SpaceNet 9—Cross-Sensor Alignment of Optical and SAR Imagery

Ronny Hänsch, Jacob Arndt, Abhishek Potnis, Philipe Dias, Peter Novotný, Fabio Pacifici, and Todd M. Bacastow

***Abstract*—Precise registration of high-resolution synthetic aperture radar (SAR) and optical imagery is necessary for realizing the full potential and benefits of multimodal image analysis. However, two significant challenges presently exist. First, there is a lack of annotated datasets and benchmarks available for high-resolution SAR–optical image registration. Second, an assessment of efficient and reliable image registration methods that can precisely align these modalities is lacking. Here, we present a holistic description of the SpaceNet 9 Challenge and its results. We present a description of the dataset and baseline algorithm along with the results of the challenge, including a description of the winning algorithms. We release the SpaceNet 9 dataset along with open-sourcing the winning algorithms and baseline. The objective of SpaceNet 9 was to compute a dense displacement map that indicates the shift needed to align pixels in an optical image to the pixels in a SAR image. The challenge launched in April 2025 and was active for approximately two months. The top five solutions reduced image alignment error from approximately 34 m to under 13 m for public and private test data, with the best results obtaining a registration error of only 8.5 and 6.7 m on the public testing and private testing dataset, respectively. Usage of pretrained image matching models, robust outlier rejection with RANSAC, and estimating local displacement were common among the top solutions. The results of this challenge provide insight into high-resolution SAR–optical image registration and offer opportunities for future benchmarking in this domain. The baseline algorithm, winning solutions, and datasets are available at https://spacenet.ai/sn9-challenge/.**

***Index Terms*—Benchmark datasets, cross-modal, image registration, multimodal learning, optical, synthetic aperture radar (SAR).**

I. Introduction

SPACENET is a collaborative initiative with the goal of accelerating open-source machine learning for geospatial applications. Since 2016, SpaceNet has provided a repository of openly available satellite imagery with coregistered labels and baseline algorithms. Nine innovative challenges have been hosted as part of this initiative where the winning algorithms for each SpaceNet challenge were open sourced. The SpaceNet consortium consists of contributors from Vantor (previously Maxar Intelligence), AWS, IEEE GRSS, Oak Ridge National Laboratory, the Open Geospatial Consortium, Topcoder, and UMBRA.

Building on its foundational mission, earlier SpaceNet challenges explored key tasks, such as road extraction, building footprint detection, and routing from satellite imagery, each expanding the community's understanding of how machine learning can support real-world geospatial problems. These efforts progressively introduced more complex scenarios and multimodal sensing conditions, setting the stage for challenges that move beyond single-sensor interpretation toward integrated analysis. As the diversity of remote sensing data grows, SpaceNet 9 extends the initiative into the domain of cross-modal image registration, addressing a crucial gap in aligning disparate data sources for time-critical applications, such as disaster response.

Swift and effective disaster response often relies on the integration and analysis of diverse remote sensing data sources, such as optical and synthetic aperture radar (SAR). However, the coregistration of multimodal imagery remains a major challenge due to the inherent differences in acquisition methods and data characteristics. SpaceNet 9 aims to address this limitation by focusing on cross-modal image registration, a critical preprocessing step for disaster analysis and recovery.

The objective of the SpaceNet 9 Challenge was to advance cross-modality image registration enabling better downstream analytics, such as computer vision algorithms for damage assessment, change detection, or feature extraction. The scope of SpaceNet 9 was to create algorithms to compute pixelwise spatial transformations between optical and SAR imagery, specifically in earthquake-affected regions. These algorithms were evaluated for their ability to align tie-points across modalities by leveraging an optical and a SAR image as input and output a two-channel transformation map that transforms the geometry of one into the other.

A dataset covering three urban areas in Türkiye impacted by earthquakes in 2023 consisting of optical and SAR imagery from Vantor and UMBRA's open data programs, respectively, along with roughly 800 hand-labeled tie-points for training and testing (public) created for SpaceNet 9 was released. The tie-points were labeled for distinguishable intersections and features across both modalities, while avoiding features on elevated structures, such as buildings, that do not align due to different look angles and image geometry (i.e., layover in SAR).

Received 20 November 2025; revised 2 February 2026; accepted 13 February 2026. Date of publication 18 February 2026; date of current version 9 April 2026. We acknowledge that this manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. *(Corresponding author: Ronny Hänsch.)*

Ronny Hänsch is with German Aerospace Center (DLR), 82234 Oberpfaffenhofen-Weßling, Germany (e-mail: ronny.haensch@dlr.de).

Jacob Arndt, Abhishek Potnis, and Philipe Dias are with the Geospatial Science and Human Security Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830 USA.

Peter Novotný is with the Faculty of Management Science and Informatics, University of Zilina, Zilina 01026, Slovakia, and also with Topcoder, USA.

Fabio Pacifici and Todd M. Bacastow are with IEEE Geoscience and Remote Sensing Society, USA.

Digital Object Identifier 10.1109/JSTARS.2026.3666103

SpaceNet 9 was hosted from April to May 2025 and had 406 registered participants/teams on Topcoder. Of these teams, 74 participants submitted 690 provisional scored solutions, 573 of which were valid submissions. In total, 35 of the valid submissions were used for final scoring on test-private data to determine the challenge winners.

The results of SpaceNet 9 consist of winning solutions that leverage pretrained keypoint matchers, such as LightGlue, excluded certain buildings, and estimated local offsets instead of performing global transformations. Interesting findings include observing that deep-learning based ad hoc keypoint matchers perform well even for optical–SAR data, building detectors are shown to be sufficiently mature to be leveraged in downstream tasks, and incorporating local height data [digital surface model (DSM) or predicted) helps improve model performance.

II. Related Works

A. SAR–Optical Image Matching Methods

Image registration methods are broadly categorized as area based, feature based, or learning based, with several hybrid variants combining elements of these approaches. An exhaustive review of SAR–optical registration methods is beyond the scope of this article, but can for example be found in [1].

1) Area-Based Methods: Area-based approaches evaluate a similarity metric over a grid (ranging from the dense pixel grid to sparse, nonuniform grids defined by salient structures) in combination with a geometric transformation. To limit computational load, the transformation is usually restricted to a small number of degrees of freedom (e.g., translation only), estimating displacements of a few dozen pixels. For well-georeferenced images, this may be sufficient. However, when geolocation is missing or inaccurate, the limited search range becomes inadequate and the registration process may fail. Larger displacements are generally handled with coarse-to-fine search strategies [2].

A variety of similarity metrics are available from general image registration pipelines, including the sum of absolute differences, the sum of squared differences, normalized cross-correlation, mutual information, and phase correlation based on frequency-domain representations. Many of these metrics are sensitive to nonlinear radiometric differences and thus often fail for SAR–optical matching unless specifically adapted.

Among them, mutual information has been most widely used [3]. It has proven effective for registering high-resolution TerraSAR-X and Ikonos images over urban areas [4], Landsat and ALOS PALSAR images [5], and Landsat with Radarsat imagery [6]. Mutual information has also been combined with feature-based approaches (discussed in Section II-A2) [7], [8].

Other similarity measures are typically only applied in combination with higher level image features that remain comparable across modalities, such as histograms of oriented gradients [9] or histograms of oriented phase congruency [10].

2) Feature-Based Methods: Feature-based image registration aligns multimodal data by detecting and matching distinctive image features. The process typically involves feature detection, description, and matching, with robust performance requiring features that are both prominent and repeatable in SAR and optical imagery. Common choices include closed boundary regions, such as water bodies or forests, as well as salient points, such as corners and intersections. These are encoded into vector descriptors and compared to establish correspondences.

Classical operators, such as SIFT [11], [12], have been widely used, and many adaptations have been proposed for multimodal scenarios. SAR-SIFT [13] modifies gradient calculations to mitigate speckle, while RIFT [14] employs phase congruency-based descriptors to address radiometric differences. Further refinements include ILS-SIFT [12], which integrates level-set segmentation with an improved RANSAC, and OS-SIFT [15], which introduces gradient operators and GLOH-like descriptors for more consistent and stable matches. Other descriptors, such as CFOG [16] and its multiresolution extensions with SAR-based edge masking [17], as well as pipelines combining improved Harris detectors with phase-congruency descriptors [18], have also demonstrated enhanced robustness.

