

RBE-Flow: Recurrent Bayesian Estimation on Feature Manifolds for Cross-Modal Registration

Mengzhu Ding¹, Xin Song¹, Xiaoke Ding¹, Hongwei Ding¹, and
Xuecong Liu^{1*}

Northeastern University, China

dingmengzhu@mails.neu.edu.cn, liuxuecong@qhd.neu.edu.cn

Abstract. Cross-modal image registration is essential for multi-sensor perception but remains fundamentally challenging due to severe non-linear radiometric discrepancies and geometric distortions. Existing deterministic matching methods lack uncertainty awareness, struggling to navigate the resulting highly non-convex optimization landscape and frequently accumulating errors in ambiguous regions. In this paper, we propose RBE-Flow, a novel framework that reformulates dense cross-modal flow estimation as a closed-loop recurrent Bayesian estimation problem on learned feature manifolds. Diverging from standard feed-forward regression, RBE-Flow establishes a robust self-correcting mechanism by deeply coupling feature-metric non-linear optimization with probabilistic state updates. Specifically, a Recurrent Manifold Optimization (RMO) block iteratively generates flow observations and their associated uncertainties, which are then optimally assimilated into the prior state via an Uncertainty-Adaptive Probabilistic Update (UAPU) using deterministic sigma-point projection. Crucially, the resulting calibrated posterior covariance is fed back to adaptively regularize the damping of subsequent optimization steps, allowing the system to modulate its convergence based on predictive confidence. To ensure stable probabilistic training, we introduce a hybrid supervision scheme featuring a geometry-aware rectified NLL loss that structurally prevents variance collapse. Extensive experiments on challenging OSdataset, WHU-OPT-SAR, and RoadScene benchmarks demonstrate that RBE-Flow consistently achieves state-of-the-art performance, outperforming existing methods by a significant margin, particularly under strict sub-pixel criteria. Project page: <https://github.com/NEU-Liuxuecong/RBE-Flow>

Keywords: Cross-modal image registration · Recurrent Bayesian Estimation · Uncertainty Estimation · Dense Correspondence · Manifold Optimization

1 Introduction

Cross-modal image registration aims to establish reliable geometric correspondences between images acquired by heterogeneous sensors (e.g., optical-SAR,

* Corresponding author.

visible-infrared) [3, 19, 24, 58]. This capability supports a wide range of downstream applications, including remote sensing analytics [26, 33, 38, 42], 3D scene understanding [6, 20], and autonomous navigation [9], where complementary modalities provide robustness under challenging illumination or weather conditions. Compared to alignment within the same modality, cross-modal registration is fundamentally more challenging because the underlying sensing mechanisms induce non-linear radiometric discrepancies (e.g., intensity reversals, speckle noise, and spectral gaps) [52, 59]. These discrepancies violate the assumptions behind standard photometric alignment and often weaken local visual evidence.

In particular, for optical-SAR pairs, intensity reversals and speckle noise often lead to inconsistent feature statistics, while for visible-infrared pairs, the spectral gap further reduces cross-modal feature consistency [35, 49]. Beyond appearance gaps, cross-modal registration frequently involves substantial geometric variations caused by viewpoint changes, scale differences, or local deformations [25, 40, 62]. These spatial deformations complicate pixel-level alignment and render traditional feature-based or intensity-driven signals unreliable [4, 48]. These coupled factors give rise to a rugged, highly non-convex matching landscape riddled with local minima. Existing deterministic regression or standard optimization methods struggle to maintain stable refinement, succumbing to overconfident erroneous updates when faced with ambiguous cues [13, 60].

To address these challenges, we propose a novel multi-modal registration network, RBE-Flow, that formulates cross-modal flow estimation as a closed-loop recurrent Bayesian estimation on feature manifolds. RBE-Flow starts from a coarse global initialization and performs refinement by minimizing a feature-metric non-linear least-squares energy, producing an iterative flow observation with associated uncertainty. We then apply an uncertainty-adaptive Bayesian posterior update via deterministic sigma-point projection, which fuses the new likelihood evidence with the prior belief and feeds back calibrated the posterior covariance to regulate subsequent optimization.

Our contributions are summarized as follows:

- We propose RBE-Flow, a unified closed-loop recurrent Bayesian estimation framework that integrates mathematical minimization with empirically derived uncertainty ensuring the robustness of the overall framework..
- We introduce the RMO block to generate local likelihood evidence by solving a feature-metric non-linear least-squares problem on the manifold, with uncertainty-aware damping to stabilize updates under cross-modal noise.
- We devise UAPU, an uncertainty-adaptive Bayesian posterior update that fuses the prior belief with RMO’s observation and uncertainty via deterministic sigma-point projection, yielding a Minimum Mean Square Error (MMSE)-optimal fusion gain and posterior covariance to down-weight unreliable evidence in ambiguous, occluded regions.
- We design a hybrid supervision strategy with an L_1 loss for coarse initialization and a geometry-aware rectified NLL for recurrent refinement, where a

dynamic variance lower bound prevents variance collapse and enforces uncertainty calibration against geometric error.

2 Related Work

2.1 Feature Matching and Cross-Modal Registration

Cross-modal registration entails establishing robust correspondences under severe radiometric differences, evolving through three paradigms:

Hand-Crafted Matching. Classical methods employ modality-invariant descriptors (e.g., RIFT [21], OS-SIFT [51], CFOG [59]) using phase congruency or oriented gradients to handle intensity variations. Recent advanced geometric and intensity-invariant designs include POS-GIFT [12], MSG [64], MMAR [30], MIRD [34], and shape-adaptive MIRD [32]. However, these fixed descriptors inherently lack distinctiveness in low-texture regions and struggle with complex local deformations.

Sparse Learnable Matching. Deep learning shifted the focus toward data-driven sparse matching to overcome fixed descriptor limitations. SuperPoint [5] and SuperGlue [39] pioneered graph neural networks for context-aware point matching. Advances, such as LightGlue [27], OmniGlue [16], and XFeat [37], have further enhanced efficiency and generalization. However, their reliance on repeatable keypoint detection frequently becomes a critical bottleneck across modalities with massive appearance gaps.

(Semi-)Dense Learnable Matching. To solve the detection bottleneck, detector-free methods infer correspondences directly from feature maps. LoFTR [41] and its variants [2, 10] use Transformers to capture long-range dependencies for semi-dense matching. Recently, models like XoFTR [47] and RoMa [7] have further adapted this paradigm for robust cross-modal and dense feature matching. Although powerful, extending them to dense cross-modal flow remains challenging due to the highly non-convex optimization landscape.

2.2 Dense Flow Estimation and Geometric Refinement

Dense flow estimation has become a dominant registration paradigm. Pioneered by RAFT [43], architectures like GMFlow [56], GMA [18], FlowFormer [14], and DPFlow [36] utilize recurrent correlation updates. To tackle cross-modal geometric transformations, recent frameworks like GAFF [31] and GDROS [42] introduce geometry-guided correlations while CRFT [29] employs a consistent-recurrent Transformer for iterative spatial calibration. However, their implicit geometric constraints and learned updates limit robustness under large appearance gaps.

