P2G: FROM POINT-WISE MATCHING TO RELATIONAL GRAPHS FOR TRAINING-FREE CROSS-MODAL IMAGE REGISTRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-modal image registration is challenging because modality discrepancies and large geometric transformations weaken the reliability of point-wise matching evidence. Rather than further strengthening individual point descriptors, we observe that intra-image relations among corresponding points remain consistent across modalities, providing complementary evidence of compatibility. Based on this observation, we propose P2G (Points-to-Graph), a task-training-free framework that progressively transforms corresponding point matches into a weighted correspondence relational graph. P2G first estimates a coarse affine transformation and preserves multiple plausible matches under geometric guidance, and then resolves their ambiguity using complementary structural and frozen foundation-model features to reach one-to-one correspondences as weighted graph nodes. Graph edges are constructed from multi-layer features of a frozen foundation model and encode how consistently intra-image query feature relations are preserved across modalities. Node reliability and edge compatibility jointly guide precise correspondence-triplet sampling for robust affine estimation. Experiments across multiple cross-modal datasets and geometric difficulty settings demonstrate that P2G achieves strong registration accuracy and robustness to shifts in scene distribution and geometric transformation than dataset-specific alternatives, offering a new graph perspective for cross-modal feature representation. Our source codes will be released.

1 INTRODUCTION

Cross-modal image registration aims to establish geometric correspondences between images acquired by different sensing modalities, supporting downstream tasks such as multi-modal fusion, change detection, and remote sensing interpretation Liu et al. (2024a); Zhou et al. (2024); Zhang et al. (2025a). However, substantial nonlinear radiometric discrepancies and geometric variations make cross-modal registration inherently challenging Jena et al. (2026); Zhang et al. (2025b). Learning-based methods provide a powerful solution by learning modality-robust representations from data. Nevertheless, their performance remains sensitive to sufficient access to modality and geometric distributions, restricting their generalization across datasets and modalities, particularly when deployed without task-specific adaptation. This motivates an important question: can cross-modal registration remain accurate under distribution shifts without task-specific optimization?

Existing cross-modal image registration methods can be broadly categorized into handcrafted Li et al. (2019; 2022a); Ye et al. (2022b); Zhang et al. (2024a; 2025c); Peng et al. (2026) and learning-based Ye et al. (2022a); Sun et al. (2024); Zhang et al. (2024b); Liu et al. (2024b); Sun et al. (2025) ones. Handcrafted methods are training-free but often suffer from limited descriptor discriminability, whereas learning-based methods provide stronger feature representations at the cost of training dependency. Despite these differences, both paradigms ultimately assign confidence to individual correspondence hypotheses and estimate geometry from these point-wise matches. Even when contextual interactions are used during feature extraction or matching, the final correspondence evidence is often dominated by unary compatibility between individual localized point pairs. Such point-wise evidence can become ambiguous when modality discrepancies weaken appearance similarity or when large geometric transformations invalidate learned geometric priors.

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.

As an attempt to address above issue, we seek a richer representation for cross-modal image registration. Our key observation is that cross-modal invariance need not reside in individual point features. Even when point responses vary substantially across modalities, their intra-image relations formed with other points within each image remain comparatively stable. Such relational consistency provides complementary evidence beyond point-wise features and offers a more robust basis for cross-modal matching. Importantly, the formulation of feature relations does not require features from different modalities to be directly similar. Instead, it asks whether the similarity between two features within one modality is preserved between their corresponding features in the other modality.

Based on this observation, we propose **P2G**, a training-free framework that progressively lifts point-wise matches into a relational graph for robust cross-modal image registration. P2G first performs coarse registration to restrict the correspondence search space and retain multiple plausible candidates for each feature point. These candidates are then re-described and re-matched using complementary structural and semantic evidence, producing reliable one-to-one correspondences as graph nodes. Relations between nodes are further constructed as edges by their cross-modal stability. Finally, node reliability and relational compatibility jointly guide correspondence-triplet sampling, favoring reliable and mutually compatible correspondences for robust affine estimation.

- We introduce a **point-to-graph representation framework** for training-free cross-modal image registration. Instead of making registration decisions simply based on point descriptors, P2G progressively organizes ambiguous candidates into reliable nodes and further connects them through relational edges. Such representation jointly exploits point-level evidence and inter-point consistency, providing a more robust basis for cross-modal correspondence and geometric estimation.
- We develop a **reliability-aware node construction** scheme that re-describes and re-matches feature points using complementary structural and semantic evidence, both of which are extracted without task-specific training and are adaptively fused according to their point-wise reliability. This process resolves candidate ambiguity and produces reliable one-to-one correspondences as graph nodes, enabling robust point representation across modality and geometric variations.
- We propose a **relational edge construction** scheme that aggregates multi-layer semantic evidence to measure the cross-modal consistency of intra-image feature relations. The resulting edge weights complement node reliability to guide correspondence-triplet sampling toward mutually compatible correspondences for robust affine estimation.

2 Related Work

2.1 Handcrafted Cross-Modal Image Registration

Handcrafted methods establish cross-modal correspondences through modality-robust descriptors without task-specific training. Early methods mainly exploit phase and orientation information to reduce sensitivity to nonlinear radiometric differences Heinrich et al. (2012); Ye & Shen (2016); Ye et al. (2019); Xiang et al. (2020a). RIFT2 uses phase-congruency features and a maximum index map for rotation-invariant matching Li et al. (2023), while HOWP further improves phase-based robustness through weighted phase orientation and log-polar description Zhang et al. (2023). Subsequent methods incorporate richer structural cues to strengthen local representation. WSSF integrates edge, phase, and steerable-filter responses to enhance structural saliency Wan et al. (2023), whereas MSG preserves multiscale edge structures through scale-space and gradient-based description Zheng et al. (2025). More recently, LPSD Liu et al. (2025) combines orientation-based structural representation with log-polar sampling to improve robustness to scale and rotation variations. These methods progressively improve the robustness of individual point descriptors, but correspondence determination still relies primarily on matching isolated keypoints. As a result, point-wise ambiguity remains difficult to resolve under large modality gaps.

2.2 Learning-Based Cross-Modal Image Registration

Learning-based cross-modal registration methods learn discriminative feature representations and correspondence models from data. Inspired by recent advancements in natural image matching DeTone et al. (2018); Sarlin et al. (2020); Sun et al. (2021), leading baselines adopt Transformer-based feature interaction and adapt learned representations to recognize modal differences. XoFTR

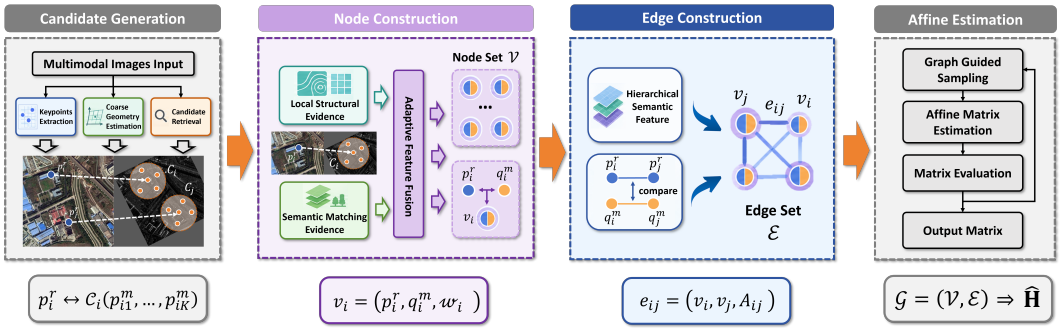


Figure 1: **Overview of P2G.** We first estimates coarse geometry and retains multiple plausible matches for each keypoint. In node construction, each node $v_i = (\mathbf{p}_i^r, \mathbf{q}_i^m, w_i)$ represents a selected correspondence, and its weight w_i measures the reliability of that match. Then, we connects pairs of nodes for relational edge description. Each edge $e_{ij} = (v_i, v_j, A_{ij})$ has a weight A_{ij} that measures compatibility between nodes.

builds upon LoFTR and adapts its matching framework to visible-thermal images through cross-modal training and coarse-to-fine refinement Tuzcuoğlu et al. (2024). As for the challenging optical-SAR registration under large geometric transformations, GDROS combines cross-modal feature interaction with geometric constraints on dense flow correspondences Sun et al. (2025). RoMaV2 advances general-purpose dense matching by incorporating foundation-model features, diverse training data, and a dedicated refinement stage Edstedt et al. (2025). More recently, CRFT learns feature flow for cross-modal registration and refines correspondences recurrently in a coarse-to-fine framework Liu et al. (2026). These methods demonstrate the value of learned feature interaction and correspondence refinement, while their performance remains tied to the representations and training distributions on which they rely.