Beyond point-based features, line features have proven effective, particularly under noisy conditions. Edge-based strategies combined with improved shape context matching and affine transformation estimation [19], as well as approaches based on line extraction combined with Hough transforms [20] or intersection matching [21], provide reliable primitives for SAR–optical alignment.

3) Learning-Based Methods: A growing body of work leverages deep learning to bridge the modality gap between SAR and optical imagery, typically by generating modality-invariant representations, synthesizing pseudoimages, or learning adaptive similarity measures to enhance registration accuracy.

GAN-based frameworks have been introduced to reduce the modality gap in optical–SAR registration. By training conditional GANs to generate pseudo-SAR patches from optical data, conventional monomodal methods, such as normalized cross-correlation and SIFT, can be applied more effectively [22]. Another line of work focuses on improving feature discriminability. SFcNet [23] introduces a novel loss function that maximizes the separation between positive and negative samples. SFcNet selects the largest mismatch as the negative sample and the correct point as the positive, thereby enhancing robustness in distinguishing true correspondences. A different strategy is the heterogeneous SuperPoint network [24], which processes full-size images and extracts both interest points and fixed-length descriptors in a single forward pass, enabling efficient and reliable SAR–optical matching.

Most deep learning approaches for direct SAR–optical registration are based on (pseudo-)Siamese networks. These extract features through dual-stream architectures and measure similarity using dot products, convolutional layers, or traditional metrics. Early work introduced a Siamese network for multimodal registration [25], while later pseudo-Siamese variants use independent, nonparameter-sharing streams to account for modality differences [26]. Generative models have also been integrated into this framework. The generative matching network synthesizes pseudo-SAR or pseudooptical patches to train a deep matching model [27], later refined with an end-to-end learning strategy for better efficiency [28].

TABLE I
SUMMARY OF OPTICAL–SAR IMAGE REGISTRATION, MATCHING, AND DATA FUSION DATASETS

Dataset	Task	Sensors	Spatial Resolution (m)	Num. of images	Image size
SEN1-2 [31]	Matching	S-1 versus S-2	10	282 384	256×256
SARptical [32]	Matching	TSX vs aerial UltraCAM	0.2–1	17 680	112×112
SEN12MS [33]	Matching	S-1 versus S-2	10	180 662	256×256
WHU-SEN-City [34]	Matching	S-1, S-2	10	32	1885×1733–8925×4611
OS [35]	Matching	GF-3 versus GE	1	10 692	256×256
QXS-SAROPT [36]	Matching	GF-3 versus GE	1	20 000	256×256
SOPatch [37]	Matching	S-1, GF-3 versus S-2, GE	1, 10	665 600	64×64
OSEval [38]	Registration	GF-3 versus WV-2/3, SV-1, GE	(0.56,0.33) versus 0.3–0.5	1232	1200×3600 versus 5500×5200
SpaceNet 9 (ours)	Registration	Umbra, WV-2	0.31–0.46	3	7192×7477–12322×12304

S-1 and S-2 denote Sentinel-1 and -2, respectively, TSX denotes TerraSAR-X, GF-3 stands for GaoFeng-3, WV for WorldView, SV for Superview, and GE for "Google Earth."[1]

Beyond Siamese and GAN-based approaches, more specialized frameworks have emerged, such as a component-based framework with three subnetworks for goodness evaluation, correspondence estimation, and outlier removal [29]. Another example is MU-Net [30], an unsupervised multiscale architecture that learns transformation parameters directly in a coarse-to-fine manner by combining multiple scales with a loss function based on structural similarity.

4) Existing Datasets: The datasets summarized in Table I represent the main resources currently available for multimodal SAR–optical matching and registration research. A large fraction of them, such as SEN1-2, SEN12MS, QXS-SAROPT, and SOPatch, are patch-based datasets created by pairing SAR (e.g., Sentinel-1 and Gaofen-3) and optical (e.g., Sentinel-2 and "Google Earth"[1] imagery). SOPatch recombines samples from multiple sources (WHU-SEN-City, OSDataset, and SEN1-2). These datasets provide paired samples in the sense that each SAR patch is associated with a corresponding optical patch extracted at the same geolocation. While invaluable for training and benchmarking deep learning methods, for e.g., patch descriptors,[2] it must be noted that the pairs are typically already well aligned through metadata-based orthorectification. Local misregistrations might exist but are generally minor compared to the large geometric and radiometric discrepancies encountered in real-world registration tasks. Furthermore, most of these datasets are (semi)manually curated to filter patch pairs with strong misalignments. The SARptical dataset goes even further in this direction, as it contains TerraSAR-X and aerial optical patches that were coregistered to pixel precision. This makes it an excellent benchmark for testing algorithms under ideal conditions, but at the same time even less representative of real-world scenarios, where large misalignments, terrain-induced distortions, and sensor geometry differences dominate.

Other datasets have been explicitly designed with full-scene images to support registration research. WHU-SEN-City provides city-scale SAR–optical pairs but aligned only through geometric metadata, which results in unaddressed registration errors ranging from a few pixels to several dozen pixels. The more recent OSEval dataset moves closer to realistic evaluation settings but is limited by the very small number of reference tie points, based on light poles visible in both modalities, which are provided per image. This makes it difficult to assess overall registration performance, since misalignments between these sparse points remain unmeasured.

Finally, our proposed SpaceNet 9 dataset is designed to overcome these limitations. By providing very high-resolution SAR and optical image pairs with dense, reliable reference correspondences, it bridges the gap between well-aligned patch-based datasets and sparse, tie point-based evaluation benchmarks. This enables more realistic training and, crucially, more comprehensive and meaningful evaluation of multimodal registration methods.

TABLE II
SPACENET 9 DATASET SPLITS

AOI	City	Section	Split	Area (km^2)	Tiepoints
1	-	1	Private Test	8.42	149
2	Gaziantep	1	Train	3.03	151
2	Gaziantep	2	Public Test	2.51	163
2	Gaziantep	3	Train	3.76	104
3	Adıyaman	1	Train	5.62	161
3	Adıyaman	2	Public Test	2.27	66

III. SN9 DATA AND BASELINE

A. Dataset

Imagery: The SpaceNet 9 dataset consists of high-resolution optical and SAR imagery. The dataset includes scenes captured over three areas of interest (AOIs) located in Türkiye. Two AOIs are selected for public training and public testing, and one AOI is withheld for private testing. The public training and public testing AOIs were divided into sections to provide a challenging and diverse coregistration dataset with variable local and global alignment scenarios. Table II describes the AOIs and the dataset splits. Fig. 1 illustrates the splits and AOIs for the public training and public testing data.

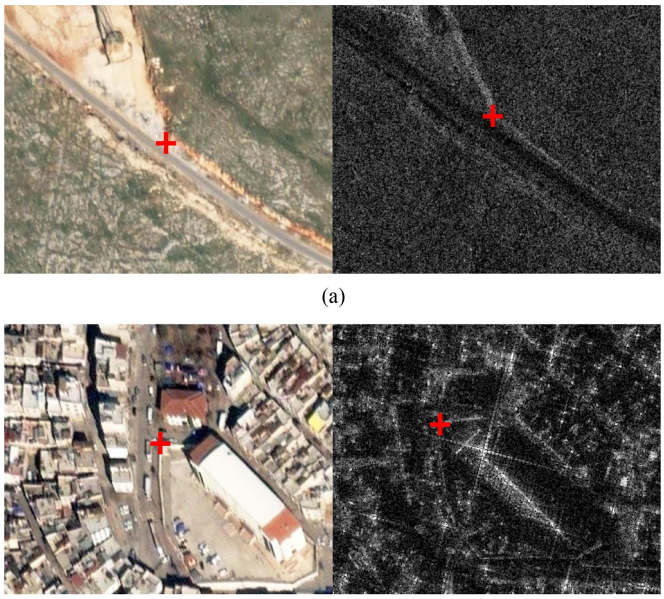
The optical imagery included in the dataset, collected by the WorldView-2 sensor, are 8-bit, three-channel (RGB) scenes with spatial resolution of 0.31 m. The SAR data, collected by Umbra, are 8-bit single channel amplitude images ranging in spatial resolution from 0.36 to 0.45 m. All imagery is stored in the GeoTIFF image format and projected in a UTM coordinate reference system.

