GeoCFR: Multimodal Remote Sensing Image Registration via Structural Representation Learning and Posterior Aggregation

Anonymous authors
Paper under double-blind review

Abstract

High-precision registration of multimodal remote sensing images (MRSIs) is fundamental to multi-source information fusion and Earth observation applications. However, cross-modal appearance differences combined with large rotations, translations, and scale changes make reliable correspondences difficult to establish. Sparse methods can handle large initial displacements, but cross-modal differences can result in too few reliable correspondences or uneven spatial coverage; dense methods provide more candidates but do not fully exploit candidate distributions in local responses for accurate global geometry estimation. We therefore propose GeoCFR, which combines cross-modal structural representation learning with local matching posterior aggregation to progressively refine a global affine transformation. The regional moment–relational spectrum descriptor (RMSD) uses regional statistics and multi-ring relationships to obtain an affine initialization; the Posterior-Aggregating Tokens for Affine Refinement (PATAR) module aggregates local candidate distributions and geometric information to predict global affine residuals at two feature scales. We also construct WH-M3R, a dataset covering Wuhan with 343,758 valid cross-modal image pairs across 13 data layers at spatial resolutions of 0.5, 2, 5, and 10 m. Match-30K is used for training and validation, and Match-1K for testing. Under the High protocol on WH-M3R-Match-1K, GeoCFR improves the correct matching rate (CMR) at the 1-, 2-, and 3-pixel thresholds by 7.7, 6.9, and 5.8 percentage points over the strongest baseline at each threshold. Under the High protocol, GeoCFR achieves leading overall zero-shot performance on SOMA-1M-Test, RoadScene, and OS-dataset, demonstrating transfer across datasets under large geometric changes.

1 Introduction

Optical and SAR imagery, maps, and elevation data provide complementary information for multi-source fusion, change detection, and disaster monitoring, all of which rely on high-precision registration (Jiang et al., 2021). However, the same scene structure can differ substantially in brightness, texture, and edge appearance across modalities, making reliable correspondences difficult to establish. Large rotations, translations, and scale differences further complicate matching.

Sparse methods handle large initial displacements through keypoint matching and robust estimation, but cross-modal differences reduce keypoint repeatability, leading to too few reliable correspondences or uneven spatial coverage (Li et al., 2020; Gao et al., 2025a; Lindenberger et al., 2023). Semi-dense and dense methods provide more candidates (Sun et al., 2021; Teed & Deng, 2020; Sun et al., 2025), yet appearance differences and repeated structures can still produce multi-peaked local responses; retaining only the highest response discards alternative candidates and their relative probabilities. High-precision registration under large geometric changes therefore requires both reliable initialization and aggregation of ambiguous local evidence to constrain global geometry estimation.

To this end, we propose GeoCFR, which combines cross-modal structural representation learning with local matching posterior aggregation to progressively refine a global affine transformation. RMSD describes scene structure using regional moments and multi-ring relational spectra to support feature matching and robust affine initialization; PATAR constructs local matching posteriors

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

around locations predicted by the current transformation and aggregates candidate distributions with geometric information across locations to predict affine residuals at two feature scales.

Large registration datasets predominantly focus on SAR–optical pairs, with limited coverage of data types such as maps and elevation (Table 1). To support training and evaluation across modalities and resolutions, we construct WH-M3R by pairing optical imagery, SAR, maps, elevation, and derived data covering the same areas of Wuhan. Training, validation, and independent test splits support comparisons across modality pairs and geometric transformation levels under the same setup.

Our main contributions are:

- We introduce GeoCFR for MRSI registration under large geometric changes, combining cross-modal structural representation learning with global affine refinement. RMSD combines regional statistics and multi-ring relationships for correspondence estimation and affine initialization.
- We develop PATAR to retain matching candidates and their relative probabilities, aggregate local posteriors with geometric information across locations, and predict global affine residuals. Two-scale updates refine the global transformation using ambiguous evidence.
- We construct WH-M3R with 343,758 valid cross-modal image pairs across 13 data layers and four spatial resolutions to support registration across modality combinations and ranges of geometric variation.

2 Related Work

2.1 Multimodal Remote Sensing Datasets

Existing aligned multimodal remote sensing datasets support multi-source representation learning and cross-modal registration. SEN1-2 and SEN12MS provide Sentinel-1/2 imagery of the same areas, with SEN12MS also including land-cover data. Globe230k combines RGB, SAR, NDVI, and DEM data. RST-1M brings together RGB, SAR, NIR, PAN, and multispectral data from multiple datasets. MaRS-16M increases the scale of SAR–optical data, while SOMA-1M provides SAR–optical pairs at multiple spatial resolutions (Schmitt et al., 2018; 2019; Shi et al., 2023; Chen et al., 2026; Yang et al., 2026; Wu et al., 2026a).

For registration, OSdataset provides high-resolution SAR–optical pairs; MultiResSAR and 3MOS span multiple resolutions and data sources; MapData targets digital-map–optical matching; and GCZ covers diverse cross-modal matching settings (Xiang et al., 2020; Zhang et al., 2025; Ye et al., 2025; Wu et al., 2026b; Gao et al., 2025a). These datasets have different emphases. WH-M3R combines 13 data layers, including optical imagery, SAR, maps, and elevation, at spatial resolutions of 0.5, 2, 5, and 10 m to support registration across modality pairs. Table 1 compares representative datasets by data composition, resolution, and size.

2.2 Multimodal Image Matching

Handcrafted feature matching uses descriptors and geometric verification to establish correspondences. SIFT and ORB rely on gradients or intensity comparisons, which can reduce their reliability under nonlinear radiometric differences (Lowe, 2004; Rublee et al., 2011). RIFT, RIFT2, and HOWP use phase and orientation information or polar-coordinate representations to improve cross-modal matching (Li et al., 2020; 2023; Zhang et al., 2023). HOMO-Feature and HIMO combine orientation information, polar-coordinate descriptors, and multi-scale strategies for matching across modalities (Gao et al., 2025a; 2026). MSG constructs a side-window Gaussian scale space and matches descriptors at densely distributed edge keypoints (Zheng et al., 2025). These handcrafted methods produce sets of point correspondences rather than pixel-wise displacement fields.

Learned sparse matchers such as SuperGlue, LightGlue, and MambaGlue aggregate contextual features and assign matches to pre-detected keypoints (Sarlin et al., 2020; Lindenberger et al., 2023; Ryoo et al., 2025). MapGlue and SFA-Net incorporate semantic or structural priors into MRSI matching (Wu et al., 2026b; Gao et al., 2025b). For methods that operate on detected keypoints, cross-modal repeatability and spatial coverage remain important for robust geometric estimation.

Table 1: Comparison of data composition and spatial organization among representative multimodal remote sensing datasets.

Dataset	Year	Modalities	Data layers	GSD (m)	Size	Multi-res.
SEN1-2	2018	SAR, optical	2	10	282K pairs	×
SEN12MS	2019	SAR, optical, LC	3	SAR/optical 10; LC 500 (native)	180K triplets	×
OSdataset	2020	SAR, optical	2	1	13K pairs	×
Globe230k	2023	RGB, SAR, NDVI, DEM	4	RGB 1; SAR 10; NDVI/DEM 30	232K images	×
3MOS	2025	SAR, optical	2	3.5–12.5	113K pairs	✓
GCZ	2025	RGB, PAN, NIR, SWIR, MSI, HSI, SAR, LiDAR, digital maps, etc.	10+	–	4,752 samples	×
MultiResSAR	2025	SAR, optical	2	0.16–10	11K pairs	✓
MapData	2026	Digital maps, optical	2	1	122K pairs	×
MaRS-16M	2026	SAR, optical	2	0.35	16.8M pairs	×
RST-1M	2026	RGB, SAR, NIR, PAN, MS	5	–	1.2M pairs	✓
SOMA-1M	2026	SAR, optical	2	0.5/3/10	1.3M pairs	✓
WH-M3R (ours)	–	RGB, PAN, MS-RGB/NIR, HS-RGB/NIR, SAR, Map, DEM, Depth, etc.	13	0.5/2/5/10	343,758 pairs	✓

Note: Multi-res. indicates aligned image pairs at more than one GSD; differences in native resolution alone do not constitute multi-resolution coverage.

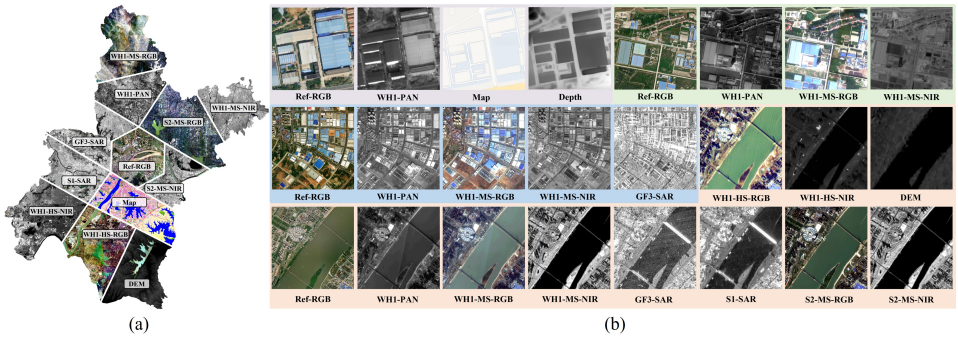
Detector-free methods establish correspondences on regular feature grids. LoFTR, MatchFormer, EfficientLoFTR, and XoFTR obtain semi-dense matches through coarse-to-fine feature interaction and local refinement (Sun et al., 2021; Wang et al., 2022; 2024; Tuzcuoğlu et al., 2024). These methods can yield more matches in weakly textured regions. However, when matching optical imagery with SAR, maps, or elevation data, large appearance differences can still lead to incorrect coarse matches and limit subsequent refinement.

Learned dense methods predict pixel-level correspondences or displacement fields. GMFlow estimates optical flow through global matching. RAFT iteratively refines flow using feature correlations. WAFT replaces cost volumes with high-resolution feature warping and recurrent flow updates. DKM and RoMa estimate dense correspondences and their confidence for geometric estimation (Xu et al., 2022; Teed & Deng, 2020; Wang & Deng, 2026; Edstedt et al., 2023; 2024). For cross-modal remote sensing registration, GDROS introduces explicit geometric guidance, CRFT refines correspondences through recurrent feature flow, and RBE-Flow optimizes dense flow using recursive Bayesian estimation (Sun et al., 2025; Liu et al., 2026; Ding et al., 2026).

GeoCFR targets cross-modal image pairs whose relative transformation can be approximated by an affine model. It combines structural representation learning with local matching posterior aggregation to address appearance differences and local matching ambiguity during affine refinement.

3 WH-M3R Dataset

WH-M3R contains aligned MRSIs of Wuhan at four spatial resolutions.



(a) (b)

Figure 1: Overview of WH-M3R. (a) Spatial coverage of the principal data layers across Wuhan. (b) Representative multimodal samples at spatial resolutions of 0.5, 2, 5, and 10 m. Background colors indicate different spatial resolutions.

Table 2: Statistics of valid image pairs and matching subsets in WH-M3R at each spatial resolution.

