

# Deep Learning for Radargrammetric DSM Generation: A StereoSAR Dataset and Multi-Scale Fusion Network

Yuting Dong, Yaozu Li, Ji Zhao, Yao Sun, Mingsheng Liao

**Abstract**—Radargrammetry is an important technique for digital surface model (DSM) reconstruction, but accurate disparity estimation from synthetic aperture radar (SAR) stereo images remains challenging due to speckle noise and geometric distortions. Despite the success of deep learning in disparity estimation for stereo matching, its application in stereoscopic SAR (StereoSAR) is still limited due to the lack of high-quality training data and task-specific models. To address this issue, this study develops a deep learning framework for radargrammetric DSM generation, integrating dataset construction and a multi-scale SAR stereo matching network. The StereoSAR4DSM dataset is developed using TerraSAR-X imagery and high-resolution aerial DSMs, with enhanced epipolar rectification and three SAR-driven augmentation strategies: multi-looking variation, random pixel sampling, and random elevation perturbation. These strategies enrich data diversity and support robust deep learning model training. Based on this dataset, we design the Multi-Scale StereoSAR Fusion Network (MSSFNet) that constructs pyramid cost volumes and progressively integrates multi-scale cost information. An attention-guided fusion mechanism and a disparity refinement module further enhance matching accuracy and restore fine terrain structures. Experimental results on two test areas with different imaging modes demonstrate that the deep learning model trained with the StereoSAR4DSM dataset outperforms traditional approaches, and the proposed MSSFNet achieves the highest accuracy compared with other deep learning methods. In addition, comparative experiments show that the SAR-driven enhancement strategies significantly improve data diversity and lead to more accurate disparity estimation. Overall, these findings confirm the effectiveness of the proposed framework and highlight the potential of deep learning-based StereoSAR methods for efficient and accurate DSM reconstruction.

**Index Terms**—DSM generation; Radargrammetry; Deep Learning; StereoSAR; Spaceborne SAR

This work was supported by National Natural Science Foundation of China under Grant Nos. 42474047 and 42471416. (Corresponding author: Ji Zhao)

Y. Dong and Y. Li are with the School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China.

J. Zhao is with the School of Computer Science, China University of Geosciences, Wuhan, 430074, China (e-mail: zhaoji@cug.edu.cn).

Y. Sun is with the Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling 82234, Germany.

M. Liao is with State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China.

## I. INTRODUCTION

DIGITAL Surface Models (DSMs) are three-dimensional geospatial data model representing surface features, and are fundamental to a wide range of geospatial applications, such as disaster assessment [1], glacier monitoring [2], and soil erosion [3]. Among the existing remote sensing technologies, Synthetic Aperture Radar (SAR) has the ability of cloud penetration and all-weather imaging, and has unique advantages in generating DSM in mountainous areas with poor meteorological conditions. Stereoscopic SAR (StereoSAR) relies on the amplitude information of SAR images to reconstruct DSM through stereo matching. Compared with phase-based methods [4] [5], StereoSAR is not limited by temporal decorrelation [6-8] and thus is particularly suitable for vegetation-covered or rugged terrain regions where phase coherence is difficult to maintain. Therefore, StereoSAR serves as a valuable and complementary technique for DSM generation [9].

Image matching in StereoSAR is a key step in generating DSM, which is to identify the homologous points between SAR stereo image pairs to obtain disparity. The disparity is subsequently converted into elevation information. Considering the unique challenges posed by SAR images, such as speckle noise and geometric distortion, it is important to develop matching algorithms for SAR stereo images to ensure accurate and reliable results. Initial attempts in SAR stereo matching relied primarily on window-based and feature-based methods such as normalized cross-correlation (NCC) [10-12], scale-invariant feature transform (SIFT) [13, 14], and least squares matching (LSM) [15, 16]. These techniques compare pixel intensity patterns within a predefined window to detect corresponding regions in the SAR stereo image pair. Although these methods provide a straightforward matching framework, their performance is easily affected by speckle noises.

Compared to local window methods that determine the disparity on a per-window basis, global matching algorithms aim to optimize the disparity map of the entire image by considering global energy minimization. Global methods formulate stereo matching as an energy minimization problem, which combine a data fidelity term and a smoothness term. A notable and widely used global method is semi-global matching (SGM) [17, 18]. SGM combines local matching costs with aggregated path-wise smoothness constraints across multiple directions to make a balance between computational efficiency and matching accuracy. For example, the Digital Automatic

TGRS-2025-09171

Terrain Extractor (DATE) workflow for radargrammetric DSM generation is proposed using the SGM algorithm integrated in the Open Source Computer Vision Library (OpenCV) [19]. The hierarchical SGM method refines disparities layer by layer by constructing a multi-scale pyramid to reconstruct a higher quality DSM [20]. These methods incorporate spatial continuity constraints into the matching process, resulting in smoother and more consistent disparity maps.

To more effectively apply local and global matching algorithms for SAR stereo images, SAR images can be rectified to reduce the geometric dissimilarities between two images of a stereo pair before the matching process. As early as 1996, Raggam and Almer [21] introduced the affine transformation to roughly align the right image to the coordinate space of the left image and limit the search range of corresponding points to a rectangular area along the range direction. To further narrow the search area to one-dimensional lines, the researchers introduced the concept of pseudo-epipolar line generation from aerial photogrammetry into SAR processing. This is achieved by projecting the two SAR images onto a common local elevation plane, which is usually defined by the mean elevation of the study area [22]. This process, known as pseudo epipolar image generation or planar epipolar rectification [23-26]. To enable more accurate geometric alignment, the enhanced epipolar rectification that incorporates external elevation information [16, 27] is developed to eliminate the azimuth disparity and allow the search to focus predominantly on the range direction. After geometric distortion compensation, local or global matching algorithms are applied to determine the exact correspondence. Local algorithms usually follow a winner-takes-all strategy to select the candidate with the lowest matching cost, while global methods incorporate spatial continuity constraints to generate smoother disparity maps.

Despite significant progress achieved through local and global matching methods, deep learning remains largely unexplored in the field of SAR stereo matching. In optical stereo matching, deep learning techniques have largely advanced the field by learning robust feature representations and end-to-end disparity estimation from large datasets. Convolutional neural network (CNN)-based methods have demonstrated strong generalization capabilities and robustness by effectively extracting feature information from images. For instance, StereoNet [28] employs a lightweight architecture that directly regresses disparities through a compact cost volume and hierarchical refinement layers, enabling efficient inference. PSMNet [29] further incorporates spatial pyramid pooling to capture multi-scale contextual cues, constructs a dense cost volume via feature concatenation, and applies 3D convolutions for cost aggregation. To handle challenges in optical remote sensing imagery, HMSMNet [30] adopts a hierarchical multi-scale framework that jointly preserves coarse global structures and fine local details. In addition, Transformer-based models such as STTR [31] replaces cost volume construction with self- and cross-attention to perform dense matching. These deep learning methods achieve impressive matching results in optical applications. However, applying deep learning methods to SAR stereo matching faces unique challenges due to the inherent

speckle noise and geometric distortions of SAR images. These factors, coupled with the scarcity of annotated StereoSAR matching datasets, have so far hampered the development of deep learning-based StereoSAR matching methods.

To address these gaps, this study develops a deep learning approach for radargrammetric DSM generation from the construction of StereoSAR dataset to a multi-scale StereoSAR fusion network. To advance deep learning methods for DSM reconstruction from SAR stereo images, the StereoSAR4DSM dataset is developed from TerraSAR-X high-resolution SAR images and high-precision aerial DSM, serving as a strong foundation for deep learning. To enhance model generalization and data diversity, SAR-driven enhancement strategies are introduced, combining domain-relevant transformations to simulate diverse imaging conditions. These strategies include multi-looking with variable window sizes to mitigate the effects of speckle noise, random sampling within pixel windows to enrich local texture variation, and elevation perturbation to emulate terrain-induced disparity variability. Overall, the dataset contains nearly 1,000 training and validation  $1024 \times 1024$  pixel patches, as well as carefully selected test regions for evaluating in-domain and out-of-domain generalization. To accurately predict the disparity from SAR stereo image pairs, we designed a multi-scale StereoSAR fusion network (MSSFNet) for radargrammetric DSM generation, which fully leverages both structural and textural information across multiple resolutions. MSSFNet extracts multi-scale features to construct pyramid cost volumes to exploit both coarse terrain structure and fine-grained image details. To further enhance robustness, a progressive cost fusion strategy is employed. Cost volumes from different scales are regularized independently and then fused progressively from low to high resolution via the Cost Information Fusion (CIF) module, which adaptively weights and combines features from adjacent scales. To refine the final disparity prediction and recover fine structural details, the disparity refinement module based on residual learning is introduced by integrating SAR image intensity and gradient features to recover high-frequency terrain structures. The refined disparity is converted into 3D coordinates through radargrammetric forward intersection and interpolated into a regular grid to generate the final DSM. Overall, MSSFNet effectively combines deep learning with StereoSAR domain priors to achieve high-resolution DSM reconstruction with improved accuracy and generalization ability. The dataset and code will be made publicly available at <https://github.com/Mango-Mars/MSSFNet>. The contributions of the study are as follows:

- 1) According to the characteristics of stereo SAR images, we propose a deep learning framework for radargrammetric DSM generation from the construction of StereoSAR dataset to a multi-scale StereoSAR fusion network.
- 2) To enrich the training diversity, we introduce three SAR-driven enhancement strategies in the StereoSAR4DSM dataset (multi-looking variation, random pixel sampling, and random elevation perturbation).

TGRS-2025-09171

- 3) To achieve robust SAR stereo matching, we design the MSSFNet, which progressively aggregates disparity information across scales via attention-guided cost fusion.

The remainder of this paper is organized as follows. Section II introduces the construction of the StereoSAR4DSM dataset. Section III describes the proposed method, including feature extraction, cost volume fusion, disparity regression, and DSM reconstruction. Experimental results and corresponding discussions are in Section IV and Section V respectively. Finally, Section VI concludes the paper.

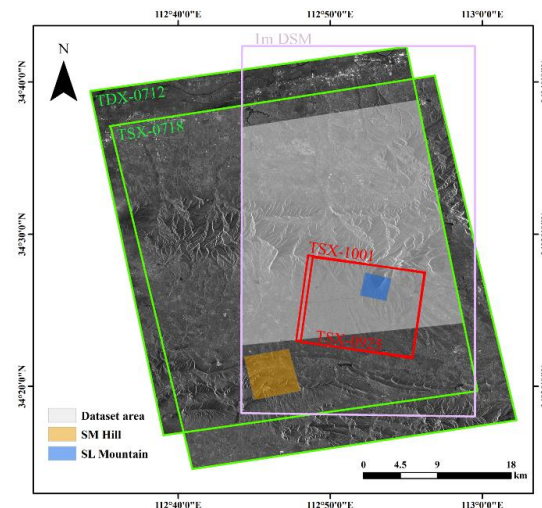
## II. STEREO SAR4DSM: A STEREO SAR DATASET FOR DEEP LEARNING-BASED DSM GENERATION

### A. Study Area and Data Sources

Mount Song, located in Henan Province, central China, was selected as the experimental site for this study. Elevation in the region varies significantly from about 150 m to more than 1500 m, featuring steep slopes, rugged mountains, and densely forested hills. Such diverse terrain features often result in severe geometric distortions and shadow effects in SAR images, especially in side-looking geometry. Furthermore, dense vegetation cover leads to temporal and phase decorrelation, which further complicates SAR-based surface reconstruction. These complex terrain features and vegetation conditions in the region provide a challenging but ideal test platform for generating DSM using StereoSAR.