3 METHOD

We propose P2G, a training-free method that progressively organizes cross-modal point matches into a correspondence graph. Given a pair of cross-modal images, P2G first retains multiple candidate matches for each reference keypoint instead of making an early one-to-one decision (Sec. 3.1). It then combines local structural and semantic evidence to select one-to-one correspondences, acting as weighted correspondence graph nodes (Sec. 3.2). Each pair of correspondence nodes is connected by an edge measuring cross-modal consistency of intra-modal feature relations (Sec. 3.3). Finally, the resulting correspondence graph guides robust affine estimation (Sec. 3.4). The overall framework is illustrated in Fig. 1.

3.1 AMBIGUITY-PRESERVING CANDIDATE GENERATION

Rotation- and scale-invariant structural descriptors remain stable under large geometric transformations, while such invariance may impact their ability to distinguish keypoints in locally similar regions. To mitigate this dilemma and avoid premature decision, P2G first estimates a coarse affine transformation and then uses the estimated geometry to construct a high-recall Top-K candidate set.

Given a reference image I^r and a query image I^m, we use GRID-FAST Liu et al. (2025) to extract spatially distributed keypoints and LPSD Liu et al. (2025) to compute their local descriptors. The two keypoint sets are denoted by $\mathcal{P}^r = \{\mathbf{p}_i^r\}_{i=1}^{N_r}$ and $\mathcal{P}^m = \{\mathbf{p}_j^m\}_{j=1}^{N_m}$, where $\mathbf{p} = [x, y]^\top$ denotes its 2D image coordinate. The descriptors of $\mathbf{p}_i^r$ and $\mathbf{p}_j^m$ are denoted as $\mathbf{d}_i^r$ and $\mathbf{d}_j^m$ respectively. Fast sample consensus (FSC) from Wu et al. (2014) is further applied to estimate a coarse affine transformation from the initial LPSD matches:

$$\mathbf{H}_0 = \begin{bmatrix} \mathbf{A}_0 & \mathbf{t}_0 \\ \mathbf{0}^\top & 1 \end{bmatrix}, \tag{1}$$

where $\mathbf{A}_0 \in \mathbb{R}^{2\times 2}$ is the rotation-scale mixed matrix and $\mathbf{t}_0 \in \mathbb{R}^2$ is the translation vector. $\mathbf{H}_0$ maps reference image coordinates to their corresponding query image coordinates via $\mathbf{p}_j^m = \mathbf{H}_0 \mathbf{p}_i^r$. The

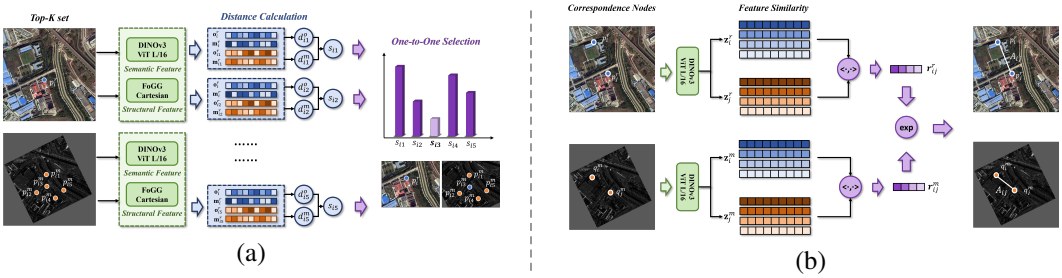


Figure 2: Node and edge construction in P2G. (a) Reliability-aware node construction through structural and semantic evidence, (b) Relational edge construction through cross-modal comparison of intra-image feature similarities.

initial matches provide a coarse geometric estimate but are not sufficiently precise for direct one-to-one matching. Then, we apply the inverse transformation to the query image using matrix $\mathbf{H}_0$ and recompute LPSD descriptors of the keypoints in the query image. This transformation-guided descriptor recomputation is expressed as $\widetilde{\mathbf{d}}_j^m = \mathcal{T}(\mathbf{d}_j^m; \mathbf{A}_0)$. For each reference keypoint $\mathbf{p}_i^r$, we rank the query image keypoints according to the ℓ_2 distance between $\mathbf{d}_i^r$ and $\widetilde{\mathbf{d}}_j^m$.

Rather than retaining only the nearest neighbor, we construct the candidate keypoint set $\mathbf{p}_i^r \longleftrightarrow \mathcal{C}_i = \{\mathbf{p}_{i1}^m, \mathbf{p}_{i2}^m, \ldots, \mathbf{p}_{iK}^m\}$, where $\mathcal{C}_i$ contains the K highest-ranked query image keypoints for $\mathbf{p}_i^r$. Each pair $(\mathbf{p}_i^r, \mathbf{p}_{ik}^m)$ is referred to as a candidate match. The Top-K set preserves plausible alternatives that may not be ranked first by the LPSD distance alone.

3.2 Reliability-Aware Node Construction

The Top-K set preserves plausible matches but does not determine which candidate should be selected. As shown in Fig. 2(a), for each candidate match, P2G evaluates local structural and semantic evidence, and adaptively fuses above two evidence to select one-to-one correspondence. Each retained correspondence then acts a weighted graph node.

Local Structural Evidence. We construct a Cartesian structural descriptor from the FoGG orientation map to characterize fine local structures. Let $\mathbf{o}_i^r$ and $\mathbf{o}_{ik}^m$ denote the ℓ_2-normalized structural descriptors of reference keypoint $\mathbf{p}_i^r$ and its k-th candidate $\mathbf{p}_{ik}^m$, respectively. Their structural distance d_{ik}^o is defined in Appendix A.1. A smaller d_{ik}^o indicates greater local structural similarity.

Semantic Matching Evidence. Similar local structures can occur at different image locations, making structural evidence alone insufficient to distinguish ambiguous candidates. We therefore extract semantic features using a frozen DINOv3 ViT-L/16 model. Let $\mathbf{m}_i^r$ and $\mathbf{m}_{ik}^m$ denote the semantic descriptors of $\mathbf{p}_i^r$ and its k-th candidate, respectively. Their semantic distance d_{ik}^m is defined in Appendix A.1. A smaller d_{ik}^m indicates greater semantic similarity.

