally be classified into two categories according to their processing strategy. The first category incorporates PS and DS measurements directly within a multi-baseline interferometric framework, performing joint estimation under an SBAS-type observation model. The second category initially processes DS observations and subsequently transforms them into equivalent single-reference measurements. These are then integrated with PS data and analyzed using the conventional PS-InSAR inversion workflow. The corresponding processing schemes are illustrated in 10(a) and (b).

For the first category, Lanari et al. extended the traditional SBAS framework. Two interferogram datasets are generated from the same SAR image stack: multi-look interferograms and full-resolution single-look interferograms. The multi-looked dataset is first processed using SBAS to estimate regional-scale deformation, elevation residuals, and atmospheric phase components [LMM⁺04]. Subsequently, the multi-looked phase is removed from the corresponding full-resolution interferometric phase to obtain residual phases. Pixels exhibiting stable phase behavior are then identified as high-coherence targets and treated as PS candidates for the estimation of local deformation and topographic pa-

rameters. This strategy combines the capability of SBAS to capture large-scale deformation patterns with the advantage of full-resolution interferograms in preserving deformation signals associated with strong local scatterers such as buildings or exposed rock surfaces. Later, Hooper proposed a related approach in which both single-master and multi-looked multi-baseline interferometric sequences are constructed from the same SAR dataset [Hoo08]. PS and DS pixels are identified separately according to spatial phase stability, after which the single-master phase observations of PS targets are converted into equivalent multi-baseline measurements and jointly processed with DS observations within the SBAS framework.

A representative example of the second category is the SqueeSAR technique [FFN⁺11]. This method addresses the limitation of sparse PS targets in non-urban environments by incorporating DS observations into the analysis. Initially, DS targets are processed using a strategy similar to SBAS to estimate their parameters and derive stable phase values based on their statistical characteristics. The resulting multi-baseline observations are then transformed into an equivalent single-master reference framework, effectively compressing the multi-baseline information into a single-reference interferometric representation. These DS observations are subsequently integrated with PS data and processed within the PS-InSAR inversion scheme. By jointly exploiting both PS and DS scatterers, this approach substantially increases the spatial density of effective observation targets while maintaining high deformation retrieval accuracy, thereby improving robustness in complex terrain environments.

1.3.5 Chapter Summary

This section presents a systematic analysis and rigorous derivation of the imaging geometry of InSAR and its principle error sources, based on the fundamental theory of SAR. Building upon this theoretical framework, the scattering characteristics of DSs and PSs are comprehensively examined. To this end, two representative multi-temporal interferometric techniques-PSI and the SBAS approach are considered as canonical frameworks. The corresponding parameter inversion procedures associated with these scattering mechanisms are analyzed in detail, including the estimation and modeling of key geophysical parameters such as RT, surface deformation, and Atmospheric Phase Delay (APD).

2 Theoretical Foundations of Deep Learning-based Semantic Segmentation

Artificial Neural Networks (ANNs) are computational frameworks inspired by the information-processing mechanisms of the human brain, formulated as abstract mathematical models of inter-neuronal signal propagation. They comprise densely interconnected processing units (neurons), each associated with a nonlinear activation function, while the connections between units are parameterized by weights that encode the magnitude and direction of information flow. A canonical ANN architecture consists of an input layer, one or more hidden layers, and an output layer. The input layer serves as the interface for raw data ingestion. Hidden layers, formed by multiple neurons, perform affine transformations on incoming signals followed by nonlinear activation, enabling hierarchical feature extraction and progressive abstraction of data representations. The output layer produces task-specific predictions based on the learned feature mappings. Model training is formulated as an optimization problem driven by a predefined loss function. Parameters are iteratively adjusted via the backpropagation algorithm, which propagates error gradients through the network to refine connection weights, thereby reducing prediction discrepancies and improving generalization performance.

2.1 Fundamental Components of Convolutional Neural Networks

Convolution Neural Networks (CNNs) constitute a principal class of ANNs specifically developed for image analysis, with the capacity to model data exhibiting inherent spatial structure. Unlike conven-

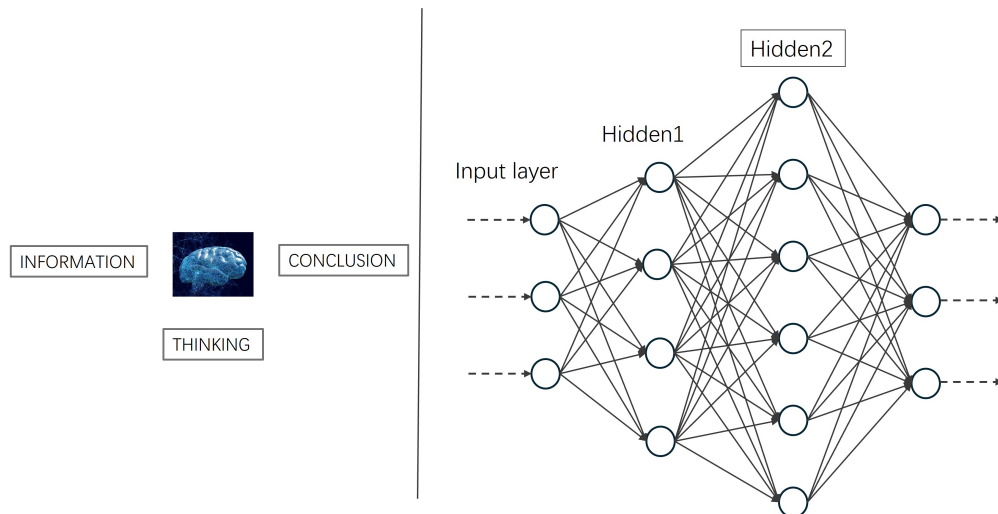


Figure 11: Architecture of the ANN.

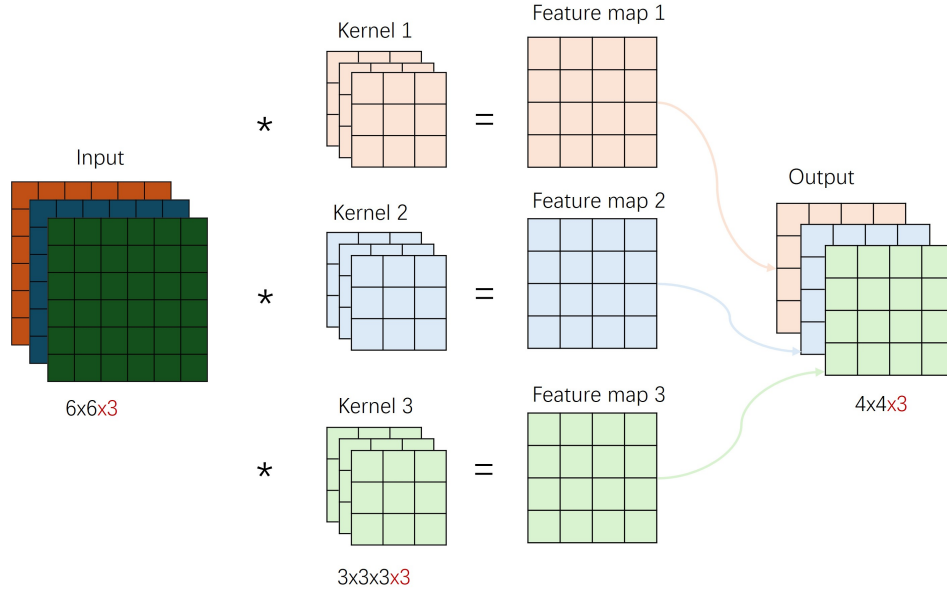


Figure 12: Convolutional layer computation without padding.

tional fully connected architectures, CNNs encode intermediate representations as three-dimensional tensors defined by spatial dimensions (height and width) and channel depth, thereby enabling structured, locality-preserving mappings between inputs and outputs.

From an architectural standpoint, CNNs comprise a set of canonical building blocks, including convolutional, pooling, normalization, and fully connected layers. Convolutional layers exploit local receptive fields and weight sharing mechanisms to capture spatially correlated patterns. The pooling layers perform spatial downsampling, reducing feature resolution while enhancing translational invariance and computational efficiency. Normalization techniques (e.g., batch normalization) stabilize the gradient flow and accelerate convergence during training. When included, fully connected layers act as global aggregators for high-level feature integration and task-specific inference.

For dense prediction tasks such as semantic segmentation, CNNs are commonly instantiated as Fully Convolutional Networks (FCNs), in which fully connected layers are removed. This configuration preserves spatial resolution throughout the network and enables end-to-end pixel-wise prediction, making it particularly suitable for fine-grained annotation tasks.

2.2 Basic Components of Convolutional Neural Networks

CNNs are among the most widely adopted DL architectures for visual representation learning. By exploiting local receptive fields, parameter sharing, and hierarchical feature extraction, CNNs effectively capture the spatial dependencies inherent in image data while substantially reducing computational complexity and the number of trainable parameters. A typical CNN comprises a sequence of convolutional layers, nonlinear activation functions, pooling operations, and normalization layers.

1. Convolutional Layer

The convolutional layer is the core computational unit of a CNN, primarily responsible for local feature extraction. Its parameters consist of multiple learnable convolutional kernels, each with a finite spatial extent and spanning all channels of the input feature map, enabling the joint modeling of local spatial patterns, and cross channel information. During the forward pass, each kernel slides across the input feature map and performs convolution over local receptive fields to produce a corresponding feature map. The resulting feature maps are stacked along the channel dimension to form the output of the convolutional layer, as illustrated in Fig. 12. Convolutional layers possess two key characteristics: local connectivity, which limits the receptive field to preserve local spatial structures while reducing computational complexity, and parameter sharing, whereby the same convolution kernel is reused across different spatial locations, significantly reducing the number of trainable parameters and improving feature learning efficiency.

The convolution operation can be formulated as a process that computes a linear mapping at each spatial location of an image. This is typically implemented via inner product operations, involving the computation of the dot product between the input feature map and a convolution kernel. Assume that the input feature map has spatial dimensions $H \times W$ with C channels and that N convolution kernels are applied, each with spatial size $K_H \times K_W$. The response at spatial location (i, j) for the n th convolution kernel, obtained by sliding the kernel across the input feature map, can be expressed as:

$$y_{i,j,o} = \sum_{m=1}^{K_H} \sum_{n=1}^{K_W} w_{m,n,o} \cdot x_{i-\lfloor \frac{K_H}{2} \rfloor + m, j-\lfloor \frac{K_W}{2} \rfloor + n} + bias_{m,n,o} \quad (2.1)$$

In the above formulation, $w_{m,n,o}$ denotes the convolution kernel(weights) associated with the o -th output channel, while x represents the input feature map. The bias term is denoted by b_o , and $y_{i,j,o}$ corresponds to the output feature response at spatial location (i, j) in the o -th channel. During the convolution operation, the spatial dimensions of the output feature map are determined by kernel size, stride, and zero-padding. Assuming a stride of S and a padding size of P , the spatial dimensions of the output feature map $H_{out} \times W_{out}$ and the number of output channels C_{out} are given by:

$$H_{out} = \frac{H - K_H + 2P}{S} + 1 \quad (2.2)$$

$$W_{out} = \frac{W - K_W + 2P}{S} + 1 \quad (2.3)$$

$$C_{out} = N \quad (2.4)$$

CNNs typically consist of a hierarchy of convolutional layers. By combining convolutional operations with nonlinear activation functions, CNNs learn complex nonlinear feature mappings and progressively construct hierarchical feature representations. Shallow layers primarily capture low level visual features, such as edges, corners, and textures, whereas deeper layers progressively aggregate and abstract these representations into high level semantic features. This hierarchical abstraction enables CNNs to effectively capture complex visual patterns and semantic relationships in image data.

2. Pooling Layer

CNNs typically include pooling layers following several convolutional layers. These layers perform spatial downsampling on feature maps. Pooling operations are applied independently to each channel (depth slice). A common configuration uses a 2×2 pooling window, which reduces the spatial resolution of the feature map by half while significantly decreasing computational cost and the number of redundant features by reducing the spatial dimensions (height and width) by a factor of two. Common pooling methods include average pooling and max pooling. Average pooling computes the mean of a local region to aggregate features, whereas max pooling selects the maximum activation within a local region to preserve the most salient feature responses. Max pooling generally exhibits stronger discriminative capability and robustness in practical applications, making it the more widely adopted choice in contemporary deep learning models. Average pooling, in contrast, is often used in specific scenarios or as a supplementary mechanism.

1)Enhanced scale invariance: Pooling operations reduce spatial resolution by downsampling feature maps. If salient information is preserved during this process, the resulting representations exhibit a degree of scale invariance, thereby improving the model's robustness to variations in input scale.

2)Mitigation of overfitting: For instance, a 2×2 pooling window with a stride of 2 effectively compresses feature responses, substantially reducing the number of activations in the feature map (approximately 75% are reduced or discarded). This decreases the parameter scale and representational redundancy in subsequent layers, while preserving critical discriminative information, thereby improving generalization performance.

3)Reduction of computational and memory overhead: By compressing the spatial dimensions of feature maps, pooling layers significantly reduce both memory footprint and computational complexity associated with intermediate representations. This facilitates the construction of deeper network architectures under resource constraints, thereby enhancing the overall representational capacity of the model.

3. Batch Normalization Layer

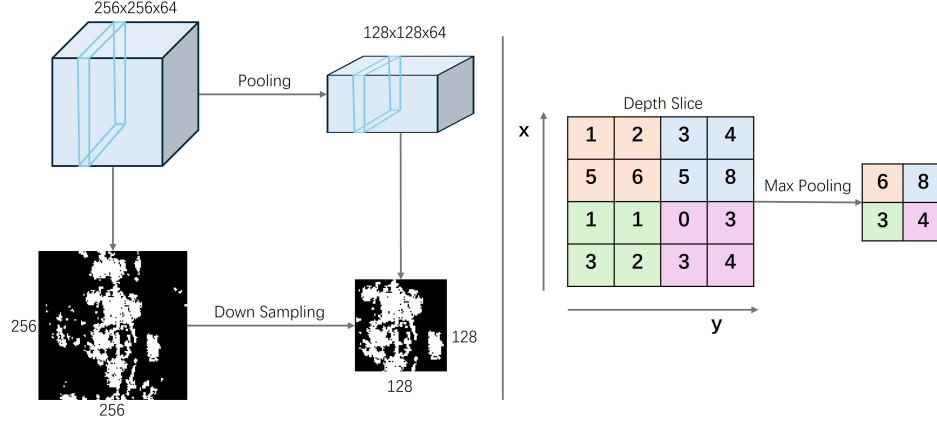


Figure 13: An example of the max pooling process chain.

In practical network architectures, the convolutional layer, Batch Normalization Layer (BN), and nonlinear activation function are typically integrated as a single convolutional layer and the activation function. During the training of deep neural network, updates to model parameters can induce shifts in the distribution of intermediate feature representations across layers, a phenomenon known as covariate shift. The Covariate shift violates the Independent and Identically Distributed (IID) assumption commonly adopted in machine learning, thereby increasing the optimization difficulty and affecting convergence stability. BN addresses this issue by normalizing the outputs of each layer to have zero mean and unit variance, thereby approximating a standard normal distribution. This normalization stabilizes feature distributions and encourages activations to fall within regions where nonlinear activation functions exhibit higher gradient sensitivity. Consequently, BN helps mitigate the problem vanishing gradients and accelerates convergence during training. Furthermore, to preserve the representational capacity of the network, BN introduces a learnable scaling γ and shifting β parameters after normalization. These parameters apply an affine transformation to the normalized output, enhancing the model's flexibility while maintaining numerical stability. The detailed formulation of batch normalization is given as follows:

$$\text{BN} = \gamma \frac{x - \mu}{\sigma} + \beta \quad (2.5)$$

$$\mu = \frac{1}{M} \sum x_i, \quad \sigma = \sqrt{\frac{1}{M} \sum (x_i - \mu)^2 + \epsilon} \quad (2.6)$$

In this formulation, x denotes the convolutional output for each sample, and M denotes the batch size. The parameters μ and σ represent the mean and variance of the convolutional outputs, respectively, with γ and β corresponding to the learnable scaling and shifting parameters. The term ϵ is introduced as a small constant to ensure numerical stability and prevent division by zero.

4. Nonlinear Activation Layer

Convolution is inherently a linear transformation; even when multiple convolutional layers are stacked, the overall operation remains equivalent to a single linear mapping. Consequently, CNNs incorporate nonlinear activation layers to transform their representations, thereby enhancing the model's capacity to capture complex, nonlinear features. These layers typically consist of activation functions that take the feature maps produced by convolutional layers as input and generate corresponding activation maps as output. This process is applied element wise, preserving the spatial dimensions of the input. As activation functions do not involve learnable parameters, they are generally not treated as independent layers but rather as essential transformations following convolution.

An effective activation function is expected to satisfy the following properties:

1) Nonlinearity, to improve the model's representations capacity; 2) Continuity and differentiability, ensuring compatibility with gradient based optimization methods; 3) A sufficiently wide non-saturated regime, to prevent activations from entering regions where gradients vanish, thereby mitigating the vanishing gradient problem in deep networks; 4) Monotonicity, which contributes to favorable optimization characteristics of the loss landscape in shallow models; 5) Approximate linearity near the

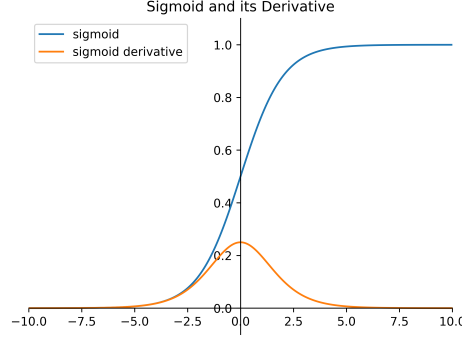


Figure 14: Sigmoid activation function curve and its gradient curve.

origin, which facilitates stable training and accelerates convergence during parameter initialization.

Common activation functions in CNNs include the sigmoid function, The hyperbolic tangent (Tanh) function, and the Rectified Linear Unit (ReLU), whose mathematical formulations are given as follows:

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (2.7)$$

$$\tanh(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}} \quad (2.8)$$

$$\text{ReLU}(x) = \max(x, 0) \quad (2.9)$$

(2.7), (2.8) and (2.9) illustrate the activation functions and their corresponding gradient profiles. Based on their functional characteristics, the performance differences among these activation functions can be analyzed as follows:

The sigmoid function exhibits a relatively narrow non-saturated regime, with significant gradients primarily within the interval $[-6, 6]$. Outside this range, i.e., in the saturated regions $(-\infty, -6)$ and $(6, +\infty)$, the gradient rapidly approaches zero. Consequently, in deep neural networks, it is prone to inducing vanishing gradients, which hinders effective parameter updates. In addition its output range $(0, 1)$ is not zero-centered, leading to bias shift and adversely affecting optimization dynamics. These limitations significantly constrained training performance in early neural network research.

The Tanh function shares a functional similar to that of the sigmoid function and likewise suffers from a limited non-saturates regime. However, its output range $(-1, 1)$ is zero centered, partially alleviating bias shift and facilitating more stable gradient propagation during optimization.

In contrast, the ReLU introduced by Nair and Hinton [NH10], has become one of the most widely adopted activation functions due to its superior empirical performance. The function outputs zero for negative inputs and exhibits linear behavior for positive inputs, with a constant gradient of 1 in the positive regime. Its zero response property on the negative half axis suppresses inactive neurons, thereby inducing sparsity in the learned representations. This sparsity improves the discriminability of features and improves the generalization performance. Moreover, the absence of saturation in the positive regime enables ReLU to effectively mitigate the vanishing gradient problem, thereby accelerating training and improving convergence behavior.

5. Upsampling

A CNN is typically composed of successive convolutional blocks, each consisting of a convolutional layer, a normalization layer(e.g., BN), and a nonlinear activation function. These components are often treated as a unified convolutional unit. In standard architectures, pooling layers are periodically inserted after convolutional blocks to progressively expand the effective receptive field while simultaneously reducing the spatial resolution of the feature maps. The alternation between convolutional feature extraction stages and downsampling operations enables the network to learn hierarchical representations, evolving from low level features(such as edges and textures) to high level semantic abstractions. As a result, the spatial resolution of feature maps decreases substantially with increasing network depth. VGG-16 is a representative backbone architecture for feature extraction, organized into multiple convolutional stages. Within each stage, the spatial resolution is preserved, while transitions between stages are achieved via max pooling layers with a 2×2 kernel and a stride of 2, which

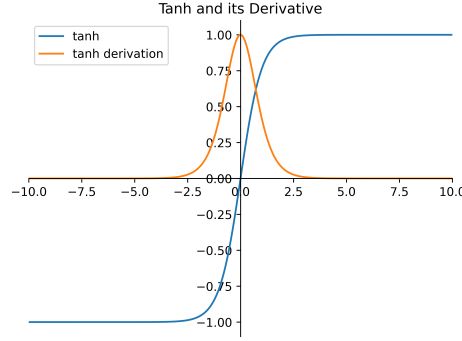


Figure 15: Tanh activation function curve and its gradient curve.

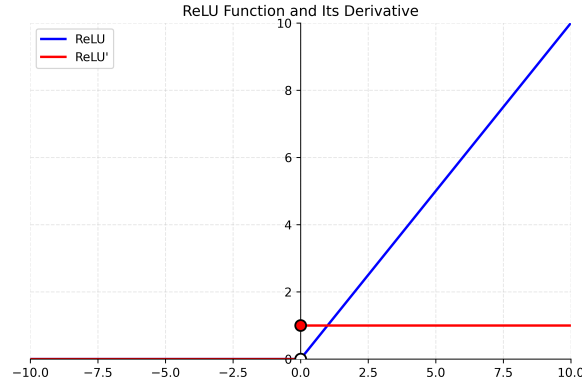


Figure 16: ReLU activation function and its derivative (curves overlap in the negative region).

reduce the spatial dimensions by a factor of two at each downsampling step. Consequently, the final feature maps constitute a highly compressed spatial representation of the input image. However, for dense prediction tasks such as semantic segmentation, operating directly on low resolution feature maps limits the ability to recover fine grained spatial details. Therefore, it is common to employ upsampling strategies, such as transposed convolutions, interpolation based upsampling, or decoder based architectures to restore spatial resolution and improve pixel level prediction accuracy.

The most commonly used feature map upsampling techniques in CNNs include nearest neighbor interpolation, bilinear interpolation, pooling index based upsampling, and transposed convolution. Nearest neighbor interpolation is the most straightforward method. It increases spatial resolution by directly replicating the value of the nearest pixel at the target location, resulting in a computationally efficient but relatively coarse approximation. In contrast, bilinear interpolation accounts for the relative spatial position of a target pixel within the original feature map. It leverages the four spatially adjacent pixels and computes a distance weighted average, yielding smoother and more continuous interpolation results. Pooling index based upsampling is typically paired with max pooling operations. During the downsampling phase, these stored indices are used to place the corresponding activation values back into their original spatial locations, while the remaining positions are filled with zeros. This mechanism is commonly employed in architectures such as SegNet [BKC17]. The aforementioned methods are parameter free, as they do not introduce additional learnable weights, and therefore incur low computational overhead. However, their representational capacity is limited. By contrast,

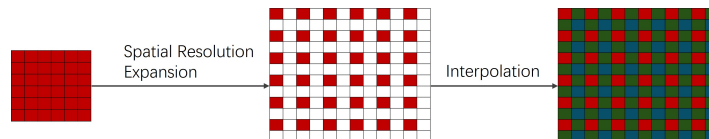


Figure 17: Schematic diagram of the upsampling process.

transposed convolution is a learnable upsampling operator that increase spatial resolution by reversing the forward convolution process with a specified kernel size and stride. Initially introduced in the context of CNN visualization, transposed convolution has since been widely adopted in dense prediction tasks.

For example, in the FCN proposed by [LSD15], transposed convolutions are used to unsample feature maps, enabling the recovery of spatial resolution and improving pixel wise prediction accuracy. Despite its effectiveness, transposed convolution introduces additional learnable parameters, thereby increasing model complexity. Moreover, its performance depends on hyperparameters such as kernel size and stride, which are typically determined empirically, making architectural design and optimization more challenging.

2.3 Research on Deep Learning-Based Semantic Segmentation Methods for Remote Sensing Images

With the rapid advancement of deep learning in computer vision, DL-based semantic segmentation of remote sensing images has emerged as a prominent research topic and attracted widespread attention. Compared with traditional segmentation approaches that rely on manually designed features, CNNs are capable of automatically learning feature representations from data and extracting high level semantic information due to their strong nonlinear representation capacity. This enables a more precise characterization of semantic categories in images and significantly improves segmentation performance in remote sensing applications.

In recent years, a variety of advanced network architectures have been proposed along with the evolution of deep learning models. In addition to conventional CNN-based approaches, Transformer-based methods, which emphasize global context modeling, and Mamba-based methods, which leverage state space modeling frameworks, have been introduced into semantic segmentation tasks, enhancing the capability to model long range dependencies. Meanwhile, the increasing diversity of remote sensing data acquisition modalities enables multi-source data to be collected from the same geographic region using different sensors, such as optical imagery, DEM, and SAR data, thereby providing richer and complementary information for segmentation tasks.

This section focuses on recent advances in CNN-based semantic segmentation for remote sensing images. First, the development of classical CNN-based segmentation models is reviewed. Building upon this foundation, existing improvement strategies are further categorized according to their technical characteristics, including multi scale feature fusion strategies, attention based mechanisms, and edge aware modeling approaches.

1. Classic CNN-based Semantic Segmentation Methods

In 2015, Jonathan Long et al. proposed FCNs and applied them to semantic segmentation, marking a milestone in the application of CNNs to dense prediction tasks [LSD15]. By replacing the fully connected layers in conventional image classification networks with convolutional layers, FCNs enable the processing of input images with arbitrary spatial dimensions and support end-to-end pixel wise prediction. This paradigm establishes a fundamental framework for subsequent research on semantic segmentation of remote sensing imagery using deep convolutional architectures.

Fig.18 presents the architecture of the FCN model. The network hierarchically extracts high level semantic representations through successive convolutional and pooling operations, resulting in a progressive reduction of the feature map resolution to $1/32$ of the input. Subsequently, the low resolution feature maps are gradually upsampled via learnable transposed convolutions (also known as deconvolutions) and interpolation techniques (e.g., bilinear interpolation) to recover spatial resolution. During this decoding process, skip connections are incorporated to fuse pixel level information from intermediate feature maps at $1/16$ and $1/8$ scales with the upsampled features, thereby facilitating effective multi scale feature aggregation and enhancing localization accuracy.

Despite its notable success, FCN still exhibits several limitations. On the one hand, repeated down-sampling operations lead to the loss of fine grained spatial details in high level feature representations, resulting in coarse object boundaries and suboptimal delineation of small or thin structures. On the other hand, the relatively simplistic upsampling strategy constrains the model’s ability to capture rich local context and long range dependencies, thereby limiting overall segmentation performance.

To address these limitations, Olaf Ronneberger et al [RFB15], proposed the U-Net architecture Fig.19. The model adopts a symmetric encoder-decoder framework, in which the encoder hierarchically

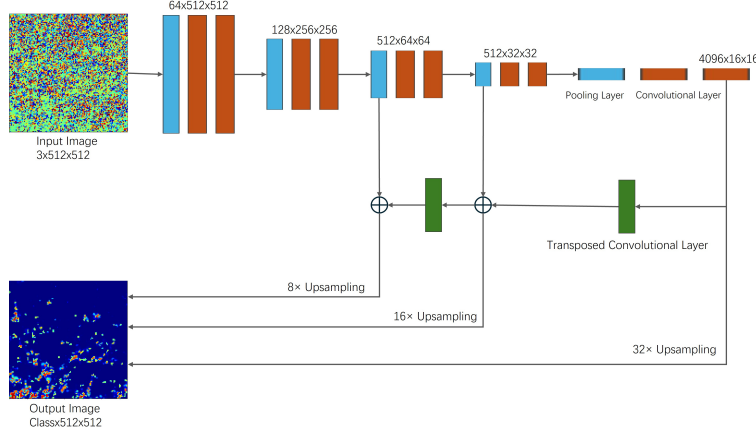


Figure 18: Architecture of the FCN model.

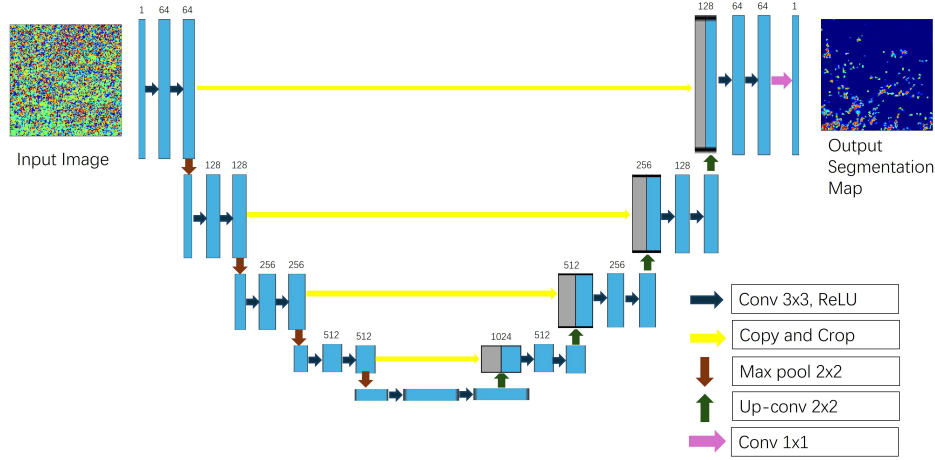


Figure 19: Architecture of the U-Net model.

learns high level semantic representations through cascaded convolutional and pooling operations, while the decoder progressively reconstructs spatial resolution via multi stage upsampling. Distinct from earlier approaches, U-Net incorporates skip connections between feature maps at corresponding resolutions, enabling the fusion of low level spatial details from the encoder with high level semantic information from the decoder. This design facilitates effective multi scale context aggregation and substantially improves boundary delineation accuracy as well as fine grained segmentation quality. Owing to its structural efficiency and strong extensibility, the encoder-decoder paradigm has since become a dominant architectural framework in semantic segmentation.