To support radargrammetric DSM generation, we selected two pairs of high-resolution SAR stereo images acquired by the TerraSAR-X and TanDEM-X satellites. One pair was obtained in stripmap (SM) mode from an ascending orbit, and the other in spotlight (SL) mode from a descending orbit, as illustrated in Fig. 1. These two configurations acquired under different imaging modes and orbital geometries are designed to evaluate the generalization ability of the deep learning model under different SAR acquisition conditions. These SAR images have a high resolution and can capture fine structural details of the terrain surface. Detailed acquisition parameters, such as acquisition dates, incidence angles, orbit directions, and modes, are summarized in Table I.

Two types of elevation data were employed to support dataset generation and algorithm evaluation. The first is the Shuttle Radar Topography Mission (SRTM) digital elevation model (DEM) [32], which provides global elevation coverage at a resolution of 30 m. In this study, the SRTM DEM was used to generate epipolar rectified image pairs and served as a baseline comparison. The second and more critical component is a high-precision DSM derived from aerial photogrammetry, which offers a spatial resolution of 1 m and an absolute vertical accuracy better than 1 m. Although the photogrammetric DSM was acquired in 2009 (two years after the SAR imagery) the relatively stable topography of the area ensures its suitability as a reliable ground-truth reference for StereoSAR evaluation. The aerial DSM serves as the primary ground truth for generating disparity labels and verifying the accuracy of the final DSM reconstruction. The extent of this aerial DSM, covering the core study area, is marked in Fig. 1 by a purple rectangular region.



**Fig. 1.** Overview of study area and SAR acquisition in the Mount Song region. The green and red rectangles indicate the coverage of the ascending stripmap and descending spotlight TerraSAR/TanDEM-X image pairs, respectively. The purple rectangle indicates the extent of the high-precision aerial photogrammetric DSM used for ground truth generation and validation.

TABLE I

BASIC PARAMETERS OF THE SELECTED TERRASAR-X/TANDEM-X SAR IMAGE PAIRS

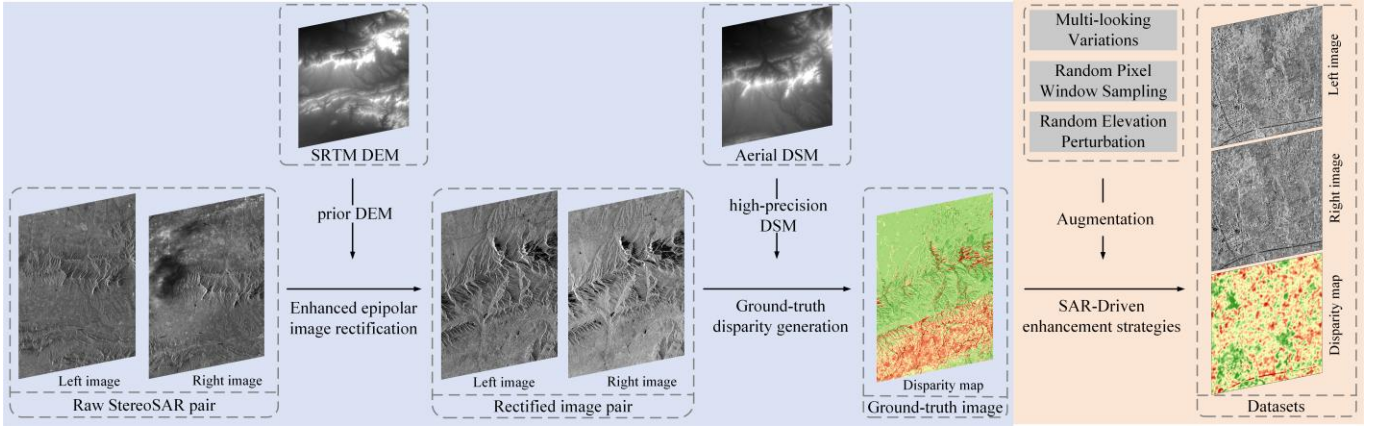
ID	Satellite	Mode	Acquisition Time	Orbit Direction	Incidence Angel (°)	Resolution of slant range/azimuth (m)
0712	TDX-1	SM	2011-07-12	Ascending	44.5	1.8/3.3
0718	TSX-1	SM	2011-07-18	Ascending	28.9	1.2/3.3
0925	TSX-1	SL	2011-09-25	Descending	39.8	1.2/1.6
1001	TSX-1	SL	2011-10-01	Descending	47.1	1.2/1.6

### B. Workflow for StereoSAR4DSM Generation

The construction of the StereoSAR4DSM dataset consists of three key stages: enhanced epipolar image rectification with prior DEM, ground-truth disparity map generation, and SAR-driven enhancement strategies for dataset diversity. These steps aim to build a high-quality benchmark dataset that links raw SAR images to accurate ground-truth disparity maps. The overall pipeline is illustrated in Fig. 2. The workflow exploits SAR image geometry information to ensure spatial consistency and facilitates the efficient training of deep learning models for radargrammetric DSM generation.

#### 1) Enhanced epipolar image rectification with prior DEM

Epipolar geometry plays a strong constraint in stereo matching, reducing the search space of corresponding pixels from two dimensions to one dimension. In optical photogrammetry, the epipolar line is the intersection of the image plane and the epipolar plane, defined by three-dimensional ground points and a stereo baseline [33]. The key property is that corresponding points in the second image must lie on the same epipolar line as the points in the first image. Although SAR systems differ from optical sensors in imaging geometry, previous studies [16, 26] have demonstrated that an epipolar-like constraint can be established for SAR stereo image pairs, enabling effective epipolar rectification and one-dimensional correspondence search.



**Fig. 2.** Overall workflow for generating the StereoSAR4DSM dataset.

However, direct rectification of SAR images is challenging due to the side-looking nature of SAR and the complex imaging geometry governed by the Range-Doppler (RD) model. Traditional rectification methods often use a constant elevation plane to construct pseudo-epipolar images, which neglect terrain variations and may lead to residual geometric distortions, especially in areas with steep terrain. To address this limitation, we employ an enhanced epipolar rectification strategy that incorporates external elevation information from the prior DEM, such as SRTM DEM. This approach enables more accurate geometric alignment between stereo pairs, while significantly reducing parallax ambiguity caused by terrain-induced distortions [16, 27].

The rectification process starts with radar encoding, in which the SRTM DEM as prior DEM is projected into the coordinate system of the SAR image. Specifically, for each pixel  $(r_m, c_m)$  in the left SAR image, its corresponding elevation  $h$  is extracted from the SRTM DEM. Using the precise orbital parameters and the RD model, these image coordinates in the left SAR image are mapped to geographical positions in the WGS84 system. This transformation is governed by range sphere equation and Doppler cone equation, expressed in (1) and (2) [34], respectively:

$$f_{range}(X_S(t_A), X_T, t_R) = |X_S - X_T| = c_0 \cdot t_R / 2 \quad (1)$$

$$f_D(X_S(t_A), V_S(t_A), X_T) = \frac{V_S \cdot (X_T - X_S)}{|V_S| \cdot |X_T - X_S|} = \sin\alpha \quad (2)$$

where  $X_T$  is the 3D coordinate of the ground target, and  $X_S$  and  $V_S$  are the SAR sensor's position and velocity vectors, which are determined by the azimuth imaging time  $t_A$ .  $c_0$  is the speed of light,  $t_R$  is the range imaging time.  $\alpha$  is the squint angle of the radar signal, and is zero for zero-Doppler processed SAR data.

These equations link SAR image coordinates to geographic coordinates through orbital information and radar signal geometry. Once the ground coordinates are obtained from the left SAR image, they are transformed back to right SAR image coordinates via RD model, using the same elevation reference  $h$ . The corresponding value in the right image is then interpolated to form the rectified stereo pair. This process compensates for terrain-induced distortions and aligns the SAR stereo images to reduce the disparity search from two

dimensions to one. The resulting rectified image pair exhibits reduced geometric distortion and improved alignment for accurate disparity estimation.

## 2) Ground-truth Disparity generation with high-precision DSM

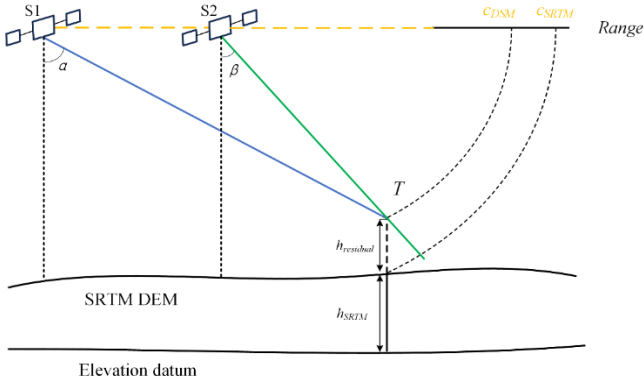
SAR Stereo image pairs and corresponding disparity maps are essential components for training deep learning models in radargrammetric DSM generation. After obtaining rectified SAR stereo image pairs through the enhanced epipolar rectification process, we derive the disparity ground truth by aligning the two images using high-precision elevation references.

The disparity generation process first projects the high-precision DSM obtained with sub-meter accuracy via aerial photogrammetry into the image coordinates of the stereo SAR pair using the rigorous RD model. For a rectified SAR stereo pair, by identifying the corresponding pixel on the right image for each pixel in the left image, homologous points can be determined and the residual disparity can be subsequently calculated. The ground truth disparity is obtained by projecting the high-precision aerial DSM and the SRTM DEM into the coordinate system of the right image, and computing the difference in column indices as defined in Eq. (3). Because both projections rely on the same SAR imaging geometry, the resulting disparity maps are geometrically consistent with the stereo image pair, ensuring that no mismatch is introduced during supervision. As illustrated in Fig. 3, this disparity is primarily expressed along the range (column) direction.

$$d = c_{SRTM} - c_{DSM} \quad (3)$$

where  $d$  is the ground truth disparity,  $c_{SRTM}$ ,  $c_{DSM}$  are the column indices by projecting the SRTM DEM and the high-precision aerial DSM, respectively, onto the right image. This process effectively quantifies the residual disparity caused by the elevation differences between the coarse SRTM and the high-precision reference DSM. The resulting disparity maps represent the exact pixel-wise disparities required for training deep learning models to reconstruct dense and accurate DSMs from raw SAR stereo images.

