OSCAR: TOWARDS GENERALIZABLE OPTICAL-SAR CROSS-MODAL IMAGE REGISTRATION WITHOUT GEOMETRIC SHORTCUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Optical–SAR cross-modal image registration is a fundamental yet challenging problem in remote sensing. Existing methods often perform well on individual datasets but generalize poorly across sensors, scenes, and geometric transformations. In this work, we show that part of the apparent success on current registration benchmarks can arise from an overlooked geometric shortcut rather than genuine cross-modal correspondence learning. We identify an overlooked shortcut in common affine-warping protocols: zero padding creates black-border patterns that are deterministically related to the transformation, allowing models to infer geometry without learning genuine cross-modal correspondences. To address this issue, we introduce a shortcut-free data construction protocol that removes transformation-dependent padding artifacts while preserving exact dense supervision. Leveraging on this rigorous setting, we propose OSCAR, a generalizable optical-SAR registration framework designed to address the two challenges exposed after shortcut removal: severe cross-modal appearance discrepancy and large global displacement. OSCAR combines foundation-model semantic features with trainable fine-grained representations, performs long-range bidirectional cross-modal interaction, explicitly estimates the dominant affine motion, and recurrently refines the remaining displacement. Jointly trained on seven heterogeneous datasets, our model consistently outperforms all jointly/separately trained baselines and improves zero-shot performance on two unseen datasets, highlighting the importance of eliminating geometric shortcuts and cross-domain trainging of our framework. All codes and datasets will be open-sourced.

1 INTRODUCTION

Optical and synthetic aperture radar (SAR) sensors provide highly complementary observations of the Earth's surface: optical imagery captures rich spectral and semantic information, whereas SAR remains insensitive to changes in illumination and adverse weather conditions. Optical-SAR image registration underpins a wide range of multi-modal remote-sensing applications, yet it remains challenging because the two modalities arise from fundamentally different imaging mechanisms and exhibit substantial discrepancies in appearance, structure, and noise. Recent learning-based methods have substantially improved optical-SAR registration accuracy on individual benchmarks whose training and testing share similar sensors, scenes, and transformation ranges. However, models trained on one dataset often deteriorate under cross-dataset and large-displacement settings. This limitation is particularly consequential for remote sensing, where sensor configurations and scene distributions vary continuously in real-world deployment. It motivates a central question of this work: what prevents current optical–SAR registration models from learning correspondences that robustly generalize beyond the datasets on which they are trained?

Investigating this question reveals an issue that is largely orthogonal to model architecture. The scarcity of manually aligned optical-SAR pairs has led existing datasets to generate self-supervised training samples by applying synthetic affine transformations to co-registered images. This provides controllable displacement and precise geometric supervision, but it also introduces an unintended source of information: Pixels falling outside the transformed image boundaries are commonly zero-padded, producing border patterns whose geometry is deterministically linked to the transformation

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.

parameters. A sufficiently expressive model may therefore partially solve registration by learning the mapping from boundary geometry to transformation, rather than by establishing correspondences between optical and SAR scene content. We experimentally show that registration models can exploit these borders: they perform well when the same artifacts are present at test time, but degrade substantially on border-free images with nearly identical scene content. Most strikingly, they can still predict plausible registration when most valid image content is masked and only the border patterns remain. These findings indicate that conventional protocols may overestimate genuine cross-modal correspondence learning.

This diagnosis leads us to revisit the problem from both the data and model perspectives. We first develop a shortcut-free construction protocol that removes padding-induced boundary cues while retaining exact dense supervision. Removing this shortcut exposes the persistent difficulty of optical-SAR registration for existing baselines including severe optical–SAR modality gap and geometric displacement, and also provides a more reliable benchmark for evaluating cross-modal correspondence learning. We further propose OSCAR, a generalizable optical–SAR registration framework built on two principles: modality-robust representation and geometry-aware initialization. OSCAR fuses complementary semantic and fine-grained representations with explicit global geometric reasoning and estimates the dominant global motion before registration refinement. Trained on 7 heterogeneous optical-SAR domains, OSCAR achieves the best endpoint registration accuracy and improved generalization across both in-domain and unconstrained images.

We summarize our key contributions as follows:

- **Benchmark Diagnosis.** We identify and experimentally verify a geometric shortcut in affine-warped optical–SAR benchmarks, where padding boundaries reveal the applied transformation without requiring genuine cross-modal correspondence.
- **Shortcut-Free Benchmark.** We introduce a border-free data-generation protocol that removes padding artifacts while preserving exact dense geometric supervision through analytical transformation propagation. Based on public optical–SAR datasets, we construct a large-scale benchmark for training and evaluation.
- **Generalizable Registration Framework.** We develop OSCAR, a Transformer-based optical–SAR registration framework that combines semantic and local dense features with global cross-modal attention and parameterized affine-motion initialization. A single jointly trained model achieves accurate registration across heterogeneous in-domain and held-out datasets.