Concurrently, Vijay Badrinarayanan et al [BKC17], introduced the SegNet model to mitigate limitations of FCN, particularly the insufficient modeling of inter pixel contextual dependencies and global object coherence during downsampling. SegNet also follows a symmetric encoder-decoder design, where the encoder leverages the first thirteen convolutional layers of a Pre-trained VGG-16 network for feature extraction. In the decoding stage, feature maps are upsampled using the max-pooling indices recorded during the encoding process, followed by convolutional layers to refine spatial representations. This strategy reduces parameter redundancy and computational cost while preserving structural consistency of segmented objects. Nevertheless, due to its reliance on discrete pooling indices for upsampling, the model exhibits limited capability in recovering fine-grained semantic information, particularly for small scale targets, where representation fidelity remains suboptimal.

Early semantic segmentation approaches predominantly relied on hierarchical feature extraction via cascaded convolutional and max pooling operations. However, repeated downsampling inevitably leads to the loss of fine grained spatial details, which degrades boundary delineation accuracy and limits the fidelity of segmentation outputs. In addition, the inherently limited receptive field of standard

convolutions restricts the model’s capacity to capture long range contextual dependencies. To mitigate these limitations, Chen et al. proposed the DeepLabv1 framework [CPK⁺14]. This model employs a pre-trained VGG-16 backbone for feature extraction and integrates a Conditional Random Field (CRF) for post-processing refinement. More importantly, it introduces atrous (dilated) convolution to replace standard convolution, enabling an enlarged receptive field without increasing parameter complexity, thereby improving contextual representation learning. Building upon this design, DeepLabv2 incorporates the Atrous Spatial Pyramid Pooling (ASPP) module, which utilizes parallel atrous convolutions with multiple dilation rates to capture multi scale contextual information [CPK⁺17]. This mechanism enhances robustness to significant object scale variation. Subsequently, DeepLabv3 eliminates the Conditional Random Field (CRF)-based post-processing stage and adopts a more powerful backbone network, such as ResNet, enabling a fully end-to-end semantic segmentation pipeline [CPSA17, HZRS16]. Meanwhile, the Atrous Spatial Pyramid Pooling (ASPP) module is further refined with the integration of Batch Normalization, improving optimization stability and feature representation capacity. Further extending this framework, DeepLabv3+ introduces an encoder–decoder architecture to enhance spatial detail recovery [CZP⁺18]. Specifically, it combines high-level semantic features extracted by ASPP with low level features from earlier layers, leading to improved boundary refinement and more precise fine grained segmentation. These developments collectively demonstrate the effectiveness of atrous convolution in multi-scale semantic representation learning. In contrast to the DeepLab family, Zhao et al. proposed PSPNet, which models global contextual information through a Pyramid Pooling Module (PPM) [ZSQ⁺17]. By aggregating features at multiple spatial scales via pooled representations, the model improves its ability to handle objects of varying sizes. Furthermore, Xiao et al. introduced UPerNet, which integrates PPM with a Feature Pyramid Network (FPN) [XLZ⁺18, LDG⁺17]. This hybrid design effectively combines global context with hierarchical feature representations, enabling more comprehensive multi scale feature modeling and achieving strong performance in natural scene semantic segmentation tasks.

2. Multiscale fusion based optimization method

high-resolution remote sensing imagery is characterized by a wide spatial coverage and rich structural details, but it also introduces pronounced scale variation challenges. On the one hand, numerous land cover objects—such as individual trees, vehicles, and pedestrians—occupy only a limited number of pixels, exhibiting extremely small spatial extents. On the other hand, instances within the same semantic category (e.g., buildings or background regions) may present substantial scale diversity across different scenes. This heterogeneous scale distribution poses significant challenges for conventional semantic segmentation methods based on convolutional neural networks. Incorporating multiscale representations can effectively enlarge the receptive field and enhance the model’s capability to capture small and densely distributed targets. Meanwhile, cross scale feature fusion facilitates complementary information exchange across different feature hierarchies, thereby alleviating semantic ambiguity. Consequently, designing an effective multiscale feature fusion mechanism is critical for semantic segmentation of high-resolution remote sensing imagery.

In recent years, multiscale feature fusion strategies have been extensively explored across various computer vision tasks. For example, Luo et al. integrated a region proposal mechanism with ResNet-50 for general semantic segmentation. Specifically, candidate regions at multiple scales are first generated to extract region-level multiscale features and foreground cues. These multiscale inputs are then used to train multiple sub-models, and segmentation performance is further improved through an ensemble strategy. Yang et al. proposed the EMFFN model for hyperspectral image classification, which consists of two complementary subnetworks: a cascaded atrous convolutional network for spectral feature extraction and a parallel multi-branch architecture for capturing multiscale spatial representations [YWDZ21]. The fused outputs of these subnets lead to enhanced classification performance. Ding et al. introduced the MAINet model for cloud and snow segmentation, where a feature interaction module is designed to enable cross scale information exchange [DXLH23]. In this framework, shallow features are fed into a multiscale boundary aware convolution module to extract edge sensitive representations, while deeper features are refined via a spatial attention module to enhance both spatial and channel wise responses. The resulting features are subsequently fused at the pixel level to produce more discriminative representations, effectively mitigating boundary ambiguity and coarse segmentation artifacts. In addition, Sun et al. proposed the MFINEA model for image dehazing, which employs a multilevel feature interaction mechanism to facilitate feature fusion between encoder and decoder stages, thereby improving edge detail reconstruction [SLD⁺23]. Furthermore, a non-local

channel attention module is incorporated to strengthen global contextual modeling capability.

In the domain of remote sensing image semantic segmentation, numerous studies have extended the U-Net architecture to enhance its multiscale representation capability, primarily through improved feature fusion strategies and attention mechanisms. For example, Li et al. proposed MACU-Net, which enhances the integration of multiscale representations by introducing multiscale skip connections, enabling effective fusion of high level semantic information and low level spatial details[L^{DZ}20]. In addition, an asymmetric convolutional module—comprising parallel 1×3 , 3×1 , and 3×3 convolutions, is incorporated to capture anisotropic contextual features, thereby improving the discriminative capacity of feature representations. Building upon this, Li et al. further developed MAResUNet, which restructures the skip connection pathways to strengthen cross scale feature aggregation and introduces a linear attention mechanism to refine encoder feature maps, leading to improved representation quality[L^{ZD}+21].

Despite these advancements, most existing approaches still rely on direct concatenation to fuse shallow, fine grained features with deep semantic representations, lacking an explicit guidance mechanism. Such a naive fusion strategy is insufficient to fully exploit the complementary characteristics of multilevel features and may introduce semantic inconsistency, ultimately constraining further performance gains. To address this limitation, Bai et al. proposed the High-Capacity Attention Network (HCANet), which replaces conventional cross layer concatenation with two compact spatial pyramid pooling modules, thereby embedding multiscale contextual cues into the fusion process[B^{CH}+21]. Zhang et al. proposed ESDINet, which adopts a dual-branch design consisting of a detail branch and a semantic branch, and introduces a differentiated cross branch interaction mechanism to improve the utilization of hierarchical spatial information[Z^{WZ}+24]. Xu et al. developed FFPNet, which performs multiscale feature fusion guided by attention by integrating a region aware pyramidal attention module and an adaptive ASPP module. Moreover, a flexible feature fusion pyramid is constructed to better capture complex spatial distributions[X^{YOZ}20]. Zheng et al. proposed the , which leverages a feature pyramid network to extract multiscale representations and employs global average pooling on high level features to generate scene embedding vectors[Z^{ZWM}20]. These embeddings are used to model the foreground with relational dependencies of the background, and feature maps on different scales are adaptively weighted, significantly enhancing discriminative capacity and achieving strong performance on the iSAID dataset[W^{ZAG}+19]. Although these methods have advanced multiscale semantic representation learning, their reliance on convolutional operations inherently limits them to local receptive fields, restricting the modeling of long range dependencies. As a result, their ability to capture global contextual information and represent high level semantic relationships in complex scenes remains constrained.

3. Attention mechanism based optimization method

Accurate modeling of both local details and global contextual information is essential for reliable semantic understanding in image segmentation tasks. CNNs are effective at capturing local patterns and exhibit translation invariance; however, their inherently limited receptive fields constrain their ability to model long range dependencies and global context. Attention mechanisms provide an effective means to enhance feature representation by explicitly capturing inter feature dependencies. Jie Hu et al. introduced the Squeeze-and-Excitation (SE) module, a channel attention mechanism designed to model channel wise interdependencies and improve representational capacity [H^{SS}18]. As shown in Fig.20, specifically, given an input feature map U , global average pooling is first applied to aggregate spatial information into channel wise descriptors. These descriptors are then passed through an excitation function typically implemented using fully connected layers followed by a Sigmoid activation to generate channel wise attention weights. The original feature map is subsequently recalibrated via channel wise reweighting, allowing semantically informative channels to be emphasized while suppressing less relevant responses, thereby enhancing discriminative feature learning.

Nevertheless, channel attention mechanisms primarily focus on channel-wise importance and overlook spatial variability. In practical scenarios, informative regions are often sparsely distributed, while standard convolution treats all spatial locations uniformly, limiting its ability to highlight salient regions. To overcome this limitation, spatial attention mechanisms have been proposed to model spatial dependencies and assign adaptive importance to different spatial locations. As shown in Fig.21, concretely, for an input feature map X , global max pooling and average pooling are first applied along the channel dimension to generate two spatial context descriptors. These descriptors are concatenated to integrate complementary information, and a convolutional layer (commonly 1×1 or 7×7) followed

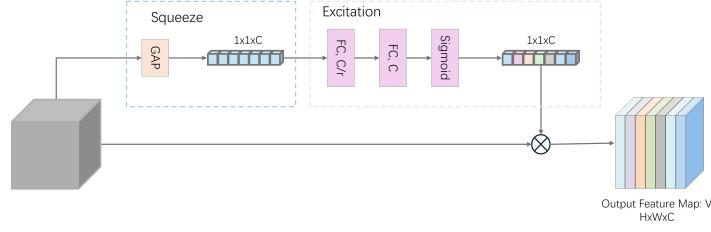


Figure 20: Architecture of the Channel Attention Module.

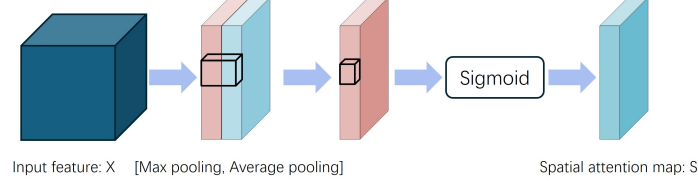


Figure 21: Architecture of the Spatial Attention Module.

by a Sigmoid activation is used to produce a single channel spatial attention map S . This attention map is then applied to the input via element wise multiplication, enabling adaptive spatial reweighting. As a result, responses in salient regions are amplified, while irrelevant background information is suppressed, leading to improved feature discriminability [ZCZ⁺19].

Existing attention mechanisms typically focus on modeling dependencies along a single dimension, which limits their representational capacity. Specifically, channel attention derives channel wise weights through global spatial aggregation. While this operation effectively captures inter channel dependencies, it tends to overlook local spatial correlations among pixels. In contrast, spatial attention emphasizes the importance of different spatial locations by aggregating information across the channel dimension; however, it remains insufficient for modeling long-range dependencies. Consequently, a single attention paradigm is inadequate for jointly capturing both global context and local structural details.

To address these limitations, Sanghyun Woo et al. proposed the CBAM [WPLK18], which sequentially integrates channel and spatial attention. Specifically, channel attention is first applied to perform channel wise feature recalibration by exploiting global contextual cues, followed by spatial attention to refine salient regions. This sequential design enables joint modeling across channel and spatial dimensions, thereby enhancing both global semantic representation and local detail preservation. In addition, Jongchan Park et al. introduced the Bottleneck Attention Module (BAM) [PWLK18], which adopts a parallel fusion strategy to combine channel and spatial attention. This design improves representational capacity while maintaining computational efficiency. Building upon these developments, more sophisticated hybrid attention mechanisms have been explored. For instance, the Dual Attention Network (DANet) incorporates both position attention and channel attention modules in parallel [FLT⁺19]. This architecture not only models inter channel dependencies but also establishes pairwise relationships among pixels across all spatial locations, thereby enabling effective global contextual modeling. Nevertheless, DANet requires the computation of high dimensional affinity matrices, resulting in substantial computational and memory overhead. Moreover, due to its limited capability

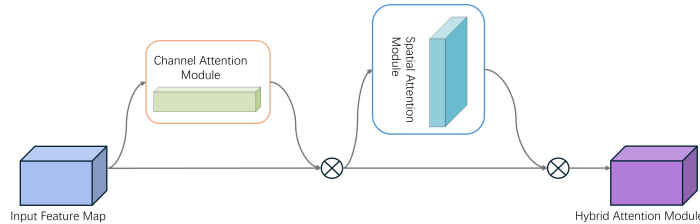


Figure 22: Architecture of the CBAM Module.

in capturing fine grained texture details, its performance on small object segmentation tasks remains suboptimal.

Thanks to the effectiveness of attention mechanisms in enhancing feature representations, a large body of attention-based approaches for remote sensing image semantic segmentation has emerged in recent years. Ding et al. identified substantial discrepancies between low level and high level features in terms of physical properties and spatial distributions, indicating that naive fusion strategies are insufficient to bridge the semantic gap [DTB20]. To address this issue, they proposed LANet, which applies localized attention to low level and high level features separately to mitigate information interference during contextual aggregation. In addition, an attention embedding module is introduced to inject high level semantic cues into low level representations, thereby improving cross level feature integration. Peng et al. proposed the Cross-Fusion Network (CFNet), which generates multiscale feature maps in the encoder and employs channel attention to perform adaptive feature selection and enhancement across scales [PZMM21]. Cross scale feature interaction is then conducted to facilitate feature fusion, effectively enlarging the receptive field of shallow features and improving the detection of small scale objects. Zhu et al. introduced the Multi-Path Attention Network (MAP-Net) for building segmentation in remote sensing imagery [ZLH⁺20]. This model adopts a multi-branch parallel architecture to extract multiscale representations at a fixed resolution, and leverages attention mechanisms for adaptive channel wise fusion, thereby improving small object recognition and boundary delineation of large structures. To alleviate the high computational cost associated with attention mechanisms, Li et al. proposed a lightweight architecture, SCAttNet, which sequentially integrates spatial attention and channel attention to model feature dependencies [LQC⁺20]. This design reduces computational overhead while enhancing segmentation performance, particularly for small objects. Fan et al. introduced MANet, which combines position-aware attention with multiscale self-attention to strengthen contextual representation learning [LZZ⁺21]. Sun et al. proposed SPANet, a dual-branch framework that extracts local and global features in parallel and incorporates a cascaded pooling attention module to enable multiscale feature aggregation, effectively alleviating boundary ambiguity. Furthermore, Li et al. [SCZ⁺22]. [LZD⁺21] developed a multi stage lightweight feature fusion network, where spatial and channel attention modules are integrated in parallel to construct an efficient hybrid attention mechanism. Hu et al. proposed a fast attention mechanism by reformulating the self-attention computation, significantly reducing computational complexity while preserving the ability to capture global contextual dependencies [HPH⁺20]. In addition, Zhao et al. proposed SSAtNet, which is built upon the ResNet-101 backbone and incorporates a pyramid attention pooling module to perform adaptive refinement of multiscale feature representations [ZLLZ21].

2.4 multi-source data fusion based semantic segmentation method

Improvements in remote sensing technology mean that multi-source imagery can now provide observations of the same geographic area from different angles, offering complementary information on surface features. The different types of data have distinct advantages when it comes to representing features: spectral imagery contains rich spectral and textural information that helps characterize the visual features of surface elements; Digital Elevation Models (DEMs) provide elevation data for surface targets, reflecting the three dimensional structure of a scene; and SAR uses microwave imaging mechanisms to capture surface scattering characteristics, offering excellent characterization capabilities for complex terrain and environments. Since single data sources are limited in their ability to represent features, making it difficult to comprehensively describe complex scenes, fusing multi-source remote sensing data provides richer feature support for semantic segmentation tasks. Integrating data features from different sources enhances the completeness and discriminative power of feature representation, thereby improving the model’s ability to capture the distribution of complex objects.

Early studies on multi-source remote sensing data processing predominantly adopted straightforward fusion schemes, such as element wise addition or channel wise concatenation of feature maps extracted by separate encoders. However, these approaches lack explicit cross modal interaction modeling, which limits their capacity to fully leverage complementary information across heterogeneous data sources. To address this issue, Xu et al. proposed a dual-branch convolutional neural network that jointly learns feature representations from hyperspectral imagery and LiDAR data for classification tasks [XLR⁺17]. Meanwhile, Caner et al. systematically explored fusion strategies at different stages of the network pipeline, namely early stage fusion and late stage fusion, for to integrate LiDAR and multispectral imagery [ALSL18]. To further improve the multimodal representation feature

, subsequent research introduced attention mechanisms and architectural enhancements. Reference [LLS⁺20] employed a channel attention mechanism to adaptively recalibrate features derived from optical imagery and SAR, thereby enhancing the discriminative power of the fused representations in land cover classification. Jin et al. incorporated data from the Digital Surface Model (DSM) into a fully convolutional framework using densely connected architectures, facilitating more effective propagation of contextual information [QF14]. Kang et al. proposed CFNet, which integrates a multiscale feature extraction module to better capture inter-modal discrepancies, resulting in improved classification performance [KXWY22]. Building upon these advances, Li et al. introduced MCANet, which leverages a mechanism of cross-attention to enable a deep cross modal interaction between optical and SAR feature representations [LZC⁺22]. CroFuseNet also proposes a dedicated cross modal fusion module to enhance the learning of synergistic representation and to take advantage of complementary information more effectively [WGS23]. Yang et al. proposed AFNet, which adopts a multi path encoder to simultaneously extract features from optical imagery and DSM data, and employs a multi-branch attention module for aggregation of cross-modal features [YLC⁺21]. To address the limitation of insufficient hierarchical feature modeling, Zhao et al. proposed MFPNet, which utilizes optical and DSM data as input, incorporates a PPM to extract multiscale representations, and performs hierarchical fusion across multiple levels and scales, thereby improving the model’s ability to capture objects with diverse spatial scales.

Overall, the integration of complementary information from heterogeneous data sources can effectively enhance the performance of semantic segmentation models for multi-source remote sensing imagery. Nevertheless, existing approaches still exhibit notable limitations in this domain. These challenges primarily arise from substantial distribution discrepancies across modalities, inherent noise contamination within individual data sources, and the intrinsic complexity of remote sensing scenes, including heterogeneous backgrounds and pronounced multiscale variations.

2.5 Chapter Summary

This chapter presents a systematic review of recent advances in DL-based semantic segmentation for remote sensing imagery. First, CNN-based approaches are examined from three perspectives, including classical architectures, multiscale feature fusion strategies, and attention enhanced models. Subsequently, methods for semantic segmentation of multi-source remote sensing data are comprehensively reviewed. These analyses provide a structured understanding of the current research landscape and highlight the key challenges in this domain.

3 Proposed DL-based PS Detection

3.1 The Limitation of Conventional PS Detection

Based on the analysis of SAR and DInSAR imaging mechanisms presented in Chapter 2, MT-DInSAR has emerged as a key technique for monitoring ground deformation in recent years. By jointly processing multi-temporal SAR observations, MT-DInSAR effectively mitigates error sources such as spatio-temporal decorrelation, atmospheric delay, and system noise, thereby enabling deformation retrieval with millimeter level precision. As a result, it has been widely applied in landslide monitoring, ground subsidence assessment, infrastructure stability evaluation, and urban deformation analysis[CMCG⁺16, WB14].

From a methodological perspective, existing MT-DInSAR processing frameworks can generally be divided into two categories: PSI and the SBAS method[FPP⁺24]. The primary differences between these approaches lie in their modeling scale, interferogram network construction, and underlying assumptions regarding scattering mechanisms. PSI-based methods, exemplified by PS-InSAR [FPR01b, FPR01a], are typically implemented at full or high spatial resolution. Their fundamental principle is to identify pixels exhibiting long-term phase stability-referred to as PS from SAR time-series data. In practice, Persistent Scatterers Candidates (PSC) are first selected using criteria such as the amplitude dispersion index or phase stability indicators. Subsequently, a phase observation model is established over the PSC set, and parameter estimation is performed to decompose the observed phase into multiple components, including APD, orbital errors, and deformation signals. Owing to their reliance on highly environments or areas dominated by man-made structures with stable scattering properties.

In contrast, the SBAS approach [BFLS03, Usa03] constructs a redundant interferogram network based on short temporal and spatial baselines, thereby reducing phase decorrelation effects associated with large baselines. This framework typically operates at a lower spatial resolution and is better suited for natural terrains characterized by DS. Ground deformation time series are retrieved by inverting the redundant interferometric network using techniques such as least squares estimation or singular value decomposition. Compared with PSI, SBAS exhibits stronger capability in modeling spatially continuous deformation; however, it generally provides lower spatial resolution and has limited ability to capture high precision point wise deformation signals.

Following the estimation and compensation of the APS, the core objective of MT-DInSAR shifts to the identification of high-coherence scatterers and the inversion of their associated geophysical parameters. This process involves selecting pixels exhibiting temporally stable phase behavior from a large set of candidates and modeling their phase evolution over time. In practice, simplified parametric phase models are commonly adopted to characterize the phase history, typically involving a limited number of key parameters, such as the DMV and RT. Consequently, it is essential to distinguish PSs from pixels severely affected by temporal decorrelation within a unified statistical framework. Under high-resolution imaging conditions, PSs are generally associated with man made structures possessing stable geometric and electromagnetic scattering properties, such as buildings and bridges.

In conventional PSI approaches, pixel stability is typically quantified using Multi-Temporal Coherence (MTC) [FPR01b, FPR01a], which measures the consistency between observed and modeled

phases by analyzing phase residuals across a temporal sequence. PS selection is commonly performed via empirical thresholding, whereby a pixel is classified as a persistent scatterer only if it exhibits low phase residuals and high temporal stability. However, this approach is inherently sensitive to threshold selection and relies exclusively on phase information, which may result in false detections or missed targets, particularly in low SNR conditions or structurally complex scenes.

To enhance the robustness of PS detection and parameter estimation, recent studies have reformulated the problem within a statistical signal detection framework. In particular, the GLRT, originally developed in SAR tomography, has been introduced to jointly exploit both amplitude and phase information for a more comprehensive characterization of scattering stability[DFP09]. Compared with traditional coherence based methods, GLRT provides improved detection performance in theory, especially for weak or low-coherence scatterers.

In addition, multi-look processing strategies are frequently adopted to improve the detectability of weakly coherent target[FFN+11, VPRF20, DPRF23]. By performing spatial averaging, these methods effectively increase the SNR, thereby enhancing detection performance. However, this improvement comes at the cost of reduced spatial resolution. To mitigate this limitation, non-local filtering approaches based on Statistical Homogeneous Pixels (SHP) selection[SRP+15], as well as adaptive multi-look strategies, have been proposed. These methods exploit the redundancy among pixels with similar statistical characteristics to suppress noise while preserving fine spatial details, achieving a more favorable trade-off between detection accuracy and spatial resolution.

This thesis proposes a two stages MT-DInSAR processing framework derived from the aforementioned theoretical analysis. In the first stage, the SBAS method is employed to perform preliminary deformation retrieval and atmospheric phase compensation at a reduced spatial resolution, thereby obtaining stable initial estimates. In the second stage, after data calibration, the GLRT is applied at full spatial resolution to achieve refined scatterer detection and geophysical parameter estimation. By integrating the robustness of SBAS with the superior detection capability of GLRT, the proposed framework achieves an effective balance between estimation accuracy and computational efficiency.

Despite the significant advancements of MT-DInSAR in surface deformation monitoring, existing approaches still exhibit notable limitations in PS detection and parameter inversion. First, conventional methods(e.g., PSI and SBAS) rely heavily on handcrafted statistical indicators, such as the amplitude dispersion index and multi-temporal coherence and empirical thresholding schemes. This reliance limits their adaptability to complex scenarios, particularly in low SNR environments or regions with severe temporal decorrelation, where false detections and missed detections are more likely to occur. Second, model driven approaches generally assume specific statistical distributions for phase noise and employ simplified parametric models to describe phase evolution. Such assumptions restrict their ability to accurately characterize then complex, nonlinear spatiotemporal dynamics present in real word scenes. Furthermore, although statistical detection methods such as GLRT offer improved theoretical detection performance, their computational burden remains substantial, posing challenged for large-scale, high-resolution data processing. Moreover, traditional MT-DInSAR frameworks predominantly rely on local features or information within limited neighborhoods, which constrains their capability to capture long range spatiotemporal dependencies. This limitation becomes particularly pronounced in heterogeneous environments or complex terrains. Consequently, enhancing the representational capacity of models to effectively capture rich spatiotemporal characteristics in multi-temporal SAR data has become a critical research focus.

3.2 DL Approaches for SAR Data Processing

Advances in computer vision have driven the widespread adoption of DL techniques in remote sensing applications. In the field of SAR, early studies have explored the potential for DL across a range of tasks, including terrain classification, target detection, parameter inversion, speckle noise reduction, InSAR processing, and SAR-Optical data fusion[LFW+18, BDT+23, BJSV19, ZMA+21, HK20].

Effective feature representation is fundamental to terrain classification using PolSAR imagery. Conventional approaches typically rely on handcrafted features derived from domain specific expertise; however, such features often exhibit limited generalization and poor adaptability across datasets. In contrast, DL-based methods enable automatic feature learning and have demonstrated superior robustness and generalization capability. For SAR target detection, traditional methods are largely based on template matching, where manually designed templates are used for category discrimination as well as classical machine learning techniques such as support vector machines (SVMs). These

approaches are inherently constrained by prior assumptions and limited representational capacity. By comparison, DL-based methods leverage deep CNNs to learn discriminative feature representations in a data driven manner, leading to improved detection performance. In parameter inversion tasks, DL models are employed to learn nonlinear mappings from SAR observations to geophysical parameters, such as sea ice concentration. Compared with conventional model driven approaches, these methods offer greater flexibility in capturing complex data relationships. For SAR despeckling, traditional techniques typically rely on handcrafted filtering operations, which may lead to the loss of structural details, particularly along edges. In contrast, DL-based approaches provide data driven solutions that achieve more effective noise suppression while preserving fine-scale structures. Moreover, the joint analysis of SAR and optical images has attracted increasing attention due to their complementary characteristics, enabling more comprehensive feature representation [PPP20, ISWY96, ZP02, CW14, WSXC16].

Specifically, DL-based solutions provided interesting advancements in different SAR multi channel processing, including interferometric and tomographic SAR. With reference to the PS identification problem, DL-based solutions have been proposed in the recent literature [AKH⁺25a, TND20, ZWD⁺22].

They are all aimed at building the starting grid of PS candidates over which MT-DInSAR processing chains carries out the separation between APD and displacement components[FPR01b, FFN⁺11]. Other DL-based solution have been proposed for selection of statistical homogeneous pixels [HWG⁺22] and for the identification of specific trends in the estimated time series [WEK22]. Several DL-based solutions have been in particular proposed also for super-resolution SAR tomography [FPP⁺24]. These include γ -Net and Hyper-LISTA [QWSZ22, QWJ⁺24]. In [QWSZ22] the authors focused on the exploitation of Recurrent Neural Networks (RNNs) in Compressive Sensings (CS) SAR Tomography. γ -Net adopts a Complex-Valued Learned Iterative Shrinkage Thresholding Algorithm (CV-LISTA) to effectively simulate the optimization steps in sparse reconstruction, thus enhancing computational efficiency. γ -Net was shown to outperform existing, higher order scatterers detection CS methods, on both simulated and real data. A different DL-based solution to CS SAR tomography has been provided in [QWJ⁺22]. Here an innovative super-resolution scheme is proposed to solve the Basis Pursuit Denoising (BPD) problem via Recurrent Neural Networks (RNN). Beyond approaches based on RNNs, recent studies have increasingly explored CNNs and hybrid CNN with RNN architectures for PS analysis. Ashutosh Tiwari et al. proposed two DL frameworks, CNN-ISS and CLSTM-ISS, which adopt PS candidates derived from the StaMPS method as supervisory signals. By explicitly modeling the spatiotemporal characteristics of the interferometric phase, these frameworks enable accurate identification of PS pixels [TND20]. S.Azadnejad et al. introduced a DL-based framework to temporal consistency estimation, aimed at coherent pixel selection in TS-InSAR analysis. Specifically, features extracted from SAR amplitude time series are fed into a to predict pixel wise temporal consistency, facilitating robust identification of stable scatterers[AKH⁺25b]. Zihan Hu et al. developed the Multi-Temporal Feature Network (MFN), which integrates a 3D U-Net with a Convolutional Long Short-Term Memory (ConvLSTM) module to jointly capture spatiotemporal dependencies in SAR data. This architecture effectively exploits both temporal dynamics and spatial context under complex terrain conditions, leading to improved accuracy in stable points detection and filtering [HLL⁺25].