[1]While the corresponding literature just lists "Google Earth" as data source, it should be noted that it is not a sensing platform but merely a data host using imagery from multiple satellites.

[2]Note that most SAR-optical datasets can be used in this way even if designed for other learning tasks.

(a) (b)

Fig. 1. Public training and testing AOIs. Green sections indicate public training data and yellow sections indicate public testing data. (a) AOI 2 (gaziantep). (b) AOI 3 (adıyaman).



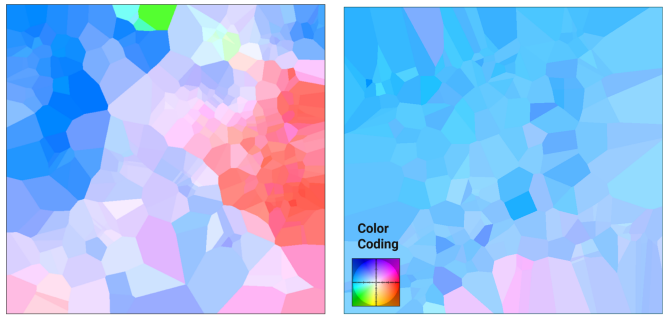
(a)

(b)

Fig. 2. Examples of annotated reference tiepoints included in SpaceNet 9. Red crosses indicate point of correspondence between modalities. (a) Annotated reference tiepoint example 1. (b) Annotated reference tiepoint example 2.

Tiepoints: For each AOI and section, we created reference tiepoints between the optical and SAR data. Tiepoint data are provided as a CSV with each row defining a single tiepoint indicating the corresponding row and column between the optical and SAR image. Our annotation process consisted of two rounds of labeling, including an initial labeling effort followed by a final review and cleaning.

Identifying corresponding locations in high-resolution SAR and optical imagery is an extremely challenging task. This is especially the case in dense urban areas with tall structures. We prioritized annotating locations least affected by common distortions seen in SAR data, such as layover, shadow, and foreshortening. We avoided annotating on or near elevated structures, such as buildings, and focused on labeling locations where there are clear and discernible corresponding ground features between the two modalities (see Fig. 2). Tiepoint labeling and data review were performed using the open-source Geographic



(a) (b)

Fig. 3. Optical-to-SAR displacement calculated from the reference tiepoints. Displacement in the x and y directions is calculated between the reference optical and SAR tiepoints and then interpolated using nearest neighbor. We visualize this image using methodology and the flow color-coding found in [39]. Saturation indicates magnitude of displacement and hue indicates direction of displacement. (a) AOI 2 (Gaziantep). (b) AOI 3 (adıyaman).

Information System (GIS) software QGIS. The total number of tiepoints included for each AOI and section are found in Table II.

Analysis: The average distance between optical and SAR tiepoints for each AOI ranges from 30 to 34 m. Displacement across each AOI is nonuniform, showing large variations in magnitude and direction of the shift. Fig. 3 illustrates the magnitude and orientation of the optical-to-SAR displacement for the public AOIs. AOI 2 exhibits larger variations in magnitude and orientation of displacement compared to AOI 3. The reference displacement maps further illustrate that perfect coregistration cannot be achieved using rigid global transformations.

The SpaceNet 9 dataset fills a severe dataset gap in the cross-modal image registration domain. The dataset includes a diversity of land cover (including both dense urban and barren lanscapes) across three AOIs and consists of high spatial resolution optical and SAR data, variable optical-to-SAR image displacement within and across AOIs, and high-quality reference tiepoints for algorithm development and evaluation.

B. Problem Statement

The objective of SpaceNet 9 is to compute a dense displacement map that indicates the x- and y-shift needed to align the pixels in the optical image to the pixels in the SAR image. Given as input an optical image $I_o \in \mathbb{R}^{3\times H_o\times W_o}$ and SAR image $I_s \in \mathbb{R}^{1\times H_s\times W_s}$, produce as output a two-channel displacement image $I_d = (I_d^{(x)}, I_d^{(y)})$, where $I_d^{(x)}, I_d^{(y)} \in \mathbb{R}^{H_o\times W_o}$. For each pixel location (h, w) in the optical image, $I_d^{(x)}(h, w)$ and $I_d^{(y)}(h, w)$ specify the shift in the x (horizontal) and y (vertical) directions, respectively, required to map the optical pixel to the corresponding pixel in I_s. The output displacement map therefore has the same spatial dimension (height and width) as the input optical image.

The accuracy of the predicted displacement image is evaluated against the displacement calculated using the annotated reference tiepoints. Specifically, the final registration accuracy for a section is defined as the average Euclidean distance in meters between the predicted and reference x- and y-shift across

all annotated tiepoints N

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\left(x_i^{\text{pred}} - x_i^{\text{ref}}\right)^2 + \left(y_i^{\text{pred}} - y_i^{\text{ref}}\right)^2}. \quad (1)$$

For reporting the public test score, we calculate this registration accuracy for both sections and then average them.

C. Baseline Algorithm

We developed a simple baseline algorithm [40] and utilities for using this data and evaluating results. The goals of this baseline are twofold: 1) to provide challenge contestants a suitable starting algorithm, code base, and evaluation procedure to improve, develop, and experiment with more advanced solutions and 2) assess the difficulty of this dense coregistration task and generalization challenges in this dataset.

The baseline algorithm is in some aspects inspired by SAR–optical image matching [29]. The baseline algorithm consists of training two keypoint detectors, one for optical keypoint detection and the other for SAR keypoint detection. For scene-level inference, we detect keypoints in both modalities using the keypoint detectors and then estimate a global affine transformation between the two point sets. A final global displacement map is created from the affine transformation.

Data preparation: We prepare two tiled datasets for training the keypoint detectors. The optical keypoint detection tiled dataset is generated by cropping SAR and optical images centered on the SAR tiepoint. We create a heatmap label for each optical–SAR tile pair with a 2-D Gaussian kernel placed at the location of the corresponding optical tiepoint. The same process is used to create the SAR keypoint detection training tiles, but cropping the SAR and optical images centered on the optical tiepoint location and placing the Gaussian kernel at the location of the corresponding SAR tiepoint.

Keypoint detection: The keypoint detector architecture consists of two separate encoders for the optical image and SAR image and a shared decoder. The architecture is similar to a U-Net [41] in that we include skip connections from blocks in the encoders to blocks in the decoder. For each skip connection we include an attention block, as found in [42]. We train the keypoint detectors to predict the keypoint location in one modality that corresponds to the center-point of the image in the other modality. The task is heatmap regression, and we use a loss similar to the one proposed in [29], i.e., a weighted mean-squared error loss to account for the imbalance between zero and nonzero pixels in the reference heatmap. A regularization term is also included to penalize predicting many nonzeros or keypoints. During training, we use data augmentation including random flips, rotations, and translations.

Inference: For inference, we start by first predicting optical keypoints in a uniform grid (e.g., every 100 m) across the image. For each predicted optical keypoint, we then predict SAR keypoints using the SAR keypoint detector.

Transformation: To obtain the final displacement map, we estimate a transformation using the detected optical and SAR keypoints from the keypoint detectors. We estimate an affine transformation via RANSAC. The estimated transformation is then used to create the displacement image.

TABLE III
BASELINE ALGORITHM RESULTS

Method	Public Test	Private Test
Zero-shift	34.81	33.89
Affine on Reference	6.51	22.37
Baseline w/ Optical Reference Tiepoints	18.92	22.20
Baseline w/ SAR Reference Tiepoints	30.63	24.18
Baseline	16.68	24.10

Scores are reported as average Euclidean distance (meters) between predicted and reference tiepoints.

