

GDROS: A Geometry-Guided Dense Registration Framework for Optical-SAR Images under Large Geometric Transformations

Zixuan Sun, Shuaifeng Zhi*, Ruize Li, Jingyuan Xia, Yongxiang Liu, Weidong Jiang

Abstract—Registration of optical and synthetic aperture radar (SAR) remote sensing images serves as a critical foundation for image fusion and visual navigation tasks. This task is particularly challenging because of their modal discrepancy, primarily manifested as severe nonlinear radiometric differences (NRD), geometric distortions, and noise variations. Under large geometric transformations, existing classical template-based and sparse keypoint-based strategies struggle to achieve reliable registration results for optical-SAR image pairs. To address these limitations, we propose GDROS, a geometry-guided dense registration framework leveraging global cross-modal image interactions. First, we extract cross-modal deep features from optical and SAR images through a CNN-Transformer hybrid feature extraction module, upon which a multi-scale 4D correlation volume is constructed and iteratively refined to establish pixel-wise dense correspondences. Subsequently, we implement a least squares regression (LSR) module to geometrically constrain the predicted dense optical flow field. Such geometry guidance mitigates prediction divergence by directly imposing an estimated affine transformation on the final flow predictions. Extensive experiments have been conducted on three representative datasets WHU-Opt-SAR dataset, OS dataset, and UBCv2 dataset with different spatial resolutions, demonstrating robust performance of our proposed method across different imaging resolutions. Qualitative and quantitative results show that GDROS significantly outperforms current state-of-the-art methods in all metrics. Our source code will be released at: <https://github.com/Zi-Xuan-Sun/GDROS>.

Index Terms—Optical Remote Sensing Images, Synthetic Aperture Radar (SAR), Optical-SAR Image Registration (OSIR), Dense Optical Flow, Least Squares Regression (LSR), Deep Learning

I. INTRODUCTION

REMOTE sensing image registration, which involves aligning images from different sensors, times, and viewing angles, is of utmost importance for improving data representation and enabling seamless multimodal data integration [13, 20]. Nowadays, with the continuous innovation in sensor technology, remote sensing images have made significant progress in both spatial and temporal resolutions. Among these, optical and SAR sensors, as vital data sources for geospatial information, possess distinct while complementary information characteristics [50, 46, 17].

Optical imagery captures the shape, color, and texture of surface objects, providing rich visual cues for object recognition and classification. However, as a passive sensing modality, its utility is restricted by solar illumination and is therefore susceptible to data degradation under cloud cover, dense vegetation, or nighttime conditions [21, 45]. In contrast,

SAR imagery, acquired through active radar pulse transmission/reception, captures the backscattering characteristics of targets, revealing subsurface features that are undetectable by optical sensors. The all-weather/day-night operational capability makes SAR sensors particularly suitable for continuous Earth observation. However, SAR image interpretation remains challenging due to complex scattering mechanisms and inherent speckle noise [33, 31, 22]. The registration of optical and SAR images can effectively enhance geospatial observation capabilities, which has far-reaching implications for tasks such as precision guidance, urban planning, environmental monitoring, and geological surveys [43, 19, 44].

Numerous studies have been carried out on image registration, including classical algorithms such as SIFT [24], SURF [2], and ORB [25], as well as learning-based methods including SuperPoint [6], SuperGlue [26] and LoFTR [30]. These approaches predominantly adopt keypoint extraction, description, and matching frameworks, demonstrating satisfactory performance in homogeneous image registration scenarios. However, the substantial modality gap between optical and SAR imagery severely compromises the stability and reliability of keypoint extraction. To address this issue, modality-robust registration algorithms have been explored for optical and SAR images, including classical methods such as OS-PC [40], RIFT2 [16], and LNIFT [15], as well as learning-based approaches like MU-Net [42], FNet [38], XoFTR [35], and CIRSNet [36]. Though demonstrating enhanced cross-modal registration accuracy, these works still persist in employing keypoint matching strategies and are fundamentally constrained by severe geometric transformations. To address this critical limitation, recent research has shifted toward dense correspondence estimation frameworks, predominantly leveraging optical flow techniques such as OSFlowNet-Ft [45] and OS³Flow [31]. Despite their potential, dense feature-based approaches remain at an early stage of development in optical-SAR image registration (OSIR). Current state-of-the-art methods perform well only within limited range of geometric transformations and have not yet achieved satisfactory performance under large transformations. Motivated by above challenges, we identify and summarize three critical and persistent challenges inherent to dense feature-based registration strategies in OSIR:

Challenges in Large Geometric Transformations. Optical-SAR image pairs with significant geometric transformations exhibit large displacement between corresponding points, which intensifies occlusion effects and leads to mismatches, especially for pixels lacking valid correspondences within image boundaries. Furthermore, extensive geometric transformations including large rotations and scaling variations

All authors are with the College of Electronic Science and Technology, National University of Defense Technology, Changsha, 410073, China.

*Shuaifeng Zhi is the corresponding author: zhishuaifeng@outlook.com.

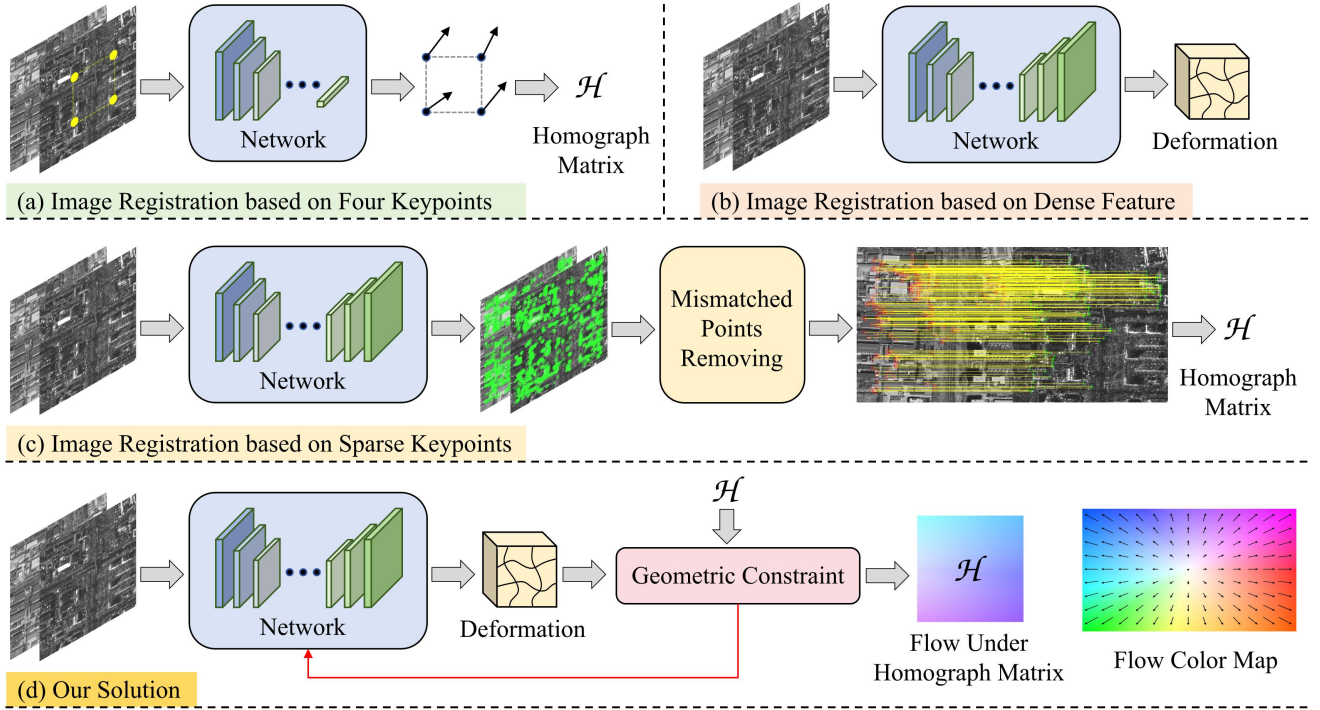


Fig. 1: Overview of learning-based optical-SAR image registration (OSIR) frameworks. (a) Predicting motion offsets of four fixed reference points to solve homography/affine matrix, typically employing an encoder-only network architecture. (b) Describing non-rigid transformations via dense optical flow, typically employing an encoder-decoder network architecture. (c) Predicting sparse(semi-dense) keypoints correspondences, filtering mismatches, and finally estimating a homography/affine matrix via geometric rectification. (d) Our proposed solution GDROS: integrating cross-modal dense optical flow with geometric constraints to achieve geometry-guided dense registration.

further widens the modality gap between optical and SAR, distorting their spatial structural relationships and similarities. To address above issues, one possible solution is to expand the potential spatial search range, which, however, results in a substantial increase in computational complexity, demanding efficient network architectures.

Challenges in Cross-Modal Image Representation. The inherent divergence in optical and SAR imaging mechanisms fundamentally restricts their feature compatibility. For instance, a mountain ridge may exhibit natural undulations in optical imagery but manifest as folded geometries in SAR due to layover effects. Perfect pixel-to-pixel correspondence between optical and SAR images is practically unattainable. Dense OSIR approaches predominantly acquire matching similarity using independently extracted image features. However, such strategy fails to capture invariant cross-modal representations or align their feature distributions. Therefore, it remains essential to design effective learning mechanisms for cross-modal representation.

Challenges in Filtering Outlier Correspondences. In terms of modeling 6-DoF affine transformations, the redundant degrees of freedom inherent in dense optical flow fields inevitably introduce registration errors. Classical outlier filtering approaches, mostly known as RANSAC [9], have greatly improved the robustness of keypoint-based registration, yet are incompatible with dense registration methods. Specifically, the large number of matching candidates and the spatial

smoothness within flow fields impose prohibitive computational costs when revealing potential inliers via randomly sampling. Consequently, an effective outliers filtering strategy tailored for dense correspondences remains as an unresolved challenge.

As an attempt to address aforementioned issues, we propose GDROS, an attention-driven dense OSIR framework to predict precise dense correspondences even under challenging large geometric transformations. In contrast to existing methods based on attention mechanism, such as XoFTR, GDROS discards the self-attention mechanism and focuses solely on cross-modal interaction to avoid excessive smoothing of intra-modal features. Furthermore, by explicitly modeling affine transformations, GDROS effectively leverages geometric constraints to suppress optical flow mismatch points. Compared to the coarse-to-fine strategies commonly employed in conventional approaches, this explicit affine modeling combined with end-to-end training enables better handling of large geometric deformations, thereby improving the accuracy and robustness of optical-SAR image registration. Extensive quantitative and qualitative experimental results show that GDROS strikes the leading optical-SAR registration accuracy without introducing much computational overhead. Our main contributions are summarized as follows:

- We propose an end-to-end flow prediction network for dense OSIR by incorporating explicit prior of affine transformation. Most notably, our method demonstrates superior

performance against leading baselines, especially on optical-SAR images with large geometric transformations.

- We introduce a cross-modal feature fusion module via a dual-level mutual attention mechanism. This design effectively bridges the domain gap between optical and SAR modalities, enhancing their inter-modal similarity and improving alignment precision.

- The proposed differentiable regression module employs a least-squares formulation by estimating 6-DoF affine transformation parameters for refined dense flow fields. It implicitly removes outliers deviating affine transformations and thus significantly enhances registration reliability.

The rest of this paper is organized as follows. Section II introduces related studies on OSIR. Section III introduces the proposed GDROS. Section IV conducts extensive experiments and analysis on the effectiveness of our architecture and its robustness to image resolution. Section V concludes the paper and discusses future work.