Inspired by the positive impacts of DL in SAR processing[ZMA⁺21], this work concentrates on a “light” use of DL, limited to the problem to detect PS starting from the outputs (Parameters of Interest (PoI) and statistic) provided by the GLRT, in order to overcome the limitations of the classical detection thresholding. The focus does not pertain to the separation of multiple components corresponding to double or higher order scatterers as in CS SAR tomography, but rather the identification of pixels containing single scatterers. The objective is to develop a PS detection algorithm that can relax the stringent constraint on the PFA by identifying pixels sharing common estimation on image features. A solution based on U-Net [RFB15] is specifically considered thanks to its capability of handling multiple input information [CHZ⁺23, HNAB21, ZSHH22].

3.3 DL-based PS detection

3.4 GLRT method

Consider a dataset consisting of N SAR images that have been precisely co-registered and phase calibrated. These observations exhibit diversity in both spatial and temporal domains. The corresponding

spatial and temporal baseline vectors are defined as:

$$\mathbf{b} = [b_1, \dots, b_N]^T, \quad \mathbf{t} = [t_1, \dots, t_N]^T \quad (3.1)$$

where the superscript T denotes the transpose operation. The spatial baseline vector $\mathbf{b}$ characterizes the geometric differences in sensor positions across acquisitions, while the temporal baseline vector $\mathbf{t}$ represents the temporal sampling intervals. Together, these baselines govern the geometric and temporal evolution of the interferometric phase, providing the fundamental basis for deformation retrieval and parameter estimation.

The PoI is defined as:

$$\boldsymbol{\xi} = [z, v]^T \quad (3.2)$$

where z denotes the RT, accounting for phase contributions induced by DEM errors, and v represents the DMV, describing the linear deformation trend over the observation period. These parameters constitute the principal unknowns in the InSAR phase model.

Assuming the presence of at most one PS within a resolution cell, the complex observation vector obtained from N acquisitions equivalently interpreted as measurements from N virtual array elements is given by:

$$\mathbf{x} \in \mathbb{C}^N \quad (3.3)$$

where $\mathbf{x}$ contains the amplitude and phase information of the multi-temporal SAR observations. Under this assumption, the PS detection problem can be formulated as a binary hypothesis testing problem:

$$\mathcal{H}_0 : \mathbf{x} = \mathbf{n} \quad (3.4)$$

$$\mathcal{H}_1 : \mathbf{x} = \gamma \mathbf{a}(\boldsymbol{\xi}) + \mathbf{n} \quad (3.5)$$

where γ denotes the unknown complex backscattering coefficient of the scatterer, which is typically treated as a nuisance parameter in the detection problem, and $\mathbf{n}$ represents the additive noise vector, commonly assumed to follow a circular complex Gaussian distribution. The steering vector $\gamma(\mathbf{a})$, which characterizes the deterministic phase response of a PS across multiple observations, is defined as (3.5):

$$\mathbf{a}(\boldsymbol{\xi}) = \frac{1}{\sqrt{N}} \exp(-j[k_z, k_v]\boldsymbol{\xi}) \quad (3.6)$$

where $\boldsymbol{\xi} = [z, v]^T$ denotes the parameter vector of interest, including RT and deformation velocity. The normalization factor $1/\sqrt{N}$ ensures that the steering vector has unit energy, which is beneficial for subsequent statistical detection and parameter estimation.

The phase coefficients k_z and k_v are defined as:

$$k_z = \frac{4\pi}{\lambda} \frac{\mathbf{b}}{r_0 \sin \vartheta} \quad (3.7)$$

$$k_v = \frac{4\pi}{\lambda} \mathbf{t} \quad (3.8)$$

where $\mathbf{b}$ and $\mathbf{t}$ denote the spatial and temporal baseline vectors, respectively. The parameter r_0 represents the slant range distance between the radar sensor and the target, ϑ is the incidence angle, and λ denotes the radar wavelength.

Physically, the term k_z corresponds to the phase contribution induced by residual topography errors, while k_v capture the phase variation caused by linear deformation over time. Therefore, the steering vector $\mathbf{a}(\boldsymbol{\xi})$ encodes the joint spatiotemporal phase evolution of a coherent scatterer, linking the underlying geophysical parameters to the observed interferometric phase measurements.

This formulation establishes a parametric relationship between the observed signal and the unknown parameters, providing a foundation for subsequent parameter estimation and statistical detection method.

The GLRT detector is derived by formulating PS identification as a binary hypothesis testing problem. Specifically, under the null hypothesis, the observation signal contains noise only, whereas under the alternative hypothesis, the observation consists of a deterministic scatterer response embedded in

noise. Since the complex backscattering coefficient and the parameter vector of interest are unknown, the corresponding likelihood functions cannot be evaluated directly and must be replaced by their ML estimates. This leads to the GLRT framework, in which the likelihood ratio is evaluated after substituting the unknown parameters with their estimated values.

Under the assumed single-scatterer model, the GLRT detector can be expressed in terms of the following normalized test statistic:

$$T(\mathbf{x}) = \frac{|\mathbf{a}^H(\hat{\boldsymbol{\xi}})\mathbf{x}|^2}{\|\mathbf{x}\|^2} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} T, \quad T \in [0, 1] \quad (3.9)$$

where $(\cdot)^H$ denotes the Hermitian transpose, $\|\mathbf{x}\|^2$ is the energy of the observed data vector, and $\mathbf{a}(\hat{\boldsymbol{\xi}})$ is the steering vector evaluated at the estimated parameter vector. The quantity $|\mathbf{a}^H(\hat{\boldsymbol{\xi}})\mathbf{x}|^2$ measures the matched response between the observed signal and the hypothesized phase model, while the denominator provides an energy normalization that makes the detector insensitive to absolute signal amplitude. As a result, the statistic $T(\mathbf{x})$ can be interpreted as a normalized coherence-like measure reflecting the degree of alignment between the observation and the PS model.

The unknown parameter vector is estimated according to the ML principle as:

$$\hat{\boldsymbol{\xi}} = \arg \max_{\boldsymbol{\xi}} |\mathbf{a}^H(\boldsymbol{\xi})\mathbf{x}| \quad (3.10)$$

The above formulation illustrates that the GLRT detector operates by evaluating the maximum response of the tomographic reconstruction process, which is effectively implemented via classical beamforming, as defined in (3.10). Specifically, the beamforming operation projects the observed multi-temporal SAR data onto a parametric steering vector manifold, thereby enabling coherent integration of signal components that conform to the assumed phase model. This procedure can be interpreted as a matched filtering process in the joint spatial-temporal domain, with the objective of identifying the parameter vector that yields the strongest constructive interference across all observations.

The resulting maximum projection value reflects the degree of consistency between the observed data and the assumed single scatterer model. To ensure robustness against variations in signal amplitude and noise power, this quantity is normalized by the total energy of the observation vector, as defined in (3.9). Consequently, the resulting statistic $T(\mathbf{x})$ can be interpreted as a normalized measure of model conformity (or coherence), with values bounded within the interval $[0, 1]$.

Subsequently, the GLRT decision rule compares this normalized statistic with a predefined threshold T , which is typically determined according to a desired PFA. This thresholding procedure establishes a statistically principled criterion for discriminating between noise dominated pixels and those exhibiting stable scattering behavior. Pixels satisfying $T(\mathbf{x}) > T$ are considered to exhibit sufficient agreement with the parametric signal model and are thus classified as PS.

For pixels that pass the detection stage, the corresponding parameter vector $\hat{\boldsymbol{\xi}}$, obtained via the maximization in (3.10), provides estimates of the RT and DMV. The RT component accounts for phase contributions induced by elevation related errors, while the DMV characterizes the linear deformation trend over the observation period. These estimated parameters enable precise three dimensional geo-location of the scatterers and form the basis for reconstructing high-resolution deformation time series.

3.5 U-Net

This thesis proposes the direct integration of DL techniques into the PS detection stage, aiming to overcome the inherent limitations of traditional detection frameworks based on statistical hypothesis testing. Conventional approaches typically rely on the GLRT statistic, as defined in (3.9), to perform pixel wise decision under a strictly controlled probability of PFA. However, such methods depend heavily on idealized statistical assumptions and empirically defined threshold. In complex remote sensing scenarios-characterized by low SNR, multiple scattering mechanisms, and pronounced spatial heterogeneity, these approaches often fail to adequately capture spatial correlations and structural dependencies among pixels, thereby limiting detection performance.

To address these challenges, this thesis introduces a data driven paradigm to replace the conventional explicit statistical decision making process. Specifically, the PS detection problem is reformulated as the identification of pixel structures exhibiting spatial measurement consistency, allowing

the model to implicitly learn discriminative features from the data rather than relying on predefined statistical criteria.

Under this framework, PS detection is cast as a pixel wise semantic segmentation task, enabling precise localization of stable scatterers through the exploitation of spatial contextual information and multiscale feature representations. To this end, the U-Net architecture[RFB15] is adopted as the backbone network. Originally developed for biomedical image segmentation, U-Net is well known for its ability to achieve high-resolution, fine grained pixel level predictions. Given that SAR imagery is characterized by complex textural patterns, strong speckle noise, and significant spatial variability, encoder-decoder architectures such as U-Net have been extensively validated in SAR image segmentation and change detection tasks, making them particularly suitable for the PS detection problem considered in this study.

The U-Net architecture follows a symmetric encoder-decoder paradigm and takes a two dimensional image patch of size $W \times H$ as input. The encoder progressively maps the input into a high dimensional feature space through successive convolutional and downsampling operations, enabling a hierarchical transition from low level spatial features to high level semantic representations. Each encoding stage consists of two consecutive 3×3 convolutional layers, each followed by BN and a ReLU activation function, which together improve training stability and enhance representational capacity. As the network depth increases, the number of feature channels doubles at each stage, while the spatial resolution is reduced via 2×2 max pooling operations. Consequently, the feature map size is reduced from $W \times H$ to $W/8 \times H/8$ at the bottleneck layer, effectively enlarging the receptive field and strengthening the model’s ability to capture global contextual information.

The bottleneck layer, located at the deepest level of the network, operates at the lowest spatial resolution and is responsible for extracting highly abstract semantic features through two convolutional layers. This stage produces a compact yet informative representation of the input data and serves as a critical bridge between the encoding and decoding pathways, providing essential contextual information for subsequent feature reconstruction.

The decoder progressively restores the spatial resolution of the feature maps through a series of upsampling operations, typically implemented via transposed convolutions or equivalent interpolation based methods. At each decoding stage, the spatial dimensions are doubled while the number of feature channels is reduced accordingly, enabling gradual recovery of fine spatial details. A key component for the U-Net architecture is the use of skip connections, which concatenate feature maps from the encoder with those in the corresponding decoder stage at the same resolution. This cross scale feature fusion mechanism effectively compensates for the loss of spatial detail during downsampling, allowing the network to jointly exploit low level structural information and high level semantic features.

Following each fusion operation, the combined feature maps are further refined using two 3×3 convolutional layers with BN and ReLU activation functions, facilitating effective integration of multi-scale contextual information. Finally, a convolutional layer 1×1 is used to project the representation of the characteristic onto the target class space, followed by a Sigmoid activation function to generate a pixel wise probability map indicating the probability that each pixel belongs to the PS class. This completes the end-to-end semantic segmentation process.

3.6 Proposed Solution

As outlined above, the core idea of the proposed method is to leverage DL techniques to construct a novel discriminative statistic for PS detection. This learned statistic replaces the conventional GLRT-based statistic in (3.9), thereby enhancing the detection capability for coherent surface scattering structures and spatial patterns. In contrast to classical approaches, which rely on predefined statistical models and threshold based decision rules, the proposed method adopts a data driven paradigm to automatically learn discriminative feature representations from multi-temporal SAR data. This enables more robust and accurate PS detection, particularly in challenging scenarios characterized by low SNR, pronounced spatial heterogeneity, and the coexistence of multiple scattering mechanisms.

From an implementation perspective, the proposed framework consists of several key components. First, the construction of the input dataset for training, validation, and testing is described in detail. Given the inherent characteristics of SAR data, to this end, the GLRT derived feature maps are first preprocessed and normalized to ensure numerical stability and consistency across samples. A sliding window strategy is then employed to extract fixed size image patches, which serve as inputs to the network. These patches inherently encode local spatial context and structural patterns associated with

PS distribution. Supervision labels are generated based on reference PS detection results obtained from GLRT method, ensuring statistical consistency between the training targets and the underlying detection framework. Subsequently, the dataset is partitioned into training, validation, and test subsets, enabling robust model training and unbiased performance evaluation.

In terms of network architecture, a multi-branch model, termed 3U-Net, as shown in Fig.23, is proposed, consisting of three parallel U-Net sub-networks. Each branch is designed to extract complementary feature representations from the input data, such as spatial structural information, multiscale contextual features, and alternative feature embeddings. This multi perspective feature extraction mechanism enables a more comprehensive characterization of PS related discriminative cues. Each branch outputs a probability map indicating the probability of each pixel belonging to the PS class. The multi-brach design significantly enhances the model’s representational capacity, thereby improving robustness and generalization performance in complex scattering environments.

To effectively integrate the predictions from different branches, a WSM strategy is introduced. This fusion mechanism jointly considers the confidence of each branch and the inter-branch consistency to perform adaptive aggregation of multiple probability maps. Specifically, WSM assigns higher weights to mutually consistent predictions while suppressing unreliable or noisy responses, thereby reinforcing stable detections and improving the overall reliability and accuracy of the PS identification results.

Furthermore, a post-processing filtering step is incorporated to mitigate isolated false positives and outliers that may arise from pixel wise segmentation. This step aims to remove spatially inconsistent PS candidates by enforcing local consistency constraints. In particular, pixel whose estimated DMV significantly deviates from that of neighboring PS pixels are identified as outliers and discarded. This filtering strategy is grounded in the physical assumption that true PS typically exhibit locally coherent deformation behavior, thereby improving both the physical interpretability and spatial consistency of the final results.

Finally, the network training strategy is detailed, including the selection of the loss function and the definition of performance evaluation metrics. To address class imbalance and boundary ambiguity inherent in pixel level segmentation tasks, loss, Dice coefficient (Dice) loss, or hybrid loss formulations can be employed. For performance evaluation, standard metrics in semantic segmentation including Accuracy (ACC) and Intersection over Union (IoU) are adopted to quantitatively assess the model’s effectiveness on the validation set.

1. Training Set Construction

The design of input features plays a critical role in determining model performance during DL training. In this work, three types of features derived from SAR time-series data are employed as network inputs, namely the velocity feature X_V , the statistical feature X_T , and the entropy feature X_E , all defined over a spatial domain of size $W \times H$. Specifically, X_V corresponds to the DMV estimated from (3.10), while X_T represents the GLRT-based detection statistic defined in (3.9). These components are not further elaborated here; instead, the focus is placed on the construction of the entropy feature X_E .

The entropy feature X_E is designed to characterize the scattering stability of each pixel from the perspective of local deformation consistency. For a given pixel i , a set of neighboring pixels with similar deformation behavior referred to as a similarity set L , it is identified using a Kolmogorov-Smirnov (KS) statistical test applied to the DMV. Specifically, the KS test is employed to evaluate whether the DMV values of neighboring pixels are drawn from the same underlying distribution as that of the reference pixel i . Pixels that satisfy the KS criterion are considered statistically consistent in terms of deformation behavior and are therefore include in the similarity set.

To account for the spatial heterogeneity inherent in SAR data, an adaptive selection strategy is adopted. The number of similar pixels is not fixed but varies dynamically according to local conditions, ranging from 0 to 25 pixels. In regions with smooth and continuous deformation, a larger number of neighboring pixels typically satisfy the KS test, enabling robust estimation of local statistics. In contrast, in areas with abrupt deformation changes or complex scattering behavior, fewer pixels are selected, effectively avoiding the inclusion of inconsistent samples. This adaptive mechanism ensures that the similarity set reflects local deformation continuity rather than purely spatial proximity.

Based on the constructed similarity set, the entropy feature is introduced to quantify the uncertainty and dispersion of deformation behavior within the local neighborhood. From an information theoretic perspective, entropy measures the degree of uncertainty associated with a random variable. When the DMV values within the similarity set are highly concentrated, indicating strong local deformation

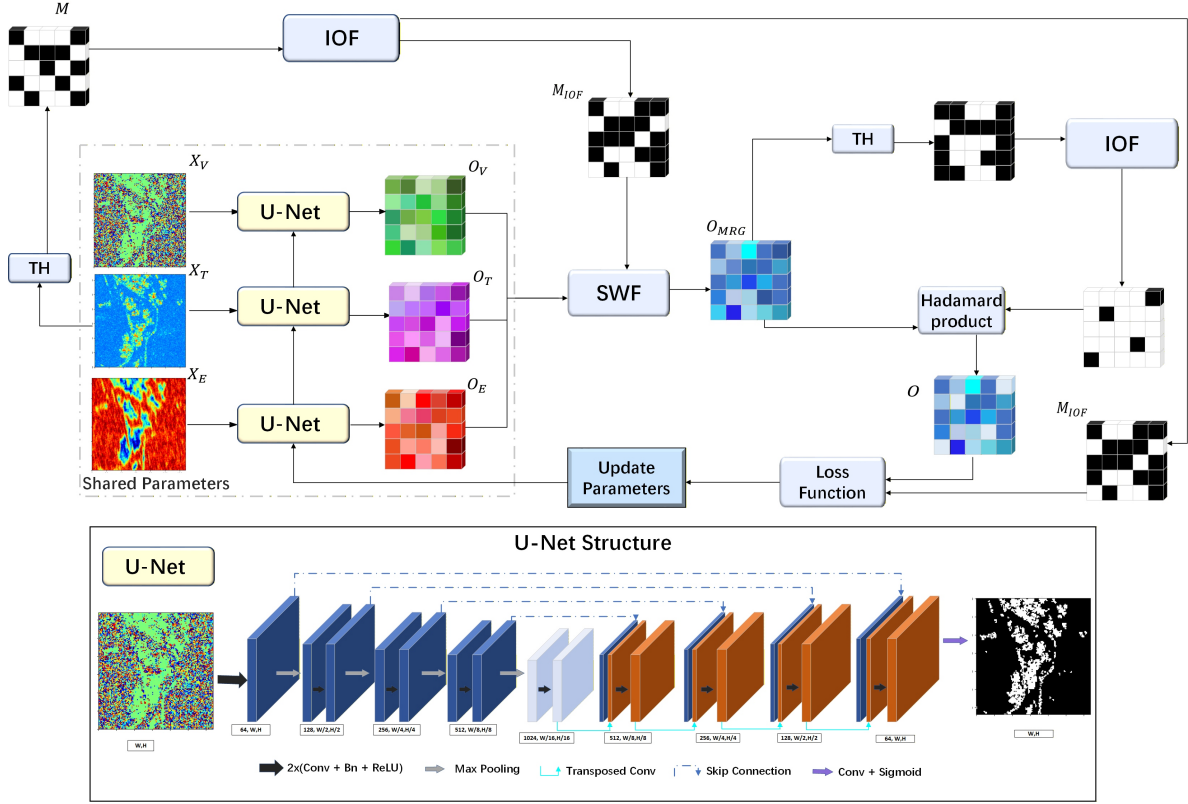


Figure 23: Structure of the proposed 3U-Net algorithm based on the use of the GLRT statistic and the corresponding DMV (upper image). U-Net architecture (lower image).

consistency that the resulting distribution exhibits a pronounced peak structure, leading to a low entropy value. In contrast, when the DMV values are widely distributed without a dominant mode, the entropy increases significantly, reflecting higher uncertainty and variability.

From a physical perspective, low entropy regions typically correspond to areas with stable scattering mechanisms and spatially coherent deformation patterns, which are more likely to contain PS. In such regions, the deformation behavior is smooth and can be effectively described by simple parametric models. Conversely, high entropy regions are often associated with deformation discontinuities or complex scattering environments, such as vegetation, water bodies, or areas affected by decorrelation. In these regions, the estimated deformation is unstable and more susceptible to noise and modeling error.

Therefore, the entropy feature X_E provides a deformation consistency driven reliability indicator, complementing the pixel wise GLRT statistic X_T . While X_T evaluates the agreement between observations and a parametric scattering model, X_E captures the spatial continuity of deformation, introducing an additional structural constraint. The combination of these features enhances the model's ability to discriminate PS, particularly in challenging scenarios involving weak scatterers, low SNR, or heterogeneous deformation fields.

Formally, based on the constructed similarity set of size L , the entropy is computed directly from the DMV values of the selected pixels. Specifically, the probability associated with the j -th similar pixel is defined as:

$$p_{j_i} = \frac{v_{j_i}}{\sum_{j=1}^L v_{j_i}}, \quad j = 1, \dots, L \quad (3.11)$$

where v_{j_i} denotes the DMV value of the j -th pixel in the similarity set corresponding to the reference pixel i . Based on this definition, the entropy E_i is computed as:

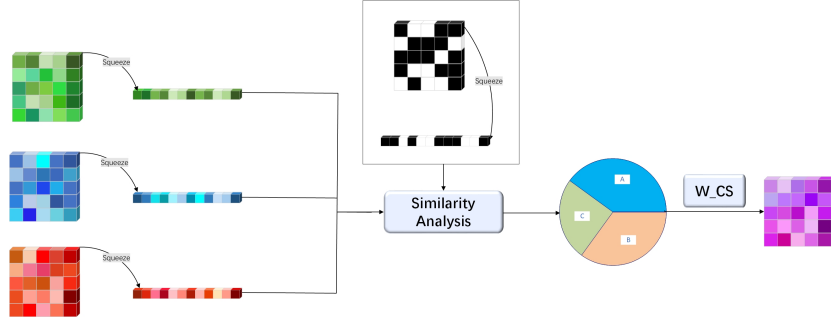


Figure 24: Structure of the MSE module.

$$E_i = - \sum_{j=1}^L p_{j_i} \log p_{j_i} \quad (3.12)$$

As discussed in the previous part, PS detection is formulated as a semantic segmentation task. Accordingly, the ground truth is represented as a binary segmentation map indicating the presence or absence of PS. Specifically, the initial ground truth map Fig.23 M is obtained by applying a threshold to the GLRT statistic T , computed to conventional GLRT-based detection, a lower threshold is adopted, corresponding to a higher probability of PFA. The final PS mask is then obtained through clustering and outlier filtering, and is used as the ground truth label, denoted as M_{IoF} . The proposed method exploits the capability of neural networks to handle noisy and imperfect labels, enabling the identification of PS exhibiting spatial consistency.

2. 3U-Net for PS detection

Fig.23 illustrates the overall architecture of the proposed 3U-Net. The backbone of the 3U-Net is based on the classical U-Net architecture. For each data stream, a U-Net is employed as a symmetric encoder-decoder network, which is specifically designed for pixel wise semantic segmentation tasks. The encoder progressively extracts hierarchical feature representations through successive downsampling operation, while the decoder gradually restores the spatial resolution via upsampling. In particular, the network emphasizes the recovery of high-resolution feature details through skip connection between corresponding encoder and decoder layers.

The proposed 3U-Net consists of three parallel data streams, each driven by a distinct input feature, namely the velocity feature X_V , the statistical feature X_T , and the entropy feature X_E . Each stream generates a corresponding probability map (O_V , O_T , O_E), representing the probability that each pixel belongs to a PS under the respective feature constraints. These three probability maps are subsequently fused through the module to produce a unified segmentation output. As illustrated in Fig.24, the fusion process is formulated based on similarity measures between the outputs of each stream and the reference label. Specifically, O_V , O_T and O_E are combined via a weighted aggregation scheme to generate a single probability map, denoted as O_{MRG} , which represents the final PS detection probability. The fusion weights are adaptively estimated using the Cosine Similarity Algorithm (CSA), which is based on the classical cosine similarity measure for quantifying the correlation between two vectors. In this framework, the reliability of each feature branch is evaluated by computing the similarity between its output and the reference label. The resulting similarity scores, denoted as S_V , S_T and S_E , are then used to guide the weighted fusion process. Accordingly, the final fused probability map is computed as follows:

$$O_{MRG} = \frac{S_V O_V + S_T O_T + S_E O_E}{S_V + S_T + S_E} \quad (3.13)$$

A notable characteristic of the proposed 3U-Net framework is that, in addition to achieving high pixel wise segmentation accuracy, it produces detection results that exhibit strong spatial consistency in deformation velocity estimation. Following the strategy adopted during the ground truth generation stage, an IOF module is further incorporated into the framework, as shown in Fig.25. This module operates as a post-processing step to refine the preliminary detection results. By exploiting the spatial continuity property of deformation velocities associated with PS points, the IOF module identifies and

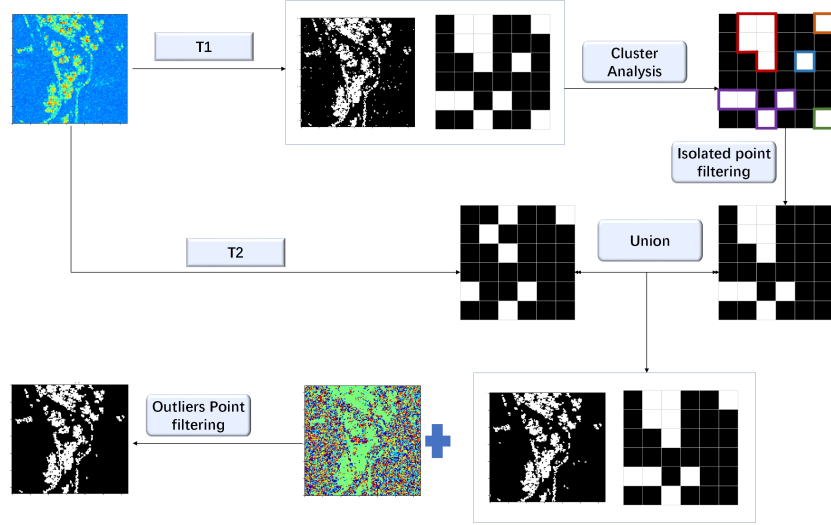


Figure 25: Structure of the IOF module.

suppresses outliers whose estimated velocities deviate significantly from those of neighboring pixels within a local spatial neighborhood. This filtering process effectively mitigates the influence of spatially inconsistent predictions and enhances the robustness and reliability of the final PS detection results.

Specifically, to eliminate spatially isolated noise points, clustering analysis is performed on the initial label map M . This is achieved through a C-connectivity analysis, where neighborhood relationships among foreground pixels are defined to group them into connected components. The connectivity parameter C is typically set to 4 or 8, corresponding to 4-neighborhood connectivity (top, down, left, right) and 8-neighborhood connectivity (including diagonal adjacency), respectively. During the connectivity scanning process, foreground pixels satisfying the neighborhood criteria are recursively or iteratively aggregated into the same connected component, resulting in a partition of the label map into non-overlapping regions. Within this framework, true PS regions tend to form large, spatially continuous components, whereas noise or false detections typically appear as isolated pixels or small connected components. By explicitly modeling spatial adjacency, the C-connectivity analysis naturally separates such noise into small independent regions. These can be filtered using an area threshold, where components smaller than a predefined size are classified as noise and removed. This procedure effectively suppresses isolated outliers in both the initial label map and the probability output after the *WSM* Module, thereby improving the spatial consistency and structural reliability of the detection results. In the sense that:

$$M = \bigcup_{k=1}^K C_k, \quad C_i \cap C_j = \emptyset, \text{ for } i \neq j \quad (3.14)$$

and then the output label map is generated as follows:

$$M_I(i, j) = \begin{cases} 1, & (i, j) \in C_k \text{ and } |C_k| > T_p \\ 0, & \text{otherwise} \end{cases} \quad (3.15)$$

where $|C_k|$ is the cardinality of the k -th cluster. In our case we set $T_p = 1$.

During the connectivity based filtering stage, small connected components are typically regarded as noise. However, such components may still contain a limited number of true positive pixels. To mitigate the risk of erroneously discarding these valid detections, a compensation strategy is introduced to reincorporate isolated pixels with low PFA into the label map after filtering. This approach effectively balances isolated noise suppression and target preservation, thereby improving the completeness and reliability of the final labels, as illustrated in Fig. 25.

The second filtering feature is designed to suppress outliers and enhance the spatial consistency of the velocity field. Specifically, for the foreground mask M_{IOF} obtained after preliminary filtering, the DMV is extracted from the velocity map for each pixel, and a local statistical reference is constructed

within the neighborhood $\mathcal{N}(i)$. To characterize local velocity consistency, the median is adopted as a robust estimator. For each pixel i , its velocity v_i is compared with the median of its neighborhood, and only those satisfying the following consistency constraint are retained:

$$|v_i - \text{median}\{v_j \mid j \in \mathcal{N}(i)\}| \leq \tau \quad (3.16)$$

where τ denotes a predefined threshold. This criterion defines a median value based consistency filtering mechanism, which effectively suppresses outliers while preserving the spatial coherence of the velocity field.

The two filtering strategies connectivity based filtering and outlier filtering, those are applied both during label construction for similarity evaluation and to the fused network outputs prior to loss computation. By incorporating these constraints into the training process, the model is explicitly encouraged to suppress spatially isolated PS and pixels exhibiting significant deviations in DMV relative to their local neighborhood, thereby improving the spatial coherence and physical reliability of the detection results.