GSD	Target data layers paired with Ref-RGB	Total pairs	Train pairs	Test pairs	Match-30K	Match-1K
0.5 m	WH1-PAN, Map, Depth	258,996	255,996	3,000	11,485	300
2 m	WH1-PAN, WH1-MS-RGB, WH1-MS-NIR	48,105	46,305	1,800	8,000	200
5 m	WH1-PAN, WH1-MS-RGB, WH1-MS-NIR, GF3-SAR	21,957	20,757	1,200	5,000	100
10 m	WH1-MS-RGB, WH1-MS-NIR, WH1-PAN, GF3-SAR, S1-SAR, S2-MS-RGB, S2-MS-NIR, WH1-HS-RGB, WH1-HS-NIR, DEM	14,700	13,700	1,000	5,515	400
Total	–	343,758	336,758	7,000	30,000	1,000

Note: Each valid pair uses a Ref-RGB tile at the same GSD as its reference. The target-layer composition of each matching subset is reported in Appendix Table 7.

Data organization. WH-M3R includes reference RGB imagery (Ref-RGB), Wuhan-1, Gaofen-3, and Sentinel-1/2 imagery, rasterized maps, elevation, and derived depth (Figure 1). Each target layer is paired with Ref-RGB imagery covering the same area. Using georeferencing information, we first resample the valid data layers to target GSDs of 0.5, 2, 5, and 10 m and obtain an initial alignment. We then correct residual misalignment with a coarse-to-fine registration procedure following SOMA-1M (Wu et al., 2026a) to produce aligned 512×512-pixel tiles. Resolution-dependent retention rules are given in Appendix A.

Data layers and dataset size. The 13 data layers are Ref-RGB, WH1-PAN, WH1-MS-RGB/NIR, WH1-HS-RGB/NIR, GF3-SAR, S1-SAR, S2-MS-RGB/NIR, Map, DEM, and Depth. They are not 13 independent imaging modalities: some RGB/NIR products share source sensors, and Depth is derived from Ref-RGB. WH-M3R comprises 111,814 reference tiles and 343,758 target tiles, yielding 455,572 images and 343,758 reference–target pairs. Table 2 reports counts at each output GSD; Appendix A provides further dataset details.

Splits and matching subsets. The training partition and independently constructed test partition contain 336,758 and 7,000 pairs, respectively. We sample 30,000 pairs from the training partition to form Match-30K, which is randomly split at the image-pair level into training and validation sets in a 9:1 ratio. Match-1K consists of 1,000 pairs sampled from the independent test partition, with 100 pairs for each of ten Ref-RGB–target-layer pairings, and is used only for final evaluation. Match-1K is geographically separated from Match-30K by latitude and longitude bounds, with no spatial overlap between the two subsets. Appendix Table 7 gives the composition; Appendix A.3 describes the detector supervision.

4 GeoCFR

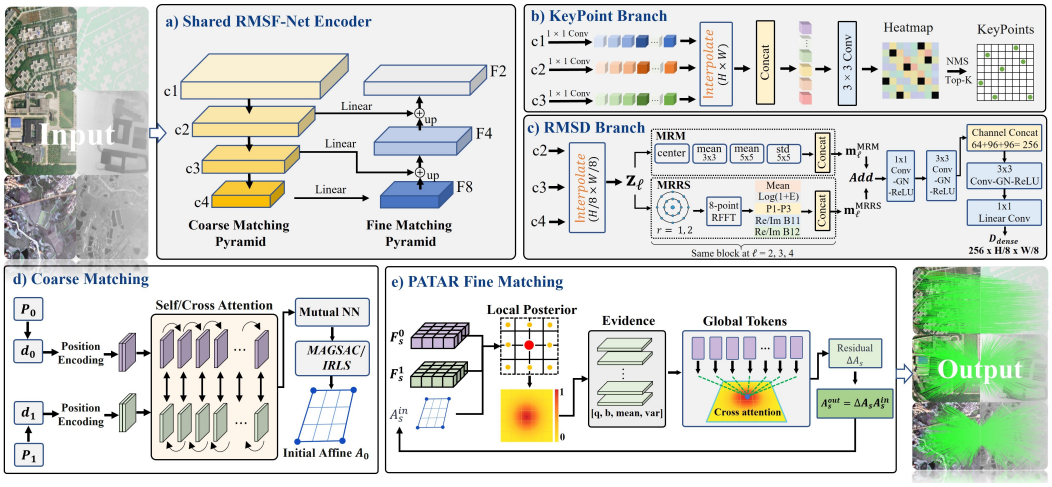


Figure 2: Overview of GeoCFR. (a–c) RMSF-Net extracts structural keypoints, RMSD descriptors, and fine-matching features; (d) coarse matching estimates $\mathbf{A}_0$; and (e) PATAR aggregates local matching posteriors at P4 and P2, the stride-4 and stride-2 feature grids, for global affine refinement.

4.1 OVERVIEW

Given a source image I_0 and a target image I_1, GeoCFR estimates an affine transformation $\mathbf{A}^*$ from source to target. The corresponding full-resolution displacement field is $\mathbf{u}(\mathbf{x}) = \mathbf{A}^*[\mathbf{x}^\mathsf{T}, 1]^\mathsf{T} - \mathbf{x}$. In Figure 2, the Regional Moment-Spectral Feature Network (RMSF-Net) extracts structural keypoints, descriptors, and multi-scale fine-matching features. Attention-based coarse matching estimates $\mathbf{A}_0$; PATAR then aggregates local matching posteriors at P4 and P2 to predict global affine residuals, yielding $\mathbf{A}^* = \mathbf{A}_2$.

4.2 CROSS-MODAL STRUCTURAL FEATURES

Backbone and keypoints. RMSF-Net applies local contrast normalization and encodes features $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4$ at strides 1, 2, 4, and 8. The keypoint head projects $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3$, upsamples the latter two to full resolution, and concatenates them to predict a score map. Non-maximum suppression (NMS) and top-K filtering select keypoints. A top-down pyramid supplies fine-matching features; fusion and sampling details are in Appendix B.

Regional moment–relational spectrum descriptor (RMSD). Encoder features at levels $\ell \in \{2, 3, 4\}$ are aligned to the one-eighth-resolution grid. We describe a fixed level and center location, omitting the level index and writing position arguments only for neighborhood sampling. RMSD combines multi-scale regional moments (MRM) with a multi-ring relational spectrum (MRRS). MRM concatenates the center response, local means, and standard deviation:

$$\mathbf{m}^{\mathrm{MRM}} = [\mathbf{z}, \boldsymbol{\mu}_3, \boldsymbol{\mu}_5, \boldsymbol{\sigma}_5]. \tag{1}$$

Here, $\mathbf{z}$ is the aligned feature, $\boldsymbol{\mu}_3, \boldsymbol{\mu}_5$ are 3×3 and 5×5 neighborhood means, and $\boldsymbol{\sigma}_5$ is the 5×5 standard deviation. MRRS measures center–neighbor similarity at eight uniformly spaced angular offsets $\boldsymbol{\delta}_{r,k}$ on each ring of radius $r \in \{1, 2\}$, then applies an orthonormal discrete Fourier transform:

$$\begin{aligned} a_{r,k} &= \langle \widehat{\mathbf{z}}(\mathbf{x}), \widehat{\mathbf{z}}(\mathbf{x} + \boldsymbol{\delta}_{r,k}) \rangle, \qquad k = 0, \dots, 7, \\ F_{r,m} &= \frac{1}{\sqrt{8}} \sum_{k=0}^{7} a_{r,k} e^{-\mathrm{j}2\pi mk/8}, \qquad m = 0, \dots, 4. \end{aligned} \tag{2}$$

Here, $\widehat{\mathbf{z}}(\mathbf{x})$ denotes $\mathbf{z}(\mathbf{x})$ after L_2 normalization along the channel dimension; j is the imaginary unit, with $\mathrm{j}^2 = -1$. Conjugate symmetry of the real-valued sequence leaves five nonredundant coefficients, $m = 0, \dots, 4$, including four nonzero-frequency terms. We define the nonzero-frequency energy, normalized power, and bispectrum as

$$\begin{aligned} E_r &= \sum_{m=1}^{4} |F_{r,m}|^2, \qquad P_{r,m} = \frac{|F_{r,m}|^2}{E_r + \epsilon}, \\ B_r^{m,n} &= \frac{F_{r,m} F_{r,n} F_{r,m+n}^*}{|F_{r,m}||F_{r,n}||F_{r,m+n}| + \epsilon}. \end{aligned} \tag{3}$$

Power uses $m \in \{1, 2, 3\}$, and bispectra use $(m, n) \in \{(1, 1), (1, 2)\}$. Here, $(\cdot)^*$ denotes complex conjugation and $\epsilon > 0$ ensures numerical stability. Power describes energy distribution; bispectra retain inter-frequency phase relationships. Each ring contributes its mean relation, log energy, three power terms, and the real and imaginary parts of two bispectra. Concatenating both nine-channel vectors gives $\mathbf{m}^{\mathrm{MRRS}} \in \mathbb{R}^{18}$. These statistics promote rotation stability; Appendix B.1 gives the derivation and its scope. MRM and MRRS features from each level are projected and fused within and across levels into an L_2-normalized, 256-dimensional descriptor map, sampled at keypoints.

4.3 COARSE MATCHING AND AFFINE INITIALIZATION

Keypoint coordinates are centered and normalized by half the larger image dimension. Encoded position, radius, and polar angle are added to the projected RMSD descriptors. Stacked self-attention and bidirectional cross-attention produce descriptors $\widetilde{\mathbf{d}}_i^0, \widetilde{\mathbf{d}}_j^1$ and matchability scores m_i^0, m_j^1, yielding

$$S_{ij} = \langle \widetilde{\mathbf{d}}_i^0, \widetilde{\mathbf{d}}_j^1 \rangle / \tau, \qquad \pi_{ij} = \mathrm{Softmax}_j(S_{ij}) \, \mathrm{Softmax}_i(S_{ij}) m_i^0 m_j^1. \tag{4}$$

Here, τ is learnable. Mutual-nearest-neighbor selection and confidence thresholding produce sparse correspondences. Confidence-initialized iteratively reweighted least squares (IRLS) estimates $\mathbf{A}_0$ during training; inference uses USAC-MAGSAC. Coordinate encoding and matching-feature construction are detailed in Appendix B.

4.4 Posterior-Based Affine Refinement

Local posterior. At feature stride $s \in \{4, 2\}$, a shared projection produces fine-matching features for the source and target images. We first describe a single feature scale, omitting its index and denoting the feature maps by $\mathbf{G}^0, \mathbf{G}^1$. The current affine estimate maps each source-grid location $\mathbf{x}_i$ to a target search center $\mathbf{b}_i$; both are expressed in coordinates of the current feature grid. Integer offsets $\mathcal{O} = \{-r, \dots, r\}^2$ define the local search:

$$C_i(\mathbf{o}) = \langle \mathbf{G}^0(\mathbf{x}_i), \mathbf{G}^1(\mathbf{b}_i + \mathbf{o}) \rangle, \qquad q_i(\mathbf{o}) = \frac{\chi_i(\mathbf{o}) \exp(C_i(\mathbf{o})/\tau)}{\sum_{\mathbf{o}' \in \mathcal{O}} \chi_i(\mathbf{o}') \exp(C_i(\mathbf{o}')/\tau)}. \tag{5}$$

Here, τ is the posterior temperature for the current feature scale, specified separately from the coarse-matching temperature. The validity mask χ_i excludes out-of-bounds candidates; positions without valid candidates are excluded from aggregation. The posterior entries are arranged into a vector $\mathbf{q}_i$ in a fixed offset order, with zero probability assigned to invalid candidates.

