# SOMA: Feature Gradient Enhanced Affine-Flow Matching for SAR-Optical Registration

**Haodong Wang**[1,2,3*], **Tao Zhuo**[4*],
**Xiuwei Zhang**[1,2,3†], **Hanlin Yin**[1,2,3†], **Wencong Wu**[1,2,3], **Yanning Zhang**[1,2,3]

[1]School of Computer Science, Northwestern Polytechnical University
[2]Shaanxi Provincial Key Lab. of Speech and Image Information Processing
[3]National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology
[4]College of Information Engineering, Northwest A&F University
traslauc@mail.nwpu.edu.cn, xwzhang@nwpu.edu.cn, iverlon1987@nwpu.edu.cn

## Abstract

Achieving pixel-level registration between SAR and optical images remains a challenging task due to their fundamentally different imaging mechanisms and visual characteristics. Although deep learning has achieved great success in many cross-modal tasks, its performance on SAR-Optical registration tasks is still unsatisfactory. Gradient-based information has traditionally played a crucial role in handcrafted descriptors by highlighting structural differences. However, such gradient cues have not been effectively leveraged in deep learning frameworks for SAR-Optical image matching. To address this gap, we propose SOMA, a dense registration framework that integrates structural gradient priors into deep features and refines alignment through a hybrid matching strategy. Specifically, we introduce the Feature Gradient Enhancer (FGE), which embeds multi-scale, multi-directional gradient filters into the feature space using attention and reconstruction mechanisms to boost feature distinctiveness. Furthermore, we propose the Global-Local Affine-Flow Matcher (GLAM), which combines affine transformation and flow-based refinement within a coarse-to-fine architecture to ensure both structural consistency and local accuracy. Experimental results demonstrate that SOMA significantly improves registration precision, increasing the CMR@1px by 12.29% on the SEN1-2 dataset and 18.50% on the GFGE_SO dataset. In addition, SOMA exhibits strong robustness and generalizes well across diverse scenes and resolutions.

**Code** — https://github.com/traslauc/SOMA

## Introduction

Multi-modal remote sensing image registration is a fundamental step in multi-source data fusion and analysis, supporting tasks such as object detection, change detection, 3D reconstruction, and joint classification (Jiang et al. 2021). Among various modalities, Synthetic Aperture Radar (SAR) and optical imagery have received particular attention due to their complementary sensing characteristics. SAR provides all-weather, all-day imaging with strong sensitivity to surface structure and roughness, even through cloud cover. Optical imagery, in contrast, offers high-resolution visual details and semantically rich information. Their combination is especially valuable in emergency response and terrain interpretation under adverse weather. Accurate alignment between SAR and optical images is thus essential for effective joint analysis.

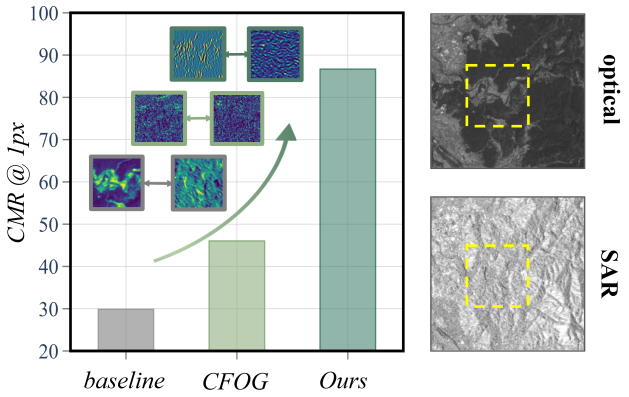


Figure 1: Comparison of CMR@1px among methods leveraging different types of feature representations. The proposed SOMA, which incorporates feature gradient enhancement, significantly outperforms both the conventional CNN baseline and the CFOG that leverages image gradients.

Recent years, learning-based SAR-Optical image registration methods have demonstrated remarkable improvement in robustness (Dong et al. 2019; Quan et al. 2022; Li et al. 2023). However, in practical applications, the substantial disparity in imaging mechanisms and image characteristics between SAR and optical images presents a significant challenge for accurate registration. In particular, SAR images often contain regions severely affected by nonlinear radiometric distortions and speckle noise, where the pixel-level features lack distinctiveness from their surroundings. This makes it difficult for existing methods to sufficiently extract reliable dense features, ultimately limiting the fol-

*These authors contributed equally.
†Corresponding authors.

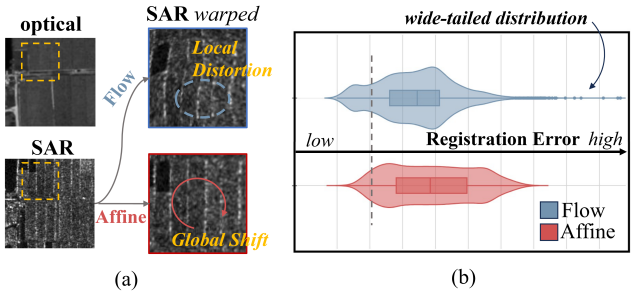


Figure 2: (a) Flow fields prioritize local warp, but leave distortion. Affine fields maintain structural consistency via global transforms but lost local precision. (b) Violin plots further reveal distinct error profiles.

lowing matching precision.

In fact, it is observed that gradient information plays a critical role in the design of handcrafted features. Handcrafted features (Yao et al. 2022; Zhang et al. 2024) utilize multi-directional and multi-scale gradient extraction strategies to effectively capture structural information, allowing local differences to be distinguished from their surroundings. However, when gradient extraction is directly applied to raw images, these features are easily overwhelmed by noise, leading to unstable responses and disturbed semantic information, as shown in Figure 1.

Building on this analysis, we explore the explicit integration of multi-directional and multi-scale gradient features into learning-based feature extraction process. Unlike traditional methods that compute gradients directly from images, we apply directional gradient filters to the feature pyramid to better capture structural cues. However, high-level abstract features often contain redundant and entangled information across spatial and channel dimensions. Direct gradient filtering in such feature spaces can severely degrade feature integrity. To address this, we adopt a feature reconstruction mechanism (Li, Wen, and He 2023) that reduces redundancy in spatial and channel dimensions, retaining features crucial for effective gradient filtering and accurate correspondence estimation. We further incorporate the attention mechanism and dilated convolutions to effectively integrate and reinforce the extracted gradient cues. Additionally, a final Gaussian smoothing step is applied to reduce early residual noise. We encapsulate this gradient enhancement into a dedicated module named Feature Gradient Enhancer (FGE).