II. RELATED WORKS

In this section, we first systematically summarize the evolution and inherent limitations of existing sparse-keypoints-based registration methodologies in Section II-A. Subsequently, we critically analyze prevailing dense-feature-based registration approaches and their performance bottlenecks in cross-modal scenarios in Section II-B. Finally, we rigorously outline geometric prior-based outlier filtering strategies for mitigating mismatches under geometric discrepancies in Section II-C.

A. OSIR based on Sparse Keypoints

Sparse keypoint-based registration methods typically involve keypoint extraction, description, matching, and outlier rejection, as illustrated in Fig. 1(a) and (b). Many existing registration methods largely adhere to this paradigm. However, in the task of OSIR, due to modal differences such as geometric distortions, nonlinear radiometric variations, and speckle noise, keypoint detectors often struggle to extract a sufficient number of robust and reliable interest points between images.

To address this challenge, RIFT [14] employs phase congruency features instead of traditional amplitude- or gradient-based features, enhancing robustness to NRD. RIFT2 [16] significantly improves computational efficiency by replacing the Gabor filter module with the Fast Fourier Transform (FFT). LNIFT [15] further mitigates the modality gap in optical-SAR pairs through its proposed normalization operation. HOWP [47] adopted a feature aggregation strategy to optimize keypoints by separately extracting corner and blob features. MOSS [49] leveraged multidimensional oriented self-similarity features to progressively improve registration performance. The SOFT [48] method enhanced rotational invariance in matching by constructing a novel second-order tensor orientation descriptor. Nevertheless, these traditional sparse keypoint-based optical-SAR registration methods exhibit inherent limitations, as they fail to fully exploit local texture information through the combination of phase, amplitude, and gradient.

In contrast, learning-based methods demonstrate superior performance in OSIR tasks, leveraging their powerful feature extraction capabilities. TS-Net [13] introduces a three-stage framework for sparse image registration between SAR and optical images. It utilizes deep neural networks (DNNs) to encode region selection, correspondence heatmap generation, and outlier removal. MU-Net [42] employs a coarse-to-fine registration pipeline by stacking multiple DNN models. It directly computes affine transformation parameters by learning correspondences for four fixed keypoints, as shown in Fig. 1(a). However, while four keypoints suffice for affine parameter calculation, their inherent instability significantly compromises registration accuracy. LoFTR [30] establishes coarse matches between grids and subsequently refines them using fine features, implementing a coarse-to-fine matching strategy. Building upon LoFTR, XoFTR [35] integrates masked image modeling pre-training, fine-tuning, and image enhancement techniques to address the modality gap. However, the matching process in this method heavily relies on feature similarity based on the attention mechanism, rather than explicit geometric constraints. As a result, it is prone to ambiguous matches in areas with repetitive textures, symmetrical structures, or weak textures. Although it relies on RANSAC post-processing to estimate the geometric model, this post-processing step cannot correct the inherent ambiguity in the underlying feature matches.

In summary, sparse keypoint-based optical-SAR registration techniques have achieved considerable progress, characterized by strong feature robustness and high extraction efficiency. However, in large-scale transformation scenarios, the difficulty of extracting transformation invariance features increases substantially, bringing significant challenges to these methods. Dense feature-based optical-SAR registration techniques provide a viable solution to address these challenges.

B. OSIR based on Dense Feature

Dense optical flow estimation computes motion displacements for all pixels in an image. By establishing pixel-wise dense correspondences, it circumvents the challenges associated with keypoint extraction, demonstrating significant potential in optical-to-SAR image registration tasks.

Since its inception in the 1950s [10], the field of optical flow estimation has witnessed substantial progress. Traditional methods, which rely on the brightness constancy assumption, such as the Lucas-Kanade [23] and DeepFlow [37], are ill-suited for optical-SAR image pairs due to their significant radiometric and structural differences. Deep learning-based approaches have overcome this limitation through powerful feature extraction capabilities. FlowNet[8] served as a pioneering milestone, being the first deep learning model to outperform classical algorithms. Subsequently, network architecture design has emerged as a pivotal factor in enhancing optical flow precision, spurring ongoing research and innovation. PWC-Net [29] introduced an enhanced spatial pyramid network that combines traditional stereo matching, feature extraction, and cost volume mechanisms with deep learning methodologies. RAFT [32] innovatively incorporated a Gated Recurrent Unit

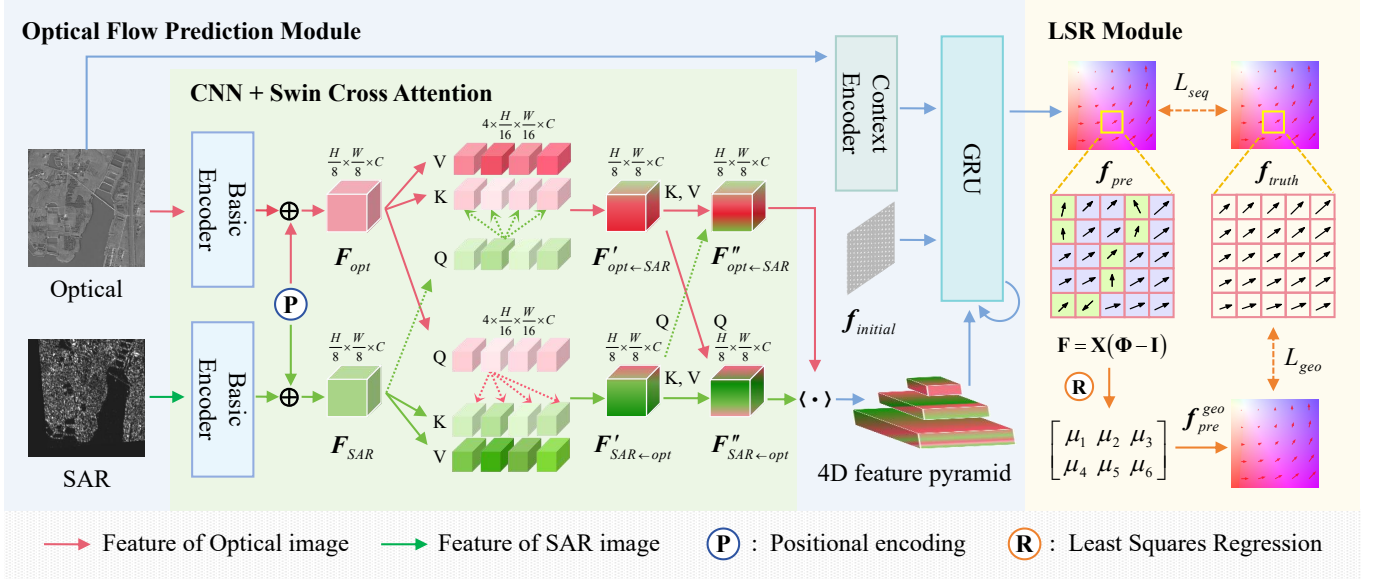


Fig. 2: Framework of our method GDROS. The input optical-SAR image pairs undergo attention mechanism-enabled feature extraction to obtain two distinct deep feature spaces, $F''_{opt \leftarrow SAR}$ and $F''_{SAR \leftarrow opt}$, with enhanced inter-modal information interaction, as depicted in the green-highlighted region. By leveraging these deep feature spaces, we construct a multi-scale 4D feature pyramid that enables GRU-based iterative refinement to generate dense optical-to-SAR flow fields. Subsequently, in the LSR-based geometric consistency enforcement module (yellow-highlighted region), geometric consistency constraints are systematically applied to correct mismatches in the initial flow field, ultimately yielding an accurate radiometric transformation model.

(GRU) module for iterative updates, mimicking the iterative refinement process of conventional optimization methods, and marked another major milestone in optical flow estimation.

Following these advances, and inspired by the success of Transformers in computer vision, recent studies have begun leveraging the global modeling capacity of Transformers to tackle large-displacement optical flow estimation. Transformer-based models such as GMFlow [41] focus on global feature similarity by replacing GRU modules with stacked Transformer blocks, achieving performance superior to RAFT. FlowFormer [12] further improves registration accuracy by utilizing self-attention mechanisms to effectively capture long-range dependencies and spatial relationships among pixels. FlowFormer++ [28] proposed a masked cost volume auto-encoding scheme to pre-train the cost volume encoder more efficiently. MemFlow [7] draws on the attention mechanism of Transformers to achieve effective aggregation of historical information, significantly enhancing estimation accuracy and generalization while maintaining real-time performance.

Currently, OSIR methods based on sparse and dense features have achieved significant development, with their continuously increasing performance. However, most of them focus on extracting features from the amplitude and scattering properties of optical and SAR images, lacking geometric constraints from prior knowledge and physical properties, which leads to instability and insufficient robustness in practical scenarios.

C. OSIR with Prior Knowledge Constraint

Outlier filtering, commonly referred to as mismatch removal or correspondence selection, serves to identify geometrically consistent correspondences (inliers) while rejecting spurious matches (outliers) within candidate sets. Within the context of optical-SAR image registration tasks, the fundamental geometric model is typically formulated as a global affine transformation. Under such constraints, outlier rejection predominantly relies on geometric consistency criteria for establishing robust feature correspondences. The most classical and widely adopted methodology in this domain remains the RANSAC (Random Sample Consensus) [9] algorithm, which has undergone extensive investigation for decades [5]. As an iterative hypothesis-testing framework, RANSAC repeatedly samples minimal subsets of correspondences to hypothesize provisional parameter models, subsequently evaluating model quality through the computation of inlier coherence counts. The algorithm's performance exhibits sensitivity to critical parameters including iteration count and threshold settings. To address these limitations, numerous RANSAC variants have emerged, each introducing innovative sampling strategies and reliability metrics. Notable advancements include MLESAC (Maximum Likelihood Estimation Sample Consensus) [34] that incorporates probabilistic correspondence weights, PROSAC (Progressive Sample Consensus) [4] that utilizes spatial coherence for guided sampling, and MAGSAC (Motion-Aware Generalized Sample Consensus) [1] that integrates motion estimation priors.

These global consensus-based methodologies upon, while theoretically guaranteeing geometric model correctness, are

often associated with significant computational overhead that conflicts with real-time image processing requirements. Additional alternative frameworks including Neighborhood Consensus Methods, Descriptor-Based Approaches, and Graph-Based Algorithms [27]. However, these technologies exhibit inherent sensitivities to local features and possess fundamental limitations in addressing heterogeneous image registration tasks involving large geometric deformations, and are therefore excluded from further discussion in this paper.

III. METHODOLOGY

In this section, we present the architectural details of our proposed method GDROS. The overall pipeline is shown in Fig. 2 and comprises two primary components: a dense optical flow prediction module and a Least Squares Regression (LSR) module. In Section III-A, we first introduce the dense flow prediction process, followed by a particular emphasis on the proposed cross-modal feature extraction module bridging the domain gap between heterogeneous optical-SAR modalities. Subsequently, Section III-B presents the LSR module for refined flow predictions under affine transformation constraints, as well as its outlier rejection mechanism. Finally, Section III-C describes the complete network training strategy.

A. Dense Optical Flow Prediction

Large geometric transformations exacerbate the modality discrepancies between optical-SAR image pairs, making the extraction of reliable and stable keypoints particularly challenging. In contrast, our method opts to predict a per-pixel dense displacement flow field $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, which establishes pixel-wise correspondence between source I_s and target I_t images. Specifically, for a pixel located at $(x_1, y_1) \in \mathbb{R}^2$ in I_s , the corresponding position $(x_2, y_2) \in \mathbb{R}^2$ in I_t satisfies: $x_2 = x_1 + f_x(x_1, y_1)$, $y_2 = y_1 + f_y(x_1, y_1)$.