While GLU-Net [45] handles large displacements via global-local correlations, its refinement implicitly approximates optimization. Conversely, embedding explicit differentiable least-squares solvers into deep networks (e.g., OptNet [1], DROID-SLAM [44]) proves highly effective. Motivated by the underexplored

application of explicit second-order solvers to dense cross-modal flow, we incorporate a differentiable Levenberg-Marquardt (LM)-style solver for robust refinement.

2.3 Uncertainty-Aware Refinement

Estimating uncertainty is crucial for reliable registration [3]. Existing approaches, such as PDC-Net [46] and probabilistic networks [8], predict confidence maps or aleatoric uncertainty to identify unreliable regions. However, most systems treat uncertainty as a passive output for post-hoc analysis rather than as an active component of the inference process. Recent studies have revisited Bayesian filtering principles in neural inference to regulate iterative dynamics [28]. While variational Bayesian techniques (e.g., VBReg [17]) have been explored for 3D point cloud outlier rejection, their application in dense 2D spatial alignment is hindered by the highly non-convex nature of image matching. Our work aims to bridge this gap, evolving from traditional 3D outlier rejection to a fully recurrent 2D feature-manifold estimation under large appearance gaps.

3 Methodology

RBE-Flow constructs a closed-loop recurrent Bayesian estimation framework, aiming to calculate feature flows in complex cross-modal scenarios. We formulate flow estimation as a dynamic state estimation problem rather than a direct regression. The overall pipeline begins with a multi-scale feature encoder to provide hierarchical semantic representations for subsequent optimization. The core recurrent process consists of two coupled modules: RMO, which generates the flow observation by solving a local non-linear optimization problem on the feature manifold; UAPU, which assimilates the observation with the prior belief to compute the posterior distribution. The posterior uncertainty output by UAPU is fed back to RMO to modulate the trajectory of the next optimization, establishing a robust self-correction mechanism. An overview of our framework is shown in Fig. 1.

3.1 Multi-scale Feature Extraction (MFE)

To better capture the texture features of cross-modal image pairs, we utilize a weight-shared ResNet-based CNN encoder to extract multi-scale features at $1/2$, $1/4$, and $1/8$ scales. The $1/8$ scale features capture global semantic context and remain modality-independent. Therefore, we pass the $1/8$ scale features into the Iteration Initialization Module to generate a robust initial flow. The features at $1/2$ and $1/4$ scales retain rich spatial details and texture information, which effectively compensate for the detail loss in deeper features.

These multi-scale features constitute the foundation of the system: coarse-scale features are used for the initialization stage, while fine-scale features serve as local observation evidence in the subsequent recurrent Bayesian inference.

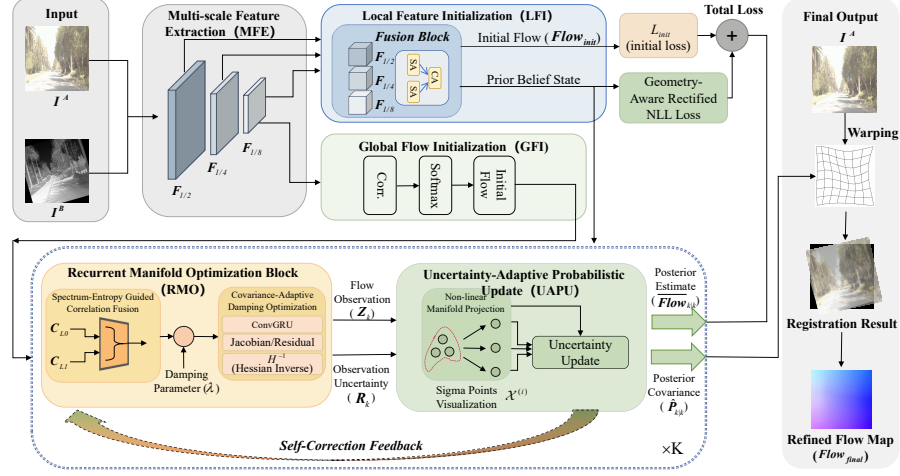


Fig. 1: Overview of the proposed RBE-Flow architecture. RBE-Flow formulates the alignment task as a closed-loop recurrent Bayesian estimation on feature manifolds. (a) **MFE**. A weight-shared CNN encoder extracts features at 1/2, 1/4, 1/8 scale from the input image pair. (b) **LFI**. A combination of SA and CA mechanisms fuses the features from different scales. (c) **GFI**. In this module, we use 4D correlation volume on 1/8 scale feature to generate the initial flow field. (d) **RMO**. This module iteratively generates flow residuals by solving a damped feature-metric least-squares problem, where the damping factor is adaptively regularized by the recirculated posterior uncertainty. (e) **UAPU**. This module propagates deterministic Sigma points through the manifold to compute an MMSE-optimal update. It assimilates optimization-derived likelihoods into the prior state, while updating the posterior covariance to guide subsequent RMO.

3.2 Iteration Initialization Module

Before entering the recurrent state estimation loop, constructing a reliable Prior Belief State is paramount. In this module, we aim to initialize both the feature representations and the flow field.

Local Feature Initialization (LFI). To generate features containing both high-level semantics and rich low-level textures, we design a multi-scale feature fusion strategy. We fuse the 1/8 scale and 1/4 scale features via self-attention (SA) to inject global context into mid-level representations. Subsequently, cross-image Cross-Attention (CA) is applied to enhance cross-modal correspondences. This process is repeated with the 1/2 scale features, yielding a distilled feature manifold that preserves high-resolution structural evidence for the likelihood computation.

Global Flow Initialization (GFI). Although ground-truth geometric displacements in cross-modal datasets are often synthesized via global affine matrices, we deliberately adopt a dense, pixel-wise flow formulation. Directly estimat-

ing low-dimensional global parameters makes the regression highly sensitive to cross-modal noise; a minuscule variance in parameter space inevitably triggers a drastic amplification of geometric errors across peripheral regions. By modeling pixel-wise dense correspondences on the manifold instead, our framework robustly absorbs local radiometric anomalies and mitigates error propagation. So to provide a robust initial estimation for subsequent iterative refinement, we compute the initial flow utilizing the $1/8$ scale features $\mathbf{F}_0$ and $\mathbf{F}_1$. We estimate the dense correlation matrix $\mathbf{C} \in \mathbb{R}^{HW \times HW}$ between feature maps $\mathbf{F}_0$ and $\mathbf{F}_1$:

$$\mathbf{C}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{F}_0(\mathbf{u})^\top \mathbf{F}_1(\mathbf{v})}{\sqrt{d}}, \quad (1)$$

where $\mathbf{u}, \mathbf{v}$ denote pixel coordinates and d is the channel dimension.