TGRS-2025-09171



**Fig. 3.** Illustration of disparity generation based on StereoSAR imaging geometry.

### 3) SAR-Driven Enhancement Strategies for Dataset Diversity

To enrich the diversity of the StereoSAR4DSM dataset, we develop a set of SAR-specific data enhancement strategies to simulate the diversity of real-world imaging conditions. These methods are designed to capture the inherent variations caused by different orbital geometries, terrain conditions, and SAR imaging parameters. Specifically, we implement flipping operations to simulate the change of satellite viewpoints, effectively generating both ascending and descending orbit perspectives from a single acquisition. In addition, we develop three enhancement techniques that are closely related to the disparity formation process of SAR stereo image pairs: multi-looking variations, random pixel window sampling, and simulated elevation perturbations. By integrating these enhancements, the dataset covers a wider range of SAR stereo matching scenarios, which helps in constructing more comprehensive benchmarks and promotes the development of robust SAR stereo matching algorithms.

(1) Multi-looking Variations. A spatial multi-looking was applied to both SAR amplitude images to reduce speckle noise and enhance the radiometric consistency between stereo pairs. Given the original amplitude image  $I(x, y)$ , the multi-looking was implemented in the image domain through local spatial averaging within an  $m \times n$  window. The multi-looked amplitude  $I_{m,n}(x, y)$  was then computed as [35]:

$$I_{m,n}(x, y) = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} I(x+i, y+j) \quad (4)$$

We generate versions with varying window sizes to simulate images with different levels of detail and scale representation. These variations expose the model to different speckle conditions and structural textures, thus enhancing the robustness of the matching.

(2) Random Pixel Window Sampling. To capture subtle textural variations, we introduce a random pixel sampling mechanism within the multi-looking window and extract not only the central pixel of each multi-looking window but also randomly sample several surrounding pixels. If the window is  $m \times n$ , we randomly choose  $K$  pixels to simulate matching inconsistencies and sub-pixel shifts.

$$I_{rand} = I(x + \delta x_k, y + \delta y_k) \quad (5)$$

where  $\delta x_k$  and  $\delta y_k$  are the  $k$ -th random offsets within the multi-looking window. This introduces subtle local variations in multi-looking, allowing the model to become more robust to intra-window variations caused by speckle or layover effects.

(3) Random Elevation Perturbation. Although the SRTM DEM is employed to assist the enhanced epipolar rectification process by providing a coarse elevation prior, it is not required to have high accuracy. The elevation values in SRTM mainly reflect the overall terrain trend, but may contain significant residual errors at finer scales, especially in complex or rapidly changing terrain. To simulate these uncertainties and enrich the disparity distribution in the dataset, we introduce a random elevation perturbation mechanism. This approach generates variant height inputs by injecting controlled random elevation shifts into the SRTM elevation values, thereby simulating real-world elevation deviations and producing different disparity outcomes. The perturbed elevation  $h'$  is calculated by applying a random perturbations  $\Delta h$  to the SRTM elevation  $h$ :

$$h' = h + \Delta h, \quad \Delta h \in U(-\rho, \rho) \quad (6)$$

where  $U(-\rho, \rho)$  denotes a uniform distribution used to simulate residual elevation shifts, and  $\rho$  is an adjustable parameter determined based on the observed local elevation variation of the SRTM DEM in the study area. These perturbed elevations are then used to regenerate the rectified SAR stereo images, introducing diverse disparity conditions into the dataset.

### C. StereoSAR4DSM Dataset Structure and Splits

After generating the epipolar rectified SAR stereo images and corresponding disparity maps described in Section II.B, we proceeded to construct the StereoSAR4DSM dataset for supervised deep learning-based DSM generation. To further enrich the dataset, SAR-driven enhancement strategies was employed. Specifically, SAR images were processed using four different multi-looking window sizes:  $2 \times 2$ ,  $3 \times 3$ ,  $4 \times 4$ , and  $5 \times 5$ . Each original image under the specific multi-looking window configuration generates three additional enhanced views based on the other enhancement strategies, resulting in a four-fold increase in the amount of data for each setting. To accommodate GPU memory limitations, considering the high resolution and large spatial extent of the original multi-look SAR data, all images and disparity maps were cropped into  $1024 \times 1024$  pixel patches with 50% overlap between adjacent regions to ensure spatial continuity and sufficient sample diversity. Table II shows the total number of image patches before and after augmentation and their distribution in the training set, validation set, and test set.

TABLE II

DETAILED STATISTICS OF THE STEREO SAR4DSM DATASET AFTER APPLYING SAR-DRIVEN ENHANCEMENT STRATEGIES

Multi-looking	Initial quantity	The number of Augmentation	Total	Training	Validation	Testing
$2 \times 2$	112	336	448	400	48	\
$3 \times 3$	60	180	240	220	20	\
$4 \times 4$	40	120	160	145	15	\
$5 \times 5$	32	90	122	105	15	2
Total	242	726	968	870	98	2

TGRS-2025-09171

As reported in Table II, the StereoSAR4DSM dataset comprises 968 SAR stereo image pairs across multi-looking configurations, of which 870 pairs are used for training, 98 pairs are used for validation, and 2 representative image pairs in a 5×5 multi-looking configuration are reserved for testing in challenging terrains and imaging scenarios. The test regions were carefully selected to assess model performance and generalization.

**SM Hill.** As shown in the yellow area in Fig. 1, this region has the same imaging conditions as the training data but is geographically separated. It covers an area of 6.6 km×6 km, with elevation ranging from 392 to 828 m. The SM Hill data can be used to evaluate spatial generalization of deep learning models since it is located in another area of the same image as the training data.

**SL Mountain.** As shown by the blue area in Fig. 1, the SL Mountain area was acquired in spotlight mode during a descending orbit, featuring different incidence angles and finer azimuth resolution compared to the ascending stripmap mode data used for training. Its area is 3 km×2.8 km and its elevation ranges from 446 m to 1174 m, presenting a more complex topography. Therefore, the SL Mountain data can be served as a reliable test of the model's ability to generalize across acquisition conditions.

### III. MULTI-SCALE STEREOSAR FUSION NETWORK

As shown in Fig. 4, the proposed framework consists of four main components: pyramid cost volume construction with multi-scale features, progressive cost aggregation, disparity regression and refinement, and DSM reconstruction. MSSFNet starts by extracting hierarchical features at multiple scales to construct pyramid cost volumes. These cost volumes are then regularized and fused in a coarse-to-fine manner, enabling robust matching under diverse imaging conditions. The predicted disparity map is progressively refined by integrating intensity and gradient information from the left SAR image to recover high-frequency terrain details. Finally, the refined disparity is converted into 3D coordinates and interpolated into a regular grid to generate the final DSM. In the following subsections, we describe our approach in detail.

#### A. Pyramid Costs Volume Construction with Multi-Scale Features

Due to its unique imaging mechanism, SAR images often exhibit speckle noise and large texture-free areas. These characteristics make ground objects retain their overall structure from a global perspective, but have less texture details from a local perspective. To address this challenge, we adopt a multi-scale feature extraction strategy that can capture both coarse structures and fine-grained details at multiple spatial scales. Initially, three 3×3 2D convolutions are used to extract basic representations. This is followed by 13 residual blocks [36] with varying dilation rates, enabling the network to capture a wider range of contextual information without losing resolution. To obtain multi-scale representations, four parallel average pooling layers with kernel sizes of 1×1, 2×2, 4×4, and 8×8 are applied, each followed by a 1×1 convolution for dimensionality

adjustment. These layers generate feature maps at 1/4, 1/8, 1/16, and 1/32 of the original resolution, enabling the model to effectively perceive both global structures and local variations. The architecture of the feature extraction module is reported in Fig. 5.

Using the extracted multi-scale features, we construct pyramid cost volumes at four levels (1/4, 1/8, 1/16, and 1/32 of the input resolution) to encode potential pixel correspondences. Feature concatenation and feature difference are two commonly used cost volume construction methods. Feature concatenation preserves comprehensive appearance and contextual information, while difference-based encoding highlights differences that are crucial for disparity estimation, especially in low-texture SAR regions. Therefore, instead of relying solely on one cost volume construction method, we adopt a hybrid scheme that alternates between feature concatenation and feature difference to balance representation richness and computational cost. At each scale, the cost volume is constructed using either the concatenation of left and right features or their element-wise difference. This combination enables the network to retain necessary contextual clues while reducing the computational burden.

#### B. Progressive Cost Aggregation

Based on the construction of pyramid cost volumes at multiple scales, effective aggregation and fusion of cost information is an important part of robust disparity estimation, especially in the presence of SAR specific noise (e.g., speckle and structural distortion). In our study, we adopt a progressive cost aggregation strategy that jointly performs intra-scale regularization and inter-scale fusion, enabling the network to fully exploit spatial context across scales while mitigating noise-induced disparity errors.

(1) *Intra-Scale Cost Regularization.* For each constructed cost volume, we adopt different regularization strategies depending on the constructed method. The cost volumes generated by feature difference [37] are processed through consecutive 3D convolutions in Fig. 6(b) to enforce local consistency and smoothness. For the cost volumes generated by feature concatenation [38], we adopt the Hourglass module in Fig. 6(c), which captures multi-scale contextual dependencies via an encoder-decoder. This multi-level context modeling facilitates robust matching in texture-less or noisy regions that are common in SAR images.

(2) *Inter-Scale Cost Information Fusion.* To further enhance disparity estimation, we introduce a progressive multi-scale cost fusion mechanism, as shown in Fig. 6(a). Starting from the lowest resolution scale, the aggregated cost volume is upsampled using bilinear interpolation and fused with the adjacent higher resolution cost volume. This fusion is conducted recursively from coarse to fine scales. However, directly summing the cost volumes across scales may propagate inconsistent or noisy responses. Therefore, we adopt a cost information fusion (CIF) module inspired by the channel attention mechanism [39]. As shown in Fig. 6(d), this CIF module generates a scale-specific attention vector by computing adaptive weights of two cost volumes through a

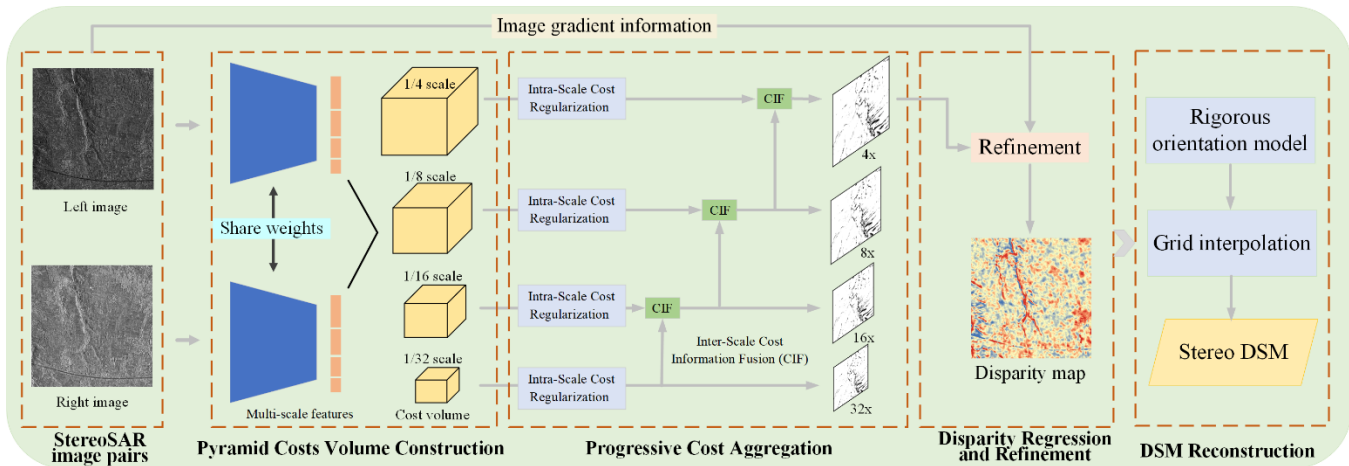


Fig. 4. The proposed multi-scale StereoSAR fusion network (MSSFNet) for radargrammetric DSM generation.