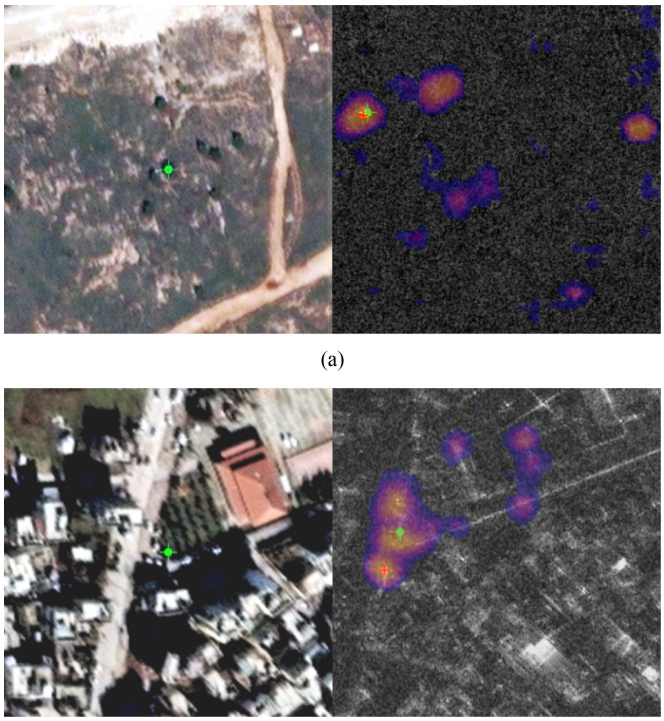
Implementation details: We created a tiled dataset from the annotated tiepoints in the public training sections and randomly split the tiles into 80% training and 20% validation. We create SAR and optical tiles so that they have consistent spatial extent. Each tile is approximately 150 × 150 m in size, and to create the label heatmap, we use a 2-D Gaussian kernel with a sigma of 5 m. This results in tiles with different pixel dimensions (height and width in number of pixels) due to the variable pixel resolution of the SAR (0.36–0.45 m) and optical imagery (0.31 m). For training, we resample the tiles to have consistent pixel dimensions (384 × 384 pixels).

We trained two keypoint detector models, including one for detecting SAR keypoints and one for detecting optical keypoints. The detectors were trained for 175 epochs using the Adam optimizer with an initial learning rate of 1e-4. We used a step learning rate schedule to drop the learning rate by a factor of 1e-2 after 30 epochs. Models were trained on a single NVIDIA V100 GPU (32 GB). During inference, we compute the location of the detected keypoint from the predicted heatmap using the argmax operation.

Baseline algorithm results: Results obtained with the baseline approach are shown in Table III. As a sanity check and to understand if the baseline is improving the coregistration, we also include the score if we were to predict no shift (i.e., the displacement map is all zeros). The baseline approach improved registration accuracy from 35 to 17 m for the public test set and from 34 to 24 m for the private test set.

Table III further breaks down the performance of several baseline variants to better understand the contribution of different design choices. Applying a simple affine transformation estimated on the reference data yields the best performance on the public test set, achieving an average error of 6.51 m, while still providing a notable improvement on the private test set (22.37 m). This indicates that a large fraction of the misregistration can be explained by a global geometric transform, at least for the public scenes.

When incorporating external tiepoints, the baseline using optical reference tiepoints performs comparably to the standard baseline on the private test set (22.20 m versus 24.10 m), but is less effective on the public test set. In contrast, using SAR reference tiepoints results in only a marginal improvement over the zero-shift baseline on the public test set and a moderate improvement on the private test set, suggesting that SAR-based tiepoints are noisier or less consistent across scenes. Overall, the



(a)

(b)

Fig. 4. Examples of two SAR keypoint detector predictions. The predicted heatmap from the SAR keypoint detector is displayed over the SAR image. The green cross indicates the reference optical and SAR keypoint locations and the red cross indicates the argmax on the predicted heatmap. (a) SAR keypoint detection example 1. (b) SAR keypoint detection example 2.

full baseline method provides a balanced improvement across both test sets, outperforming the zero-shift baseline and most ablated variants, while highlighting the potential upper bound achievable with simple global alignment models.

We notice several limitations in the proposed baseline. First, outputs from both keypoint detectors often have multiple modes, resulting in some ambiguity in the matching keypoint. It was very difficult to train the keypoint detectors to output a single peak in the heatmap. This is seen in Fig. 4.

IV. Challenge Overview

A. Challenge Submission Phase

The submission phase on the Topcoder platform lasted from 3 April to 26 May 2025, a total of seven and a half weeks. During this period, any member of the Topcoder community could register individually or as part of a team, thereby gaining access to the competition dataset and the supporting discussion forum. The challenge was actively promoted through various channels, particularly during the first half of the submission phase. As a result, the number of registered participants increased steadily throughout the period, as shown in Fig. 5.

The final evaluation of the submitted solutions on the previously hidden private set dataset took place only after the submission phase was completed. To provide participants with feedback during development, they were required to submit not only their full code but also the output generated by their solution

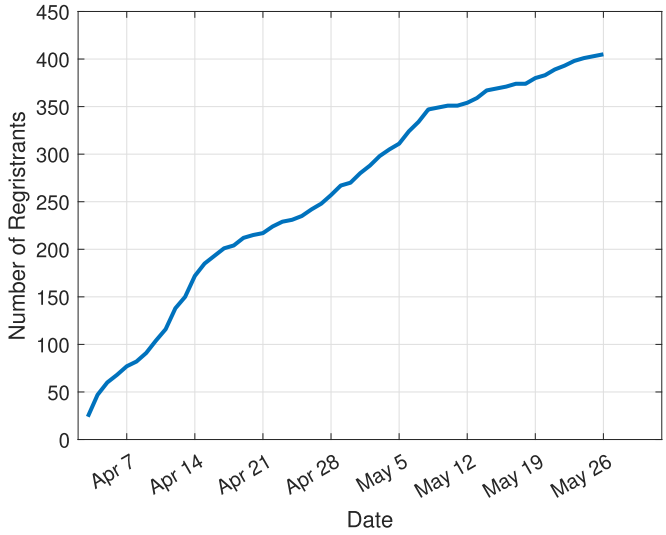


Fig. 5. Cumulative number of registered participants.

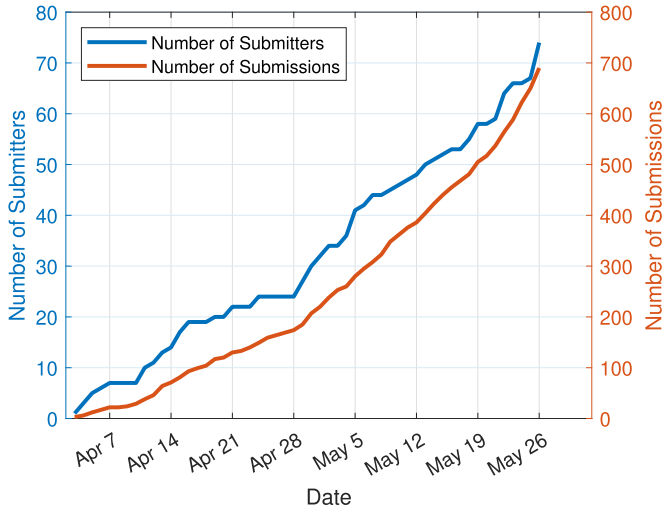


Fig. 6. Cumulative number of submitters and submissions.

on the public test dataset (described in Table II and Fig. 1). This output was automatically scored using the evaluation metric described in Section III-B, and the results were displayed on a *provisional* leaderboard. Participants were able to update their code, resubmit, and obtain new provisional scores, resulting in a dynamic leaderboard that evolved with each submission. Fig. 6 illustrates the cumulative number of submissions, as well as the number of participants/teams that submitted at least once. A surge in submission activity near the deadline is typical for such competitions.