Adaptive Fusion and One-to-One Selection. To select one candidate for each reference keypoint, we adjust the contribution of semantic evidence according to its discriminability relative to structural evidence. For $x \in \{o, m\}$, let d_{i1}^x and d_{i2}^x denote the smallest and second-smallest distances among the K candidates. The normalized distance margin and relative semantic discriminability are:

$$\rho_i = \frac{\gamma_i^m}{\gamma_i^m + \gamma_i^o + \epsilon}, \qquad \text{where} \qquad \gamma_i^x = \frac{d_{i2}^x - d_{i1}^x}{\max(d_{i2}^x, \epsilon)}. \tag{2}$$

Here, $\epsilon > 0$ prevents numerical instability. A larger ρ_i indicates that semantic distance distinguishes the candidates more clearly relative to structural distance. We then define the semantic fusion weight and fused matching cost as

$$\lambda_i = \lambda_{\min} + (\lambda_{\max} - \lambda_{\min})\rho_i, \qquad s_{ik} = (1 - \lambda_i)d_{ik}^o + \lambda_i d_{ik}^m. \tag{3}$$

where $\lambda_{\min} = 0.05$ and $\lambda_{\max} = 0.30$. A smaller s_{ik} indicates a better candidate match. For each reference keypoint, we select the candidate with the smallest fused matching cost. If multiple reference keypoints select the same query image keypoint, only the match with the lowest cost



(a) (b) (c)

Figure 3: Graph representation in P2G: (a) a weighted correspondence node, (b) a weighted relational edge, and (c) the obtained graph model.

(denoted as $\widetilde{s}_i$) is retained. Then, a robust, monotonically decreasing mapping based on the median absolute deviation converts $\widetilde{s}_i$ into a node weight $w_i \in (0, 1]$. The resulting node set is

$$\mathcal{V} = \{v_i = (\mathbf{p}_i^r, \mathbf{q}_i^m, w_i)\}_{i=1}^N . \tag{4}$$

Here, N is the number of retained one-to-one correspondences, $\mathbf{q}_i^m$ is the query image keypoint paired with $\mathbf{p}_i^r$, and w_i quantifies the reliability of this correspondence. These weighted correspondence nodes form the vertex set of the graph constructed next.

3.3 Relational Edge Construction

The nodes constructed above represent one-to-one correspondences, and their weights measure individual matching reliability. We connect each pair of nodes with an edge to capture relations between correspondences. The edge weight measures the cross-modal consistency of intra-image feature point similarities, complementing node reliability with pairwise relational evidence. The edge construction process is illustrated in Fig. 2(b).

Hierarchical Intra-Image Feature Similarities. For nodes v_i and v_j, we compare their within-image feature relationships in the reference and query images, respectively, using multi-scale DINOv3 features. Let $r_{ij}^{r,(\ell)}$ and $r_{ij}^{m,(\ell)}$ denote the feature similarity of reference image points and query image points at DINOv3 layer ℓ, respectively. Each similarity compares two locations within the same image, please refer to A.2 for more details.

Cross-Modal Pairwise Compatibility. At each layer, we propose a layer-wise compatibility indicator to measure whether the reference image feature similarity is preserved in the query image:

$$a_{ij}^{(\ell)} = \exp\left(-\left|r_{ij}^{r,(\ell)} - r_{ij}^{m,(\ell)}\right| / \tau_r\right) . \tag{5}$$

where τ_r is a hyperparameter. A larger $a_{ij}^{(\ell)}$ indicates greater cross-modal compatibility between correspondences i and j. We further calculate a fused compatibility indicator to aggregate per-layer compatibility:

$$A_{ij} = \bar{a}_{ij} \exp\left(-\delta_{ij}/\tau_s\right) , \qquad A_{ij} \in [0, 1], \tag{6}$$

where $\bar{a}_{ij}$ is the average value of $a_{ij}^{(\ell)}$. δ_{ij} is the standard deviation of $a_{ij}^{(\ell)}$, and τ_s is another hyperparameter. A large A_{ij} indicates that the intra-image feature similarity is consistently preserved across modalities and feature levels. In this way, the resulting edge set is:

$$\mathcal{E} = \{e_{ij} = (v_i, v_j, A_{ij}) \mid 1 \le i < j \le N\} \tag{7}$$

Finally, as illustrated in Fig. 3, we establish the correspondence graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each node links a matched point pair across images and carries weight w_i, while each edge links two nodes through their points within each image and carries weight A_{ij}. The weighted nodes and edges of $\mathcal{G}$ guide the following affine estimation.

3.4 Relation-Guided Affine Estimation

Given the correspondence graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, P2G uses the node and edge weights to guide minimal-set sampling and affine hypothesis evaluation.

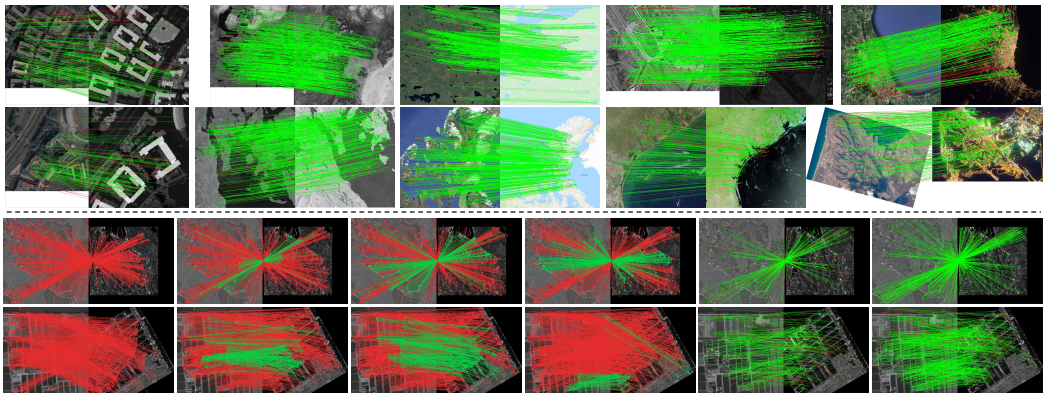


Figure 4: Visualization of registration correspondences. The first two rows show P2G results on five real-world modality pairs, while the last two rows compare different methods on the WHU-OPT-SAR dataset and OSDataset, respectively. For learning-based methods, we apply their estimated transformations to warp 5,000 fixed SIFT keypoints for the ease of visualization. Green and red lines indicate correspondences with registration errors below and above 2 pixels, respectively.

Relation-Guided Minimal-Set Sampling. An affine transformation is uniquely determined by three correspondences whose reference image points are non-collinear. We refer to this minimal correspondence set as a *correspondence triplet*. Uniform random sampling can require many iterations to obtain a proper triplet when outliers dominate the correspondence set. However, P2G can sample the correspondence triplet according to specific probabilities, guided by $\mathcal{G}$, which is more efficient than traditional methods. The sampling algorithm is introduced as follows:

The first node is the anchor of the correspondence triplet and the sampling probability is:

$$\Pr(i_1 = i) = w_i / \left(\sum\nolimits_{j\in\mathcal{V}} w_j\right), \qquad i \in \mathcal{V}. \tag{8}$$

Thus, a correspondence with a larger node weight is more likely to be selected as the anchor. After selecting anchor i_1, the available node set is $\mathcal{A}_1 = \mathcal{V} \setminus \{i_1\}$. The second node is sampled according to a log-probability function:

$$\Pr(i_2 = i \mid i_1) \propto -\log\left(e^{-w_i} + e^{-A_{i,i_1}}\right), \qquad i \in \mathcal{A}_1. \tag{9}$$

Based on this criterion in equation 9, we tend to select a correspondence where both w_i and A_{i,i_1} are high, rather than correspondences where only a single indicator is high. Similarly, after selecting i_1 and i_2, the remaining set is $\mathcal{A}_2 = \mathcal{V} \setminus \{i_1, i_2\}$. The third node is sampled according to a probability shown as:

$$\Pr(i_3 = i \mid i_1, i_2) \propto -\log\left(e^{-w_i} + e^{-A_{i,i_1}} + e^{-A_{i,i_2}}\right), \qquad i \in \mathcal{A}_2. \tag{10}$$

Because all three nodes are sampled probabilistically, P2G favors reliable and mutually compatible triplets while continuing to explore different minimal sets across iterations.