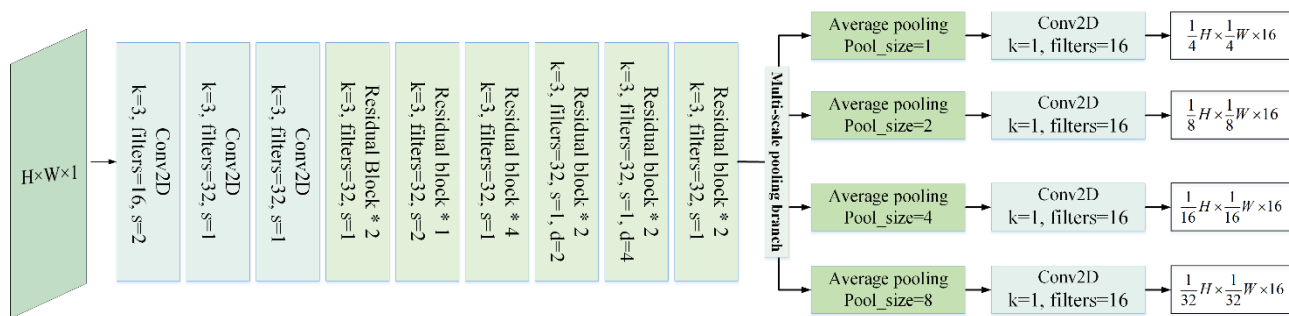


Fig. 5. The architecture of the feature extraction module. "k" represents the size of the convolution kernel, "s" indicates stride, and "d" denotes dilation rate.

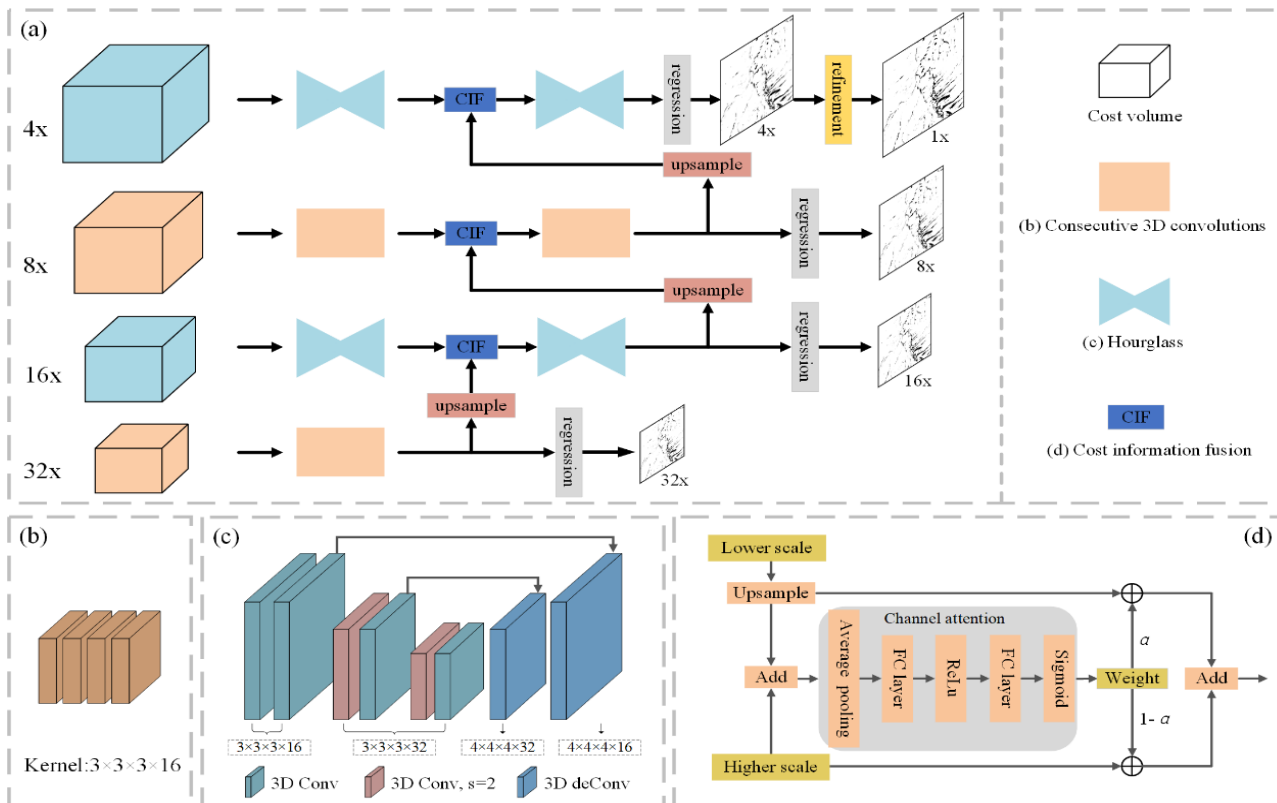


Fig. 6. Illustration of the progressive cost aggregation strategy.

TGRS-2025-09171

global average pooling (GAP) operation followed by a fully connected (FC) layer and an activation function. The weighted cost volumes are then summed to obtain the fused cross-scale cost. This hierarchical aggregation ensures that coarse global context and fine-grained local details are gradually integrated, thus enhancing the model's ability to handle structural diversity and noise in SAR stereo image pairs.

### C. Disparity Regression and Refinement

After progressive aggregation and fusion of the multi-scale cost volumes, we adopt a differentiable soft-argmin approach [40] to regress the disparity value from the cost volume, which enables the model to learn sub-pixel accurate estimation in an end-to-end manner. The softmax function is first applied along the disparity dimension to normalize the cost values into probabilities. The predicted disparity  $\hat{d}$  is then computed as a weighted sum of all candidate disparity levels:

$$\hat{d} = \sum_{d=D_{min}}^{D_{max}} d \cdot P(d) \quad (7)$$

where  $P(d)$  denotes the normalized probability corresponding to disparity  $d$  using softmax function.  $D_{min}$  and  $D_{max}$  define the valid disparity range.

To obtain a detail-preserving disparity output with the same high resolution as the input SAR stereo images, we adopt an edge-preserving refinement strategy inspired by prior works [28, 30], which progressively restores resolution while incorporating structural and edge cues from the left input image. Specifically, the disparity map generated at the highest scale is progressively upsampled to the full image resolution using bilinear interpolation. We then extract both the image intensity and gradient cues of the left image to guide refinement. As shown in Fig. 7, the gradient feature, image intensity, and the upsampled disparity are concatenated and passed through a refinement module composed of six dilated convolutional layers. This module learns a residual disparity map, which is added to the coarse prediction to recover high-frequency details, generating the final refined disparity output. The edge-preserving refinement enhances detail recovery in regions with abrupt elevation changes or blurred textures.

To supervise training across the multi-scale architecture, we compute loss functions at each output scale. For the predicted disparity map, we adopt the Smooth L1 loss [41] to reduce the influence of outliers:

$$L = \frac{1}{N} \sum_{(i,j)} \text{smooth}_{L1}(d_{(i,j)} - \hat{d}_{(i,j)}) \quad (8)$$

$$\text{smooth}_{L1}(z) = \begin{cases} 0.5z^2, & \text{if } |z| < 1 \\ |z| - 0.5, & \text{otherwise} \end{cases} \quad (9)$$

where  $N$  is the number of valid pixels,  $d_{(i,j)}$  is the ground-truth disparity, and  $\hat{d}_{(i,j)}$  is the predicted disparity. The total loss combines the predictions of all scales and the final output:

$$L_{total} = \lambda_4 L_4 + \lambda_3 L_3 + \lambda_2 L_2 + \lambda_1 L_1 + \lambda L \quad (10)$$

where  $\lambda_i$  are weight coefficients that balance the contributions of losses at different scales. This multi-level supervision enhances stability during training and encourages the network

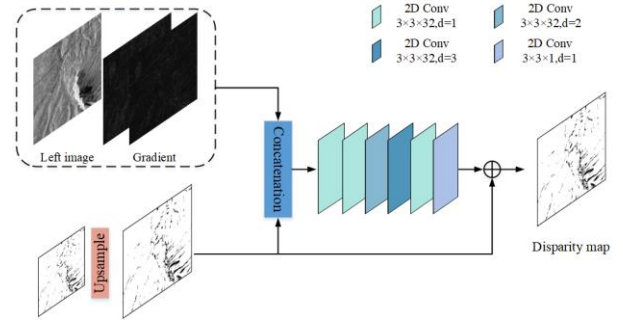


Fig. 7. Illustration of the edge-preserving refinement strategy.

learns to estimate disparities effectively at both coarse and fine resolutions.

### D. DSM Reconstruction from Disparity Maps

Once the disparity map is estimated, the next step is to reconstruct DSM by determining the geographic coordinates of each homologous point. In SAR imaging, each pixel in the image corresponds to a pair of specific range and azimuth imaging times of a ground target. For a matched point across stereo pairs, two independent RD model equations expressed as (1) and (2) can be constructed, each defining the slant-range geometry from a different viewing geometry. Given a stereo pair, each homologous point generates two sets of RD models, resulting in four nonlinear constraint equations. To obtain the precise 3D location of each point, we solve these nonlinear equations using the least squares method, which minimizes the sum of squared residuals between the observed and modeled measurements. Through this process, dense 3D coordinates corresponding to each matching pixel pair are obtained.

To generate the final DSM, the irregular 3D points must be transformed into a gridded surface representation through resampling and interpolation. Resampling ensures consistent pixel spacing in the output DSM, while interpolation fills in missing elevation values to ensure surface continuity. We adopt a triangulation-based interpolation method, which effectively preserves terrain features and supports accurate elevation estimation across complex terrains. The result is a georeferenced high-resolution DSM that provides detailed surface elevation information from SAR stereo pairs.