To extract dense correspondences, we normalize the correlation into a matching probability distribution $\mathbf{P}$ via softmax, and then compute the expected coordinate $\hat{\mathbf{v}}$ on the target image:

$$P(\mathbf{u}, \mathbf{v}) = \frac{\exp(\mathbf{C}(\mathbf{u}, \mathbf{v}))}{\sum_{\mathbf{v}'} \exp(\mathbf{C}(\mathbf{u}, \mathbf{v}'))}, \quad (2)$$

$$\hat{\mathbf{v}}(\mathbf{u}) = \sum_{\mathbf{v}} P(\mathbf{u}, \mathbf{v}) \cdot \mathbf{v}. \quad (3)$$

The initial flow $\mathbf{Flow}_{init}$ is derived as:

$$\mathbf{Flow}_{init}(\mathbf{u}) = \hat{\mathbf{v}}(\mathbf{u}) - \mathbf{u}. \quad (4)$$

This initial optical flow is passed into RMO, constituting the initial prior mean for the recurrent Bayesian estimation process and marking the starting point of the closed-loop estimation.

3.3 Recurrent Manifold Optimization Block (RMO).

Based on the initialized flow, the system enters the recurrent loop. In the k -th iteration, this module serves as the observation generator of our Bayesian estimation. It solves a local non-linear optimization on the manifold to yield a flow observation $\mathbf{Z}_k$, and its associated uncertainty $\mathbf{R}_k$. The data pair $(\mathbf{Z}_k, \mathbf{R}_k)$ constitutes the complete parameters of the likelihood distribution and serves as input for probabilistic fusion in the subsequent UAPU module.

Our method is based on a core assumption: high-dimensional image features are not randomly scattered throughout the ambient space, but rather reside on a lower-dimensional non-linear manifold. Therefore, flow estimation is essentially finding the shortest path on this feature manifold that connects the corresponding features between the image pair.

Spectrum-Entropy Guided Correlation Fusion. To capture the multi-scale geometric structure of the manifold, we construct a two-level correlation volume: $\mathbf{C}_{L0}$ (local window) for high-frequency details; and $\mathbf{C}_{L1}$ (pooled features)

for low-frequency global consistency. To suppress redundant noise, we perform feature fusion via Spectrum Entropy-Guided Fusion (SGCF). First, we calculate the entropy $\mathcal{H}$ for $\mathbf{C}_{L0}$ and $\mathbf{C}_{L1}$ respectively from the correlation probability distribution $\mathbf{p}$:

$$\mathcal{H}(\mathbf{u}) = - \sum_i p_i \log p_i, \quad (5)$$

where $\mathbf{p} = \text{softmax}(\mathbf{C}_L)$.

Subsequently, adaptive fusion weights are computed through a lightweight SpecNet Φ_{Spec} :

$$[\omega_{L0}, \omega_{L1}] = \text{softmax}(\Phi_{\text{Spec}}([\mathbf{C}_{L0}, \mathbf{C}_{L1}, \mathcal{H}_{L0}, \mathcal{H}_{L1}])). \quad (6)$$

The final fused correlation :

$$\mathbf{C}_{\text{fused}} = \omega_{L0} \cdot \mathbf{C}_{L0} + \omega_{L1} \cdot \mathbf{C}_{L1}. \quad (7)$$

This mechanism enables the network to dynamically adjust its reliance on different scale information.

Covariance-Adaptive Damping Optimization (CDO). The core task of RMO is to find $\Delta \mathbf{Flow}$ to minimize the locally linearized least-squares objective function. We utilize a multi-head ConvGRU to directly predict the Jacobian $\mathbf{J}_k$, residual gradient $\mathbf{g}_k$, and damping measurement λ_k required for the subsequent LM computation. For λ , we introduce the posterior covariance from the previous timestep to adjust the damping coefficient:

$$\lambda'_k = \lambda_k + \beta \cdot \text{tr}(\mathbf{P}_{k-1|k-1}). \quad (8)$$

When the prior uncertainty is high, λ'_k increases, adaptively switching the update from a Gauss-Newton step toward a more conservative Gradient Descent. The flow increment $\Delta \mathbf{Flow}_k$ is solved via:

$$\left(\mathbf{J}_k^\top \mathbf{J}_k + \lambda'_k \mathbf{I} \right) \Delta \mathbf{Flow}_k = -\mathbf{J}_k^\top \mathbf{g}_k. \quad (9)$$

We define the observation result:

$$\mathbf{Z}_k = \hat{\mathbf{Flow}}_{k-1} + \Delta \mathbf{Flow}_k. \quad (10)$$

3.4 Uncertainty-Adaptive Probabilistic Update (UAPU)

The observations of RMO are accurate on a local cost surface, but may exhibit errors (such as occlusions and motion boundaries) from a global perspective. The core idea of UAPU is to model the flow refinement as a Non-linear Belief Propagation problem: at each iteration step, the system takes the posterior distribution of the previous timestep as the prior belief, fuses it with the likelihood evidence provided by RMO, and computes the current posterior distribution via Bayes' theorem, thereby achieving self-correction of errors.

Uncertainty-Diffused Prior Belief Propagation. We model the flow state transition as a random walk. The prior mean $\hat{\mathbf{Flow}}_{k|k-1}$ is:

$$\hat{\mathbf{Flow}}_{k|k-1} = \hat{\mathbf{Flow}}_{k-1|k-1}. \quad (11)$$

Considering that flow uncertainty exhibits significant spatial heterogeneity across different regions, we introduce QNet Ψ_Q to adaptively predict the process noise $\mathbf{Q}_k$, yielding the prior covariance $\mathbf{P}_{k|k-1}$:

$$\mathbf{P}_{k|k-1} = \mathbf{P}_{k-1|k-1} + \mathbf{Q}_k. \quad (12)$$

To ensure numerical stability and positive semi-definiteness, we apply factor $\mathbf{S}$ as the Cholesky decomposition result of $\mathbf{P}$.

Non-linear Manifold Distribution Projection. Due to the strong non-linearity of the feature manifold, performing local linearization on the mapping from state to observation space would introduce non-negligible truncation errors. We generate $2L + 1 = 5$ deterministic sampling points $\mathbf{X}^{(i)}$ around the prior mean $\hat{\mathbf{Flow}}_{k|k-1}$: 1 center point, and $2L$ perturbed points symmetrically distributed along the axes of the state manifold. These are projected into the observation space to obtain the predicted observation point set $\mathbf{Z}_k^{(i)}$:

$$\mathbf{Z}_k^{(i)} = \text{ObsNet}(\mathbf{X}_k^{(i)}), \quad (13)$$

where $i \geq 1$ denotes the i -th Sigma point. The observation network $\text{ObsNet}(\cdot)$ acts as a learnable non-linear mapping function to bridge the state space and the feature-metric observation space, which is implemented as a lightweight MLP network composed of 1×1 convolutions.

The predicted observation mean is calculated as:

$$\hat{\mathbf{Z}}_k = \sum_{i=0}^4 W_m^{(i)} \mathbf{Z}_k^{(i)}, \quad (14)$$

where $W_m^{(0)} = \frac{\lambda_{\text{scale}}}{L + \lambda_{\text{scale}}}$, $L = 2$ characterizes the spatial dimensionality of the optical flow vector, λ_{scale} acts as a scaling parameter governing the manifold spread of the generated Sigma points.