3. Loss Function

The loss function is designed to quantify the consistency between the network predictions and the corresponding labels, which consist of $N_p = WH$ pixels. In binary semantic segmentation tasks, Binary Cross Entropy (BCE) is a commonly adopted pixel wise loss function, defined as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N_p} \sum_{i=1}^{N_p} \left[M_{\text{IOF},i} \log O_{\text{MRG},i} + (1 - M_{\text{IOF},i}) \log(1 - O_{\text{MRG},i}) \right] \quad (3.17)$$

where $O_{\text{MRG},i}$ denotes the predicted probability at pixel i from the fused output, and $M_{\text{IOF},i}$ represents the corresponding label. BCE performs fine grained optimization by penalizing discrepancies between predicted probabilities and ground truth labels at the pixel level. However, due to its symmetric treatment of positive and negative samples, BCE is susceptible to class imbalance. In scenarios where background pixels dominate, the loss can be biased toward the majority class, thereby weakening the learning of minority foreground(PS) pixels.

To alleviate this limitation and enhance the model's sensitivity to foreground regions, this paper incorporates the Dice loss, defined as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_{I=1}^{N_p} M_{\text{IOF},I} O_{\text{MRG},I}}{\sum_{I=1}^{N_p} M_{\text{IOF},I} + \sum_{I=1}^{N_p} O_{\text{MRG},I}} \quad (3.18)$$

Dice loss evaluates the agreement between predictions and labels from a region overlap perspective, aiming to maximize the overlap between predicted and true foreground regions. By normalizing the extent of the foreground, it effectively mitigates the influence of the dominant background class. This property makes it particularly suitable for handling class imbalance and sparse positive samples.

Within the proposed self-supervised framework, the M_{IOF} labels may be incomplete or contaminated by noise, due to their derivation from heuristic or statistical methods. Under such conditions, relying solely on pixel wise losses(e.g., BCE) may lead to false suppression of potential positive samples. In contrast, Dice loss enforces regional consistency constraints, enabling the model to recover structurally coherent foreground regions even when labels are partially missing or inaccurate.

Based on these considerations, a composite loss function is adopted as a weighted combination of BCE and Dice losses:

$$\mathcal{L}_{\text{combine}} = \alpha \mathcal{L}_{\text{BCE}} + \beta \mathcal{L}_{\text{Dice}} \quad (3.19)$$

where α and β are hyperparameters that balance the contributions of the two components. This hybrid formulation jointly optimizes the accuracy of the classification in pixels and the structural consistency of the region level, thereby improving the robustness and generalization capacity of the model in the presence of class imbalance and imperfect labels.

4. valuation Index

In DL-based evaluation, the performance of the proposed method is assessed using two widely adopted metrics: IoU and ACC. Their definitions are given as follows:

$$\text{IoU} = \frac{TP}{FN + FP + TP} \quad (3.20)$$

$$\text{ACC} = \frac{TP + TN}{TP + FP + FN + TN} \quad (3.21)$$

where TP , FP , TN and TN denote true positives, false positives, false negatives, and true negatives, respectively.

ACC quantifies the overall classification accuracy by measuring the proportion of correctly classified pixels over the entire image, considering both positive (PS) and negative (non-PS) classes. As such, it provides a global assessment of the model’s classification capability. However, in semantic segmentation tasks particularly in foreground sparse scenarios such as PS detection ACC can be dominated by the majority negative class due to the prevalence of background pixels, potentially masking the model’s actual performance on foreground regions.

In contrast, IoU evaluates model performance from a region overlap perspective, defined as the ratio between the intersection and union of the predicted and ground truth positive regions. This metric directly reflects the localization accuracy and segmentation quality of the target (PS) regions. Since IoU focuses exclusively on the consistency of the positive class, it is inherently more robust to class imbalance and provides a more faithful assessment of performance in detection and segmentation tasks.

From an evaluation standpoint, ACC and IoU capture complementary aspects of model performance. While ACC reflects global classification accuracy, IoU emphasizes spatial alignment and structural consistency of the predicted target regions. Their joint use forms a more comprehensive evaluation framework: ACC ensures overall classification stability, whereas IoU enforces accurate delineation of PS regions, preventing artificially inflated performance caused by background-dominated predictions.

Consistent with the design philosophy of the proposed loss function balancing pixel wise accuracy (BCE) and region level consistency (Dice), the combined use of IoU and ACC during evaluation provides a more reliable and holistic assessment of model performance. This is particularly important in complex remote sensing scenarios characterized by class imbalance and imperfect pseudo-labels, where such a dual-metric evaluation strategy enhances both the robustness and interpretability of the results.

3.7 Experimental Results

3.7.1 Overview of the CSK Satellite System and Data

The data used in this study are acquired from the CSK constellation and its second-generation system, CSG, operated by the Italian Space Agency (ASI). The system operates in the X-band, with a central frequency of approximately 9.6 GHz, and provides all-weather, day-and-night Earth observation capabilities. The CSK constellation consists of four satellites launched between 2007 and 2010, while the CSG system, as an upgraded generation, has been deployed since 2019. Both generations operate in near-polar sun-synchronous orbits, with an orbital altitude of approximately 619 km and an inclination of about 97.86°, and each satellite has a revisit cycle of 16 days. Under a multi-satellite cooperative configuration, the system enables high temporal sampling density for data acquisition.

In terms of imaging modes, the CSK system supports several acquisition modes, including Stripmap, PingPong, and ScanSAR. The Stripmap mode provides meter-level spatial resolution, making it suitable for detailed structural observation. The PingPong mode achieves dual-polarization acquisition through polarization multiplexing, while the ScanSAR mode enables wide-area coverage by employing multiple sub-swath scanning. Building upon this framework, the CSG system further enhances imaging capabilities by improving spatial resolution and radiometric performance, and by supporting more flexible polarization configurations, including full-polarization (QuadPol) acquisition. These imaging modes allow a flexible trade-off among spatial resolution, swath coverage, and polarimetric information, depending on specific application requirements.

Regarding data products, the CSK system provides multi level standard products, including SLC, Multi-Look Intensity (MLI), and Geocoded Ellipsoid Corrected (GEC) products. Among these, the SLC product preserves the full complex information, including both amplitude and phase, and accurately characterizes the electromagnetic scattering behavior of ground targets. It serves as the fundamental data type for InSAR and time-series analysis. The MLI product reduces speckle noise through multi-looking and enhances image interpretability, while the GEC product further applies geometric

correction, making it suitable for geospatial analysis and visualization. In this study, the analysis is primarily based on SLC data to fully exploit the phase information and scattering characteristics.

A detailed description of the system parameters and acquisition modes of the CSK and CSG constellations is provided in the following tables.

Table 1: CSK acquisition modes.

Parameter	Stripmap Himage	Stripmap PingPong	ScanSAR Wide	ScanSAR Huge
Resolution (m)	2.6×3	9.7×11	23×13.5	38×13.5
Scene Size (km)	40×40	30×30	100×100	200×200
Polarisation	HH, HV, VH, VV	HH/VV, HH/HV, VV/VH	HH, HV, VH, VV	HH, HV, VH, VV
Duration (s)	7	6	15	30
Number of Looks	3	3	4–9	25–66
multi-look Resolution (m)	5	20	30	100
Geolocation Accuracy (m)	25	25	30	100

3.7.2 Experiments on Real Data

The towns of Lungro and Salerno, locate in the mountainous and coastal urban regions of southern Italy, respectively, represent two markedly different types of surface environments. Lungro is characterized by a mountainous landscape with dense vegetation cover and pronounced topographic variability. In such environments, the scattering mechanism is predominantly governed by volume scattering and mixed volume-surface scattering. Due to factors such as vegetation dynamics, soil moisture variability, and terrain induced geometric distortions, these areas are highly susceptible to spatio-temporal decorrelation. As a result, only a limited number of pixels exhibit stable backscattering behavior over time, leading to sparse and fragmented distribution of PS, as shown in Fig.26(a). In contrast, Salerno is situated in a highly urbanized coastal region, where man-made structures, such as buildings, roads, and infrastructure, are densely distributed. The scattering behavior in such areas is dominated by strong, deterministic scatterers associated with stable geometric configurations. These targets typically exhibit consistent electromagnetic responses across acquisitions, enabling them to preserve high interferometric coherence over long temporal baselines and effectively suppress spatio-temporal decorrelation effects. Consequently, PS are densely distributed and exhibit stability, as shown in Fig.26(b). The proposed method is evaluated on datasets acquired over these two representative terrain types, thereby enabling a comprehensive assessment of its robustness and adaptability under heterogeneous scattering conditions.

1. Dataset

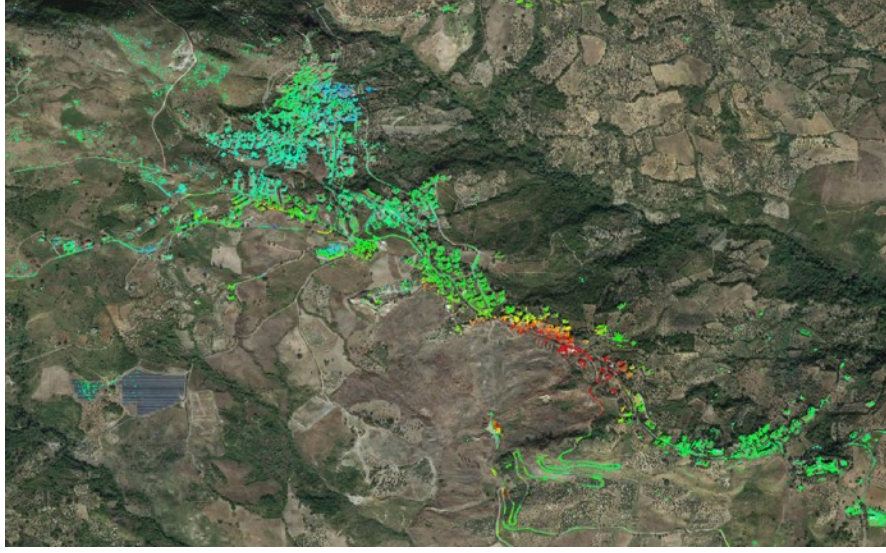
This study utilizes ultra-high-resolution X-band SAR data acquired by the CSK and CSK satellite constellations. The first dataset covers a coastal region located south of Salerno in southern Italy. It consists of 148 CSK SAR acquisitions with a spatial resolution of $3 \text{ m} \times 3 \text{ m}$, spanning the period from October 2012 to March 2022. The perpendicular baseline extent is approximately 2,660 m. For this scene, 17 image patches were extracted from the original large-scale SAR imagery, each with dimensions of $3,840 \times 3,840$ pixels in both the azimuth and range directions, to support subsequent training and evaluation tasks, the details shown in Table.4. The second dataset corresponds to the central Apennine mountainous region in the Lungro area of Calabria, southern Italy. It comprises 126 CSK acquisitions and 4 CSK acquisitions, all with a spatial resolution of $3 \text{ m} \times 3 \text{ m}$. The observation period extends from October 2011 to November 2021, with a perpendicular baseline span of approximately 2,850 m. In

Table 2: CSG acquisition modes.

Parameter	Stripmap	Stripmap PingPong	QuadPol	ScanSAR 1	ScanSAR 2
Resolution (m)	3×3	12×5	3×3	20×4	40×6
Scene Size (km)	40×40	30×30	40×15	100×100	200×200
Polarisation	HH, VV, HV, VH / Dual	HH/VV, HH/HV, VV/VH	HH+HV+VV+VH	HH, HV, VH, VV / Dual	HH, HV, VH, VV / Dual
Duration (s)	7	6	N/A	15	30
Number of Looks	2×2 / 4×4	1×2 / 2×5	2×2 / 4×4	1×3 / 1×5 / 2×8	1×4 / 1×7 / 3×16
multi-look Resolution (m)	6×7 / 11×14	12×10 / 23×26	6×7 / 11×14	20×14 / 23×27 / 35×40	40×27 / 47×54 / 115×135
Geolocation Accuracy (m)	3.75	12	3.75	12	12

Table 3: COSMO-SkyMed system and orbital parameters.

Parameter	Value
Orbit type	Sun-synchronous, near-polar
Altitude	~ 619 km
Inclination	$\sim 97.86^\circ$
Repeat cycle	16 days (per satellite)
Constellation revisit time	1 day (4-satellite constellation)
Radar frequency	X-band (9.6 GHz)
Wavelength	~ 3.1 cm
Orbit control accuracy	High-precision (cm-level, improved in CSG)
Incidence angle range	$\sim 20^\circ - 50^\circ$
Orbit direction	Ascending / Descending



(a) Lungro: DMV image.



(b) Salerno: DMV image.

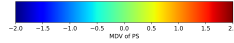


Figure 26: Study areas: Lungro and Salerno.

this case, a SAR image of size $2,816 \times 3,584$ pixels (in azimuth and range directions, respectively) was utilized for analysis, the details shown in Table.5. These two datasets exhibit pronounced differences in land cover characteristics, interferometric phase processing complexity, and the availability of effective training samples, thereby providing complementary test conditions. During the preprocessing stage, all datasets were processed using the classical SBAS technique [FPP+24] for phase calibration and time-series reconstruction. This workflow follows the standard MT-DInSAR processing paradigm, in which multi-temporal interferometric phases are first modeled and corrected prior to GLRT-based detection. Specifically, the procedure includes the estimation and removal of the APS, as well as the compensation of low-resolution nonlinear deformation components, thereby ensuring reliable inputs for subsequent scatterer detection.

2. Implementation Aspects

Both datasets were further processed using a sliding window based patch extraction strategy, where each patch has a spatial size of 128×128 pixels. The samples were randomly partitioned into training, validation, and test sets, accounting for 65%, 15%, and 20% of the total data, respectively. All

Table 4: SAR data description for Salerno area.

Area	Salerno, Campania, South Italy
SAR Sensor	CSK
Acquisition Mode	Stripmap
Spatial Resolution	$\sim 3 \times 3$ m
Acquisition Interval	October 2012 – March 2022
# of Images	148 CSK
Incidence Angle	$\sim 27.4^\circ$
Orbit Direction	Ascending

Table 5: SAR data description for Lungro area.

Area	Lungro, Calabria, South Italy
SAR Sensor	CSK / CSG (CSK/CSG)
Acquisition Mode	Stripmap
Spatial Resolution	$\sim 3 \times 3$ m
Acquisition Interval	October 2011 – November 2021
# of Images	126 CSK + 4 CSG
Incidence Angle	$\sim 32.3^\circ$
Orbit Direction	Descending

experiments were conducted on a single NVIDIA GeForce GTX 1080 Ti with 2×12 GB RAM, and implemented using the PyTorch framework.

During training, model parameters were optimized using the Adam optimizer with an initial learning rate of 0.001. To enhance convergence stability and training efficiency, a Polynomial Learning Rate Decay(Poly schedule) was adopted. The learning rate at training step t is defined as:

$$\eta_t = \eta_0 \left(1 - \frac{t}{T}\right)^\gamma \quad (3.22)$$

where η_0 denotes the initial learning rate, t is the current training iteration, T represents the total number of training iterations, and γ is the power factor controlling the decay rate. This schedule progressively decreases the learning rate throughout training, enabling a smooth transition from rapid initial optimization to fine-grained parameter refinement, thereby improving both convergence stability and generalization performance. In the experimental setup, the input patch size for the network was set to 32×32 , and the maximum number of training iterations was fixed at 100.

3. Results on Real Data

We compare the results of the traditional, GLRT based PS detector, and the detector provided by the proposed 3U-Net methods.

1)Salerno area: In the Salerno dataset, PS detection results based on the GLRT were obtained by setting the decision threshold $T = 0.37$ in (3.9), corresponding to a PFA of 1.6×10^{-6} . However, to ensure consistency with the label generation process, the results shown in the figure were produced using a lower threshold $T = 0.35$, corresponding to a higher PFA of 2.0×10^{-5} . Fig.27(a), (b), and (g) illustrate the PS distributions detected by the GLRT across three image patches, color-coded according to the estimated DMV. The corresponding 3U-Net based detection results are presented in Fig.27 (c), (f), and (I), using the same DMV based color encoding. Quantitative comparisons demonstrate that the 3U-Net method achieves substantial performance gains across all test regions. Specifically, the GLRT detects 1699262, 1427897, and 1429003 PSs in subgraphs (a), (d), and (g), respectively, whereas the 3U-Net identifies 2606805, 2394296, and 2418987 PSs in subgraphs (c), (f), and (I), corresponding

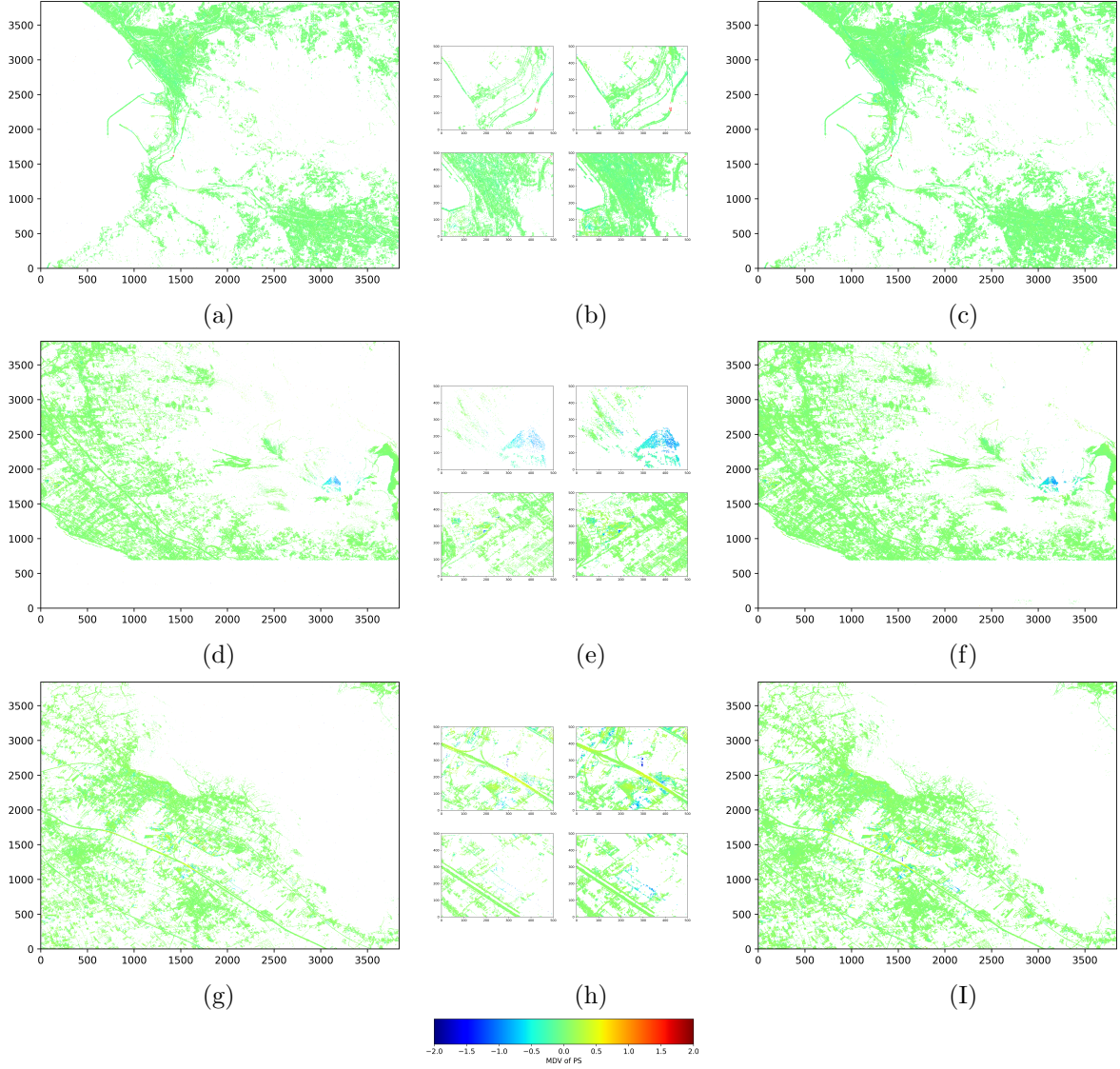


Figure 27: PS for the Salerno dataset: Low-threshold with GLRT after IOF filtering (leftmost column) and 3U-Net (rightmost column) results. Zooms on selected areas are shown in the mid columns with GLRT after IOF on the left and 3U-Net on the right.

to relative improvements of 53.4%, 67.7%, and 69.3%, respectively. For enhanced visual comparison, Fig. 27 (b), (e), and (h) provide magnified views of selected regions, where the left column shows GLRT results and the right column shows those obtained with 3U-Net.

From both visual inspection and structural analysis, the 3U-Net produces PS maps with a significantly denser and more spatially coherent distribution. In regions characterized by pronounced deformation gradients, such as building edges or deformation transition zones for the GLRT, which relies on strict phase model consistency, is particularly sensitive. This often results in coherence degradation and consequently leads to missed detection. In contrast, 3U-Net leverages multi-scale spatial context and learned feature representations to identify pixels exhibiting spatial consistency, even when local interferometric coherence is weak. This capability substantially improves detection completeness in deformation affected areas. Moreover, the 3U-Net demonstrates enhanced robustness in suppressing noise. The resulting PS maps exhibit smoother and more continuous spatial patterns, with a noticeable reduction in isolated spurious detections. This behavior can be attributed to the incorporation of spatial consistency constraints during training (via IOF module), as well as the modeling of neighborhood based statistical features such as DMV consistency. These mechanisms enable the network to effectively discriminate true scatterers with noise. By contrast, the GLRT being a pixel-wise statistical

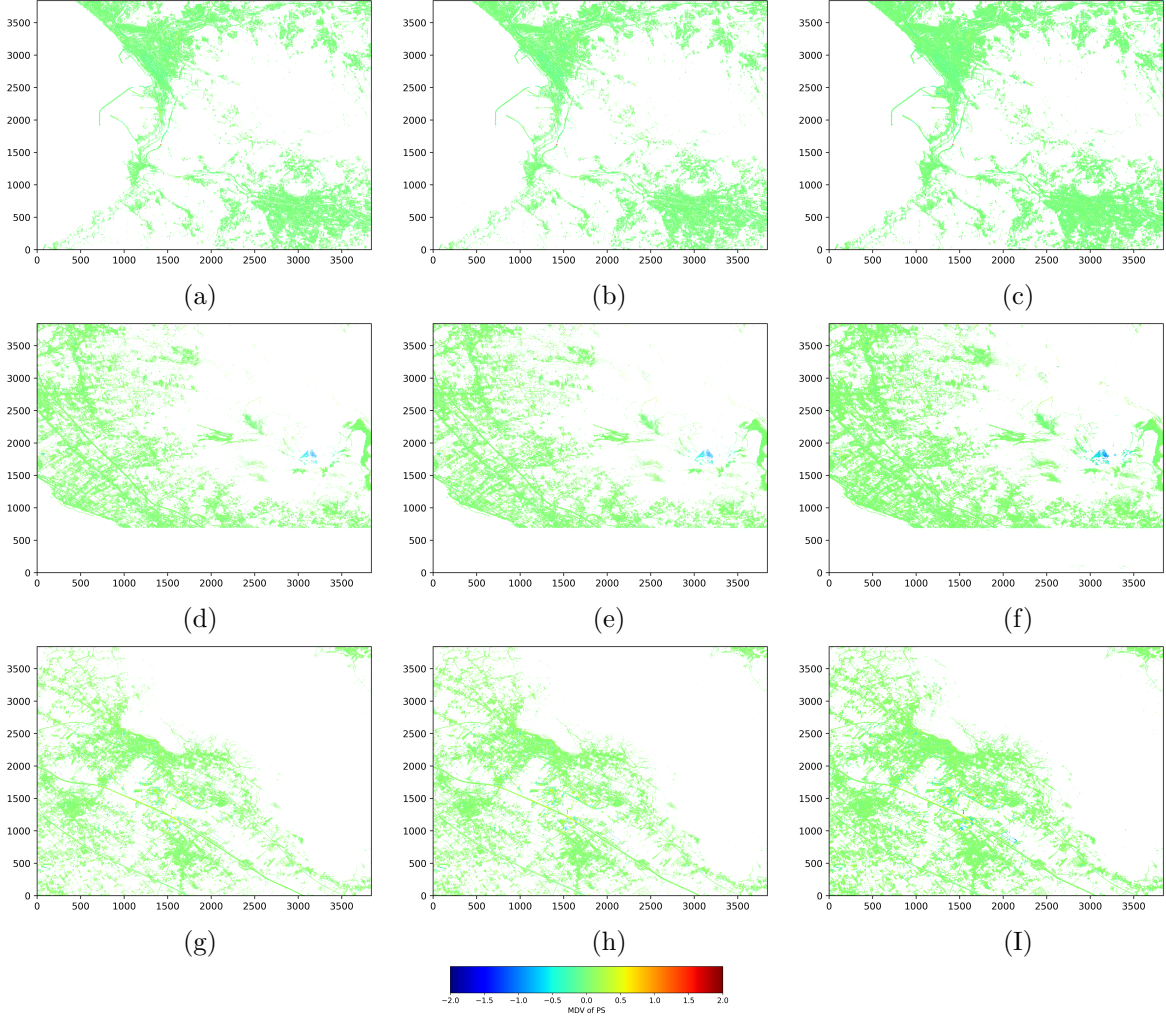


Figure 28: PS for the Salerno dataset: Standard threshold with GLRT (mid column), low-threshold with GLRT after IOF filtering (leftmost column) and 3U-Net (rightmost column) results.

detector, lacks the ability to exploit spatial dependencies and is therefore more prone to generating isolated false response, particularly in low SNR conditions or structurally complex scenes.

Under stricter GLRT settings (I.e. a lower PFA of 1.6×10^{-10}), the detection becomes increasingly conservative, resulting in a substantial reduction in the number of identified PSs. Under this benchmark, the relative improvements achieved by 3U-Net increase further to 66.7%, 85.5%, and 87.5%, respectively. A clear hierarchical relationship emerges among the three configurations in terms of detection capability: the standard GLRT with strict thresholds yields the lowest PS detection rate; the relaxed threshold GLRT used for label generation achieves a moderate detection rate; and the 3U-Net consistently attains the highest detection rate. These observations indicate that conventional GLRT based approaches tend to suffer from a high false negative rate under stringent statistical constraints. Although lowering the detection threshold (I.e. increasing the PFA) partially alleviates this issue, it remains insufficient for recovering spatially consistent yet weakly coherent PS candidates. In contrast, the proposed 3U-Net framework effectively identifies a larger set of physically meaningful scatterers while preserving spatial consistency, thereby achieving a superior trade off between detection completeness and reliability.

1) Lungro area:

The generalization capability is a critical performance criterion for DL-based methods. To further evaluate the robustness and cross-scene transferability of the proposed approach, additional experiments were conducted on the Lungro dataset. Specifically, the 3U-Net model trained on the Salerno dataset was directly applied to the Lungro dataset without retraining. For a fair comparison with

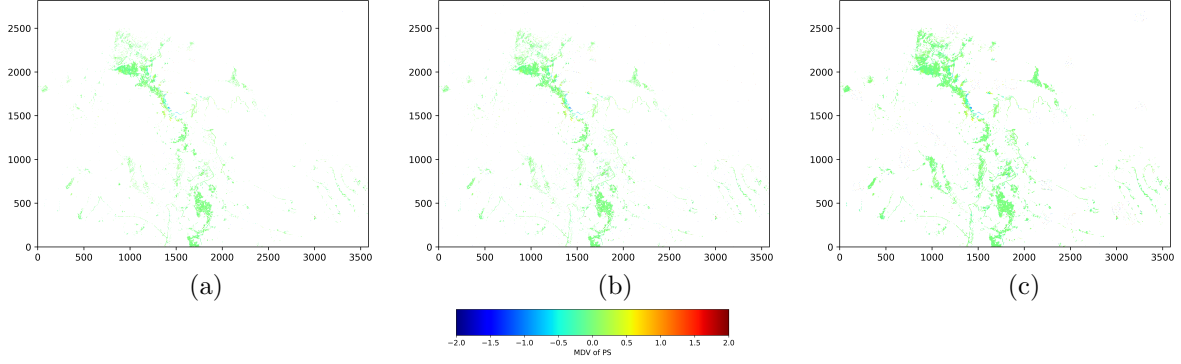


Figure 29: PS for the Lungro dataset: Standard-threshold with GLRT (left column), low-threshold with GLRT after IOF filtering (mid column) and 3U-Net (right column) results.

the GLRT-based detector, the PFA was first set to 1×10^{-5} , corresponding to a detection threshold of $T = 0.37$, yielding GLRT results under a relaxed threshold configuration. In contrast, the standard processing scheme for this dataset typically adopts a more stringent threshold of $T = 0.4$, corresponding to a lower PFA of 2×10^{-7} , in order to mitigate the impact of vegetation induced decorrelation and complex scattering behaviors. Fig. 29 presents the PS detection results obtained using different methods. Specifically, Fig. 29 (a) corresponds to the standard GLRT results (high threshold, low PFA), Fig. 29 (b) shows the GLRT results under a relaxed threshold (high PFA) after IOF module, and Fig. 29 (c) illustrates the results obtained using the proposed 3U-Net model. As observed, the standard GLRT approach produces conservative detections due to the stringent thresholding, resulting in the lowest number of identified PS points. Relaxing the threshold increases the detection density. In contrast, the 3U-Net approach achieves the highest detection density while effectively suppressing spurious responses, leading to a more favorable trade off between detection completeness and reliability.

4. Ablation Study

The 3U-Net architecture in Fig. 23 has been derived according to a strategy and includes specific modules for the merging and for the rejection of noise by filtering isolated PS and outliers. We have performed an ablation study to discuss the effects of the filtering and of using 3U-Net in place of 1U-Net.