## IV. EXPERIMENTAL RESULTS AND ANALYSIS

To comprehensively evaluate the effectiveness of deep learning-based stereo matching for SAR imagery, we conducted a series of experiments using the StereoSAR4DSM dataset. We compare the proposed MSSFNet with traditional methods and state-of-the-art deep learning methods, analyzing the resulting disparity maps and reconstructed DSMs.

### A. Experimental Setup

The proposed MSSFNet was trained using the Adam optimizer with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . As a preprocessing step, all input SAR intensity images were normalized to the range  $[-1, 1]$  using min-max scaling per patch to standardize the dynamic range and stabilize training. The disparity maps used as supervised learning targets and network outputs remain in their original pixel units without normalization, preserving the

geometric relationship required for radargrammetric elevation reconstruction. To preserve spatial and structural information that is critical to disparity estimation, the input image patches were kept at their original resolution of 1024×1024 pixels without resizing. The model was trained from scratch for 70 epochs on the StereoSAR4DSM dataset. The initial learning rate was set to 0.001 and reduced by half every 10 epochs. A disparity search range of [−64, 64] was adopted based on the dataset’s disparity distribution characteristics. Loss function weights were empirically set to  $\lambda_4 = 0.3$ ,  $\lambda_3 = 0.5$ ,  $\lambda_2 = 0.7$ ,  $\lambda_1 = 1.0$ , and  $\lambda = 0.6$  following common practice in hierarchical stereo matching networks (e.g., PSMNet and HMSMNet) to emphasize coarse-to-fine supervision. Training was conducted on an Ubuntu 20.04 system using the TensorFlow framework and an NVIDIA RTX 4090 GPU.

To comprehensively evaluate the performance of the proposed framework, this study uses the disparity evaluation metric and the DSM evaluation metric to evaluate the disparity map and the final radargrammetric DSM, respectively. For the DSM evaluation metrics, we employ three statistical indicators: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and 90% Linear Error (LE90) to evaluate the accuracy of the reconstructed DSMs. For the disparity evaluation metrics, we adopt two commonly used metrics in optical remote sensing: End-Point Error (EPE) and the percentage of erroneous pixels (D1). They are defined as:

$$EPE = \frac{1}{N} \sum_{i=1}^N |d_{est}(i) - d_{gt}(i)| \quad (11)$$

$$D1 = \frac{1}{N} \sum_{i=1}^N 1(|d_{est}(i) - d_{gt}(i)| > \delta) \quad (12)$$

where  $N$  is the number of valid pixels, and  $d_{est}$  and  $d_{gt}$  represent the estimated and ground-truth disparities, respectively. Due to the 5×5 multi-looking applied to the SAR data, a threshold  $\delta=0.6$  is used for  $D1$  based on the typical value of 3 in optical stereo benchmarks.

To validate the effectiveness of MSSFNet, we compare it against seven representative stereo matching methods: NCC, SGM [17], StereoNet [28], PSMNet [29], HMSMNet [30] and STTR [31]. NCC and SGM are classical methods widely adopted in stereo matching for their high efficiency and stability. StereoNet and PSMNet are representative deep learning approaches developed for computer vision. HMSMNet is a hierarchical multi-scale architecture specifically designed for satellite image matching and has shown excellent performance on optical datasets. STTR is included as a Transformer-based stereo matching baseline, which captures long-range dependencies via attention mechanisms.

### B. Results on SM Hill

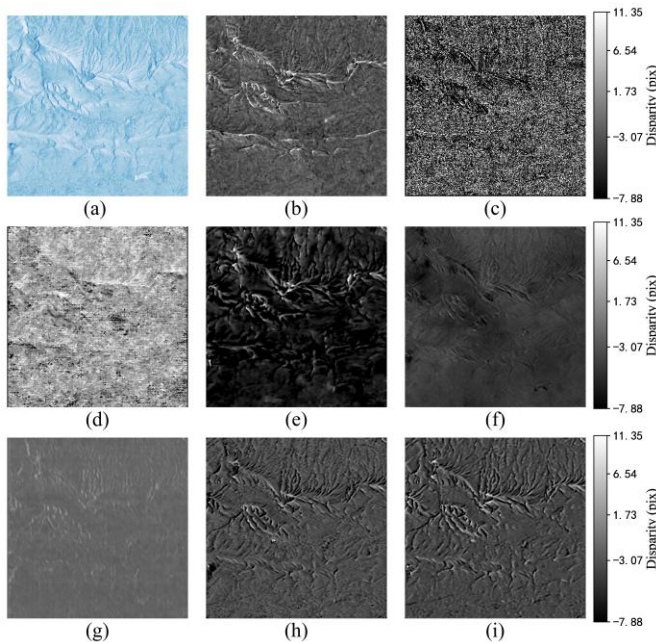
This section presents the results of experiments conducted on the SM Hill test area, which has similar imaging geometry and viewpoints as the training dataset, but is geographically separated. As reported in Table III, the disparity error of traditional methods is significantly higher. The NCC algorithm based on local window matching has difficulty in handling the radiation and geometric distortion unique to SAR, resulting in

a  $D1$  error of up to 79.49%. SGM introduces a global smoothness constraint to enhance robustness, but its performance is still insufficient with an EPE of 1.07 pixels. The deep learning-based models significantly outperform classical methods in disparity accuracy. HMSMNet and the proposed MSSFNet achieve the best accuracy, with MSSFNet reporting the EPE of 0.26 pixels and  $D1$  error of 6.64%. A visual comparison of disparity maps (Fig. 8) further highlights the advantages of deep learning methods, which is consistent with the quantitative improvements shown in Table III. Specifically, deep learning methods produce disparity maps with higher spatial continuity, improved noise suppression, and stronger structural fidelity. Although PSMNet uses spatial pyramid pooling for multi-scale feature extraction, the lack of effective cross-scale cost aggregation limits its accuracy. STTR employs the Transformer architecture to capture long-range dependencies for pixel matching. However, inherent speckle noise in SAR images may interfere with attention calculation, leading to suboptimal matching performance in areas with topographic edges and weak textures. Among deep learning-based methods, HMSMNet and MSSFNet adopt multi-scale architectures with progressive cost fusion strategies. These designs effectively balance the extraction of coarse terrain structures and fine-grained details, thereby improving accuracy.

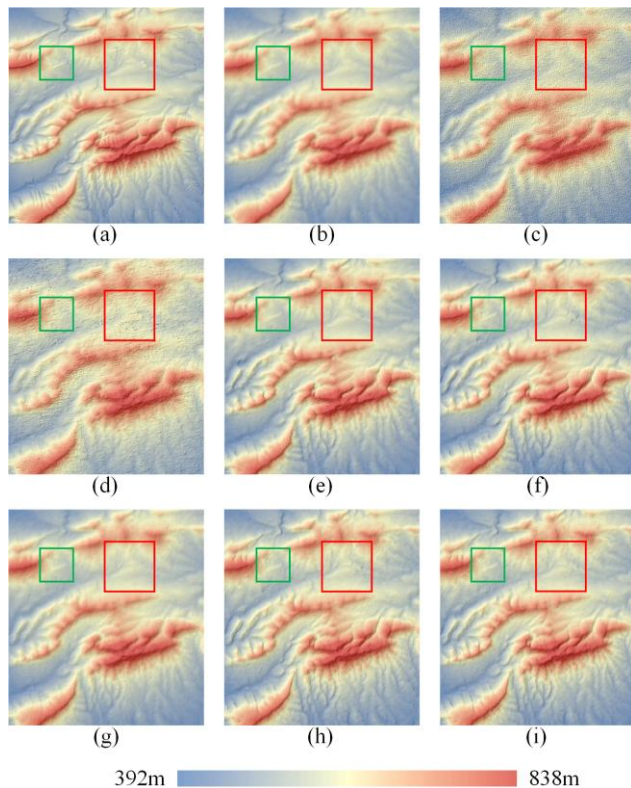
As shown in Table III, the accuracy of the DSM generated by the deep learning model is better than that generated by the traditional image matching algorithm, which is consistent with the conclusion of the disparity estimation accuracy. Among all methods, MSSFNet achieves the best overall performance, with an RMSE of 4.29 m, MAE of 3.27 m, and LE90 of 6.73 m. These quantitative advantages are visually confirmed in the hillshaded DSMs shown in Fig. 9. As shown in a typical area highlighted by red and green boxes in Fig. 9, deep learning methods (especially MSSFNet) are able to capture more terrain details and preserve the structure more effectively than classical approaches. These improvements are mainly attributed to the multi-scale fusion of cost volumes and the refinement strategy combining SAR intensity and gradient cues. Overall, the results on SM Hill demonstrate that MSSFNet achieves robust and accurate disparity estimation under imaging conditions similar to the training data. This leads to high-quality DSM reconstruction with improved terrain detail preservation and elevation accuracy.

TABLE III  
ACCURACY EVALUATION OF DISPARITY MAPS AND RECONSTRUCTED DSMs IN THE SM HILL AREA

Method	Disparity		Stereo DSM		
	EPE(pix)	$D1$ (%)	RMSE(m)	MAE(m)	LE90(m)
SRTM	\	\	4.66	3.38	7.0
NCC	1.36	79.49	15.28	12.67	24.12
SGM	1.07	78.77	18.19	15.52	28.22
StereoNet	0.36	12.64	5.26	3.86	8.12
PSMNet	0.38	17.31	5.65	4.20	8.78
STTR	0.62	39.16	7.31	6.24	11.20
HMSMNet	<b>0.25</b>	6.70	4.39	3.35	6.85
MSSFNet	0.26	<b>6.64</b>	<b>4.29</b>	<b>3.27</b>	<b>6.73</b>



**Fig. 8.** Visual comparison of disparity maps generated by different methods on the SM Hill test area. (a) the left SAR image, (b) ground-truth disparity, (c) NCC, (d) SGM, (e) PSMNet, (f) StereoNet, (g) STTR, (h) HMSMNet and (i) MSSFNet.



**Fig. 9.** Hillshade visualizations of DSMs reconstructed by different methods in the SM Hill test area. The red and green boxes highlight typical areas to emphasize differences in terrain detail recovery. (a) reference DSM, (b) SRTM DEM, (c) NCC, (d) SGM, (e) PSMNet, (f) StereoNet, (g) STTR, (h) HMSMNet and (i) MSSFNet.