From an observer's perspective, it is often interesting to follow how the best provisional score evolves during the submission phase. This is shaped both by the participants' ability to improve their methods over time and by when the top competitors choose to join. In this challenge, the provisional leaderboard was dominated for much of the competition by a single participant (who ultimately placed second in the final ranking). Consequently,

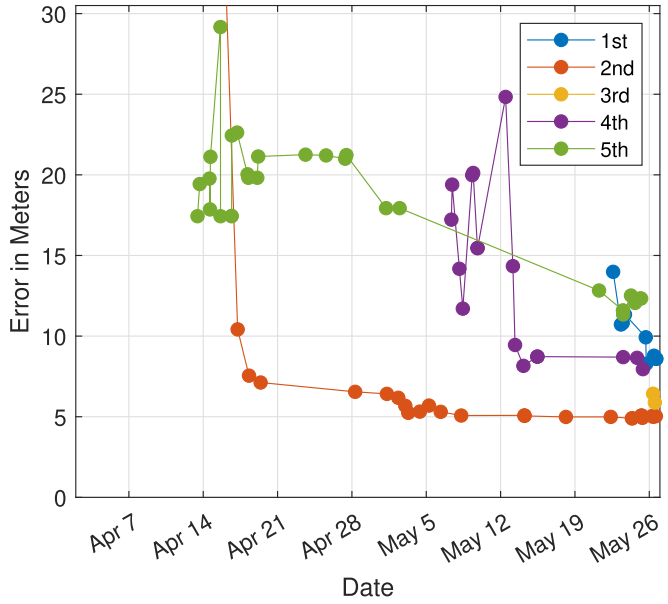


Fig. 7. Provisional error progression for top five participants (in the final ranking).

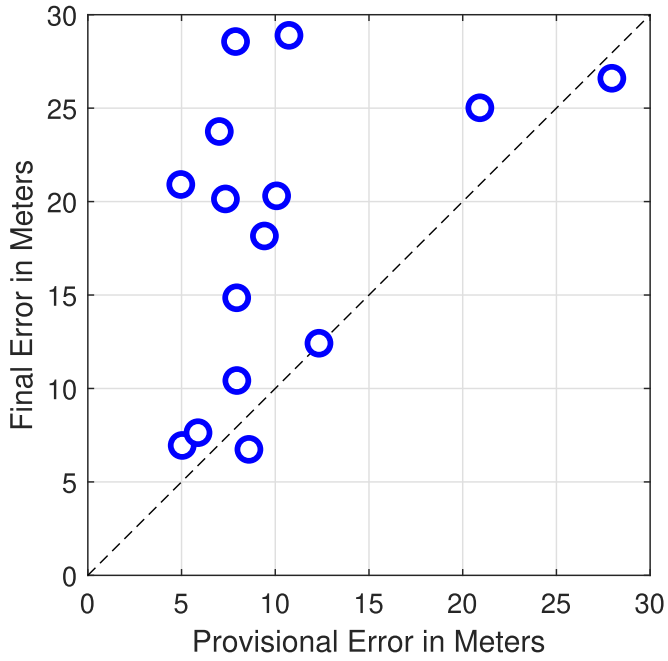


Fig. 8. Comparison of provisional and final errors for top 15 participants.

the trajectory of the lowest error closely mirrored the provisional results of this participant, which can be seen in Fig. 7. The figure shows the evolution of the provisional errors for the top five competitors (ranked by final results).

As shown in Fig. 7, the final ranking did not align fully with the provisional leaderboard. This outcome is common in machine learning competitions, where provisional scores are based on a public test dataset. Intensive tuning to maximize provisional performance often leads to overfitting. In addition, the variance in the performance of individual solutions in different datasets can play an important role. In this case, the final evaluation used imagery from a region not included in the training or public test datasets (see Table II and Fig. 1), and solutions varied in their ability to generalize to this new area. This effect is further illustrated in Fig. 8, which compares the errors of each of the top 15 participants in the provisional (x-axis) and final (y-axis) datasets. Although ten of the leading solutions achieved errors between 5 and 10 m in the public dataset, only the top three maintained this level of performance in the final data set.

B. Final Testing

During the submission phase, only the outputs generated within participants' local test environments were evaluated. Although the submitted packages also included the corresponding code, executing these solutions continuously was not feasible. The runtime required for a single solution, combined with the large number of submissions arriving within short intervals—particularly near the deadline—would have risked overloading the evaluation hardware. Moreover, continuous execution would have necessitated the maintenance of GPU-enabled infrastructure throughout the entire submission phase, significantly increasing costs. In contrast, evaluation of the submitted outputs could be reliably performed on a smaller CPU-based instance.

For the final evaluation, participants were required to provide their solutions in a Dockerized format to enable automation of the testing process. While containerization is designed to ensure reproducibility—such that a solution functioning in a participant's local environment should also function in the final evaluation—specific issues often emerge that are environment-dependent. Furthermore, despite detailed submission guidelines, minor structural errors can prevent automated execution. To reduce the risk of such failures and to allow participants to validate their submissions, two preliminary test rounds were conducted on 5 May and 19 May. In these rounds, the most recently submitted solution of each participant was executed on the same hardware designated for the final evaluation. In cases where execution failed, participants were provided with feedback identifying the error, allowing them to correct their submissions in advance of the final test.

To ensure a fair comparison of the solutions submitted, the evaluation methodology was specified prior to the launch of the challenge. This included defining which portions of the dataset would be used for development, provisional scoring, and final scoring, as well as establishing the hardware environment for evaluation and the maximum allowed inference time.

All evaluations were conducted on an AWS `g4dn.12xlarge` instance equipped with 192 GiB of RAM, 48 virtual CPUs, and four NVIDIA T4 Tensor Core GPUs (each with 16 GiB of GPU memory). This configuration was selected based on the size of the input data, with the intention of avoiding hardware-related bottlenecks during solution development. Potential end users of these algorithms were assumed to have access to sufficiently powerful hardware.

The runtime limit for producing results on a single pair of images of approximately $10\,000 \times 10\,000$ pixels was set at 30 min. This threshold was short enough to allow the evaluation

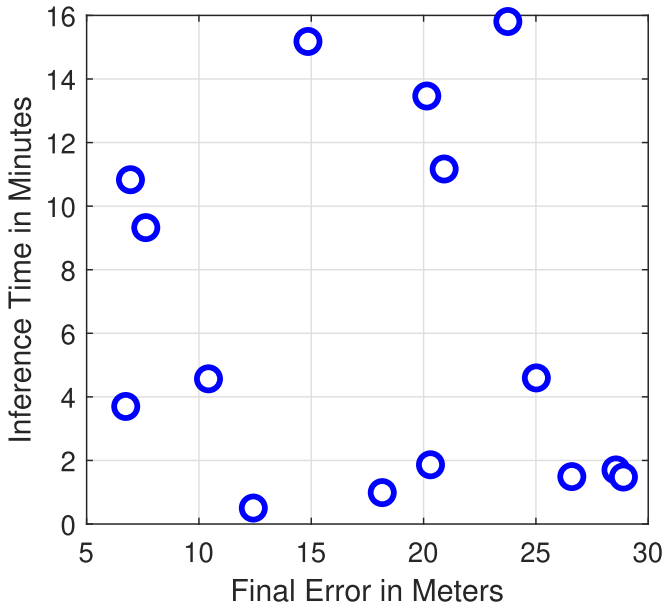


Fig. 9. Comparison of final errors and running times for top 15 participants.

TABLE IV
AVERAGE ERRORS (M) OF TOP-SCORING PARTICIPANTS

Place	Public Test	Private Test
1st	8.59	6.74
2nd	5.04	6.95
3rd	5.88	7.63
4th	7.95	10.42
5th	12.33	12.42
10th (Best graduate)	4.96	20.91
14th (Best undergraduate)	7.88	28.56

process to be completed within several days of the conclusion of the challenge submission period, even in worst-case scenarios, and lenient enough not to restrict the practical applicability of the winning solutions. As shown in Fig. 9, this runtime limit did not impose constraints on solution development: None of the top 15 entries required more than 16 min for inference, and the winning solution produced results in less than 4 min. These results indicate that participants were not forced to perform excessive code optimizations in order to remain within the limit.