Uncertainty-Gated Posterior Belief Fusion. To optimally fuse the prior belief with the observation evidence, we compute the innovation covariance $\mathbf{P}_{zz}$ and the state-measurement cross-covariance $\mathbf{P}_{xz}$ via the weighted outer product of the sampling points:

$$\mathbf{P}_{zz} = \sum_{i=0}^4 W_c^{(i)} (\mathbf{Z}_k^{(i)} - \hat{\mathbf{Z}}_k)(\mathbf{Z}_k^{(i)} - \hat{\mathbf{Z}}_k)^\top + \mathbf{R}_k, \quad (15)$$

where $\mathbf{R}_k$ is the observation noise covariance estimated from a light ConvNet.

$$\mathbf{P}_{xz} = \sum_{i=0}^4 W_c^{(i)} (\mathbf{X}_k^{(i)} - \hat{\mathbf{Flow}}_{k|k-1})(\mathbf{Z}_k^{(i)} - \hat{\mathbf{Z}}_k)^\top. \quad (16)$$

Under the Gaussian approximation assumption of the posterior distribution, minimizing the mean square error $\text{tr}(\mathbf{P}_{k|k})$ of the posterior estimation is equivalent to solving the following unconstrained optimization for $\mathcal{K}_k$. Thus, under the MMSE criterion, we obtain the optimal Bayesian fusion gain $\mathcal{K}_k$:

$$\frac{\partial \text{tr}(\mathbf{P}_{k|k})}{\partial \mathcal{K}_k} = \mathbf{0} \implies \mathcal{K}_k = \mathbf{P}_{xz} \mathbf{P}_{zz}^{-1}. \quad (17)$$

The posterior estimate is updated following Bayes' theorem, fusing the prior and likelihood:

$$\hat{\mathbf{Flow}}_{k|k} = \hat{\mathbf{Flow}}_{k|k-1} + \mathcal{K}_k(\mathbf{Z}_k - \hat{\mathbf{Z}}_k), \quad (18)$$

and the posterior covariance are updated as:

$$\mathbf{P}_{k|k} = \mathbf{P}_{k|k-1} - \mathcal{K}_k \mathbf{P}_{zz} \mathcal{K}_k^\top. \quad (19)$$

This update restricts the information gain by contracting the prior uncertainty space.

Finally, the posterior statistical moments $(\hat{\mathbf{Flow}}_{k|k}, \mathbf{S}_{k|k})$ output by UAPU are recirculated, forming a complete closed-loop recurrent system. This feedback mechanism establishes a Bayesian inference chain in the temporal dimension, enabling the system to progressively converge the uncertain initial belief to an accurate posterior approximation of the true feature flow through iterative evidence accumulation.

3.5 Loss Function

To effectively train this recurrent probabilistic framework, we adopt a multi-stage loss, bridging the deterministic global alignment and the stochastic local refinement. We construct an Initialization Loss to guide the network for preliminary alignment at coarse resolution, and a Geometry-Aware Rectified NLL Loss.

Initialization Loss. For the initial flow $\mathbf{Flow}_{\text{init}}$ generated in GFI, we employ a simple L1 loss to supervise it. We downsample the optical flow ground truth $\mathbf{Flow}_{\text{gt}}$ to the same resolution as $\mathbf{Flow}_{\text{init}}$ and compute the L1 distance between them:

$$\mathcal{L}_{\text{init}} = \frac{1}{N} \sum_{\mathbf{p}} |\mathbf{Flow}_{\text{init}} - \mathbf{Flow}_{\text{gt}}|, \quad (20)$$

where N is the number of pixels. This loss ensures that the model has a reasonable global alignment foundation before entering the fine matching stage.

Geometry-Aware Rectified NLL Loss. The vanilla NLL objective frequently exhibits numerical instability, often leading to degenerated variance estimates; models exploit gradients by predicting practically zero variances ($\sigma \rightarrow 0$) to minimize the penalty, leading to exploding gradients. To enforce a meaningful uncertainty estimate, we introduce a Geometry-Aware Rectification. We impose

a dynamic lower bound on the predicted variance, scaling with the magnitude of the geometric error:

$$(\hat{\sigma}_k^{(j)})^2 = \text{softplus}(s_k^{(j)}) + \alpha \cdot (\hat{\mathbf{Flow}}_k^{(j)} - \mathbf{Flow}_{\text{gt}}^{(j)})^2 + \epsilon, \quad (21)$$

here, $\mathbf{S}$ is the Cholesky factor output by UAPU, i.e., the square root of the covariance matrix. This mechanism ensures that high-error regions are forced to have high uncertainty, preventing the "overconfident but wrong" pathology. The Negative Log-Likelihood loss in iteration k is:

$$\mathcal{L}_{\text{NLL}}^{(k)} = \frac{1}{2} \sum_{j \in \{x, y\}} \left(\log(\hat{\sigma}_k^{(j)})^2 + \frac{(\hat{\mathbf{Flow}}_k^{(j)} - \mathbf{Flow}_{\text{gt}}^{(j)})^2}{(\hat{\sigma}_k^{(j)})^2} \right). \quad (22)$$

Finally, we apply deep supervision on the outputs of the N_{iter} iterations, and compute their weighted sum using a decay coefficient $\gamma = 0.9$, realizing a loss function that is both result-oriented and process-oriented.

$$\mathcal{L}_{\text{NLL}} = \frac{1}{\sum_{k=1}^{N_{\text{iter}}} \gamma^{N_{\text{iter}}-k}} \sum_{k=1}^{N_{\text{iter}}} \gamma^{N_{\text{iter}}-k} \mathcal{L}_{\text{NLL}}^{(k)}, \quad (23)$$

Total Loss. Finally, we perform a weighted sum of the Initialization Loss and the Geometry-Aware Rectified NLL Loss to obtain the total loss function:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{init}} \mathcal{L}_{\text{init}} + \lambda_{\text{NLL}} \mathcal{L}_{\text{NLL}}. \quad (24)$$

where λ_{init} and λ_{NLL} are the respective weights for the two losses.

4 Experiments

4.1 Implementation Details

RBE-Flow is initialized using the publicly released XofTR pretrained weights (trained at 640 resolution), and then its entire framework is fine-tuned end-to-end. We train this framework on RoadScene [53, 55], WHU-OPT-SAR [15], and OSdataset [50], with a learning rate of 3×10^{-4} and a batch size of 8.

Datasets. We conduct experiments on OSdataset, an Optical-SAR dataset covering a wide variety of terrain types such as forests, farmlands, and rivers; WHU-OPT-SAR dataset, an Optical-SAR dataset focused on complex urban scenes; and RoadScene dataset, a RGB-IR dataset capturing street views and traffic scenarios under varying lighting conditions. RoadScene is derived from the MURF [54] through rigorous cropping and expansion to ensure robust correspondence learning. We follow the standard splits of 1172/153/288 (train/val/test) for OSdataset, 7000/700/700 for WHU-OPT-SAR dataset, 5120/640/640 for RoadScene dataset. Our data augmentation strategy introduces random scaling between [0.9, 1.1], rotations within $[-30^\circ, 30^\circ]$, and translations of up to 15 pixels.

During the training process, images are resized to 64×64 patches. Furthermore, to maintain consistent supervision across all stages, the ground-truth flow fields are directly derived from the provided pixel-wise alignments.