Affine Hypothesis Estimation and Evaluation. At iteration t, the sampled correspondence triplet determines an affine hypothesis $\mathbf{H}_t$. The re-projection error of correspondence i under this hypothesis is

$$e_i(\mathbf{H}_t) = \left\|\mathbf{q}_i^m - \mathbf{H}_t\mathbf{p}_i^r\right\|_2, \tag{11}$$

where multiplication by $\mathbf{H}_t$ is understood in homogeneous coordinates and only the resulting 2-D coordinates are retained. A correspondence is classified as an inlier if $e_i(\mathbf{H}_t) < \tau$, where τ is the inlier threshold, which used to be 3. The set of inlier correspondences is denoted as Φ. We only consider the error contributed by inlier correspondences denoted by $\mathcal{S}(\mathbf{H}_t) = \sum_{i\in\Phi} w_i e_i(\mathbf{H}_t)$. After T sampling iterations, the final affine transformation is:

$$\widehat{\mathbf{H}} = \underset{t=1,\ldots,T}{\arg\max}\, \mathcal{S}(\mathbf{H}_t). \tag{12}$$

Table 1: Quantitative comparison on WHU-OPT-SAR and OSDataset under the Easy and Hard settings. @1px, @3px, and @5px denote AUC at the corresponding thresholds. The best and second-best results are highlighted in **bold** and underline, respectively.

Methods	WHU-OPT-SAR(Easy)				OSDataset(Easy)				WHU-OPT-SAR(Hard)				OSDataset(Hard)			
	RMSE*↓	@1px↑	@3px↑	@5px↑	RMSE*↓	@1px↑	@3px↑	@5px↑	RMSE*↓	@1px↑	@3px↑	@5px↑	RMSE*↓	@1px↑	@3px↑	@5px↑
RIFT2Li et al. (2023)	5.45	0.00	0.00	0.19	6.46	0.00	0.07	1.45	5.96	0.00	0.00	0.05	6.56	0.00	0.01	0.98
HOWPZhang et al. (2023)	6.08	0.01	0.55	1.64	5.60	0.00	1.27	5.65	6.00	0.00	0.02	0.07	6.64	0.00	0.06	0.21
WSSFWan et al. (2023)	5.35	0.00	1.59	5.62	5.72	0.00	1.00	5.97	5.62	0.01	0.67	2.22	6.37	0.00	0.42	2.53
MSGZheng et al. (2025)	3.47	0.09	8.99	22.56	3.73	0.16	10.18	25.30	3.78	0.05	4.11	9.66	4.11	0.00	2.55	7.95
LPSDLiu et al. (2025)	2.22	1.60	30.80	47.35	2.69	0.76	22.28	44.31	1.82	2.60	36.25	52.68	2.47	1.23	26.73	48.11
XoFTRTuzcuoğlu et al. (2024)	1.54	4.84	49.83	69.33	2.18	3.15	35.14	57.43	1.90	2.81	40.93	59.66	3.24	0.48	16.63	37.63
RoMaV2Edstedt et al. (2025)	1.42	4.82	**55.68**	72.04	2.11	0.42	34.22	58.20	1.46	7.18	55.02	**71.27**	2.22	0.87	33.33	56.07
GDROSSun et al. (2025)	1.37	7.49	54.62	**72.70**	1.99	4.26	39.10	60.96	1.42	4.86	52.59	70.71	2.55	0.83	25.11	48.98
CRFTLiu et al. (2026)	2.21	7.91	41.08	54.01	**1.77**	**5.38**	**41.60**	**64.54**	3.88	0.00	6.14	16.82	5.92	0.00	0.71	3.08
P2G (Ours)	**1.14**	**10.79**	54.00	66.79	1.85	2.95	38.99	59.97	**0.98**	**15.36**	**58.16**	69.25	**1.71**	**5.33**	**44.73**	**63.89**

4 EXPERIMENTS

Datasets. We conduct experiments on four datasets: WHU-OPT-SAR Li et al. (2022b), OSDataset Xiang et al. (2020b), RoadScene Xu et al. (2020), and MRSIDataset2Zhang et al. (2023). WHU-OPT-SAR and OSDataset contain visible–SAR image pairs with different spatial resolutions, RoadScene is a small-scale visible–infrared dataset, and MRSIDataset2 contains multiple modality combinations. MRSIDataset2 is evaluated using its provided ground truth. For the other three datasets, we apply random affine transformations to the original aligned image pairs and divide the generated samples into Easy and Hard settings according to the transformation magnitude. More details are provided in Appendix B.

Evaluation Metrics. Following MatchAnything He et al. (2025) and GDROS Sun et al. (2025), we report the average root mean square error (RMSE*), success rate (SR@τ), and area under the success-rate curve (AUC@τ). Throughout the paper, RMSE* denotes the average RMSE over image pairs with errors below 10 pixels, excluding gross registration failures from dominating the mean. SR@τ measures the proportion of image pairs with registration errors below τ pixels. AUC@τ is the normalized area under the success-rate curve over the interval $[0, \tau]$, providing a stricter evaluation by accounting for performance across all thresholds within this range. Detailed definitions are provided in Appendix C.

Baselines. We compare P2G with five representative handcrafted methods (RIFT2 Li et al. (2023), HOWP Zhang et al. (2023), WSSF Wan et al. (2023), MSG Zheng et al. (2025), and LPSD Liu et al. (2025)) and six learning-based methods (XoFTR Tuzcuoğlu et al. (2024), GDROS Sun et al. (2025), RoMaV2 Edstedt et al. (2025), and CRFT Liu et al. (2026)). Each learning-based baseline is separately trained and tested under the Easy and Hard settings to ensure evaluation within the corresponding transformation ranges.

4.1 EXPERIMENTAL RESULTS

Visible–SAR Image Registration. Figure 4 and Table 1 present qualitative and quantitative comparisons of representative methods on WHU-OPT-SAR and OSDataset datasets. Under the Easy setting, P2G achieves the lowest RMSE* and highest AUC@1px on WHU-OPT-SAR, while remaining competitive on OSDataset. Its advantage becomes more distinct under the Hard setting, where reliable correspondence selection becomes more challenging. P2G obtains the lowest RMSE* on both datasets and leads at the strict thresholds on WHU-OPT-SAR. On OSDataset, it ranks first across all four metrics, reducing RMSE* from 2.22 to 1.71 pixels and improving AUC@1px, AUC@3px, and AUC@5px by 4.10, 11.40, and 7.82 percentage points over the respective second-best results.

Fig. 5 shows the SR across different RMSE thresholds, further illustrating both the strengths and limitations of P2G. On WHU-OPT-SAR, P2G shows a clear advantage under strict thresholds, whereas learning-based GDROS achieves higher success rates at more relaxed thresholds. Further inspection reveals that some low-texture mountainous scenes contain insufficient distinctive local structures for P2G's keypoint-based candidate generation, leading to registration failures. Compared with the Easy setting, P2G achieves larger gains over competing methods under the Hard setting. This is because larger geometric variations increase the discriminability of transformation-guided structural

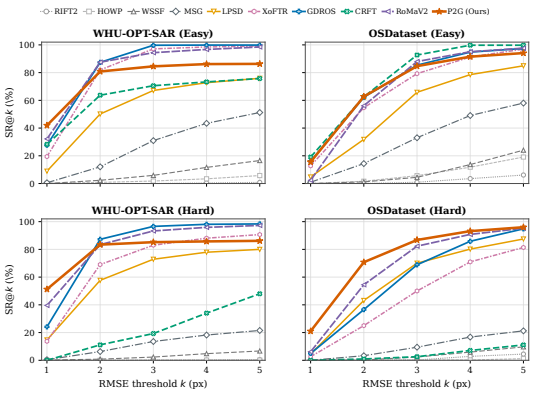


Figure 5: SR under different RMSE thresholds on WHU-OPT-SAR and OSDataset.

Table 2: Quantitative comparison on RoadScene under the Easy and Hard settings. @1px, @3px, and @5px denote AUC at the corresponding thresholds. The best and second-best results are highlighted in **bold** and underline, respectively.