Although FGE enhances the deep features, achieving accurate dense registration still requires a robust matcher. Most existing multimodal remote sensing image matching approaches focus on learning a single geometric transformation. However, global transformation lacks local precision, while dense warp tends to compromise structural consistency. As a result, these methods may suffer from local distortion under challenging conditions, as shown in Figure 2.

We propose a Global-Local Affine-Flow Matcher (GLAM) to fix this issue, which introduces affine regression in a coarse-to-fine manner. The affine regression captures large-scale structural alignment and provides consistent deformation priors, while the flow regression refines local details at progressively finer resolutions. By coupling these two transformations, our GLAM mitigates local misalignment and enhances global stability, leading to more precise and robust cross-modal matching.

Finally, to give the entire pipeline a stable coarse warp, we adopt a frozen DINOv2 (Oquab et al. 2024; Darcet et al. 2024) vision backbone as the coarse-level feature encoder, inspired by RoMa (Edstedt et al. 2024). This global perceptual context from DINOv2 complements the FGE, providing reliable input to GLAM.

To this end, we propose a dense matching framework for SAR and optical image registration, termed SOMA (**S**AR-**O**ptical **MA**tching), which integrates the proposed FGE and GLAM modules within an end-to-end architecture to achieve robust and accurate SAR-Optical image registration.

Our main contributions are summarized as follows:

- We propose FGE, a feature gradient enhancer that leverages multi-scale, multi-directional gradients to model distinctive representations, improving matching precision and robustness between SAR and Optical images.
- We propose GLAM, a global-local affine-flow matcher that forms a bidirectional coupling between local flow and global affine estimation, improving local accuracy while avoiding distortion.
- We introduce a frozen DINOv2 to stable init alignment and improve robustness in registration.
- We benchmark SOMA on diverse public datasets and it consistently surpasses state-of-the-art baselines, boosting CMR@1px by **12.29%** on SEN1-2 and **18.50%** on GFGE_SO dataset, while retaining strong robustness and generalization in challenging scenes.

## Related Work

### SAR-Optical Registration Methods

SAR-Optical image registration approaches can be broadly categorized into handcrafted methods and deep learning-based models. Handcrafted methods typically employ statistical similarity measures (Sedaghat and Mohammadi 2019; Hong et al. 2022) or gradient-based descriptors such as CFOG and SAR-SIFT (Dellinger et al. 2014; Xiang, Wang, and You 2018; Yao et al. 2022; Zhang et al. 2024) to extract structural cues. While these methods are effective in capturing geometry, they often suffer from limited robustness when confronted with large modality gaps in texture and radiometry. In contrast, recent learning-based methods leverage Siamese networks (Liu, Qi, and Peng 2023), attention mechanisms (Zhang et al. 2021; Quan et al. 2022), or unsupervised learning strategies (Ye et al. 2022) to extract modality-invariant features, improving generalization across diverse scenarios. However, such learning-based methods may still underperform in terms of fine-grained matching accuracy. This limitation has also been noted in prior works (Zhang et al. 2020), suggesting the importance of highly distinctive feature representations for accurate pixel-level alignment.

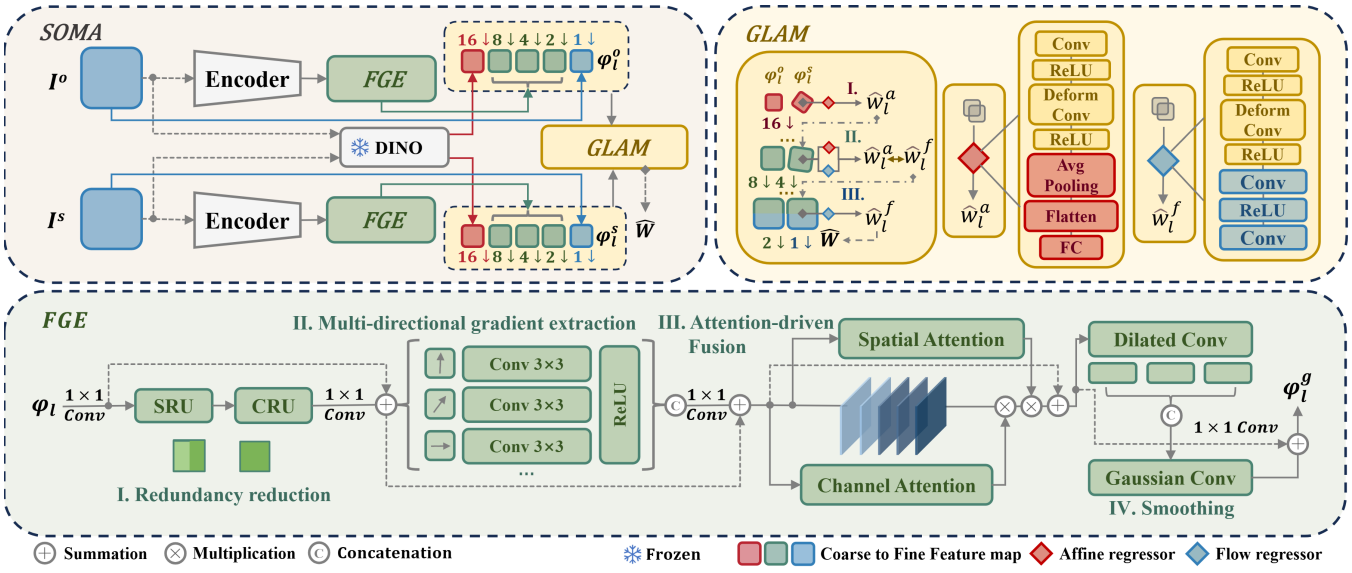


Figure 3: SOMA framework with FGE and GLAM. Given a SAR-Optical image pair, SOMA extracts multi-scale features by two separated ResNet50 encoders and a frozen DINOv2 branch, then enhances representations with Feature Gradient Enhancer (FGE), and performs hierarchical alignment using Global-Local Affine-Flow Matcher (GLAM). All convolution layers use a kernel size of $3 \times 3$ by default, unless stated otherwise.