Posterior aggregation. Selecting only the highest-scoring candidate discards information about alternative matches and their relative probabilities. PATAR retains the full local matching posterior $\mathbf{q}_i$, its mean offset $\boldsymbol{\mu}_i = \sum_{\mathbf{o} \in \mathcal{O}} q_i(\mathbf{o})\mathbf{o}$, and the trace of its covariance $\nu_i = \sum_{\mathbf{o} \in \mathcal{O}} q_i(\mathbf{o}) \|\mathbf{o} - \boldsymbol{\mu}_i\|_2^2$. Together with the normalized target search center $\overline{\mathbf{b}}_i$, these form

$$\mathbf{e}_i = \mathrm{Concat}\left[\mathbf{q}_i, \overline{\mathbf{b}}_i, \boldsymbol{\mu}_i/r, \nu_i/(2r^2)\right]. \tag{6}$$

Context encoding and pooling yield an evidence sequence $\mathbf{E}$. Standard multi-head cross-attention aggregates local matching posteriors and geometric cues into learnable global tokens. Self-attention then exchanges information among these tokens (Vaswani et al., 2017). An FFN, mean pooling, and an affine head predict one six-parameter residual $\boldsymbol{\xi}$ for the whole image. Dense local matching information thus refines a single global affine transformation rather than independent pixel displacements.

Affine update and cascade. Restoring the feature-scale index s, let $\boldsymbol{\xi}_s = (t_x, t_y, \omega, \alpha, \beta, \gamma)$ and $\mathbf{t}_s = (t_x, t_y)^\mathsf{T}$. All affine matrices are expressed in full-resolution pixel coordinates, and $\mathbf{t}_s$ is measured in pixels. With image center $\mathbf{c} = \frac{1}{2}(W - 1, H - 1)^\mathsf{T}$, the residual transformation is

$$\mathbf{K}_s = \begin{bmatrix} \alpha + \beta & \gamma - \omega \\ \gamma + \omega & \alpha - \beta \end{bmatrix}, \quad \mathbf{L}_s = \exp(\mathbf{K}_s), \quad \Delta\mathbf{A}_s = [\mathbf{L}_s, \mathbf{c} + \mathbf{t}_s - \mathbf{L}_s\mathbf{c}]. \tag{7}$$

Here, exp denotes the matrix exponential. P4 searches around $\mathbf{A}_0$; P2 searches around the updated $\mathbf{A}_4$. Bars denote homogeneous affine matrices, giving

$$\overline{\mathbf{A}}_4 = \overline{\Delta\mathbf{A}}_4 \overline{\mathbf{A}}_0, \qquad \overline{\mathbf{A}}_2 = \overline{\Delta\mathbf{A}}_2 \overline{\mathbf{A}}_4, \tag{8}$$

with $\mathbf{A}^* = \mathbf{A}_2$ defining the final displacement field.

4.5 Supervision

In training stage 1, RMSF-Net and the coarse matcher are optimized jointly: $\mathcal{L}_c$ supervises positive and negative assignments, unmatched points, and matchability; $\mathcal{L}_p$ is weighted binary cross-entropy on keypoint score maps. Training stage 2 freezes the shared encoder and coarse matcher and optimizes the fine-matching pyramid and PATAR:

$$\mathcal{L}^{(1)} = \mathcal{L}_c + \lambda_p \mathcal{L}_p, \qquad \mathcal{L}^{(2)} = \lambda_{\mathrm{corr}}\mathcal{L}_{\mathrm{corr}} + \lambda_{\mathrm{off}}\mathcal{L}_{\mathrm{off}} + \lambda_{\mathrm{stage}}\mathcal{L}_{\mathrm{stage}} + \lambda_{\mathrm{final}}\mathcal{L}_{\mathrm{final}}. \tag{9}$$

Local matching posteriors are supervised by KL divergence from Gaussian targets, and their mean offsets by a Smooth L1 loss. The stage loss compares affine estimates on a fixed grid, while the final loss applies a Charbonnier penalty to the displacement error. Appendix C defines the losses, coordinate units, and valid supervision regions; keypoint pseudo-labels are described in Appendix A.3.

Table 3: Geometric transformation ranges for the Low, Medium, and High protocols.

Protocol	Rotation	Translation	Scale
Low	$[-36^\circ, 36^\circ]$	$\pm 0.1W,\ \pm 0.1H$	$[0.9, 1.1]$
Medium	$[-90^\circ, 90^\circ]$	$\pm 0.2W,\ \pm 0.2H$	$[0.8, 1.2]$
High	$[-180^\circ, 180^\circ]$	$\pm 0.3W,\ \pm 0.3H$	$[0.7, 1.3]$

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We train and validate on WH-M3R-Match-30K and test on WH-M3R-Match-1K, using the model trained for each of the Low, Medium, and High protocols. To assess cross-dataset transfer, we conduct zero-shot evaluations on three external datasets. SOMA-1M-Test (Wu et al., 2026a) contains 1,000 optical–SAR image pairs at spatial resolutions of 0.5, 3, and 10 m; RoadScene (Xu et al., 2020) contains 221 optical–infrared image pairs; and OSdataset (Xiang et al., 2020) contains 424 optical–SAR image pairs. All zero-shot evaluations use the High protocol, with each learning-based method using its WH-M3R-trained High model without target-dataset fine-tuning.

Geometric transformation protocols. We define three geometric transformation protocols, Low, Medium, and High, by varying rotation, translation, and isotropic scaling. The sampled transformations belong to the similarity-transform subset of the affine model and exclude shear and anisotropic scaling. Both views are sampled entirely from valid pixels of the original aligned image pair, without synthetic black borders or padding. The source view I_0 uses a centered reference crop, and the target view I_1 uses a transformed moving crop; both share an adaptively chosen context scale. Appendix D specifies the sampling procedure, validity checks, and ground-truth construction.

The three evaluation sets are pre-generated with fixed random seeds so that every method uses the same samples. All metrics are computed only over valid pixels whose ground-truth correspondences remain within the target image.

Training strategy. All learning-based methods follow the same Low-to-Medium-to-High curriculum. Each successive curriculum level starts from the best validation checkpoint of the preceding level, and each protocol is evaluated with its corresponding model. GeoCFR's two training stages, coarse matching and affine refinement, are defined in Section 4.5.

Evaluation metrics. We report the correct matching rate at threshold δ (CMR@δ). It is the percentage of test pairs whose mean endpoint error (EPE) does not exceed δ pixels. We use thresholds of 1, 2, and 3 pixels; higher values indicate better performance.

Table 4: Quantitative comparison on WH-M3R-Match-1K (CMR, %).

	Low			Medium			High		
Method	1 px	2 px	3 px	1 px	2 px	3 px	1 px	2 px	3 px
MSG	1.9	15.7	29.3	2.6	11.1	22.1	0.6	4.4	8.7
HIMO	0.1	1.3	1.9	0.0	1.3	2.6	0.2	1.1	1.8
LG	1.6	27.2	51.8	0.9	19.1	39.3	0.6	11.0	25.0
ELoFTR	20.9	45.7	60.9	16.2	37.1	52.5	12.4	30.1	44.0
XoFTR	20.3	46.3	57.8	17.3	36.6	46.1	8.1	26.5	35.4
RAFT	34.7/42.6	66.9/68.8	79.0/<u>80.5</u>	14.3/29.2	51.3/<u>58.3</u>	65.9/69.7	2.5/18.3	26.5/49.7	43.9/<u>62.9</u>
GMFlow	1.2/5.1	18.0/23.0	40.0/44.2	0.0/0.0	0.1/1.9	1.9/5.7	0.0/0.0	0.0/0.3	0.0/0.9
GDROS	<u>43.1</u>	<u>68.9</u>	79.6	28.8	57.3	<u>70.4</u>	10.4	36.1	53.7
WAFT	36.7/41.3	67.1/67.9	77.2/77.6	21.7/<u>32.6</u>	54.4/55.6	66.9/67.4	11.4/<u>26.6</u>	46.9/<u>50.1</u>	58.3/59.8
GeoCFR	**48.5**	**71.0**	**80.7**	**40.2**	**63.5**	**74.1**	**34.3**	**57.0**	**68.7**

Entries separated by a slash denote original flow / affine-fitted results. Best and second-best values across all reported variants are bold and underlined, respectively.

Baselines. We compare two handcrafted methods, MSG and HIMO, and the learned sparse matcher SuperPoint+LightGlue (LG). We also include the semi-dense matchers ELoFTR and XoFTR and the dense displacement estimators RAFT, GMFlow, GDROS, and WAFT. All learning-based methods are retrained on WH-M3R-Match-30K with 512×512 inputs. Outputs from MSG, HIMO, LG, ELoFTR, and XoFTR are converted into dense displacement fields induced by global affine transfor-

mations using a common USAC-MAGSAC procedure with a 0.5-pixel reprojection threshold. For RAFT, GMFlow, and WAFT, we report both raw-flow and affine-projected results; full settings are given in Appendix B.

5.2 RESULTS

Table 4 shows that GeoCFR achieves the highest CMR at every threshold under all three protocols, with larger margins at stricter thresholds. At the 1-pixel threshold, it reaches 48.5%, 40.2%, and 34.3%, outperforming the strongest baselines by 5.4, 7.6, and 7.7 percentage points, respectively. Under the High protocol, it achieves 57.0% and 68.7% at the 2- and 3-pixel thresholds, with margins of 6.9 and 5.8 points. These comparisons include affine-fitted dense baselines and show that GeoCFR enables more image pairs to meet strict accuracy requirements under larger geometric transformations. Appendix Tables 11 and 12 report detailed results for Ref-RGB paired with each of the ten target data layers and for eight additional modality combinations, respectively.

Table 5: Zero-shot transfer from WH-M3R-Match-30K to SOMA-1M-Test, RoadScene, and OSdataset under the High protocol (CMR, %).

	SOMA-1M-Test			RoadScene			OSdataset		
Method	1 px	2 px	3 px	1 px	2 px	3 px	1 px	2 px	3 px
MSG	0.2	1.1	2.2	0.0	1.4	5.4	0.0	0.2	0.7
HIMO	0.0	0.1	0.3	0.0	0.0	0.0	0.0	0.0	0.2
LG	0.0	0.1	1.2	0.0	0.5	4.5	0.0	1.9	5.2
ELoFTR	0.6	6.6	9.9	0.0	0.0	4.1	0.0	0.9	4.5
XoFTR	0.6	3.6	6.0	0.0	0.0	0.5	0.0	0.0	0.9
RAFT	0.2/3.9	8.4/24.1	20.1/<u>38.4</u>	0.0/<u>0.5</u>	1.8/<u>10.0</u>	12.2/**27.6**	0.0/1.4	2.8/10.4	9.2/23.1
GMFlow	0.0/0.0	0.0/0.0	0.0/0.2	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.2
GDROS	2.8	18.1	33.5	<u>0.5</u>	8.6	21.3	0.2	5.4	17.7
WAFT	3.3/<u>10.6</u>	25.3/<u>28.2</u>	35.8/37.2	0.0/0.0	4.1/8.6	20.8/<u>24.9</u>	0.2/**3.1**	9.9/<u>11.1</u>	25.5/<u>27.4</u>
GeoCFR	**13.9**	**31.3**	**41.0**	**0.9**	**10.4**	<u>24.9</u>	<u>1.9</u>	**13.0**	**30.2**

Slash-separated entries denote original/affine-fitted results; bold and underlined values are best and second-best across all variants.