### C. Results on SL Mountain

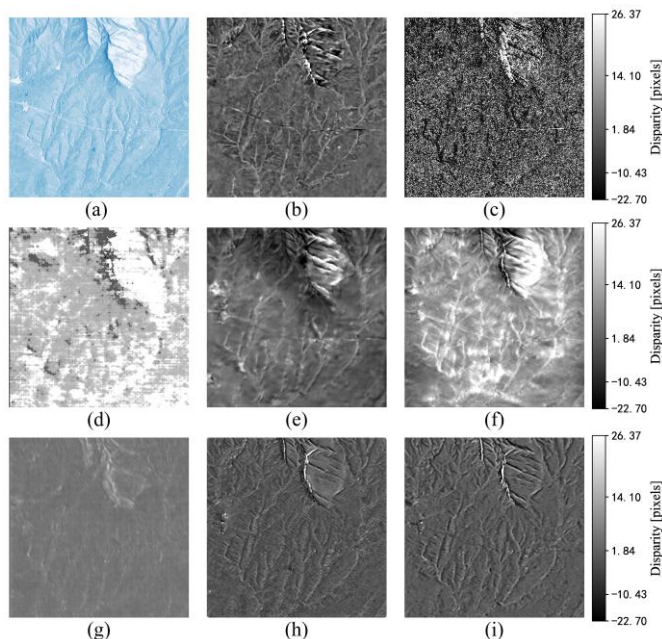
To evaluate the generalization ability of SAR stereo matching models under different imaging conditions, we

conducted experiments in the SL Mountain test area. Unlike the SM Hill region, SL Mountain was imaged using a different SAR acquisition mode and observation geometry, resulting in severe terrain-induced distortions and steep mountainous features. These challenging conditions test the robustness of disparity estimation models when applied to previously unseen data. Table IV presents the quantitative accuracy of the disparity and DSM results, while Fig. 10 and Fig. 11 provide qualitative comparisons of the disparity maps and hillshaded DSMs, respectively.

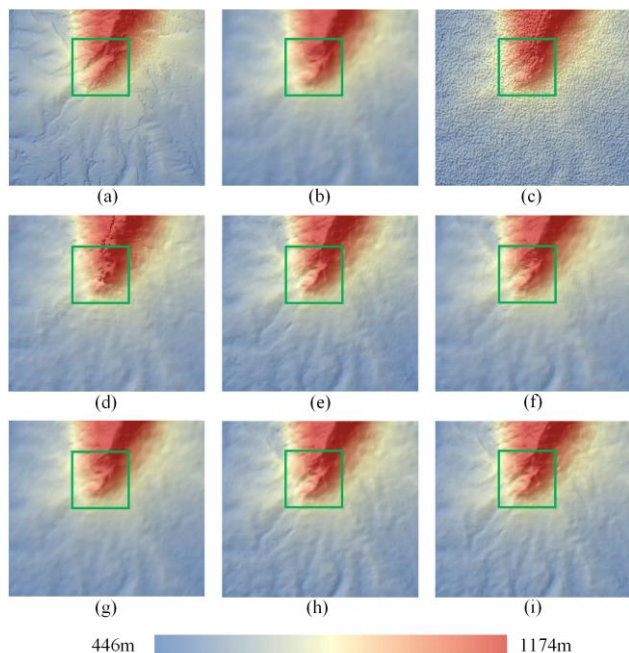
As reported in Table IV, compared to the SM Hill area, the performance of all methods decreases in SL Mountain due to the increased geometric challenges. However, deep learning methods still outperform traditional baselines and show better generalization ability under cross-domain conditions. Among the deep learning-based approaches, StereoNet and PSMNet show the most significant performance degradation. StereoNet's  $D1$  error increases by more than 50%, indicating that the mismatch is more severe in textureless regions. This may be due to its reliance on a single-scale feature representation that is difficult to adapt to different SAR imaging conditions. PSMNet's EPE also drops significantly, indicating that its cost aggregation is not robust enough to domain shift due to insufficient cross-scale fusion. By modeling global relationships among pixels and using relative position encoding, STTR is insensitive to variations in data distribution across different regions, resulting in relatively stable matching outcomes. The proposed MSSFNet achieves the best performance among all methods in both disparity accuracy and DSM reconstruction. Although its accuracy is slightly lower than in the SM Hill test area, it still outperforms all baselines, which confirms the effectiveness of its progressive cost fusion strategy and attention-based refinement in handling structural variations. The quantitative results in Table IV and the clearer ridge structures visible in Fig. 10 and Fig. 11 demonstrate its strong generalization capability.

TABLE IV  
ACCURACY EVALUATION OF DISPARITY MAPS AND RECONSTRUCTED DSMs IN THE SL MOUNTAIN AREA

Method	Disparity		Stereo DSM		
	EPE(pixel)	$D1$ (%)	RMSE(m)	MAE(m)	LE90(m)
SRTM	\	\	5.56	3.58	7.93
NCC	1.27	78.91	20.07	17.27	29.82
SGM	0.94	88.45	23.10	17.33	37.79
StereoNet	0.76	61.68	12.21	8.70	16.61
PSMNet	0.65	44.99	7.53	5.07	10.22
STTR	0.69	51.82	9.61	8.01	13.81
HMSMNet	0.26	9.00	5.92	4.64	9.40
MSSFNet	<b>0.25</b>	<b>8.89</b>	<b>5.43</b>	<b>3.50</b>	<b>7.85</b>



**Fig. 10.** Visual comparison of disparity maps generated by different methods on the SL Mountain test area. (a) the left stereo image, (b) ground-truth disparity, (c) NCC, (d) SGM, (e) PSMNet, (f) StereoNet, (g) STTR, (h) HMSMNet and (i) MSSFNet.



**Fig. 11.** Hillshade visualizations of DSMs reconstructed by different methods in the SL Mountain test area. The red and green boxes highlight typical areas to emphasize differences in terrain detail recovery. (a) reference DSM, (b) SRTM DEM, (c) NCC, (d) SGM, (e) PSMNet, (f) StereoNet, (g) STTR, (h) HMSMNet and (i) MSSFNet.

## V. DISCUSSION

### A. Effectiveness of SAR-driven enhancement strategies

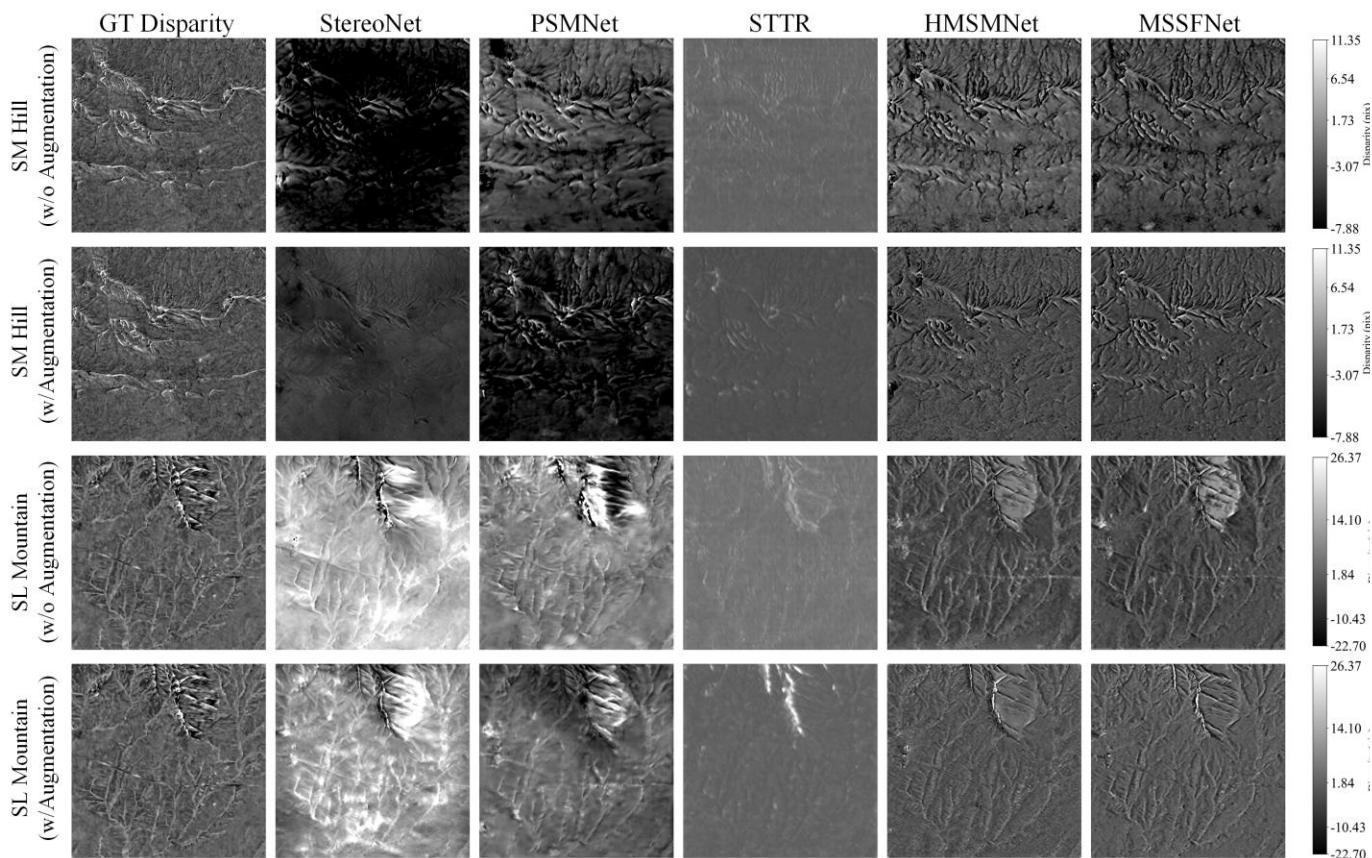
To improve the performance of deep learning models in SAR stereo matching, we introduce a set of SAR-driven enhancement strategies based on the imaging characteristics of SAR data. These strategies are designed to enrich the diversity

of training data by addressing key challenges such as speckle noise, local texture variation, and elevation prior inaccuracies. To evaluate the effectiveness of the proposed enhancement strategies, we compared the disparity estimation and DSM reconstruction results with and without enhancement strategies for the deep learning methods. Before augmentation, the dataset contained only 32 SAR stereo image pairs in  $5 \times 5$  multi-looking for training. Due to the limited availability of SAR-specific training data, the deep learning models were pre-trained on optical US3D stereo matching datasets and then fine-tuned using this limited StereoSAR dataset. After applying the SAR-driven enhancement strategies, the dataset was expanded to nearly 1,000 samples, significantly increasing both its diversity and capacity to support effective model training.

Table V reports the quantitative evaluation results for SM Hill and SL Mountain, while Fig. 12 shows the qualitative comparison of disparity maps. As shown in Table V, all models benefit from the augmentation strategies, and achieve more significant improvements. In particular, for deep learning methods such as StereoNet and PSMNet, the disparity accuracy is improved by more than 60%. The results of STTR were relatively stable, with only minor improvements. More advanced models, including HMSMNet and MSSFNet, already achieve high baseline accuracy and still benefit from further performance gains under augmented training. MSSFNet achieves an EPE of 0.26 pixels and RMSE of 4.29 m in the SM Hill region after applying SAR-driven data augmentation, consistently outperforming all baseline methods. In the more challenging SL Mountain area, it maintains strong performance with an EPE of 0.25 pixels and an RMSE of 5.43 m, demonstrating excellent generalization capabilities under different imaging conditions. The visual results in Fig. 12 further show that disparity maps after augmentation are more consistent, with suppressed speckle noise and clearer edge transitions. Overall, these results confirm the effectiveness of the proposed SAR-driven enhancement strategies in improving the accuracy of disparity estimation and DSM under different SAR imaging conditions.