2 Related Work

2.1 Dense Correspondence Estimation

Dense correspondence estimation mainly follows two directions: optical flow and image matching. PWC-Net Sun et al. (2018) combines feature pyramids, warping, and cost volumes, while RAFT Teed & Deng (2020) uses all-pairs correlations and recurrent refinement. FlowFormer Huang et al. (2022) applies Transformers to cost-volume encoding and flow decoding. SEA-RAFT Wang et al. (2024) introduces a Mixture-of-Laplace loss and direct initial-flow prediction, whereas DPFlow Morimitsu et al. (2025) uses adaptive dual pyramids to model global context and fine details. WAFT Wang & Deng (2026) replaces explicit cost volumes with high-resolution feature warping and provides the optical-flow framework adopted by OSCAR. Image matching has evolved from sparse feature association to detector-free dense estimation. SuperGlue Sarlin et al. (2020) combines attention-based graph reasoning with differentiable optimal transport, and LoFTR Sun et al. (2021) performs coarse matching and fine refinement through self- and cross-attention. RoMa Edstedt et al. (2024) integrates foundation-model and convolutional features, while RoMaV2 Edstedt et al. (2025) further incorporates DINOv3 representations Siméoni et al. (2025). Despite their success on natural images, these methods remain difficult to transfer to optical–SAR imagery because of the substantial cross-modal appearance gap.

2.2 Cross-Modal Optical–SAR Image Registration

Optical-SAR registration is challenging because of substantial radiometric and structural differences. Traditional methods, including SIFT Lowe (2004), HOPC Ye et al. (2017), CFOG Ye et al.

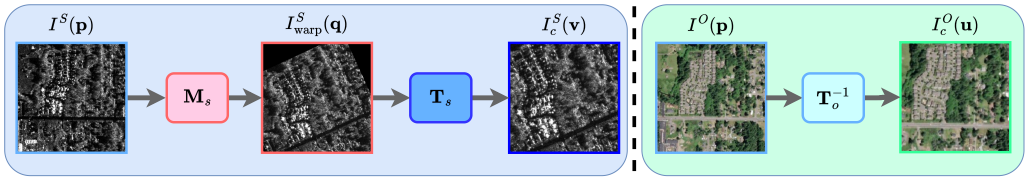

Figure 1: Optical–SAR Matching Dataset Generation with Exact Optical Flow Ground Truth and No Geometric or Black-Border Shortcuts.

(2019), RIFT Li et al. (2020), and LNIFT Li et al. (2022a), rely on handcrafted gradient, phase-congruency, or structural descriptors, but often lack robustness in complex scenes and under large geometric transformations.Learning-based methods include general cross-modal matchers, such as XoFTR Tuzcuoglu et al. (2024) and MINIMA Ren et al. (2025), and optical–SAR-specific approaches, such as GDROS Sun et al. (2025) and TAR Cai et al. (2026). However, limited training data and domain variation often restrict their generalization to unseen resolutions and scenes. Our method addresses these limitations through joint training on heterogeneous datasets while modeling both cross-modal representation discrepancies and large global displacements.

2.3 Optical-SAR Datasets and Geometric Supervision

Existing optical–SAR benchmarks mainly provide co-registered or approximately aligned image pairs for matching, fusion, classification, and segmentation Ye et al. (2025); Xiang et al. (2025); Huang et al. (2023); Li et al. (2022b). Although they cover diverse sensors, resolutions, and scenes, most lack the large displacements and dense geometric labels required for registration training. Synthetic affine transformations are therefore applied to aligned pairs to generate controlled displacements and dense supervision for self-supervised training Wang et al. (2026); Sun et al. (2024). However, zero padding creates boundary patterns deterministically linked to the applied transformation, allowing models to infer geometry from dataset artifacts rather than genuine cross-modal correspondences. This constitutes a form of shortcut learning Geirhos et al. (2020); Lapuschkin et al. (2019). We address this issue with a border-free pipeline that combines affine warping, valid-region cropping, and resizing while preserving exact dense geometric supervision.

3 Method

3.1 Dataset Construction Pipeline without Geometric Shortcut

Conventional optical–SAR matching datasets simulate geometric misalignment by warping one image of a co-registered pair. However, constant padding during warping introduces artificial borders whose shapes depend on the applied transformation, potentially providing unintended geometric shortcuts (Fig. 3). We therefore construct image pairs that remove these padding-induced boundary cues while preserving analytically recoverable geometric correspondences.

Optical–SAR Data Generation. Given a co-registered optical–SAR pair (I^O, I^S) of resolution $H \times W$, we keep the optical image unchanged and perturb the SAR image using a randomly sampled affine transformation $\mathbf{M}_s$. The perturbation combines rotation θ and isotropic scaling s about the image center with translation $\mathbf{t} = (t_x, t_y)^\top$. The warped SAR image I^S_{warp} is generated by inverse warping with bilinear interpolation. Sampling outside the original image requires constant padding, which creates the artificial boundaries targeted by our construction.

Geometric Shortcut-free Data Generation. As illustrated in Fig. 1, we first identify the largest axis-aligned square C_s fully contained within the valid, non-padded region of I^S_{warp} using maximal-square dynamic programming. Let k denote its side length. We crop the warped SAR image within C_s and extract an equally sized square C_o centered in the original optical image. Both crops are then resized to $H \times W$, yielding the final pair (I^O_c, I^S_c).