### Pretrained Vision Models

Robust cross-modal matching hinges on strong, transferable feature representations. Large self-supervised vision models have made this feasible: models such as DINOv2 (Oquab et al. 2024), CLIP (Radford et al. 2021), and iBOT (Zhou et al. 2022)—trained with contrastive or reconstruction objectives on massive image data—show remarkable generalization (Cai et al. 2025). Recent work across a wide range of vision tasks further confirms that keeping these models frozen boosts cross-domain performance: RoMa (Edstedt et al. 2024) and OmniGlue (Jiang et al. 2024) exploit frozen DINOv2 features to stabilize semantic correspondences, while ChangeCLIP (Dong et al. 2024) achieves comparable gains with CLIP embeddings. Nevertheless, effectively utilizing their representations for robust initialization in SAR-Optical matching remains an open issue. Motivated by this, we adopt a frozen DINOv2 backbone as an encoder component in our registration framework, providing a reliable foundation for subsequent coarse-to-fine alignment.

### Sparse and Dense, Local and Global Deformation Estimation

Within multi-modal image matching, traditional keypoint-based methods are effective but often struggle in textureless or repetitive regions due to sparse and unstable correspondences. Dense matching approaches, which estimate pixel-wise deformations, offer higher precision but lack global constraints, making them prone to structural inconsistencies (Melekhov et al. 2019; Truong, Danelljan, and Timofte 2020). Affine transformation (Bebis et al. 1999) methods (Xu, Yuan, and Ma 2023; Xu et al. 2022) provide efficient global alignment but are often limited by parameter coupling and sensitivity to local variations. To overcome these limitations, our GLAM establishes a cooperative optimization between global affine estimation and local flow refinement, achieving both structural consistency and fine-grained adaptability across diverse scenarios.

## Method

### Framework Overview

As illustrated in Figure 3, SOMA predicts a dense deformation field $\hat{W} \in \mathbb{R}^{H \times W \times 2}$ that warps an optical image $I^{o} \in \mathbb{R}^{H \times W}$ to align with a SAR image $I^{s} \in \mathbb{R}^{H \times W}$.

Dual encoders respectively extract five-level feature pyramids $\varphi_l^{o}, \varphi_l^{s} \in \mathbb{R}^{C \times H_l \times W_l}$, where $l \in \{1, 2, 4, 8, 16\}$ denotes the downsampling factor with respect to the original resolution. At the coarsest scale ($l = 16$), a frozen DINOv2 is incorporated for robust features. At intermediate scales, we apply the Feature Gradient Enhancer (FGE) to explicitly inject multi-directional gradient into the feature space. FGE first reduces spatial and channel redundancy, then applies fixed directional filters followed by attention-based fusion and Gaussian smoothing, yielding enhanced features with improved structural distinctiveness.

These features are fed into the Global-Local Affine-Flow Matcher (GLAM), which estimates a hierarchy of deformation fields $\hat{W}_l$. In a coarse-to-fine manner, an affine regressor predicts global transformation parameters $\theta_l \in \mathbb{R}^{2 \times 3}$, which are then converted into a full-resolution displacement field $\hat{W}_l^{a} \in \mathbb{R}^{H_l \times W_l \times 2}$, while a flow regressor estimates residual flow $\hat{W}_l^{f} \in \mathbb{R}^{H_l \times W_l \times 2}$, refining the warped result from the previous level. In addition, a pixel-wise certainty map $\hat{p} \in \mathbb{R}^{H \times W}$ is predicted at the finest level to weight supervision during training and is omitted during inference.

## Feature Extraction

We extract multi-scale features at four levels $l \in \{1, 2, 4, 8\}$ using two modality-specific ResNet50 backbones trained from scratch. However, deep CNN features are sensitive to modality imbalance: for instance, SAR images may exhibit severe sudden intensity changes in specular or noisy regions, leading to unstable responses (Zhang, Wang, and Liu 2022).

To stabilize coarse alignment, we incorporate a frozen DINOv2 at $l = 16$. Pretrained via self-distillation, DINOv2 produces contrast-aware embeddings that remain perceptually meaningful across modalities (Cai et al. 2025). These features serve as a robust structural anchor at the coarsest level, helping the hierarchical matcher establish reliable correspondences before refinement at higher resolutions.

## Feature Gradient Enhancer (FGE)

Accurate fine-scale registration requires reliable feature description ability with high distinctiveness across SAR and optical modalities. Although gradient cues are naturally suited for capturing such representation, directly applying fixed gradient filters to high-dimensional CNN feature maps often amplifies aliased noise (Ribeiro and Schön 2021). Moreover, how to effectively integrate gradient information into deep neural networks remains an open issue. To address this, we propose the Feature Gradient Enhancer (FGE), which first reduces feature redundancy, and then explicitly integrates multi-directional gradient cues using attention and dilated convolutions, resulting in robust and distinctive feature representations. The details of FGE's four main components are elaborated as follows.

**Redundancy reduction.** FGE begins by reconstructing the input feature map $\varphi_l$. We adapt two lightweight modules from SCConv (Li, Wen, and He 2023): a spatial reconstruction unit $\mathrm{SRU}(\cdot)$ and a channel reconstruction unit $\mathrm{CRU}(\cdot)$, suppressing noisy and aliased responses. The reconstructed feature $F_{\mathrm{recon}}$ serves as a foundation that facilitates effective gradient extraction:

$$F_{\mathrm{recon}} = \mathrm{CRU}\big(\mathrm{SRU}(\varphi_l)\big) + \varphi_l. \tag{1}$$

Residual connection is employed to maintain stable information throughout the reduction process and subsequent enhancement stages.