V. SN9 Winning Approaches

Prizes were awarded to the top five solutions, as well as to the best solution submitted by a graduate and an undergraduate student. Table IV reports the final errors of the winning solutions, whose main characteristics are summarized in the following sections. In addition, Fig. 10 illustrates the distribution of final errors for the top-scoring participants.

In Figs. 11 and 12, we visualize the reference and predicted displacement maps from the top three solutions for the public testing AOIs. Using the displacement map, we warp the optical image and show notable improvements in alignment to the SAR image. This is displayed for the first place solution in Fig. 13.

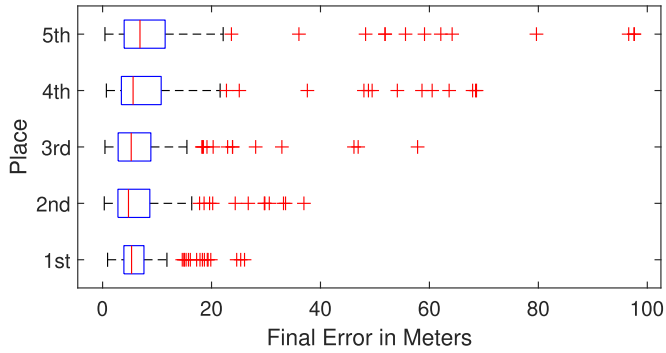


Fig. 10. Boxplot of errors on the private test AOI for top-scoring participants.

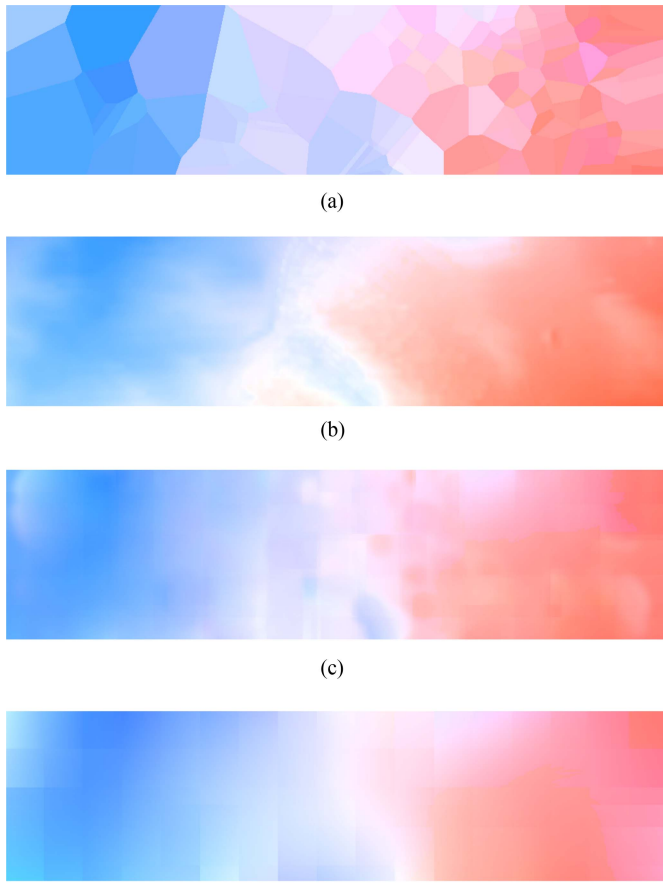


Fig. 11. Reference and predicted displacement maps for public test AOI 2 Section 2. (a) Reference. (b) 1st. (c) 2nd. (d) 3rd.

A. First Place

The winning solution combined geometric priors with lightweight learning components. The pipeline first reprojected the SAR image to match the grid of the optical image. Candidate correspondences were extracted using the LightGlue feature matcher [43], and terrain information was introduced through depth maps generated with DepthAnything v2 [44] as well as coarse digital surface models (DSM) from Copernicus data. Displacement vectors (dX, dY) between the modalities were

Fig. 12. Reference and predicted displacement maps for public test AOI 3 Section 2. (a) Reference. (b) 1st. (c) 2nd. (d) 3rd.

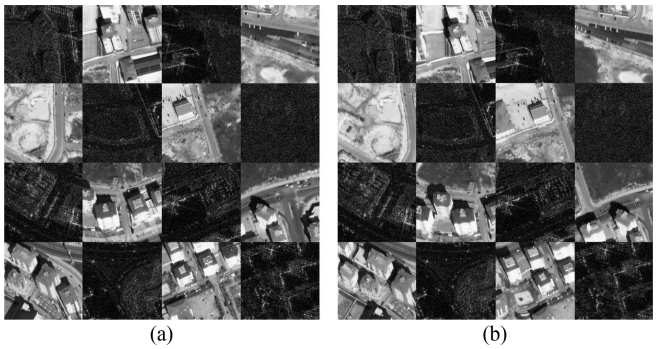


Fig. 13. Mosaic of prealignment and postalignment in a portion of public test AOI 2. (a) Before alignment. (b) After alignment (1st place solution).

then estimated using a linear regression model fitted with a RANSAC variant.

A distinctive aspect was the use of terrain and depth information as proxies for missing sensor geometry, enabling robust displacement estimation without overfitting to the very limited dataset. The approach deliberately avoided complex convolutional neural networks, focusing instead on interpretability and robustness to noisy tie points.

Because the method did not depend on training with the provided dataset and instead relied on general-purpose depth priors and global elevation models, the solution generalized well to unseen geographic regions and imaging conditions.

B. Second Place

The runner-up designed a two-stage pipeline. The *base registration stage* performed patchwise feature extraction and matching using several publicly available pretrained models. Feature descriptors were obtained from MINIMA [45], ALIKED [46], DISK [47], and SIFT [48], while matching was carried out using transformer-based networks, such as LoFTR [49], XoFTR [50], and LightGlue [43]. Different detector–matcher combinations were evaluated to improve robustness across varying image modalities. To improve the reliability of keypoints, the system also incorporated auxiliary models for building segmentation and above-ground height estimation, which were used for filtering correspondences in densely built-up areas. Both global homography and local homographies were estimated and combined using Gaussian-weighted confidence maps. The *refinement stage* rewarped the optical image according to the estimated offsets and repeated the matching process to capture residual errors.

The approach introduced several notable features: 1) local homography estimation with Gaussian-based weight maps to reflect tie point reliability, 2) keypoint filtering in urban areas by combining a building segmentation model with an above-ground height estimation model, and 3) a refinement loop to improve alignment after the initial registration.

The use of multimodal feature matchers trained on diverse datasets, combined with explicit handling of urban versus rural structures, improved robustness across terrain types. The patchwise design with overlapping windows further supported generalization to unseen AOIs.

C. Third Place

This participant constructed a modular pipeline based on dividing the images into overlapping patches. For each patch, feature correspondences were extracted using the `image-matching-models` library [51], [52], and local transformations were estimated using RANSAC. These local transformations were then aggregated into a global displacement map.

The method emphasized stability over complexity. Special treatment of border patches and filtering of unreliable matches improved the robustness of the results. Although simpler than higher ranked solutions, the focus on aggregation of local transformations provided a stable outcome.

The approach made minimal assumptions about the data and thus could be applied to a wide range of optical–SAR image pairs.

D. Fourth Place

The fourth-place solution used a matcher-based pipeline centered on the RoMA algorithm [53], which uses DinoV2 [54] features. The optical image was converted to grayscale, while the SAR image was smoothed with a Gaussian blur. Matching keypoints were filtered by certainty score, reduced using the KMeans++ clustering method, and interpolated into dense displacement maps through thin-plate spline interpolation.

Three original contributions distinguished this solution: 1) certainty-based filtering of matching keypoints to discard unreliable matches, 2) clustering of keypoints to remove redundancy while preserving spatial coverage, and 3) sparse-to-dense interpolation with radial basis functions to produce the final dense displacement field required by the challenge.

Although RoMA was originally trained on natural images, the approach transferred successfully to multimodal SAR–optical registration. This indicates good cross-domain adaptability. The reliance on pretrained general-purpose matchers made the pipeline potentially useful beyond the competition dataset.