1) IOF filtering: Fig. 30 illustrates the effect of the IOF module on both label refinement and noise suppression within the network. First, the role of IOF in label generation is reflected in the PS detection mask and the corresponding DMV map. Fig. 30 (a) and (c) present the initial detection mask and the associated DMV, respectively, obtained under a relaxed threshold setting. Owing to the elevated PFA, these results contain a considerable number of false positives. The corresponding filtered outputs are shown in Fig. 30 (b) and (d). It can be observed that the IOF module effectively removes the majority of spurious detections while preserving true label structures, it also contributes to improving PS prediction by using local velocity consistency constraints. Specifically, by leveraging the spatial continuity of the deformation field, the IOF module suppresses outliers that are inconsistent with neighboring velocity estimates. Beyond its role in label construction, Fig. 30 (e) and (f) show the PS detection results obtained using the proposed 3U-Net and the DMV map derived from the classical GLRT method under a low PFA (high threshold) configuration with IOF filtering, respectively.

2) WSM merging: This thesis analyses the fusion process in which three feature streams are integrated to generate the final segmentation probability map. The fusion weights are adaptively determined based on similarity measures between the outputs of each branch and the reference labels. Fig. 31 (a), (b), and (c) present the probability maps derived from the GLRT statistic, DMV, and entropy features, respectively, while the fused result is shown in Fig. 31 (d). As expected, the outputs of the GLRT and velocity branches in Fig. 31 (a) and 7 (b) exhibit higher spatial resolution, enabling more accurate representation of fine-scale structural details. In contrast, the entropy branch in Fig. 31 (c), constructed from local statistical descriptors, produces spatially smoother responses with relatively lower resolution. The information provided by the DMV feature is complementary to that of the GLRT statistic. Specifically, the velocity branch exhibits enhanced spatial continuity and yields a more homogeneous representation of deformation patterns, effectively capturing smooth variations in the

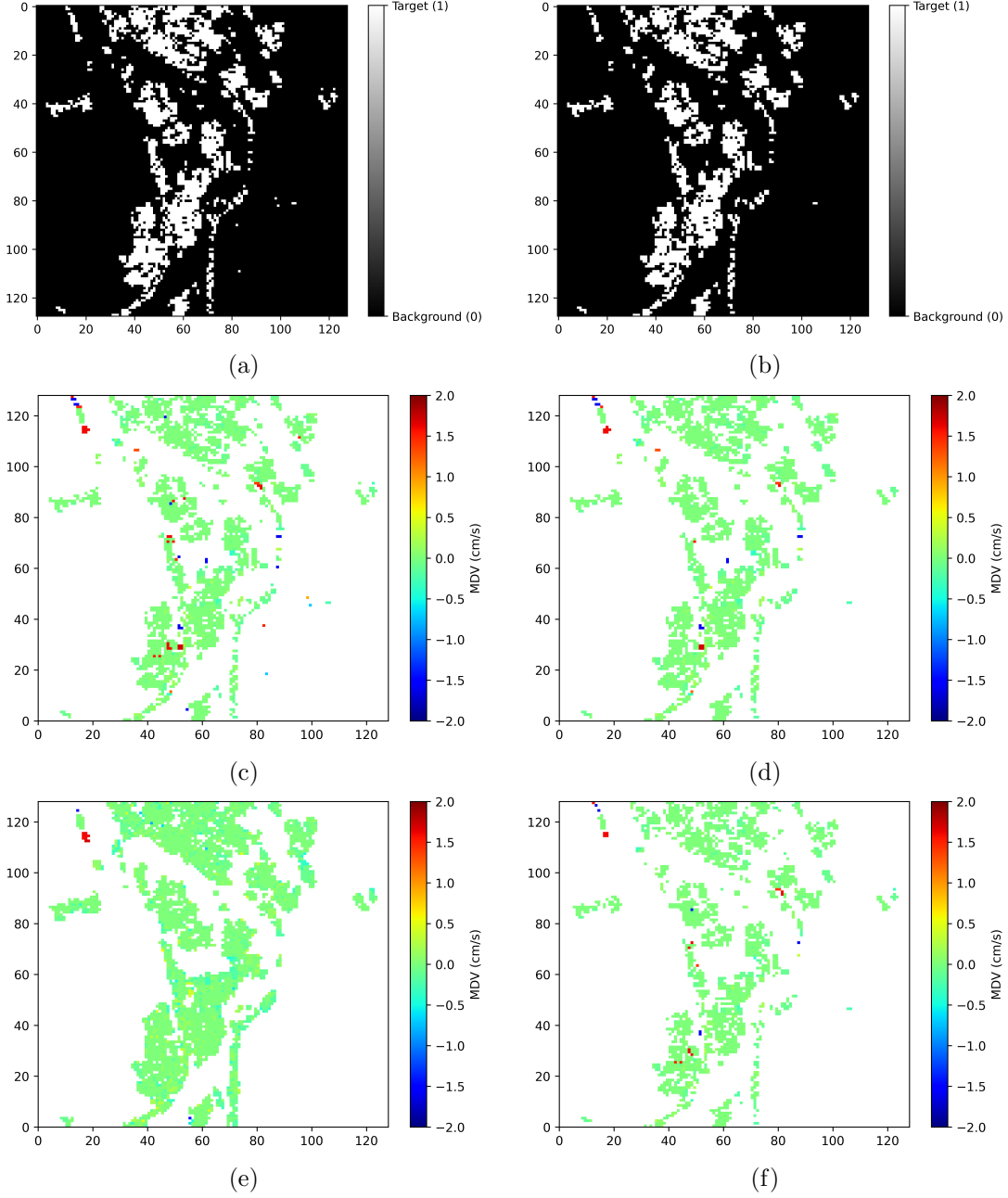


Figure 30: Initial label generated with a low threshold (a). Label after initial label with IOF filtering (b). PS velocity corresponding for the initial label (c), for the label after the IOF filtering (d), for the label predicted by 3U-Net (e) and detected with classical GLRT thresholding, I.e. with a high threshold with low PFA (f).

deformation field. Conversely, the GLRT branch operates at the pixel level and focuses on coherence-based discrimination, making it more sensitive to noise and local perturbations. Furthermore, the entropy feature in Fig.31(c) contributes by encoding local statistical consistency, which facilitates the identification of regions characterized by weak scattering or low SNR. This provides additional contextual support, particularly in challenging areas where conventional coherence-based indicators may be unreliable. During the weighted fusion stage (Fig.31(d)), the proposed WSM mechanism exploits the similarity relationships between each branch output and the reference labels to perform adaptive feature aggregation. This process effectively leverages the complementary properties of the three feature streams: it suppresses spurious responses, preserves high-resolution structural details,

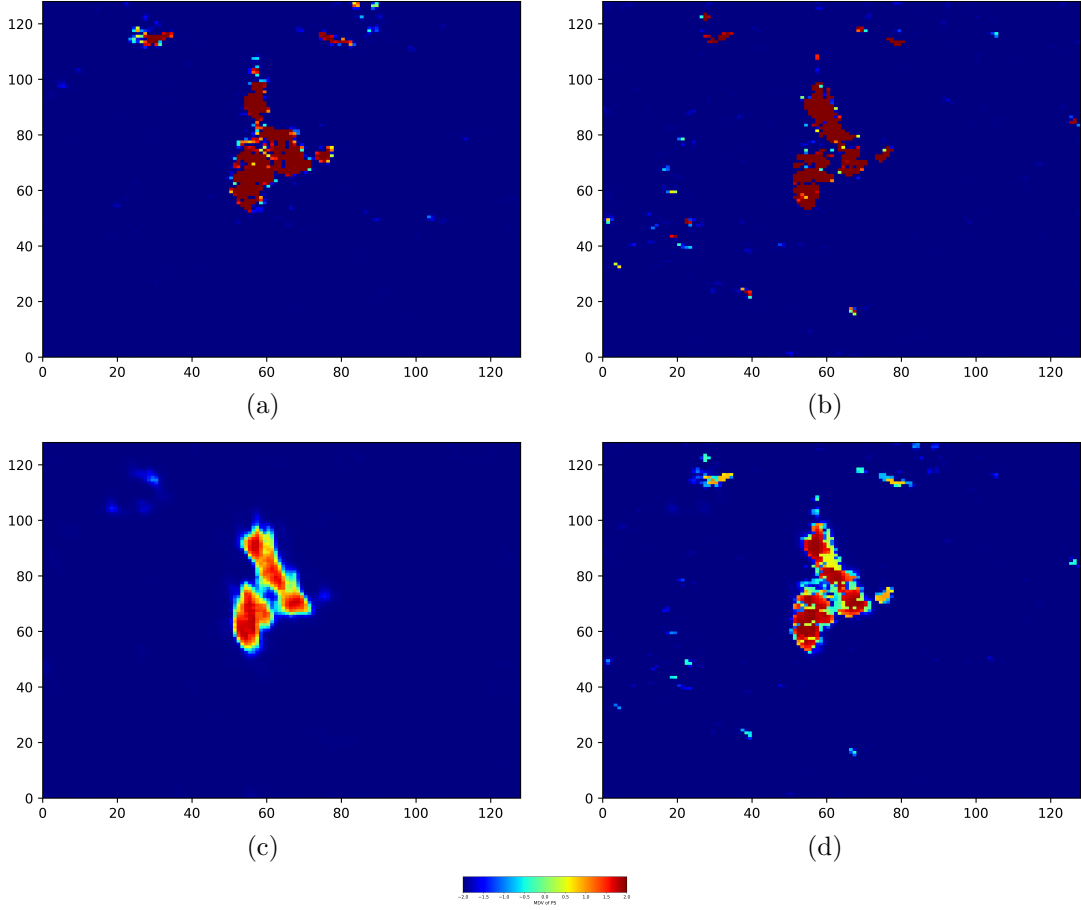


Figure 31: Probabilities associated with the three-streams outputs: displacement mean velocity DMV, GLRT statistics (b) and entropy (c). The probability to 3U-Net after WSM module merging is shown in (d).

and enhances regional consistency.

3) IOF after the WSM:

Next, it further analyzes the training dynamics of the network after introducing the IOF module into the fused probability map (right-hand side of Fig.31). Specifically, we investigate the effect of applying IOF filtering to the weighted fusion output (i.e. the WSM output) prior to loss computation, and compare it with the baseline setting where IOF is applied only during label generation. Fig.32 (a) and (b) present the PS detection results of the 3U-Net without IOF constraints and with IOF filtering applied to the WSM output, respectively. The results indicate that incorporating the IOF module directly into the optimization pipeline introduces a pronounced regularization effect on the learning process. In particular, the integration of IOF constraints leads to a more homogeneous and spatially continuous distribution of detected PS, while significantly reducing local outliers and spurious responses. This behavior suggests that enforcing velocity consistency constraints on the fused probability map during training enables the network to better capture physically consistent scattering patterns. As a result, both the robustness and reliability of PS detection are improved. Moreover, Compared with applying IOF solely at the label preprocessing stage, embedding this constraint within the training loop allows spatial consistency to be enforced throughout the optimization process. This yields a more effective supervision signal and ultimately enhances the overall performance of the model.

4) Loss function selection: PS and noise pixels typically exhibit a highly imbalance spatial distribution, and the proportion of PS pixels relative to the total pixel count can vary substantially across different scenes. Such class imbalance and inter scene variability can significantly impact classification performance. To enhance robustness under these conditions, this thesis investigates two widely used loss functions and their combinations. Table 6 reports a quantitative comparison of different loss formulations using IoU

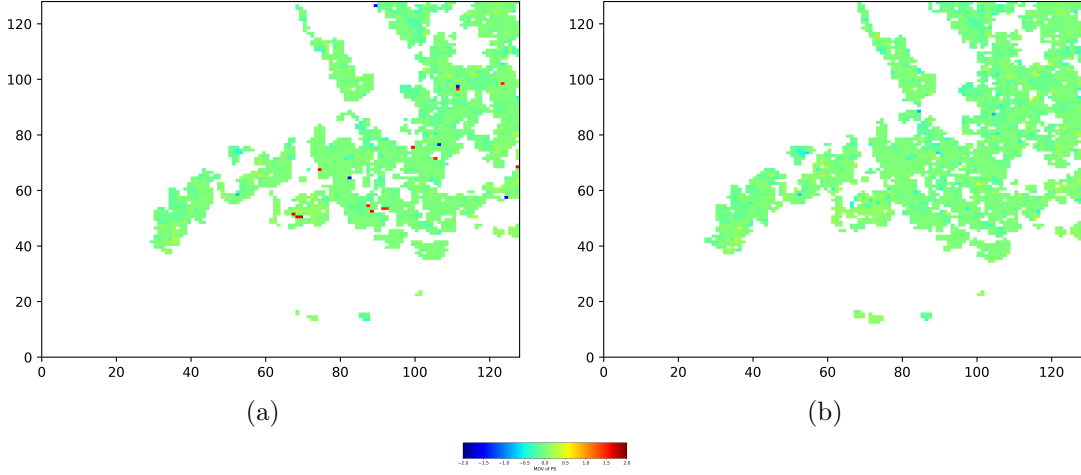


Figure 32: PS detection of the proposed 3U-Net architecture without (a) and with the IOF filtering module (b).

and ACC, as defined in (3.20) and (3.21). In the BCE formulation, positive and negative samples contribute symmetrically at the pixel level. However, due to the predominance of negative samples, the optimization process is often biased towards the background class, which limits the model’s sensitivity to the minority foreground class. In contrast, Dice loss is a region overlap driven objective that emphasizes the agreement between predicted and ground truth foreground regions. By directly maximizing the overlap ratio, Dice loss is particularly effective in improving recall and enhancing the detection of foreground structures. Nevertheless, its inherent bias towards positive samples may weaken the penalization of false positive, leading to over segmentation of the foreground. This, in turn, can degrade the IoU metric(see Eq.(3.20)) and increase the false positive rate. In other words, although Dice loss improves coverage of true PS regions, its standalone use may introduce spurious detections. To address these complementary limitations, this study adopts a combined loss formulation integrating BCE and Dice loss. BCE enforces pixel-wise discrimination, thereby effectively suppressing false negatives. Their combination enables a balance optimization between precision and recall, leading to improved performance under class imbalance and enhancing the robustness of PS detection.

Loss Type	IoU	ACC
BCE	0.7545	0.9593
Dice	0.7458	0.9565
BCE + Dice	0.8001	0.9604

Table 6: Performance comparison between the results obtained with the three different loss function combinations.

5). Difference between 1U-Net and 3U-Net:

This thesis further investigates the impact of incorporating the WSM mechanism and a multi-branch architecture on network performance. To this end, comparative experiments were conducted between a signal branch 1U-Net and the proposed 3U-Net. In the 1U-Net configuration, only the branch using DMV as input was retained. This design is motivated by the following considerations: entropy features, being derived from local statistical descriptors, inherently introduce a reduction in spatial resolution; meanwhile, the GLRT statistic is primarily utilized for label generation via thresholding and provides limited additional discriminative value when directly used as an input feature. The top and bottom rows of Fig.33 present a visual comparison of PS detection results obtained by 1U-Net and 3U-Net on two different image patches (left and right). The results indicate that the DMV driven 1U-Net is already capable of identifying a substantial number of PS pixels that are typically detected by the conventional GLRT thresholding approach, highlighting the strong discriminative capability of velocity features. However, after introducing the additional feature branches, named the GLRT statistic and entropy features, and integrating them through the WSM module, the 3U-Net

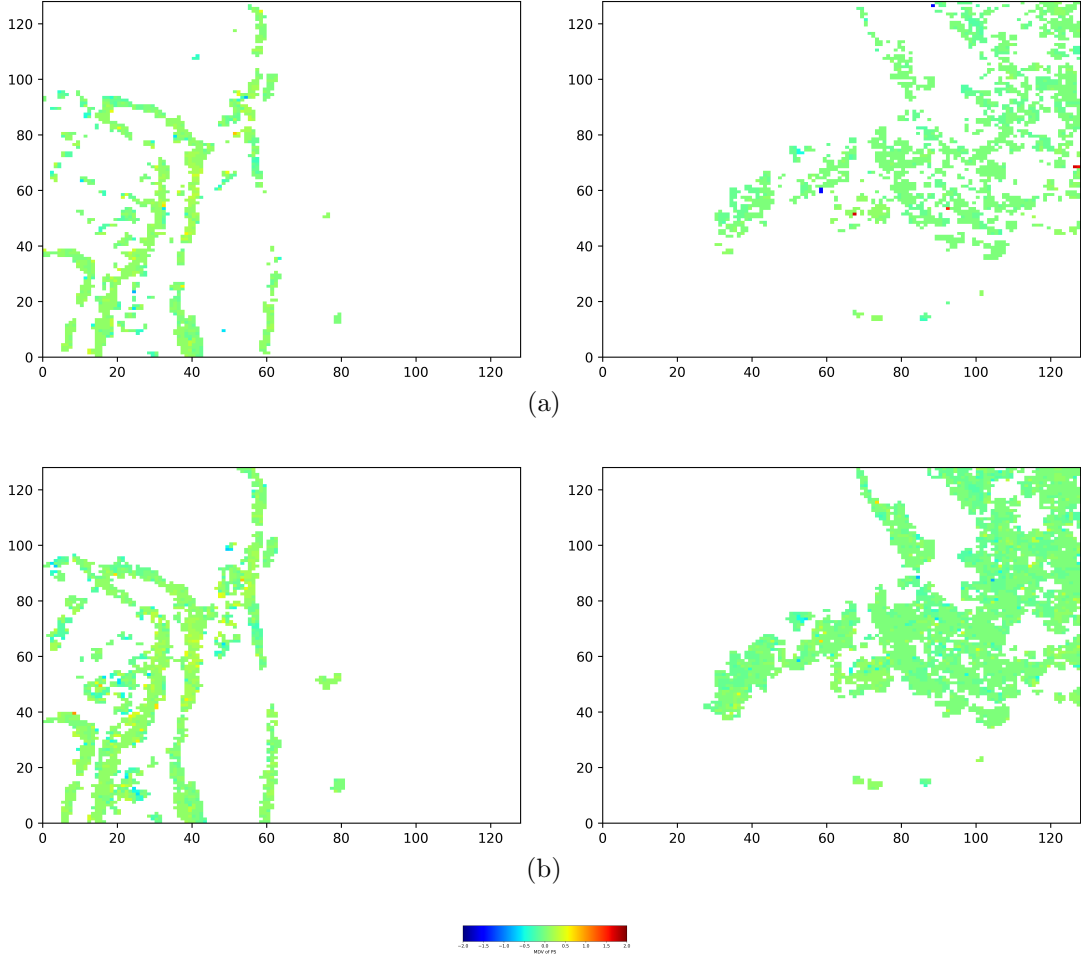


Figure 33: PS corresponding with the DMV detected by a 1U-Net (a), by 3U-Net (b).

achieves markedly improved detection performance. The multi-branch architecture effectively exploits complementary feature representations, enabling the network to jointly leverage pixel-level statistic based discrimination, spatial continuity of the deformation field, and local statistical consistency. As a result, the proposed model demonstrates enhanced capability in capturing PS as well as related spatial structural patterns.

3.7.3 Experiments on Simulated Data

To demonstrate the improvement of the proposed 3U-Net to the state-of-the-art GLRT detector, this Section considers a controlled, I.e. simulated scenario. In particular, for the CSK and CSG case, two simulated datasets were generated according to the characteristics in terms of spatial and temporal baseline and incidence angle of the real Salerno and Lungro data. These datasets are referred to as the Salerno simulated and Lungro simulated datasets.

Fig. 34 depicts a representative scene common to both datasets. Specifically, Fig. 34(a) and (b) present the DMV field and the corresponding topography, respectively. To simulate varying observation conditions, additive noise was introduced into the data. An appropriate amplitude distribution was designed such that the horizontal SNR varies continuously within the range of $-16dB$ to $-4dB$, as illustrated in Fig. 34(c). The three highlighted regions in the figure correspond to SNR levels of $-11dB$, $-8dB$, and $0dB$, respectively, enabling a comparative analysis of detection performance under different noise regimes. The triangular area located in the upper-right corner of Fig. 34(c) represents a signal-absent region, I.e. a highly decorrelated area where interferometric cannot be retrieved. The proposed gradient SNR configuration allows the simulated scenario to systematically span a wide range of observation conditions, from low to relatively high SNR levels. This controlled setup provides

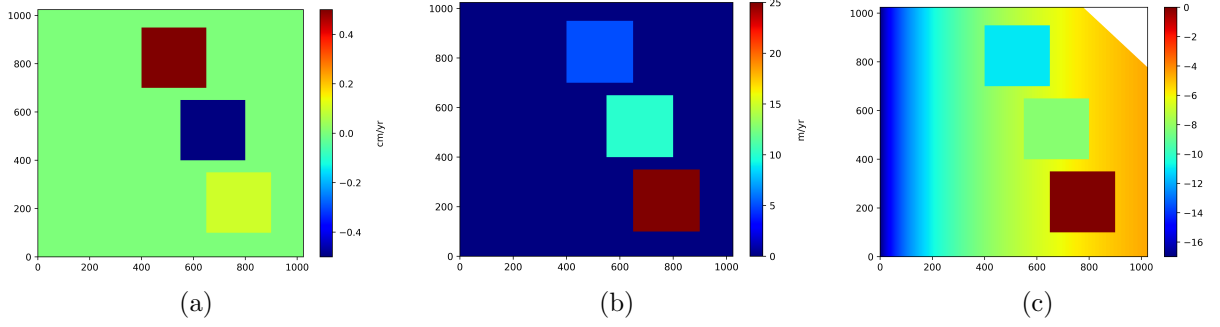


Figure 34: DMV (a), RT (b), SNR (c) scenario for the simulated datasets.

a representative and reproducible framework for analyzing detection performance curves. Furthermore, it facilitates a more rigorous assessment of the robustness and performance trends of different methods under noise and decorrelation effects.

1. Salerno simulation results

The simulated dataset constructed from Salerno measurements is processed using the conventional GLRT method and two 3U-Net models trained on real data from Salerno and Lungro, respectively, for comparative analysis. Specifically, Fig. 35(a) and (b) present the DMV estimated by the GLRT method before and after the thresholding operation, respectively. The detection threshold is set to $T = 0.35$, corresponding to a PFA of 1.5×10^{-4} . This relatively relaxed thresholding condition aims to increase the detection rate of potential PS candidates. Fig. 35 (c) and (d) illustrate the PS detection results obtained from the 3U-Net models trained on Lungro and Salerno datasets, respectively. From a visual perspective, the DL-based 3U-Net approach demonstrates clear advantages over the conventional GLRT method. In particular, it yields a denser and more spatially coherent distribution of PS, while exhibiting enhanced robustness to noise contamination. Moreover, even under cross scene condition, I.e. applying the Lungro trained model to Salerno based simulated data, the 3U-Net framework remains capable of effectively identifying spatially consistent scatterers, highlighting its strong generalization capability. When trained and tested on data from the same scene(Salerno), the model further improves in terms of structural detail preservation and completeness of PS distribution. This observation indicates that domain consistency between training and testing data can further enhance detection performance.

This work evaluates the performance improvement of the proposed method quantitatively by calculating the Probability of Detection (PD) and MSE for the estimated DMV using the SNR as the independent variable. The corresponding PD and MSE curves are shown in Fig. 36(a) and (b). Fig. 36(a) clearly shows that 3U-Net exhibits a higher PD than the classical GLRT detector across the entire SNR range, particularly under low SNR conditions. This suggests that DL models have stronger discriminatory capabilities and are more robust in environments with weak signals and strong noise interference. Further analysis of the MSE curves in Fig. 36(b) reveals that although the model trained on the Lungro real dataset shows a slight improvement in detection probability, its DMV estimation error is significantly higher than that cross scenario training may introduce estimation bias, impacting the accuracy of parameter inversion. Conversely, when the GLRT detector achieves stable detection performance in regions with a high SNR, the MSE curve of the 3U-Net model trained on the Salerno dataset gradually converges towards the GLRT results. This indicates that under high SNR conditions, the two methods are consistent in their accuracy of parameter estimation. Furthermore, in the highly uncorrelated region shown in Fig. 36(the signal free area in the top right corner), the GLRT detector produced three false positives, which is consistent with its theoretical PFA at a given threshold. In contrast, the 3U-Net model trained on the Lungro dataset produced six false positives, whereas the model trained on the Salerno dataset produced none. To conclude, from both the qualitative and quantitative point of view, the results achieved by 3U-Net trained on the corresponding real dataset are significantly better than those corresponding to the classical GLRT.

2. Lungro simulation results

Consistent with the experimental protocol adopted for the simulated Salerno dataset, the simulated Lungro dataset is processed using the conventional GLRT method and two 3U-Net models trained on real data from Salerno and Lungro, respectively, for comparative analysis. Specifically, Fig. 37(a) and

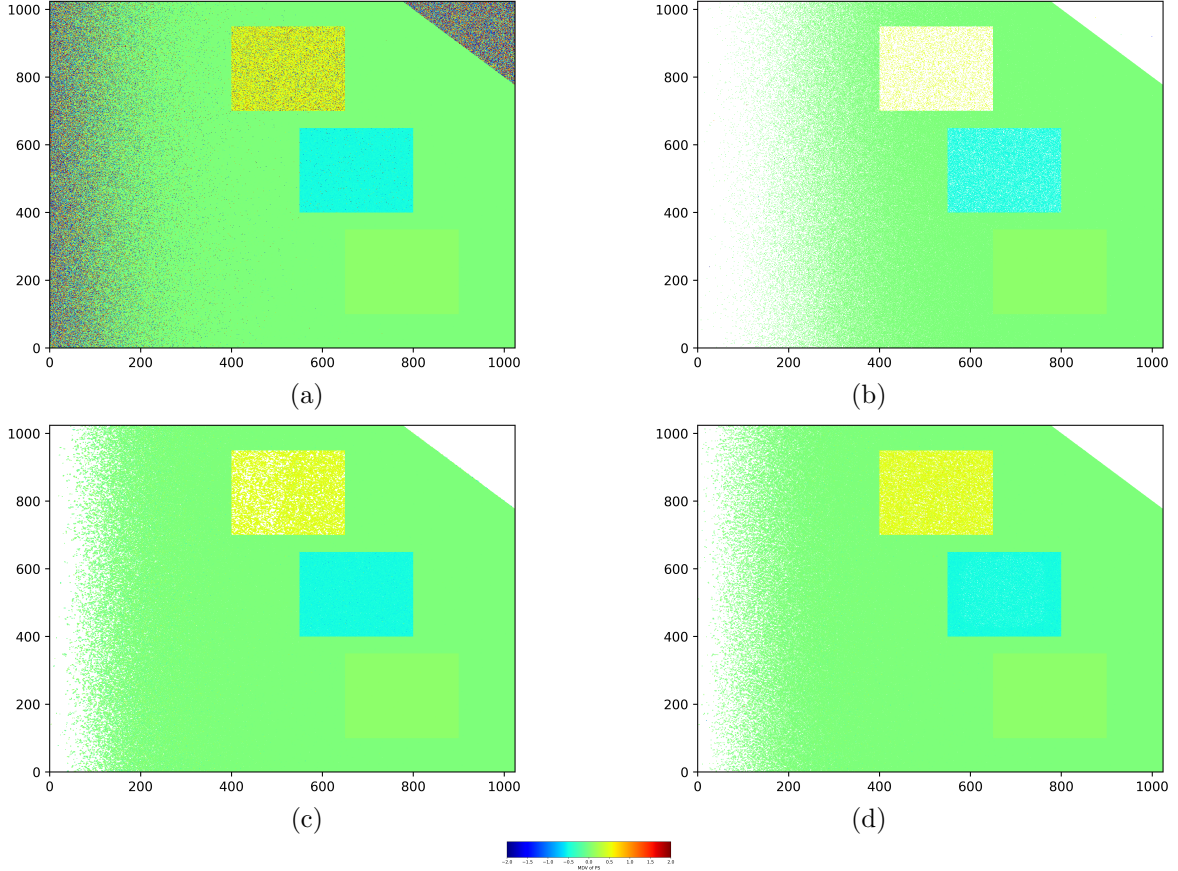


Figure 35: PS results on the Salerno simulated dataset; Subgraphs (a) and (b) shows the PS corresponding DMV provided by the GLRT before and after the detection based on thresholding. Subgraphs (c) and (d) provide the estimated DMV of the PS detected by 3U-Net model trained on Lungro and Salerno real data, respectively.

(b) present the DMV estimated by the GLRT method before and after the thresholding operation, respectively. The detection threshold is set to $T = 0.37$, which is consistent with the configuration used in real data processing, corresponding to a PFA of 2.5×10^{-5} . This setting provides a balanced trade off between detection stability and sensitivity. Fig. 37(c) and (d) illustrate the PS detection results obtained using the 3U-Net models trained on the Lungro and Salerno datasets, respectively. A visual comparison indicates that, consistent with the observations in the Salerno case, the proposed 3U-Net approach significantly outperforms the conventional GLRT detector. It produces a denser and more spatially coherent distribution of PS, while exhibiting improved robustness to noise contamination.