Network Backbone. Our dense optical flow prediction backbone, as shown in blue-highlighted region of Figure 2, inherits the RAFT [32] meta-architecture comprising three core components: feature extraction module, 4D volumetric space construction, and a GRU-based iterative refinement module. It is noteworthy that the context encoder aims to provide essential contextual information for the GRU blocks, and we choose optical images as input given their richer textural and structural details compared to SAR images. The GRU update block incorporates the conditioned iterative optimization paradigm from classical approaches, generating a sequence of optical flow $\{f^1, \dots, f^N\}$ starting from an initial flow field estimation $f_0 = \mathbf{0}$. Through this process, each estimate undergoes progressive refinement via iterative MSE minimization against the ground-truth optical flow field.

Feature Extraction. In optical-SAR image registration tasks with significant geometric transformations, a fundamental challenge lies in bridging the modality gap between optical and SAR images while effectively extracting their shared latent structural features. To address this, we propose a hybrid feature extractor (highlighted in green in Fig. 2) that synergistically integrates convolutional neural networks (CNNs) with Transformer architectures. This configuration preserves

intrinsic fine-grained spatial information through CNNs' local receptive fields while enabling long-range cross-modal information exchange via Transformers' attention mechanisms.

Specifically, we employ a weight-sharing ResNet architecture pre-trained on ImageNet as the base encoder to extract domain-specific features $\mathbf{F}_{\text{opt}}$ and $\mathbf{F}_{\text{sar}}$ from optical and SAR images respectively. Consistent with RAFT, we perform $8\times$ spatial downsampling during feature extraction to maintain computational tractability. However, the CNN-derived features operate in isolation, insufficient to overcome the fundamental modality disparity between heterogeneous image domains. To resolve this limitation, we innovatively design a Cross-Attention-Only Transformer module that completely eliminates self-attention operations, structured as follows:

Positional Encoding. We embed fixed 2D sinusoidal positional embeddings into CNN-extracted features $\mathbf{F}_{\text{opt}}$ and $\mathbf{F}_{\text{sar}}$, endowing the system with explicit spatial awareness, following standard practice in DETR [3]:

$$\mathbf{F}_{\text{opt}}^{\text{pos}} = \text{Pos}(\mathbf{F}_{\text{opt}}), \quad \mathbf{F}_{\text{sar}}^{\text{pos}} = \text{Pos}(\mathbf{F}_{\text{sar}}), \quad (1)$$

where $\text{Pos}(\cdot)$ denotes the positional encoding operation. We have found this design effectively enhances feature similarity under large-scale geometric variations while resolving ambiguities induced by significant deformations, as quantitatively demonstrated in Table IV of the ablation study section.

Cross-Attention-Only Interaction. The positionally encoded features subsequently undergo cross-attention operations where queries originate from one modality while keys and values derive from the other, formulated as:

$$\begin{aligned} \mathbf{Q}_x &= \mathbf{F}_x^{\text{pos}} \cdot \mathbf{W}_q, \quad \mathbf{K}_x = \mathbf{F}_x^{\text{pos}} \cdot \mathbf{W}_k, \quad \mathbf{V}_x = \mathbf{F}_x^{\text{pos}} \cdot \mathbf{W}_v, \\ \mathbf{F}'_{\text{opt} \leftarrow \text{sar}} &= \text{CrossAttn}(\mathbf{Q}_{\text{sar}}, \mathbf{K}_{\text{opt}}, \mathbf{V}_{\text{opt}}), \\ \mathbf{F}'_{\text{sar} \leftarrow \text{opt}} &= \text{CrossAttn}(\mathbf{Q}_{\text{opt}}, \mathbf{K}_{\text{sar}}, \mathbf{V}_{\text{sar}}), \end{aligned} \quad (2)$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ denote learnable weight matrices, $x \in \{\text{opt}, \text{sar}\}$ specifies the modality type, and $\mathbf{F}'_{\text{opt} \leftarrow \text{sar}}/\mathbf{F}'_{\text{sar} \leftarrow \text{opt}}$ represent the refined optical/SAR deep features after cross-modal interaction. This hierarchical process selectively aggregates knowledge from potential matching candidates in another image by measuring cross-view feature similarity, achieving selective inter-modal information aggregation. This process generates modality-interacted independent features $\mathbf{F}'_{\text{opt} \leftarrow \text{sar}}$ and $\mathbf{F}'_{\text{sar} \leftarrow \text{opt}}$. To further aggregate cross-modal latent information, we recursively apply the cross-attention mechanism:

$$\begin{aligned} \mathbf{F}''_{\text{opt} \leftarrow \text{sar}} &= \text{CrossAttn}(\mathbf{Q}'_{\text{sar}}, \mathbf{K}'_{\text{opt}}, \mathbf{V}'_{\text{opt}}) \\ \mathbf{F}''_{\text{sar} \leftarrow \text{opt}} &= \text{CrossAttn}(\mathbf{Q}'_{\text{opt}}, \mathbf{K}'_{\text{sar}}, \mathbf{V}'_{\text{sar}}) \end{aligned} \quad (3)$$

These doubly refined features $\mathbf{F}''_{\text{opt} \leftarrow \text{sar}}$ and $\mathbf{F}''_{\text{sar} \leftarrow \text{opt}}$ serve as inputs for subsequent optical flow prediction. Notably, to mitigate the computational complexity inherent in pairwise attention operations, we adopt a shifted local window attention strategy consistent with GMFlow [41], where the number of windows is fixed at 4. The proposed two-stage cross-attention-only architecture demonstrates superior efficacy over conventional self- and cross-attention frameworks, as cross-modal information interaction plays a more pivotal role than

single-modality feature depth in heterogeneous image registration tasks, which is validated by ablation experiments in Section IV-D.

B. LSR Module

Optical flow fields inherently possess multiple degrees of freedom and can naturally simulate non-rigid deformations. However, due to the unique imaging characteristics of SAR, strict pixel-wise alignment between SAR and optical images is fundamentally unattainable. Precise extraction of non-rigid transformations may amplify localized errors. For instance, building structures in optical imagery may exhibit layover distortion (top-bottom inversion) in SAR images. Precise local non-rigid registration in such areas risks introducing geometric misalignments, which may compromise overall registration accuracy. Consequently, constraining the mathematical model to an affine transformation, rather than pursuing non-rigid deformation, better captures the global registration relationship between optical and SAR images. The affine transformation matrix Φ , encompassing translation, scaling, and rotation, is mathematically expressed as:

$$\Phi = \begin{bmatrix} \mu_1 & \mu_2 & \mu_3 \\ \mu_4 & \mu_5 & \mu_6 \end{bmatrix} = \begin{bmatrix} S_x * \cos(\theta) & -S_x * \sin(\theta) & T_x \\ S_y * \sin(\theta) & S_y * \cos(\theta) & T_y \end{bmatrix}, \quad (4)$$

where S_x, S_y, T_x, T_y denote the scaling factors and translational offsets along the x and y axes, respectively; θ represents the rotation angle.

Compared to the six degrees of freedom in affine transformations, the redundant degrees of freedom inherent to optical flow fields inevitably introduce additional registration errors. Mathematically, three corresponding points suffice to uniquely solve the six parameters of an affine transformation matrix. The mismatch filtering strategies widely employed in keypoint matching, such as RANSAC, operate through random sampling of triple-point subsets to iteratively estimate optimal models. However, this approach fails to leverage the inherent density and smoothness characteristics of optical flow fields. To address this limitation, we innovatively propose the LSR network module, as illustrated in the yellow highlighted region of Fig. 2. The LSR module adaptively regresses affine transformation parameters by exploiting dense correspondences rather than sparse subsets, which enhances robustness without requiring laborious parameter tuning procedures.

For any pixel position $[x_o, y_o]$ in the image, the coordinate displacement vector under an affine transformation can be computed via the affine transformation matrix Φ . Specifically, each unique Φ uniquely defines a distinct optical flow field, mathematically expressed as:

$$\mathbf{F}(x, y) = \begin{bmatrix} flow_x \\ flow_y \end{bmatrix} = \begin{bmatrix} \mu_1 - 1 & \mu_2 & \mu_3 \\ \mu_4 & \mu_5 - 1 & \mu_6 \end{bmatrix} \begin{bmatrix} x_o \\ y_o \\ 1 \end{bmatrix}, \quad (5)$$

where $flow_x, flow_y$ denote the vector magnitudes of the optical flow field along the x - and y -directions at coordinate $[x_o, y_o]$, respectively. When extending this transformation to all pixels in the image $I \in \mathbb{R}^{H \times W}$, the expression can be generalized as:

$$\mathbf{F} = \mathbf{X}(\Phi - \mathbf{I}), \quad (6)$$

where $\mathbf{F} \in \mathbb{R}^{N \times 2}$ denotes the optical flow field, $\mathbf{X} \in \mathbb{R}^{N \times 3}$ represents the set of original pixel coordinates in the image, $\Phi \in \mathbb{R}^{2 \times 3}$ corresponds to the 6-parameter affine transformation matrix (comprising translation, scaling, and rotation), $\mathbf{I} \in \mathbb{R}^{2 \times 3}$ is the identity matrix, and $N = H \times W$ denotes the total number of pixels. We formulate the parameter estimation as a least squares problem. By minimizing the residual sum of squares:

$$\mathcal{L}(\Phi) = \|\mathbf{F} - \mathbf{X}(\Phi - \mathbf{I})\|_2^2, \quad (7)$$

where the notation $\|\cdot\|_2$ denotes the Euclidean norm. The optimal solution is obtained through solving the normal equations:

$$\mathbf{X}^T \mathbf{F} = \mathbf{X}^T \mathbf{X}(\Phi - \mathbf{I}), \quad (8)$$

which yields the closed-form solution:

$$\Phi = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{F} + \mathbf{I}, \quad (9)$$

where Φ denotes the final network output, representing the affine transformation matrix optimized from the predicted optical flow.

C. Training Configurations

Our training objective combines two complementary loss terms: an aggregated sequence loss and a geometric constraint loss. The aggregated sequence loss supervises the iterative flow refinement process of the GRU module by progressively weighting the flow estimates across iterations:

$$L_{\text{seq}} = \sum_{i=1}^N \omega^{N-i} \|f_{\text{os}}^i - f_{\text{os}}^{\text{gt}}\|_1, \quad (10)$$

where f_{os}^i denotes the estimated optical flow at i -th iteration, $f_{\text{os}}^{\text{gt}}$ represents the ground-truth flow from the optical image to the SAR image, and ω controls the temporal weighting decay. The exponentially decaying weights emphasize later iterations while maintaining gradient flow to earlier predictions, forming a coarse-to-fine optimization process.

