Evaluating Deep Matching Models for SAR-Optical Image Pairs using the SpaceNet9 Dataset

Constantin Günzel, Michael Schmitt

Department of Aerospace Engineering, University of the Bundeswehr Munich
(constantin.guenzel, michael.schmitt)@unibw.de

Keywords: SAR-Optical Registration, Multi-Modal-Matching, SAR, SpaceNet9, MINIMA, LightGlue

Abstract

This paper focuses on cross-modal image matching between Synthetic Aperture Radar (SAR) and optical imagery, a longstanding challenge due to fundamental differences in imaging geometry and radiometry. Beyond applicational needs in satellite data fusion and downstream mapping, this study is motivated by the rapid advances in the field of Computer Vision. Thus, this work evaluates classical and modern learning-based feature matching methods on the renowned SpaceNet9 dataset using a unified evaluation framework. The results show that classical methods such as SIFT fail to produce reliable correspondences, while learning-based approaches, particularly MINIMA, significantly improve performance without additional retraining. However, matching accuracy is strongly influenced by scene structure and SAR-specific geometric effects, therefore robust SAR-optical correspondence remains an open challenge.

1. Introduction

Cross-modal feature matching between Synthetic Aperture Radar (SAR) and optical imagery has become an essential topic in remote sensing, as a rapidly increasing number of satellites, diverse sensor modalities and open data programs offer novel opportunities for data fusion and information extraction. While optical sensors capture spectral and textural information, SAR provides geometry-sensitive, all-weather imaging. Combining both modalities is essential for applications such as change detection, mapping, and multi-sensor data integration.

However, establishing SAR-optical correspondence remains challenging due to their modality gap resulting from radiometric differences and geometric effects (Hughes and Schmitt, 2021). Traditional keypoint detectors and descriptors, such as SIFT (Lowe, 1999) or ORB (Rublee et al., 2011) rely on radiometric consistency and therefore often fail in cross-modal scenarios. Recent advances in the field of Computer Vision have introduced a new generation of learning-based matchers, such as SuperGlue and its successor LightGlue (Lindenberger et al., 2023), which exploit neural networks and transformer architectures, as well as modality-invariant frameworks such as MINIMA (Ren et al., 2025), which have shown promising results for SAR-optical data.

This study systematically evaluates classical and modern matching methods on the SpaceNet9 dataset under a unified evaluation pipeline. Specifically, LightGlue (with SuperPoint), MINIMA, and a classical SIFT baseline are compared with respect to their ability to establish reliable correspondences without modality-specific adaptation. In summary, the main contributions of this paper are therefore as follows:

- A unified evaluation pipeline for SAR-optical matching,
- A comparative analysis of classical and learning-based matching methods under similar conditions, considering scene structure and SAR distortions.

2. Dataset and Study Area

2.1 SpaceNet9 Dataset

The experiments are conducted on the publicly available SpaceNet9 (Hänsch et al., 2024) dataset, which provides co-registered SAR and optical image pairs together with manually annotated tie-points linking features between both modalities. These annotated tie-points are used as Control Points (CPs) for the evaluation. In contrast to geodetic Ground Control Points, they correspond to manually annotated 2D points in both image domains, typically describe clearly identifiable features, such as road intersections, and are used to evaluate cross-modal alignment accuracy.

In this study, only the training split of the dataset is used, as annotations are exclusively available for this subset. Consequently, all available annotated scenes are included, resulting in a total of three SAR-optical image pairs. This limited number of scenes presents a common constraint in SAR-optical datasets, where high-quality cross-modal annotations are scarce. The optical imagery originates from the Maxar Open Data Program with a ground sampling distance (GSD) of approximately 0.30 m, while the SAR data is acquired by the Umbra constellation with a GSD between 0.40 m and 0.45 m. The SAR images are provided as single-channel amplitude data.

Due to differences in image cropping and sensor geometry, the spatial extents of both modalities differ slightly. For reference, Table 1 states the approximate extent and GSD of each scene:[1]

Scene	$\sim$ Extent	GSD SAR	GSD Opt.
02_*x*_train_01	3.08×0.95 km	0.41 m	0.30 m
02_*x*_train_02	3.08×1.20 km	0.41 m	0.30 m
03_*x*_train_01	2.36×3.36 km	0.45 m	0.30 m

Table 1. Overview of the three SAR-Optical Image Pairs provided by SpaceNet9 (Training Data)

[1] Scene identifiers in Table 1 are abbreviated, original file names differ by modality (“sar” / “optical”, here: “x”) and include a “.tif” extension.