Fig. 38(a) and (b) present the PD and MSE curves, respectively. The quantitative results indicate that the proposed 3U-Net consistently outperforms the conventional GLRT detector in terms of PD, confirming its superior detection capability. Compared with the model trained on the Salerno dataset, the 3U-Net trained on Lungro data achieves a slight improvement in PD and a lower MSE(see Fig. 38), indicating a more favorable trade off between detection performance and parameter estimation accuracy in this simulation scenario. This observation further suggests that domain consistency, i.e. alignment between the statistical properties and scattering characteristics of the training and testing data, plays a crucial role in achieving optimal performance. However, in signal absent(highly decorrelated) regions, the methods exhibit notable differences in false alarm behavior. Specifically, the threshold based GLRT detector produces only two false positives, whereas the 3U-Net models trained on Lungro and Salerno datasets generate 10 and 11 false positives, respectively. This indicates that, although DL based approaches significantly enhance detection completeness, they are more sensitive to noise in low-coherence or decorrelated areas, which may lead to increased spurious detection. Overall, these results highlight the inherent trade off between detection sensitivity and false alarm suppression in DL

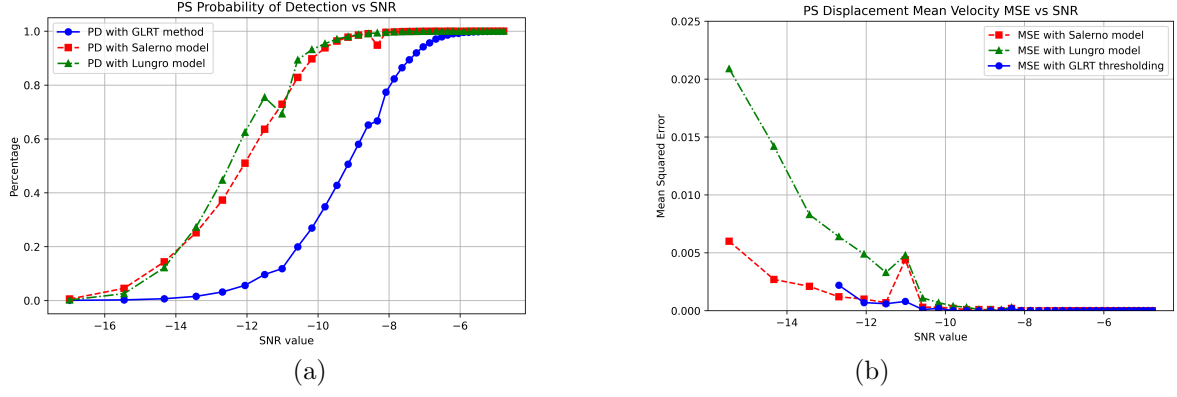


Figure 36: Quantitative analysis of the PS detection for the Salerno simulation dataset achieved with the GLRT method based on thresholding and the 3U-Net models trained by Salerno and Lungro real data, separately. PS Probability of Detection vs the SNR (left); Displacement Mean Velocity vs the SNR (right).

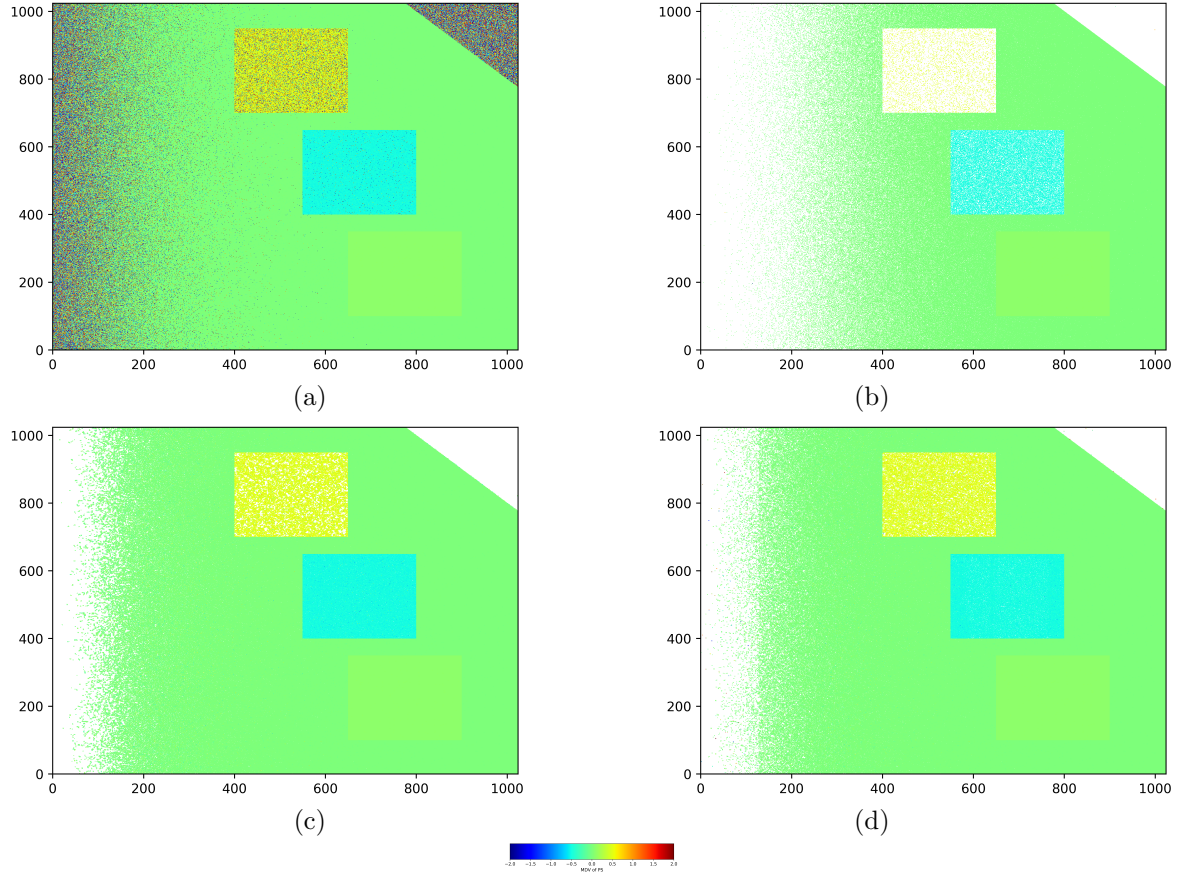


Figure 37: As for Fig. 35 but for the Lungro simulated dataset.

based PS detection. While the proposed method substantially improves PD and maintains competitive estimation accuracy, its behavior in decorrelated regions suggests the need for further refinement.

3.7.4 Chapter Summary

In this chapter, a novel DL based PS detection method has been proposed to improve the performances of classical coherent detectors based on the simple thresholding of a suitable statistic, such as the MIC of and the GLRT. U-Net, widely used in image segmentation, has been selected thank to its ability to

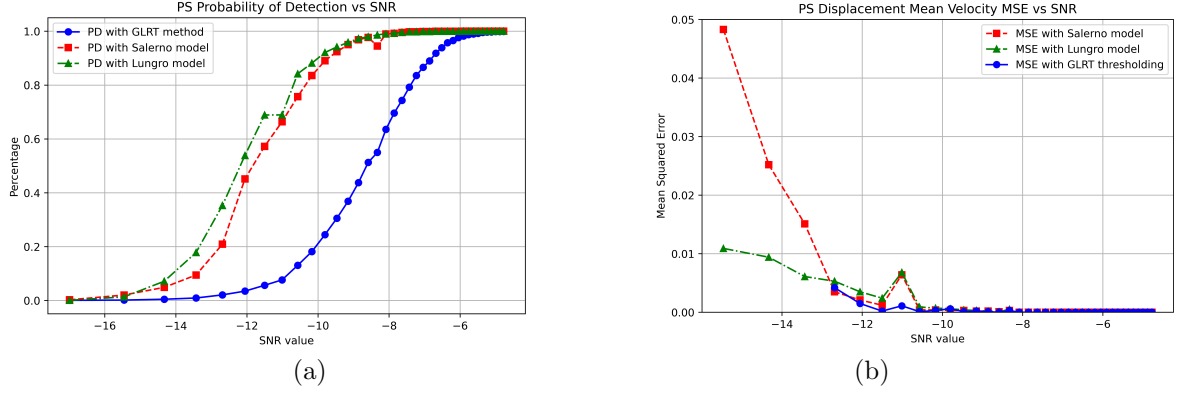


Figure 38: As in Fig. 36 but for the Lungro simulation dataset.

capture image features. Specifically we have considered a 3U-Net able to update the statistic generally used for the PS detection. Proper filtering for isolated and outliers has been integrated in the process starting from the training stage. The network is self-trained by labels generated from the classical GLRT. The results on real data acquired by the sensors of the CSK and CSG constellation clearly show the improved performances of the proposed solution. The improvements of the proposed detector over the classical approach has been also quantified on simulated data.

4 Extension of the proposed DL based PS detector on dual-pol sentinel-1 data

Chapter 2 presents the fundamental principles of SAR. Building upon this foundation, it introduces the basic concepts and theoretical framework of PolSAR, followed by a systematic Polarimetric Interferometric Synthetic Aperture Radar (PolInSAR). In particular, this chapter reviews the polarimetric scattering mechanisms in PolSAR, these theoretical foundations provide essential support for subsequent analysis and applications.

4.1 Fundamentals of PolSAR

1. Scattering Mechanisms of PolSAR

During electromagnetic scattering, radar targets exhibit distinct responses to incident waves with different polarization states, commonly referred to as polarimetric scattering characteristics[LP17]. The polarization state characterizes the orientation of the electric fields vector within the plane perpendicular to the propagation direction. For linearly polarized waves, vertical polarization(V) corresponds to an electric field oscillating in the incidence plane identified by the propagation direction and the normal to the surface, while for horizontal polarization(H) the oscillations are perpendicular to the incidence plane. These two orthogonal linear polarizations form the fundamental polarization basis in PolSAR systems.

In practical SAR observations, different polarization channels are obtained according to the transmit, receive polarization combinations. When the transmit and receive polarizations are identical(e.g. HH or VV), the observation is referred to as co-polarization; when they are orthogonal(e.g. HV or VH), it is termed cross-polarization. As illustrated in Fig.39, these different polarization combinations provide complementary information about the underlying scattering mechanisms of the observed targets. From a physical standpoint, each polarization channel is associated with specific scattering responses.

Broadly speaking, co-polarization channels(HH, VV) are primarily sensitive to surface scattering and double bounce scattering, whereas cross-polarization channels(HV, VH) are dominated by volume scattering, such as multiple scattering within vegetation canopies [LP17]. Owing to the diversity of scattering mechanisms across different land cover types, polarimetric information significantly enhances the discriminative capability for target classification and identification. In practical applications, different polarization modes exhibit distinct sensitivities to geophysical properties. For example, VV polarization is particularly responsive to backscatter variations from vegetation structures and ocean surfaces(e.g. wave dynamics)[LCZ21], while HH polarization is more suitable for applications such as soil moisture retrieval, ship detection, and water-ice boundary delineation[DKKP⁺21, GDR16, LDD21]. Cross-polarization channels(HV, VH), being sensitive to volume scattering, are especially effective for distinguishing vegetated from non-vegetated areas, delineating water bodies, and characterizing surface toughness[SGN⁺20, Eld22].

Therefore, the joint utilization of multiple polarization channels(e.g. dual-pol or full-pol configurations) enables a more comprehensive characterization of target scattering behavior. This multi-

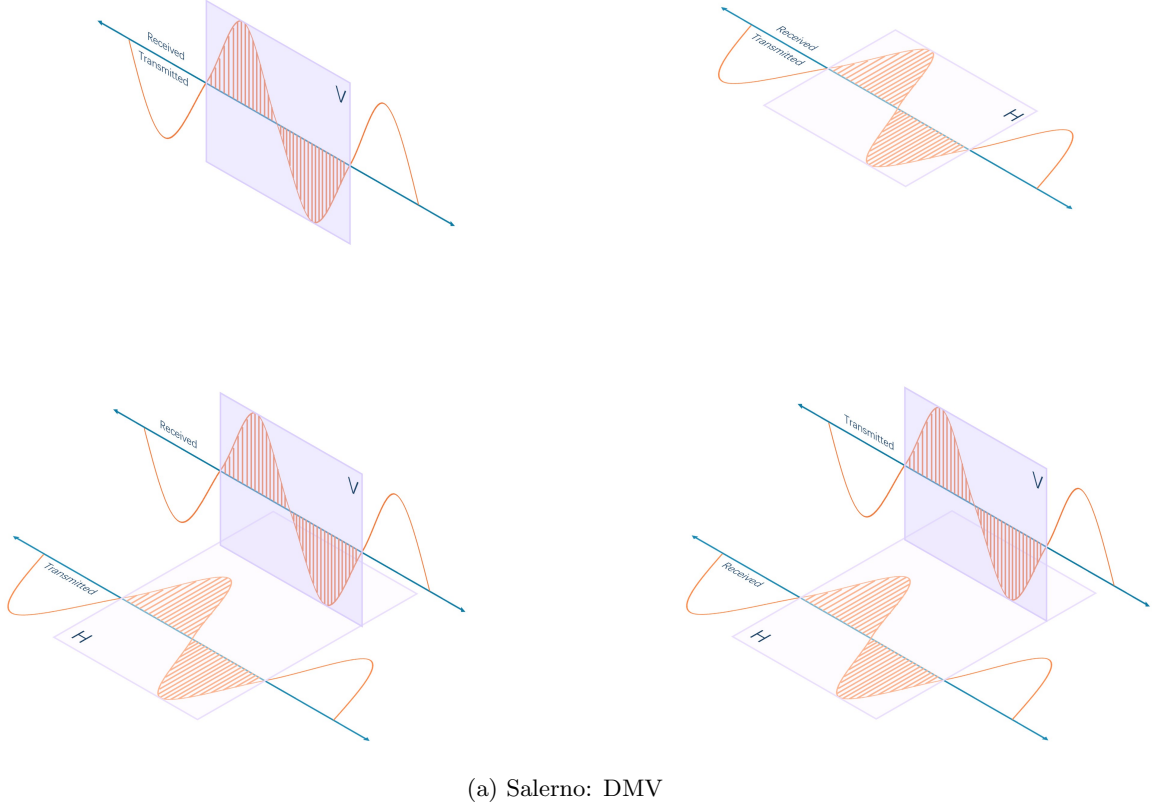


Figure 39: Schematic of Different Polarization States

dimensional representation provides physically meaningful information that supports a wide range of remote sensing applications, including land cover classification, geophysical parameter inversion, and change detection.

The sensitivity of electromagnetic waves to target properties, including dielectric characteristics, physical attributes, geometric structure, and orientation, varies with the polarization state. Consequently, the incorporation of polarimetric information significantly enhances the capability of imaging radar systems to characterize scattering behavior, thereby improving the accuracy and reliability of target classification and geophysical parameter retrieval. In PolSAR systems, this polarization dependent behavior is described through the target's polarimetric scattering response. In full-polarimetric acquisitions, the responses corresponding to all transmit, receive polarization combinations are recorded, enabling a comprehensive representation of polarization conversion effects and the underlying scattering mechanisms. This behavior is conventionally modeled using the Sinclair Scattering Matrix[LP17], which is a complex values 2×2 matrix:

$$\mathbf{S} = \begin{bmatrix} S_{HH} & S_{HV} \\ S_{VH} & S_{VV} \end{bmatrix} \quad (4.1)$$

where $\mathbf{S}_{pq}(p, q \in H, V)$ denotes the complex scattering coefficient corresponding to transmit polarization p and receive polarization q , containing both amplitude and phase information.

Based on the previously introduced Sinclair scattering matrix $\mathbf{S}$, a vectorized representation is commonly adopted to facilitate the analysis and processing of polarimetric scattering information. In the Lexicographic representation, the scattering matrix is expanded over a set of orthogonal basis matrices ψ_L :

$$\psi_L = \left\{ \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} \right\} \quad (4.2)$$

Under the assumption of monostatic radar, the reciprocity condition $\mathbf{S}_{HV} = \mathbf{S}_{VH}$ holds, imply-

ing that the scattering matrix contains only three independent components. Accordingly, $\mathbf{S}$ can be vectorized into a three dimensional complex vector, referred to as the Lexicographic scattering vector $\mathbf{K}_L$:

$$\mathbf{K}_L = [S_{HH} \quad \sqrt{2}S_{HV} \quad S_{VV}]^T \quad (4.3)$$

The normalization factor $\sqrt{2}$ is introduced to ensure energy consistency among different polarization channels, thus preserving the total scattering power in the vector representation. The Lexicographic scattering vector provides a compact and standardized vector form of the Sinclair scattering matrix, retaining the essential polarimetric information.

While the Lexicographic representation provides a direct vectorization of the Sinclair scattering matrix based on physical polarization channels, its components are not fully decorrelated and do not explicitly correspond to canonical scattering mechanisms. To address this limitation, an alternative orthogonal basis, referred to as the Pauli basis ψ_p , is introduced to achieve a more physically interpretable representation.

In the Pauli representation, the scattering matrix $\mathbf{S}$ is expanded to the following set of orthogonal basis matrices:

$$\psi_p = \left\{ \sqrt{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \sqrt{2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \sqrt{2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \sqrt{2} \begin{bmatrix} 0 & -j \\ j & 0 \end{bmatrix} \right\} \quad (4.4)$$

Under the monostatic assumption with the reciprocity condition $\mathbf{S}_{HV} = \mathbf{S}_{VH}$, the scattering matrix can be vectorized into the Pauli scattering vector $\mathbf{K}_p$:

$$\mathbf{k}_p = \left[\frac{S_{HH}+S_{VV}}{\sqrt{2}} \quad \frac{S_{HH}-S_{VV}}{\sqrt{2}} \quad \sqrt{2}S_{HV} \right]^T \quad (4.5)$$

Compared with the Lexicographic scattering vector $\mathbf{K}_L$, the Pauli representation provides a more structured decomposition of the scattering process.

Based on the previously defined scattering vector representations, further statistical characterization of polarimetric SAR data can be achieved by constructing second order matrices. Specifically, by forming the outer product between the scattering vector and its conjugate transpose, and performing ensemble averaging, one obtains the polarimetric covariance matrix and the coherency matrix.

In the Lexicographic representation, the covariance matrix $\mathbf{C}_3$ is defined as:

$$\mathbf{C}_3 = \langle \mathbf{k}_L \mathbf{k}_L^H \rangle = \begin{bmatrix} \langle |S_{HH}|^2 \rangle & \sqrt{2} \langle |S_{HH} S_{HV}^*|^2 \rangle & \langle |S_{HH} S_{VV}^*|^2 \rangle \\ \sqrt{2} \langle |S_{HV} S_{HH}^*|^2 \rangle & 2 \langle |S_{HV}|^2 \rangle & \sqrt{2} \langle |S_{HV} S_{VV}^*|^2 \rangle \\ \langle |S_{VV} S_{HH}^*|^2 \rangle & \sqrt{2} \langle |S_{VV} S_{HV}^*|^2 \rangle & \langle |S_{VV}|^2 \rangle \end{bmatrix} \quad (4.6)$$

where $(\cdot)^H$ denotes the Hermitian transpose and $\langle \cdot \rangle$ represents ensemble or multi-look averaging. The matrix $\mathbf{C}_3$ describes the second order statistical properties and correlation structure among different polarization channels in the Lexicographic basis.

Similarly, in the Pauli representation, the corresponding coherency matrix $\mathbf{T}_3$ is defined as:

$$\mathbf{T}_3 = \langle \mathbf{k}_p \mathbf{k}_p^H \rangle = \frac{1}{2} \begin{bmatrix} \langle |S_{HH} + S_{VV}|^2 \rangle & \langle (S_{HH} + S_{VV})(S_{HH} - S_{VV})^* \rangle & 2 \langle (S_{HH} + S_{VV}) S_{HV}^* \rangle \\ \langle (S_{HH} - S_{VV})(S_{HH} + S_{VV})^* \rangle & \langle |S_{HH} - S_{VV}|^2 \rangle & 2 \langle (S_{HH} - S_{VV}) S_{HV}^* \rangle \\ 2 \langle S_{HV} (S_{HH} + S_{VV})^* \rangle & 2 \langle S_{HV} (S_{HH} - S_{VV})^* \rangle & 4 \langle |S_{HV}|^2 \rangle \end{bmatrix} \quad (4.7)$$

where $\mathbf{K}_p$ is the Pauli scattering vector. Compared with the covariance matrix $\mathbf{C}_3$ the coherency matrix $\mathbf{T}_3$ provides a more physically meaningful representation, as its components are directly related to fundamental scattering mechanisms.

These second order matrices serve as the foundation for polarimetric decomposition techniques and statistical analysis, enabling the extraction of physically interpretable features from PolSAR data.

In this study, deformation estimation is performed independently on different polarimetric channels (VV and VH) within the conventional InSAR framework. No modifications are introduced to the

standard interferometric model. Instead, the approach exploits the differences in scattering responses across polarization channels to derive deformation estimates under distinct polarimetric conditions. Since different polarizations exhibit varying sensitivities to scattering mechanisms, this strategy provides complementary insights into surface deformation characteristics.

4.2 Experimental results

4.2.1 Overview of the Sentinel-1A Satellite System and Data

This thesis employs SAR data acquired by the Sentinel-1A satellite, which is part of the Copernicus program operated by the European Space Agency (ESA). Sentinel-1A operates in the C-band, with a central frequency of approximately 5.405 GHz, and provides all-weather, day-and-night earth observation capabilities. The satellite was launched in 2014 and operates in a near-polar sun-synchronous orbit at an altitude of approximately 693 km, with an inclination of about 98.18. The cycle of re-visiting a single satellite is 12 days. Under constellation operation, a higher temporal sampling rate can be achieved, providing stable support for multi-temporal observations.

In terms of image geometry, Sentinel-1A adopts a side-looking SAR configuration. High azimuth resolution is achieved through synthetic aperture processing based on platform motion. The system transmits linearly frequency-modulated (chirp) signals and performs range compression and Doppler processing to form SAR images, resulting in complex-valued data containing both amplitude and phase information. This data representation fully characterizes the electromagnetic scattering response of ground targets and serves as the fundamental data type for interferometric analysis and related applications.

Regarding acquisition modes, Sentinel-1A supports multiple imaging modes, including Interferometric Wide Swath (IW), Extra Wide Swath (EW), and Stripmap. The IWIW mode, which is the most commonly used, is based on the Terrain Observation with Progressive Scans SAR (TOPS) technique. It achieves a balance between wide swath coverage and spatial resolution through periodic azimuth beam steering. The typical spatial resolution is approximately $5 \text{ m} \times 20 \text{ m}$ (azimuth $\times$ range), with a swath width of about 250 km. The EW mode is designed for larger-area coverage, whereas the Stripmap mode provides higher spatial resolution for detailed local observations. In terms of polarization, the system supports single-polarization (HH or VV) and dual-polarization (VV+VH or HH+HV) acquisition modes, enabling the characterization of scattering behavior under different polarization configurations.

With respect to data products, Sentinel-1A provides standardized multi-level products, primarily including SLC and Ground Range Detected (GRD) products. The SLC product preserves the full complex signal, including both amplitude and phase, and is expressed in slant-range geometry, making it the fundamental data type for interferometric SAR and time-series analysis. The GRD product is generated through multi-looking and geometric projection, producing amplitude-only imagery suitable for general observation and feature extraction. In this thesis, SLC products are adopted as the primary data source to ensure the integrity of phase information and the accuracy of subsequent analyses.

Sentinel-1A data are characterized by a standardized product structure and acquisition strategy. The consistency in imaging modes, polarization configurations, and data formats ensures a high level of uniformity, providing reliable and consistent data support for multi-temporal SAR analysis.

A detailed description of the system parameters and acquisition modes of the Sentinel-1A is provided in the following tables.

4.2.2 Experiments on Dual-pol SAR Data

This thesis selected two areas for investigation: Val d’Agri area in southern Italy and Moncenisio area in northern Italy. These areas were chosen because they are representative of regions in Italy that are susceptible to geological hazards.

The Val d’Agri region is located in the southern Italian region of Basilicata. The center of this typical intermontane basin, which lies within the Southern Apennines, generally range from 500m to 1500m, as shown in Fig.40(a). The terrain consists of plains, hills, and mountains, exhibiting moderate to significant topographic variation. The landforms were primarily shaped by tectonic movements and fluvial erosion, resulting in an elongated basin structure and a well developed river network. Vegetation cover within the region is relatively continuous and dominated by farmland, forests and shrubs. Overall,

Table 7: Sentinel-1 acquisition modes.

Parameter	SM	IW	EW
Resolution (m)	$\sim 5 \times 5$	$\sim 5 \times 20$	$\sim 20 \times 40$
Scene Size (km)	$\sim 80 \times 80$	$\sim 250 \times 250$	$\sim 400 \times 400$
Polarisation	HH, VV / HH+HV	VV / VV+VH	HH, VV / HH+HV
Duration (s)	~ 25	~ 25	~ 25
Number of Looks	1×1	5×1	10×2
multi-look Resolution (m)	$\sim 5 \times 5$	$\sim 20 \times 22$	$\sim 40 \times 40$
Geolocation Accuracy (m)	~ 7	~ 7	~ 10

Table 8: Sentinel-1 system and orbital parameters.

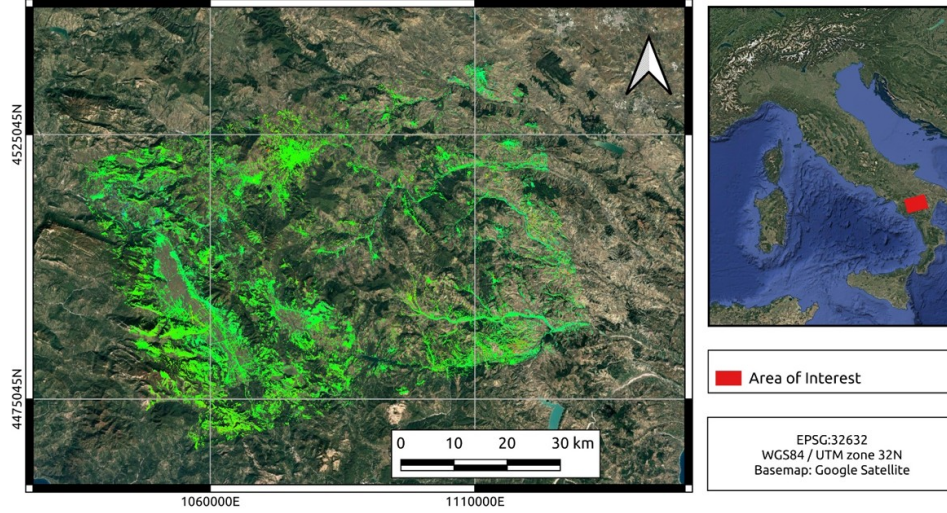
Parameter	Value
Radar band	C-band (5.405 GHz)
Wavelength	~ 5.6 cm
Orbit type	Near-polar, sun-synchronous
Altitude	~ 693 km
Inclination	$\sim 98.18^\circ$
Repeat cycle	12 days (Sentinel-1A)
Orbit accuracy	< 5 cm
Incidence angle (IW)	$\sim 29^\circ - 46^\circ$
Swath width (IW)	~ 250 km

it is moderately high. Geologically, the region lies within a fold-and-fault zone of the Apennine orogenic belt and consists primarily of carbonate and clastic sedimentary rocks. The region features a complex tectonic pattern, a well developed fault system, and relatively active crustal activity. Against this backdrop, the region is prone to various geological hazards, including earthquakes, landslides, and ground subsidence. These hazards are exacerbated by heavy rainfall, human activities such as oil and gas extraction, and long-term tectonic evolution.

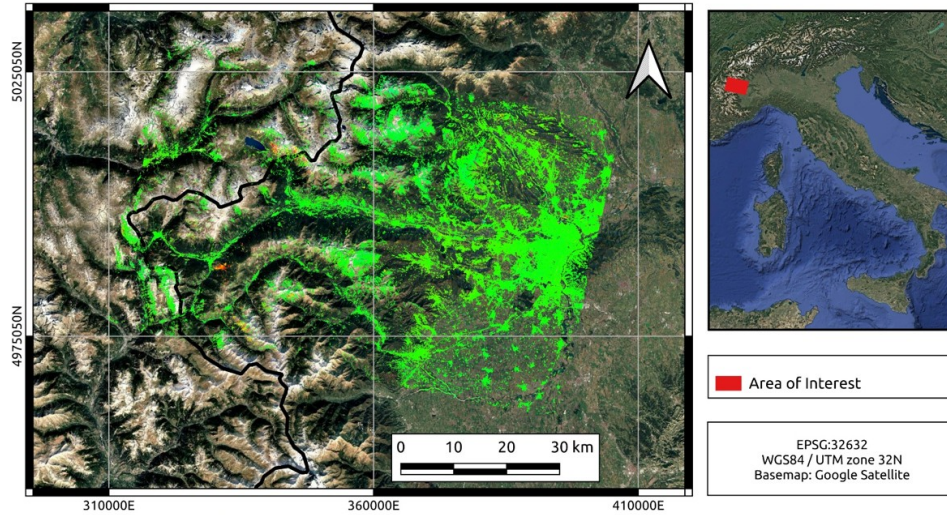
The Moncenisio region is situated in the western Alps along the Italy-France border, as shown in Fig.40(b). The area is characterized by high relief mountainous terrain, with elevations generally exceeding 1500m and surrounding peaks rising above 3000m. The topography is highly rugged, featuring steep slopes and pronounced elevation gradients. Glacial processes have extensively shaped the landscape, giving rise to characteristic geomorphological features such as the Moncenisio plateau and associated reservoir systems. Vegetation exhibits a clear altitudinal zonation: forested areas and alpine meadows dominate at lower to mid elevations, whereas higher elevations are primarily composed of bare rock and scree, resulting in overall sparse and spatially heterogeneous vegetation cover. Geologically, the region forms part of the Alpine orogenic belt and is predominantly composed of metamorphic and sedimentary formations, with a complex tectonic framework and well developed fault systems, indicate of ongoing geodynamic activity. The interplay between steep topography and complex geological conditions leads to reduced slope stability, making the area highly susceptible to geohazards such as landslides, rockfalls, debris flows, and snow avalanches. These processes are particularly intensified under trigger factors such as intense precipitation, snowfall, and freeze thaw cycles.