E. Fifth Place

The fifth-place solution combined SuperPoint [55] for feature extraction and LightGlue [43] for matching, with a central role played by adaptive preprocessing of SAR images. Depending on metrics, such as edge strength, texture variance, and black border ratio, the pipeline dynamically selected between a "simple" mode (bilateral filtering and histogram equalization) and an "enhanced" mode (contrast-limited adaptive histogram equalization, sharpening, and Gaussian blur). Matches were filtered and fitted to an affine model using RANSAC.

The most distinctive element was the adaptive preprocessing scheme tailored to SAR tile quality. This design improved robustness by dynamically adapting to the specific characteristics of each image.

The ability to adapt preprocessing steps per tile allowed the solution to handle heterogeneous SAR imagery across urban and rural settings, enhancing generalization to varied real-world conditions.

F. Best Graduate

The top graduate solution approached SAR–optical registration with a multistage pipeline. A coarse-to-fine hierarchical matching strategy handled very large images and initial misalignments, using RoMA [53] deep feature matchers to identify correspondences. A parallel ensemble of multiple affine estimations was then run to reduce instability, selecting the transformation with the lowest reprojection error. Finally, a hybrid transformation model combined a global affine alignment with a local nonrigid B-spline transformation to capture both large-scale shifts and fine local distortions.

Several key elements characterized this approach. The hierarchical "align-and-crop" strategy broke down the matching process into progressively smaller patches, increasing both density and accuracy of correspondences. The parallel ensemble method addressed the randomness and instability of single affine estimations, ensuring robust outcomes. The hybrid affine–B-spline transformation balanced global orientation correction with local warping adjustments, achieving a higher precision than either model could alone.

The pipeline was designed for scalability to massive images and adaptability to different geographic conditions. The hierarchical and hybrid structure supported generalization across varied terrain and imaging conditions, while reliance on robust, pretrained matchers ensured resilience to appearance differences between SAR and optical imagery. However, the computational cost was high, requiring parallelization across multiple graphics processing units (GPUs), which may limit applicability in resource-constrained settings. Despite this, the solution demonstrated strong potential for general use in multimodal registration tasks.

G. Best Undergraduate

The best undergraduate student built a pipeline based on the MINIMA framework [45]. The system employed the RoMA matcher [53], a transformer-based model trained on a synthetic multimodal dataset. Sparse correspondences between optical and SAR images were extracted, converted into geographic coordinates, and used to fit a robust affine transformation via RANSAC. A dense pixel-level offset map was then generated by applying this transformation to every pixel.

The central feature was the combination of a modality-invariant matcher with a robust global affine model to achieve dense alignment. By leveraging generative pretraining rather than dataset-specific learning, the approach avoided overfitting. The solution also preserved geospatial metadata in its output, ensuring compatibility with GIS workflows.

The reliance on a modality-invariant matcher provided strong cross-domain adaptability, making the approach transferable to diverse satellite imagery scenarios. However, the single global affine model limited its ability to handle complex terrain distortions, suggesting that future extensions (e.g., piecewise transformations or iterative refinement) would further enhance generalizability.

VI. Discussion

Feature-based registration using modern matchers proved to be one of the most effective strategies. In particular, LightGlue, often paired with keypoint detectors, such as SIFT, SuperPoint, ALIKED, or DISK, played a central role in several top-performing pipelines. Transformer-based and multimodal-pretrained matchers, such as RoMA and MINIMA, demonstrated excellent cross-domain adaptability, successfully bridging the gap between optical and SAR modalities.

In terms of geometric modeling, RANSAC and its variants were consistently used to handle noisy correspondences. Approaches combining affine or hybrid global–local transformations achieved robust alignment even under challenging conditions. Moreover, sparse-to-dense interpolation techniques, such as thin-plate or B-splines, were particularly effective in producing smooth displacement fields from sparse matches.

Several teams improved robustness by incorporating auxiliary data and priors. The winning solution, for instance, utilized terrain information, such as DSMs, while the runner-up enhanced performance in urban areas through urban-aware keypoint filtering.

Another recurring pattern among successful methods was the use of multistage and hierarchical strategies. Coarse-to-fine alignment, iterative refinement, and patchwise decomposition stabilized convergence and increased registration precision. In addition, adaptive preprocessing of SAR imagery—including the dynamic selection of filters, such as contrast enhancement and blurring—significantly boosted robustness across heterogeneous SAR tiles, as demonstrated by the fifth-place solution.

A clear trend among the top-ranking solutions was the avoidance of heavy end-to-end training. None of the leading methods relied on large CNNs trained specifically for the challenge dataset. Instead, they leveraged general-purpose pretrained matchers, such as RoMA, LightGlue, or SuperPoint, emphasizing transferability over dataset-specific optimization.

Most pipelines employed a patch-based decomposition strategy, dividing images into smaller regions for local alignment and subsequently aggregating results into a consistent global transformation. This approach improved flexibility and

robustness in the presence of large displacements or spatially varying distortions.

All winning teams also implemented robust outlier rejection mechanisms, most commonly using RANSAC or certainty-based filtering, to ensure that noisy or mismatched tie points did not degrade the final transformation. In addition, hybrid global–local strategies were popular: several methods combined global transformations, such as affine or homography models with local refinements using splines or patchwise models.

Finally, the top-performing solutions placed strong emphasis on generalization. Rather than optimizing for a specific dataset, they achieved robustness across unseen regions by leveraging external priors (e.g., DSMs and depth maps) and modality-invariant matchers trained on diverse datasets.

Several unexpected insights emerged from the challenge results. Perhaps most notably, simplicity often outperformed complexity. The first-place solution achieved top performance with a lightweight regression model that incorporated depth priors, deliberately avoiding complex deep learning architectures. Similarly, the third-place method—based on a straightforward patchwise RANSAC aggregation—produced remarkably stable and accurate results.

Another noteworthy observation was that general-purpose matchers generalized well across domains. Models, such as RoMA, originally trained on natural or synthetic imagery, transferred effectively to the SAR–optical registration task without retraining.

In addition, adaptive preprocessing proved highly impactful. The fifth-place pipeline, which dynamically adjusted SAR filtering parameters, achieved greater robustness improvements than competing methods that introduced additional deep learning components.

Lastly, it was observed that high computational cost did not necessarily correlate with top performance. A graduate-level solution employing multi-GPU hybrid affine–B-spline transformations achieved strong accuracy but was ultimately outperformed by simpler, more efficient pipelines.

VII. Conclusion

First responders are commonly faced with the challenge of integrating and analyzing optical pre-event imagery and postevent SAR, which have inherent alignment issues due to the different modalities. SpaceNet 9 marked the first prize challenge in which we explored cross-modal image registration algorithms to accelerate downstream analytics and uses. To facilitate this challenge, the SpaceNet team developed a dataset comprised of three urban areas in Türkiye impacted by earthquakes in 2023 consisting of optical and SAR imagery from Vantor and UMBRA's open data programs, respectively, along with roughly 800 hand-labeled tie-points for training and testing (public and private).

SpaceNet 9 identified a variety of techniques that worked well to help improve model performance for this challenge, including deep learning-based ad hoc keypoint matchers, incorporating auxiliary map data, processing data patches, and adaptive preprocessing. Simple models that emphasized generalization generally outperformed complex, specialized models. The size of the dataset was a primary limitation, including the square kilometers of each AOI, the number of tie points, and the geographic diversity of the locations.

Future work will include additional SpaceNet challenges that build off SpaceNet 9's multimodal data characteristics with an emphasis on practical applications of this research in fields, such as disaster relief and recovery. The SpaceNet 9 open dataset and winning models can be leveraged by the community to further improve upon this work and enable downstream uses. Our team is in the process of planning SpaceNet 10 and looks forward to building on this effort in an upcoming challenge.

Acknowledgment

We acknowledge the contributions of all of the SpaceNet partners including AWS, IEEE GRSS, Oak Ridge National Laboratory, OGC, Topcoder, UMBRA, and Vantor.

We acknowledge that this manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

References

[1] W. Zhang et al., "Multi-resolution SAR and optical remote sensing image registration methods: A review, datasets, and future perspectives," 2025, *arXiv: 2502.01002*.