Table 5 shows that GeoCFR achieves leading overall zero-shot performance under the High protocol. It ranks first at all three thresholds on SOMA-1M-Test and at two thresholds each on RoadScene and OSdataset; the remaining two results rank second or tied for second. Without target-dataset fine-tuning, these results demonstrate transfer across the evaluated optical–SAR and optical–infrared pairs under large geometric transformations.

Figure 3 shows that GeoCFR retains more inliers with broader spatial coverage than several baselines. The visualization compares sampled correspondences from each method's final displacement field rather than raw keypoint counts.

5.3 ABLATION STUDIES

Table 6: Ablation of the RMSD components and refinement strategies on WH-M3R-Match-1K (CMR, %). Peak-Affine uses P4→P2 refinement; (5a)/(5b) report PATAR outputs after the P4/P2 refinement stages.

		Low			Medium			High		
Stage	Configuration	1 px	2 px	3 px	1 px	2 px	3 px	1 px	2 px	3 px
Coarse	(1) Baseline	20.7	55.7	69.8	15.1	47.5	64.1	6.7	37.0	56.8
	(2) +MRM	28.0	61.1	73.3	13.4	51.6	67.4	8.9	39.9	58.0
	(3) +MRM+MRRS	28.7	62.2	74.1	18.9	53.6	68.2	11.3	44.4	61.6
Fine	(4) Peak-Affine	23.2	58.1	74.1	18.5	53.0	67.8	10.0	47.4	63.4
	(5a) PATAR after P4	44.0	69.7	79.2	36.1	61.0	72.4	21.2	51.7	64.5
	(5b) PATAR after P2	**48.5**	**71.0**	**80.7**	**40.2**	**63.5**	**74.1**	**34.3**	**57.0**	**68.7**

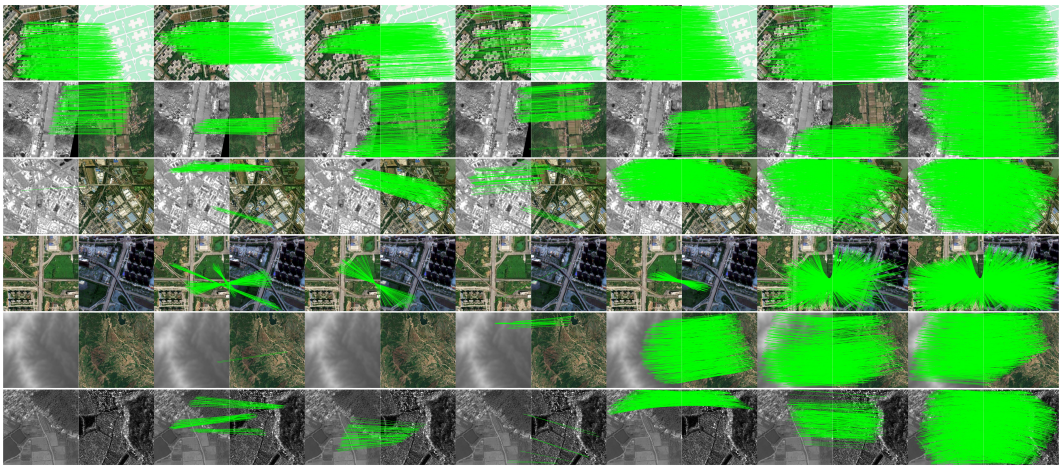


Figure 3: Qualitative comparison. For each image pair, 2,048 source pixels are uniformly sampled from the ground-truth valid region. Green lines denote correspondences from each method's final displacement field with endpoint error no greater than 1.5 pixels.

Structural descriptor. The baseline uses the center feature without MRM or MRRS. MRM raises CMR@1 under the Low protocol from 20.7% to 28.0%. Adding MRRS improves every metric, with CMR@1 gains of 0.7, 5.5, and 2.4 points under the Low, Medium, and High protocols, respectively.

Posterior aggregation. Peak-Affine, which uses local correlation peaks, yields lower CMR@1 than the complete coarse model under all three protocols. PATAR aggregates full local posteriors and geometric cues: its P4 output already outperforms Peak-Affine on every metric, while its P2 output exceeds Peak-Affine in CMR@1 by 25.3, 21.7, and 24.3 points under the Low, Medium, and High protocols.

Successive refinement. Adding P2 improves every metric, including a 13.1-point gain in CMR@1 under the High protocol. We omit full-resolution P1 refinement to limit computational cost and inference time.

6 Discussion

Dataset value. This work evaluates WH-M3R for cross-modal registration; its potential for multi-modal fusion, cross-modal generation, and pretraining remains to be investigated.

Method limitations and future work. GeoCFR aligns images using a global affine transformation, which cannot fully correct parallax and complex local deformations caused by terrain relief or differences in viewing geometry. Future work could combine global alignment with local deformation modeling and develop structural representations that are more robust to differences in sensor characteristics and imaging conditions.

7 Conclusion

We propose GeoCFR, an MRSI registration method that learns cross-modal structural descriptors to obtain an initial affine transformation. GeoCFR then aggregates local matching posteriors to progressively refine the transformation. We also introduce the WH-M3R dataset to support registration research across modality combinations and spatial resolutions. Experiments on WH-M3R-Match-1K show that GeoCFR achieves leading registration performance under all three geometric transformation protocols, with larger margins at stricter accuracy thresholds. Zero-shot evaluations on SOMA-1M-Test, RoadScene, and OSdataset further show that GeoCFR transfers well across datasets under large geometric transformations.

AI USE STATEMENT

Generative AI assisted with translation, language polishing, LaTeX formatting, literature searches, the generation and review of portions of the code, and checking the algebraic consistency and presentation of mathematical derivations. Qwen3-VL-8B-Instruct was also used to generate automatic scene labels for the descriptive statistics in Appendix A.1. Generative AI was not used to generate registration evaluation results. All AI-assisted mathematical content and code were independently reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Pierre Charbonnier, Laure Blanc-Féraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of the IEEE International Conference on Image Processing*, volume 2, pp. 168–172, 1994. doi: 10.1109/ICIP.1994.413553.

Haoyang Chen, Jing Zhang, Hebaixu Wang, Shiqin Wang, Pohsun Huang, Jiayuan Li, Haonan Guo, Di Wang, Zheng Wang, and Bo Du. Any2Any: Unified arbitrary modality translation for remote sensing. *arXiv preprint arXiv:2603.04114*, 2026. URL https://arxiv.org/abs/2603.04114.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, 2018. doi: 10.1109/CVPRW.2018.00060.

Mengzhu Ding, Xin Song, Xiaoke Ding, Hongwei Ding, and Xuecong Liu. RBE-Flow: Recurrent Bayesian estimation on feature manifolds for cross-modal registration. *arXiv preprint arXiv:2606.30492*, 2026. URL https://arxiv.org/abs/2606.30492.

Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17765–17775, 2023. doi: 10.1109/CVPR52729.2023.01704.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19790–19800, 2024. doi: 10.1109/CVPR52733.2024.01871.

Chenzhong Gao, Wei Li, and Desheng Weng. HOMO-Feature: Cross-arbitrary-modal image matching with homomorphism of organized major orientation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10538–10548, 2025a. doi: 10.1109/ICCV51701.2025.00981.

Chenzhong Gao, Wei Li, Desheng Weng, Ran Tao, Xiang-Gen Xia, and Qian Du. HIMO: Cross-arbitrary-modality image invariant feature transform with hierarchical intrinsic major orientation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(8):9001–9018, 2026. doi: 10.1109/TPAMI.2026.3675028.

Tian Gao, Chaozhen Lan, Wenjun Huang, and Sheng Wang. SFA-Net: A SAM-guided focused attention network for multimodal remote sensing image matching. *ISPRS Journal of Photogrammetry and Remote Sensing*, 223:188–206, 2025b. doi: 10.1016/j.isprsjprs.2025.02.032.

Xingyu Jiang, Jiayi Ma, Guobao Xiao, Zhenfeng Shao, and Xiaojie Guo. A review of multimodal image matching: Methods and applications. *Information Fusion*, 73:22–71, 2021. doi: 10.1016/j.inffus.2021.02.012.

Jiayuan Li, Qingwu Hu, and Mingyao Ai. RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing*, 29:3296–3310, 2020. doi: 10.1109/TIP.2019.2959244.

Jiayuan Li, Pengcheng Shi, Qingwu Hu, and Yongjun Zhang. RIFT2: Speeding-up RIFT with a new rotation-invariance technique. *arXiv preprint arXiv:2303.00319*, 2023. URL https://arxiv.org/abs/2303.00319.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e4cd50120b6d7e8daff1749d6bbaa889-Abstract-Conference.html.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17627–17638, 2023. doi: 10.1109/ICCV51070.2023.01616.

Xuecong Liu, Mengzhu Ding, Zixuan Sun, Zhang Li, and Xichao Teng. CRFT: Consistent-recurrent feature flow transformer for cross-modal image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34784–34794, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_CRFT_Consistent-Recurrent_Feature_Flow_Transformer_for_Cross-Modal_Image_Registration_CVPR_2026_paper.html.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. doi: 10.1023/B:VISI.0000029664.99615.94.

Jiangwei Ren, Xingyu Jiang, Zizhuo Li, Dingkang Liang, Xin Zhou, and Xiang Bai. MINIMA: Modality invariant image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23059–23068, 2025. doi: 10.1109/CVPR52734.2025.02147.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2564–2571, 2011. doi: 10.1109/ICCV.2011.6126544.

Kihwan Ryoo, Hyungtae Lim, and Hyun Myung. MambaGlue: Fast and robust local feature matching with Mamba. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 5758–5765, 2025. doi: 10.1109/ICRA55743.2025.11128473.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4938–4947, 2020. doi: 10.1109/CVPR42600.2020.00499.

Michael Schmitt, Lloyd Haydn Hughes, and Xiao Xiang Zhu. The SEN1-2 dataset for deep learning in SAR-optical data fusion. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-1:141–146, 2018. doi: 10.5194/isprs-annals-IV-1-141-2018.

Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. SEN12MS: A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W7:153–160, 2019. doi: 10.5194/isprs-annals-IV-2-W7-153-2019.

Qian Shi, Da He, Zhengyu Liu, Xiaoping Liu, and Jingqian Xue. Globe230k: A benchmark dense-pixel annotation dataset for global land cover mapping. *Journal of Remote Sensing*, 3:0078, 2023. doi: 10.34133/remotesensing.0078.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8922–8931, 2021. doi: 10.1109/CVPR46437.2021.00881.

Zixuan Sun, Shuaifeng Zhi, Ruize Li, Jingyuan Xia, Yongxiang Liu, and Weidong Jiang. GDROS: A geometry-guided dense registration framework for optical-SAR images under large geometric transformations. *IEEE Transactions on Geoscience and Remote Sensing*, 63:5650315, 2025. doi: 10.1109/TGRS.2025.3627132.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020*, pp. 402–419, 2020. doi: 10.1007/978-3-030-58536-5_24.

Önder Tuzcuoğlu, Aybora Köksal, Buğra Sofu, Sinan Kalkan, and A. Aydın Alatan. XoFTR: Cross-modal feature matching transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4275–4286, 2024. doi: 10.1109/CVPRW63382.2024.00431.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *Asian Conference on Computer Vision*, pp. 256–273. Springer, 2022.

Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21666–21675, 2024. doi: 10.1109/CVPR52733.2024.02047.

Yihan Wang and Jia Deng. Waft: Warping-alone field transforms for optical flow. In *International Conference on Learning Representations*, volume 2026, pp. 157389–157402, 2026.

Peihao Wu, Yongxiang Yao, Yi Wan, Wenfei Zhang, Ruipeng Zhao, Jiayuan Li, and Yongjun Zhang. SOMA-1M: A large-scale SAR-optical multi-resolution alignment dataset for multi-task remote sensing. *arXiv preprint arXiv:2602.05480*, 2026a. URL https://arxiv.org/abs/2602.05480.

Peihao Wu, Yongxiang Yao, Wenfei Zhang, Dong Wei, Yi Wan, Haoyu Guo, Yansheng Li, and Yongjun Zhang. MapGlue: Multimodal remote sensing image matching. *ISPRS Journal of Photogrammetry and Remote Sensing*, 240:166–183, 2026b. doi: 10.1016/j.isprsjprs.2026.07.007.

Yuming Xiang, Rongshu Tao, Feng Wang, Hongjian You, and Bing Han. Automatic registration of optical and SAR images via improved phase congruency model. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:5847–5861, 2020. doi: 10.1109/JSTARS.2020.3026162.

Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. FusionDN: A unified densely connected network for image fusion. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. GMFlow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8121–8130, 2022. doi: 10.1109/CVPR52688.2022.00795.

Ruoyu Yang, Yinhe Liu, Heng Yan, Yiheng Zhou, Yihan Fu, Han Luo, and Yanfei Zhong. MaRS: A multi-modality very-high-resolution remote sensing foundation model with cross-granularity meta-modality learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14):11685–11693, 2026. doi: 10.1609/aaai.v40i14.38153.

Yibin Ye, Xichao Teng, Hongrui Yang, Shuo Chen, Yuli Sun, Yijie Bian, Tao Tan, Zhang Li, and Qifeng Yu. 3MOS: A multi-source, multi-resolution, and multi-scene optical-SAR dataset with insights for multi-modal image matching. *Visual Intelligence*, 3:19, 2025. doi: 10.1007/s44267-025-00091-0.

Wenfei Zhang, Ruipeng Zhao, Yongxiang Yao, Yi Wan, Peihao Wu, Jiayuan Li, Yansheng Li, and Yongjun Zhang. Multi-resolution SAR and optical remote sensing image registration methods: A review, datasets, and future perspectives. *arXiv preprint arXiv:2502.01002*, 2025. URL https://arxiv.org/abs/2502.01002.

Yongjun Zhang, Yongxiang Yao, Yi Wan, Weiyu Liu, Wupeng Yang, Zhi Zheng, and Rang Xiao. Histogram of the orientation of the weighted phase descriptor for multi-modal remote sensing image matching. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:1–15, 2023. doi: 10.1016/j.isprsjprs.2022.12.018.

Chongyue Zheng, Shanshan Li, Chengyou Wang, and Bing Zhang. MSG: Robust multimodal remote sensing image matching using side window Gaussian space. *IEEE Transactions on Geoscience and Remote Sensing*, 63:4706223, 2025. doi: 10.1109/TGRS.2025.3591504.

A Additional Dataset Details

Derived layers. WH1-MS-RGB and S2-MS-RGB are RGB composites of the corresponding multispectral products; WH1-MS-NIR and S2-MS-NIR contain their near-infrared bands. WH1-HS-RGB/NIR are generated from Wuhan-1 hyperspectral data. Depth is generated from 0.5 m Ref-RGB imagery using Depth Anything 3 (Lin et al., 2026).

Resolution-dependent retention. At 0.5 m and 2 m, we retain only spatial groups in which all data layers available at that GSD are valid. At 5 m and 10 m, we retain each target–Ref-RGB pair whenever the target layer and Ref-RGB have valid overlapping coverage, without requiring every other layer to be present. This preserves valid pairs while reducing sample loss caused by missing layers at coarser resolutions.

Table 7: Target-layer composition and sample counts of the WH-M3R-Match subsets.

Target layer	GSD (m)	Match-30K	Match-1K
Map	0.5	4,500	100
Depth	0.5	4,500	100
DEM	10	1,183	100
GF3-SAR	5	5,000	100
S1-SAR	10	1,433	100
WH1-MS-RGB	2	4,000	100
WH1-MS-NIR	2	4,000	100
WH1-HS-RGB	10	1,468	100
WH1-HS-NIR	10	1,431	100
WH1-PAN	0.5	2,485	100
Total	–	30,000	1,000

Note: Each pair uses a Ref-RGB image resampled to the same GSD as the target layer as its reference.

Table 7 lists the composition of WH-M3R-Match-30K and WH-M3R-Match-1K. Both subsets pair Ref-RGB with ten target data layers spanning maps, depth, elevation, SAR, panchromatic, multispectral, near-infrared, and hyperspectral imagery. Match-30K contains 30,000 pairs sampled from the training partition, with counts for each pairing determined by data availability; a random 9:1 split at the image-pair level yields 27,000 training pairs and 3,000 validation pairs. Match-1K contains 100 pairs per target layer from the independent test partition, for a total of 1,000 pairs. Its geographic extent is separated from that of Match-30K by latitude and longitude bounds, with no spatial overlap. These subsets cover representative Ref-RGB–target-layer pairings rather than every possible data-layer combination in WH-M3R.

A.1 Scene Composition

To obtain an approximate description of scene composition in the training partition, we use Qwen3-VL-8B-Instruct (Bai et al., 2025) to assign 109,814 training Ref-RGB images to eight scene categories. Figure 4(a) shows that Village and Farmland account for the largest shares, at 24.7% and

22.0%, respectively. The remaining predicted categories are Water, Urban, Industrial, Construction Land, Vegetation, and Transportation. Figure 4(b) shows differences in predicted scene composition across spatial-resolution groups, with larger shares of Farmland and Water in the coarser-resolution groups. These statistics provide only a rough description of scene composition in the training partition; they are neither precise scene-classification ground truth nor estimates of land-area proportions. Differences between resolution groups may reflect both sample composition and biases in automatic classification.

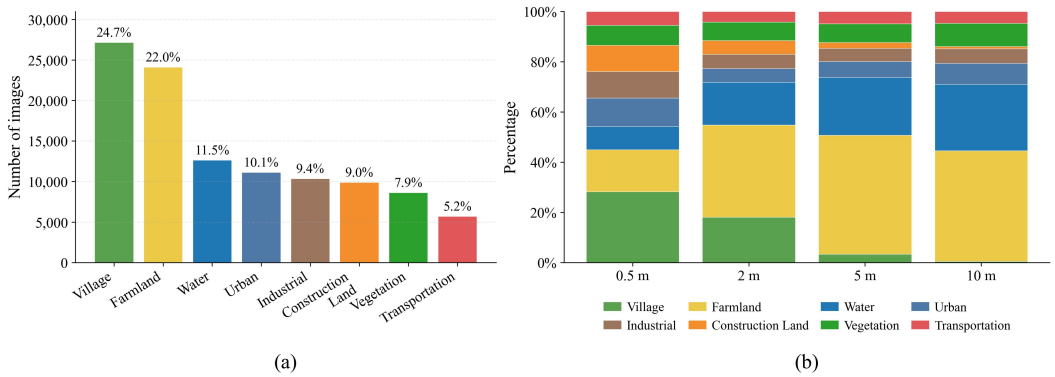


Figure 4: Automatically estimated scene composition of Ref-RGB imagery in the training partition. (a) Overall counts and proportions of predicted categories. (b) Predicted category proportions within each spatial-resolution group. The statistics use automatic image-level labels and provide only a rough description of scene composition.

A.2 Accuracy Assessment

We assess the initial alignment of Match-1K using a stratified random sample of ten original, untransformed pairs from each Ref-RGB–target-layer combination, for a total of 100 pairs. We manually mark ten spatially distributed correspondences per pair; because DEM contains fewer reliably localizable structures, we mark five per Ref-RGB–DEM pair, yielding 950 correspondences in total. The mean error for pair k is $\bar{e}_k = N_k^{-1} \sum_{n=1}^{N_k} \|\mathbf{p}_{k,n}^1 - \mathbf{p}_{k,n}^0\|_2$. PCK@$\delta$ denotes the percentage of annotated correspondences with a coordinate discrepancy no greater than δ pixels. The overall mean is computed over the 100 pair-level errors, giving every pair equal weight despite the smaller number of DEM points.

Table 8: Manual checkpoint assessment of the initial alignment in the stratified random sample. PCK is computed over the annotated points of each target-layer pairing category; the overall values average the ten categories equally.

Target layer	Pairs	Points/pair	Mean error (px)	PCK@3 (%)	PCK@5 (%)
Map	10	10	2.50	70.0	96.0
Depth	10	10	1.38	93.0	99.0
DEM	10	5	5.00	34.0	52.0
GF3-SAR	10	10	1.36	98.0	100.0
S1-SAR	10	10	2.16	78.0	99.0
WH1-MS-RGB	10	10	1.52	93.0	100.0
WH1-MS-NIR	10	10	1.62	94.0	100.0
WH1-HS-RGB	10	10	1.40	96.0	100.0
WH1-HS-NIR	10	10	1.85	89.0	100.0
WH1-PAN	10	10	2.82	63.0	88.0
Overall	100	950 total	2.16	80.8	93.4

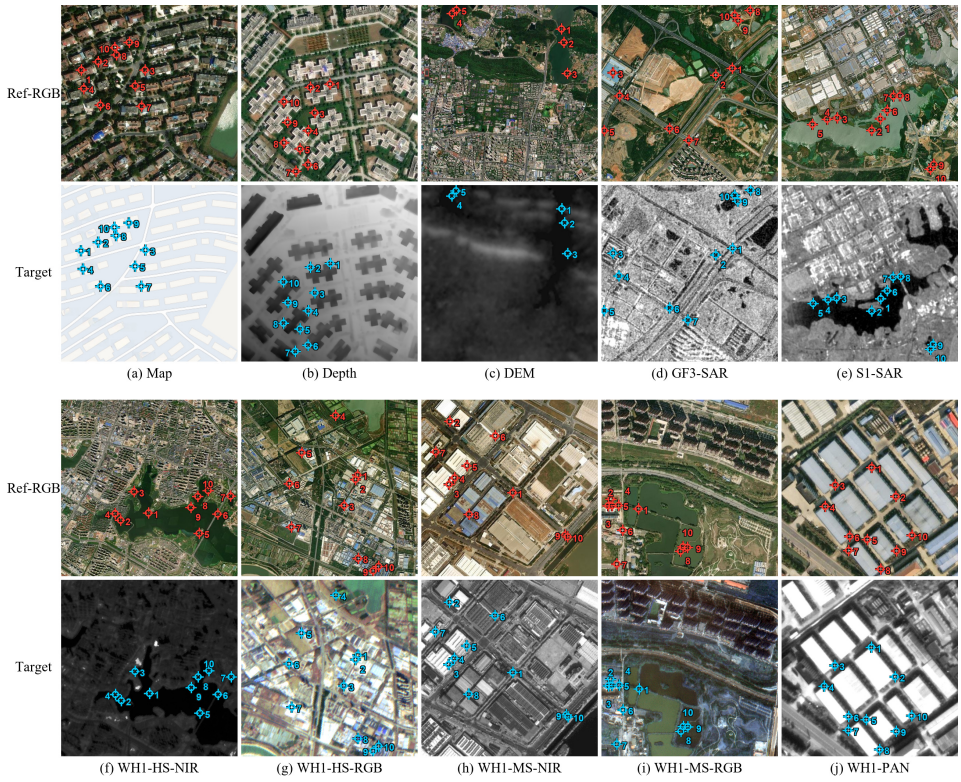


Figure 5: Representative manual correspondence annotations across the ten target data layers. Red and green markers with the same index denote manually selected corresponding locations.