Evaluation Metrics. We quantify the matching efficacy using the Average End-Point Error (AEPE) and the Correct Match Rate (CMR). While AEPE computes the expected flow displacement divergence, CMR indicates the percentage of samples that successfully meet the particular error constraint. Guided by existing evaluation protocols, CMR is computed over an interval of thresholds (between 0.1 and 5 pixels) to yield a multi-angle geometric precision under different constraints.

Baselines. Inspired by MatchAnything [11] and CoMatch [61], RBE-Flow is compared with such representative methods from the following categories:

- *Sparse matching methods:* LNIFT [22], RIFT2 [21,23], HOWP [63], MSG [64]. Such handcrafted keypoint-descriptor methods provide stable alignments under structural noise. But their vulnerability to cross-modal appearance gaps and large-scale geometric shifts severely limits their applicability.
- *Semi-dense matching methods:* XoFTR [47]. Serving as a prime example of contemporary Transformer-driven, detector-free architectures, its matching capability noticeably deteriorates when faced with severe textural discrepancies across modalities.
- *Dense matching methods:* RAFT [43], GMFlow [57], ADRNet [52], GDROS [42]. Such methods generate dense pixel-level flows. However, general flow estimators struggle profoundly with massive modality gaps, whereas adapted models (e.g., GDROS and ADRNet) remain restricted by domain-specific priors and low-resolution outputs.

4.2 Experiment Results

We compare RBE-Flow with 9 representative handcrafted, sparse, semi-dense, and dense correspondence methods and fine-tuned learning-based methods on the three datasets respectively. All of the experiments show our RBE-Flow achieves the best result even under strict thresholds. This demonstrates the effectiveness of our method in cross-modal registration tasks.

Results on OSdataset. As shown in Table 1 and Fig. 3(a), our RBE-Flow demonstrates highly competitive performance on the OSdataset. Compared to the previous state-of-the-art method ADRNet, our approach significantly reduces the AEPE from $1.41px$ to $0.77px$, achieving a remarkable 45.4% error reduction. In terms of CMR, RBE-Flow has the highest accuracy at both normal and fine thresholds, achieving 78.5% and 57.3% at $1px$ and $0.7px$, respectively. Although our method ranks second at the most stringent threshold of $0.3px$ (11.6%) trailing slightly behind GDROS (16.6%), the qualitative line chart in Fig. 3(a) indicates that RBE-Flow consistently maintains a superior overall trend across a continuous range of error tolerances.

Results on WHU-OPT-SAR. In this dataset, our method shows the most prominent accuracy. RBE-Flow outperforms all methods in all thresholds. Specifically, it achieves an AEPE of 0.53, nearly a $3\times$ improvement over the second-best

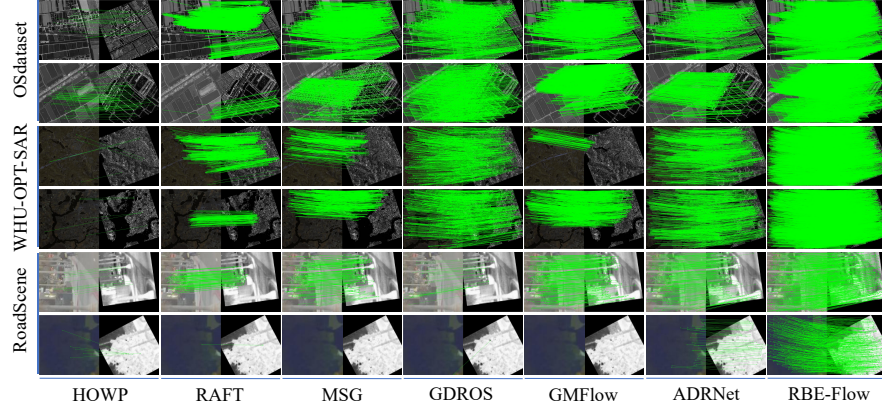


Fig. 2: Inlier correspondence visualization with state-of-the-art methods.

For each image pair, we uniformly sample 5000 candidate correspondences, and only the matches with registration error less than 2 pixels are visualized. As illustrated by the inlier visualizations, RBE-Flow consistently achieves the most comprehensive and geometrically stable correspondences across all tested benchmarks. The dense matching results underscores the efficacy of our recurrent Bayesian framework in resolving profound modality gaps and structural distortions.

ADNet (1.43). For the CMR metric, while traditional methods (e.g., HOWP, RIFT2) and some deep learning approaches completely fail on this challenging dataset, our RBE-Flow achieves extreme robustness with a CMR of 91.0% at $1px$ and 80.0% at $0.7px$ (outperforming ADNet’s 52.9% and 33.0%). Fig. 3(b) visually confirms this dominance, where the RBE-Flow retains a high matching ratio even when the threshold becomes extraordinarily strict.

Results on RoadScene. RBE-Flow achieves the best overall performance on the RoadScene dataset. As presented in Table 1, it shows the state-of-the-art AEPE of 0.49px. The steep and rapid decay of the baseline curves in Fig. 3(c) further underscores the difficulty of this dataset, yet RBE-Flow remains notably stable against varying error constraints.

Overall Discussion. The series of experiments show the two main advantages of RBE-Flow: strong cross-domain generalization and high sub-pixel accuracy. While recent baselines such as ADNet and RAFT perform reasonably well on the OSdataset, they both struggle with the WHU-OPT-SAR and RoadScene dataset. Moreover, as shown in the line charts (Fig. 3), existing methods share a common limitation: their performance drops sharply when the error threshold is strictly reduced ($t < 1px$). RBE-Flow largely mitigates this issue, maintaining a robust matching ratio even under tight tolerances. This stability implies that our method provides more reliable fine-grained alignments across varying real-world conditions.

Table 1: Quantitative comparison on OSdataset, WHU-OPT-SAR, and RoadScene datasets. AEPE and CMR at thresholds $t = \{3, 1, 0.7, 0.3\}$. Best results are in **bold**, and second-best are underlined.

Methods	OSdataset					WHU-OPT-SAR					RoadScene				
	AEPE ↓	CMR ↑ (%)				AEPE ↓	CMR ↑ (%)				AEPE ↓	CMR ↑ (%)			
		@3px	@1px	@0.7px	@0.3px		@3px	@1px	@0.7px	@0.3px		@3px	@1px	@0.7px	@0.3px
HOWP [63] ISPRS'23	20.39	16.7	0.3	0.0	0.0	157.14	0.7	0.0	0.0	0.0	16.67	15.0	0.2	0.1	0.0
LNIFT [22] TGRS'22	46.98	31.3	0.3	0.0	0.0	167.00	3.0	0.0	0.0	0.0	28.89	0.8	0.1	0.0	0.0
MSG [64] TGRS'25	44.50	35.5	0.3	0.0	0.0	120.97	12.0	0.1	0.0	0.0	43.59	6.5	0.0	0.0	0.0
RIFT2 [21] TIP'20	29.83	23.3	0.7	0.3	0.0	214.08	1.1	0.0	0.0	0.0	29.45	0.5	0.0	0.0	0.0
GMFlow [57] TPAMI'23	5.13	38.9	2.1	0.3	0.0	3.57	27.1	2.7	0.6	0.0	4.42	44.8	0.5	0.1	0.0
XoFTR [47] CVPR'24	5.41	6.9	0.0	0.0	0.0	18.05	4.3	0.0	0.0	0.0	6.00	0.4	0.0	0.0	0.0
RAFT [43] ECCV'20	2.89	67.0	14.0	7.0	1.3	2.07	77.3	25.4	14.6	3.0	1.79	85.9	31.3	17.9	4.0
ADRNNet [52] TGRS'24	<u>1.41</u>	<u>92.5</u>	<u>50.5</u>	31.9	7.5	<u>1.43</u>	<u>91.7</u>	<u>52.9</u>	<u>33.0</u>	7.7	0.98	95.1	71.3	53.1	15.6
GDROS [42] TGRS'25	1.51	78.0	48.8	<u>36.2</u>	16.6	2.87	70.9	28.7	19.8	<u>8.5</u>	<u>0.68</u>	<u>98.5</u>	<u>79.0</u>	<u>65.1</u>	<u>32.4</u>
RBE-Flow (ours)	0.77	99.7	78.5	57.3	<u>11.6</u>	0.53	99.9	91.0	80.0	24.0	0.49	99.9	90.8	85.8	37.7