The GRU-generated flow sequence $\{f^1, \dots, f^N\}$ is geometrically regularized through proposed LSR module, producing corresponding affine transformations $\{\Phi^1, \dots, \Phi^N\}$. We compute geometrically constrained flows $\{f_{\text{lsr}}^1, \dots, f_{\text{lsr}}^N\}$ via Eq. 5, and evaluate the geometric constraint loss as:

$$L_{\text{geo}} = \sum_{i=1}^N \omega^{N-i} \|f_{\text{lsr}}^i - f^{\text{gt}}\|_1. \quad (11)$$

The geometric constraint loss L_{geo} imposes an affine transformation constraint on the predicted optical flow, which encourages to filter diverging mismatched points. The final training loss is a linear combination of the two loss terms:

$$L_{\text{total}} = \lambda_{\text{seq}} \cdot L_{\text{seq}} + \lambda_{\text{geo}} \cdot L_{\text{geo}}. \quad (12)$$

In our experiments, the best overall registration accuracy and stability were achieved when $\lambda_{\text{seq}} = 0.5$ and $\lambda_{\text{geo}} = 0.5$. Setting $\lambda_{\text{seq}} = 1$ and $\lambda_{\text{geo}} = 0$ resulted in a slight degradation in performance, though the network remained relatively stable. Conversely, when $\lambda_{\text{seq}} = 0$ and $\lambda_{\text{geo}} = 1$, performance decreased significantly. Under the configuration with $\lambda_{\text{seq}} = 0.5$

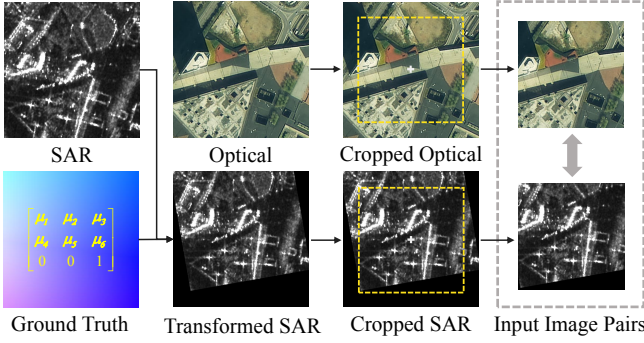


Fig. 3: Example of network input image pairs generation.

and $\lambda_{\text{geo}} = 0.5$, the geometric loss helps refine the output structure and improves model performance by incorporating additional geometric constraints, building upon the foundation provided by the sequential loss. Additionally, the weighting coefficient ω was empirically set to 0.85, and the number of GRU iterations N was set to 12 to achieve a smooth convergence with moderate computational costs.

IV. EXPERIMENTS

A. Benchmarks and Data Preparation

Benchmark Datasets: Our method was comprehensively validated on three publicly available datasets with different spatial resolutions: *WHU-OPT-SAR dataset* [18] (5-meter resolution), *OS dataset* [39] (1-meter resolution), and *UBCv2 dataset* [11] (0.5-meter resolution). The WHU-OPT-SAR dataset contains 100 large-scale optical-SAR image pairs with different terrains, which are further segmented into 512×512 sub-images for the sake of efficiency: training (5,600 pairs), validation (700 pairs), and testing (700 pairs). The OS dataset spans multiple geographic regions and contains 2,673 aligned optical-SAR images, divided into splits of training (2,011 pairs), validation (238 pairs), and testing (424 pairs). The UBCv2 dataset, initially utilized for building detection and classification, contains 7,170 pairs of high-resolution optical and SAR satellite images. The high-resolution nature of UBCv2 data amplifies the texture differences of optical-SAR images and also results in much smaller field of view (FOV), serving as a representative benchmark for latest imaging resource. However, we notice that there are incomplete, cloud-occluded, or textureless image pairs within the UBCv2 dataset, which are unsuitable for registration tasks. Thus, we remove such data by a pre-screening step, and finally reach a split of training (3,517 pairs), validation (1,437 pairs), and test (1,447 pairs).

Generation of Optical-SAR Image Pairs: The data preparation pipeline for optical-SAR image pairs used in network training is illustrated in Fig. 3. For each precisely registered image pair, we apply random affine transformations to the SAR image within specified parameter ranges to generate transformed SAR images, and compute its corresponding optical flow field as ground-truth supervision. Our experimental setup employs the following transformation bounds: The translation parameters were limited within the range of $[-30, 30]$ with a precision of 1 pixel, the scaling parameter was limited within

the range of $[0.8, 1.2]$ with a precision of 0.05, and the rotation parameters were limited within the range of $[-20^\circ, 20^\circ]$ with a precision of 1° . We believe such a transformation range presents significant technical challenges to optical-SAR registration, as most leading approaches are limited to translation-only or small-scale rotational/scaling transformations.

Notably, our experiments reveal a positive correlation between the richness of shared structural information in image pairs and the registration accuracy. Although the original full-size input enriches contextual information, it substantially increases the computational load for flow predictions. To further preserve shared image content and avoid interference from invalid black-border artifacts during applied transformations, we center-crop the 512×512 pixel input to a size of 400×400 , as illustrated in Fig. 3.

B. Metrics and Experimental Settings

1) *Evaluation Metrics:* To comprehensively assess registration performance, we employ three popular metrics and additionally propose a novel metric to evaluate the overall registration accuracy across multiple error tolerance levels:

Average Endpoint Error (AEPE) computes the mean Endpoint Error (EPE) across all image pairs in the test set, where for each image pair the EPE computes averaged pixel-wise Euclidean distance l_2 between predicted keypoints and their ground-truth correspondences:

$$\text{EPE}(k) = \frac{1}{M} \sum_{(x,y) \in I_{opt}^k} l_2(x,y), \quad (13)$$

$$l_2 = \sqrt{(f_u^{\text{pre}} - f_u^{\text{gt}})^2 + (f_v^{\text{pre}} - f_v^{\text{gt}})^2},$$

where $f_u^{\text{pre}}, f_v^{\text{pre}}$ and $f_u^{\text{gt}}, f_v^{\text{gt}}$ denote the predicted and ground-truth optical flow vectors at pixel position (x,y) , respectively. M is the total number of valid pixels within the reference optical image I_{opt} . As $\text{EPE}(k)$ is computed for the k -th image pair, we could further derive AEPE using all per-pair EPE value of the test data:

$$\text{AEPE} = \frac{1}{N} \sum_{k=1}^N \text{EPE}(k), \quad (14)$$

where N is the total number of image pairs in test set.

Root Mean Square Error (RMSE) measures the global variance of displacement magnitude across all image pairs, thereby evaluating the dispersion of registration accuracy:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{k=1}^N (\text{EPE}(k) - \overline{\text{EPE}})^2}, \quad (15)$$

where $\text{EPE}(k)$ is the endpoint error of the k -th image pair described above.

Correct Match Rate@ τ (CMR@ τ) quantifies the proportion of correctly matched image pairs under predefined precision thresholds:

$$\text{CMR@}\tau = \frac{N_\tau}{N_{\text{total}}} \times 100\%, \quad (16)$$

where τ denotes the precision threshold. $N_\tau = N_{\{k | \text{EPE}(k) < \tau\}}$ represents the number of image pairs

TABLE I: Comparative results of different methods on three test sets of the **WHU-OPT-SAR dataset (5-meter resolution)** in ‘mean $\pm$ std’ format. **Bold** indicates the best result, and underline indicates the second best result.

Category	Method	$\tau \leq 1px$		$\tau \leq 2px$		$\tau \leq 5px$		All	
		CMR@ $\tau\uparrow$	AEPE@ $\tau\downarrow$	CMR@ $\tau\uparrow$	AEPE@ $\tau\downarrow$	CMR@ $\tau\uparrow$	AEPE@ $\tau\downarrow$	AEPE $\downarrow$	RMSE $\downarrow$
Sparse	RIFT2 [14, 16]	0.14(± 0.12)%	0.58(± 0.41)	1.53(± 0.24)%	1.49(± 0.05)	8.52(± 0.44)%	3.04(± 0.05)	192.58(± 1.93)	15524.13(± 374.22)
	LNIFT [15]	0.14(± 0.00)%	0.84(± 0.07)	1.33(± 0.47)%	1.45(± 0.06)	10.00(± 0.35)%	3.19(± 0.10)	159.10(± 3.76)	11479.10(± 95.99)
	XoFTR [35]	5.27(± 1.01)%	0.79(± 0.01)	24.38(± 0.96)%	1.30(± 0.02)	30.91(± 1.11)%	1.53(± 0.03)	53.00(± 1.58)	2874.67(± 49.59)
Dense	FlowFormer [12]	14.29(± 1.52)%	0.83(± 0.01)	64.81(± 0.99)%	1.31(± 0.01)	91.81(± 0.41)%	1.75(± 0.03)	2.98(± 0.20)	28.35(± 7.88)
	FlowFormer++ [28]	16.86(± 0.71)%	0.81(± 0.01)	66.33(± 0.75)%	1.28(± 0.02)	90.00(± 0.84)%	1.67(± 0.02)	3.22(± 0.18)	37.00(± 5.50)
	GMFlow [41]	0.53(± 0.24)%	0.93(± 0.06)	13.24(± 0.94)%	1.59(± 0.01)	66.48(± 1.48)%	3.01(± 0.00)	6.40(± 0.54)	249.73(± 120.83)
	RAFT [32]	13.91(± 0.44)%	0.81(± 0.01)	61.86(± 0.20)%	1.31(± 0.01)	98.33(± 0.18)%	1.88(± 0.02)	2.04(± 0.07)	4.25(± 3.22)
	OS ³ Flow [31]	21.14(± 0.12)%	0.72(± 0.01)	57.19(± 1.98)%	1.19(± 0.00)	93.38(± 0.37)%	1.89(± 0.04)	2.35(± 0.12)	6.65(± 3.09)
	Ours	72.05(± 1.06)%	<u>0.60(± 0.01)</u>	96.86(± 0.65)%	0.78(± 0.01)	99.57(± 0.11)%	0.83(± 0.01)	0.90(± 0.04)	0.62(± 0.15)

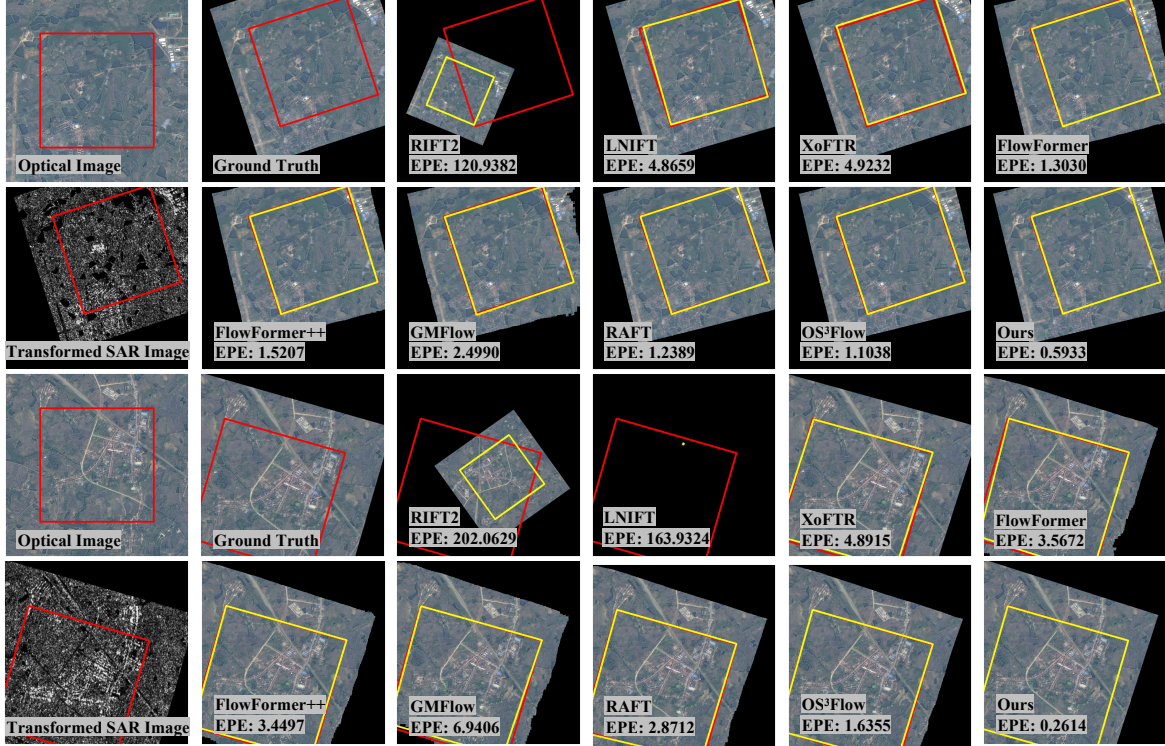


Fig. 4: Registration results on the WHU-OPT-SAR dataset. The yellow line represents the ground truth registration result, and the red line represents the experimental registration result.

satisfying $EPE < \tau$ in the test set. Following the standard practice in classical registration benchmarks [14, 33], we employ a multi-threshold strategy to assess registration performance at different precision levels: for coarse-level matching evaluation, we adopt thresholds of $\tau = 3px$ and $\tau = 5px$, while for fine-level accuracy assessment, we utilize more strict thresholds of $\tau = 1px$ and $\tau = 2px$.