Because cropping and resizing change both coordinate systems, we update the geometric supervision by composing these operations with the original affine warp. Specifically, the final optical-to-SAR mapping is $\bar{\mathbf{M}}_f = \mathbf{T}_s \bar{\mathbf{M}}_s \mathbf{T}_o$, where $\bar{\mathbf{M}}_s$ is the homogeneous extension of $\mathbf{M}_s$, $\mathbf{T}_o$ maps final optical coordinates back to the original optical image, and $\mathbf{T}_s$ maps warped SAR coordinates to the final SAR image. This composition accounts for the different crop origins and the resizing

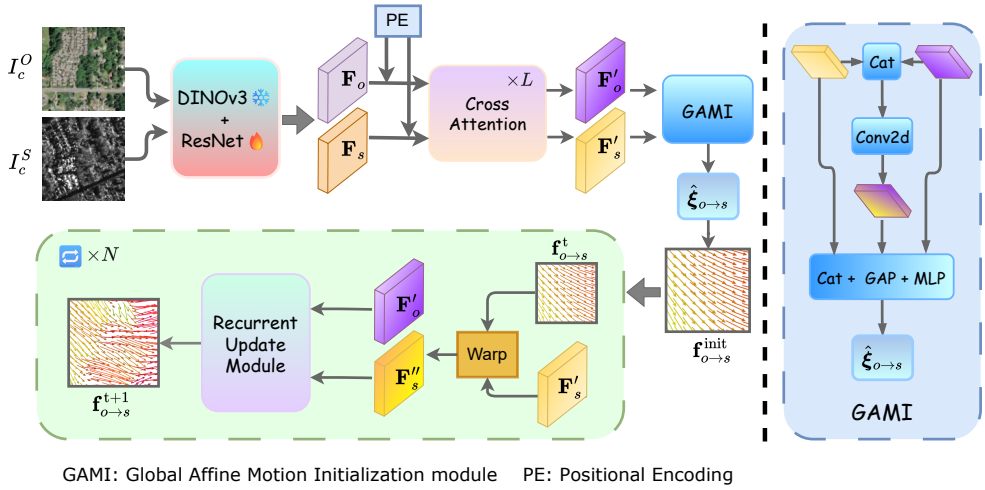

Figure 2: Optical–SAR Cross-modal Image Registration Model.We use cross-attention to extract long-range cross-modal features and employ a Global Affine Motion Initialization module to predict a prior flow field as the starting point for iterative refinement.

operation, retaining exact geometric supervision after the padded boundaries are removed. Sampling distributions, crop definitions, and the complete coordinate derivation are provided in Appendix A.1.

3.2 OSCAR: Generalizable Optical-SAR Image Registration Framework

This section introduces the proposed OSCAR model. Following recent advances in optical-flow estimation, we reconstruct the model based on WAFT, which provides a simple architecture and efficient flow estimation.

Feature Extractor. We retain the basic architecture of WAFT while freezing the DINOv3 backbone and training its DPT head. In parallel, a fully trainable ResNet18 is employed to extract fine-grained local features. The features obtained from the two branches are subsequently fused to produce the optical and SAR image features, denoted by $\mathbf{F}_o$ and $\mathbf{F}_s$, respectively. Both feature maps have a spatial resolution equal to half of the original image resolution.

Optical–SAR Global Cross-Attention Module (OSCA). The substantial modality gap between optical and SAR images makes direct flow estimation from their features difficult. To improve cross-modal interaction, OSCA first augments the extracted optical and SAR features with sinusoidal positional encodings and then applies L bidirectional linear cross-attention layers. At each layer, the optical features attend to the SAR keys and values, while the SAR features are updated symmetrically using the optical features. Residual connections and learnable projections are used throughout the updates.

The resulting enhanced features, denoted by $\mathbf{F}'_o$ and $\mathbf{F}'_s$, provide modality-aware representations for subsequent registration. Linear attention avoids explicitly constructing the full quadratic attention matrix, thereby reducing memory usage and computational cost. The complete formulation is provided in Appendix A.2.

Global Affine Motion Initialization (GAMI). The original WAFT-like framework receives a full-zero initial flow, which is appropriate when complex motion is mainly localized to foreground regions. In optical-SAR registration, however, displacement is typically global, making zero-flow initialization a less favorable initial guess before recurrent flow refinement. We therefore use a global affine flow as the initial estimate.

Specifically, GAMI takes the enhanced features $\mathbf{F}'_o$ and $\mathbf{F}'_s$ as input and predicts a six-dimensional restricted affine motion vector:

$$\hat{\boldsymbol{\xi}}_{o\to s} = \mathrm{GAMI}\left(\mathbf{F}'_o, \mathbf{F}'_s\right) = \left[\hat{t}_x, \hat{t}_y, \widehat{\cos\theta}, \widehat{\sin\theta}, \widehat{\Delta\log s_x}, \widehat{\Delta\log s_y}\right]^{\mathsf{T}}, \tag{1}$$

where $\hat{t}_x$ and $\hat{t}_y$ denote horizontal and vertical translations, $\widehat{\cos\theta}$ and $\widehat{\sin\theta}$ encode rotation, and $\widehat{\Delta\log s_x}$ and $\widehat{\Delta\log s_y}$ represent logarithmic scale variations along the two spatial axes.

The predicted parameters are deterministically rendered into a dense initial flow field:

$$\mathbf{f}_{o\to s}^{\text{init}} = \mathcal{G}_{\text{aff}}\left(\hat{\boldsymbol{\xi}}_{o\to s}\right), \tag{2}$$

where $\mathcal{G}_{\text{aff}}(\cdot)$ maps the affine parameters to the corresponding dense flow. This global motion prior captures the dominant displacement before recurrent refinement, allowing subsequent iterations to focus on local residual flow.

The resulting initial flow provides a global motion prior for the subsequent recurrent refinement module. It enables the model to capture the dominant displacement at the beginning of the iterative process, allowing subsequent updates to focus primarily on estimating local residual flow.