Methods	Easy				Hard			
	RMSE* ↓	@1px↑	@3px↑	@5px↑	RMSE* ↓	@1px↑	@3px↑	@5px↑
RIFT2	5.63	0.00	2.00	8.36	5.48	0.00	1.07	4.76
HOWP	4.77	0.03	4.83	15.98	5.91	0.00	0.07	0.65
WSSF	4.89	0.00	3.62	13.07	5.42	0.00	1.74	7.23
MSG	3.61	0.21	13.45	29.50	3.73	0.25	6.93	15.07
LPSD	2.53	0.22	22.45	46.03	2.39	0.29	24.60	46.60
XoFTR	5.13	0.00	4.30	11.29	4.07	0.00	0.34	4.62
GDROS	2.85	0.63	22.52	40.56	3.03	0.00	9.28	23.10
CRFT	–	0.00	0.00	0.00	–	0.00	0.00	0.00
RoMaV2	5.49	0.00	2.75	10.07	6.40	0.00	0.23	2.00
P2G (Ours)	**1.78**	**3.54**	**44.45**	**61.99**	**1.64**	**6.28**	**46.79**	**62.98**

descriptors among competing candidates, providing more reliable correspondences for subsequent node construction and relation-guided estimation.

Furthermore, we evaluate the reliability of feature graph representation and generalization across geometric difficulty settings on both datasets. Detailed results and analyses are provided in Appendices D and E.1, respectively.

Visible-Infrared Image Registration. RoadScene contains only 221 image pairs, so the learning-based baselines are evaluated using weights trained on the corresponding Easy or Hard setting of WHU-OPT-SAR, without retraining on RoadScene. Table 2 shows that P2G achieves the best results across all reported metrics in both settings. Its advantage is particularly clear in AUC@3px: under the Easy setting, P2G reaches 44.45%, exceeding the strongest baseline by 21.93 percentage points; under the Hard setting, it reaches 46.79%, exceeding the strongest baseline by 22.19 percentage points. P2G also attains the lowest RMSE* in both settings. These results show that its training-free point-to-graph representation transfers effectively from visible–SAR to visible–infrared registration while retaining accuracy under larger geometric transformations.

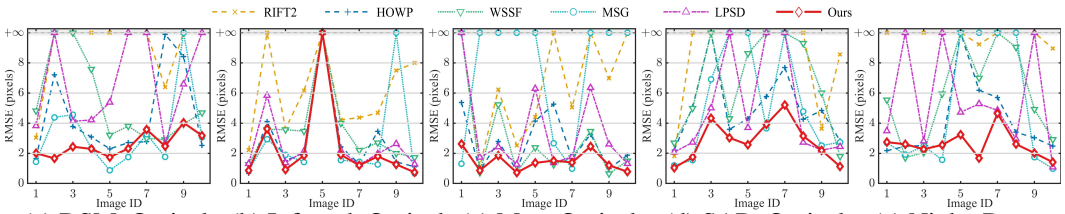
Registration across Diverse Modality. To further evaluate cross-modal generalization, we compare P2G with the handcrafted baselines on MRSIDataset2, which contains five modality combinations with 10 image pairs each and provides ground-truth transformations for evaluation.

As shown in Fig. 6, P2G maintains low registration errors across DSM–optical, infrared–optical, map–optical, SAR–optical, and night–day image pairs, whereas the competing methods exhibit frequent failures on different modality combinations. Among the 50 evaluated pairs, P2G registers 49 pairs with an RMSE below 10 pixels and achieves an Average RMSE* of 2.22 pixels. It is particularly accurate on the map–optical pairs, attaining an average RMSE of 1.48 pixels. Except for one infrared–optical pair, all remaining results are below 5.21 pixels. These results demonstrate that the progressive point-to-graph representation is not restricted to a specific modality combination and generalizes effectively to diverse cross-modal registration tasks without dataset-specific training.

4.2 Ablation Studies

Contribution of Core Components. Table 3(*Left*) shows that each stage of P2G addresses a distinct source of registration error. CG substantially improves performance on both datasets by retaining plausible correspondences instead of committing to the nearest match. NC then resolves this ambiguity: relative to CG alone, it raises AUC@3px from 32.00% to 48.61% on WHU-OPT-SAR and from 22.79% to 37.59% on OSDataset. Finally, relation-guided estimation improves all reported metrics over the constructed nodes. Its gain is especially clear at the strictest threshold, where AUC@1px increases from 6.74% to 15.36% and from 3.16% to 5.33% on the two datasets, respectively. The gains at each stage support the proposed point-to-graph construction, which progressively turns ambiguous candidates into reliable nodes and uses their relations for geometric estimation.

Effect of Adaptive Fusion. Table 3(*Right*) examines how structural and semantic evidence contribute to node construction. Structural features outperform semantic features alone across all met-



(a) DSM–Optical (b) Infrared–Optical (c) Map–Optical (d) SAR–Optical (e) Night–Day

Figure 6: Pair-wise RMSE comparison on MRSIdataset2 over five modalities. RMSE values above 10 pixels are displayed as $+\infty$ for visualization.

Table 3: Ablation studies under the Hard setting. *Left:* Contributions of candidate generation (CG), node construction (NC), and relation guidance (RG) on WHU-OPT-SAR and OSDataset. FSC is used as the geometric estimator when RG is disabled. *Right:* Comparison of different structure-semantic fusion strategies on WHU-OPT-SAR, where Fixed denotes a structural-to-semantic weight ratio of $0.8 : 0.2$. @1px, @3px, and @5px report AUC values at corresponding thresholds.

Setup			WHU-OPT-SAR				OSDataset			
CG	NC	RG	RMSE*↓	@1px↑	@3px↑	@5px↑	RMSE*↓	@1px↑	@3px↑	@5px↑
			4.32	0.04	6.39	19.31	4.29	0.19	5.99	22.17
✓			2.22	2.72	32.00	46.36	2.90	0.80	22.79	42.91
✓	✓		1.32	6.74	48.61	63.44	1.98	3.16	37.59	59.39
✓	✓	✓	**0.98**	**15.36**	**58.16**	**69.25**	**1.71**	**5.33**	**44.73**	**63.89**

Fusion	RMSE*↓	@1px↑	@3px↑	@5px↑
Struct.-only	1.17	12.99	53.23	66.87
Sem.-only	1.39	6.37	48.72	62.80
Fixed	1.10	13.32	55.52	67.48
Adaptive	**0.98**	**15.36**	**58.16**	**69.25**

rics, while fixed fusion improves upon both, confirming that semantic features provide complementary information. However, the reliability of these two cues varies with scene content and local texture, so a fixed fusion weight cannot reflect their relative usefulness for every reference point. Adaptive fusion accounts for this variation through feature discriminability, reducing RMSE* from 1.10 to 0.98 pixels and increasing AUC@1px from 13.32% to 15.36% over fixed fusion.

Effect of Relation Guidance.

Figure 7 evaluates whether relation-guided estimation makes effective use of a limited hypothesis budget. The advantage is most pronounced at 20 iterations: RG achieves an RMSE* of 1.22 pixels and an AUC@1px of 13.77%, compared with 2.05 pixels and 4.98% for FSC and 3.14 pixels and 1.79% for RANSAC Fischler & Bolles (1981). Increasing the budget improves all three methods, but RG retains its lead at 1,000 iterations, reaching an RMSE* of 0.98 pixels and an AUC@1px of 15.36%. This convergence pattern shows that the correspondence graph guides sampling toward useful affine hypotheses early, improving estimation efficiency as well as final accuracy.