4.3 Ablation Study

To validate the effectiveness of our proposed recurrent Bayesian estimation framework, we ablate the core components on the same dataset with Sec.4.2. We define our baseline (A) as the multi-scale CNN encoder appended with LFI, and calculated following an effective, coarse-to-fine, semi-dense point matching. Then we gradually add GFI, our loss function (loss), RMO, and UAPU. Results are shown in Table 2.

Table 2: Ablation experiments on three datasets. AEPE and CMR at thresholds $t = \{1, 0.7, 0.3\}$. Best results are in **bold**, and second-best are underlined.

Method	OSdataset				WHU-OPT-SAR				RoadScene			
	AEPE ↓	CMR ↑ (%)			AEPE ↓	CMR ↑ (%)			AEPE ↓	CMR ↑ (%)		
		@1px	@0.7px	@0.3px		@1px	@0.7px	@0.3px		@1px	@0.7px	@0.3px
(1) A+L1	1.32	22.9	2.8	0.0	5.52	2.5	1.2	0.0	0.69	87.4	76.0	1.0
(2) A+L1+GFI	0.98	61.8	32.6	0.3	1.10	47.3	13.9	0.3	0.64	87.3	67.0	13.0
(3) A+GFI+loss	0.91	66.7	36.8	0.7	0.76	82.9	58.6	1.0	0.59	89.0	78.9	12.6
(4) A+GFI+loss+RMO	<u>0.84</u>	<u>71.9</u>	<u>49.0</u>	<u>3.1</u>	<u>0.61</u>	<u>89.3</u>	<u>75.4</u>	<u>11.9</u>	<u>0.54</u>	<u>89.8</u>	<u>81.4</u>	<u>26.9</u>
(5) A+GFI+loss+RMO+UAPU	0.77	78.5	57.3	11.6	0.53	91.0	80.0	24.0	0.49	90.8	85.8	37.7

(1) **Baseline (A+L1).** The baseline performs poorly when the CMR within 1px, on the WHU-OPT-SAR dataset, it results in an AEPE of 5.52px and low accuracy at strict sub-pixel thresholds (0.7px is 1.2%), showing its limitation to estimate reliable cross-modal fine-grained correspondences.

(2) **A+L1+GFI.** Adding the GFI module provides a reliable initial flow by computing dense feature correlations, which helps resolve large spatial displacements. As a result, the performance improves significantly. The AEPE on the WHU-OPT-SAR dataset drops from 5.52 to 1.10, and the CMR(1px) on the OSdataset increases from 22.9% to 61.8%. This shows that a global correlation step is necessary to provide a good starting point for subsequent refinement.

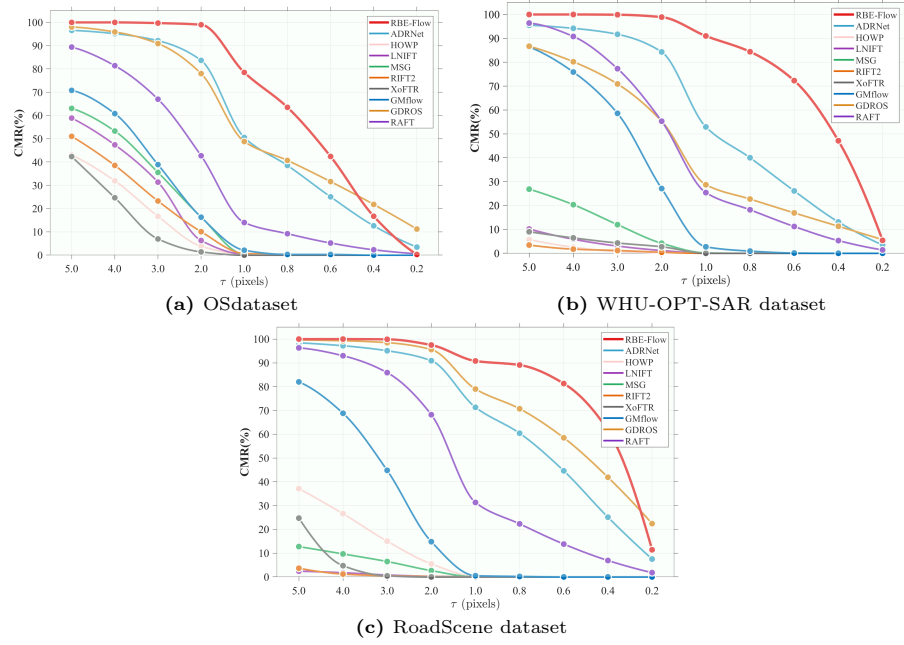


Fig. 3: CMR curves under varying thresholds on (a) OSdataset, (b) WHU-OPT-SAR, and (c) RoadScene. We compare RBE-Flow with representative hand-crafted, sparse, and (semi-)dense matching methods by evaluating the CMR across a range of pixel thresholds. RBE-Flow consistently achieves the highest CMR under both strict and relaxed thresholds, demonstrating its robustness to cross-modal appearance discrepancies and geometric variations.

(3) A+GFI+loss. In this step, we replace the standard L1 refinement loss with our Geometry-Aware Rectified NLL Loss. The metrics reflect this improvement: the CMR under 1px on the WHU-OPT-SAR dataset increases from 47.3% to 82.9%, demonstrating that properly constrained probabilistic supervision leads to better alignment.

(4) A+GFI+loss+RMO. Introducing the RMO block allows the network to iteratively solve a local non-linear optimization problem on the feature manifold, computing flow increments dynamically rather than guessing them through a black-box layer. This module particularly benefits fine-grained matching. On the RoadScene dataset, the strict sub-pixel accuracy (CMR:0.3px) improves from 12.6% to 26.9%, showing the effectiveness of the covariance-adaptive damping mechanism in local detail refinement.

(5) A+GFI+loss+RMO+UAPU. Finally, integrating the UAPU block forms our complete RBE-Flow framework. This closed-loop uncertainty feedback yields the best performance across all datasets. The improvement is especially notable at the strictest 0.3px threshold, increasing from 3.1% to 11.6% and from 11.9% to 24.0% on the OSdataset and WHU-OPT-SAR respectively.