Average Endpoint Error@ τ (AEPE@ τ) is newly proposed to evaluate the overall registration accuracy under varying tolerance thresholds, complementing the limitation of CMR@ τ which focuses solely on the image quantity within threshold τ instead of their absolute registration precision. Thus the AEPE@ τ is defined as:

$$AEPE@ \tau = \frac{1}{N_{\tau}} \sum_{k \in \mathcal{M}_{\tau}} EPE(k), \quad (17)$$

where $\mathcal{M}_{\tau} = \{k \mid EPE(k) < \tau\}$. This allows us to characterize how matching accuracy evolves with precision thresholds, revealing further performance details and resilience under

varying CMR@ τ value. It should be noted that since AEPE@ τ is threshold-dependent and influenced by the number of image pairs meeting the condition, it should not be interpreted in isolation. A meaningful evaluation of the registration performance for the subset of correctly matched samples can only be achieved by jointly analyzing AEPE@ τ with the CMR.

2) *Experimental Settings*: Our experiments utilized the AdamW optimizer for network training with an initial learning rate of $1.2e-5$, a batch size of 12, and a maximum iteration of 120,000 steps. The GRU module underwent 12 iterations during training and 32 iterations during testing. When processing image pairs of size 512×512 pixels, the training process requires 16,672 MB of GPU memory. All experiments were implemented in PyTorch using a single NVIDIA GeForce RTX 4090 GPU and an Intel Core i9-14900k 24-core CPU.

All baseline methods are implemented by retraining their original pre-trained models using publicly released codebases with default training configurations. Notably, our cropping

TABLE II: Comparative results of different methods on three test sets of the **OS dataset (1-meter resolution)** in ‘mean $\pm$ std’ format. **Bold** indicates the best result, and underline indicates the second best result.

Category	Method	$\tau \leq 1\text{px}$		$\tau \leq 2\text{px}$		$\tau \leq 5\text{px}$		All	
		CMR@ $\tau\uparrow$	AEPE@ $\tau\downarrow$	CMR@ $\tau\uparrow$	AEPE@ $\tau\downarrow$	CMR@ $\tau\uparrow$	AEPE@ $\tau\downarrow$	AEPE $\downarrow$	RMSE $\downarrow$
Sparse	RIFT2 [14, 16]	0.55(± 0.29)%	0.92(± 0.06)	4.40(± 0.40)%	1.48(± 0.14)	28.69(± 1.22)%	3.13(± 0.03)	72.24(± 8.13)	12540.89(± 517.13)
	LNIFT [15]	0.32(± 0.29)%	0.49(± 0.38)	3.85(± 1.28)%	1.52(± 0.05)	32.55(± 0.33)%	3.29(± 0.05)	95.79(± 0.84)	12777.92(± 282.86)
	XoFTR [35]	4.83(± 1.02)%	0.81(± 0.00)	27.16(± 0.59)%	1.36(± 0.05)	56.23(± 1.33)%	2.22(± 0.06)	25.00(± 1.66)	2173.35(± 316.47)
Dense	FlowFormer [12]	0.86(± 0.11)%	0.91(± 0.02)	21.39(± 1.54)%	1.57(± 0.02)	60.90(± 0.78)%	2.61(± 0.06)	6.87(± 0.21)	103.70(± 14.15)
	FlowFormer++ [28]	0.16(± 0.11)%	<u>0.64(± 0.45)</u>	17.37(± 1.11)%	1.62(± 0.01)	62.26(± 0.77)%	2.76(± 0.06)	7.76(± 0.26)	112.10(± 9.78)
	GMFlow [41]	0.00(± 0.00)%	—	4.40(± 0.87)%	1.65(± 0.06)	49.77(± 1.02)%	3.36(± 0.05)	6.17(± 0.22)	22.42(± 7.30)
	RAFT [32]	2.20(± 0.22)%	0.85(± 0.03)	32.86(± 1.24)%	1.54(± 0.02)	90.49(± 0.11)%	2.51(± 0.02)	2.95(± 0.07)	5.44(± 3.94)
	OS ³ Flow [31]	5.15(± 1.20)%	0.72(± 0.01)	31.70(± 3.94)%	1.42(± 0.04)	86.29(± 1.42)%	2.50(± 0.11)	3.17(± 0.18)	7.07(± 4.60)
	Ours	33.88(± 0.73)%	0.73(± 0.00)	80.19(± 1.77)%	1.13(± 0.00)	99.45(± 0.11)%	1.46(± 0.02)	1.48(± 0.02)	0.75(± 0.02)

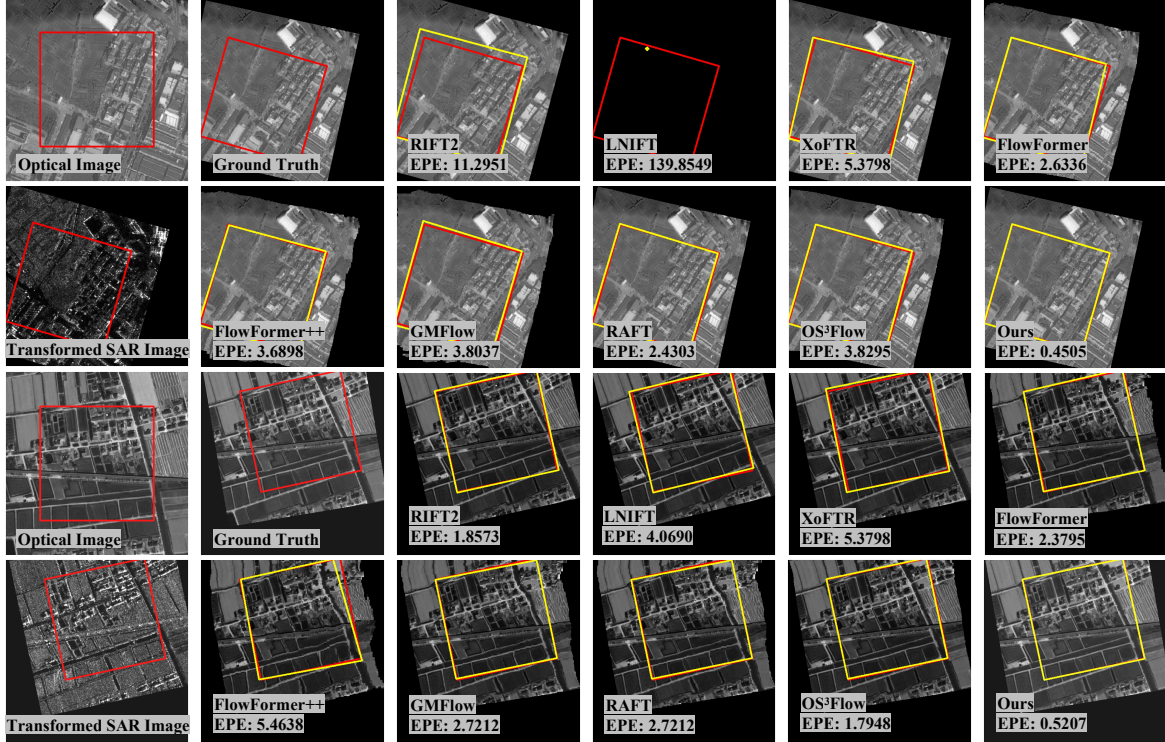


Fig. 5: Registration results on the OS dataset. The yellow line represents the ground truth registration result, and the red line represents the experimental registration result.

strategy, as illustrated in Fig. 3, was found to benefit both registration accuracy and robustness when applied to RAFT [32] and OS³Flow [31]. Therefore, we incorporate this strategy into the implementations of RAFT and OS³Flow under identical experimental conditions to ensure fair comparisons.

Furthermore, to eliminate potential bias from the test dataset, we generate three independent test sets within pre-defined affine transformation ranges using different random seeds. All quantitative experimental results are reported as ‘mean $\pm$ standard deviation’ across all three test sets.

C. Comparisons to Baseline Methods

To validate the effectiveness of our method, we compare it against seven SOTA baseline methods, including three keypoint-based registration methods (RIFT2 [14, 16], LNIFT [15], XoFTR [35]) and five dense-based registration methods (FlowFormer [12], FlowFormer++ [28], GMFlow [41], RAFT [32], OS³Flow [31]). Qualitative and quantitative analyses

were conducted on three distinct datasets to evaluate the generalization capabilities of these methods under varying spatial resolutions.

1) *Results on the WHU-OPT-SAR dataset:* Qualitative comparisons on the WHU-OPT-SAR dataset (5m spatial resolution) are illustrated in Fig. 4, where red bounding boxes denote ground-truth correspondences formed by connecting four reference points, and yellow boxes represent predictions from evaluated methods. Our method demonstrates exceptional alignment accuracy, with predicted yellow boxes nearly overlapping the red ground-truth boxes across diverse terrain scenarios.

Quantitative results in Table I further validate our approach’s superiority. Our method achieves **sub-pixel-level registration precision** with an overall AEPE of 0.90 pixels, surpassing the second-best method RAFT by 1.1 pixels. The advantages are particularly pronounced in high-precision matching metrics: at a threshold of $\tau \leq 1$ px, our method attains a CMR of

TABLE III: Comparative results of different methods on three test sets of the **UBCv2 dataset (0.5-meter resolution)** in ‘mean $\pm$ std’ format. **Bold** indicates the best result, and underline indicates the second best result.