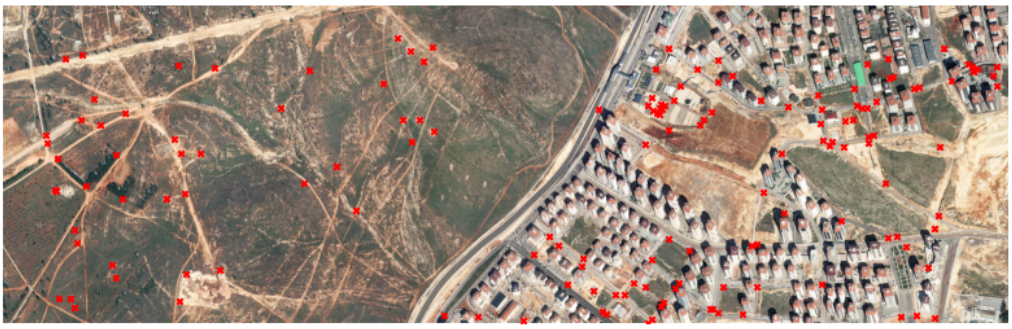

Figure 1. Locations of the provided labels (*red marker*) used as Control Points within one representative scene (*02_optical_train_01.tif*) in the electro-optical part of the image pair.

The selected scenes are located in urban and semi-urban environments and exhibit varying structural characteristics, ranging from sparsely built areas to densely urbanised regions in Türkiye.

2.2 Control Points

As previously mentioned, the annotated correspondences, defining a coordinate pair in each modality, are used as Control Points (CPs) to evaluate matching accuracy. Each CP represents a manually defined point pair linking a location in the SAR image to its counterpart in the optical image.

The number of CPs per scene differs, with the three scenes containing 151, 104, and 161 labeled correspondences, respectively. Their spatial distribution is not uniform, as annotations are concentrated in areas with distinctive structures, leading to varying coverage across the scenes. This reflects the manual annotation strategy of the dataset, which prioritizes clearly identifiable cross-modal features. Figure 1 demonstrates this non-uniform spatial distribution of CP for one representative scene. The remaining scenes exhibit similar characteristics.

2.3 Structural Characteristics

All scenes contain a mixture of ground-level areas and elevated structures such as buildings. Since elevation differences significantly affect SAR-optical correspondence, the proportion of built-up areas and the distribution of CPs across ground-level regions are analysed.

Table 2 summarises the fraction of elevated areas within each scene, as well as the proportion of CPs located on ground surfaces based on building masks of the presented scenes. The scenes display increasing building density, while the proportion of ground-based CPs decreases accordingly.[2]

Scene	Elevated area [%]	CPs on ground [%]
02_x_train_01	18.49	84.11
02_x_train_02	68.38	80.77
03_x_train_01	87.40	37.89

Table 2. Distribution of elevated structures within the scenes, as well as the distribution of the (non-uniformly placed) CPs located on ground level.

[2] Note on Table 2: The percentages of elevated area and ground-based CPs do not sum to 100%, as the spatial distribution of CPs is non-uniform. Although both quantities are correlated, they are not directly comparable.

These structural differences are expected to influence the matching performance, particularly due to SAR-specific geometric effects such as layover and foreshortening.

3. Methodology

3.1 Matching Methods

This study evaluates one classical and two learning-based feature matching approaches.

SIFT: Scale-Invariant Feature Transform (SIFT) (Lowe, 1999) is a classical handcrafted method that detects keypoints in scale space and describes them using gradient-based descriptors. Correspondences are established via nearest-neighbour matching with Lowe’s ratio test (Lowe, 2004). As it relies on radiometric consistency, its performance is limited in cross-modal scenarios.

SuperPoint + LightGlue: SuperPoint (DeTone et al., 2018) is a learning-based keypoint detector and descriptor trained via self-supervision. LightGlue (Lindenberger et al., 2023) performs matching using a transformer-based architecture with attention mechanisms. As it is not explicitly trained for multimodal data, its performance depends on the generalization capability of the learned features.

MINIMA (RoMa): MINIMA (Ren et al., 2025) is a modality-invariant matching framework trained on diverse imaging modalities. The RoMa variant predicts dense correspondences using a transformer-based architecture. As MINIMA is specifically designed for multi-modal applications, it is thus expected to generalize across different sensor domains.

Overall, these methods represent a progression from handcrafted descriptors to modern multi-modal matching approaches.