**Multi-directional gradient extraction.** We apply a bank of rotated Sobel-like kernels $\{K_{\theta_i}\}$ on $F_{\mathrm{recon}}$, each capturing a directional gradient:

$$G_i = \mathrm{ReLU}\big(K_{\theta_i} * F_{\mathrm{recon}}\big), \quad i = 1, \ldots, N. \tag{2}$$

In practice, we adopt $N = 8$ different directions with $\theta_i \in \{0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°\}$. The outputs $G_i$ of these $N$ branches are concatenated and fused into $F_{\mathrm{grad}}$ via $1 \times 1$ convolution,

$$F_{\mathrm{grad}} = \mathrm{Conv}_{1\times1}\big([G_1, \ldots, G_N]\big) + F_{\mathrm{recon}}. \tag{3}$$

**Attention-driven fusion.** To adaptively enhance the most informative gradient cues under varying scene conditions, we introduce simple attentions (Woo et al. 2018) following the directional filtering. Specifically, we apply a lightweight channel-attention block $\mathrm{CA}(\cdot)$ and a spatial-attention block $\mathrm{SA}(\cdot)$ to selectively emphasize salient gradient responses. This mechanism compensates for the fixed nature of gradient kernels and enhances the discriminability of structural features crucial for matching. Formally, the attention-enhanced feature $F_{\mathrm{att}}$ is computed as:

$$F_{\mathrm{att}} = F_{\mathrm{grad}} \otimes \mathrm{CA}(F_{\mathrm{grad}}) \otimes \mathrm{SA}(F_{\mathrm{grad}}) + F_{\mathrm{grad}}, \tag{4}$$

where $\otimes$ denotes element-wise multiplication.

To enrich structural representation with multi-scale context, we apply dilated convolutions with dilation rates $\{1, 2, 3\}$ to $F_{\mathrm{att}}$. The resulting feature maps $F_{\mathrm{ms}}$ are fused via $1 \times 1$ convolution as well:

$$F_{\mathrm{ms}} = \mathrm{Conv}_{1\times1}\big([\mathrm{Dilate}_d(F_{\mathrm{att}})]_{d=1}^{3}\big). \tag{5}$$

**Smoothing.** To stabilize the early stages of training, we apply a depthwise convolution initialized with a Gaussian low-pass kernel $\mathrm{Gauss}(\cdot)$. The final enhanced feature map $\varphi_l^g$ is obtained by adding the attention-enhanced feature:

$$\varphi_l^g = F_{\mathrm{att}} + \mathrm{Gauss}\big(F_{\mathrm{ms}}\big). \tag{6}$$

FGE is applied to the levels $l \in \{2, 4, 8\}$. Its output $\varphi_l^g$ retains the same shape as $\varphi_l$ and is directly consumed by the hierarchical matcher. For consistency of notation, we hereafter still denote $\varphi_l^g$ as $\varphi_l$.

In addition, how the proposed FGE affects cross-modal feature matchability is discussed in the Supplement. Interestingly, we observe that the feature gradient enhancement leads to a notable decrease in cosine similarity between corresponding SAR and optical features, suggesting that high similarity is not a necessary condition for effective cross-modal matching.

## Global-Local Affine-Flow Matcher (GLAM)

Most existing multimodal remote sensing image matching approaches rely on regressing a single geometric transformation to align cross-modal inputs. However, due to the inherent modality gap between SAR and optical images, relying solely on either global or local transformations often proves insufficient.

GLAM estimates two complementary deformation fields, an affine field $\hat{W}_l^{\mathrm{a}}$ and a flow field $\hat{W}_l^{\mathrm{f}}$, which further correct local misalignments remaining from the previous level. These two are computed using shared features but separate decoders shown in Figure 3.

**Affine-Flow Warping.** Let $\varphi_l^{\mathrm{o}}$ and $\varphi_l^{\mathrm{s}}$ denote the optical and SAR feature maps at level $l$. To facilitate hierarchical refinement, the SAR features are first warped using the estimated transformation from the previous scale $\hat{W}_l^{\mathrm{prev}}$, yielding $\tilde{\varphi}_l^{\mathrm{s}}$. The aligned pair $[\varphi_l^{\mathrm{o}}, \tilde{\varphi}_l^{\mathrm{s}}]$ is then passed to two decoders:

$$\hat{W}_l^{\mathrm{a}} = \mathcal{A}([\varphi_l^{\mathrm{o}}, \tilde{\varphi}_l^{\mathrm{s}}]), \quad \hat{W}_l^{\mathrm{f}} = \mathcal{F}([\varphi_l^{\mathrm{o}}, \tilde{\varphi}_l^{\mathrm{s}}]), \tag{7}$$

where $\mathcal{A}(\cdot)$ and $\mathcal{F}(\cdot)$ denote the affine and flow regressors. The outputs $\hat{W}_l^{\mathrm{a}}, \hat{W}_l^{\mathrm{f}}$ are used to update the current transformation $\hat{W}_l$.

**Multi-stage matching pipeline.** The matching proceeds through three stages:

**Coarse Initialization ($l = 16$):** At the coarsest level, only the affine regressor is used, allowing the model to estimate an initial warp $\hat{W}_{16}$ based on global affine. This serves as a robust starting point for finer-scale refinement.

**Coupled Registration ($l = 8, 4$):** At intermediate levels, both affine and flow regressors are active. SAR features are warped using the previous warp $\hat{W}_l^{prev}$ then concatenated with optical features. GLAM simultaneously predicts $\hat{W}_l^{\mathrm{a}}, \hat{W}_l^{\mathrm{f}}$ at this stage, then updates $\hat{W}_l$ using $\hat{W}_l^{\mathrm{f}}$. These two deformation fields are jointly optimized through a consistency loss $\mathcal{L}_{\text{cons}}$, which encourages mutual guidance between global alignment and fine-grained warping. Details of the loss are provided in the Loss Functions section.

**Refinement ($l = 2, 1$):** At this stage, large misalignment has been resolved, allowing the following stages to focus on fine-grained refinement. Thus, the affine regressor is omitted. Only the flow is estimated to capture residual displacements. In addition, a pixel-wise certainty map $\hat{p}$ is predicted as a certainty logit used for the final refinement step.

Thus, by incorporating both affine and flow transformations within a coarse-to-fine framework, GLAM achieves global structural consistency and local precision.