Summary. The ablations validate a clear coarse-to-fine strategy: while a robust global flow initialization from GFI is essential for resolving large displacements, it is the synergistic RMO-UAPU loop that enables superior sub-pixel precision by empowering the system to reason about its own posterior uncertainty, leading to the best overall performance across all datasets.

Efficiency Analysis. RBE-Flow is trained end-to-end using a single NVIDIA RTX 4090 GPU with 14.3M parameters. During inference, RBE-Flow achieves an average runtime of 1.48 ms per image pair. Each iteration adds a cost of 0.13 ms. Specifically, during the training phase (batch size of 8, 7 recurrent iterations), our full probabilistic framework (RMO+UAPU) adds a modest temporal overhead of only 29.42 ms per step (+12.9%) compared to the deterministic RMO-only baseline. Crucially, the memory overhead introduced by the UAPU block is virtually negligible (+0.002 GB), keeping the total peak GPU memory strictly bounded at 10.16 GB. This thoroughly validates that our recurrent Bayesian paradigm achieves high-precision multi-modal alignment without incurring prohibitive computational or memory overhead.

5 Conclusion

We present RBE-Flow, a closed-loop recurrent Bayesian estimation framework specifically designed for robust cross-modal registration. By reformulating flow estimation as a dynamic state estimation problem rather than a standard regression, our approach explicitly addresses the inherent ambiguities in cross-modal feature matching. The architecture is fundamentally driven by the tight coupling of localized manifold optimization, which acts as a source of flow observations, and a recurrent Bayesian update mechanism, which sequentially refines the state distribution through non-linear belief propagation. Experimental results on the OSdataset, WHU-OPT-SAR, and RoadScene datasets demonstrate that RBE-Flow consistently achieves state-of-the-art performance particularly under strict sub-pixel thresholds. Looking forward, the principles of recurrent Bayesian estimation and uncertainty-driven feedback introduced in RBE-Flow offer a path toward more reliable spatial alignment in high-stakes applications like long-term autonomous navigation and cross-modality medical diagnosis. The flexibility of our modular design ensures that RBE-Flow remains compatible with evolving large-scale vision models and advanced geometric reasoning techniques, facilitating more robust cross-modal perception in increasingly complex environments. In future work, we plan to extend this probabilistic manifold optimization strategy from 2D dense matching to 3D multi-modal scene reconstruction.

Acknowledgment. This work was supported by Hebei Natural Science Foundation: F2025501003, and the Fundamental Research Funds for the Central Universities: N2523014. Shijiazhuang-Northeastern University Science and Technology Cooperation Special Project: NEUS2025-01-003.

References

1. Amos, B., Kolter, J.Z.: Optnet: Differentiable optimization as a layer in neural networks. In: International conference on machine learning. pp. 136–145. PMLR (2017)
2. Chen, H., Luo, Z., Zhou, L., Tian, Y., Zhen, M., Fang, T., Mckinnon, D., Tsin, Y., Quan, L.: Aspanformer: Detector-free image matching with adaptive span transformer. In: European conference on computer vision. pp. 20–36. Springer (2022)
3. Chen, J., Liu, Y., Wei, S., Bian, Z., Subramanian, S., Carass, A., Prince, J.L., Du, Y.: A survey on deep learning in medical image registration: New technologies, uncertainty, evaluation metrics, and beyond. *Medical Image Analysis* **100**, 103385 (2025)
4. Dai, K., Wang, K., Xie, T., Sun, T., Zhang, J., Kong, Q., Jiang, Z., Li, R., Zhao, L., Omar, M.: Dsap: Dynamic sparse attention perception matcher for accurate local feature matching. *IEEE Transactions on Instrumentation and Measurement* **73**, 1–16 (2024)
5. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops. pp. 224–236 (2018). <https://doi.org/10.1109/CVPRW.2018.00060>
6. Edstedt, J., Mateus, A., Jaenal, A.: Colabsfm: Collaborative structure-from-motion by point cloud registration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6573–6583 (2025)
7. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19790–19800 (2024)
8. Gast, J., Roth, S.: Lightweight probabilistic deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3369–3378 (2018)
9. Gehrig, D., Scaramuzza, D.: Low-latency automotive vision with event cameras. *Nature* **629**(8014), 1034–1040 (2024)
10. Giang, K.T., Song, S., Jo, S.: Topicfm: Robust and interpretable topic-assisted feature matching. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2447–2455 (2023)
11. He, X., Yu, H., Peng, S., Tan, D., Shen, Z., Bao, H., Zhou, X.: Matchanything: Universal cross-modality image matching with large-scale pre-training. arXiv preprint arXiv:2501.07556 (2025)
12. Hou, Z., Liu, Y., Zhang, L.: Pos-gift: A geometric and intensity-invariant feature transformation for multimodal images. *Information Fusion* **102**, 102027 (2024)
13. Hu, D., Peng, L., Chu, T., Zhang, X., Mao, Y., Bondell, H., Gong, M.: Uncertainty quantification in depth estimation via constrained ordinal regression. In: European Conference on Computer Vision. pp. 237–256. Springer (2022)
14. Huang, Z., Shi, X., Zhang, C., Wang, Q., Cheung, K.C., Qin, H., Dai, J., Li, H.: Flowformer: A transformer architecture for optical flow. In: European conference on computer vision. pp. 668–685. Springer (2022)
15. Ji, S., Wei, S., Lu, M.: Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience and remote sensing* **57**(1), 574–586 (2018)
16. Jiang, H., Karpur, A., Cao, B., Huang, Q., Araujo, A.: Omniglue: Generalizable feature matching with foundation model guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19865–19875 (2024)

17. Jiang, H., Dang, Z., Wei, Z., Xie, J., Yang, J., Salzmann, M.: Robust outlier rejection for 3d registration with variational bayes. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1148–1157 (June 2023)
18. Jiang, S., Campbell, D., Lu, Y., Li, H., Hartley, R.: Learning to estimate hidden motions with global motion aggregation. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 9772–9781 (2021)
19. Jiang, X., Ma, J., Xiao, G., Shao, Z., Guo, X.: A review of multimodal image matching: Methods and applications. *Information Fusion* **73**, 22–71 (2021)
20. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: *European Conference on Computer Vision*. pp. 71–91. Springer (2024)
21. Li, J., Hu, Q., Ai, M.: Rift: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing* **29**, 3296–3310 (2019)
22. Li, J., Hu, Q., Ai, M.: LNIFT: Locally normalized image for rotation invariant multimodal feature matching. *IEEE Transactions on Geoscience and Remote Sensing* **60**, 1–14 (2022)
23. Li, J., Shi, P., Hu, Q., Zhang, Y.: Rift2: Speeding-up rift with a new rotation-invariance technique. *arXiv* (2023)
24. Li, W., Chen, Q., Gu, G., Sui, X.: Object matching of visible–infrared image based on attention mechanism and feature fusion. *Pattern Recognition* **158**, 110972 (2025)
25. Li, X., Yang, W., Deng, J., Cheng, Z., Zhou, X., Zhang, T.: Implicit correspondence learning for image-to-point cloud registration. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 16922–16931 (2025)
26. Li, Y., Li, X., Li, W., Hou, Q., Liu, L., Cheng, M.M., Yang, J.: Sardet-100k: Towards open-source benchmark and toolkit for large-scale sar object detection. *Advances in Neural Information Processing Systems* **37**, 128430–128461 (2024)
27. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 17627–17638 (2023)
28. Liu, K., Yu, Y.: Revisiting the domain gap issue in non-cooperative spacecraft pose tracking. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6864–6873 (2024)
29. Liu, X., Ding, M., Sun, Z., Li, Z., Teng, X.: Crft: Consistent-recurrent feature flow transformer for cross-modal image registration. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 34784–34794 (2026)
30. Liu, X., Kou, Y., Chen, L., Teng, X., Li, Z., Luo, J.: Multi-orientation multi-scale aggregated registration with cyclic back-projection for multimodal images. *Image and Vision Computing* p. 106089 (2026)
31. Liu, X., Sun, Z., Ding, H., Song, X., Zhang, S., Sun, Y.: Gaff: Global attention feature flow network for optical and sar image registration under geometric transformations. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2026)
32. Liu, X., Teng, X., Bian, Y., Li, Z., Yu, Q.: Shape-adaptive modality independent region descriptor for multimodal remote sensing image matching. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **17**, 18139–18155 (2024)
33. Liu, X., Teng, X., Li, Z., Yu, Q., Bian, Y.: A fast algorithm for high accuracy airborne sar geolocation based on local linear approximation. *IEEE Transactions on Instrumentation and Measurement* **71**, 1–12 (2022)