3.2 Evaluation Strategy

To quantify the geometric accuracy of the predicted correspondences, we derive an expected match location from the annotated Control Points (CPs) provided in the SpaceNet9 dataset and compare it to the location predicted by the matching algorithm. Let $\mathbf{s}_i = (x_i^{\text{SAR}}, y_i^{\text{SAR}})$ denote the coordinates of a CP in the SAR image and $\mathbf{o}_i = (x_i^{\text{opt}}, y_i^{\text{opt}})$ the corresponding coordinates in the optical image. The set of CP pairs defines a sparse mapping between both modalities.

A global geometric transformation between SAR and optical imagery is generally insufficient due to local distortions caused

by viewing geometry, relief displacement, and layover effects. Instead, we approximate the mapping locally using a piecewise affine transformation.

First, a Delaunay triangulation is constructed over the CP coordinates in the source image (SAR). This partitions the image into a set of non-overlapping triangles whose vertices correspond to three neighboring CPs. For an arbitrary point $\mathbf{p}$ detected in the SAR image, the triangle containing this point is identified. The point is then expressed using barycentric coordinates with respect to the triangle vertices:

$$\mathbf{p} = \lambda_1 \mathbf{s}_1 + \lambda_2 \mathbf{s}_2 + \lambda_3 \mathbf{s}_3 \tag{1}$$

with

$$\lambda_1 + \lambda_2 + \lambda_3 = 1. \tag{2}$$

The same barycentric weights are applied to the corresponding vertices in the optical image:

$$\hat{\mathbf{q}} = \lambda_1 \mathbf{o}_1 + \lambda_2 \mathbf{o}_2 + \lambda_3 \mathbf{o}_3. \tag{3}$$

This yields the predicted position $\hat{\mathbf{q}}$ of the point in the optical image. The procedure corresponds to a piecewise affine transfer defined by the CP correspondences.

Let $\mathbf{q}$ denote the match predicted by the evaluated feature matcher. The geometric error is defined as the Euclidean pixel distance between the predicted and observed correspondence:

$$e = \|\hat{\mathbf{q}} - \mathbf{q}\|_2 . \tag{4}$$

This error measure represents how closely the predicted match aligns with the locally interpolated geometric mapping defined by the CPs. The steps described were visualised in Figure 2.

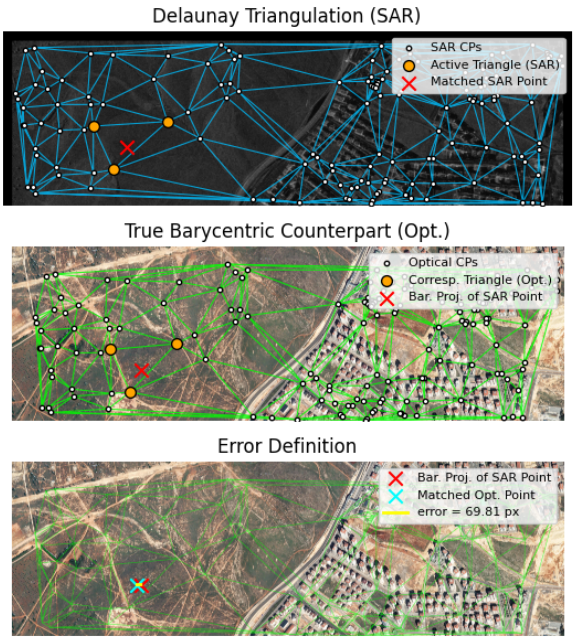

Figure 2. Error estimation using CP-based Delaunay Triangulation. *Top:* A SAR keypoint $\mathbf{p}$ within a Delaunay triangle defined by neighbouring CPs. *Center:* The point is mapped to the optical image via barycentric coordinates. *Bottom:* The error $\mathbf{e}$ is computed as the Euclidean distance between the projected position ($\hat{\mathbf{q}}$) and the predicted match $\mathbf{q}$.

An alternative evaluation in georeferenced coordinates would require accurate sensor models and orthorectification. To avoid additional sources of uncertainty, a pixel-based evaluation is adopted, relying solely on the relative geometry defined by the CPs.

3.3 Processing Pipeline

All methods are evaluated using the same processing pipeline to ensure comparability. The only varying component is the feature detection and matching algorithm.

Given a SAR-optical image pair, keypoints and correspondences are extracted using the respective matching method. Each match is represented as a coordinate pair $(\mathbf{p}, \mathbf{q})$. The expected correspondence $\hat{\mathbf{q}}$ is computed using the CP-based piecewise affine interpolation and the resulting geometric error defined as described in Section 3.2.