Recurrent Update Module. Following the iterative refinement strategy used in optical flow estimation, we employ a recurrent module to progressively refine the optical-to-SAR flow. The initial flow is defined as $\mathbf{f}_{o\to s}^{(0)} = \mathbf{f}_{o\to s}^{\text{init}}$. At iteration t, the module takes the optical feature, the SAR feature warped by the current flow, and the previous hidden state as inputs, and predicts a residual flow update together with the next hidden state:

$$\left(\Delta\mathbf{f}_{o\to s}^{(t)}, \mathbf{h}_{t+1}\right) = \mathcal{R}\left(\mathbf{F}'_o, \text{Warp}\left(\mathbf{F}'_s, \mathbf{f}_{o\to s}^{(t)}\right), \mathbf{h}_t\right). \tag{3}$$

Here, $\mathcal{R}$ is implemented using a modified DPT-style architecture. This update is iterated to refine the flow, while $\mathbf{h}_t$ propagates accumulated matching information across iterations.

3.3 Training Objective

Iterative Flow Prediction Loss. For the recurrent flow predictions, we adopt the Mixture-of-Laplace (MoL) loss from SEA-RAFT Wang et al. (2024). At iteration t, an MoL head maps the hidden state $\mathbf{h}_t$ to a six-channel tensor:

$$\boldsymbol{\eta}_t = \Phi_{\text{MoL}}\left(\mathbf{h}_t\right), \qquad \boldsymbol{\eta}_t \in \mathbb{R}^{h\times w\times 6}, \tag{4}$$

which parameterizes the predicted flow and its MoL uncertainty distribution at the feature resolution. The prediction is then upsampled to the original image resolution using convex upsampling (Wang et al., 2024). For N recurrent iterations, the flow loss is

$$\mathcal{L}_{\text{flow}} = \sum\nolimits_{t=1}^{N} \gamma^{N-t}\mathcal{L}_{\text{MoL}}^{t}, \tag{5}$$

where $\mathcal{L}_{\text{MoL}}^{t}$ is the MoL loss at iteration t and $\gamma < 1$ is the decay factor.

Initial Flow Loss. The GAMI module produces the deterministic initial flow $\mathbf{f}_{o\to s}^{\text{init}}$ without MoL uncertainty parameters. We therefore supervise it with a masked pixel-wise L_1 loss:

$$\mathcal{L}_{\text{init}} = \left[\sum\nolimits_{\mathbf{x}} \mathcal{V}(\mathbf{x})\left|\mathbf{f}_{o\to s}^{\text{init}}(\mathbf{x}) - \mathbf{f}_{o\to s}^{\text{gt}}(\mathbf{x})\right|_1\right] \Big/ \left[\sum\nolimits_{\mathbf{x}} \mathcal{V}(\mathbf{x}) + \epsilon\right]. \tag{6}$$

Here, $\mathbf{f}_{o\to s}^{\text{gt}}$ is the ground-truth flow, $\mathcal{V}$ is the valid-region mask, and ϵ ensures numerical stability.

The overall objective combines the two terms:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{flow}}\mathcal{L}_{\text{flow}} + \lambda_{\text{init}}\mathcal{L}_{\text{init}}, \tag{7}$$

where λ_{flow} and λ_{init} control their relative weights.

4 Experiments

4.1 Benchmark Datasets

Dataset Collection. We combine seven existing optical–SAR datasets—3MOS, OSdataset, OSdataset2.0, QXS-SAROPT, SOPatch, UBCv2, and WHU-OPT-SAR—to construct the benchmark.

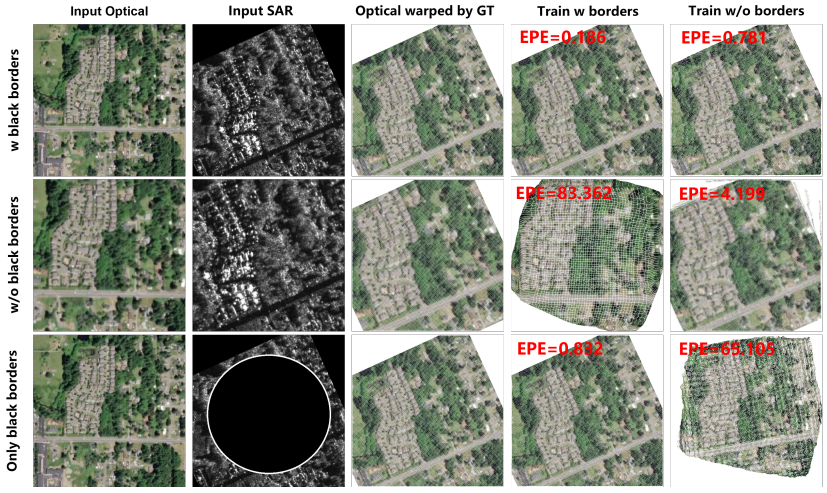

Figure 3: Evidence of black-border geometric shortcut learning in optical–SAR registration under border removal and content-masking conditions.

They provide pixel-level registered image pairs from diverse geographic regions, satellite platforms, land-cover types, and spatial resolutions. Following their original splits, the benchmark contains approximately 130K training pairs and 30K test pairs. We additionally evaluate zero-shot generalization on DFC2020 and YESeg-OPT-SAR, which exhibit different imaging and scene characteristics.