The median pair-level mean error is 1.76 pixels. Among the sampled pairs, 64%, 84%, and 94% have mean errors no greater than 2, 3, and 5 pixels, respectively. Accuracy varies across data layers: GF3-SAR, Depth, WH1-HS-RGB, and the two WH1-MS layers have mean errors of 1.36–1.62 pixels, whereas Map and WH1-PAN reach 2.50 and 2.82 pixels. DEM has the largest mean error, 5.00 pixels, and also presents the greatest difficulty for manual point localization.

Manual correspondences do not constitute absolute geometric ground truth. SAR scattering centers and optical object boundaries need not coincide, map symbols do not define unique physical boundary locations, and smoothly varying DEM regions often lack precisely localizable structures. The measured differences therefore combine residual image misalignment with manual localization uncertainty, which cannot be separated by this assessment. For highly dissimilar modalities, uncertainty in manual single-point localization is typically more pronounced. Automatic methods estimate global geometry by aggregating image-wide structural evidence, which helps reduce the effect of locally ambiguous positions and provides a complement to manual assessment. We therefore use manual annotations for a sampled assessment of relative alignment quality and combine them with model estimates as auxiliary consistency evidence, without treating either as a calibration of the absolute geometric accuracy of WH-M3R.

Model-assisted dense consistency. We further evaluate all 1,000 original Match-1K pairs without synthetic transformations. GeoCFR and WAFT use their Low models trained on Match-30K, whereas MINIMA$_{\mathrm{RoMa}}$ uses released weights trained without WH-M3R (Ren et al., 2025). GeoCFR predicts a global affine transformation directly. For WAFT and MINIMA$_{\mathrm{RoMa}}$, we fit an affine transformation with USAC-MAGSAC at a 0.5-pixel reprojection threshold, using dense flow and cached correspondences, respectively; MINIMA$_{\mathrm{RoMa}}$ retains the highest-certainty correspondence in each 8-pixel source-grid cell. Using the zero-displacement field induced by identity mapping in the coordinate system of each original pair as a consistency reference, we compute the mean displacement magnitude over the full 512×512 grid. CMR at threshold δ is the proportion of all pairs whose mean residual does not exceed δ; unsuccessful estimates count as failures.

Table 9: Model-assisted dense residual consistency on the 1,000 original Match-1K pairs. Residuals are in pixels. Mean and median values for $\text{MINIMA}_{\text{RoMa}}$ use its successful estimates; CMR uses all pairs.

Method	WH-M3R	Succ. (%)	Mean	Med.	1 px	2 px	3 px	5 px
$\text{MINIMA}_{\text{RoMa}}$	No	72.4	1.59	0.78	45.8	60.4	64.7	68.5
WAFT	Yes	100.0	1.84	0.78	62.7	83.2	89.9	94.1
GeoCFR	Yes	100.0	1.43	0.68	65.6	85.7	91.7	95.5

GeoCFR and WAFT give similar residual estimates for the original pairs. Their residual EPEs have a Spearman correlation of 0.828 and a median absolute difference of 0.239 pixels. Both methods simultaneously satisfy the 3-pixel criterion on 88.0% of the pairs. Their decisions differ on only 5.6% of the pairs at this threshold.

On the 100 manually annotated pairs, the correlations between manual errors and the residuals estimated by GeoCFR, WAFT, and $\text{MINIMA}_{\text{RoMa}}$ are 0.543, 0.560, and 0.570, respectively. These values indicate moderate positive associations. The $\text{MINIMA}_{\text{RoMa}}$ correlation is computed over the 82 pairs for which it produces valid estimates. Across the full test set, $\text{MINIMA}_{\text{RoMa}}$ succeeds on 72.4% of the pairs. Most failures occur on Map and DEM, showing that its applicability varies across data layers. The model-based results provide a supplementary consistency check over the complete Match-1K set.

Together, these checks support the use of WH-M3R for training and evaluating global affine registration in this work. The pre-aligned pairs provide a reasonably stable coordinate basis for evaluation under controlled synthetic geometric transformations. Alignment quality nevertheless varies across data layers. Map, DEM, and some PAN pairs remain more challenging under strict pixel thresholds.

A.3 Keypoint Pseudo-Labels

The detector is supervised with automatically generated binary keypoint masks. Following Homographic Adaptation (DeTone et al., 2018), a pretrained SuperPoint detector extracts dense scores before non-maximum suppression (NMS) from randomly homography-warped views of each preprocessed grayscale image. These heatmaps are bilinearly warped back and averaged per pixel over valid observations:

$$S_k = D_{\text{SP}}(\mathcal{W}_{H_k} I), \qquad \overline{S}(\mathbf{x}) = \frac{\sum_{k=1}^{N_T} M_k(\mathbf{x})\,[\mathcal{W}_{H_k^{-1}} S_k](\mathbf{x})}{\max\Big(1, \sum_{k=1}^{N_T} M_k(\mathbf{x})\Big)}. \tag{A.1}$$

Here, D_{SP} returns the pre-NMS scores, $\mathcal{W}_H$ denotes warping by homography H, and N_T is the number of sampled transforms. The mask M_k is obtained by inverse-warping an all-ones image and thresholding it at 0.9. After averaging, the outer five-pixel border of $\overline{S}$ is set to zero.

After aggregation, keypoints are selected by gradient-guided multi-radius NMS adapted from Saliency-Enhanced SuperPoint (SES) in MapGlue (Wu et al., 2026b). Retained locations are set to one and all other pixels to zero, independently for each image. During training, these binary masks are transformed with the images and validity masks. Correspondence and displacement targets are derived separately from the known relative geometry in Appendix D.

B Implementation and Spectral Stability

Encoder and fine-matching pyramid. Each encoder block uses convolution, GroupNorm, and ReLU. Features $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4$ have spatial sizes $H \times W$, $H/2 \times W/2$, $H/4 \times W/4$, and $H/8 \times W/8$, respectively. Fine-matching features are fused top-down:

$$\begin{aligned} \mathbf{F}_8 &= S_8(L_8(\mathbf{c}_4)), \\ \mathbf{F}_4 &= S_4(L_4(\mathbf{c}_3) + \text{Up}_{8\to4}(\mathbf{F}_8)), \\ \mathbf{F}_2 &= S_2(L_2(\mathbf{c}_2) + \text{Up}_{4\to2}(\mathbf{F}_4)). \end{aligned} \tag{B.1}$$

Here, L_s is a lateral 1×1 projection, $\text{Up}_{a\to b}$ combines channel projection with bilinear upsampling, and S_s is a 3×3 smoothing convolution. Features $\mathbf{F}_8, \mathbf{F}_4, \mathbf{F}_2$ have 256, 128, and 64 channels, respectively.

Multi-scale keypoint detection. The detector projects each of the first three encoder levels to 32 channels and upsamples the latter two to $H \times W$:

$$\begin{aligned} \mathbf{H}_f &= \mathrm{Concat}[\Pi_1(\mathbf{c}_1), \mathrm{Up}(\Pi_2(\mathbf{c}_2)), \mathrm{Up}(\Pi_3(\mathbf{c}_3))]\,, \\ p &= \mathrm{Sigmoid}(\mathrm{Conv}_{3\times 3}(\mathbf{H}_f))\,. \end{aligned} \tag{B.2}$$

Each Π_ℓ is a 1×1 projection; $\mathbf{H}_f$ has 96 channels, and p is a single-channel score map. Non-maximum suppression, border removal, thresholding, and top-K selection produce keypoints. Both views share this detector.

MRRS channel composition. Using the main text's fixed-level, fixed-location notation, define the ring mean $\bar{a}_r = \frac{1}{8}\sum_{k=0}^{7} a_{r,k}$ and bispectral vector $\mathbf{b}_r = [B_r^{1,1}, B_r^{1,2}]^\mathsf{T}$. Each ring supplies nine channels, and the two rings are concatenated:

$$\begin{aligned} \mathbf{g}_r &= \mathrm{Concat}\big[\bar{a}_r, \log(1+E_r), \{P_{r,m}\}_{m=1}^{3}, \mathrm{Re}\,\mathbf{b}_r, \mathrm{Im}\,\mathbf{b}_r\big]\,, \\ \mathbf{m}^{\mathrm{MRRS}} &= \mathrm{Concat}[\mathbf{g}_1, \mathbf{g}_2] \in \mathbb{R}^{18}. \end{aligned} \tag{B.3}$$

The five nonredundant Fourier coefficients $F_{r,m}$ comprise the DC term $m = 0$ and four nonzero-frequency terms. Energy E_r sums modes $m = 1, \ldots, 4$, while only modes $m = 1, 2, 3$ supply separate power channels. Mode $m = 4$ contributes to energy but does not add a fourth power channel.

Descriptor fusion and sampling. Restoring level index $\ell \in \{2, 3, 4\}$, projected MRRS features are concatenated with MRM at each level, followed by cross-level fusion:

$$\begin{aligned} \mathbf{h}_\ell &= \Phi_\ell\big(\mathrm{Concat}[\mathbf{m}_\ell^{\mathrm{MRM}}, \Psi_\ell(\mathbf{m}_\ell^{\mathrm{MRRS}})]\big)\,, \\ \mathbf{D}_{\mathrm{dense}} &= \mathrm{Norm}_2(\Phi_d(\mathrm{Concat}[\mathbf{h}_2, \mathbf{h}_3, \mathbf{h}_4]))\,. \end{aligned} \tag{B.4}$$

Here, Ψ_ℓ projects MRRS channels, and Φ_ℓ fuses features within a level, producing 64, 96, and 96 channels. Cross-level mapping Φ_d uses a 3×3 convolution and a linear 1×1 convolution; Norm_2 performs L_2 normalization along the channel dimension. A keypoint $\mathbf{p}_i$ is mapped to coordinate $\mathbf{p}_i^{(8)}$ on the stride-8 descriptor grid and sampled as

$$\mathbf{d}_i = \mathrm{Norm}_2\Big(\mathrm{Bilinear}(\mathbf{D}_{\mathrm{dense}}, \mathbf{p}_i^{(8)})\Big) \in \mathbb{R}^{256}. \tag{B.5}$$

The operator $\mathrm{Bilinear}$ denotes bilinear sampling. The detector fuses $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3$, whereas the descriptor fuses $\mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4$.