34. Liu, X., Teng, X., Luo, J., Li, Z., Yu, Q., Bian, Y.: Robust multi-sensor image matching based on normalized self-similarity region descriptor. *Chinese Journal of Aeronautics* **37**(1), 271–286 (2024)
35. Mao, S., Lu, S., Du, Z., Jiao, L., Gou, S., Mou, L., Lu, X., Xiong, L., Zhang, Y.: Cross-rejective open-set sar image registration. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 23027–23036 (2025)
36. Morimitsu, H., Zhu, X., Cesar, R.M., Ji, X., Yin, X.C.: Dpflow: Adaptive optical flow estimation with a dual-pyramid framework. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 17810–17820 (June 2025)
37. Potje, G., Cadar, F., Araujo, A., Martins, R., Nascimento, E.R.: Xfeat: Accelerated features for lightweight image matching. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2682–2691 (2024)
38. Qin, H., Xu, T., Li, T., Chen, Z., Feng, T., Li, J.: Must: The first dataset and unified framework for multispectral uav single object tracking. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 16882–16891 (2025)
39. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* pp. 4938–4947 (2020). <https://doi.org/10.1109/CVPR42600.2020.00499>
40. Schusterbauer, J., Gui, M., Fundel, F., Ommer, B.: Diff2flow: Training flow matching models via diffusion model alignment. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 28347–28357 (2025)
41. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* pp. 8922–8931 (2021). <https://doi.org/10.1109/CVPR46437.2021.00881>
42. Sun, Z., Zhi, S., Li, R., Xia, J., Liu, Y., Jiang, W.: Gdros: A geometry-guided dense registration framework for optical-sar images under large geometric transformations. *arXiv preprint arXiv:2511.00598* (2025)
43. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: *Proc. Eur. Conf. Comput. Vis.* (2020). https://doi.org/10.1007/978-3-030-58536-5_24
44. Teed, Z., Deng, J.: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems* **34**, 16558–16569 (2021)
45. Truong, P., Danelljan, M., Timofte, R.: Glu-net: Global-local universal network for dense flow and correspondences. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6258–6268 (2020)
46. Truong, P., Danelljan, M., Van Gool, L., Timofte, R.: Learning accurate dense correspondences and when to trust them. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 5714–5724 (2021)
47. Tuzcuoglu, Ö., Köksal, A., Sofu, B., Kalkan, S., Alatan, A.A.: Xoftr: Cross-modal feature matching transformer. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* pp. 4275–4286 (2024). <https://doi.org/10.1109/CVPRW63382.2024.00431>
48. Wu, S., Wang, Y., Liu, X., Yang, Y., Wang, R., Guo, G., Doermann, D., Zhang, B.: Dfm: Differentiable feature matching for anomaly detection. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 15224–15233 (2025)
49. Wu, Z., Zheng, J., Ren, X., Vasluianu, F.A., Ma, C., Paudel, D.P., Van Gool, L., Timofte, R.: Single-model and any-modality for video object tracking. In: *Proceed-*

- ings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19156–19166 (2024)
50. Xiang, Y., Tao, R., Wang, F., You, H., Han, B.: Automatic registration of optical and sar images via improved phase congruency model. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **13**, 5847–5861 (2020)
 51. Xiang, Y., Wang, F., You, H.: Os-sift: A robust sift-like algorithm for high-resolution optical-to-sar image registration in suburban areas. *IEEE Trans. Geosci. Remote Sens.* **56**(6), 3078–3090 (2018). <https://doi.org/10.1109/tgrs.2018.2790483>
 52. Xiao, Y., Zhang, C., Chen, Y., Jiang, B., Tang, J.: Adrnet: Affine and deformable registration networks for multimodal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **62**, 1–13 (2024)
 53. Xu, H., Ma, J., Le, Z., Jiang, J., Guo, X.: FusionDn: A unified densely connected network for image fusion. In: *proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence* (2020)
 54. Xu, H., Ma, J., Yuan, J., Le, Z., Liu, W.: 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)*. pp. 19679–19688 (2022)
 55. Xu, H., Yuan, J., Ma, J.: MURF: Mutually reinforcing multi-modal image registration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(10), 12148–12166 (2023)
 56. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Tao, D.: Gmflow: Learning optical flow via global matching. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 8121–8130 (2022)
 57. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Tao, D.: Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(11), 13961–13975 (2023)
 58. Yang, B., Chen, J., Ye, M.: Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 11069–11079 (2023)
 59. Ye, Y., Bruzzone, L., Shan, J., Bovolo, F., Zhu, Q.: Fast and robust matching for multimodal remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing* **57**(11), 9059–9070 (2019)
 60. Zeng, J., Gu, Z., Liu, W., Cai, L., Cheng, J.: Uncertainty aware interest point detection and description. In: *WACV*. pp. 2144–2153 (2025)
 61. Zhang, J., Xia, Z., Dong, M., Shen, S., Yue, L., Zheng, X.: Comatcher: Multi-view collaborative feature matching. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 21970–21980 (2025)
 62. Zhang, K., Deng, Y., Ma, J., Favaro, P.: Adapting dense matching for homography estimation with grid-based acceleration. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 6294–6303 (2025)
 63. Zhang, Y., Yao, Y., Zhang, W., Zhang, Y.: Histogram of the orientation of the weighted phase descriptor for multi-modal remote sensing image matching. *ISPRS Journal of Photogrammetry and Remote Sensing* **196**, 1–15 (2023)
 64. Zheng, C., Li, S., Wang, C., Zhang, B.: Msg: Robust multimodal remote sensing image matching using side window gaussian space. *IEEE Transactions on Geoscience and Remote Sensing* (2025)