Optical–SAR Data Preparation. Following the shortcut-free procedure in Sec. 3.1, we transform each pair to obtain (I_c^O, I_c^S). Given the affine transformation $\bar{\mathbf{M}}_f$, we analytically compute the ground-truth optical-to-SAR flow $\mathbf{F}_{\text{gt}}$ and its validity mask $\mathcal{V}$. Flow errors are evaluated only on valid pixels. The applied transformations consist of translation, rotation, and scaling with ranges $[-20\%, 20\%]$, $[-45°, 45°]$, and $[0.8, 1.2]$, respectively. All image pairs are resized to 512×512 during training and testing.

4.2 Evaluation Protocols

To ensure a fair comparison among different types of matching methods, we consistently treat the optical image as the source image and the SAR image as the target image, and evaluate all matching results in the source-to-target direction. All statistics are computed only over valid pixels satisfying $\mathcal{V}_k(\mathbf{x}) = 1$, where $\mathcal{V}_k$ denotes the validity mask for the k-th image pair.

Dense Matching Methods. For dense matching models, we directly obtain a dense predicted flow field from the source image to the target image and compare it with the ground-truth flow at all valid pixel locations. The mean endpoint error (EPE) for the k-th image pair is defined as:

$$\text{EPE}_k = \Big[\sum\nolimits_{\mathbf{x}} \mathcal{V}_k(\mathbf{x}) e_k(\mathbf{x})\Big] \Big/ \Big[\sum\nolimits_{\mathbf{x}} \mathcal{V}_k(\mathbf{x})\Big]. \tag{8}$$

The mean EPE over the entire dataset is computed as the average of EPE_k across all image pairs.

Semi-Dense and Sparse Matching Methods. For semi-dense and sparse matching models, we first extract their predicted correspondences and adopt the same RANSAC parameters for all methods to fit an affine transformation matrix $\hat{\mathbf{M}}$. We convert $\hat{\mathbf{M}}$ to its corresponding dense optical flow field, followed by the same procedure used for dense optical flow methods.

Evaluation Metrics. Following previous work Sun et al. (2024); Xiang et al. (2025); Ye et al. (2025); Sun et al. (2025), we primarily report three representative evaluation metrics of (1) Mean endpoint error (EPE): the average endpoint error over all image pairs. (2) The proportion of image pairs whose mean EPE is less than 1, 2, and 5 pixels, denoted as 1px, 2px, and 5px, respectively.(3) RMSE: the root mean square error of endpoint error.

4.3 Diagnosing the Geometric Shortcut

We investigate whether black borders introduced by conventional affine warping serve as predictive cues for registration. Under otherwise identical settings, we train two registration baselines Wang &

Table 1: EPE and RMSE on seven optical–SAR matching datasets under different training settings.

Training	Method	3MOS		Osdataset		QXS_SAROPT		OSDataset2.0	
		EPE ↓	RMSE ↓	EPE ↓	RMSE ↓	EPE ↓	RMSE ↓	EPE ↓	RMSE ↓
No training	XoFTR (2024)	147.194	172.495	146.498	181.597	158.561	186.217	99.531	145.702
	SEA-RAFT-M (2024)	126.117	139.999	177.012	192.345	185.488	199.894	146.930	161.412
	RoMaV2 (2025)	232.060	237.294	206.861	221.056	224.809	229.209	230.494	239.942
	WAFT (2026)	173.685	201.987	178.008	206.910	191.347	198.161	164.544	187.780
Separate training	RoMaV2 (2025)	7.484	15.371	19.868	25.699	24.295	32.086	13.854	18.032
	WAFT (2026)	8.009	26.910	16.115	30.875	25.277	35.879	14.156	30.015
Joint training	GDROS (2025)	5.604	15.644	10.123	16.540	23.160	34.270	4.898	8.354
	RoMaV2 (2025)	7.695	14.963	12.484	15.728	20.101	27.719	6.966	9.291
	WAFT (2026)	5.269	16.976	8.940	**15.189**	20.289	30.531	4.290	9.883
	OSCAR	**3.560**	**11.681**	**8.471**	15.221	**17.327**	**27.419**	**3.710**	**6.649**

Training	Method	SOPatch(SEN1–2)		SOPatch(WHU-SEN)		UBC v2		WHU-OPT-SAR	
		EPE ↓	RMSE ↓	EPE ↓	RMSE ↓	EPE ↓	RMSE ↓	EPE ↓	RMSE ↓
No training	XoFTR (2024)	130.115	169.858	118.560	171.376	164.348	192.024	126.108	171.929
	SEA-RAFT-M (2024)	97.924	103.391	108.555	112.597	179.327	198.965	144.130	152.111
	RoMaV2 (2025)	243.909	248.200	235.799	243.903	225.722	231.225	238.249	242.043
	WAFT (2026)	180.253	205.456	153.125	174.409	163.405	184.781	170.394	192.165
Separate training	RoMaV2 (2025)	15.058	24.665	17.278	21.460	25.228	32.613	6.956	9.461
	WAFT (2026)	21.917	52.934	22.230	49.651	26.035	43.179	5.846	13.753
Joint training	GDROS (2025)	2.638	**10.355**	5.806	14.926	17.570	23.773	3.700	6.449
	RoMaV2 (2025)	4.927	11.975	5.853	8.042	16.400	**22.051**	4.121	5.476
	WAFT (2026)	2.648	11.524	3.899	5.603	14.827	23.634	3.588	7.998
	OSCAR	**2.347**	11.094	**3.675**	**5.233**	**13.691**	22.598	**3.205**	**5.043**

Deng (2026): *Train w/ borders*, using conventionally warped pairs with zero-padded boundaries, and *Train w/o borders*, using our shortcut-free protocol. We evaluate both models under three controlled conditions (Fig. 3): bordered images, border-free images with largely unchanged scene content, and content-masked images that retain the border geometry while removing most valid scene content.