Coordinate normalization and positional encoding. Image $v \in \{0, 1\}$ has width W_v and height H_v. Keypoint $\mathbf{p}_i^v$ is centered and normalized using the same scale for both axes:

$$\overline{\mathbf{p}}_i^v = \frac{\mathbf{p}_i^v - \frac{1}{2}(W_v, H_v)^\mathsf{T}}{\frac{1}{2}\max(W_v, H_v)}. \tag{B.6}$$

Writing $\overline{\mathbf{p}}_i^v = (\bar{x}_i^v, \bar{y}_i^v)^\mathsf{T}$, the radius and angle are

$$\rho_i^v = \sqrt{(\bar{x}_i^v)^2 + (\bar{y}_i^v)^2 + \epsilon_\rho}, \qquad \phi_i^v = \mathrm{atan2}(\bar{y}_i^v, \bar{x}_i^v). \tag{B.7}$$

Here, $\epsilon_\rho > 0$ is a numerical stabilizer. Positional encoding and the initial matching features are

$$\begin{aligned} \mathbf{e}_i^v &= f_{\mathrm{pos}}([\bar{x}_i^v, \bar{y}_i^v, \rho_i^v, \sin\phi_i^v, \cos\phi_i^v])\,, \\ \mathbf{h}_i^{v,(0)} &= W_{\mathrm{in}}\mathbf{d}_i^v + \mathbf{e}_i^v. \end{aligned} \tag{B.8}$$

The positional MLP f_{pos} has two linear layers with GELU. Input mapping W_{in} aligns descriptor and matching dimensions and may be the identity when the descriptor and matching dimensions are equal. Self-attention exchanges information within each image, while bidirectional cross-attention exchanges information between the two images. Output projection and L_2 normalization yield $\widetilde{\mathbf{d}}_i^v$.

Token aggregation. As in the main text, we omit scale indices; each scale uses its corresponding tokens and prediction head. For query, key, and value sequences Q, K, V, the heads of standard multi-head attention are computed and combined as

$$\begin{aligned} \mathrm{head}_h(Q, K, V) &= \mathrm{Softmax}\left(\frac{(QW_Q^h)(KW_K^h)^\mathsf{T}}{\sqrt{d_h}}\right)VW_V^h, \\ \mathrm{MHA}(Q, K, V) &= \mathrm{Concat}[\mathrm{head}_1, \ldots, \mathrm{head}_{N_h}]W_O. \end{aligned} \tag{B.9}$$

Here, N_h is the number of heads, d_h is the head dimension, and the W's are projection matrices. Evidence sequence $\mathbf{E}$ is obtained by encoding and pooling $\mathbf{e}_i$ from the main text. Cross-attention uses the tokens as queries, $Q = \mathbf{T}$, and the evidence sequence as keys and values, $K = V = \mathbf{E}$. The resulting tokens interact to predict a global affine residual:

$$\begin{aligned} \mathbf{T}^c = \mathbf{T} + \mathrm{MHA}(\mathbf{T}, \mathbf{E}, \mathbf{E}), \quad \widetilde{\mathbf{T}} = \mathrm{Transformer}(\mathbf{T}^c), \\ \boldsymbol{\xi} = f^{\mathrm{aff}}\Big(\mathrm{Mean}(\widetilde{\mathbf{T}})\Big). \end{aligned} \tag{B.10}$$

Transformer self-attention uses the token sequence as queries, keys, and values, together with an FFN, normalization, and residual connections. The operator Mean averages over tokens, and f^{aff} maps the result to six residual parameters that jointly refine one global affine transformation.

Training configuration. Table 10 lists the key parameters in GeoCFR's final training configuration. Scale-specific values are listed in P4, P2 order.

Table 10: Key parameters in GeoCFR's final training configuration.

Component	Setting	Value
Input	Patch size	512×512
Keypoints	Maximum number	2,048
Coarse matcher	Feature dimension / heads / layers	256 / 4 / 9
Refinement pyramid	Scales / metric dimensions	P4, P2 / 32, 24
Local posterior	Search radii (feature-grid pixels)	6, 2
Global tokens	Tokens per scale / token dimension	8, 8 / 96
Attention	Heads / aggregation layers per scale	4 / 2, 2
Loss weights	Correlation / offset / stage / final	1.0 / 0.25 / 0.5 / 0.5
Optimization	Optimizer / batch size / epochs	AdamW / 4 / 20
Optimization	Learning rate / weight decay	10^{-4} / 10^{-4}

Baseline geometric post-processing. For MSG, HIMO, LG, ELoFTR, and XoFTR, USAC-MAGSAC fits a source-to-target affine matrix to the sparse or semi-dense correspondences with a 0.5-pixel reprojection threshold. The fitted matrix is converted to a dense displacement field over the full source grid. For RAFT, GMFlow, and WAFT, source positions are sampled every 8 pixels and paired with target positions obtained by adding the predicted flow. Correspondences with non-finite predictions or target positions outside the image are discarded. USAC-MAGSAC then uses a 0.5-pixel reprojection threshold, 0.999 confidence, and at most 10,000 iterations. Fitting uses predictions only, without ground-truth filtering. If no valid affine estimate is obtained, the image pair is counted as a registration failure at every CMR threshold.

B.1 Rotation Stability Analysis

We again fix the feature level and spatial location. For a real-valued sequence of length eight, $F_{r,8-m} = F^*_{r,m}$ for $m = 1, 2, 3$, while $F_{r,0}$ and $F_{r,4}$ are real. Thus $m = 0, \ldots, 4$ contains all nonredundant frequency information. Consider the cyclic shift $a'_{r,k} = a_{r,(k-\Delta k) \bmod 8}$. Substituting $t = (k - \Delta k) \bmod 8$ yields

$$\begin{aligned} F'_{r,m} &= \frac{1}{\sqrt{8}} \sum_{k=0}^{7} a_{r,(k-\Delta k) \bmod 8} e^{-\mathrm{j}2\pi mk/8} \\ &= e^{-\mathrm{j}2\pi m\Delta k/8} \frac{1}{\sqrt{8}} \sum_{t=0}^{7} a_{r,t} e^{-\mathrm{j}2\pi mt/8} = e^{-\mathrm{j}2\pi m\Delta k/8} F_{r,m}. \end{aligned} \tag{B.11}$$

The zero-frequency term and all magnitudes are unchanged, so the mean, energy, and normalized power terms are invariant. For the bispectrum, the numerator's phase factor is

$$e^{-\mathrm{j}2\pi m\Delta k/8} e^{-\mathrm{j}2\pi n\Delta k/8} e^{+\mathrm{j}2\pi (m+n)\Delta k/8} = 1, \tag{B.12}$$

and the magnitude-based denominator is unchanged. Thus both real and imaginary bispectral channels are invariant to this discrete shift. This property concerns the sampled relation sequence, not exact rotation invariance of the entire network: arbitrary rotations alter interpolation and encoder responses. Regional statistics and spectral relations should therefore be understood as complementary structural cues, not a proof of invariance to arbitrary modality changes.

C Loss Definitions

We omit sample, view, and position indices, retaining the feature-scale index s and the ground-truth superscript gt; $q_s, \boldsymbol{\mu}_s$ correspond to $q_i, \boldsymbol{\mu}_i$ in the main text's single-scale description. Let $\langle\cdot\rangle_{\mathcal{S}}$ denote the mean over a valid set $\mathcal{S}$, with empty sets contributing zero. Detector and matchability losses are computed per view and averaged over both views. Local refinement losses are first averaged over valid positions within each image with sufficient supervision locations, then over eligible images and feature scales.

Detector supervision. Keypoint score maps are supervised with weighted binary cross-entropy:

$$\mathcal{L}_p = \mathrm{WBCE}(p, y, V). \tag{C.1}$$

Here, p, y, V are the sigmoid score map, binary keypoint pseudo-labels, and pixel validity mask. WBCE uses only valid pixels, weights positive labels by w_p, and normalizes by the number of valid pixels plus one.

Coarse assignments. For a candidate point pair $(\mathbf{p}, \mathbf{p}')$ from the two images, the known transform H defines the symmetric reprojection distance

$$d = \max\left(\|\mathbf{p}' - H\mathbf{p}\|_2, \|\mathbf{p} - H^{-1}\mathbf{p}'\|_2\right). \tag{C.2}$$

Valid pairs with $d < \delta_+$ and $d > \delta_-$ form the positive set $\mathcal{P}$ and negative set $\mathcal{N}$, respectively; pairs between the thresholds are ignored. For assignment probability π,

$$\begin{aligned} \mathcal{L}_{\mathrm{pos}} &= -\langle \log(\pi + \epsilon)\rangle_{\mathcal{P}}, \\ \mathcal{L}_{\mathrm{neg}} &= -\langle \log(1 - \pi + \epsilon)\rangle_{\mathcal{N}_h} - \eta\langle \log(1 - \pi + \epsilon)\rangle_{\mathcal{N}}. \end{aligned} \tag{C.3}$$

Here, $\mathcal{N}_h$ contains the highest-probability negatives, and η weights all negatives. Mutual nearest neighbors within δ_+ are labeled as matched; valid points whose nearest correspondence exceeds δ_- form the unmatched set $\mathcal{U}$; other points are ignored. The unmatched probability $p_\varnothing$ and matchability m are supervised by

$$\mathcal{L}_{\mathrm{dustbin}} = -\langle \log p_\varnothing\rangle_{\mathcal{U}}, \qquad \mathcal{L}_{\mathrm{match}} = \mathrm{BCE}(m, z). \tag{C.4}$$

Here, z is the binary match label, and BCE averages over non-ignored points only. The dustbin loss is averaged only over views containing unmatched points and is zero if neither view contains any. The total coarse loss is

$$\mathcal{L}_c = \lambda_{\mathrm{pos}}\mathcal{L}_{\mathrm{pos}} + \lambda_{\mathrm{neg}}\mathcal{L}_{\mathrm{neg}} + \lambda_d\mathcal{L}_{\mathrm{dustbin}} + \lambda_m\mathcal{L}_{\mathrm{match}}. \tag{C.5}$$

Intermediate attention layers are supervised using normalized layer weights.

Posterior and offset losses. We construct a discrete Gaussian target q_s^{gt}, centered at the ground-truth residual $\mathbf{o}_s^{\mathrm{gt}}$ with standard deviation σ_s, and normalize it over valid candidates. The local posterior q_s and its mean offset $\boldsymbol{\mu}_s$ receive distributional and localization supervision, respectively:

$$\begin{aligned} \mathcal{L}_{\mathrm{corr}} &= \sum_s a_s \left\langle D_{\mathrm{KL}}(q_s^{\mathrm{gt}} \| q_s)\right\rangle_{\Omega_s}, \\ \mathcal{L}_{\mathrm{off}} &= \sum_s a_s \left\langle \mathrm{SmoothL1}_\beta(\mathbf{D}_s\boldsymbol{\mu}_s, \mathbf{D}_s\mathbf{o}_s^{\mathrm{gt}})\right\rangle_{\Omega_s}. \end{aligned} \tag{C.6}$$

Residual offsets are measured from the current search center in feature-grid pixels; $\mathbf{D}_s$ converts them to full-resolution pixels. Smooth L1 uses transition threshold β and sums over both coordinates; a_s denotes normalized weights over feature scales with valid supervision. The valid position set Ω_s includes only locations whose ground-truth correspondences are valid, whose residual offsets lie within the search window, and whose valid candidates cover sufficient Gaussian probability mass. Images with too few such locations are excluded from supervision at that feature scale.

Stage and final geometric losses. The stage loss supervises affine corrections at P4 and P2, while the final loss supervises the output displacement field:

$$\mathcal{L}_{\mathrm{stage}} = \sum_s a_s \sqrt{d(\mathbf{A}_s, \widehat{\mathbf{A}}_s^{\mathrm{gt}})^2 + \epsilon_r^2}, \qquad \mathcal{L}_{\mathrm{final}} = \left\langle \sqrt{\|\mathbf{u} - \mathbf{u}^{\mathrm{gt}}\|_2^2 + \epsilon_r^2} \right\rangle_{\Omega}. \tag{C.7}$$

Here, $d(A, B)$ is the mean Euclidean distance between corresponding positions obtained by applying A and B to the same fixed grid points. The stage target $\widehat{\mathbf{A}}_s^{\mathrm{gt}}$ is constructed by linear interpolation between the incoming affine estimate and ground truth.