These resulting errors are aggregated both per scene and across all scenes, and analysed using statistical measures and spatial distributions as seen in Section 5. The pipeline is applied identically to all scenes and matching methods, ensuring a consistent and reproducible evaluation framework.

4. Experimental Setup

4.1 Parameter Selection

To ensure a fair comparison between the evaluated matching methods, all models are applied with minimal modification and without any additional geometric verification such as RANSAC or homography estimation. This allows assessing the intrinsic matching performance of each method in a purely data-driven setting.

Key parameters are selected through small-scale preliminary experiments to balance match density and accuracy. The chosen configurations are fixed across all scenes and not tuned per scene. Overall, the selected settings represent a consistent and practically relevant trade-off across all evaluated methods.

4.2 Evaluation Dimensions

The performance of the matching methods is evaluated along three complementary dimensions.

First, overall matching accuracy is assessed using statistical measures such as median error and empirical cumulative distribution functions (ECDF). Second, the influence of scene structure is analysed by separating matches into ground-based and elevated targets. Last, spatial and directional error characteristics are examined through error decomposition and spatial distributions.

Together, these dimensions provide a comprehensive assessment of accuracy and model behaviour.

5. Results

5.1 Overall Matching Performance

Table 3 summarises the overall matching performance across all scenes. Significant differences can be observed between classical and learning-based approaches.

	SIFT	LightGlue	MINIMA
#Matches	80	364	**12886**
Med. Error [px]	3028.95	43.38	**27.66**
$\leq$ 10 px [%]	0.0	6.87	**7.20**
$\leq$ 30 px [%]	0.0	31.04	**55.61**
$\leq$ 50 px [%]	0.0	61.26	**84.59**

Table 3. Comparison between all models over the three existing SN9 training scenes.

SIFT fails to establish reliable correspondences, producing very few matches and extremely large errors, with a median error exceeding 3000 pixels.

In contrast, both learning-based methods achieve substantially lower errors and a significantly higher proportion of accurate matches. MINIMA achieves the best overall performance, combining the lowest median error of 27.66 pixels with a high number of correspondences. LightGlue produces fewer matches with slightly higher error, possibly attributed by a more selective matching strategy.

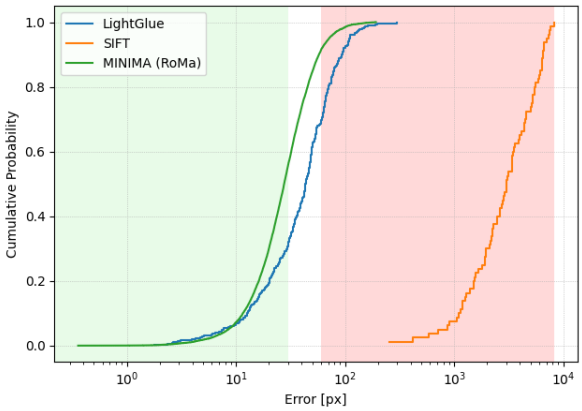

Figure 3. Empirical Cumulative Distribution Function (ECDF) per Model with the green area representing errors $\leq$ 30 px and the red area representing errors $>$ 60 px.

Figure 3 (ECDF) further illustrates the distribution of matching errors. The curve for MINIMA is consistently shifted towards lower error values, confirming its superior accuracy. LightGlue shows a similar but slightly less favourable distribution, while SIFT exhibits a strongly shifted curve with predominantly large errors.

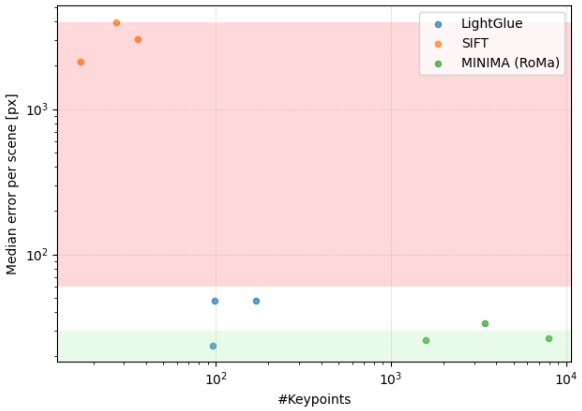

Figure 4. Median error [px] over the number of keypoints in all three SN9 scenes (one dot each).

Figure 4 visualises the relationship between match density and accuracy. While SIFT produces few but unreliable correspondences, both learning-based methods achieve substantially better performance. Notably, MINIMA demonstrates that high match density can be achieved without sacrificing accuracy.