### B. Ablation experiments

To evaluate the contribution of each component in MSSFNet, we conducted a series of ablation experiments by selectively removing key modules from the full model in this section. Specifically, we analyzed the effectiveness of three architectural strategies: (1) multi-scale progressive cost aggregation, (2) channel attention-based cross-scale cost fusion (CIF), and (3) the disparity refinement module. Four network variants were constructed by selectively removing specific modules from the complete MSSFNet pipeline. Net-v1 is a simplified single-scale model that excludes multi-scale cost aggregation and adaptive fusion, and only retains the disparity refinement module. Net-v2 removes the CIF module responsible for adaptive fusion of adjacent scales, while retaining progressive cost aggregation and the refinement module. Net-v3 removes the disparity refinement module but retains the multi-scale cost aggregation and CIF. MSSFNet is the complete version that incorporates all three modules.



**Fig. 12.** Visual comparison of disparity maps generated by different methods before and after SAR-driven data enhancement in SM Hill and SL Mountain.

TABLE V  
PERFORMANCE COMPARISON BEFORE AND AFTER THE SAR-DRIVEN ENHANCEMENT STRATEGY FOR TWO TEST AREAS

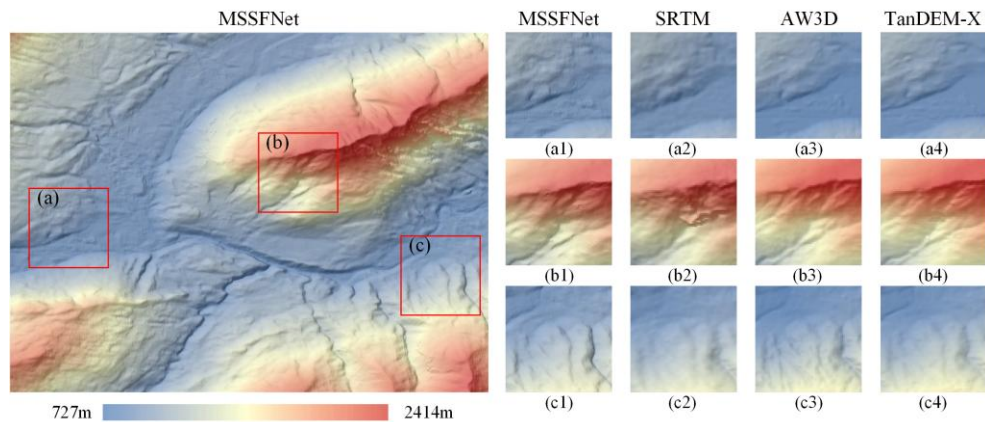
Method	Augmentation	SM Hill					SL Mountain				
		EPE (pix)	D1 (%)	RMSE (m)	MAE (m)	LE90 (m)	EPE (pix)	D1 (%)	RMSE (m)	MAE (m)	LE90 (m)
StereoNet	w/o	0.84	71.57	10.72	8.65	16.00	1.21	90.01	14.78	12.08	21.82
	w/	0.36	12.64	5.26	3.86	8.12	0.76	61.68	12.21	8.70	16.61
PSMNet	w/o	0.68	53.63	8.00	5.88	11.74	0.88	38.31	15.86	10.10	21.94
	w/	0.38	17.31	5.65	4.20	8.78	0.65	44.99	7.53	5.07	10.22
STTR	w/o	0.71	50.12	8.44	6.53	11.94	0.74	53.48	10.53	8.98	15.01
	w/	0.62	39.16	7.31	6.24	11.20	0.69	51.82	9.61	8.01	13.81
HMSMNet	w/o	0.30	10.74	4.71	3.54	7.35	0.34	16.64	6.84	5.33	9.38
	w/	0.25	6.70	4.39	3.35	6.85	0.26	9.00	5.92	4.64	9.40
MSSFNet	w/o	0.29	10.16	5.94	4.73	9.31	0.33	15.90	6.52	5.06	8.99
	w/	0.26	6.64	4.29	3.27	6.73	0.25	8.89	5.43	3.50	7.85

Quantitative evaluation results on SM Hill and SL Mountain are listed in Table VI, highlighting the performance impact of each module. Net-v1 shows the weakest performance, confirming the importance of multi-scale processing. While Net-v2 and Net-v3 achieve better accuracy, both suffer in either geometric consistency or detail recovery when compared with the full model. The refinement module plays a crucial role in preserving terrain structure and suppressing residual noise, as evidenced by the significant performance improvement of MSSFNet over Net-v3. These results clearly demonstrate that

the full MSSFNet achieves the lowest error metrics in both regions, validating the effectiveness of the three modules.

TABLE VI  
QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS ON MSSFNET

Method	Progressive Cost Aggregation	CIF Refinement	SM Hill		SL Mountain	
			EPE (pixel)	D1 (%)	EPE (pixel)	D1 (%)
Net-v1		✓	1.12	72.69	1.54	90.14
Net-v2	✓	✓	0.30	6.98	0.27	10.68
Net-v3	✓	✓	0.32	7.16	0.42	30.12
MSSFNet	✓	✓	<b>0.26</b>	<b>6.64</b>	<b>0.25</b>	<b>8.89</b>



**Fig. 13.** The DSM results generated from Sentinel-1 stereo pair using the proposed MSSFNet and comparison with SRTM, AW3D and TanDEM-X DEM in three representative subregions (a–c). Subplots (a1–c1) show DSMs reconstructed by MSSFNet, (a2–c2) the corresponding SRTM DEMs, (a3–c3) the AW3D DEMs, and (a4–c4) the TanDEM-X DEMs.

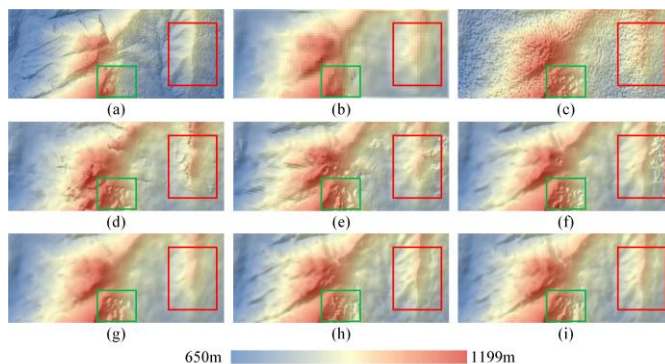
### C. Generalization across different regions and SAR sensors

To further evaluate the generalization capability of the proposed MSSFNet across different regions and SAR sensors, we conducted cross-region and cross-platform experiments using Sentinel-1 SAR imagery. Specifically, a pair of stripmap-mode Sentinel-1B images acquired on 12 February and 17 February 2017 over the European Alps were selected. The selected region features diverse topography including mountain peaks, hilly terrain and plains, with elevations ranging from 727 m to 2414 m and covering an area of approximately 12 km × 9 km. In the experiments, we applied the model trained on the StereoSAR4DSM dataset (derived from TerraSAR-X imagery) directly to the Sentinel-1B image pair without fine-tuning to assess its out-of-domain generalization ability. Since no high-resolution ground truth DSM is available for this region, we used AW3D [42] and TanDEM-X 30m Edited DEM [43] as reference benchmarks to evaluate the performance of the reconstructed DSMs. The accuracy is summarized in Table VII. It should be noted that both AW3D and TanDEM-X are global-scale elevation products and contain inherent uncertainties, particularly in complex mountainous terrain. To provide a reference baseline, the discrepancy between AW3D and TanDEM-X was also analyzed, yielding an RMSE of 6.28 m, MAE of 3.61 m, and LE90 of 7.9 m. This indicates that meter-level differences already exist between these two state-of-the-art DEM products. In this context, the quantitative results reported in Table VII are considered acceptable. The DSM result generated by MSSFNet is shown in Fig. 13. We show the hillshaded terrain generated by MSSFNet and the corresponding SRTM, AW3D and TanDEM-X data in the selected area (a)–(c). As shown in Fig. 13, the reconstructed DSM by MSSFNet remains consistent with AW3D and TanDEM-X reference products, and successfully captures more detailed and continuous terrain features in different landforms, compared with SRTM, including clearer ridge lines in mountainous areas and finer textures in transition slopes. These results demonstrate that MSSFNet, trained solely on TerraSAR-X data, can generalize well to Sentinel-1 data from different regions.

TABLE VII

QUANTITATIVE EVALUATION OF DSM GENERATED BY MSSFNET USING SENTINEL-1B DATA, COMPARED WITH AW3D AND TANDEM-X

Reference DEM	REFERENCE PRODUCTS		
	RMSE(m)	MAE(m)	LE90(m)
AW3D	13.76	9.98	21.73
TanDEM-X	10.76	8.14	17.41



**Fig. 14.** Hillshade visualizations of DSMs reconstructed by different methods in a challenging mountainous area. The red and green boxes highlight typical layover- and steep slope-dominated regions. (a) reference DSM, (b) SRTM DEM, (c) NCC, (d) SGM, (e) PSMNet, (f) StereoNet, (g) STTR, (h) HMSMNet and (i) MSSFNet.

### D. Strengths and Limitations of MSSFNet

Based on the experimental results in both SM Hill and SL Mountain regions, MSSFNet demonstrates clear advantages in SAR stereo matching, including improved spatial continuity, enhanced noise suppression, and better preservation of terrain structures compared with classical and existing deep learning baselines. To further analyze limitations of MSSFNet in challenging terrain, Fig. 14 presents a representative mountainous area characterized by severe foreshortening, layover, and shadowing effects. In these regions, MSSFNet produces DSMs that are visually closest to the reference overall, with clearer ridge lines and more continuous terrain surfaces than other methods. However, some limitations remain in areas dominated by extreme SAR imaging distortions. As shown in the red-boxed layover regions of Fig. 14, strong signal superposition and compressed geometry result in systematic

TGRS-2025-09171

height overestimation and localized elevation artifacts in the reconstructed DSM. In the green-boxed regions, steep slopes and foreshortening cause complex multi-layered terrain to be compressed into very few pixels or completely lost, preventing the correct reconstruction of very steep, step-like terrain structures for all methods, including MSSFNet. These effects reflect fundamental physical constraints of stereo SAR imaging. While MSSFNet mitigates many of these issues through multi-scale fusion and contextual inference, accurate reconstruction in such blind zones remains challenging. Future work will therefore focus on multi-view stereo fusion to further reduce geometry-induced information loss and improve DSM reconstruction in such challenging regions.