1. Dataset

This thesis utilizes C-band SAR data acquired by Sentinel-1A. The first dataset consist of 101 ascending scenes of resampled Sentinel-1A imagery (VV and VH channel, respectively), collected between June 2022 and July 2025, with an effective spatial resolution of 2.33×13.93 m (in azimuth and range directions, respectively). This dataset covers the entire southern Val d’Agri study area. The perpendicular baseline spans approximately 700 m, and each SAR image has dimensions of 22961×4186 pixels in the azimuth and range directions, respectively. Detailed acquisition parameters are provide in Table.9



(a) Val d'Agri



(b) Moncenisio

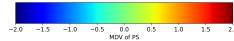


Figure 40: Study areas: Val d'Agri and Moncenisio areas in Italy. In both images, standard high-resolution DMV product provided by GLRT overlaid to Google Map images.

The second dataset comprises 112 descending Sentinel-1A scenes acquired over the Moncenisio region between November 2022 and November 2025, with the same spatial resolution of 2.33×13.93 m (VV and VH channel, respectively). The maximum perpendicular baseline is approximately 800 m, and each image consists of 22935×4192 pixels in the azimuth and range directions, respectively. Further details are summarized in Table.10

The two datasets correspond to geohazard prone regions in southern and northern Italy, respectively, characterized by markedly different topographic and scattering conditions. The use of such heterogeneous datasets provides a robust basis for evaluating the generalization capability of the proposed model. During preprocessing, all datasets undergo a phase calibration and time-series reconstruction processing carried out with the classical SBAS approach implemented at low-resolution. The processing pipeline follows the standard multi-temporal coherent interferometric phases are modeled and corrected prior to GLRT based detection. Specifically, the workflow includes the estimation and removal of the APS, as well as the compensation of low-spatial resolution linear and nonlinear de-

Table 9: SAR data description for Val d’Agri area.

Area	Val d’Agri, Basilicata, South Italy
SAR Sensor	Sentinel-1A
Acquisition Mode	TOPSAR-Interferometric Wide Swath(IW)
Spatial Resolution	$\sim 2.33 \times 13.93$ m
Acquisition Interval	June 2022 – July 2025
# of Images	101
Incidence Angle	$\sim 34^\circ$
Orbit Direction	Ascending

Table 10: SAR data description for Moncenisio area.

Area	Moncenisio, Turin, North Italy
SAR Sensor	Sentinel-1A
Acquisition Mode	TOPSAR-Interferometric Wide Swath(IW)
Spatial Resolution	$\sim 2.43 \times 13.93$ m
Acquisition Interval	November 2022 – November 2025
# of Images	112
Incidence Angle	$\sim 34^\circ$
Orbit Direction	Descending

formation components. This ensures the generation of physically consistent and reliable inputs for subsequent PS detection based on the high-resolution temporal coherence analysis implemented in our framework with GLRT, see in Fig. 40.

2. Implementation Aspects

Both datasets were further processed using a sliding window based patch extraction strategy, where each patch has a spatial size of 128×128 pixels. The samples were randomly partitioned into training, validation, and test sets, accounting for 65%, 15%, and 20% of the total data, respectively. All experiments were conducted on a single NVIDIA GeForce GTX 1080 Ti with 2×12 GB RAM, and implemented using the PyTorch framework.

During training, model parameters were optimized using the Adam optimizer with an initial learning rate of 0.001. To enhance convergence stability and training efficiency, a Polynomial Learning Rate Decay(Poly schedule) was adopted. At the same time, an early stopping strategy was employed, whereby the training process was terminated once the loss function converged and exhibited on further significant improvement in 10 epochs.

3. Results on Val d’Agri

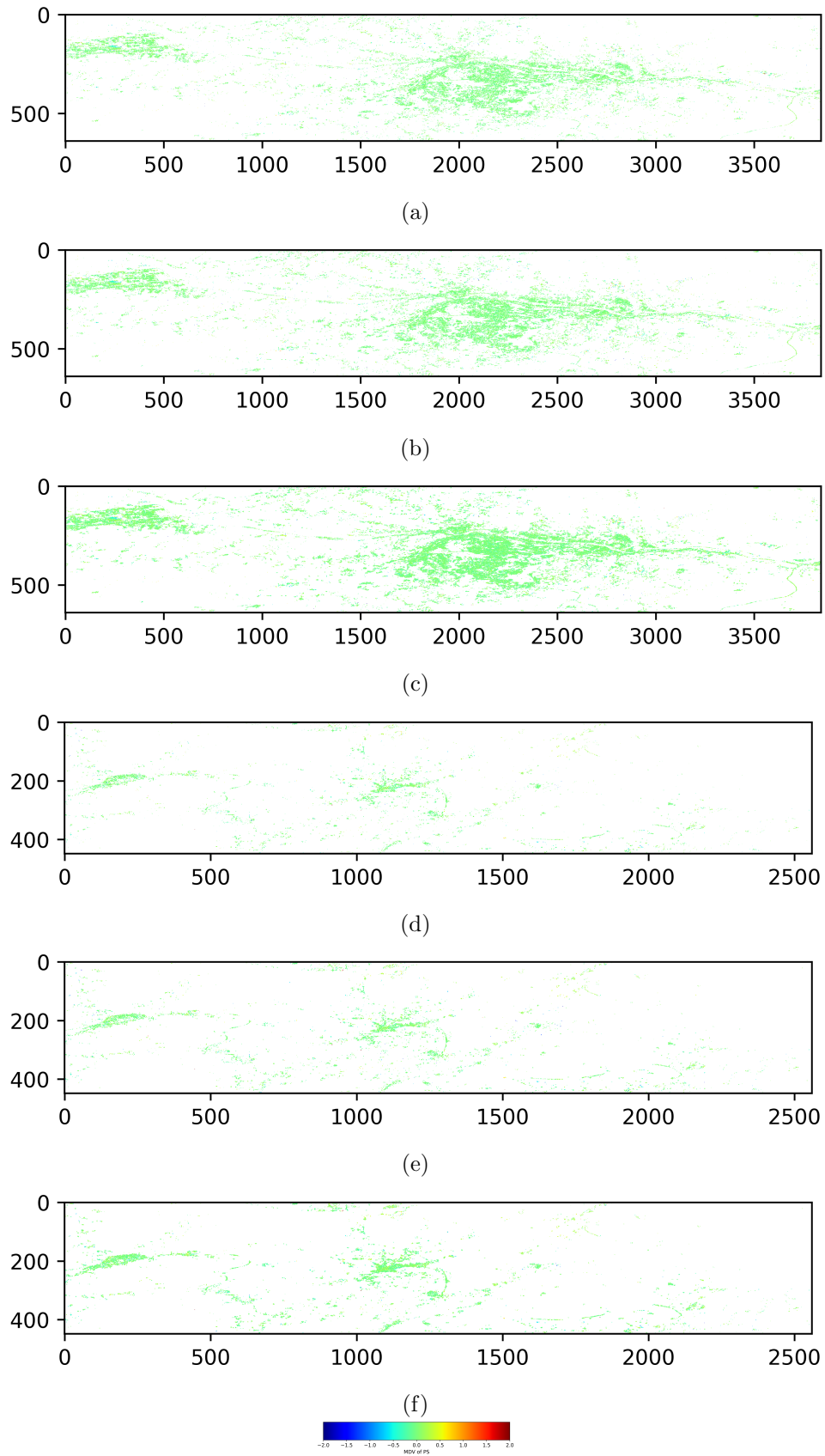


Figure 41: PS detection in two areas (top and bottom rows) of Val d'Agri using VV-channel data. The left column shows the results obtained using the GLRT detector, the middle column shows the results obtained using the GLRT with IOF detector, and the right column shows the results obtained using the 3U-Net detector.

In this experiment, the VV and VH channels were processed independently. Following an analysis of the study area, four representative Regions of Interest (ROIs), accounting for approximately 10% of the total dataset, were selected as the test set. From the remaining data, 80% was allocated for training and 10% for validation.

Figure ?? presents the PS detection results for the VV channel, with two representative test regions illustrated. Subfigures (a) and (d) show the PS detection results for test regions 1 and 2, respectively, obtained using the GLRT method with a detection threshold of 0.5. Subfigures (b) and (e) present the corresponding results after applying the IOF filtering module with a reduced threshold of 0.47. The subfigures (c) and (f) display the PS detection 3U-Net for the same regions. From a visual perspective, the PS density increases progressively from the GLRT method with a high threshold, to the GLRT method with a relaxed threshold combined with IOF filtering, and finally to the proposed 3U-Net approach. Notably, the 3U-Net results exhibit a clear improvement in segmentation accuracy while maintaining effective suppression of spurious detections, without introducing additional noise points. Quantitative comparisons are provided in Table.11, which reports the number of detected PS points for each method. In test one, the GLRT method identified 186,236 PS points, while the GLRT combined with IOF filtering detected 201,850 PS. In contrast, the proposed 3U-Net detected 273,981 PS. This corresponds to relative increases of 47.1% and 35.7% compared to the GLRT and GLRT combine IOF module approaches, respectively, demonstrating the superior detection capability of the proposed method. In test two, the GLRT method identified 24287 PS, while the GLRT combined with IOF filtering detected 27063 PS. In contrast, the proposed 3U-Net detected 35948 PS. This corresponds to relative increases of 48% and 32.8% compared to the GLRT and GLRT combine IOF module approaches, respectively, demonstrating the superior detection capability of the proposed method.

Table 11: The number of PS detection for Val d’Agri data with VV channel.

Method	Test one	Test two
GLRT	186,236	24287
GLRT with IOF filtering	201,850	27063
3U-Net	273,981	35948

In this experiment, the model was trained using a joint dataset comprising both VV and VH channels. Accordingly, in addition to the VV results, the detection performance on the VH channel is also of interest. Fig. 42 presents the PS detection results for the VH channel in test regions one and two. Subfigures (a) and (d) show the PS detection results obtained using the GLRT method with a high threshold; subfigures (b) and (e) present the results from the GLRT combined with IOF filtering module under a relaxed threshold; and subfigures (c) and (f) illustrate the detection results produced by the proposed 3U-Net approach. Notably, the proposed method achieves the highest detection density while maintaining effective suppression of spurious responses. Table.12 summarizes the quantitative comparison of PS counts across different methods. In test region one, the high threshold GLRT detector identified 134140 PS, the GLRT combined with IOF filtering detected 146791 PS, and the 3U-Net detected 208627 PS. This corresponds to relative increases of 55.5% and 42.1% compared to the GLRT and GLRT combine IOF module approaches, respectively. In test region two, the corresponding numbers are 14550, 16238, and 22645 PS, respectively. This corresponds to relative increases of 55.6% and 39.5% compared to the GLRT and GLRT combine IOF module approaches. These results indicate that the proposed model maintains consistent detection capability across different polarimetric channels, while effectively balancing detection sensitivity and noise suppression.

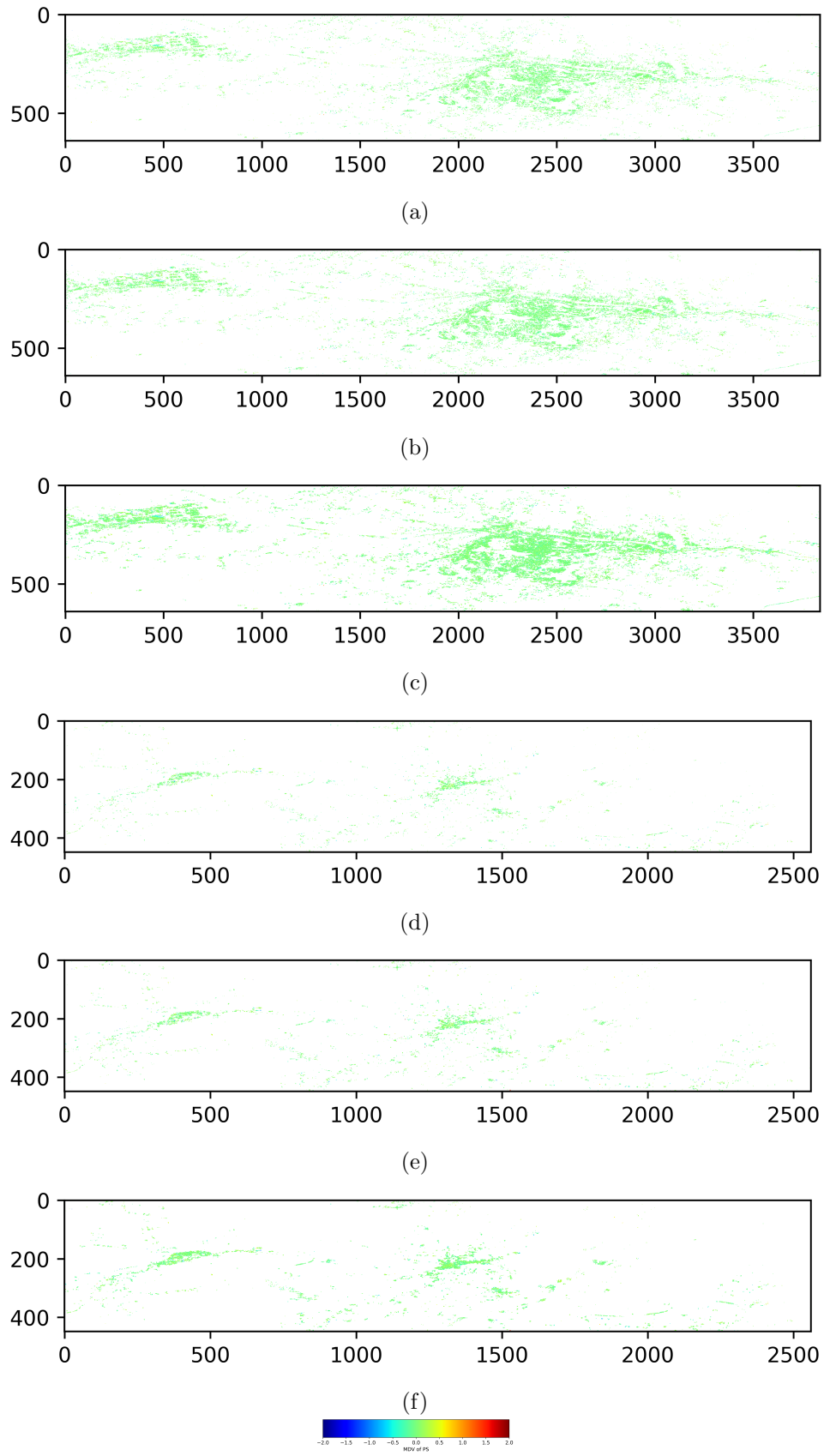


Figure 42: PS detection of Val d'Agri data with VH channel. The left column is for the GLRT detector, the middle column is for the GLRT with IOF detector and the right column is for the 3U-Net detector.

Table 12: The number of PS detection for Val d’Agri data with VH channel.

Method	Test one	Test two
GLRT	134140	14550
GLRT with IOF filtering	146791	16238
3U-Net	208627	22645

As demonstrated by the above experimental results, the number of PS detected in the VV channel is significantly higher than that in the VH channel across all methods, including the conventional GLRT approach, the GLRT combined with the IOF filtering strategy, and the proposed DL based method. This observation is consistent with the underlying scattering mechanisms, as co-polarized channels (VV) are generally more sensitive to stable surface and double bounce scattering, whereas cross-polarized channels (VH) are dominated by volume scattering and are more susceptible to decorrelation effects. Notably, despite the lower absolute number of detected PS in the VH channel, the DL based approach exhibits a higher relative improvement compared to conventional methods in both test area one and two. The improvement in the PS detection of 3U-Net over GLRT is always above the 30% with a peak of 55.5% achieved on the VH channel. This indicates that the proposed 3U-Net is particularly effective at recovering potential PS in challenging conditions, highlighting its robustness and superior capability in handling low-coherence and complex scattering environments.

Next this thesis crosses compare with more details the results of VV and VH channels. Fig.?? presents both the VH and VV channel results in the left and right parts, respectively. The subfigures (a) and (c), in the top row show the PS detection results for the VH and VV channels, respectively, obtained using the GLRT method with a detection threshold of 0.5. The subfigures (d) and (f) illustrate the PS distributions for the VV and VH channels, respectively, generated by the proposed 3U-Net. It can be observed that the 3U-Net method produces a denser distribution of PS points in both VV and VH channels while effectively suppressing spurious detections, without introducing additional noise. For a more detailed comparison, a local region is magnified and centered in Fig.?? (b) and (e) for GLRT and 3U-Net, respectively. Again, VH channel is shown on the left and the VV channel on the right. From Subfigure (b), it can be observed that, under the GLRT framework, the VV channel generally yields a higher PS detection density than the VH channel across most regions, which is consistent with the dominant scattering mechanisms present in this area. However, within the highlighted red box regions, the PS detection performance of the two channels becomes comparable. In contrast, Subfigure (e) shows that the 3U-Net method maintains the overall trend of higher PS density in the VV channel, while exhibiting a notable enhancement in the VH channel within the highlighted regions. Specifically, the VH channel on the left part of Fig.?? (e) shows local accumulations of PS that are not present in the right image corresponding to the VV channel. This observation highlights the potential of the VH channel for PS extraction when appropriate feature learning and spatial modeling are employed. It also indicates that the proposed 3U-Net is capable of effectively exploiting complementary scattering information across polarimetric channels, thus overcoming the limitations of conventional statistical detectors in low-coherence or complex scattering environment.

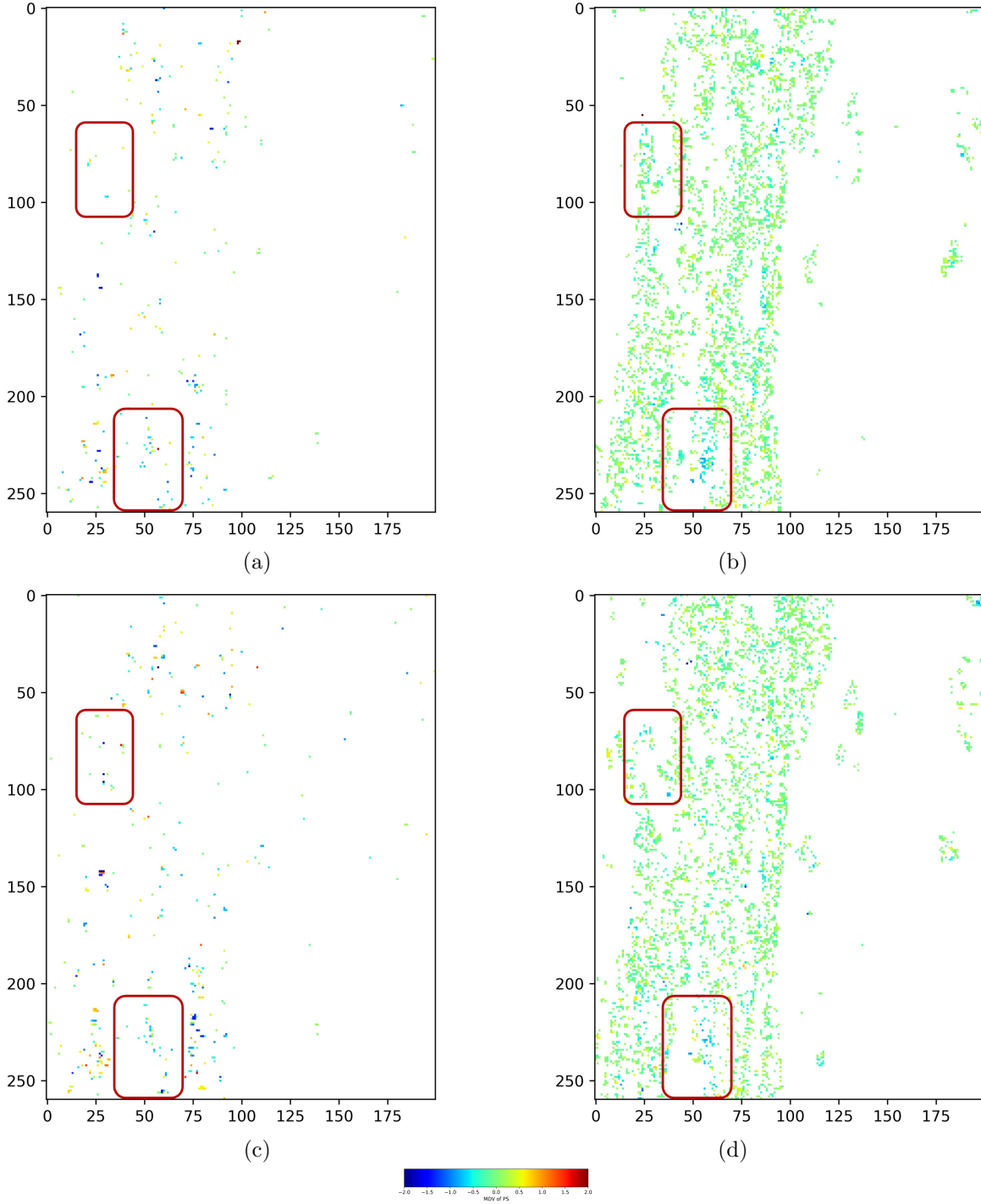


Figure 44: Differences in PS Point Distribution Between GLRT and 3U-Net Methods Across VV and VH Channels. (a) and (c) show the PS points that can be detected using only the GLRT method for the VH and VV channels, respectively. (b) and (d) show the PS points that can be detected using only the 3U-Net method for the VH and VV channels, respectively.

Figure. 44 presents a specific reciprocal analysis of PS detection capabilities between the GLRT and 3U-Net methods, that is PD detected by one detection scheme over the other. The subfigures (a) and (c) illustrate PS detected exclusively by the GLRT methods but missed by 3U-Net, corresponding to the VH and VV channels, respectively. Conversely, subfigures (b) and (d) depict PS identified solely by the 3U-Net method in the VH and VV channels. As observed, PS points uniquely detected by the GLRT method exhibit a spatially sparse and discontinuous distribution, often appearing as

isolated “salt-and-pepper” patterns, with corresponding velocity estimates lacking spatial coherence. In contrast, PS points exclusively detected by the 3U-Net method demonstrate a higher degree of spatial connectivity and continuity, with smoother and more physically consistent deformation (DMV) fields, as shown in subfigures (b) and (d). Moreover, PS points identified by 3U-Net in the VH channel exhibit a more clustered and structured spatial distribution compared to those in the VV channel. Taking the region highlighted by the red box as an example, this area has previously been identified as a representative region where the VH channel contributes significantly under GLRT-based detection. Quantitative comparison further reveals that, as shown in Table.13, in the VH channel, only 292 PS are uniquely detected by GLRT, whereas 4479 PS are uniquely identified by 3U-Net, representing a 14 times increase. Similarly, in the VV channel, 389 PS are detected exclusively by GLRT, while 4167 are identified solely by 3U-Net, corresponding to a 9 times increase. These results further highlight the capability of the proposed 3U-Net framework to recover spatially consistent scatterers that are missed by traditional statistical detectors, particularly in low-coherence scenarios such as the VH channel, thereby demonstrating its effectiveness in enhancing PS detection completeness.

Table 13: The number of difference PS detection for Val d’Agri data with GLRT and 3U-Net for VH and VV channela.

Method	VH channel	VV channel
GLRT only	292	389
3U-Net only	4479	4267

3. Results on Moncenisio

In addition to validating the potential of the proposed model 3U-Net in the southern study area Val d’Agri, this experiment will test the network’s ability to generalize by applying the model trained in the southern study area to the northern study area of Moncenisio.

Figure45 presents the cross-dataset evaluation results, where the model is trained on dataset Val d’Agri and directly applied to dataset Moncenisio. The subfigures (a) and (d) show the PS detection results for the VH and VV polarization channels, respectively, obtained using the GLRT detector. The subfigures (b) and (e) present the corresponding PS detection results derived from the proposed 3U-Net model for the VH and VV channels. The subfigures (c) and (f) further illustrate the PS candidates identified by 3U-Net only in the VH or VV channels, respectively. Qualitative analysis indicates that, despite the domain shift between dataset Val d’Agri and dataset Moncenisio, the 3U-Net model maintains strong generalization capability and achieves a substantially higher PS density in both polarization channels compared to the GLRT-based approach. Furthermore, although the VV channel yields a higher number of detected PS points overall, the VH channel still contributes a significant proportion of complementary PS detections, highlighting its importance in capturing scattering mechanisms in low-coherence environments. As reported in Table.14, the GLRT method detects 3017976 and 5160464 PS points in the VH and VV channels, respectively, whereas the 3U-Net model identifies 4447085 and 7140394 PS points. This corresponds to relative improvements of 47.4% and 38.4% in PS detection for the VH and VV channels, respectively, compared with GLRT. These results demonstrate that the proposed 3U-Net framework provides enhanced detection capability, particularly under low-coherence conditions, with notable gains observed in the VH polarization channel.

Table 14: The PS detection for Moncenisio data with GLRT and 3U-Net for VH and VV channela.

Method	GLRT Detector	3U-Net detector
VH channel	3017976	4447085
VV channel	5160464	7140394

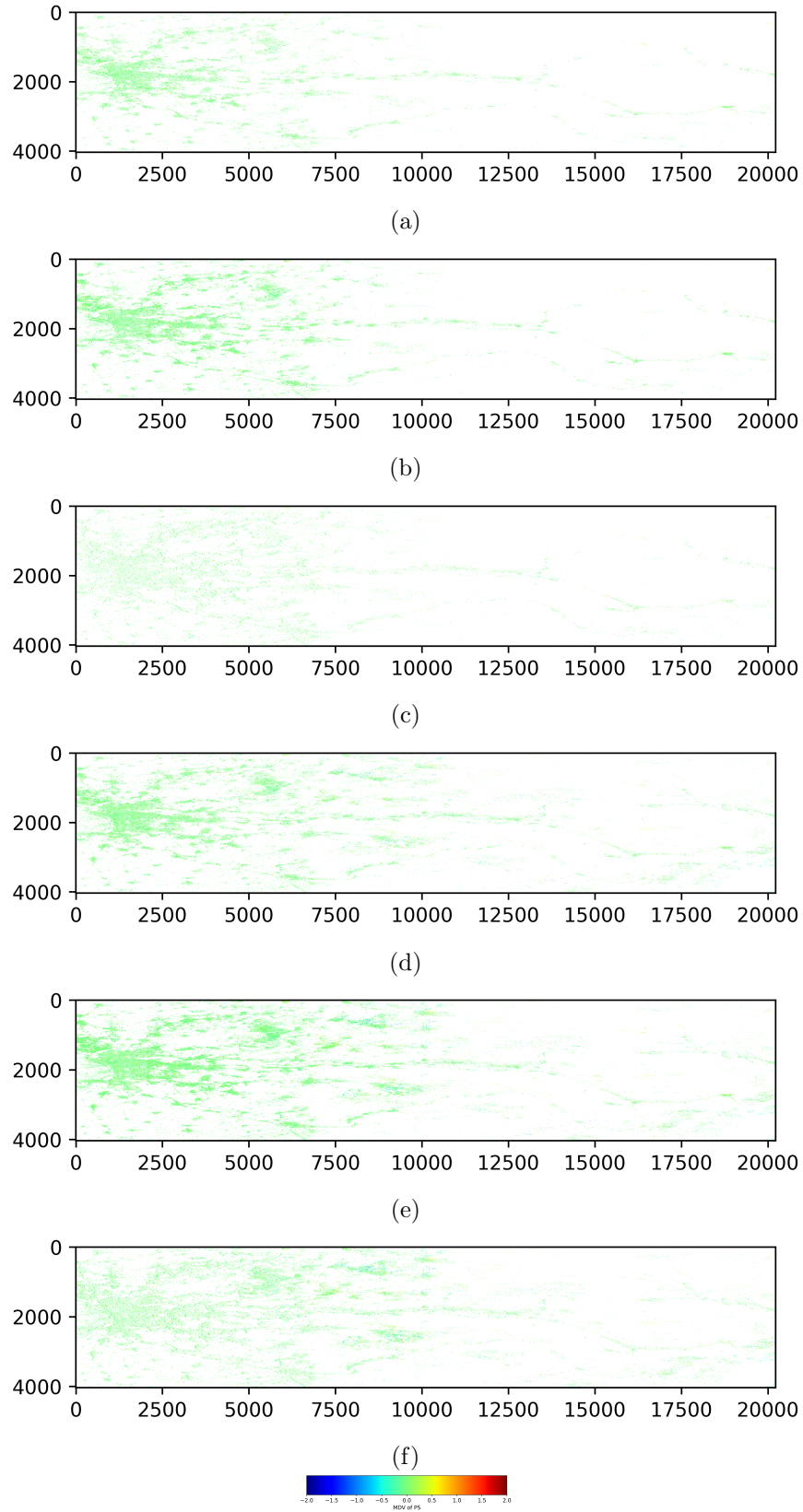


Figure 45: PS detection results for the Moncenisio dataset using a model trained on the Val d’Agri dataset for VV and VH channels. (a) GLRT-based PS detection for the VH channel; (b) 3U-Net-based PS detection for the VH channel; (c) 3U-Net-based PS detection only for the VH channel compared to VV channel; (d) GLRT-based PS detection for the VV channel; (e) 3U-Net-based PS detection for the VV channel; (f) 3U-Net-based PS detection only for the VV channel compared to VH channel.