5.2 Qualitative Examples

Figure 5 shows zoomed examples of individual correspondences with increasing error, providing an intuitive understanding of the visual characteristics of successful and unsuccessful matches.

Figures 6 and 7 illustrate representative matching results for different methods. MINIMA produces dense and spatially consistent correspondences with largely coherent matching directions, whereas SIFT yields sparse and visually inconsistent matches, indicating unreliable correspondences.

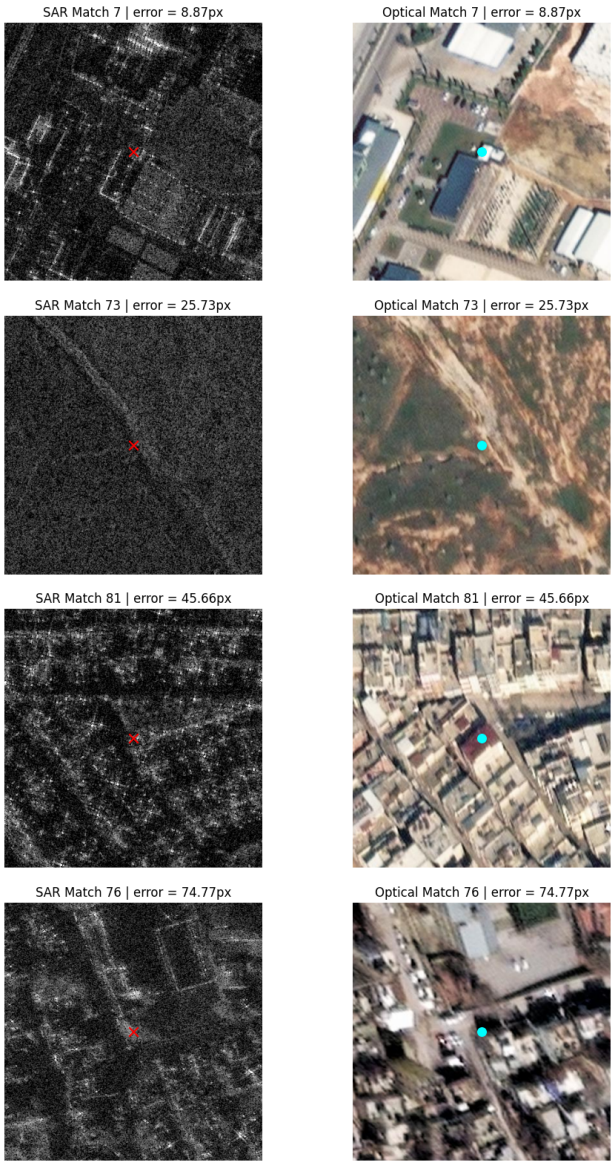

Figure 5. Zoomed-in examples of matches from different scenes with increasing pixel error, showing both positive and negative matches in SuperGlue results.

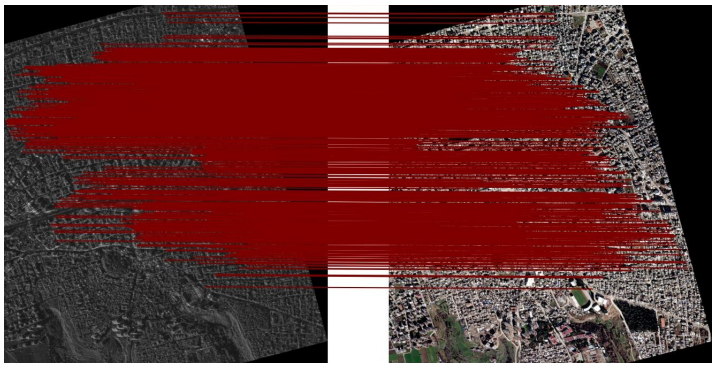

Figure 6. Correspondences across a SAR-optical pair after feature matching using MINIMA’s RoMa model.

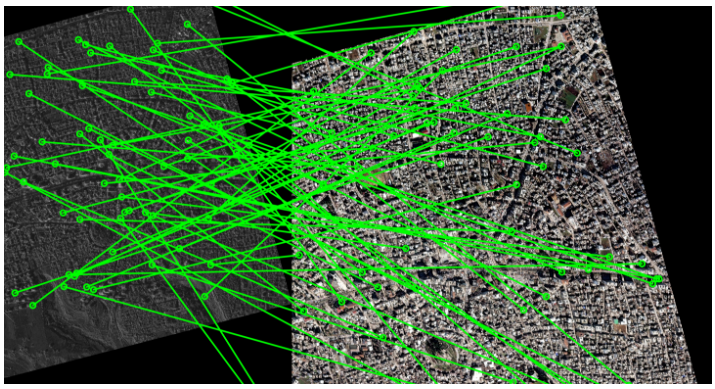

Figure 7. Visualisation of the (mostly) incorrect matches produced by SIFT on the same scene.