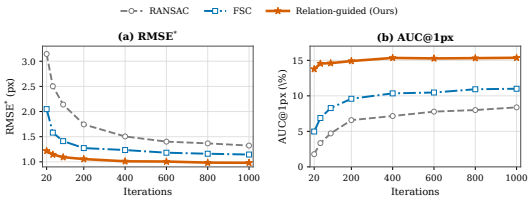


Figure 7: Convergence comparison of RANSAC, FSC, and RG on WHU-OPT-SAR (Hard).

5 Conclusion

This work addresses the ambiguity of isolated point representations under substantial modality gaps and proposes P2G, a training-free cross-modal image registration framework that progressively extends point-level matching to feature graph representation. P2G first preserves multiple plausible correspondences, then constructs reliable one-to-one nodes through complementary structural and semantic representations, and further models the cross-modal stability of intra-image relations as graph edges to guide robust geometric estimation. Experiments on several modality combinations demonstrate that P2G achieves strong registration accuracy and generalization, particularly under large geometric transformations. The ablation studies further indicate that cross-modal registration can benefit not only from improving individual point descriptors, but also from explicitly incorporating stable inter-point relations into the representation, suggesting a promising direction for improving cross-modal feature representations.

6 AI USE STATEMENT

Generative AI tools were used to assist with research discussions, including feedback on the experimental design, as well as the interpretation of experimental results. They were also used to assist with code development, table preparation, literature searches, English translation, and manuscript drafting and editing. For example, AI tools helped format comparison tables using experimental results provided by the authors; they did not generate the reported measurements. The authors checked the tables against the original results and reviewed and revised all AI-assisted code and text. The authors take responsibility for the methods, results, claims, and final content of this paper.

REFERENCES

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

Johan Edstedt, David Nordström, Yushan Zhang, Georg Bökman, Jonathan Astermark, Viktor Larsson, Anders Heyden, Fredrik Kahl, Mårten Wadenbäck, and Michael Felsberg. Roma v2: Harder better faster denser feature matching. *arXiv preprint arXiv:2511.15706*, 2025.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Xingyi He, Hao Yu, Sida Peng, Dongli Tan, Zehong Shen, Hujun Bao, and Xiaowei Zhou. Matchanything: Universal cross-modality image matching with large-scale pre-training. *arXiv preprint arXiv:2501.07556*, 2025.

Mattias P Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V Gleeson, Michael Brady, and Julia A Schnabel. Mind: Modality independent neighbourhood descriptor for multimodal deformable registration. *Medical image analysis*, 16:1423–1435, 2012.

Rohit Jena, Vedant Zope, Pratik A Chaudhari, and James Gee. A scalable distributed framework for multimodal gigavoxel image registration. In *Proceedings of the International Conference on Learning Representations (ICLR)*, volume 2026, pp. 119573–119592, 2026.

Jiayuan Li, Qingwu Hu, and Mingyao Ai. Rift: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing (TIP)*, 29:3296–3310, 2019.

Jiayuan Li, Wangyi Xu, Pengcheng Shi, Yongjun Zhang, and Qingwu Hu. Lnift: Locally normalized image for rotation invariant multimodal feature matching. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 60:1–14, 2022a.

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.

Xue Li, Guo Zhang, Hao Cui, Shasha Hou, Shunyao Wang, Xin Li, Yujia Chen, Zhijiang Li, and Li Zhang. Mcanet: A joint semantic segmentation framework of optical and sar images for land use classification. *International Journal of Applied Earth Observation and Geoinformation (INT J APPL EARTH OBS)*, 106:102638, 2022b.

Junyi Liu, Qingsong Wang, Diling Liao, Junjie Huang, Tao Lai, and Haifeng Huang. Lpsd: Log-polar sampling descriptor for efficient and robust multimodal image matching. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 63:1–19, 2025.

Rui Liu, Jing Ling, and Hongsheng Zhang. Softformer: Sar-optical fusion transformer for urban land use and land cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS)*, 218:277–293, 2024a.

Xuecong Liu, Mengzhu Ding, Zixuan Sun, Zhang Li, and Xichao Teng. Crft: Consistent-recurrent feature flow transformer for cross-modal image registration. *arXiv preprint arXiv:2604.05689*, 2026.

Yijun Liu, Mingxin Lin, Yuanhui Mo, and Qingsong Wang. SAR–optical image matching using self-supervised detection and a transformer–CNN-based network. *IEEE Geoscience and Remote Sensing Letters (GRSL)*, 21:1–5, 2024b.

Tao Peng, Zhiqiang Han, Tengfeng Tang, Xianping Ma, and Yuanxin Ye. Ripc: A novel fft-based multimodal image matching framework with radiometric, scale and rotation invariance. *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS)*, 240:671–689, 2026.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4938–4947, 2020.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8922–8931, 2021.

Zixuan Sun, Shuaifeng Zhi, Kai Huo, Xuecong Liu, Weidong Jiang, and Yongxiang Liu. Os3flow: Optical and sar image registration using symmetry-guided semi-dense optical flow. *IEEE Geoscience and Remote Sensing Letters (GRSL)*, 2024.

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 (TGRS)*, 63:1–15, 2025.

Önder Tuzcuoğlu, Aybora Köksal, Buğra Sofu, Sinan Kalkan, and A Aydin Alatan. Xoftr: Cross-modal feature matching transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4275–4286, 2024.

Genyi Wan, Zhen Ye, Yusheng Xu, Rong Huang, Yingying Zhou, Huan Xie, and Xiaohua Tong. Multimodal remote sensing image matching based on weighted structure saliency feature. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 62:1–16, 2023.

Yue Wu, Wenping Ma, Maoguo Gong, Linzhi Su, and Licheng Jiao. A novel point-matching algorithm based on fast sample consensus for image registration. *IEEE Geoscience and Remote Sensing Letters (GRSL)*, 12(1):43–47, 2014.

Yuming Xiang, Rongshu Tao, Ling Wan, Feng Wang, and Hongjian You. Os-pc: Combining feature representation and 3-d phase correlation for subpixel optical and sar image registration. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 58(9):6451–6466, 2020a.

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 (J-STARS)*, 13:5847–5861, 2020b.

Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

Yuanxin Ye and Li Shen. Hopc: A novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching. *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS)*, 3:9–16, 2016.

Yuanxin Ye, Lorenzo Bruzzone, Jie Shan, Francesca Bovolo, and Qing Zhu. Fast and robust matching for multimodal remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 57(11):9059–9070, 2019.

Yuanxin Ye, Tengfeng Tang, Bai Zhu, Chao Yang, Bo Li, and Siyuan Hao. A multiscale framework with unsupervised learning for remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 60:1–15, 2022a.

Yuanxin Ye, Bai Zhu, Tengfeng Tang, Chao Yang, Qizhi Xu, and Guo Zhang. A robust multimodal remote sensing image registration method and system using steerable filters with first-and second-order gradients. *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS)*, 188:331–350, 2022b.

Haiming Zhang, Guorui Ma, Di Wang, and Yongxian Zhang. M3icnet: A cross-modal resolution preserving building damage detection method with optical and sar remote sensing imagery and two heterogeneous image disaster datasets. *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS)*, 221:224–250, 2025a.

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*, 2025b.

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 (ISPRS)*, 196:1–15, 2023.

Yongjun Zhang, Wenfei Zhang, Yongxiang Yao, Zhi Zheng, Yi Wan, and Mingtao Xiong. Robust registration of multi-modal remote sensing images based on multi-dimensional oriented self-similarity features. *International Journal of Applied Earth Observation and Geoinformation (INT J APPL EARTH OBS)*, 127:103639, 2024a.

Yongjun Zhang, Peihao Wu, Yongxiang Yao, Yi Wan, Wenfei Zhang, Yansheng Li, and Xiaohu Yan. Multi-modal remote sensing image robust matching based on second-order tensor orientation feature transformation. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 2025c.

Yongxian Zhang, Chaozhen Lan, Haiming Zhang, Guorui Ma, and Heng Li. Multimodal remote sensing image matching via learning features and attention mechanism. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 62:1–20, 2024b.