As shown in Fig. 3, the border-trained model achieves accurate registration on bordered inputs but degrades substantially when the borders are removed. The model trained without borders remains substantially more stable on border-free inputs. This contrast indicates that conventional training encourages reliance on padding geometry rather than solely on cross-modal scene correspondence.

Content masking provides further evidence: the border-trained model still produces plausible registration, including flow predictions in regions lacking usable cross-modal appearance information, whereas the model trained without borders fails when scene evidence is removed. These results suggest that border geometry alone enables the conventionally trained model to infer a substantial portion of the synthetic transformation.

Because padding patterns are deterministically coupled with the applied transformation, low registration error on bordered synthetic pairs does not necessarily indicate successful cross-modal correspondence learning and may overestimate registration ability based on scene content. Unless otherwise specified, all subsequent experiments use our shortcut-free protocol.

4.4 Comparison with Existing Methods

We compare OSCAR with representative dense and cross-modal registration methods under the proposed shortcut-free evaluation protocol. Baselines are evaluated in three settings: (1) *No training*: publicly available pretrained models are evaluated without optical–SAR adaptation to assess the modality gap; (2) *Separate training*: RoMaV2 and WAFT are trained independently on each target dataset for dataset-specific adaptation; and (3) *Joint training*: GDROS, RoMaV2, WAFT, and OSCAR are trained on the same integrated optical–SAR collection, with a single set of weights evaluated across all test subsets. Joint training provides the most directly controlled comparison under identical multi-domain supervision.

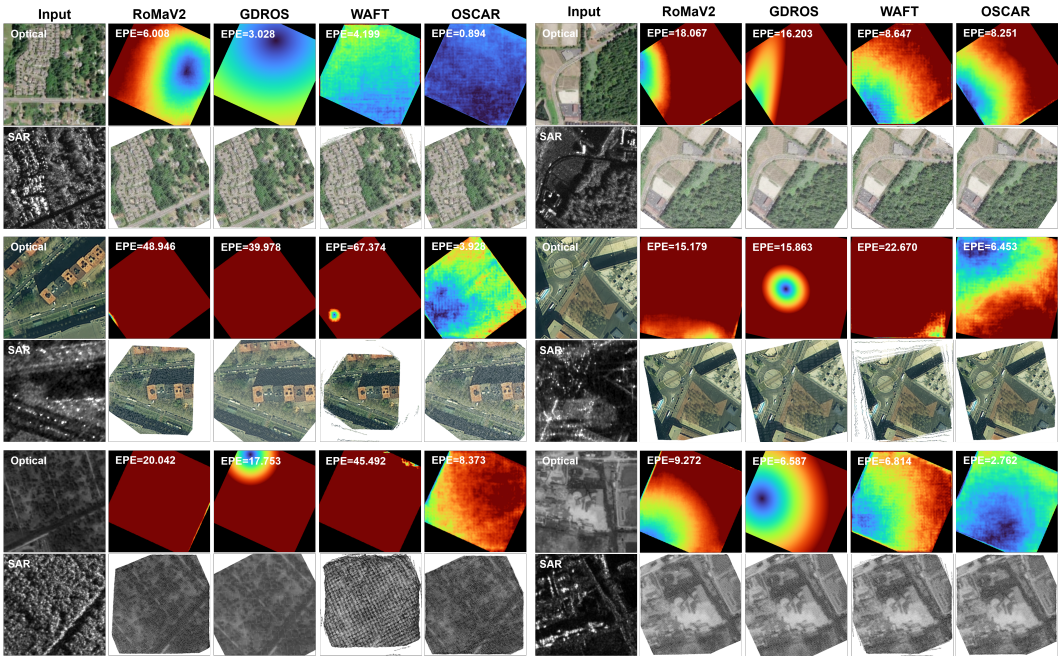

Figure 4: Qualitative Comparison of Optical–SAR Matching Results and Error Maps. In the error heatmaps, pixels with an flow error of 0 are shown in blue, while those with an error of at least 10 are shown in red. The color transition from blue to red indicates increasing error.

Table 1 reports results on eight evaluation subsets from seven datasets. Under joint training, OSCAR achieves the lowest mean EPE on every subset, reducing EPE from 5.269 to 3.560 on 3MOS and from 20.289 to 17.327 on QXS-SAROPT. It also reduces RMSE on seven subsets and performs comparably to WAFT on OSdataset, without dataset-specific fine-tuning at test time. Table 3 further shows that OSCAR achieves the highest 5-pixel accuracy across all eight subsets and improves performance on most subsets at the stricter 1- and 2-pixel thresholds. Jointly trained WAFT remains slightly better on a few strict-threshold measurements, indicating room for improvement in fine localization despite OSCAR's consistent gains in cross-domain robustness.

Table 2: EPE and RMSE metrics on two held-out zero-shot datasets.