Category	Method	$\tau \leq 2\text{px}$		$\tau \leq 3\text{px}$		$\tau \leq 5\text{px}$		All	
		CMR@ $\tau\uparrow$	AEPE@ $\tau\downarrow$	CMR@ $\tau\uparrow$	AEPE@ $\tau\downarrow$	CMR@ $\tau\uparrow$	AEPE@ $\tau\downarrow$	AEPE $\downarrow$	RMSE $\downarrow$
Sparse	RIFT2 [14, 16]	0.00(± 0.00)%	—	0.00(± 0.00)%	—	0.00(± 0.00)%	—	236.79(± 1.80)	9253.72(± 110.04)
	LNIFT [15]	0.00(± 0.00)%	—	0.00(± 0.00)%	—	0.14(± 0.00)%	4.75(± 0.07)	224.28(± 0.21)	2018.86(± 41.38)
	XoFTR [35]	0.00(± 0.00)%	—	0.25(± 0.04)%	2.89(± 0.01)	8.09(± 0.21)%	4.01(± 0.11)	84.35(± 2.09)	3366.19(± 160.87)
Dense	FlowFormer [12]	0.00(± 0.00)%	—	1.45(± 0.21)%	2.59(± 0.05)	13.72(± 0.11)%	4.05(± 0.02)	10.24(± 0.00)	40.48(± 1.11)
	FlowFormer++ [28]	0.00(± 0.00)%	—	0.36(± 0.07)%	2.56(± 0.10)	5.29(± 0.15)%	4.10(± 0.00)	24.89(± 0.69)	261.65(± 15.06)
	GMFlow [41]	0.00(± 0.00)%	—	0.04(± 0.04)%	1.39(± 1.39)	0.76(± 0.07)%	4.30(± 0.01)	23.47(± 0.13)	160.79(± 5.21)
	RAFT [32]	0.45(± 0.04)%	1.76(± 0.14)	3.29(± 0.11)%	2.51(± 0.01)	23.95(± 0.52)%	3.94(± 0.02)	8.09(± 0.04)	23.39(± 1.17)
	OS ³ Flow [31]	1.01(± 0.04)%	1.58(± 0.10)	4.84(± 0.76)%	2.37(± 0.00)	26.75(± 1.18)%	3.77(± 0.04)	7.99(± 0.08)	23.81(± 1.95)
	Ours	14.89(± 0.38)%	1.56(± 0.02)	37.08(± 0.04)%	2.13(± 0.00)	72.50(± 0.42)%	2.97(± 0.00)	4.49(± 0.00)	11.97(± 0.31)

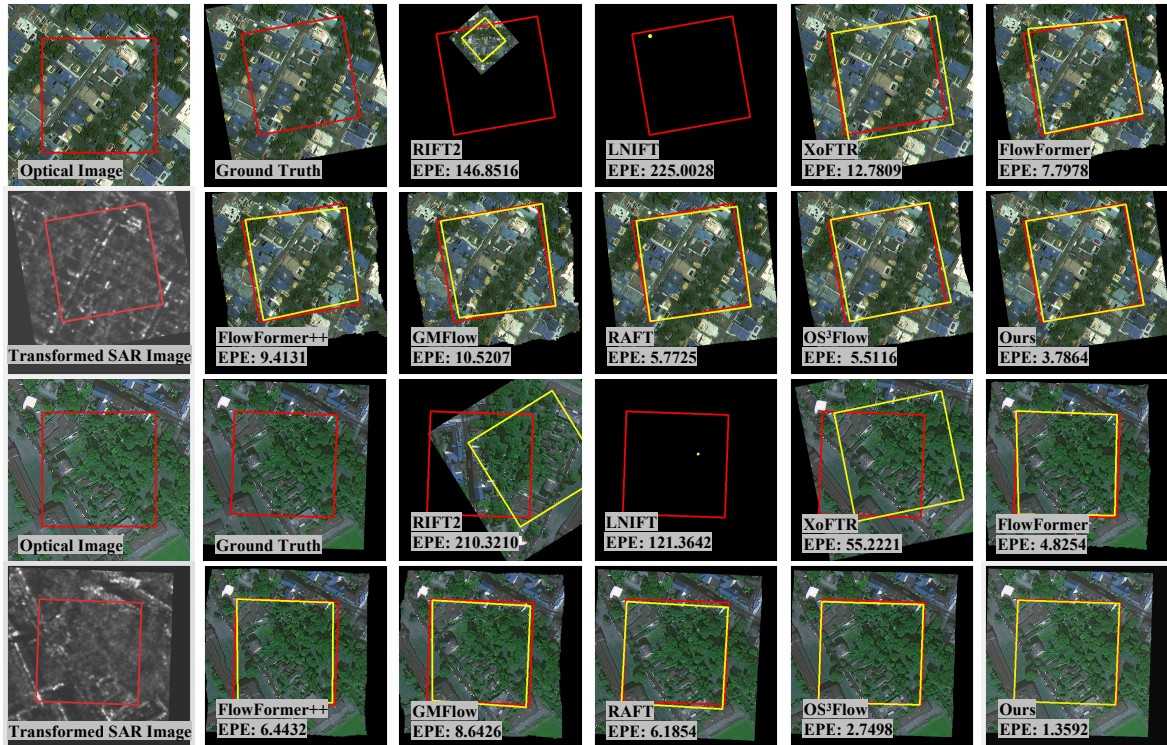


Fig. 6: Registration results on the UBCv2 dataset. The yellow line represents the ground truth registration result, and the red line represents the experimental registration result.

72.05%, surpassing the suboptimal OS³Flow by 50.91 percentage points. For $\tau \leq 2\text{px}$, our CMR reaches 96.86%, exceeding the suboptimal FlowFormer++ by 30.53 percentage points and covering nearly all test image pairs. Despite matching significantly more image pairs across thresholds, our method maintains nearly the lowest AEPE, demonstrating stable high-precision registration. Furthermore, our method achieves the lowest RMSE of 0.90, the only approach to fall below 1.0, underscoring its robustness and stability under low-resolution conditions.

2) *Results on the OS dataset:* The OS dataset, with a spatial resolution of 1 m, presents significantly greater registration challenges compared to the WHU-OPT-SAR dataset. Under identical image dimensions, its effective receptive field captures 5^2 times fewer cross-modal co-registered structural features, substantially increasing the difficulty of identifying shared correspondences. Qualitative results in Fig. 5 demon-

strate our method’s superior alignment accuracy, where predicted yellow bounding boxes closely align with ground-truth red boxes across diverse terrains, including urban areas and mountainous regions.

Table II presents the quantitative evaluation results of different methods on the OS dataset. Compared to the WHU-OPT-SAR dataset, the increased registration complexity of OS dataset leads to performance deterioration across all baseline methods. Nevertheless, our approach maintains significant superiority in both EPE and RMSE metrics. Specifically, Our method attains CMR of 33.88% at $\tau \leq 1\text{px}$, 80.19% at $\tau \leq 2\text{px}$, and 99.45% at $\tau \leq 5\text{px}$, surpassing the second-best methods by 28.73, 48.49, and 8.96 percentage points, respectively. It demonstrates that our method remains fully competent for coarse registration tasks on the moderately high spatial resolution OS dataset, while exhibiting substantially superior performance in high-precision correct registration rates

TABLE IV: Ablation study results of components in the feature extraction module on the OS dataset, including positional encoding, cross-attention, and self-attention. **Bold** indicates the best result, and underline indicates the second best result.

setup			All		$\tau \leq 1\text{px}$		$\tau \leq 2\text{px}$		$\tau \leq 3\text{px}$		$\tau \leq 5\text{px}$		Param
PE	SA	CA	AEPE $\downarrow$	RMSE $\downarrow$	CMR@ $\tau\uparrow$	AEPE@ $\tau\downarrow$	CMR@ $\tau\uparrow$	AEPE@ $\tau\downarrow$	CMR@ $\tau\uparrow$	AEPE@ $\tau\downarrow$	CMR@ $\tau\uparrow$	AEPE@ $\tau\downarrow$	(M)
		✓✓	2.45	<u>2.12</u>	8.73%	<u>0.80</u>	47.64%	<u>1.37</u>	74.06%	1.75	92.92%	2.16	8.99
✓			2.99	24.64	4.25%	0.78	39.15%	1.44	70.99%	1.88	91.51%	2.31	5.32
✓		✓	2.57	<u>2.12</u>	6.84%	0.83	42.22%	1.39	70.75%	1.80	92.45%	2.28	7.16
✓		✓✓✓	<u>2.43</u>	2.15	8.25%	<u>0.80</u>	<u>47.88%</u>	1.36	75.71%	<u>1.76</u>	<u>93.87%</u>	<u>2.17</u>	10.82
✓	✓		2.64	2.13	4.01%	0.85	40.80%	1.48	68.40%	1.88	92.92%	2.37	7.16
✓	✓✓		2.58	2.37	7.08%	0.81	41.98%	1.40	71.93%	1.85	93.63%	2.30	8.99
✓	✓	✓✓	2.56	3.58	10.38%	<u>0.80</u>	45.28%	1.38	72.88%	1.78	93.40%	2.23	10.82
✓		✓✓	2.40	1.76	<u>10.14%</u>	0.81	48.82%	1.38	<u>74.29%</u>	1.75	94.34%	2.19	8.99

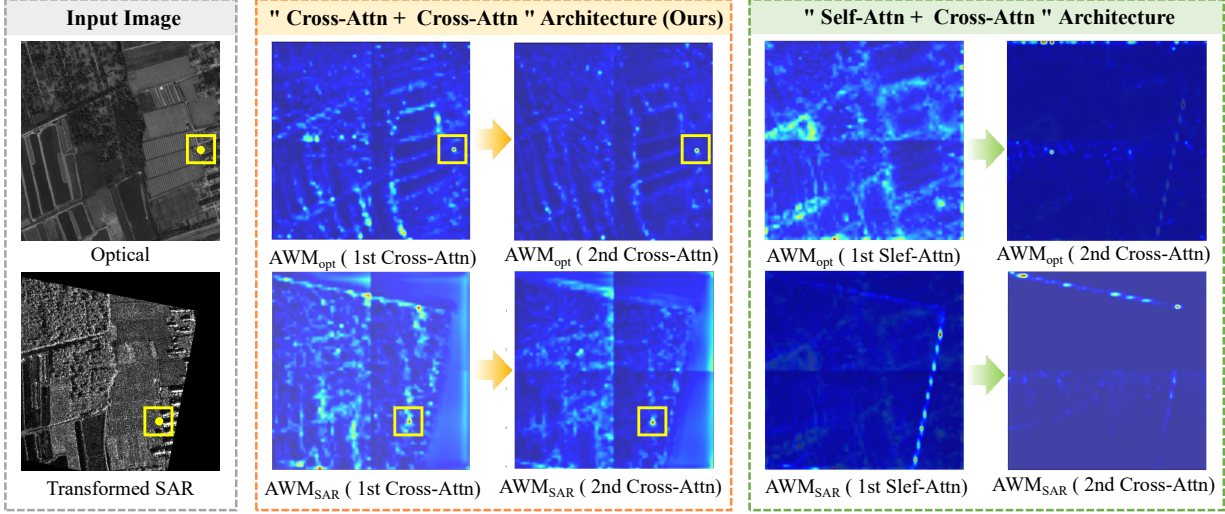


Fig. 7: Visualization of Attention Weight Matrices (AWM) from different attention architectures.

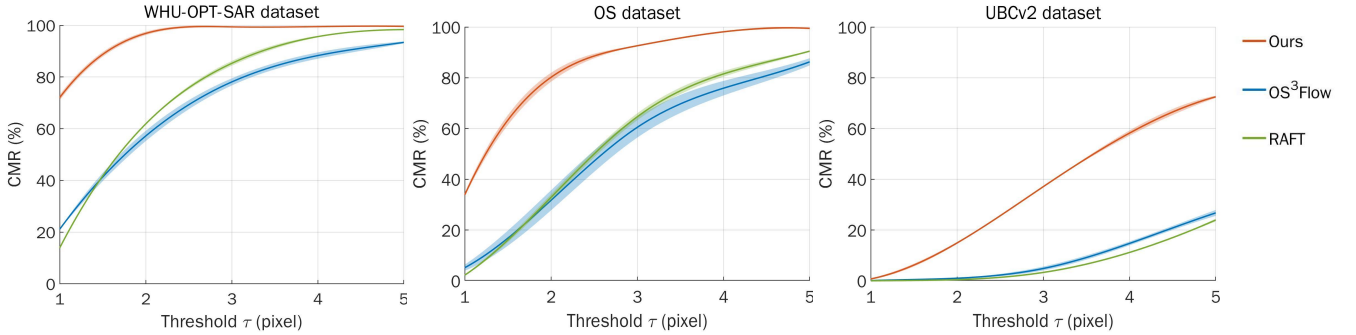


Fig. 8: CMR@ τ metric with different thresholds on three benchmark datasets.

compared to other SOTA approaches. Notably, our method attains an overall RMSE of approximately 1.5 pixels on OS dataset, further confirming its exceptional robustness in challenging registration scenarios.