5.3 Influence of Scene Structure

To analyse the impact of scene structure on matching performance, matches are separated into ground-based and elevated targets using manually created building masks.

	SIFT	LightGlue	MINIMA
GROUND			
#Matches	49	140	**3202**
Med. Error [px]	3044.88	29.63	**25.65**
ELEVATED			
#Matches	31	224	**9684**
Med. Error [px]	2740.82	48.39	**28.27**

Table 4. Comparison of matching accuracy (median error [px]) between ground-based and elevated targets in the SN9 training dataset.

Table 4 shows the number of matches and the corresponding median errors for both target types. Across all methods, ground-based matches consistently exhibit lower errors than those on elevated structures. This displays the influence of SAR-specific geometric distortions such as layover and foreshortening, which complicate the establishment of consistent correspondences.

For the learning-based methods, this effect is clearly visible. LightGlue achieves a median error of 29.63 pixels on ground targets, compared to 48.39 pixels on elevated structures. Similarly, MINIMA obtains a median error of 25.65 pixels on ground targets and 28.27 pixels on elevated regions. Notably, the performance degradation for elevated targets is less pronounced for MINIMA, demonstrating a higher robustness to geometric distortions. While specifically MINIMA remains robust across target types, the degradation in performance for elevated structures is evident. In contrast, SIFT produces large errors for both categories, with no meaningful distinction between ground and elevated targets, indicating that the method fails to capture reliable correspondences in either case.

	Model Matches on GRD [%]			GRD
	SIFT	LightGlue	MINIMA	Area [%]
Scene 0	80.56	89.58	80.14	71.51
Scene 1	82.35	24.26	25.41	31.62
Scene 2	22.22	13.13	13.49	12.60

Urbanisation

Table 5. Relative distribution of ground (*here:* GRD) matches of each model, compared to the ground area per scene.

As shown in Table 5, for LightGlue and MINIMA the proportion of ground-based matches decreases with increasing scene urbanisation, resembling a clear correlation. While this behaviour appears intuitive, it reveals that the models do not explicitly surpress matches in urban areas despite their increased geometric complexity. The trend is also reflected in the matching accuracy, which decreases from Scene 0 to Scene 2. This suggests that scene structure, and in particular the presence of elevated objects, has a direct impact on the achievable matching accuracy.

Part of the observed error differences between ground and elevated targets may also be influenced by the evaluation procedure. Due to the local interpolation based on CPs, elevation mismatches between the interpolating CPs and the evaluated point can introduce systematic errors. Consequently, even geometrically correct correspondences may exhibit non-zero errors.

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-105-2026

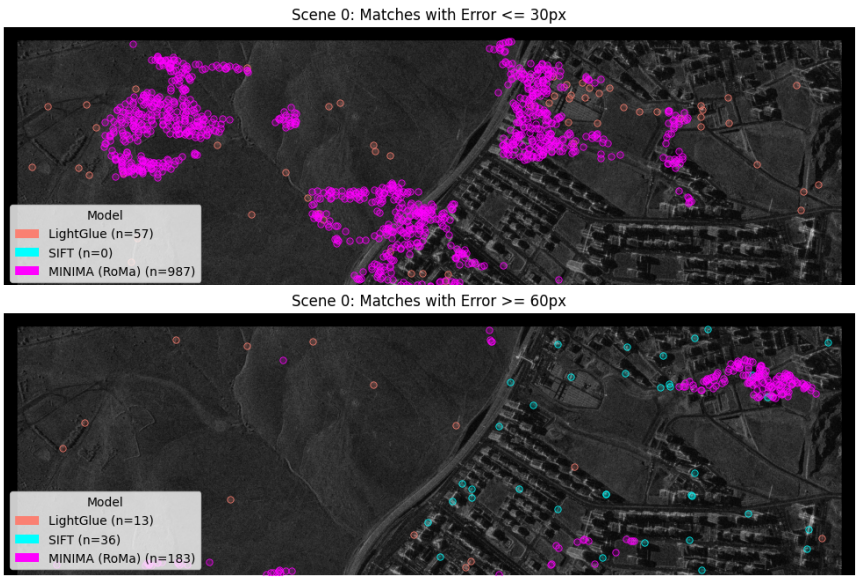

Figure 8. Distribution of positive and negative matches per matcher in one exemplary scene.