Method	DFC2020		YESeg-OPT-SAR	
	EPE ↓	RMSE ↓	EPE ↓	RMSE ↓
GDROS (2025)	6.730	<u>16.331</u>	<u>26.473</u>	39.240
RoMaV2 (2025)	12.892	21.945	31.821	41.474
WAFT (2025)	<u>5.990</u>	**15.754**	26.484	<u>37.514</u>
OSCAR	**5.231**	17.683	**24.079**	**35.624**

Qualitative results in Fig. 4 support the quantitative findings in Table 1. Competing methods exhibit large or spatially concentrated errors on challenging pairs, whereas OSCAR more consistently recovers the dominant displacement and reduces residual misalignment. Together, these results demonstrate improved registration robustness across heterogeneous optical–SAR domains without requiring dataset-specific models.

We further conducted generalization experiments on two held-out DFC2020 and YESeg-OPT-SAR datasets, as shown in Tab. 2, OSCAR achieves the lowest mean EPE on both datasets, providing evidence that the benefit of heterogeneous joint training and the proposed architecture transfers to previously unseen optical-SAR data distributions. Additionally, OSCAR achieves the lowest RMSE on YESeg-OPT-SAR but slightly higher than that of WAFT on DFC2020, despite having lower mean EPE. This indicates that OSCAR improves the average registration accuracy on DFC2020 while still occasional exhibit large-error cases. Improving robustness to such failure cases remains an important direction for future work.Overall, OSCAR achieved the best EPE performance on all 8 subsets and held-out datasets, demonstrating strong cross-dataset generalization capability and suggesting that OSCAR's gains are not restricted to the datasets used for joint training.

Table 3: Percentage of pixels with endpoint error below 1, 2, and 5 pixels on optical–SAR matching datasets. The upper and lower panels report results on two groups of datasets, respectively.

Method	3MOS			Osdataset			QXS_SAROPT			OSDataset2.0		
	1px ↑	2px ↑	5px ↑	1px ↑	2px ↑	5px ↑	1px ↑	2px ↑	5px ↑	1px ↑	2px ↑	5px ↑
GDROS (2025)	15.19%	50.99%	83.01%	0.53%	5.01%	36.85%	0.05%	1.28%	15.23%	3.29%	22.74%	73.06%
RoMaV2 (2025)	2.87%	19.33%	59.25%	0.00%	0.35%	12.85%	0.00%	0.50%	10.08%	0.21%	4.69%	44.79%
WAFT (2026)	39.39%	70.38%	86.84%	**0.88%**	**7.37%**	46.58%	**0.13%**	2.43%	21.85%	10.15%	**42.13%**	82.44%
OSCAR	**43.80%**	**73.41%**	**89.40%**	0.47%	6.84%	**47.64%**	0.10%	**3.35%**	**30.15%**	**10.78%**	41.36%	**83.48%**

Method	SOPatch(SEN1-2)			SOPatch(WHU-SEN)			UBC_v2			whu-opt-sar		
	1px ↑	2px ↑	5px ↑	1px ↑	2px ↑	5px ↑	1px ↑	2px ↑	5px ↑	1px ↑	2px ↑	5px ↑
GDROS (2025)	32.15%	79.22%	96.09%	3.25%	19.13%	64.98%	0.00%	0.48%	10.01%	9.43%	36.29%	75.57%
RoMaV2 (2025)	2.57%	23.47%	77.14%	0.00%	7.22%	53.79%	0.00%	0.21%	7.04%	2.36%	19.57%	72.50%
WAFT (2026)	61.61%	91.20%	96.82%	**13.00%**	**40.79%**	76.90%	**0.41%**	4.83%	26.93%	17.79%	46.71%	76.50%
OSCAR	**62.10%**	**92.05%**	**97.43%**	12.64%	39.71%	**77.98%**	0.28%	**5.46%**	**31.01%**	**20.14%**	**49.64%**	**77.64%**

Table 4: Ablation study of the proposed cross-attention module and GAMI module in terms of EPE.

Experiment	EPE ↓								
	All	3MOS	Osdataset	QXS	OSDataset2.0	SOP-SEN	SOP-WHU	UBC_v2	whu-opt-sar
Baseline (WAFT)	8.758	5.269	8.940	20.289	4.290	2.648	3.899	14.827	3.588
Baseline+OSCA	8.107	4.684	8.718	18.682	3.905	2.458	**3.652**	13.813	3.319
Baseline+Swin-CA	8.733	5.385	8.616	19.739	4.201	2.725	4.160	15.769	3.414
Baseline+OSCA+GAMI	**7.080**	**3.560**	**8.471**	**17.327**	**3.710**	**2.347**	3.675	**13.691**	**3.205**

4.5 Ablation Studies

We ablate the two key architectural components of OSCAR under joint-training protocol. Tab. 4 reports the mean EPE on each evaluation subset and their overall average metrics.

Necessity of Cross-Attention. Adding ViT-like global cross-attention to baseline reduces the average EPE from 8.758 to 8.107 as well as lower EPE on all eight evaluation subsets. In contrast, replacing global attention with window-based Swin-like cross-attention Liu et al. (2021) yields marginal gains or even worse accuracy. This contrast suggests that the performance gain is not simply due to the introduction of an attention module. Instead, long-range patch-wise cross-modal interaction appears particularly beneficial for establishing correspondences under the large spatial displacement encountered in optical-SAR registration.