3) *Results on the UBCv2 dataset:* The UBCv2 dataset, with an ultra-high spatial resolution of 0.5 m and fixed 512×512 image dimensions, contains very rare heterogeneous common structural features. Furthermore, there exist additional issues such as high image noise and cloud occlusion which severely degrade the quality of the input data. Compared with the WHU-OPT-SAR dataset and the OS dataset, the UBCv2 dataset exhibits substantially larger modality differences that

pose new challenges for image registration. We therefore constrained the affine transformation parameters to narrower ranges: translation within $[-15, 15]$ pixels (± 1 -pixel precision), scaling within $[0.9, 1.1]$ (± 0.05 precision), and rotation within $[-10^\circ, 10^\circ]$ ($\pm 1^\circ$ precision).

Qualitative results in Fig. 6 demonstrate that while alignment between predicted yellow boxes and ground-truth red boxes remains imperfect, our method exhibits marked improvements over competitors. Quantitative evaluations in Table III reveal that traditional methods (RIFT, LNIFT) fail entirely on this dataset, while learning-based approaches suffer significant performance degradation. Nevertheless, our method

TABLE V: Ablation results of the LSR module across three datasets. ‘LSR’ denotes the least squares regression integrated into network training, and ‘LS’ indicates the least squares regression excluded from network training. **Bold** indicates the best result, and underline indicates the second best result.

Datasets	setup	AEPE↓	RMSE↓	CMR@ τ ↑	
				$\tau \leq 1\text{px}$	$\tau \leq 2\text{px}$
WHU-OPT-SAR dataset	None	1.50(± 0.06)	4.35(± 2.36)	34.24(± 0.94)%	83.86(± 0.77)%
	w/ LS	0.90(± 0.03)	1.10(± 0.68)	71.24(± 0.94)%	96.57(± 0.31)%
	w/ LSR	<u>0.90(± 0.04)</u>	0.62(± 0.15)	72.05(± 1.06)%	96.86(± 0.65)%
OS dataset	None	2.45(± 0.03)	1.85(± 0.04)	5.19(± 0.66)%	46.62(± 1.49)%
	w/ LS	1.50(± 0.06)	1.45(± 1.07)	32.00(± 1.25)%	79.25(± 2.22)%
	w/ LSR	1.48(± 0.02)	0.75(± 0.02)	33.88(± 0.73)%	80.19(± 1.77)%
UBCv2 dataset	None	7.33(± 0.01)	30.52(± 0.94)	0.00(± 0.00)%	0.76(± 0.14)%
	w/ LS	4.67(± 0.16)	10.72(± 0.37)	0.11(± 0.11)%	12.02(± 0.11)%
	w/ LSR	4.49(± 0.00)	<u>11.97(± 0.31)</u>	0.73(± 0.18)%	14.89(± 0.38)%

achieves state-of-the-art results with CMR of 14.89% at $\tau \leq 2$ px, 37.08% at $\tau \leq 3$ px, and 72.50% at $\tau \leq 5$ px, outperforming the second-best methods by 13.88, 32.24, and 45.75 percentage points, respectively. The experimental results on the ultra-high-resolution UBCv2 dataset demonstrate that our method achieves coarse registration on the majority of image pairs and fine-grained registration on a subset of cases, despite the dataset’s extreme challenges.

Fig. 8 illustrates the trends of CMR under varying thresholds τ for different methods. The solid curves represent the mean CMR values of each method across three randomly transformed test sets, while the shaded regions around the curves indicate the variance in CMR observed across these sets. It can be observed that our method significantly outperforms both RAFT and OS³Flow in terms of both matching accuracy and stability. On the lower-resolution WHU-OPT-SAR dataset OS dataset, the advantage of our approach is most pronounced in high-precision matches with registration errors below 2 pixels. Moreover, it achieved nearly complete matching across almost all samples at thresholds of 3 and 4 pixels on the two datasets, respectively. The results on the UBCv2 dataset further highlight the limitations of current cross-modal registration paradigms when applied to ultra-high-resolution scenarios. While our hybrid CNN-Transformer architecture and LSR module partially mitigate these challenges, the substantial performance gap emphasizes the need for novel methodologies to address severe modality discrepancies, sparse shared structural features, and pervasive noise, which we leave as future research venue.

D. Ablation Study

In this section, we conduct comprehensive ablation studies to analyze the contributions of key design choices in our method, including individual components within cross-modal feature extraction module and the LSR module.

1) *Cross-Attention-Only Mechanism*: To validate the effectiveness of our proposed cross-attention-only mechanism, we evaluated the impact of different components, including positional encoding (PE), cross-attention (CA), and self-attention (SA), on registration performance using the OS dataset, as summarized in Table IV. Additionally, we compared the attention weight matrices produced by our proposed “dual-level

cross-attention” architecture against those from the conventional “self-attention + cross-attention” structure, as illustrated in Fig. 7.

PE introduces spatial awareness into the registration process, enhancing robustness to geometric deformations. CA establishes inter-modal dependencies between optical and SAR features through selective information exchange. SA facilitates intra-modal context aggregation to refine domain-specific representations. As shown in the table, PE, CA, and SA contribute to improving registration performance to varying degrees. However, the “cross-attention only” configuration, comprising two cross-attention layers, is more suitable for optical-SAR registration than the conventional “SA + CA” design. This advantage can be visually interpreted from the attention weight matrices. In our “CA + CA” architecture (orange boxes in Fig. 7), the first cross-attention layer effectively filters and aligns cross-modal features, highlighting numerous potential correspondence regions. The second cross-attention layer further refines and fuses these aligned features at a deeper level, concentrating attention on semantically consistent key areas, as indicated by the yellow boxes. In contrast, in the traditional “SA + CA” architecture (green boxes in Fig. 7), the initial self-attention layer primarily enhances intra-image contextual relationships (e.g., structural details within a single building). However, due to the lack of cross-modal guidance at this stage, the resulting activations may not align well with the SAR modality. Given the significant domain gaps, including nonlinear radiometric differences, noise patterns, and structural discrepancies, self-attention maps from optical and SAR images often fail to achieve meaningful alignment. This misalignment can amplify modal differences rather than mitigate them, thereby impairing the subsequent cross-attention performance.

2) *LSR Module*: To validate the critical role of our proposed LSR module, we conducted systematic ablation experiments across three datasets. These experiments evaluated two configurations: the proposed LSR module integrated into network training, and the classical Least Squares (LS) post-processing excluded from network training.

As detailed in Table IV, both the LSR module and LS achieve significant performance improvements gains across all three benchmark datasets, compared to the original optical flow fields. This validates that imposing geometric constraints on divergent flow fields effectively filters error-prone correspondences. Notably, our LSR module delivers breakthrough improvements in sub-pixel precision metrics, increasing CMR ($\tau \leq 1\text{px}$) by more than 72% on the WHU-OPT-SAR dataset. Controlled comparative analyses further reveal that the LSR module achieves superior RMSE performance while maintaining robust registration performance on challenging image pairs, outperforming the LS baseline. This advantage originates from the geometric constraint loss (Eq. 11), which enforces physically plausible affine transformations in optical flow predictions and suppresses outliers simultaneously. Experimental results demonstrate that the differentiable LSR module, guided by geometric priors during end-to-end network training, enhances both accuracy and robustness in cross-modal registration tasks. The synergistic optimization

TABLE VI: Comparison of Registration Efficiency Among Different Methods.

Method	RIFT2	LNIFT	XoFTR	FlowFormer	FlowFormer++	GMFlow	RAFT	OS ³ Flow	Ours
Param(M)	—	—	10.86	16.08	15.88	4.68	5.25	10.52	8.99
FPS($s/pair$)	12.9325	8.7453	0.1386	0.9271	1.5124	0.0983	0.1169	6.2885	0.1429
GFLOPs	—	—	293.77	146.96	71.07	63.45	207.01	414.03	273.62

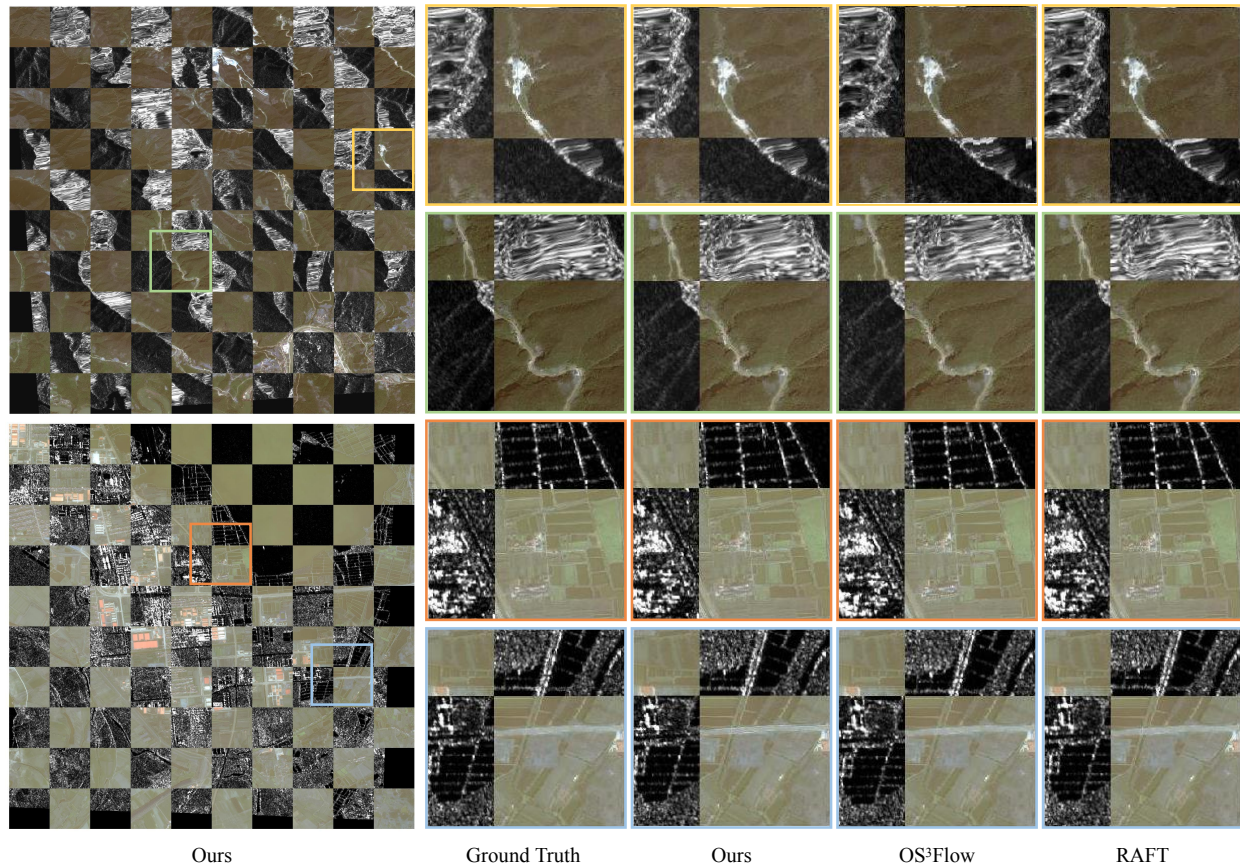


Fig. 9: Registration results of two large-scale image pairs (1600×1600 pixels) from the WHU-OPT-SAR dataset. Yellow, red, green, and blue boxes denote enlarged views of local regions covering distinct geographical features, including roads, harbors, and mountainous terrain.

of geometric consistency and feature representation through our trainable architecture accounts for its consistent metric superiority across all datasets.