## Loss Functions

As part of SOMA's hierarchical design, our loss function specifically addresses the challenges inherent in coarse-to-fine registration. In addition to the conventional RMSE-based warp loss, our optimization objective incorporates four contrapuntally designed terms to enhance certainty awareness, enforce multi-scale progressive refinement, facilitate complementary affine-flow interactions, and mitigate global drift introduced by affine transformations.

The final loss $\mathcal{L}$ is defined as:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{warp}}}_{\text{warp loss}} + \underbrace{\lambda\mathcal{L}_{\text{cons}}}_{\text{affine-flow consistency}} + \underbrace{\alpha_{\text{c}}\mathcal{L}_{\text{cert}}}_{\text{certainty supervision}} + \underbrace{\alpha_{\text{d}}\mathcal{L}_{\text{delta}}}_{\text{residual supervision}} + \underbrace{\alpha_{\text{u}}\mathcal{L}_{\text{uni}}}_{\text{patch-wise uniformity}}, \tag{8}$$

where $\alpha_{\text{c}} = \alpha_{\text{d}} = \alpha_{\text{u}} = 0.1$, and $\lambda = 0.5$.

**Warp Loss $\mathcal{L}_{\text{warp}}$.** We follow standard practice by minimizing the pixel-wise RMSE between the predicted deformation field $\hat{W}_1$ and the ground truth warp $W$:

$$\mathcal{L}_{\text{warp}} = \text{RMSE}\left(\hat{W}_1, W_{gt}\right). \tag{9}$$

This term anchors the overall registration quality. However, its performance is insufficient when faced with large modality discrepancies.

**Affine-Flow Consistency $\mathcal{L}_{\text{cons}}$.** GLAM simultaneously estimates affine and flow fields, with the goal of leveraging their complementary strengths. To achieve this interaction and mutual enhancement, we minimize the discrepancy between their predictions through a consistency loss function:

$$\mathcal{L}_{\text{cons}} = \sum_{l\in\{8,4\}} \text{RMSE}\left(\hat{W}_l^{\mathrm{f}}, \hat{W}_l^{\mathrm{a}}\right). \tag{10}$$

**Certainty Supervision $\mathcal{L}_{\text{cert}}$.** To address the increased modality differences arising at finer scales—particularly evident as the resolution approaches the original input—we introduce a supervision term on the predicted certainty logits $\hat{p}$, encouraging them to reflect the actual matching quality:

$$\mathcal{L}_{\text{cert}} = \frac{1}{|\Omega|}\sum_{\mathbf{x}\in\Omega}\left(\sigma(\hat{p}(\mathbf{x})) - e^{-\|\hat{W}_1(\mathbf{x}) - W_{gt}(\mathbf{x})\|_2}\right)^2, \tag{11}$$

where $\Omega$ denotes the image domain and $\sigma(\cdot)$ is the sigmoid function.

**Multi-Scale Residual Supervision $\mathcal{L}_{\text{delta}}$.** To stabilize hierarchical refinement under modality-induced distortions, we supervise the residual deformation at each pyramid level $l \in \{8, 4, 2, 1\}$, defined as the difference between ground-truth and the accumulated prediction from coarser levels:

$$\mathcal{L}_{\text{delta}} = \sum_{l\in\{8,4,2,1\}} w_l \left\|\hat{W}_l - (W_{gt} - \hat{W}_l^{\text{prev}})\right\|_1, \tag{12}$$

with weights $w_8 = 0.125, w_4 = 0.25, w_2 = 0.5, w_1 = 1$. $\hat{W}_l$ and $\hat{W}_l^{\text{prev}}$ are upsampled to the size of $W_{gt}$.

**Patch-Wise Uniformity $\mathcal{L}_{\text{uni}}$.** Although affine transformations offer valuable global guidance for modeling coarse alignment, in some cases involving complex local distortions, they can inadvertently introduce spatial bias across certain regions. To mitigate this effect and ensure more balanced local accuracy, we divide the predicted warp $\hat{W}_1$ and the ground truth $W_{gt}$ into four non-overlapping patches $\{\hat{W}_1^i, W_{gt}^i\}, i \in \{1, 2, 3, 4\}$, and minimize the standard deviation (Std) of patch-wise registration error:

$$\mathcal{L}_{\text{uni}} = \text{Std}(\left\{\text{RMSE}\left(\hat{W}_1^i, W_{gt}^i\right)\right\}_{i=1}^4). \tag{13}$$

Our loss formulation extends beyond conventional objectives by explicitly addressing the structural, hierarchical, and spatial challenges in SAR-Optical matching. This interplay of supervision leads to more stable and accurate convergence, as validated by our experiments.

# Experiments

## Datasets

We evaluate SOMA's registration performance on two public datasets: SEN1-2(Schmitt, Hughes, and Zhu 2018) dataset and GFGE_SO(Yang et al. 2025) dataset. The SEN1-2 dataset encompasses diverse land cover types at 10-meter spatial resolution. Since the original image pairs are not strictly aligned, we use the finely registered version provided by SOPatch(Xu et al. 2023). GFGE_SO dataset contains multi-temporal, multi-satellite SAR-Optical pairs with a higher spatial resolution of 5m. Compared to SEN1-2, it

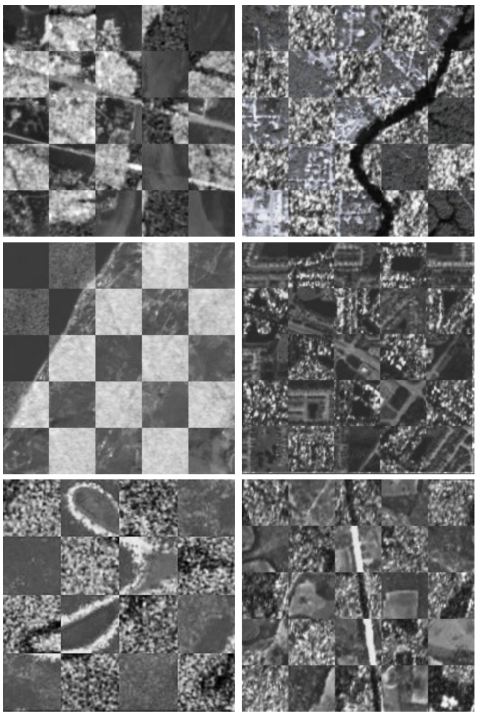

Figure 4: Registration Results of SOMA on the SEN1-2 (Left) and the GFGE_SO dataset (Right).