5.4 Spatial and Directional Error Characteristics

The error vectors are decomposed into range (x) and azimuth (y) components. Across all scenes and methods, the error in range direction is consistently larger than in azimuth direction ($|\Delta x| > |\Delta y|$), which is characteristic of SAR-specific geometric distortions such as layover. Moreover, the magnitude of the range error increases with scene complexity, as illustrated in Figure 9.

In addition to directional errors, the spatial distribution of matches reveals distinct patterns. While correspondences are generally distributed across the scene, they often appear in locally clustered regions. Figure 8 shows examples of accurate matches (≤ 30 px) and erroneous matches (≥ 60 px) for a representative scene. These spatial patterns also highlight differences in model behaviour. LightGlue produces fewer but spatially consistent correspondences, whereas MINIMA generates denser matches across a wider range of structures. Notably, regions that yield poor matches for one model can still produce accurate correspondences for another, indicating potentially complementary behaviour.

Moreover, the influence of bright scatterers was analysed using a CA-CFAR-based detection of high-intensity SAR pixels. The resulting CFAR-masks were used to identify correspondences located on strong scatterers. As shown in Table 6, SIFT produces a comparatively high proportion of matches on bright scatterers, while LightGlue and MINIMA show lower and more consistent values across scenes. This reflects that classical methods tend to focus stronger on radiometrically distinctive features.

	Bright Scatterers [%]		
	SIFT	**LightGlue**	**MINIMA**
Scene 0	36.11	3.12	7.06
Scene 1	70.59	10.06	17.67
Scene 2	11.11	8.08	4.42

Urbanisation ↓

Table 6. Proportion of matches in close proximity to bright SAR scatterers per scene and model, based on CA-CFAR detection.

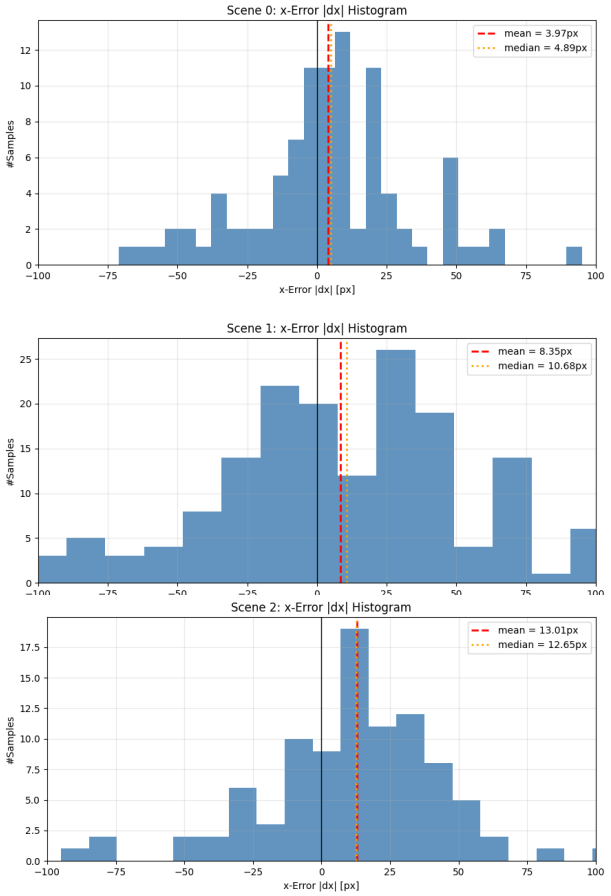

Figure 9. Visualisation of the shift of the mean and median Δx error over the three SN9 test scenes (SuperGlue). The complexity and urbanisation increases from Scene 0 to Scene 2.

However, restricting the evaluation to matches on bright scatterers does not improve the matching accuracy. Instead, a slight increase in error is observed for both LightGlue and MINIMA. This suggests that bright scatterers in the analysed test scenes are often associated with complex urban structures where geo-

metric distortions dominate. These findings challenge the common assumption that strong SAR scatterers provide favourable features for cross-modal matching.

6. Discussion

6.1 Reliability of the Evaluation Setup

Quantifying matching accuracy in SAR-optical image pairs is inherently challenging, particularly in urban environments where the 3D scene structure is not explicitly known. Due to the fundamentally different imaging geometries of SAR and optical sensors, a unique ground truth for point-to-point correspondences does not exist. Effects such as layover, foreshortening, and occlusions introduce ambiguities that cannot be fully resolved, even with manual annotation.