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 (TGRS)*, 2025.

Yan Zhou, Jinding Gao, and Xiaoping Liu. A unified feature-motion consistency framework for robust image matching. *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS)*, 218: 368–388, 2024.

A Additional Method Details

This section provides the descriptor and similarity calculations used in the node and edge construction of Secs. 3.2 and 3.3.

A.1 Reliability-Aware Node Construction

Local Structural Evidence. For each reference keypoint and its Top-K query image candidates, we construct Cartesian structural descriptors from the FoGG orientation map. A 128×128 support region is sampled with a stride of 2 and divided into an $N_{\text{block}} \times N_{\text{block}}$ grid. Within each cell, responses from N_{bin} quantized orientations are accumulated into a histogram. Concatenating the cell-wise histograms gives an $N_{\text{block}} \times N_{\text{block}} \times N_{\text{bin}}$-dimensional descriptor. We use $N_{\text{block}} = 16$ and $N_{\text{bin}} = 6$, producing a 1536-dimensional descriptor.

Let $\mathbf{o}_i^r$ and $\mathbf{o}_{ik}^m$ denote the ℓ_2-normalized descriptors of reference keypoint $\mathbf{p}_i^r$ and its k-th candidate $\mathbf{p}_{ik}^m$. Their structural distance is

$$d_{ik}^o = 1 - \langle \mathbf{o}_i^r, \mathbf{o}_{ik}^m \rangle , \tag{13}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. A smaller distance indicates greater local structural similarity.

Semantic Matching Evidence. Since $\mathbf{H}_0$ maps reference image coordinates to query image coordinates, we use $\mathbf{H}_0^{-1}$ to warp the query image into the reference coordinate frame. The reference image and the coarsely aligned query image, both of size $H \times W$, are processed by a frozen DINOv3 ViT-L/16 model. We extract the 24th-layer feature map of size $H/16 \times W/16 \times C$, where $C = 1024$. Each C-dimensional patch token is concatenated with the mean feature of its 3×3 neighborhood, giving a $2C$-dimensional representation. Descriptors at keypoint locations are obtained by bilinear interpolation on the contextual feature grid and then ℓ_2-normalized. For the semantic descriptors $\mathbf{m}_i^r$ and $\mathbf{m}_{ik}^m$, the distance is

$$d_{ik}^m = 1 - \langle \mathbf{m}_i^r, \mathbf{m}_{ik}^m \rangle . \tag{14}$$

A.2 Relational Edge Construction

Hierarchical Intra-Image Feature Similarities. After one-to-one correspondences have been selected, features from four DINOv3 layers describe the image locations represented by each node. The selected layer set is $\mathcal{L} = \{0, 1, 2, 3\}$. At layer ℓ, $\mathbf{z}_i^{r,(\ell)}$ and $\mathbf{z}_i^{m,(\ell)}$ are the normalized feature descriptors of $\mathbf{p}_i^r$ and $\mathbf{q}_i^m$, respectively. For nodes i and j, their within-image similarities are

$$r_{ij}^{r,(\ell)} = \left\langle \mathbf{z}_i^{r,(\ell)}, \mathbf{z}_j^{r,(\ell)} \right\rangle , \qquad r_{ij}^{m,(\ell)} = \left\langle \mathbf{z}_i^{m,(\ell)}, \mathbf{z}_j^{m,(\ell)} \right\rangle . \tag{15}$$

Both inner products compare locations within the same image. The layer-wise compatibility in Sec. 3.3 then compares the reference and query image similarities.

B Training and Test Data

The WHU-OPT-SAR dataset has a spatial resolution of 5 m and contains 100 large-size visible–SAR image pairs covering diverse terrain types. To improve processing efficiency, the original images are cropped into 512×512-pixel patches, resulting in 5,300 image pairs for training and 730 image pairs for testing. OSDataset has a spatial resolution of 1 m and contains 2,673 aligned visible–SAR image pairs from multiple geographic regions. It is divided into 2,011 training pairs, 238 validation pairs, and 424 test pairs. The RoadScene dataset contains 221 aligned visible–infrared image pairs selected from FLIR video sequences. It covers diverse traffic scenes containing roads, vehicles, pedestrians, and other representative objects, providing realistic cross-modal appearance variations for visible–infrared registration. MRSIDataset2 is a small-scale multimodal image registration benchmark covering five modality combinations: infrared–visible, SAR–visible, map–visible, DSM–visible, and night–day. Each combination contains 10 image pairs. Unlike the other three datasets, MRSIDataset2 is evaluated directly using its provided ground truth without additional synthetic transformations.

For WHU-OPT-SAR, OSDataset, and RoadScene, random affine transformations are applied to the original aligned image pairs. We define two settings according to the transformation magnitude. **Easy Transformation:** The rotation angle is randomly sampled from $[-30°, 30°]$, the horizontal and vertical translations are sampled from $[-50, 50]$ pixels, and the scaling factor is sampled from $[0.8, 1.2]$. **Hard Transformation:** The rotation angle is randomly sampled from $[-180°, 180°]$, the horizontal and vertical translations are sampled from $[-50, 50]$ pixels, and the scaling factor is sampled from $[0.6, 1.4]$.

C Evaluation Metrics

For the i-th image pair, let $(\widehat{x}_{ij}, \widehat{y}_{ij})$ and $(x_{ij}^{\mathrm{gt}}, y_{ij}^{\mathrm{gt}})$ denote the estimated and ground-truth coordinates of the j-th evaluation point, respectively. Its squared endpoint error and the RMSE of the image pair are

$$e_{ij} = \sqrt{(\widehat{x}_{ij} - x_{ij}^{\mathrm{gt}})^2 + (\widehat{y}_{ij} - y_{ij}^{\mathrm{gt}})^2}, \qquad r_i = \mathrm{RMSE}_i = \sqrt{\frac{1}{M_i}\sum_{j=1}^{M_i} e_{ij}^2} \tag{16}$$

where M_i is the number of evaluation points in the i-th image pair. Average RMSE* is the arithmetic mean of RMSE_i over image pairs with $\mathrm{RMSE}_i \le 10$ pixels. Median RMSE is computed over all evaluated image pairs. Given T test image pairs, the success rate at threshold τ is defined as

$$\mathrm{SR}\,@\tau = \frac{1}{T}\sum_{i=1}^{T} \mathbb{I}(\mathrm{RMSE}_i \le \tau), \qquad \tau \in \{1, 3, 5\}, \tag{17}$$

where $\mathbb{I}(\cdot)$ is the indicator function. SR@τ measures the proportion of image pairs satisfying the specified accuracy threshold. AUC@τ is the normalized area under the success-rate curve over $[0, \tau]$:

$$\mathrm{AUC}\,@\tau = \frac{1}{\tau}\int_0^{\tau} \mathrm{SR}(t)\,\mathrm{d}t. \tag{18}$$

Unlike SR@τ, which evaluates a single threshold, AUC@τ accounts for registration errors throughout $[0, \tau]$. Lower Average RMSE* and Median RMSE, and higher SR@τ and AUC@τ, indicate better performance.

D Why Does Feature Graph Representation Work?

We examine whether relations between correspondences provide a more reliable signal than isolated point features. This analysis uses the one-to-one correspondences produced by NC on the Hard splits of WHU-OPT-SAR and OSDataset. It does not change candidate selection or estimate an affine transformation. A correspondence is labeled correct when its error under the ground-truth transformation is at most 3 pixels.

Table 4: Reliability of selected one-to-one correspondences on the Hard splits. Results are image-pair-level averages at a 3-pixel correctness threshold. Shuffled results are averaged over ten permutations per image pair. Higher is better.