## VI. CONCLUSION

This study proposes a deep learning framework for radargrammetric DSM generation from SAR stereo imagery. To enable accurate disparity estimation, we constructed the StereoSAR4DSM dataset based on TerraSAR-X SAR imagery and high-resolution aerial DSMs. The dataset integrates enhanced epipolar rectification with prior DEMs and introduces three SAR-driven data enhancement strategies (multi-looking variation, random pixel sampling, and elevation perturbation), which enrich data diversity and simulate real-world imaging variability. To fully exploit this dataset, we design a task-specific deep network, namely Multi-Scale Stereo SAR Fusion Network (MSSFNet). MSSFNet leverages multi-scale feature extraction and pyramid cost volume construction for coarse-to-fine matching. It employs a progressive cost fusion strategy guided by a channel attention-based fusion module and a residual refinement block that integrates SAR intensity and gradient cues to restore fine structural details. These modules enhance robustness to speckle noise and improve matching consistency across different terrains and imaging conditions. Experiments conducted on two different test areas confirm that MSSFNet outperforms classical and state-of-the-art deep learning methods in terms of disparity estimation and DSM accuracy. Overall, this work demonstrates the strong potential of deep learning in radargrammetric applications and provides a high-quality dataset and network architecture to support future research in StereoSAR-based DSM reconstruction.

## REFERENCES

- [1] Y. Qin *et al.*, "High-precision flood mapping from Sentinel-1 dualpolarization SAR data," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.
- [2] E. M. Sanz, M. Stefko, and I. Hajsek, "DEM-assisted 3D reconstruction of Aletsch glacier displacements using monostatic and bistatic differential interferometry," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.
- [3] N. Kariminejad, M. Jafari, F. Domazetović, and A. Quesada-Román, "An Overview of the Importance of DEM Resolution in Soil Erosion Assessment," *Papers in Applied Geography*, vol. 10, no. 3, pp. 207-216, 2024.
- [4] T. Toutin and L. Gray, "State-of-the-art of elevation extraction from satellite SAR data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 55, no. 1, pp. 13-33, 2000.
- [5] H. Jiang, L. Zhang, Y. Wang, and M. Liao, "Fusion of high-resolution DEMs derived from COSMO-SkyMed and TerraSAR-X InSAR datasets," *Journal of Geodesy*, vol. 88, pp. 587-599, 2014.
- [6] J. H. Yu, L. Ge, and X. Li, "Radargrammetry for digital elevation model generation using Envisat reprocessed image and simulation image," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 9, pp. 1589-1593, 2014.
- [7] U. S. Guimarães, I. da Silva Narvaes, M. d. L. B. T. Galo, A. d. Q. da Silva, and P. de Oliveira Camargo, "Radargrammetric approaches to the flat relief of the amazon coast using COSMO-SkyMed and TerraSAR-X datasets," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 284-296, 2018.
- [8] R. Palamà *et al.*, "Radargrammetry DEM generation using high-resolution SAR imagery over La Palma during the 2021 Cumbre Vieja volcanic eruption," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1-5, 2023.
- [9] J. Wang, H. Chai, X. Li, and X. Lv, "Improving Mountainous DSM Accuracy Through an Innovative Opposite-Side Radargrammetry Algorithm," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.
- [10] T. Balz, L. Zhang, and M. Liao, "Direct stereo radargrammetric processing using massively parallel processing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 79, pp. 137-146, 2013.
- [11] Y. Wang, Q. Yu, and W. Yu, "An improved Normalized Cross Correlation algorithm for SAR image registration," in *2012 IEEE International Geoscience and Remote Sensing Symposium*, 2012, pp. 2086-2089: IEEE.
- [12] S. Méric, F. Fayard, and É. Pottier, "A multiwindow approach for radargrammetric improvements," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 10, pp. 3803-3810, 2011.
- [13] P. Schwind, S. Suri, P. Reinartz, and A. Siebert, "Applicability of the SIFT operator to geometric SAR image registration," *International Journal of Remote Sensing*, vol. 31, no. 8, pp. 1959-1980, 2010.
- [14] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, and F. Tupin, "SAR-SIFT: a SIFT-like algorithm for SAR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 1, pp. 453-466, 2014.
- [15] S. Xiaotian, Z. Guo, and W. Xia, "High-precision DEM production for spaceborne stereo SAR images based on SIFT matching and region-based least squares matching," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 39, pp. 49-53, 2012.
- [16] Y. Dong, L. Zhang, T. Balz, H. Luo, and M. Liao, "Radargrammetric DSM generation in mountainous areas through adaptive-window least squares matching constrained by enhanced epipolar geometry," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 137, pp. 61-72, 2018.
- [17] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans Pattern Anal Mach Intell*, vol. 30, no. 2, pp. 328-41, Feb 2008.
- [18] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, 2005, vol. 2, pp. 807-814: IEEE.
- [19] M. Di Rita, A. Nascetti, F. Fratarcangeli, and M. G. Crespi, "Upgrade of FOSS DATE plug-in: implementation of a new radargrammetric DSM generation capability," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 41, no. B7, pp. 821-825, 2016.
- [20] J. Wang *et al.*, "Radargrammetric DSM generation by semi-global matching and evaluation of penalty functions," *Remote Sensing*, vol. 14, no. 8, p. 1778, 2022.

TGRS-2025-09171

- [21] H. Raggam and A. Almer, "Assessment of the potential of JERS-1 for relief mapping using optical and SAR data," *International Archives of Photogrammetry and Remote Sensing*, vol. 31, pp. 671-676, 1996.
- [22] M. Wang, F. Hu, and J. Li, "Epipolar resampling of linear pushbroom satellite imagery by a new epipolarity model," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 3, pp. 347-355, 2011.
- [23] A. Schubert, D. Small, E. Meier, and D. Nuesch, "Robustness of wavelet-based stereo matching for variable acquisition geometries using simulated SAR images," in *IEEE International Geoscience and Remote Sensing Symposium*, 2002, vol. 5, pp. 2759-2761: IEEE.
- [24] A. Fusiello and L. Irsara, "Quasi-euclidean uncalibrated epipolar rectification," in *2008 19th International Conference on Pattern Recognition*, 2008, pp. 1-4: IEEE.
- [25] T. Toutin, "Impact of Radarsat-2 SAR ultrafine-mode parameters on stereo-radargrammetric DEMs," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 10, pp. 3816-3823, 2010.
- [26] K. Gutjahr, R. Perko, H. Raggam, and M. Schardt, "The epipolarity constraint in stereo-radargrammetric DEM generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 8, pp. 5014-5022, 2014.
- [27] R. Perko, K. Gutjahr, M. Krüger, H. Raggam, and M. Schardt, "DEM-based epipolar rectification for optimized radargrammetry," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017, pp. 969-972: IEEE.
- [28] S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, and S. Izadi, "Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 573-590.
- [29] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5410-5418.
- [30] S. He, S. Li, S. Jiang, and W. Jiang, "HMSM-Net: Hierarchical multi-scale matching network for disparity estimation of high-resolution satellite stereo images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 188, pp. 314-330, 2022.
- [31] Z. Li *et al.*, "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6197-6206.
- [32] E. Rodriguez, C. S. Morris, and J. E. Belz, "A global assessment of the SRTM performance," *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 3, pp. 249-260, 2006.
- [33] A. F. Habib, M. F. Morgan, S. Jeong, and K. O. Kim, "Epipolar geometry of line cameras moving with constant velocity and attitude," *ETRI journal*, vol. 27, no. 2, pp. 172-180, 2005.
- [34] J. C. Curlander, "Location of spaceborne SAR imagery," *IEEE Transactions on Geoscience and Remote Sensing*, no. 3, pp. 359-364, 1982.
- [35] C. Oliver and S. Quegan, *Understanding synthetic aperture radar images*. SciTech Publishing, 2004.
- [36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
- [37] R. Chabira, J. Straub, C. Sweeney, R. Newcombe, and H. Fuchs, "StereoDRNet: Dilated Residual StereoNet," presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
- [38] G. Xu, J. Cheng, P. Guo, and X. Yang, "Attention concatenation volume for accurate and efficient stereo matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12981-12990.
- [39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132-7141.
- [40] A. Kendall *et al.*, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 66-75.
- [41] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440-1448.
- [42] T. Tadono, J. Takaku, K. Tsutsui, F. Oda, and H. Nagai, "Status of "ALOS World 3D (AW3D)" global DSM generation," in *2015 IEEE international geoscience and remote sensing symposium (IGARSS)*, 2015, pp. 3822-3825: IEEE.
- [43] C. González, M. Bachmann, J.-L. Bueso-Bello, P. Rizzoli, and M. Zink, "A fully automatic algorithm for editing the TanDEM-X global DEM," *Remote Sensing*, vol. 12, no. 23, p. 3961, 2020.



**Yuting Dong** received her PhD in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, China, in 2018. From February 2019 to November 2020, she worked as a post-doc at the German Aerospace Center (DLR). She received an Alexander von Humboldt Research Fellowship in 2022. She currently works at the China University of Geosciences. Her research interests include remote sensing data processing and InSAR topographic mapping.

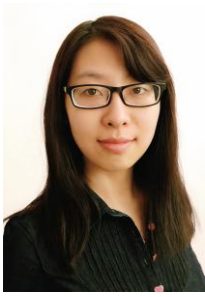


**Yaozu Li** graduated from Chang'an University with a bachelor's degree in science in 2023. He is currently pursuing a master's degree in surveying and mapping engineering at the School of Geography and Information Engineering, China University of Geosciences (Wuhan). His main research interests include stereoscopic radargrammetry.



**Ji Zhao** received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, China, in 2017. He is now an associate professor at the School of Computer Science, China University of Geosciences, Wuhan. His major research interests include remote sensing image classification, scene analysis, remote sensing applications, and machine learning algorithms.

TGRS-2025-09171



**Yao Sun** is a Research Scientist at Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. She received the bachelor's degree in cartography and geo-information system from Wuhan University, Wuhan, China, in 2012, and the M.Sc. and Dr.-Ing. degrees from the Technical University of Munich (TUM), Munich, Germany, in 2016 and 2021,

respectively. She was a Research Fellow with DLR-IMF from 2016 to 2022, and a Research Scientist with TUM from 2021 to 2024. In 2024, she was a Guest Scientist with Radar and Optical Remote Sensing for Geohazards Group at GFZ Helmholtz Centre for Geosciences, Potsdam, Germany.

Her research interests include multi-modal data integration, disaster management, artificial intelligence for Earth observation, and Open Science practices for equitable geospatial information access.



**Mingsheng Liao** received the B.S. degree in electronic engineering from the Wuhan Technical University of Surveying and Mapping (WTUSM), Wuhan, China, in 1982, the M.A. degree in electronic and information engineering from the Huazhong University of Science and Technology, Wuhan, in 1985, and the Ph.D. degree in

photogrammetry and remote sensing from WTUSM, in 2000.

He is the Principal Investigator of several projects funded by the Ministry of Science and Technology (MOST), China, and the National Natural Science Foundation of China. Since 1997, he has been a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan. Since 2004, he has been the Principal Investigator of the ESA-MOST Cooperative Dragon project. His research interests include algorithms for interferometric synthetic aperture radar, integration and fusion of multisource spatial information, and applications of remote sensing data.