The proposed evaluation relies on manually annotated Control Points (CPs) and a piecewise affine interpolation based on Delaunay triangulation. Both components introduce uncertainty. Manual annotations are subject to localisation errors, while the interpolation accuracy depends on the spatial distribution of CPs, with large or elongated triangles leading to local approximation errors. A leave-one-out validation of the CP-based interpolation yields median errors of up to 11 pixels.

In addition, the evaluation procedure itself introduces systematic bias. The piecewise affine interpolation assumes locally consistent geometry defined by the CPs. When CPs and evaluated points differ in elevation, the interpolation may produce biased projections. As a result, even geometrically correct correspondences can exhibit non-zero errors. This effect cannot be separated from genuine matching errors within the proposed framework.

Despite these limitations, the evaluation remains suitable for comparative analysis. All methods are assessed under identical conditions, and the uncertainty introduced by the CPs and interpolation is significantly smaller than the observed differences between methods. However, it is important to emphasise that this framework does not aim to determine absolute geometric accuracy, but rather to comparatively assess the performance and behaviour of different matching approaches under identical conditions in relation to each other. In this context, the evaluation framework is sufficiently robust to reveal clear trends.

6.2 Interpretation and Implications

The results imply that modern learning-based matching methods can effectively bridge parts of the SAR-optical modality gap, even without retraining. In contrast, classical approaches fail to establish geometrically meaningful correspondences, revealing the limitations of handcrafted descriptors in cross-modal settings. At the same time, matching performance is strongly constrained by scene structure and SAR-specific imaging geometry. Elevated objects and densely built environments introduce systematic errors, particularly in range direction, which cannot be resolved by feature similarity alone. This indicates that a major source of error is not solely the matching algorithm itself, but the underlying geometric inconsistencies between both modalities.

This implies that further improvements are unlikely to arise solely from better feature representations. Instead, integrating geometric constraints or combining complementary matching strategies appears to be a more promising direction. Overall, while learning-based methods significantly improve cross-modal matching, achieving geometrically consistent and reliable correspondences in SAR-optical imagery remains fundamentally limited by the imaging geometry.

7. Conclusion

This paper presented a systematic evaluation of classical and learning-based feature matching methods for SAR-optical image pairs using the SpaceNet9 dataset. A unified, geometry-aware evaluation framework based on CP-driven piecewise affine interpolation was used to ensure a consistent comparison. However, the reported errors should be interpreted as relative measures under a consistent evaluation framework rather than absolute geometric accuracies. While the evaluation method is subject to systematic limitations, the observed differences between methods indicate their relative performance.

The results show that modern learning-based approaches substantially improve matching performance without retraining, whereas classical methods fail to produce reliable correspondences. However, matching accuracy is strongly limited by scene structure and SAR-specific imaging geometry, particularly in densely built environments. These findings indicate that the main limitation of SAR-optical matching is not solely the feature representation, but the inherent geometric inconsistencies between both modalities. Achieving geometrically consistent SAR-optical matching therefore remains an open challenge that requires combining data-driven and physically informed approaches.

References

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. SuperPoint: Self-Supervised Interest Point Detection and Description. *CVPR Workshops*, 337–349.

Hänsch, R., Arndt, J., Dias, P., Potnis, A., Lunga, D., Petrie, D., Bacastow, T. M., 2024. Introducing SpaceNet 9 - Cross-Modal Satellite Imagery Registration for Natural Disaster Responses. *Proceedings of IGARSS*, 234–238.

Hughes, L. H., Schmitt, M., 2021. Comparative Evaluation of Deep Learning-Based SAR-Optical Image Matching Approaches. *Proceedings of IGARSS*, 423–426.

Lindenberger, P., Sarlin, P. E., Pollefeys, M., 2023. LightGlue: Local Feature Matching at Light Speed. *Proceedings of ICCV*, 17627–17638.

Lowe, D. G., 1999. Object Recognition from Local Scale-Invariant Features. *Proceedings of ICCV*, 1150–1157.

Lowe, D. G., 2004. Distinctive Image Features from scale-invariant Keypoints. *Proceedings of ICCV*, 91–110.

Ren, J., Jiang, X., Li, Z., Liang, D., Zhou, X., Bai, X., 2025. MINIMA: Modality Invariant Image Matching. *Proceedings of CVPR*, 23059–23068.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. ORB: An efficient Alternative to SIFT or SURF. *Proceedings of ICCV*, 2564–2571.