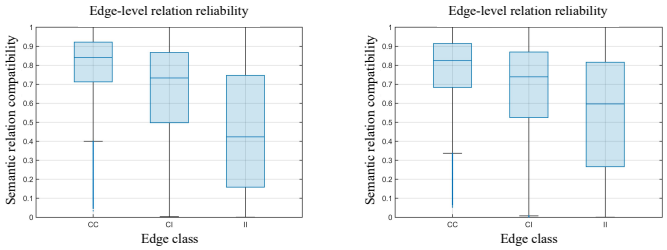
	WHU-OPT-SAR		OSDataset	
Score	ROC-AUC	AP	ROC-AUC	AP
Point-wise semantics	0.670	0.366	0.651	0.357
Shuffled graph	0.498	0.259	0.499	0.263
Feature graph	**0.729**	**0.440**	**0.688**	**0.397**

For each correspondence, the point-wise score is the negative DINOv3 feature distance of its selected point pair; a larger value indicates higher similarity. The graph-based score is the mean relational compatibility of its incident edges in the sparse feature graph. We also construct a shuffled control by permuting edge weights within each image pair while retaining the graph connectivity and the set of edge weights. ROC-AUC and average precision (AP) are computed separately for each image

pair and then averaged. All methods are evaluated on the same correspondences with valid semantic features.

Table 4 shows that the feature-graph score ranks correct correspondences more accurately than point-wise semantic similarity on both datasets. Its ROC-AUC exceeds the point-wise score by 0.059 on WHU-OPT-SAR and 0.037 on OSDataset. Permuting the edge weights removes this advantage in the image-pair-level evaluation, indicating that the result depends on how relational compatibility is assigned to the graph edges.

We further examine whether individual edges carry useful information. An edge belongs to CC if both endpoint correspondences are correct, CI if exactly one is correct, and II if neither is correct. Figure 8 shows the distributions of semantic relation compatibility for these three classes. On both datasets, CC edges have a higher median compatibility than CI and II edges. Although the distributions overlap, this tendency supports the use of relational compatibility to guide hypothesis sampling after one-to-one correspondences have been constructed.



(a) WHU-OPT-SAR (b) OSDataset

Figure 8: Edge-level relation reliability on the Hard splits of (a) WHU-OPT-SAR and (b) OSDataset. CC, CI, and II denote edges connecting two, one, and zero correct correspondences, respectively; correctness is defined by a 3-pixel ground-truth error threshold. The vertical axis shows semantic relation compatibility, rather than the final sampling probability. For visualization, at most 200,000 CI and 200,000 II edges are sampled per image pair.

E MORE STUDIES

E.1 Generalization Across Geometric Difficulty Settings

Table 5: Generalization across geometric difficulty settings on WHU-OPT-SAR and OSDataset. Matched denotes training and testing under the same Easy or Hard setting, whereas Cross denotes training on Easy and testing on Hard, or training on Hard and testing on Easy. P2G requires no setting-specific training. @1px, @3px, and @5px denote AUC at the corresponding thresholds.

Train/Test	Methods	WHU-OPT-SAR(Easy)				OSDataset(Easy)				WHU-OPT-SAR(Hard)				OSDataset(Hard)			
		RMSE*↓	@1px↑	@3px↑	@5px↑	RMSE*↓	@1px↑	@3px↑	@5px↑	RMSE*↓	@1px↑	@3px↑	@5px↑	RMSE*↓	@1px↑	@3px↑	@5px↑
Matched	XoFTRTuzcuoğlu et al. (2024)	1.54	4.84	49.83	69.33	2.18	3.15	35.14	57.43	1.90	2.81	40.93	59.66	3.24	0.48	16.63	37.63
	GDROSSun et al. (2025)	1.37	7.49	54.62	72.70	1.99	4.26	39.10	60.96	1.42	4.86	52.59	70.71	2.55	0.83	25.11	48.98
	CRFTLiu et al. (2026)	2.21	7.91	41.08	54.01	1.77	5.38	41.60	64.54	3.88	0.00	6.14	16.82	5.92	0.00	0.71	3.08
	RoMaV2Edstedt et al. (2025)	1.42	4.82	55.68	72.04	2.11	0.42	34.22	58.20	1.46	7.18	55.02	71.27	2.22	0.87	33.33	56.07
Cross	XoFTRTuzcuoğlu et al. (2024)	2.20	1.33	32.44	53.34	3.10	0.74	18.35	39.51	2.59	0.45	4.43	5.83	2.53	0.13	1.99	3.79
	GDROSSun et al. (2025)	1.72	3.32	42.22	62.72	2.72	0.71	22.00	46.26	1.71	0.63	4.69	5.99	2.34	0.13	2.25	4.07
	CRFTLiu et al. (2026)	6.08	0.00	0.34	2.70	6.45	0.00	0.50	2.94	–	0.00	0.00	0.00	6.50	0.00	0.00	0.00
	RoMaV2Edstedt et al. (2025)	1.89	2.12	42.14	61.48	2.25	0.59	30.25	55.26	1.48	0.78	8.63	11.09	3.43	0.07	10.33	22.80
Training-free	**P2G (Ours)**	1.14	10.79	54.00	66.79	1.85	2.95	38.99	59.97	0.98	15.36	58.16	69.25	1.71	5.33	44.73	63.89

Table 5 examines how learned methods transfer between the Easy and Hard settings of the same dataset. When trained and tested under the same setting, several learned methods achieve strong results. Their performance declines sharply, however, when models trained on Easy are tested on Hard. On WHU-OPT-SAR, the best learned AUC@3px under this transfer is 8.63%, compared with 58.16% for P2G; on OSDataset, the corresponding values are 10.33% and 44.73%. Transfer from Hard to Easy is less severe for some methods, indicating that the generalization gap depends on the direction of the geometric shift. P2G requires no setting-specific training and maintains the same

performance in both comparisons. These results show that its registration accuracy is less dependent on prior exposure to a particular range of geometric transformations.

E.2 Effect of the Candidate Number

We evaluate the candidate number K on WHU-OPT-SAR under the Hard setting. Table 6 reports the results for $K \in \{2, 4, 5, 6, 8\}$.

Table 6: Effect of the candidate number K on WHU-OPT-SAR under the Hard setting. SR and AUC values are reported as percentages. The best and second-best results are highlighted in **bold** and underline, respectively.

K	RMSE*↓	SR@1↑	SR@3↑	SR@5↑	AUC@1↑	AUC@3↑	AUC@5↑
2	1.22	34.12	80.82	82.19	7.87	47.36	61.89
4	1.07	48.77	84.80	<u>86.03</u>	13.23	56.10	67.89
5	**0.98**	51.23	<u>85.21</u>	**86.16**	15.36	**58.16**	**69.25**
6	<u>0.99</u>	<u>51.64</u>	<u>85.21</u>	**86.16**	<u>15.42</u>	<u>58.09</u>	<u>69.17</u>
8	1.00	**52.10**	**85.34**	**86.16**	**15.63**	57.79	69.02

Performance improves substantially as K increases from 2 to 5, with RMSE* decreasing from 1.22 to 0.98 pixels and AUC@5 increasing from 61.89% to 69.25%. Further increasing K yields only marginal gains at the strict 1-pixel threshold while slightly reducing AUC@3 and AUC@5. We therefore adopt $K = 5$ as the best overall trade-off between candidate coverage, registration accuracy, and computational cost.

F Computational Cost Analysis

P2G takes an average of 4.05 s to register a pair of 512×512-pixel images. Compared with conventional handcrafted methods, DINOv3 inference introduces additional computational overhead. Its inference speed is lower than that of the learning-based baselines, which typically require less than 1 s per pair. However, applying learning-based methods to a real scene application may require extra task-specific training or fine-tuning, introducing computational costs beyond inference. P2G avoids this additional training burden and can be applied directly to new registration tasks. Thus, despite its longer inference time, P2G offers practical deployment flexibility, particularly when training resources or task-specific data are limited.