Effect of Global Affine Motion Initialization (GAMI). Introducing GAMI for recurrent refinement reduces the average EPE from 8.107 to 7.080, yielding a further 12.7% relative improvement. Overall, the complete OSCAR model reduces average EPE by 19.2% relative to the WAFT baseline. These results support the motivation for GAMI: explicitly estimating the dominant global transformation provides a more informative initialization, allowing subsequent iterations to focus on residual correspondence errors.

Together, these ablations demonstrate that the two components are complementary: global cross-attention strengthens long-range cross-modal interaction, while GAMI provides an explicit global geometric prior before iterative refinement.

5 Conclusion

In this work, we identify a geometric shortcut in optical–SAR registration: zero-padded boundaries in conventional affine-warping protocols provide transformation-dependent cues. We introduce a border-free protocol that removes these artifacts while preserving exact dense supervision. Under this protocol, OSCAR combines heterogeneous semantic and local representations, long-range cross-modal interaction, global affine initialization, and recurrent refinement. A single jointly trained model achieves the lowest mean EPE on all seven datasets used for training and both zero-shot evaluation datasets. Ablations confirm the complementary benefits of global cross-modal attention and geometry-aware initialization. These findings underscore the importance of reliable evaluation alongside model design. Extending this framework to naturally misregistered imagery, non-affine deformation, and broader unseen sensor configurations represents an important direction for future work.

AI USE STATEMENT

We used generative AI tools to assist with the retrieval and discovery of related work, as well as to improve the drafting, language, and organization of this manuscript. All literature references, technical claims, experimental procedures, results, and interpretations were reviewed and verified by the authors. The authors reviewed and revised all AI-assisted output and take full responsibility for the final content of this paper.

REFERENCES

Zhuoyu Cai, Dou Quan, Ning Huyan, Pei He, Shuang Wang, and Licheng Jiao. TAR: Text semantic assisted cross-modal image registration framework for optical and sar images. *arXiv preprint arXiv:2605.12064*, 2026.

Johan Edstedt, Qiyu Sun, Georg B"okman, Maarten Wadenb"ack, and Michael Felsberg. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19790–19800, 2024.

Johan Edstedt, David Nordstr"om, Yushan Zhang, Georg B"okman, Jonathan Astermark, Viktor Larsson, Anders Heyden, Fredrik Kahl, Maarten Wadenb"ack, and Michael Felsberg. RoMa v2: Harder better faster denser feature matching. *arXiv preprint arXiv:2511.15706*, 2025.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Xingliang Huang, Kaiqiang Chen, Deke Tang, Chenglong Liu, Libo Ren, Zheng Sun, Ronny H"ansch, Michael Schmitt, Xian Sun, Hai Huang, and Helmut Mayer. Urban building classification (UBC) v2—a benchmark for global building detection and fine-grained classification from satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023. doi: 10.1109/TGRS.2023.3311093.

Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. FlowFormer: A transformer architecture for optical flow. In *Computer Vision – ECCV 2022*, pp. 668–685, 2022.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019. doi: 10.1038/s41467-019-08987-4.

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, 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*, 60:1–14, 2022a. doi: 10.1109/TGRS.2022.3165940.

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*, 106: 102638, 2022b. doi: 10.1016/j.jag.2021.102638.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

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.

Henrique Morimitsu, Xiaobin Zhu, Roberto M. Cesar, Xiangyang Ji, and Xu-Cheng Yin. DPFlow: Adaptive optical flow estimation with a dual-pyramid framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17810–17820, 2025. doi: 10.1109/CVPR52734.2025.01659.

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 (CVPR)*, pp. 23059–23068, 2025.

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 (CVPR)*, pp. 4938–4947, 2020.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8934–8943, 2018.

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 (CVPR)*, pp. 8922–8931, 2021.

Zixuan Sun, Shuaifeng Zhi, Kai Huo, Xuecong Liu, Weidong Jiang, and Yongxiang Liu. Os 3 flow: Optical and sar image registration using symmetry-guided semi-dense optical flow. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 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*, 63:1–15, 2025.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020*, pp. 402–419, 2020.

Onder Tuzcuoglu, Aybora Koksal, Bugra Sofu, Sinan Kalkan, and A. Aydin Alatan. XoFTR: Cross-modal feature matching transformer. *arXiv preprint arXiv:2404.09692*, 2024.

Wei Wang, Dou Quan, Ning Huyan, Chonghua Lv, Shuang Wang, Yunan Li, and Licheng Jiao. Multi-expert learning framework with the state space model for optical and sar image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

Yihan Wang and Jia Deng. WAFT: Warping-alone field transforms for optical flow. In *International Conference on Learning Representations (ICLR)*, 2026.

Yihan Wang, Lahav Lipson, and Jia Deng. SEA-RAFT: Simple, efficient, accurate RAFT for optical flow. In *Computer Vision – ECCV 2024*, 2024.

Yuming Xiang, Jinyang Chen, Zhonghua Hong, Ning Jiao, Feng Wang, Hongjian You, and Xue Tong. OSDataset2.0: Sar-optical image matching dataset and evaluation benchmark. *Journal of Radars*, 14:983–995, 2025. doi: 10.12000/JR25176.

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.

Yuanxin Ye, Jie Shan, Lorenzo Bruzzone, and Li Shen. Robust registration of multimodal remote sensing images based on structural similarity. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2941–2958, 2017. doi: 10.1109/TGRS.2017.2656380.

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*, 57(11):9059–9070, 2019. doi: 10.1109/TGRS.2019.2924684.