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

[3] H.-M. Chen and P. Varshney, "A pyramid approach for multimodality image registration based on mutual information," in *Proc. 3rd Int. Conf. Inf. Fusion*, 2000, pp. MOD3/9–MOD315.

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

[5] X. Xu, X. Li, X. Liu, H. Shen, and Q. Shi, "Multimodal registration of remotely sensed images based on Jeffrey's divergence," *ISPRS J. Photogrammetry Remote Sens.*, vol. 122, pp. 97–115, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271616304324

[6] H.-M. Chen, M. Arora, and P. Varshney, "Mutual information-based image registration for remote sensing data," *Int. J. Remote Sens.*, vol. 24, no. 18, pp. 3701–3706, 2003.

[7] W. Yang, C. Han, H. Sun, and Y. Cao, "Registration of high resolution SAR and optical images based on multiple features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2005, vol. 5, pp. 3542–3544.

[8] G. Lehureau, F. Tupin, C. Tison, G. Oller, and D. Petit, "Registration of metric resolution SAR and optical images in urban areas," in *Proc. 7th Eur. Conf. Synthetic Aperture Radar*, 2008, pp. 1–4.

[9] Q. Li, G. Qu, and Z. Li, "Matching between SAR images and optical images based on HOG descriptor," in *Proc. IET Int. Radar Conf.* 2013, pp. 1–4.

[10] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, "Robust registration of multimodal remote sensing images based on structural similarity," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2941–2958, May 2017.

[11] B. Fan, C. Huo, C. Pan, and Q. Kong, "Registration of optical and SAR satellite images by exploring the spatial relationship of the improved SIFT," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 4, pp. 657–661, Jul. 2013.

[12] C. Xu, H. Sui, H. Li, and J. Liu, "An automatic optical and SAR image registration method with iterative level set segmentation and sift," *Int. J. Remote Sens.*, vol. 36, no. 15, pp. 3997–4017, 2015.

[13] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, and F. Tupin, "SAR-SIFT: A SIFT-like algorithm for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 453–466, Jan. 2015.

[14] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Trans. Image Process.*, vol. 29, pp. 3296–3310, 2020.

[15] Y. Xiang, F. Wang, and H. You, "OS-SIFT: A robust SIFT-like algorithm for high-resolution optical-to-SAR image registration in suburban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3078–3090, Jun. 2018.

[16] Y. Ye, L. Bruzzone, J. Shan, F. Bovolo, and Q. Zhu, "Fast and robust matching for multimodal remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9059–9070, Nov. 2019, doi: 10.1109/TGRS.2019.2924684.

[17] Y. Ye, C. Yang, J. Zhang, J. Fan, R. Feng, and Y. Qin, "Optical-to-SAR image matching using multiscale masked structure features," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art no. 6509405.

[18] J. Fan, Y. Wu, M. Li, W. Liang, and Y. Cao, "SAR and optical image registration using nonlinear diffusion and phase congruency structural descriptor," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5368–5379, Sep. 2018.

[19] L. Huang, Z. Li, and R. Zhang, "SAR and optical images registration using shape context," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2010, pp. 1007–1010.

[20] B. Xiong, W. Li, L. Zhao, J. Lu, X. Zhang, and G. Kuang, "Registration for SAR and optical images based on straight line features and mutual information," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 2582–2585.

[21] H. Sui, C. Xu, J. Liu, and F. Hua, "Automatic optical-to-SAR image registration by iterative line extraction and Voronoi integrated spectral point matching," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 6058–6072, Nov. 2015.

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

[23] H. Zhang et al., "Registration of multimodal remote sensing image based on deep fully convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 3028–3042, Aug. 2019.

[24] M. Zhao, G. Zhang, and M. Ding, "Heterogeneous self-supervised interest point matching for multi-modal remote sensing image registration," *Int. J. Remote Sens.*, vol. 43, no. 3, pp. 915–931, 2022.

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

[26] L. H. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu, "Identifying corresponding patches in SAR and optical images with a pseudo-Siamese CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 784–788, May 2018.

[27] D. Quan et al., "Deep generative matching network for optical and SAR image registration," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 6215–6218.

[28] S. Wang, D. Quan, X. Liang, M. Ning, Y. Guo, and L. Jiao, "A deep learning framework for remote sensing image registration," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 148–164, 2018, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271617303891

[29] L. H. Hughes, D. Marcos, S. Lobry, D. Tuia, and M. Schmitt, "A deep learning framework for matching of SAR and optical imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 169, pp. 166–179, 2020.

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

[31] M. Schmitt, L. Hughes, and X. X. Zhu, "The SEN1-2 dataset for deep learning in SAR-optical data fusion," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, pp. 141–146, 2018.

[32] Y. Wang and X. X. Zhu, "The sarptical dataset for joint analysis of SAR and optical image in dense urban area," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 6840–6843.

[33] M. Schmitt, L. Hughes, C. Qiu, and X. Zhu, "SEN12MS-A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 42, pp. 153–160, 2019.

[34] L. Wang et al., "SAR-to-optical image translation using supervised cycle-consistent adversarial networks," *IEEE Access*, vol. 7, pp. 129136–129149, 2019.

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

[36] M. Huang et al., "The QXS-SAROPT dataset for deep learning in SAR-optical data fusion," 2021, *arXiv:2103.08259*.

[37] W. Xu, X. Yuan, Q. Hu, and J. Li, "SAR-optical feature matching: A large-scale patch dataset and a deep local descriptor," *Int. J. Appl. Earth Observation Geoinformation*, vol. 122, 2023, Art. no. 103433.

[38] Y. Xiang, X. Wang, F. Wang, H. You, X. Qiu, and K. Fu, "A global-to-local algorithm for high-resolution optical and SAR image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art no. 5215320.

[39] S. Baker et al., "A database and evaluation methodology for optical flow," in *Proc. 2007 IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[40] R. Hänsch Arndt, P. Dias, A. Potnis, D. Lunga, D. Petrie, and T. M. Bacastow, "Introducing SpaceNet 9 - cross-modal satellite imagery registration for natural disaster responses," in *Proc. 2024 IEEE Int. Geosci. Remote Sens. Symp.*, 2024, pp. 234–238.

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

[42] R. Li, S. Zheng, C. Duan, J. Su, and C. Zhang, "Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art no. 8009205.

[43] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "LightGlue: Local feature matching at light speed," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 17627–17638.

[44] L. Yang et al., "Depth anything V2," in *Proc. 38th Int. Conf. Neural Inform. Process. Syst.*, 2024, pp. 21875–21911.

[45] J. Ren, X. Jiang, Z. Li, D. Liang, X. Zhou, and X. Bai, "MINIMA: Modality invariant image matching," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2025, pp. 23059–23068.

[46] X. Zhao, X. Wu, W. Chen, P. C. Chen, Q. Xu, and Z. Li, "Aliked: A lighter keypoint and descriptor extraction network via deformable transformation," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art no. 5014016.

[47] M. Tyszkiewicz, P. Fua, and E. Trulls, "DISK: Learning local features with policy gradient," in *Proc. 34th Int. Conf. Neural Inform. Process. Syst.*, 2020, pp. 14254–14265.

[48] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, 1999, pp. 1150–1157.

[49] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8922–8931.

[50] Ö. Tuzcuoğlu, A. Köksal, B. Sofu, S. Kalkan, and A. A. Alatan, "XoFTR: Cross-modal feature matching transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 4275–4286.

[51] A. Stoken, and contributors, "image-matching-models," 2024, Accessed: Oct. 17, 2025. [Online]. Available: https://github.com/alexstoken/image-matching-models

[52] G. Berton, G. Goletto, G. Trivigno, A. Stoken, B. Caputo, and C. Masone, "EarthMatch: Iterative coregistration for fine-grained localization of astronaut photography," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2024, pp. 4264–4274.

[53] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg, "RoMa: Robust dense feature matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 19790–19800.

[54] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," 2024, *arXiv:2304.07193*.

[55] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 224–236.