| Method | CMR@Threshold (%) ↑ | | | | |
|---|---|---|---|---|---|
| | 1px | 2px | 3px | 4px | 5px |
| MI (Suri and Reinartz 2009) | 44.03 | 53.45 | 57.16 | 58.78 | 60.01 |
| CFOG (Ye et al. 2019) | 46.02 | 59.73 | 64.87 | 66.58 | 67.53 |
| DDFN (Zhang et al. 2020) | 49.64 | 65.63 | 72.76 | 74.95 | 76.00 |
| FFT U-Net (Fang et al. 2021) | 54.87 | 73.43 | 79.90 | 83.52 | 86.56 |
| MoPSI (Liu, Qi, and Peng 2023) | 60.87 | 81.71 | 90.18 | 92.27 | 93.60 |
| OSMNet (Zhang et al. 2021) | 62.58 | 83.04 | 91.13 | 93.13 | 94.46 |
| RMSO-ConvNeXt (Yang et al. 2025) | 68.67 | 86.18 | 92.46 | 94.46 | 95.60 |
| SO-ConvNeXt (Yang et al. 2025) | <u>74.38</u> | <u>90.27</u> | <u>94.36</u> | <u>95.70</u> | <u>96.27</u> |
| **SOMA** | **86.67** | **94.50** | **95.97** | **97.56** | **98.78** |
| | *(+12.29)* | *(+4.23)* | *(+1.61)* | *(+1.86)* | *(+2.51)* |

Table 1: Comparison of correctly match rate (CMR) at varying pixel thresholds on the SEN1-2 dataset.

presents more severe spectral discrepancies and radiometric noise, posing greater challenges for cross-modal registration. We further test SOMA on two external datasets for generalizability analysis, which are not used during training: WHU-SEN-City(Wang et al. 2019) and OSdataset(Xiang et al. 2020).

To replicate realistic registration scenarios, random geometric transformations are applied to the SAR images. Specifically, following the previous protocols, we introduce translations of up to 32 pixels, scale variations of 0.2, and rotations within ±5°. In ablation, we further extend the range by increasing the maximum translation to 50 pixels and the rotation span to ±20°.

## Implementation Details

We adopt an end-to-end training strategy, where the DINOv2 backbone is kept frozen. All other components are trained from scratch. Training is conducted for 100 epochs with a batch size of 4, using the AdamW optimizer with a learning rate of $5 \times 10^{-5}$. A warm-up phase of 5 epochs is applied at the beginning of training. All experiments and evaluations are performed on two NVIDIA RTX 4090 GPUs.

## Comparison with State-of-the-Art Methods

We compare SOMA against several representative methods, including traditional methods and deep learning-based methods known for their fine-grained performance. The traditional side includes MI (Suri and Reinartz 2009) and CFOG (Ye et al. 2019). And, on the learning-based side, we select a set of works—DDFN (Zhang et al. 2020), FFT U-Net (Fang et al. 2021), OSMNet (Zhang et al. 2021), MoPSI (Liu, Qi, and Peng 2023), SO-ConvNeXt and RMSO-ConvNeXt (Yang et al. 2025)—as current performance upper bounds.

We adopt the Correctly Matching Rate (CMR) as a primary evaluation metric to assess the robustness and precision of registration methods. CMR computes the proportion of image pairs whose registration error remains below a predefined threshold, thereby reflecting the method's ability to consistently achieve acceptable alignment.

Table 1 summarizes the CMR at pixel thresholds ranging from 1-5px on the SEN1-2 dataset. Our proposed SOMA consistently achieves the best performance across all thresholds. At the most stringent criterion, SOMA attains 86.67%, surpassing the previous state-of-the-art method by 12.29%. This advantage persists as the matching threshold relaxes, with SOMA reaching 98.78% at 5px, outperforming by 2.51%. As shown in Table 2, on the more challenging GFGE_SO dataset, which features higher-resolution images and stronger spectral discrepancies, SOMA also surpasses the strongest baseline across all thresholds, with improvements peaking at 18.50% under the 1px threshold.

The consistent and significant improvements across all thresholds demonstrate SOMA's superior performance in pixel-level registration. Some examples of the results are shown in Figure 4.

## Ablation Study

We perform an ablation study on the SEN1-2 dataset by progressively incorporating the DINOv2, FGE, and GLAM modules into a shared baseline. As shown in Table 3, we report results in terms of CMR, as well as the registration error $R_{\text{avg}}$, calculated as the mean RMSE across all test pairs.

Individually, the FGE module produces the most substan-

| Method | CMR@Threshold (%) ↑ | | | | |
|---|---|---|---|---|---|
| | 1px | 2px | 3px | 4px | 5px |
| MI (Suri and Reinartz 2009) | 30.21 | 40.36 | 44.25 | 46.74 | 48.33 |
| CFOG (Ye et al. 2019) | 37.98 | 46.24 | 55.01 | 59.18 | 61.87 |
| DDFN (Zhang et al. 2020) | 43.05 | 52.31 | 59.68 | 64.26 | 67.05 |
| FFT U-Net (Fang et al. 2021) | 47.93 | 57.49 | 63.07 | 67.05 | 71.23 |
| MoPSI (Liu, Qi, and Peng 2023) | 54.97 | 69.94 | 75.02 | 82.88 | 89.16 |
| OSMNet (Zhang et al. 2021) | 56.40 | 72.53 | 79.00 | 87.26 | 91.84 |
| SO-ConvNeXt (Yang et al. 2025) | 58.69 | 75.51 | 82.78 | 92.05 | 95.83 |
| RMSO-ConvNeXt (Yang et al. 2025) | 60.88 | 79.00 | 85.37 | 92.94 | 96.23 |
| **SOMA** | **79.38** | **88.82** | **93.97** | **96.68** | **98.42** |
| | *(+18.50)* | *(+9.82)* | *(+8.60)* | *(+3.74)* | *(+2.19)* |

Table 2: Comparison of correctly match rate (CMR) at varying pixel thresholds on the GFGE_SO dataset.