4.2.3 Chapter Summary

This chapter has investigated use of co and cross polarized Sentinel-1A data in the proposed 3U-Net PS detection. The experimental results demonstrate that the model achieves reliable and robust PS detection performance across both polarization channels. In particular, its capability to identify PS points under low-coherence conditions is significantly enhanced in the VH channel, where it shows clear advantages over the conventional GLRT-based approach. These findings highlight the effectiveness of the proposed method in improving PS detection completeness and robustness in challenging scattering environments.

5 Conclusion and Outlook

5.1 Conclusion

This thesis has investigated the use of an artificial neural network as a support for the detection of PS in multi-temporal interferometric SAR data processing also known as PSI. The basic idea has been to integrate a widely used 2D DL model able to identify spatial structures in images, specifically U-Net, within a classical GLRT test. In Chapter 1 the basic concepts of SAR and InSAR to explain the classical PS detection schemes has been reviewed. Chapter 2 has been dedicated to the introduction of the basic concepts of DL. Chapter 3 has presented a detailed description of the proposed deep learning framework and experimental results are there shown and extended in Chapter 4. The main achievements of the conducted research are summarized as follows:

- A 3U-Net is proposed for PS detection, leveraging the strong representation capability of U-Net architectures in semantic segmentation. PS masks generated using the GLRT method are employed as pseudo-labels for proposed network training. Specifically, during pseudo-label generation, initial PS masks are extracted by combining the GLRT detector with a PFA curve. The resulting pseudo-labels are subsequently refined using the proposed IOF module to obtain reliable reference masks for network training. Because these pseudo-labels are primarily derived from coherence-based measurements, they may fail to adequately identify PS characterized by string velocity consistency but relatively low coherence. To address this limitation, the proposed framework jointly exploits ensemble coherence and velocity consistency, using their complementary information to generate denser, more complete, and more reliable PS.
- The proposed model was trained and evaluated using high-resolution SAR data acquired by the CSK and CSG missions over the Salerno and Lungro areas in southern Italy. The experimental results demonstrate that the PS detected by the proposed method outperform those obtained using the conventional GLRT detector in terms of both the number of reliable PS detections and the overall quality of the resulting spatial distributions. Because pixel level ground-truth PS masks are generally unavailable for real SAR observations, simulated data with known reference labels were additionally employed to quantitatively assess the PS detection accuracy. These experiments provide a more objective evaluation of the effectiveness and reliability of the proposed method.
- Furthermore, the proposed 3U-Net framework was extended to dual-polarization Sentinel-1A data. Experiments conducted using VV-and VH-polarized observations over the Vald' Agri and Moncenisio study areas demonstrate that the model consistently produces higher quality PS with more complete spatial coverage than the conventional GLRT method. This advantage is maintained when topographic information is used in place of entropy as a input feature, indicating that feature configuration. The relevant PS detection metrics improve by more than 30% compared with those of the GLRT-based approach. Notably, greater performance gains are achieved for the VH channel. These findings further demonstrate the robustness, adaptability, and effectiveness of the proposed method under complex scattering conditions and in low-coherence environments.

5.2 Outlook

Despite the promising performance achieved by the proposed deep learning-based PS detection framework, several aspects warrant further investigation.

- Firstly, a more extensive testing of the method on data should be carried out on other datasets to further confirm the significant improvement in the detection performances, together with validation with independent measurements.
- Despite the testing with X-Band and C-Band data, the generalization capability of the proposed framework across different sensors, datasets, and polarization modes is an interesting aspect to be addressed in the future developments.
- Secondly, further research is required on multi-source data fusion. At present, the model is trained using single-source datasets (e.g., CSK, CSG or Sentinel-1). To enhance generalization capability across diverse acquisition conditions, it is essential to incorporate heterogeneous SAR datasets into the training process. Given the rapid evolution of SAR sensors and acquisition modes, future model designs should be adaptable to multi-platform, multi-polarization, and multi-resolution SAR data.
- Finally, the current 3U-Net framework relies on GLRT-derived PS masks as supervisory signals. However, inaccuracies inherent in the GLRT-based detection process may propagate into the training labels, potentially limiting the model’s performance. Future work will therefore focus on developing end-to-end learning strategies that directly extract PS candidates from de-entangled interferometric complex (amplitude and phase) data, thereby eliminating dependence on conventional statistical detectors.

Bibliography

- [AKE04] Nico Adam, Bert Kampes, and Michael Eineder. Development of a scientific permanent scatterer system: Modifications for mixed ers/envisat time series. In *Proceedings of the 2004 Envisat & ERS Symposium*, volume 572, pages 1–73, 2004.
- [AKH⁺25a] S Azadnejad, A Kandiri, A Hrysiewicz, F O’Loughlin, EP Holohan, S Dev, and S Donohue. Application of deep learning for coherent pixel selection in time series insar for urban area and transport infrastructure monitoring. *International Journal of Applied Earth Observation and Geoinformation*, 142:104718, 2025.
- [AKH⁺25b] S. Azadnejad, A. Kandiri, A. Hrysiewicz, F. O’Loughlin, E.P. Holohan, S. Dev, and S. Donohue. Application of deep learning for coherent pixel selection in time series insar for urban area and transport infrastructure monitoring. *International Journal of Applied Earth Observation and Geoinformation*, 142:104718, 2025.
- [ALSL18] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Beyond rgb: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS journal of photogrammetry and remote sensing*, 140:20–32, 2018.
- [AvLGH07] Mahmut Arıkan, Freek van Leijen, Liu Guang, and Ramon Hanssen. Improved image alignment under the influence of elevation. In *Proceedings of FRINGE*, 2007.
- [BCH⁺21] Haiwei Bai, Jian Cheng, Xia Huang, Siyu Liu, and Changjian Deng. Hcanet: A hierarchical context aggregation network for semantic segmentation of high-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.
- [BDT⁺23] Zoé Berenger, Loïc Denis, Florence Tupin, Laurent Ferro-Famil, and Yue Huang. A deep-learning approach for sar tomographic imaging of forested areas. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023.
- [BFLS03] Paolo Berardino, Gianfranco Fornaro, Riccardo Lanari, and Eugenio Sansosti. A new algorithm for surface deformation monitoring based on small baseline differential sar interferograms. *IEEE Transactions on geoscience and remote sensing*, 40(11):2375–2383, 2003.
- [BJSV19] Alessandra Budillon, Angel Caroline Johnsy, Gilda Schirinzi, and Sergio Vitale. Sar tomography based on deep learning. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 3625–3628. IEEE, 2019.
- [BKC17] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

- [Bro92] Lisa Gottesfeld Brown. A survey of image registration techniques. *ACM computing surveys (CSUR)*, 24(4):325–376, 1992.
- [CHZ⁺23] Hesheng Chen, Yi He, Lifeng Zhang, Wang Yang, Yaoxiang Liu, Binghai Gao, Qing Zhang, and Jiangang Lu. A multi-input channel u-net landslide detection method fusing sar multisource remote sensing data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:1215–1232, 2023.
- [CMCG⁺16] M. Crosetto, O. Monserrat, M. Cuevas-González, N. Devanthéry, and B. Crippa. Persistent scatterer interferometry: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:78–89, 2016.
- [CMM11] Mario Costantini, Fabio Malvarosa, and Federico Minati. A general formulation for redundant integration of finite differences and phase unwrapping on a sparse multidimensional domain. *IEEE Transactions on Geoscience and Remote Sensing*, 50(3):758–768, 2011.
- [CPK⁺14] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.
- [CPK⁺17] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
- [CPSA17] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
- [CW14] Sizhe Chen and Haipeng Wang. Sar target recognition based on deep learning. In *2014 International Conference on Data Science and Advanced Analytics (DSAA)*, pages 541–547. IEEE, 2014.
- [CZP⁺18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
- [DFP09] A. De Maio, G. Fornaro, and A. Pauciullo. Detection of Single Scatterers in Multidimensional SAR Imaging. *IEEE Trans. Geosci. Remote Sens.*, 47(7):2284–2297, July 2009.
- [DKKP⁺21] Rucha Dave, Gaurav Kumar, Dharmendra Kr. Pandey, Armugha Khan, and Bimal Bhattacharya. Evaluation of modified dubois model for estimating surface soil moisture using dual polarization risat-1 c-band sar data. *Geocarto international*, 36(13):1459–1469, 2021.
- [DPRF23] Cosmin Dănişor, Antonio Pauciullo, Diego Reale, and Gianfranco Fornaro. Detection of Distributed Scatterers in Multitemporal SAR Interferometry: A Comparison Between CAESAR and SqueeSAR Detectors. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.
- [DTB20] Lei Ding, Hao Tang, and Lorenzo Bruzzone. Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1):426–435, 2020.
- [DXLH23] Li Ding, Min Xia, Haifeng Lin, and Kai Hu. Multi-level attention interactive network for cloud and snow detection segmentation. *Remote Sensing*, 16(1):112, 2023.
- [DZZLD15] Francesco De Zan, Mariantonietta Zonno, and Paco Lopez-Dekker. Phase inconsistencies and multiple scattering in sar interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, 53(12):6608–6616, 2015.

- [Ela88] Charles Elachi. Spaceborne radar remote sensing: applications and techniques. *Spaceborne radar remote sensing: Applications and techniques by Elachi*, page 39692, 1988.
- [Eld22] Knut Eldhuset. Using the pauli scattering mechanisms for analysis of polarimetric sar images from radarsat constellation and tandem-x missions. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:5799–5810, 2022.
- [EM16] Mostafa Esmaili and Mahdi Motagh. Improved persistent scatterer analysis using amplitude dispersion index optimization of dual polarimetry data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117:108–114, 2016.
- [FFN⁺11] Alessandro Ferretti, Alfio Fumagalli, Fabrizio Novali, Claudio Prati, Fabio Rocca, and Alessio Rucci. A new algorithm for processing interferometric data-stacks: Squeesar. *IEEE transactions on geoscience and remote sensing*, 49(9):3460–3470, 2011.
- [FLT⁺19] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.
- [FPP⁺24] Gianfranco Fornaro, Antonio Pauciullo, Vito Pascazio, Gilda Schirinzi, and Diego Reale. *Multi-Dimensional Imaging with Synthetic Aperture Radar*. Elsevier, 2024.
- [FPR01a] A. Ferretti, C. Prati, and F. Rocca. Permanent scatterers in sar interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, 39(1):8–20, 2001.
- [FPR01b] Alessandro Ferretti, Claudio Prati, and Fabio Rocca. Permanent scatterers in sar interferometry. *IEEE Transactions on geoscience and remote sensing*, 39(1):8–20, 2001.
- [GA13] Kanika Goel and Nico Adam. A distributed scatterer interferometry approach for precision monitoring of known surface deformation phenomena. *IEEE Transactions on Geoscience and Remote Sensing*, 52(9):5454–5468, 2013.
- [GDR16] Maria Daniela Graziano, Marco D’Errico, and Giancarlo Rufino. Ship heading and velocity analysis by wake detection in sar images. *Acta astronautica*, 128:72–82, 2016.
- [GL94] J-L Gauvain and Chin-Hui Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE transactions on speech and audio processing*, 2(2):291–298, 1994.
- [GR94] Dennis C Ghiglia and Louis A Romero. Robust two-dimensional weighted and unweighted phase unwrapping that uses fast transforms and iterative methods. *Journal of the Optical Society of America A*, 11(1):107–117, 1994.
- [GZW88] Richard M Goldstein, Howard A Zebker, and Charles L Werner. Satellite radar interferometry: Two-dimensional phase unwrapping. *Radio science*, 23(4):713–720, 1988.
- [Han01] Ramon F Hanssen. *Radar interferometry: data interpretation and error analysis*. Springer, 2001.
- [HB02] Jürgen Holzner and Richard Bamler. Burst-mode and scansar interferometry. *IEEE transactions on geoscience and remote sensing*, 40(9):1917–1934, 2002.
- [HK20] Thorsten Hoeser and Claudia Kuenzer. Object detection and image segmentation with deep learning on earth observation data: A review-part i: Evolution and recent trends. *Remote Sensing*, 12(10):1667, 2020.
- [HLL⁺25] Zihan Hu, Mofan Li, Gen Li, Yifan Wang, Chuanxu Sun, and Zehua Dong. Persistent scatterer pixel selection method based on multi-temporal feature extraction network. *Remote Sensing*, 17(19), 2025.
- [HNAB21] Sebastian Hafner, Andrea Nascetti, Hossein Azizpour, and Yifang Ban. Sentinel-1 and sentinel-2 data fusion for urban change detection using a dual stream u-net. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

- [Hoo06] Andrew John Hooper. *Persistent scatter radar interferometry for crustal deformation studies and modeling of volcanic deformation*. Stanford University, 2006.
- [Hoo08] Andrew Hooper. A multi-temporal insar method incorporating both persistent scatterer and small baseline approaches. *Geophysical research letters*, 35(16), 2008.
- [HPH⁺20] Ping Hu, Federico Perazzi, Fabian Caba Heilbron, Oliver Wang, Zhe Lin, Kate Saenko, and Stan Sclaroff. Real-time semantic segmentation with fast attention. *IEEE Robotics and Automation Letters*, 6(1):263–270, 2020.
- [HSS18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
- [HSZ07] Andrew Hooper, Pa Segall, and Howard Zebker. Persistent scatterer interferometric synthetic aperture radar for crustal deformation analysis, with application to volcán alcedo, galápagos. *Journal of Geophysical Research: Solid Earth*, 112(B7), 2007.
- [HWG⁺22] Jun Hu, Wenqing Wu, Rong Gui, Zhiwei Li, and Jianjun Zhu. Deep learning-based homogeneous pixel selection for multitemporal sar interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022.
- [HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [HZSK04] Andrew Hooper, Howard Zebker, Paul Segall, and Bert Kampes. A new method for measuring deformation on volcanoes and other natural terrains using insar persistent scatterers. *Geophysical research letters*, 31(23), 2004.
- [ISWY96] Katsushi Ikeuchi, Takeshi Shakunaga, Mark D Wheeler, and Taku Yamazaki. Invariant histograms and deformable template matching for sar target recognition. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 100–105. IEEE, 1996.
- [JB94] Dieter Just and Richard Bamler. Phase statistics of interferograms with applications to synthetic aperture radar. *Applied optics*, 33(20):4361–4368, 1994.
- [JDL13] Mi Jiang, Xiaoli Ding, and Zhiwei Li. Hybrid approach for unbiased coherence estimation for multitemporal insar. *IEEE Transactions on Geoscience and Remote Sensing*, 52(5):2459–2473, 2013.
- [Kam06] Bert M Kampes. *Radar interferometry: persistent scatterer technique*. Springer, 2006.
- [KvLMH05] Gini Ketelaar, Freek van Leijen, Petar Marinkovic, and Ramon Hanssen. Initial point selection and validation in ps-insar using integrated amplitude calibration. In *INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM*, volume 8, page 5490, 2005.
- [KXWY22] Wenchao Kang, Yuming Xiang, Feng Wang, and Hongjian You. Cfnet: A cross fusion network for joint land cover classification using optical and sar images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:1562–1574, 2022.
- [LBM⁺02] H Laur, P Bally, P Meadows, J Sanchez, B Schaettler, E Lopinto, and D Esteban. Derivation of the backscattering coefficient σ_0 in esa ers sar pri products. In *Proc. of the Second International Workshop on ERS Applications*, volume 139, 2002.
- [LCZ21] Han Liu, Zezong Chen, and Chen Zhao. A new doppler model incorporated with free and broken-short waves for coherent s-band wave radar at near-grazing angles. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021.
- [LDD21] Johannes Lohse, Anthony P Doulgeris, and Wolfgang Dierking. Incident angle dependence of sentinel-1 texture features for sea ice classification. *Remote Sensing*, 13(4):552, 2021.

- [LDG⁺17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- [LDZ20] Rui Li, Chenxi Duan, and Shunyi Zheng. Macu-net semantic segmentation from high-resolution remote sensing images. *arXiv preprint arXiv:2007.13083*, 2020.
- [LFW⁺18] Yongcheng Liu, Bin Fan, Lingfeng Wang, Jun Bai, Shiming Xiang, and Chunhong Pan. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. *ISPRS journal of photogrammetry and remote sensing*, 145:78–95, 2018.
- [LLS⁺20] Xiao Li, Lin Lei, Yuli Sun, Ming Li, and Gangyao Kuang. Multimodal bilinear fusion network with second-order attention-based channel selection for land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:1011–1026, 2020.
- [LMM⁺04] Riccardo Lanari, Oscar Mora, Michele Manunta, Jordi J Mallorquí, Paolo Berardino, and Eugenio Sansosti. A small-baseline approach for investigating deformations on full-resolution differential sar interferograms. *IEEE transactions on geoscience and remote sensing*, 42(7):1377–1386, 2004.
- [LP17] Jong-Sen Lee and Eric Pottier. *Polarimetric radar imaging: from basics to applications*. CRC press, 2017.
- [LQC⁺20] Haifeng Li, Kaijian Qiu, Li Chen, Xiaoming Mei, Liang Hong, and Chao Tao. Scattnet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 18(5):905–909, 2020.
- [LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [LZC⁺22] Xue Li, Guo Zhang, Hao Cui, Shasha Hou, Shunyao Wang, Xin Li, Yujia Chen, Zhijiang Li, and Li Zhang. Mcanet: A joint semantic segmentation framework of optical and sar images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106:102638, 2022.
- [LZD⁺21] Rui Li, Shunyi Zheng, Chenxi Duan, Jianlin Su, and Ce Zhang. Multistage attention resunet for semantic segmentation of fine-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.
- [LZZ⁺21] Rui Li, Shunyi Zheng, Ce Zhang, Chenxi Duan, Jianlin Su, Libo Wang, and Peter M Atkinson. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021.
- [NH10] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.
- [NHR⁺08] DO Nitti, RF Hanssen, A Refice, F Bovenga, G Milillo, and R Nutricato. Evaluation of dem-assisted sar coregistration. In *Image and signal processing for Remote sensing XIV*, volume 7109, pages 353–366. SPIE, 2008.
- [NHR⁺10] Davide Oscar Nitti, Ramon F Hanssen, Alberto Refice, Fabio Bovenga, and Raffaele Nutricato. Impact of dem-assisted coregistration on high-resolution sar interferometry. *IEEE Transactions on Geoscience and Remote Sensing*, 49(3):1127–1143, 2010.
- [NPD02] Giovanni Nico, Gintautas Palubinskas, and Mihai Datcu. Bayesian approaches to phase unwrapping: theoretical study. *IEEE Transactions on Signal processing*, 48(9):2545–2556, 2002.

- [Oli91] CJ Oliver. Information from sar images. *Journal of Physics D: Applied Physics*, 24(9):1493–1514, 1991.
- [PISM⁺12] Pau Prats-Iraola, Rolf Scheiber, Luca Marotti, Steffen Wollstadt, and Andreas Reigber. Tops interferometry with terrasars-x. *IEEE Transactions on geoscience and remote sensing*, 50(8):3179–3188, 2012.
- [PPP20] Hemani Parikh, Samir Patel, and Vibha Patel. Classification of sar and polsar images using deep learning: A review. *International Journal of Image and Data Fusion*, 11(1):1–32, 2020.
- [PWLK18] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.
- [PZMM21] Chengli Peng, Kaining Zhang, Yong Ma, and Jiayi Ma. Cross fusion net: A fast semantic segmentation network for small-scale semantic information capturing in aerial scenes. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021.
- [QF14] Rongjun Qin and Wei Fang. A hierarchical building detection method for very high resolution remotely sensed images combined with dsm using graph cut optimization. *Photogrammetric Engineering & Remote Sensing*, 80(9):873–883, 2014.
- [QWJ⁺22] Kun Qian, Yuanyuan Wang, Peter Jung, Yilei Shi, and Xiao Xiang Zhu. Basis pursuit denoising via recurrent neural network applied to super-resolving sar tomography. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022.
- [QWJ⁺24] Kun Qian, Yuanyuan Wang, Peter Jung, Yilei Shi, and Xiao Xiang Zhu. Hyperlista-abt: An ultra-light unfolded network for accurate multi-component differential tomographic sar inversion. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- [QWSZ22] Kun Qian, Yuanyuan Wang, Yilei Shi, and Xiao Xiang Zhu. γ -net: Superresolving sar tomographic inversion via deep learning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.
- [RBN06] Alberto Refice, Fabio Bovenga, and Raffaele Nutricato. Mst-based stepwise connection strategies for multipass radar data, with application to coregistration and equalization. *IEEE transactions on geoscience and remote sensing*, 44(8):2029–2040, 2006.
- [RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [SBM⁺06] E. Sansosti, P. Berardino, M. Manunta, F. Serafino, and G. Fornaro. Geometrical sar image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 44(10):2861–2870, 2006.
- [SC94] MS Seymour and IG Cumming. Maximum likelihood estimation for sar interferometry. In *Proceedings of IGARSS’94-1994 IEEE International Geoscience and Remote Sensing Symposium*, volume 4, pages 2272–2275. IEEE, 1994.
- [SCZ⁺22] Le Sun, Shiwei Cheng, Yuhui Zheng, Zebin Wu, and Jianwei Zhang. Spanet: Successive pooling attention network for semantic segmentation of remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:4045–4057, 2022.
- [SGN⁺20] João Paulo Delapasse Simioni, Laurindo Antonio Guasselli, Victor Fernandez Nascimento, Luis Fernando Chimelo Ruiz, and Tassia Fraga Belloli. Integration of multi-sensor analysis and decision tree for evaluation of dual and quad-pol sar in l-and c-bands applied for marsh delineation: Jpd simioni et al. *Environment, development and sustainability*, 22(6):5603–5620, 2020.

- [SJPT⁺14] Rolf Scheiber, Marc Jäger, Pau Prats-Iraola, Francesco De Zan, and Dirk Geudtner. Speckle tracking and interferometric processing of terrasars-x tops data for mapping non-stationary scenarios. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(4):1709–1720, 2014.
- [SLD⁺23] Hang Sun, Bohui Li, Zhiping Dan, Wei Hu, Bo Du, Wen Yang, and Jun Wan. Multi-level feature interaction and efficient non-local information enhanced channel attention for image dehazing. *Neural Networks*, 163:10–27, 2023.
- [SLM⁺03] Tazio Strozzi, Adrian Luckman, Tavi Murray, Urs Wegmuller, and Charles L Werner. Glacier motion estimation using sar offset-tracking procedures. *IEEE Transactions on Geoscience and Remote Sensing*, 40(11):2384–2391, 2003.
- [SRP⁺15] Francescopaolo Sica, Diego Reale, Giovanni Poggi, Luisa Verdoliva, and Gianfranco Fornaro. Nonlocal adaptive multilooking in SAR multipass differential interferometry. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(4):1727–1742, 2015.
- [TLBV02] Ridha Touzi, Armand Lopes, Jérôme Bruniquel, and Paris W Vachon. Coherence estimation for sar imagery. *IEEE transactions on geoscience and remote sensing*, 37(1):135–149, 2002.
- [TND20] Ashutosh Tiwari, Avadh Bihari Narayan, and Onkar Dikshit. Deep learning networks for selection of measurement pixels in multi-temporal sar interferometric processing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:169–182, 2020.
- [UBF⁺10] LMH Ulander, A Barmettler, B Flood, P-O Frolind, A Gustavsson, T Jonsson, E Meier, J Rasmusson, and G Stenstrom. Signal-to-clutter ratio enhancement in bistatic very high frequency (vhf)-band sar images of truck vehicles in forested and urban terrain. *IET radar, sonar & navigation*, 4(3):438–448, 2010.
- [Usa03] Stefania Usai. A least squares database approach for sar interferometric data. *IEEE Transactions on Geoscience and Remote Sensing*, 41(4):753–760, 2003.
- [VPRF20] Simona Verde, Antonio Pauciullo, Diego Reale, and Gianfranco Fornaro. Multiresolution detection of persistent scatterers: A performance comparison between multilook GLRT and CAESAR. *IEEE Transactions on Geoscience and Remote Sensing*, 59(4):3088–3103, 2020.
- [WB14] J. Wasowski and F. Bovenga. Investigating landslides and unstable slopes with satellite Multi Temporal Interferometry: Current issues and future perspectives. *Engineering Geology*, 174:103–138, 2014.
- [WEK22] Duo Wang, Markus Even, and Hansjrg Kutterer. Deep learning based distributed scatterers acceleration approach: Distributed scatterers prediction net. *International Journal of Applied Earth Observation and Geoinformation*, 115:103112, 2022.
- [WGSL23] Wenfu Wu, Songjing Guo, Zhenfeng Shao, and Deren Li. Crofusenet: A semantic segmentation network for urban impervious surface extraction based on cross fusion of optical and sar images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:2573–2588, 2023.
- [WPLK18] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.
- [WSXC16] Lei Wang, K Andrea Scott, Linlin Xu, and David A Clausi. Sea ice concentration estimation during melt from dual-pol sar scenes using deep convolutional neural networks: A case study. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8):4524–4533, 2016.

- [WZAG⁺19] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 28–37, 2019.
- [XLR⁺17] Xiaodong Xu, Wei Li, Qiong Ran, Qian Du, Lianru Gao, and Bing Zhang. Multisource remote sensing data classification based on convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):937–949, 2017.
- [XLZ⁺18] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.
- [XYOZ20] Qingsong Xu, Xin Yuan, Chaojun Ouyang, and Yue Zeng. Spatial-spectral ffnnet: Attention-based pyramid network for segmentation and classification of remote sensing images. *arXiv preprint arXiv:2008.08775*, 2020.
- [YLC⁺21] Xuan Yang, Shanshan Li, Zhengchao Chen, Jocelyn Chanussot, Xiuping Jia, Bing Zhang, Baipeng Li, and Pan Chen. An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 177:238–262, 2021.
- [YWDZ21] Jiaqi Yang, Chen Wu, Bo Du, and Liangpei Zhang. Enhanced multiscale feature fusion network for hsi classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12):10328–10347, 2021.
- [ZCZ⁺19] Xizhou Zhu, Dazhi Cheng, Zheng Zhang, Stephen Lin, and Jifeng Dai. An empirical study of spatial attention mechanisms in deep networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6688–6697, 2019.
- [ZLH⁺20] Qing Zhu, Cheng Liao, Han Hu, Xiaoming Mei, and Haifeng Li. Map-net: Multiple attending path neural network for building footprint extraction from remote sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):6169–6181, 2020.
- [ZLLZ21] Qi Zhao, Jiahui Liu, Yuewen Li, and Hong Zhang. Semantic segmentation with attention mechanism for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021.
- [ZMA⁺21] Xiao Xiang Zhu, Sina Montazeri, Mohsin Ali, Yuansheng Hua, Yuanyuan Wang, Lichao Mou, Yilei Shi, Feng Xu, and Richard Bamler. Deep learning meets sar: Concepts, models, pitfalls, and perspectives. *IEEE Geoscience and Remote Sensing Magazine*, 9(4):143–172, 2021.
- [ZMLZ10] Daiyin Zhu, Xinhua Mao, Yong Li, and Zhaoda Zhu. Far-field limit of pfa for sar moving target imaging. *IEEE Transactions on Aerospace and Electronic systems*, 46(2):917–929, 2010.
- [ZP02] Qun Zhao and Jose C Principe. Support vector machines for sar automatic target recognition. *IEEE Transactions on Aerospace and Electronic Systems*, 37(2):643–654, 2002.
- [ZSHH22] Zhiqiang Zeng, Jinping Sun, Zhu Han, and Wen Hong. Sar automatic target recognition method based on multi-stream complex-valued networks. *IEEE transactions on geoscience and remote sensing*, 60:1–18, 2022.
- [ZSQ⁺17] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.
- [ZWD⁺22] Yonghong Zhang, Jujie Wei, M Duan, Yonghui Kang, Q He, Hongan Wu, and Zhong Lu. Coherent pixel selection using a dual-channel 1-d cnn for time series insar analysis. *International Journal of Applied Earth Observation and Geoinformation*, 112:102927, 2022.

- [ZWZ⁺24] Xiangrong Zhang, Zhenhang Weng, Peng Zhu, Xiao Han, Jin Zhu, and Licheng Jiao. Esdinet: Efficient shallow-deep interaction network for semantic segmentation of high-resolution aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.
- [ZZWM20] Zhuo Zheng, Yanfei Zhong, Junjue Wang, and Ailong Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4096–4105, 2020.

List of Publications

International Journals

- [J1] W. Tang, S. Vitale, S. Verde and G. Fornaro, "Persistent Scatterers Detection Supported by Deep Learning: A Solution Based on U-Net," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 64, pp. 1-12, 2026, Art no. 5202012, doi: 10.1109/TGRS.2026.3657993

International Conferences

- [C1] Weili Tang, Sergio Vitale, Diego Reale, Edinson Andres Solarte Casanova, Simona Verde, Gabriel Vasile, and Gianfranco Fornaro, "Deep-Learning Based Detection of Persistent Scatterers on Sentinel-1 data," *EUSAR 2026 - 16th European Conference on Synthetic Aperture Radar*, Baden-Baden, Germany, 9-11 June 2026 (Accepted in the Student Competition Session).
- [C2] W. Tang, S. Verde, S. Vitale, V. Pascazio and G. Fornaro, "A Deep Learning Based Solution for Persistent Scatterers Detection," *IGARSS 2025 - 2025 IEEE International Geoscience and Remote Sensing Symposium*, Brisbane, Australia, 2025, pp. 261-264, doi: 10.1109/IGARSS55030.2025.11243492.
- [C3] W. Tang et al., "Deep Learning Based Persistent Scatterers Detection: First Results," *2024 IEEE International Workshop on Technologies for Defense and Security (TechDefense)*, Naples, Italy, 2024, pp. 189-193, doi: 10.1109/TechDefense63521.2024.10863597 (Young Researcher Award - Second Price)