E. Efficiency and Generalization Capability

Computational Efficiency. Table VI compares the performance of various methods in terms of parameter count, inference speed (FPS), and floating point operations (FLOPs). The FPS metric was computed by averaging the registration time per image pair over 700 test pairs. Our method achieves a processing speed of **0.1429 seconds per pair**, ranking second only to GMFlow, RAFT and XoFTR. FLOPs were measured using the thop library with input images resized to 512×512 pixels, yielding a computational complexity of 273.62 GFLOPs for our model. Notably, traditional methods RIFT/LNIFT exhibit significantly longer computation times compared to learning-based approaches, underscoring the computational superiority of deep learning paradigms. The

results validate the practical viability of our framework for real-time cross-modal registration tasks.

Large-Scale Image Registration. In practical registration tasks, the optical-SAR image pairs may have a large image size. To show the practicability of our method, we performed a direct evaluation on large-scale optical-SAR image pairs of size 1500×1500 using models trained exclusively on cropped 512×512 patches from the WHU-OPT-SAR dataset. The affine transformation ranges in these large-scale images remain consistent with those in the training images. Without loss of generality, we adopt a size of 1500×1500 as it is the largest dimension fitting our single GPU device. As visualized in the checkerboard comparison in Fig. 9, our method achieves superior alignment accuracy in extended geographical features without any architecture modification or fine-tuning. More specifically, our method achieves exciting qualitative alignment accuracy in geographical patterns such as roads, coastal boundaries, and mountainous terrains, while competing methods exhibit obvious misalignment or blurring artifacts.

Quantitatively, the registration achieved an EPE of 1.8 pixels and 1.4 pixels on two pairs of large-scale images, respectively, which validates the effectiveness of our framework in large-scale image registration tasks.

V. CONCLUSION

In this paper, we present GDROS, a geometry-constrained end-to-end framework for optical-SAR image registration. The proposed framework employs a hybrid CNN-Transformer architecture to extract cross-modality interactive deep features, followed by a novel LSR module that geometrically rectifies predicted optical flow fields while filtering out mismatched correspondences, ultimately deriving affine transformation parameters between image pairs. Extensive experiments on the WHU-OPT-SAR dataset, the OS dataset, and the UBCv2 dataset demonstrate GDROS's exceptional capability to address registration challenges under significant geometric variations, achieving high accuracy and robust performance across different resolution scenarios. While the framework shows promising scalability to large-scale images, its current limitations in high-precision registration for ultra-high-resolution data with sparse shared features highlight critical future research directions.

ACKNOWLEDGMENTS

Research presented here has been partly supported by the National Natural Science Foundation of China under Grant Number 62201603, the Hunan Provincial Natural Science Foundation of China under Grant Number 2025JJ40060, the China Postdoctoral Science Foundation under Grant Number 2023TQ0088, the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20233539, and the Research Program of NUDT under Grant Number ZK22-04.

REFERENCES

- [1] Daniel Barath, Jiri Matas, and Jana Noskova. "MAGSAC: marginalizing sample consensus". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10197–10205.
- [2] Herbert Bay et al. "Speeded-up robust features (SURF)". In: *Computer Vision and Image Understanding (CVIU)* 110 (2008), pp. 346–359.
- [3] Nicolas Carion et al. "End-to-end object detection with transformers". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020, pp. 213–229.
- [4] Ondrej Chum and Jiri Matas. "Matching with PROSAC: progressive sample consensus". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. IEEE. 2005, pp. 220–226.
- [5] Konstantinos G Derpanis. "Overview of the RANSAC Algorithm". In: *Image Rochester NY 4.1* (2010), pp. 2–3.
- [6] 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)*. 2018, pp. 224–236.
- [7] Qiaole Dong and Yanwei Fu. "Memflow: Optical flow estimation and prediction with memory". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 19068–19078.
- [8] Alexey Dosovitskiy et al. "FlowNet: Learning optical flow with convolutional networks". In: *Proceedings of the International Conference on Computer Vision (ICCV)*. 2015, pp. 2758–2766.
- [9] Martin A Fischler and Robert C Bolles. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: *Communications of the ACM* 24.6 (1981), pp. 381–395.
- [10] James J Gibson. *The perception of the visual world*. 1950.
- [11] Xingliang Huang et al. "Urban building classification (UBC) V2—A benchmark for global building detection and fine-grained classification from satellite imagery". In: *IEEE Transactions on Geoscience and Remote Sensing (TGRS)* 61 (2023), pp. 1–16.
- [12] Zhaoyang Huang et al. "Flowformer: A transformer architecture for optical flow". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2022, pp. 668–685.
- [13] Lloyd Haydn Hughes et al. "A deep learning framework for matching of SAR and optical imagery". In: *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS)* 169 (2020), pp. 166–179.
- [14] Jiayuan Li, Qingwu Hu, and Mingyao Ai. "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform". In: *IEEE Transactions on Image Processing (TIP)* 29 (2019), pp. 3296–3310.
- [15] Jiayuan Li et al. "LNIFT: Locally normalized image for rotation invariant multimodal feature matching". In: *IEEE Transactions on Geoscience and Remote Sensing (TGRS)* 60 (2022), pp. 1–14.
- [16] Jiayuan Li et al. "RIFT2: Speeding-up RIFT with a new rotation-invariance technique". In: *arXiv preprint arXiv:2303.00319* (2023).
- [17] Weijie Li et al. "Predicting gradient is better: Exploring self-supervised learning for SAR ATR with a joint-embedding predictive architecture". In: *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS)* 218 (2024), pp. 326–338.
- [18] Xue Li et al. "MCANet: A joint semantic segmentation framework of optical and SAR images for land use classification". In: *International Journal of Applied Earth Observation and Geoinformation (INT J APPL EARTH OBS)* 106 (2022), p. 102638.
- [19] Rui Liu, Jing Ling, and Hongsheng Zhang. "SoftFormer: SAR-optical fusion transformer for urban land use and land cover classification". In: *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS)* 218 (2024), pp. 277–293.
- [20] Xuecong Liu et al. "A Fast Algorithm for High Accuracy Airborne SAR Geolocation Based on Local Linear Approximation". In: *IEEE Transactions on Instrumentation and Measurement (TIM)* 71 (2022), pp. 1–12.

- [21] Xuecong Liu et al. “Robust multi-sensor image matching based on normalized self-similarity region descriptor”. In: *Chinese Journal of Aeronautics* 37.1 (2024), pp. 271–286.
- [22] Xuecong Liu et al. “Shape-Adaptive Modality Independent Region Descriptor for Multimodal Remote Sensing Image Matching”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (J-STARS)* (2024).
- [23] Bruce D Lucas and Takeo Kanade. “An iterative image registration technique with an application to stereo vision”. In: *IJCAI’81: 7th International Joint Conference on Artificial Intelligence*. Vol. 2. 1981, pp. 674–679.
- [24] Pauline C Ng and Steven Henikoff. “SIFT: Predicting amino acid changes that affect protein function”. In: *Nucleic Acids Research* 31 (2003), pp. 3812–3814.
- [25] Ethan Rublee et al. “ORB: An efficient alternative to SIFT or SURF”. In: *International Conference on Computer Vision Systems (ICVS)*. 2011, pp. 2564–2571.
- [26] Paul-Edouard Sarlin et al. “Superglue: Learning feature matching with graph neural networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 4938–4947.
- [27] Liang Shen et al. “Extended neighborhood consensus with affine correspondence for outlier filtering in feature matching”. In: *IEEE Transactions on Geoscience and Remote Sensing (TGRS)* (2024).
- [28] Xiaoyu Shi et al. “Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 1599–1610.
- [29] Deqing Sun et al. “Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 8934–8943.
- [30] Jiaming Sun et al. “LoFTR: Detector-free local feature matching with transformers”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 8922–8931.
- [31] Zixuan Sun et al. “OS3Flow: Optical and SAR image registration using symmetry-guided semi-dense optical flow”. In: *IEEE Geoscience and Remote Sensing Letters (GRSL)* (2024).
- [32] Zachary Teed and Jia Deng. “Raft: Recurrent all-pairs field transforms for optical flow”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020.
- [33] Xichao Teng et al. “OMIRD: Orientated Modality Independent Region Descriptor for Optical-to-SAR Image Matching”. In: *IEEE Geoscience and Remote Sensing Letters (GRSL)* 20 (2023), pp. 1–5.
- [34] Philip HS Torr and Andrew Zisserman. “MLESAC: A new robust estimator with application to estimating image geometry”. In: *Computer Vision and Image Understanding (CVIU)* 78.1 (2000), pp. 138–156.
- [35] Önder Tuzcuoğlu et al. “Xoftr: Cross-modal feature matching transformer”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 4275–4286.
- [36] Peng Wang et al. “CIRSM-Net: A Cyclic Registration Network for SAR and Optical Images”. In: *IEEE Transactions on Geoscience and Remote Sensing (TGRS)* (2025).
- [37] Philippe Weinzaepfel et al. “DeepFlow: Large displacement optical flow with deep matching”. In: *Proceedings of the International Conference on Computer Vision (ICCV)*. 2013, pp. 1385–1392.
- [38] Deliang Xiang et al. “Optical and SAR image registration based on feature decoupling network”. In: *IEEE Transactions on Geoscience and Remote Sensing (TGRS)* 60 (2022), pp. 1–13.
- [39] Yuming Xiang et al. “Automatic registration of optical and SAR images via improved phase congruency model”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (J-STARS)* 13 (2020), pp. 5847–5861.
- [40] Yuming Xiang et al. “OS-PC: Combining feature representation and 3-D phase correlation for subpixel optical and SAR image registration”. In: *IEEE Transactions on Geoscience and Remote Sensing (TGRS)* 58.9 (2020), pp. 6451–6466.
- [41] Haoifei Xu et al. “Unifying flow, stereo and depth estimation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2023).
- [42] Yuanxin Ye et al. “A multiscale framework with unsupervised learning for remote sensing image registration”. In: *IEEE Transactions on Geoscience and Remote Sensing (TGRS)* 60 (2022), pp. 1–15.
- [43] Yuanxin Ye et al. “A robust multimodal remote sensing image registration method and system using steerable filters with first-and second-order gradients”. In: *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS)* 188 (2022), pp. 331–350.
- [44] Haiming Zhang et al. “M3ICNet: A cross-modal resolution preserving building damage detection method with optical and SAR remote sensing imagery and two heterogeneous image disaster datasets”. In: *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS)* 221 (2025), pp. 224–250.
- [45] Han Zhang et al. “Optical and SAR image dense registration using a robust deep optical flow framework”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (J-STARS)* 16 (2023), pp. 1269–1294.
- [46] Wenfei Zhang et al. “Multi-Resolution SAR and Optical Remote Sensing Image Registration Methods: A Review, Datasets, and Future Perspectives”. In: *arXiv preprint arXiv:2502.01002* (2025).
- [47] Yongjun Zhang et al. “Histogram of the orientation of the weighted phase descriptor for multi-modal remote sensing image matching”. In: *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS)* 196 (2023), pp. 1–15.
- [48] Yongjun Zhang et al. “Multi-modal remote sensing image robust matching based on Second-order tensor

- orientation feature transformation”. In: *IEEE Transactions on Geoscience and Remote Sensing (TGRS)* (2025).
- [49] Yongjun Zhang et al. “Robust registration of multi-modal remote sensing images based on multi-dimensional oriented self-similarity features”. In: *International Journal of Applied Earth Observation and Geoinformation (INT J APPL EARTH OBS)* 127 (2024), p. 103639.
- [50] Yan Zhou, Jinding Gao, and Xiaoping Liu. “A unified feature-motion consistency framework for robust image matching”. In: *ISPRS Journal of Photogrammetry and Remote Sensing (ISPRS)* 218 (2024), pp. 368–388.