| Setup | CMR@Threshold (%)↑ | | | | | $R_{avg}$ ↓ |
|---|---|---|---|---|---|---|
| | 1px | 2px | 3px | 4px | 5px | |
| **baseline** | 28.85 | 76.28 | 86.19 | 90.83 | 93.40 | 2.58 |
| + DINO | 27.75 | 81.05 | 89.24 | 94.38 | 96.21 | 2.36 |
| | *(-1.10)* | *(+4.77)* | *(+3.05)* | *(+3.55)* | *(+2.81)* | *(-0.22)* |
| + FGE | 49.39 | 86.31 | 91.69 | 94.50 | 97.31 | 1.95 |
| | *(+20.54)* | *(+10.03)* | *(+5.50)* | *(+3.67)* | *(+3.91)* | *(-0.63)* |
| + GLAM | 42.79 | 71.15 | 84.23 | 88.75 | 92.91 | 2.25 |
| | *(+13.94)* | *(-5.13)* | *(-1.96)* | *(-2.08)* | *(-0.49)* | *(-0.33)* |
| + DINO + FGE | 52.44 | 88.88 | 94.01 | 95.23 | 96.70 | 1.77 |
| | *(+23.59)* | *(+12.60)* | *(+7.82)* | *(+4.40)* | *(+3.30)* | *(-0.81)* |
| + DINO + GLAM | 60.88 | 84.23 | 91.81 | 95.84 | 97.92 | 1.55 |
| | *(+32.03)* | *(+7.95)* | *(+5.62)* | *(+5.01)* | *(+4.52)* | *(-1.03)* |
| + FGE + GLAM | 75.43 | 87.16 | 92.91 | 95.48 | 96.70 | 1.33 |
| | *(+46.58)* | *(+10.88)* | *(+6.72)* | *(+4.65)* | *(+3.30)* | *(-1.25)* |
| **SOMA (full)** | **87.53** | **94.87** | **96.45** | **97.19** | **97.80** | **0.94** |
| | *(+58.68)* | *(+18.59)* | *(+10.26)* | *(+6.36)* | *(+4.40)* | *(-1.64)* |

Table 3: Component analysis of SOMA.

| Dataset | Resolution | Source | CMR@5px (%) | $R_{avg}$ |
|---|---|---|---|---|
| **WHU-SEN-City** | 10m | Sentinel-1<br>Sentinel-2 | 96.75 | 1.80 |
| **OSdataset** | 1m | Google Earth<br>GaoFen-3 | 92.96 | 3.37 |

Table 4: Generalization performance of SOMA.

tial improvement over the baseline, particularly under strict alignment criteria. At the 1px threshold, it increases CMR by 20.54% compared to the baseline, demonstrating the effectiveness of feature gradient enhancement in improving fine-level characteristics. The GLAM module also contributes consistent gains, especially at 1-3px, indicating the value of coarse-to-fine affine-flow modeling in capturing hierarchical geometric transformations.

When used in combination, the benefits of each module become more pronounced. FGE + GLAM already achieves strong performance, outperforming all single-module variants. Adding DINOv2 leads to further gains, confirming the advantage of semantic regularization via frozen fundamental features. The complete framework, SOMA, integrates all three components and achieves the best results, with a CMR of 87.53% at 1px and the lowest registration error. Further analyzes are provided in the Supplement.

## Generalizability Analysis

To evaluate the generalizability of SOMA, we perform cross-dataset testing using models trained solely on SEN1-2. Two additional datasets are used for evaluation. The first, WHU-SEN-City(Wang et al. 2019), is collected from the same satellite and shares the same spatial resolution as SEN1-2, but covers 32 cities in China, exhibiting different geographical distributions. The second, OSdataset(Xiang et al. 2020), is constructed from Google Earth optical images and GaoFen-3 SAR data, which feature significantly higher spatial resolution and diverse image characteristics of multiple sources. Here, we use image pairs provided by SOPatch for both datasets, consistent with the SEN1-2 setting.

The experimental results in Table 4 show that SOMA maintains strong performance on WHU-SEN-City, indicating good generalization across scenes with similar resolution and sensor characteristics, but varying geographic content. On OSdataset, despite the lack of high-resolution training supervision, SOMA still shows reasonably good performance at 5px. This suggests that SOMA retains generalizability under cross-sensor and high-resolution domain shifts, even without explicit adaptation.

## Runtime Analysis

We evaluated the runtime of SOMA in comparison to the baseline. SOMA processes each image pair with a size of $512 \times 512$ pixels in 94 ms on average, incurring just a 6.8% overhead compared to the baseline's 88 ms, while delivering significantly improved accuracy.

# Conclusion

We have proposed SOMA, a high-precision and robust framework for SAR-optical image registration. SOMA employs the Feature Gradient Enhancer (FGE) to refine deep features, guiding the model to focus on structural cues that are more conducive to establishing accurate matching. A frozen DINOv2 further stabilize coarse alignment. The Global-Local Affine-Flow Matcher (GLAM) jointly predicts affine and flow fields, enabling mutual guidance between global and local transformations and achieving accurate dense matching in a coarse-to-fine manner. Experimental results demonstrate that SOMA delivers significant improvements in pixel-level accuracy on different datasets and generalizes well across varying scenarios.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62476220 and 61971356, as well as the Natural Science Basic Research Program of Shaanxi Province under Grant No. 2024JC-DXWT-07.

## References

Bebis, G.; Georgiopoulos, M.; da Vitoria Lobo, N.; and Shah, M. 1999. Learning affine transformations. *Pattern Recognition*, 32(10): 1783–1799.

Cai, Y.; Yin, F.; Hammou, D.; and Mantiuk, R. 2025. Do computer vision foundation models learn the low-level characteristics of the human visual system? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 20039–20048.

Darcet, T.; Oquab, M.; Mairal, J.; and Bojanowski, P. 2024. Vision Transformers Need Registers. In *International Conference on Learning Representations (ICLR)*.

Dellinger, F.; Delon, J.; Gousseau, Y.; Michel, J.; and Tupin, F. 2014. SAR-SIFT: a SIFT-like algorithm for SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 53(1): 453–466.