The final loss directly compares the predicted displacement $\mathbf{u}$ with the full ground truth $\mathbf{u}^{\mathrm{gt}}$. Here, Ω is the valid pixel set and $\epsilon_r > 0$ is a smoothing constant. Both losses use Charbonnier-type robust errors (Charbonnier et al., 1994) and are averaged over valid images. Ground-truth filtering defines training losses and evaluation masks only; it is not an inference input.

D Border-Free Geometric Sampling

Transform sampling. Starting from two original aligned images of equal size, we sample rotation, translation, and isotropic scaling in the 512×512 output coordinate system according to Table 3. The resulting similarity transform is represented by an affine matrix. A grid with 8-pixel spacing estimates overlap in both directions. The lower of the two estimates must reach 0.7/0.5/0.3 for Low/Medium/High, respectively. Samples must also meet the minimum valid grid-point count.

Crop adjustment. The reference crop initially covers the largest centered square within the valid area of the original image. We use binary search to find the smallest shared crop-reduction factor that keeps all four output corners, mapped back to each original image, inside its boundary with a one-pixel margin. Increasing this factor reduces both sampling regions without changing the relative transform. The search intervals are $[1, 1.6]$, $[1, 1.8]$, and $[1, 2.0]$ for Low, Medium, and High, respectively. Candidates that fail the overlap or boundary checks are resampled, with at most 50 attempts overall.

Images and ground truth. The resulting sampling maps produce 512×512 image pairs, and keypoint pseudo-labels undergo the same warps. The relative transform determines the ground-truth displacement field and validity mask. Border-free images need not overlap completely: evaluation uses only valid pixels whose correspondences remain inside the target image. Evaluation images, ground truth, and masks are generated once with fixed per-sample seeds and reused by all methods.

E Results by Data-Layer Pairing

Results by target data layer. Table 11 reports GeoCFR separately for the ten Ref-RGB–target-layer pairings in Match-1K. Each pairing contains 100 image pairs, and the overall row gives each target layer equal weight.

Table 11: GeoCFR results by target data layer on WH-M3R-Match-1K (CMR, %).

	Low			Medium			High		
Target layer	1 px	2 px	3 px	1 px	2 px	3 px	1 px	2 px	3 px
Map	7.0	42.0	66.0	5.0	29.0	51.0	2.0	19.0	39.0
Depth	100.0	100.0	100.0	100.0	100.0	100.0	99.0	99.0	99.0
DEM	2.0	18.0	33.0	3.0	13.0	23.0	1.0	5.0	13.0
GF3-SAR	23.0	65.0	84.0	13.0	52.0	73.0	14.0	46.0	59.0
S1-SAR	85.0	98.0	99.0	73.0	96.0	98.0	57.0	91.0	92.0
WH1-MS-RGB	56.0	87.0	92.0	45.0	78.0	87.0	38.0	63.0	82.0
WH1-MS-NIR	52.0	81.0	89.0	36.0	72.0	81.0	29.0	65.0	81.0
WH1-HS-RGB	76.0	84.0	89.0	64.0	78.0	86.0	64.0	81.0	88.0
WH1-HS-NIR	60.0	85.0	88.0	50.0	80.0	88.0	31.0	69.0	86.0
WH1-PAN	24.0	50.0	67.0	8.0	37.0	54.0	8.0	32.0	48.0
Overall	**48.5**	**71.0**	**80.7**	**40.2**	**63.5**	**74.1**	**34.3**	**57.0**	**68.7**

Performance varies substantially across target data layers. Depth achieves the highest CMR, consistent with its derivation from Ref-RGB and their closely aligned structures. S1-SAR and the WH1-HS layers also retain relatively high CMR under the High protocol. Map, DEM, GF3-SAR, and WH1-PAN are more challenging, with DEM producing the lowest results. These differences show that the aggregate results do not represent uniform performance across all data layers.

Table 12: GeoCFR results on eight other modality combinations (CMR, %).

	Low			Medium			High		
Data-layer pair	1 px	2 px	3 px	1 px	2 px	3 px	1 px	2 px	3 px
Map+Depth	3	22	54	3	18	45	0	14	34
Map+WH1-PAN	3	16	36	0	14	28	1	8	16
DEM+S1-SAR	1	9	19	0	5	14	0	1	7
DEM+WH1-HS-NIR	4	12	21	0	4	10	0	1	4
Depth+WH1-PAN	10	35	54	9	29	46	7	22	37
S1-SAR+WH1-HS-NIR	44	81	89	34	70	84	27	61	77
S1-SAR+WH1-HS-RGB	40	77	84	33	72	82	20	63	74
WH1-MS-NIR+WH1-MS-RGB	75	93	94	69	91	93	61	89	92

To evaluate other modality combinations, we construct eight pairings from spatially co-located images in WH-M3R-Match-1K, with 100 pairs each (Table 12). None of these combinations is included in training, where each target layer is paired only with Ref-RGB.

The same-sensor WH1-MS-NIR–WH1-MS-RGB pairing achieves the highest CMR: 61%, 89%, and 92% at the 1-, 2-, and 3-pixel thresholds under the High protocol. These results show transfer to a same-sensor band combination with appearance differences and large geometric transformations. S1-SAR paired with either WH1-HS layer performs next best; S1-SAR–WH1-HS-NIR reaches 27%, 61%, and 77%, respectively.

Depth–WH1-PAN performs moderately, while pairings involving Map or DEM perform poorly at strict thresholds. Under the High protocol, CMR at 3 pixels is only 7% for DEM–S1-SAR and 4% for DEM–WH1-HS-NIR. Possible contributing factors include cartographic representations and differences between elevation, backscatter, and reflectance.

F Additional Qualitative Results

Figure 6 complements the inlier match visualizations in the main text with checkerboard overlays of the registered images. In the examples shown, building outlines, roads, and water-body boundaries are generally continuous across checkerboard cell boundaries, including in the close-ups. This continuity is visible despite appearance differences arising from cartographic representations and SAR speckle, allowing alignment to be inspected across the image and in the enlarged regions.

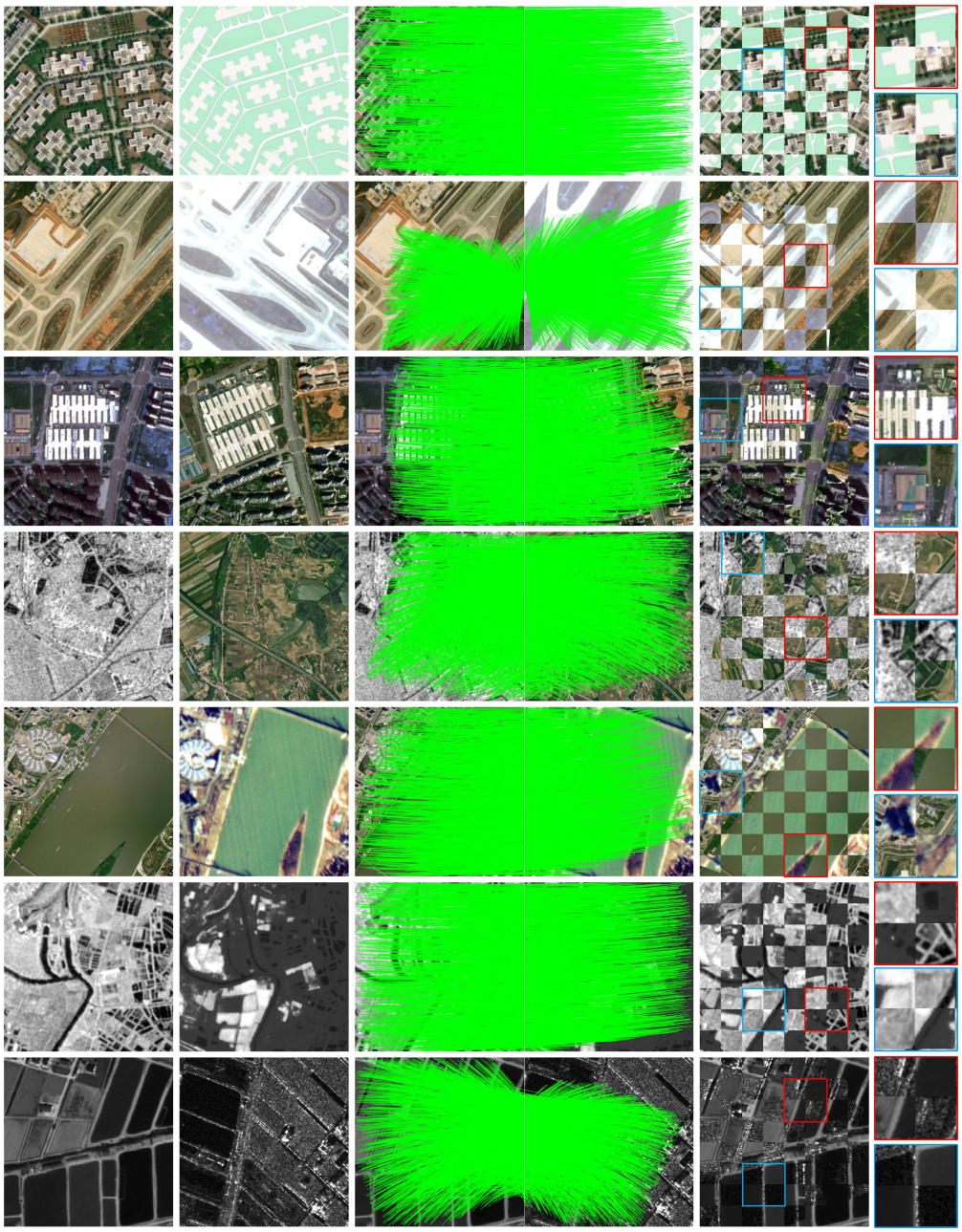


Figure 6: Additional qualitative results of GeoCFR. Each row shows, from left to right, the input image pair, an inlier match visualization, a checkerboard overlay of the registered images, and close-ups of the regions marked in red and blue. Correspondence sampling and inlier filtering follow the same criteria as in Figure 3.

G INFERENCE EFFICIENCY

Table 13 reports the average inference time of each method at a batch size of 1. GPU methods were evaluated on an NVIDIA GeForce RTX 4090 D, while MSG and HIMO used their native CPU/MATLAB implementations. GeoCFR takes an average of 107.8 ms per pair, longer than the other learning-based methods, but achieves better registration results at strict accuracy thresholds. Compared with GDROS, the baseline with the highest CMR@1 under the Low protocol, GeoCFR requires an additional 12.9 ms per pair while increasing CMR@1 from 43.1% to 48.5%. This comparison shows that, under the current implementation and evaluation settings, GeoCFR trades additional inference time for a higher registration success rate at a strict accuracy threshold.

Table 13: Comparison of inference time across methods.

Method	MSG	HIMO	LG	ELoFTR	XoFTR	RAFT	GMFlow	GDROS	WAFT	GeoCFR
Time (ms/pair)	6257.5	18698.5	25.8	25.8	52.3	98.7	12.6	94.9	79.2	107.8