Dong, S.; Wang, L.; Du, B.; and Meng, X. 2024. ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208: 53–69.

Dong, Y.; Jiao, W.; Long, T.; Liu, L.; He, G.; Gong, C.; and Guo, Y. 2019. Local Deep Descriptor for Remote Sensing Image Feature Matching. *Remote Sensing*, 11(4).

Edstedt, J.; Sun, Q.; Bökman, G.; Wadenbäck, M.; and Felsberg, M. 2024. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19790–19800.

Fang, Y.; Hu, J.; Du, C.; Liu, Z.; and Zhang, L. 2021. SAR-optical image matching by integrating Siamese U-Net with FFT correlation. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Hong, Y.; Leng, C.; Zhang, X.; Peng, J.; Jiao, L.; and Basu, A. 2022. Max-index based local self-similarity descriptor for robust multi-modal image registration. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Jiang, H.; Karpur, A.; Cao, B.; Huang, Q.; and Araujo, A. 2024. OmniGlue: Generalizable Feature Matching with Foundation Model Guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19865–19875.

Jiang, X.; Ma, J.; Xiao, G.; Shao, Z.; and Guo, X. 2021. A review of multimodal image matching: Methods and applications. *Information Fusion*, 73: 22–71.

Li, J.; Wen, Y.; and He, L. 2023. SCConv: Spatial and Channel Reconstruction Convolution for Feature Redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6153–6162.

Li, L.; Han, L.; Ding, M.; and Cao, H. 2023. Multimodal image fusion framework for end-to-end remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–14.

Liu, Y.; Qi, H.; and Peng, S. 2023. Optical and SAR images matching based on phase structure convolutional features. *IEEE Geoscience and Remote Sensing Letters*, 20: 1–5.

Melekhov, I.; Tiulpin, A.; Sattler, T.; Pollefeys, M.; Rahtu, E.; and Kannala, J. 2019. Dgc-net: Dense geometric correspondence network. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1034–1042.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research Journal*, 1–31.

Quan, D.; Wang, S.; Gu, Y.; Lei, R.; Yang, B.; Wei, S.; Hou, B.; and Jiao, L. 2022. Deep Feature Correlation Learning for Multi-Modal Remote Sensing Image Registration. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Ribeiro, A. H.; and Schön, T. B. 2021. How convolutional neural networks deal with aliasing. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2755–2759.

Schmitt, M.; Hughes, L.; and Zhu, X. 2018. The SEN1-2 dataset for deep learning in SAR-optical data fusion. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4: 141–146.

Sedaghat, A.; and Mohammadi, N. 2019. Illumination-robust remote sensing image matching based on oriented self-similarity. *ISPRS Journal of Photogrammetry and Remote Sensing*, 153: 21–35.

Suri, S.; and Reinartz, P. 2009. Mutual-information-based registration of TerraSAR-X and Ikonos imagery in urban areas. *IEEE Transactions on Geoscience and Remote Sensing*, 48(2): 939–949.

Truong, P.; Danelljan, M.; and Timofte, R. 2020. GLU-Net: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6258–6268.

Wang, L.; Xu, X.; Yu, Y.; Yang, R.; Gui, R.; Xu, Z.; and Pu, F. 2019. SAR-to-Optical Image Translation Using Supervised Cycle-Consistent Adversarial Networks. *IEEE Access*, 7: 129136–129149.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Xiang, Y.; Tao, R.; Wang, F.; You, H.; and Han, B. 2020. Automatic Registration of Optical and SAR Images Via Improved Phase Congruency Model. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13: 5847–5861.

Xiang, Y.; Wang, F.; and You, H. 2018. OS-SIFT: A robust SIFT-like algorithm for high-resolution optical-to-SAR image registration in suburban areas. *IEEE Transactions on Geoscience and Remote Sensing*, 56(6): 3078–3090.

Xu, H.; Ma, J.; Yuan, J.; Le, Z.; and Liu, W. 2022. RFNet: Unsupervised Network for Mutually Reinforcing Multi-Modal Image Registration and Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19679–19688.

Xu, H.; Yuan, J.; and Ma, J. 2023. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(10): 12148–12166.

Xu, W.; Yuan, X.; Hu, Q.; and Li, J. 2023. SAR-optical feature matching: A large-scale patch dataset and a deep local descriptor. *International Journal of Applied Earth Observation and Geoinformation*, 122: 103433.

Yang, C.; Gong, G.; Liu, C.; Deng, J.; and Ye, Y. 2025. RMSO-ConvNeXt: A Lightweight CNN Network for Robust SAR and Optical Image Matching Under Strong Noise Interference. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–13.

Yao, Y.; Zhang, Y.; Wan, Y.; Liu, X.; Yan, X.; and Li, J. 2022. Multi-modal remote sensing image matching considering co-occurrence filter. *IEEE Transactions on Image Processing*, 31: 2584–2597.

Ye, Y.; Bruzzone, L.; Shan, J.; Bovolo, F.; and Zhu, Q. 2019. Fast and robust matching for multimodal remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11): 9059–9070.

Ye, Y.; Tang, T.; Zhu, B.; Yang, C.; Li, B.; and Hao, S. 2022. A multiscale framework with unsupervised learning for remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–15.

Zhang, H.; Lei, L.; Ni, W.; Tang, T.; Wu, J.; Xiang, D.; and Kuang, G. 2020. Optical and SAR image matching using pixelwise deep dense features. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Zhang, H.; Lei, L.; Ni, W.; Tang, T.; Wu, J.; Xiang, D.; and Kuang, G. 2021. Explore better network framework for high-resolution optical and SAR image matching. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–18.

Zhang, X.; Wang, Y.; and Liu, H. 2022. Robust Optical and SAR Image Registration Based on OS-SIFT and Cascaded Sample Consensus. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Zhang, X.; Wang, Y.; Liu, J.; Wang, S.; Zhang, C.; and Liu, H. 2024. Robust Coarse-to-Fine Registration Algorithm for Optical and SAR Images Based on Two Novel Multiscale and Multidirectional Features. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–26.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; and Kong, T. 2022. iBOT: Image BERT Pre-Training with Online Tokenizer. *International Conference on Learning Representations (ICLR)*.